

EMIS – A Multilingual Information System on European Media Law

1. Document Base	2
2. Multilinguality	2
3. Conceptual Structures	2
3.1. Systematic Structure	3
3.1.1. Content Structure	3
3.2. Thematic Structure	3
4. Text Retrieval	4
4.1. Free Text Retrieval	6
4.2. Boolean Search	7
4.2.1. Simple Mode	7
4.2.2. Expert Mode	8
4.3. Keyword Search	9
5. Options	9
5.1. Search space	9
5.2. Search scope	9
6. Result	10
6.1. Result lists	10
6.2. Display of a single document	11

The document contains a preliminary system description. All pages are also available as online assistance.

1. Document Base

In EMIS, 246 documents (norms) in the languages German (94), English (97) and French (55) are processed. The retrieval operates on 8499 single documents. Further 161 norms are available in the Systematic Structure.

Some of these documents exist in several languages, but only one version is included into the search (i.e. the version which corresponds to the working language, otherwise the English documents are preferenced). With the output of the documents, the user can also display the versions in the two other languages, if available.

2. Multilinguality

The EMIS System is available in three languages: German, English and French. This is related to the interface, the communication language in which the system communicates with the user, the display of the data as well as of the results.

The language, which the user has to be select at the start, the so-called *working language* determines which the documents are searched. Documents such as the regulations of the European Union are available in all three languages, are involved only in the search only in the version which corresponds to the working language. For other documents available in different languages the version in the working language is preferred to the English or German version.

Multilinguality in EMIS means not only that the interface is available in different languages but the system provides also multilingual search facilities. The user can extend the search for a term typed in in the working language to documents which are only available in the other languages covered by the system (cf. Options). To do this, the term is translated by means of a simple translation tool. This tool uses dictionaries where the entries are annotated with domain labels. i.e. for each entry the domain for which the translation is valuable is given. For EMIS the translation marked as relevant for the domain of 'media law' are preferred. If no domain specific translation is found all others translations are involved in the search. The translation of phrases is done word for word if no translation for the whole phrase is found. The first translation found is then used for the search in the foreign language documents.

3. Conceptual Structures

In order to offer the EMIS users an easy access to the domain of 'Media Law', the EMR has developed two conceptual structures:

- The *Systematic Structure* lists all relevant norms existing in Europe, the regulations of the European Union, of the Council of Europe as well as of some international organisations like the UN and the WTO.
- The so-called *Thematic Structure* describes the domain by means of different topics, which consist in turn of different concepts. This structure is comparable to a classical thesaurus.

Both structures enable the user a directed access, where only the given information is available.

Content of the data base

The function `content of the data base` lists all norms currently available in the data base, sorted by countries. After selecting a norm, the complete text is displayed.

3.1. Systematic Structure

The Systematic Structure should provide an overview about the jurisdiction in the domain of new media in Europe. The norms in this structure are arranged according to countries and to the extend (federal law vs law of a land) and different subject areas such as *broadcasting, telecommunication, copyright or film law*.

Because not all existing norms and regulations are currently available or will be available at all in the EMIS system, a colored representation is used to indicate the state of availability of a norm:

Green means that these documents are available in the system.

Red means that the appropriate documents are available at the EMR, but not yet available on-line because there is no version in one of the system languages, or they have a minor importance in the hierarchy and will be therefore not put into the system. But these documents can be requested at the EMR (email emr@emr-sb.de)

Blue means that these documents exist but they are not available either on-line or on paper. They could be requested via the EMR Media Network.

Black marks headings and other text.

For each norm also the text version in the other languages are indicated; the text will be displayed by clicking on a language sign (EN,DE,FR).

The left part (frame) of the window facilitates the navigation within the large systematic structure. By selecting a particular country, the corresponding norms are automatically displayed in the right frame.

Selecting a norm

After selecting a reference (green title), a small window appears, in which the user can select one of the following functionalities:

- output of the information which is held for this norm in the database,
- output of a so-called *content structure* which allows to read a norm paragraph by paragraph, or
- output of the complete text of the norm, if exists in the current working language.

Search possibilities

A search within the systematic structure is only possible in the titles of the listed norms by means of the browser function `Edit/Find in page`. This kind of string search can also be carried out to search the text of a norm.

3.1.1. Content Structure

This structure corresponds to a table of contents. Here, the user has the possibility to read a norm section by section.

Depending on the type of the subdivision of a norm, the user can read individual paragraphs or sections in the right window after selecting a subdivision displayed in the left frame.

