

# Find information in large collections of multilingual text

## Text Mining to facilitate information access

### Beat the information overflow

In the electronic age, an enormous amount of information is available about almost any subject domain, but the task of finding and processing this information is getting ever more difficult. Automatic text analysis tools can help **find** potentially user-relevant documents quickly, **extract** information from them, **display** the results in an organised manner, and allow quick **access** to the most interesting text passages.



For instance, all the documents found by a search engine can be downloaded and clustered into groups of related documents. For each of the clusters, references to people, organisations, places and dates, as well as occurrences of user-defined specialist terms, can be extracted and listed. Automatically generated hyperlinks can help the users access the text passages where the names and terms were found. Users can then investigate the whole document collection efficiently and in an organised manner, focusing on those groups of documents where the most interesting names and terms were found.

**Bulgarian:** Членовете на Европейския парламент се избират чрез пъ регионална основа, като например в [Италия](#)[Italy], [Великобритания](#) национална основа, като във [Франция](#)[France], [Испания](#)[Spain], [Авс:](#) [Люксембург](#)[Luxembourg] и други, или при смесена система ([Германия](#))

**Czech:** Poslanci Evropského parlamentu jsou voleni na základě všeobecné zastoupení, a to buď na základě regionálním, jako například v [Itálii](#)[Italy], [Belgium](#), nebo národním, jako ve [Francii](#)[France], [Španělsku](#)[Spain]ku, [Ra Lucembursku](#)[Luxembourg] a dalších zemích, nebo na základě smíšeného

**Estonian:** Euroopa Parlamendi liikmed valitakse otseselt ja üldiste valimis kas süis regionaalset alusel, nagu näiteks [Itaalia](#)[Italy]s, [Ühendkuningriiki](#)

### Overcome the language barrier

The most interesting information is often written in foreign languages. Machine Translation tools have limitations and are not available for many language pairs. A viable solution to this information access bottleneck are customisable tools that display some user-relevant information via the **cross-lingual glossing** of specialist terms, place names, etc. This helps to identify the most interesting documents so that users can focus their effort on these.

## Daily news analysis in many languages

### Grouping news by subject

The JRC's *Europe Media Monitor* system (<http://emm.jrc.org>) monitors a daily average of 25,000 news articles in 30 different languages live and around the clock. As hundreds of news articles often report about the same event, JRC software organises these articles automatically into groups.



**Keywords:** Iraq, United States / Al Qaeda, Saddam Hussein / Baghdad, Sunni, Shiite, Iraqi, forces, police, violence, sectarian.  
**Importance:** 8654 articles in 466 clusters.

### Linking news reports over time and across languages

JRC's software (<http://press.jrc.it/NewsExplorer>) also identifies whether articles belong to an ongoing story and displays timelines for the biggest stories this year, month or week. A novel and unique function is the automatic linking of news across languages. The intuitive user interface publicly accessible at <http://press.jrc.it/NewsExplorer> thus allows analysts to browse the news collection, to see how stories developed over time, and to compare how the same event was reported in different countries.

## Extracting and displaying information

For each group of related articles, the software extracts and stores references to places, people and organisations, and it generates a geographical map to display the results. In a customisable version of the software, users can provide lists of terms that they are particularly interested in. If these terms are found in any of the clusters, they will be displayed next to the map and hyperlinks allow to jump directly to the text passages where these terms were mentioned.

## Detecting name variants

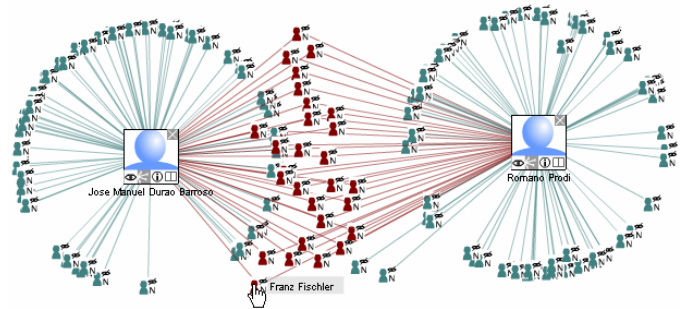
In the news, the same person is often referred to using different spelling variants, especially across languages. Approximate matching techniques help identify which spelling variants are likely to belong to the same individual. Additional name variants and photographs can often be retrieved automatically from free internet sources and added to the information about this person. When users search the news collection for all articles mentioning a certain person, articles will thus be found even if the name is spelled differently. In two years of news analysis, information about over 500,000 persons was collected, with up to 150 variants for the same name.

## Learning relationships

As names, places and keywords are identified and stored for each news cluster, the system learns over time which of these entities is frequently associated with which others. A network of connections develops between people and people, people and countries, countries and keywords, and so on. The relationships change and are updated continuously, fed by thousands of news articles every day.

## Navigating news collections over time and across languages

Due to the automatically generated news meta-data and the links identified between news, persons, places and keywords, users can browse and explore the news collection according to their needs without being hindered by the language barrier: The same story reported in Spanish media, other names related to a given one, news related to a certain keyword, ... When the searched information is found, a hyperlink to the original news item allows the users to read the corresponding articles. Part of the system is available at <http://press.jrc.it/NewsExplorer>.



Iyad Allawi  
Iyad Alauī  
Ajad Allawi  
Ιγιάντ Αλλάουι  
Ijad Alawi  
Illyad Allaoui  
Iyad Alauī  
Ayat Allawi  
اياد علاوي  
Eyad Allawi  
Iyad al-Allawi  
伊亚德·阿拉维  
...

## Subject classification across languages

### Support to European Parliaments

The libraries of the European Parliament, and of many national and regional parliaments in Europe, categorise all of their texts (legal texts, debates, resolutions, parliamentary questions, etc.) according to the 6,000 subject domain classes of the Eurovoc thesaurus (<http://europa.eu.int/celex/eurovoc>). The thesaurus is available in over twenty languages so that the Eurovoc classes identified for a given text can be displayed and searched in all the other languages. The JRC has developed a statistical system that classifies new documents automatically or interactively. The Spanish Congress of Deputies in Madrid is the first to use it in their daily work.

### Cross-lingual keyword display

The JRC system can currently Eurovoc-classify texts in 14 different EU languages. The result of the analysis is a list of the most important classes for a given document. These lists serve as keywords that show users what a text is about.

The results of the class assignment can be displayed in the text language or in any of the other twenty languages. Users can thus get an idea of the contents of a document even if they do not understand the language of the text. This technology is also a major ingredient for finding related news articles across languages in the EMM NewsExplorer.

## CONTACTS

Ralf Steinberger

Tel.: +39-0332-786271

Fax: +39-0332-785154

e-mail: [Ralf.Steinberger@jrc.it](mailto:Ralf.Steinberger@jrc.it)

website: <http://www.jrc.it/langtech>