3.2. Thematic Structure

The Thematic Structure is one of the structures the EMR has developed to describe the domain of 'Media Law'. Different topics (see left frame) are described by means of a number of relevant concepts organised in a hierarchy (right frame). To each concept, the

EMR has assigned a set of relevant documents, which are displayed according to countries and norms after the user has selected a particular concept.

Only one concept can be selected at time!

Additionally to the list of documents available on-line, another list of further documents which are not yet in the system's data base is displayed, if available. The colors used in the representation of this list have the same meaning as in the Systematic Structure. At present, such a list is only implemented for one example Events of major importance to society under the topic Broadcasting.

Search possibilities

A search in Thematic Structure is only possible by means of the browser function Edit/Find in Page which carries out a simple string search on the concept names.

4. Text Retrieval

The so-called term search or full text retrieval as realised in the EMIS system is not based on a string search, which is used in most retrieval systems, but the implemented algorithm is enriched with linguistic intelligence. This has the advantage of more precise results due to the algorithm does not search for strings as they appear in the query or the text but for normalised terms. Normalisation means that the inflections of plural forms or verb conjugations are removed. This is also the reason why the user has not to use wildcards, as for example *transm** to find all occurrences which are related to *transmission*. By this linguistically based processing, wrong hits, which can never be excluded using wildcards as for example *transmute*, are mostly avoided.

The linguistic information used in EMIS is determined by means of a morpho-syntactic analysis of the documents as well as the queried terms. The analysis for *transmission*, for instance, is:

```
{ori=transmission,wnra=1,wnrr=1,snr=1,pctr=no,last=yes,pctl=no,gra=small,
source=wf,lu=transmission,s=ation,ds=transmit~ion,ls=transmit,c=noun,
case=nom;acc,nb=sg,ehead={case=nom;acc,nb=sg}}
```

whereas only the following features are of interest here:

ori corresponds to the input term,
lu indicates the lexical basic form,
ls the morphologic derivation

The values of *lu* and *ls* are used to produce the **search patterns**. For the example above these are *transmission* and *transmit*. The patterns are then looked up in the corresponding index. There are different indices depending on which feature is used as key. For English and French, two indices are constructed from the results of the analyses of the documents using the features *lu* and *ls*. To find all possible hits, a lookup of all patterns in all indices is carried out, and a look up with the value of the *lu* feature in the *lu*-index to find the exact hits only. The pattern must accurately correspond to a key. The algorithm finds beside the exact occurrences of *transmission* also the following hits:

```
transmissions (lu value/lu-index)
transmitting time (ls value/ls-index),
transmits (ls value/ls-index),
transmitted (ls value/ls-index),
```

Because of the special German compound construction, a third feature is used for the determination of the search patterns and for the construction of the indices, the *t* feature.

The following shows the analysis of the German compound *Werbesendung* (advertising broadcast):

```
{c=noun,lu=werbesendung,s=ation,t=werben#sendung,cs=v#n,ts=werbe#sendung,
ds=werben#senden~ung,ls=werben#senden,ss=v#ation,lng=germ,lngs=germ#germ,
ori=Werbesendung,snr=1,pctr=no,pctl=no,last=yes,wnra=1,wnrr=1,gra=cap,nb=sg,g
=f, ehead={nb=sg,g=f},w=2}
```

whereas the *t* feature marks the parts of a compound. For compounds, the *ls*-features also marks the derivations of all parts of a compound. All these values are used as search patterns, for the example above this means the algorithm used the following patterns: *werbesendung*, *werben#sendung*, *werben#senden*, *werben/sendung/senden* to find the following occurrences:

```
Werbesendungen (lu/lu-index)
Fernsehwerbesendung (t/t-Index)
Werbung in Sendungen (parts of t or ls/lu-index)
Werbezeit während Kindersendungen (parts of t or ls/ls-index and t-index)
```

whereas the parts have to occur in the same sentence, which means they have the same value for *snr*.

The formation of compounds in English and French does not take place via concatenation as in German, but English or French compounds usually consist of several words. In order to detect them, a kind of thumb rule is applied: all the words of a compound have to occur within a certain environment which is computed as follows:

$$(n-1) * 3$$

where *n* is the number of words of the compound. For example, for *television advertising* or *publicité télévisée* the exact hits are:

```
...concerning radio and television advertising...
...la commission de la publicité radiophonique et télévisée...
```

and these two hits not:

```
On television the beginning and end of a block of advertisements ...
...la publicité dans les programmes sonores et de télévision...
```

To compute the environment for each word the value of the *wnrr* feature is also part of the index.

The linguistic analysis is also used to carried out a multilingual search. The linguistically analysed inputs are translated using a simple translation tool which uses the information of the analysis to determine the correct translation. In general, if there are domain specific translations found in the lexicon, these are preferred. All translations found are equally analysed and searched in the respective foreign language documents as described above.

The result of a search is sorted according to the used search patterns: Thus, the search in English and French documents results in at most three different lists and the search in German documents in up to five result lists. The most relevant hits are in any case those where a pattern extracted from a *lu*-feature matches a key in the *lu*-index. (cf. Result)

The retrieval based on linguistic information is more complex compared to a string search, but more precise. Due to this complexity, the runtime for a search is higher due to the number of sort operations to be carried out, especially when all hits are looked up. A good search strategy is therefore firstly to look up only the exact hits and then to search for all hits in the documents of one or several countries.

4.1. Free Text Retrieval

The free text retrieval as realised in EMIS is based on linguistic information, therefore the user can type in complete words (no truncations are necessary), and all occurrences are found (cf. Text Retrieval).

The following input types are permitted: *simple words*, *abbreviations* and so-called *multiword units (phrases)*:

- Verbs: *transmit*
- Adjectives: *issued*
- Nouns: *Advertisement*
- Abbreviations: *BBC*

In addition, whole phrases (including compounds) can be input such as

- *public broadcast*
- *terms and conditions*
- *allocation of satellite channels*

All the parts of such a phrase must occur in the same sentence in order to represent a hit. One should note that all parts of a phrase (function words are removed) are linguistically analysed and for each a search is carried out, i.e. the duration of the search is directly proportional to the length of the phrase.

A search for words as for example (function words):

- *those*
- *yes*
- *not*

is not permitted and will be terminated by an error message.

Remark

Umlauts, french accents and other special characters can be input for instance as *ä*, *ae* or in HTML format as *ä*.

Abbreviations should be written in capital letters. For German input, the spelling rules should be observed (i.e. the first letter of a noun has to be capitalized).

The user should input the singular form of the term, i.e. *child* and not *children*. This is highly relevant in order to find the translation of a phrase, for instance *message publicitaire interdit* and not *messages publicitaires interdits!*

Multilingual Search

For a multilingual search, the queried term is translated whereby the domain specific translations are preferred. Then a search is done for all translations and the results of their linguistic analysis, if the corresponding option is selected.

For the multilingual search of a phrase, due to complexity only the first translation found is looked up. Here also, the domain specific translation is preferred. Additionally, if there are two German translations, then the compound is used for the search, i.e. the translations of *human dignity* are *Menschenwürde* and *Würde des Menschen*.

If no translation for the whole phrase is found, a word for word translation based on a shallow parsing is carried out (function words are removed).

Runtime

The runtime of search depends on the following factors:

- Type of input (simple term vs. phrase)
- Search scope (all or exact hits only)
- Number of translations found
- Number of hits (for each hit, a database access has to be carried out).

Average Runtime

Input	Search Space	Search Scope	Number of Translations DE/FR	Time (sec)	Number of Hits
<i>Broadcasting</i>	United Kingdom	all		1	23
	Luxemburg	all	2	3	121
	English Texts	exact		30	997
	English Texts	all		70	2380
	All Texts	exact	1/2	50	1746
	All Texts	all		150	3942
<i>Youth protection</i>	All Texts	exact	1/3	15	60
	All Texts	all		20	86
<i>Digital television</i>	All Texts	exact	1/1	20	9
	All Texts	all		50	38

Maximal Runtime

Input	Search Space	Search scope	Number of Translations DE/FR	Duration (sec)	Number of hits
<i>Broadcasting programme</i>	English Texts	all		90	409
	All texts	exact	1/1	80	337
	All texts	all	1/1	240	908
<i>Broadcasting transmission</i>	All texts	exact	1/1	20	50
	All texts	all	1/1	60	352

The search for the category *Further eventually interesting documents* (cf. Result) in German documents is limited to 2000 comparisons. If there are more hits, a link is generated which delivers the missing hits after selecting.

The runtime is measured locally on a SUN SPARC ULTRA2 (512MB RAM, 300MHz). On-line access additionally depends on the speed of the connection.

4.2. Boolean Search

This function provides two different input modes, the *Simple Mode* and the so-called *Expert Mode*.

4.2.1. Simple Mode

This mode allows the combination of four simple terms by means of the indicated operators AND, OR and AND NOT.

As input, only simple words and abbreviations, and for German also compounds (not recommended!), are allowed. The input of phrases (ex.: *subliminal techniques, independant broadcasting corporation*) is not allowed and will be explicitly terminated by an error message.

For each input term a text retrieval is carried out (cf. Text Retrieval) and the respective results are combined according to the operators and according to the rule that AND more strongly binds than OR. The following examples should clarify this rule:

Advertisement AND radio OR television

Results in a list of documents in which *advertisement* and *radio* occur and all texts which contain the term *television*;

Advertisement OR times ANDNOT radio

Results in a list of documents in which *advertisement* occurs and all documents, which contain *time* but not *radio*;

Financing AND NOT advertisement AND consumer AND protection

Results in a list of documents, in which *financing* and *consumer protection* occur but not *advertisement*.

Within the boolean search, in contrast to free text retrieval, the terms have to occur in the entire document and not necessarily in the same sentence.

4.2.2. Expert Mode

This mode allows the input of boolean expressions, completely aggregated by parentheses whereas **AND** has to be used for *and*, **OR** for *or* and **ANDNOT** for *and not*.

The following example shows the evaluation of such an expression:

(Advertisement AND (Radio OR Television) AND NOT Broadcasting)

Outputs all documents in which *advertisement* occurs, and either *radio* and/or *television* but not the term *broadcasting*.

Multilingual Search

Each input term will be translated, if a multilingual search space is selected, and for each translation found a search (a free text retrieval) is carried out.

Runtime

The runtime of a boolean search depends on the following factors:

- Number of terms being combined
- Type of input (relevant only for German input)
- Type of operator (AND requires complex sort operations)
- Search scope (all or exact only)
- Number of translations found
- Number of hits.

The runtime depends on the runtime of the text retrieval for each search term as well as on the number of sort operations to be carried out to compute the boolean expression (max 76). Therefore a good search strategy is to search first only for exact hits and afterwards for all hits for a certain number of countries. Due to time constraints, for German input, the computation of the result list *Possibly interesting documents* is interrupted after 1500 comparisons (relevant for inputs like *Broadcasting programme*).

Simple Mode

Average Runtime:

Input	Search Space	Search Scope	Number of Translations DE-FR	Time	Number of Hits
<i>Television AND Youth</i>	exact	all texts	1/1 - 471	40 sec	24
	all	all texts		140 sec	47

Maximal Runtime

Broadcast AND Television AND Advertising

exact	all texts	1/1/2 – 2/4/1	80 sec	82
all	all texts		6 min	230

Expert ModeAverage Runtime:*(Television AND (Youth OR Child))*

exact	all texts	1/1/1 – 4/1/1	40 sec	111
all	all texts		200 sec	241

Maximal Runtime*(Cable AND Transmission AND (Television OR Broadcasting))*

exact	all texts	1/1/1/1 - 2/2/4/2	100 sec	77
all	all texts		8 min	143

The runtime is measured locally on a SUN SPARC ULTRA2 (512MB RAM, 300MHz). On-line access additionally depends on the speed of the connection.

4.3. Keyword Search

To each document the EMR has manually assigned keywords. The function `Complete keyword list` displays a list of all key words assigned. For organisational reasons this whole list is subdivided into partial lists organised in alphabetic order.

In order to search a keyword, first a sublist must be determined by selecting a particular letter. Then the corresponding sublist is displayed, in which *one* particular keyword can then be selected. For this keyword all assigned documents are displayed according to the selected search space.

Because this search which is not based on linguistic technologies is carried out by just looking up the keyword index, the result is displayed at once.

5. Options

The result of the text retrieval functionalities can be determined by the two following options:

5.1. Search space

This option limits the search to certain languages or countries. The search can be carried out either on the set of German, English and/or French documents or (exklusiv!) on the norms belonging to one or several countries. As many countries as desired can be selected. A combination of both options is not allowed!

Note

The documents of the European Union are involved in the search only in the selected working language (i.e. language of the interface). I.e. with a German interface, only the German-language regulations of the European Union are part of the search space. Performing a multilingual search the translated terms are in no case looked up in the corresponding English or French documents of the European Union!

5.2. Search scope

This option determined to what extent the linguistic information is used in a search.

`All hits` means that all patterns resulting from the linguistic analysis (see example) are looked up. In order to make this transparent for the user, the patterns used are displayed in the result window of a search.

`Exact Hits only` means that only the input term is looked up. Since the search is based on linguistic information, occurrences of the term in an inflected form (e.g. plural) are also found.

Example

To search for *broadcasting* the following search patterns are constructed: *broadcasting*, *broadcast*. A search for exact hits looks up only *broadcasting*; the search for all hits looked for all patterns indicated above.

For German search terms, especially compounds, the set of search patterns is extended by the parts of the compound. For instance: The input *Rundfunksendung* results in the following set of patterns: *rundfunksendung*, *rundfunk#sendung*, *rundfunk#senden*, *rundfunk/sendung/senden* whereas for the last, *rundfunk* and *sendung* or *senden* have to occur in the same sentence.

6. Result

As result of a term search a list of relevant documents is output, organised according to the selected search space option, to languages and/or countries as well as to norms. For each language, the respective search patterns are also indicated.

6.1. Result lists

consist according to the selected search scope (only accurate one or all hits) of several sublists, which are organised descendingly according to relevance of the search pattern. The relevance of the results is determined by the pattern used for the look up. As a result of the different linguistic analyses for German, English or French inputs and documents, a different number of possible result sublists are retrieved:

- a. The search in English and French documents can lead to the following result lists:
 1. List of documents, which contain exact hits; for compounds, this means that all parts must occur in a certain environment (distance 3).
Access: lu-value in lu-index
Example: *broadcasting*: *broadcasting*
television advertising: *advertising on television*
 2. List of documents, containing terms with the same derivation as the search term.
Access: ls-value in ls-index
Example: *broadcasting*: *broadcast*and/or for multiword units
 3. List of documents, in which the parts of the multiword units occur outside of the environment.
Access: lu-value in lu-index and/or ls-value in ls-index
Example: *television advertising*:...
...advertising on its own radio and television... .
...advertisements on television

- b. Search in German documents can result in the following sublists:

For all input types, the following result categories can be obtained:

1. List of documents which contain the search word exactly; this corresponds to the result of a search for exact hits.
Access: lu-value in lu-index
Example: *Werbung*: *Werbung*;
2. List of documents which contain the search word as part of a bigger compound.
Access: t-value in t-index
Example: *Werbung*: *Zigarettenwerbung*;

3. List of documents containing terms with the same derivation as the search word.
Access: t-value in t-index
Example: *Werbung; werben;*

For compounds the following additional categories can be obtained:

4. List of documents which contain the parts of a compound within the same sentence.
Access: t/lS-value in lu-index
Example: *Fernsehwerbung: Werbung im Fernsehen...;*
5. List of documents which contain terms, which contain in turn parts of the search words as part.
Access: t/lS-values in t/lS-index
Example: *Werbung: Werbezeiten in Fernsehsendungen ...;*

Within a phrase search as well as within a boolean search, the result lists of the single term searches are combined and compared crosswise, in order to obtain all hits. The result categories contain the following hits:

1. List of documents, containing only exact hits of all single searches
Example: *gemeinsames Programm: gemeinsame Programme;*
2. List of documents with hits of the 2nd category above for one or several of the search terms as well as hits of exact hits for the other terms
Example: *gemeinsames Programm: gemeinsame Programmteile;*
3. List of documents with hits of the 3rd category above for one or several of the search words as well as hits from first and the 2nd category for the other terms.
Example: *gemeinsames Programm: Gemeinschaft der Programmanbieter;*

For compounds only:

4. List of documents with hit of the 4. category above as well as hits of the categories 1 - 3.
Example: *gemeinsames Programmangebot: gemeinsames Angebot von Programmen;*
5. All other hits, not qualified for any of the former lists.
Example: *gemeinsames Programmangebot: ... anbieten von gemeinschaftlichen Programminhalten...*

6.2. Display of a single document

By selecting a certain document of the result list, the text, the country as well as the title of the current norm are displayed whereas the title provides a link into the Systematic Structure.

Additionally at the end of the text, all information stored in the data base to this document is displayed: footnotes, the date of enactment, the date of the entry into force, references to other relevant norms as well as to transpositions, a reference to DEMIS and to a list of relevant literature. If available, a reference to the English or French version of the text is also provided (EN,FR,DE).

Note

There are still some norms specified in the list of relevant norms and transpositions which are not yet available in the data base. This will be explicitly mentioned.