

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 ц 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 (**Герма**

Czech: Poslanci Evropského parlamentu jsou voleni na základě všeobec zastoupení, a to buď na základě regionálním, jako například v **Itálii**[Italy], **Belgium**], nebo národním, jako ve **Francii**[France], **Španěls**[Spain]ku, **Ra Luce mbursku**[Luxembourg] a dalších zemích, nebo na základě smíšenéh

Estonian: Euroopa Parlamendi liikmed valitakse otseselt ja üldiste valim kas süis regionaalsel alusel, nagu näiteks **Itaalia**[Italy]s, **Ühendkuningrii**

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

Linking news reports across languages

The JRC software also links news clusters with the related news published in previous days, and even with related press articles written in other languages. An intuitive user interface (<http://press.jrc.it/NewsExplorer>) allows analysts to browse the news collection, to look up the major news for a given day in the past, and to compare how the same event was reported in different countries.

Italians march for release of journalist es fr it nl
Il Manifesto was vehemently opposed to the US-led war on Iraq and Sgrena had condemned the Italian deployment following the fall of Baghdad. In her emotional appeal, Sgrena called on Italy to withdraw its soldiers. Italy has some 3000 troops in Iraq, the fourth largest foreign contingent after US, British and South Korean forces.
aljazeera-en 19/02/2005 20:19

Militants kidnap two journalists
GuFDailyNews 19/02/2005 09:05

No claim of responsibility for mosque ...
Irish-news 19/02/2005 11:09


Huge Rome rally for Iraq hostage
euronews 19/02/2005 19:40

Italians march for release of journalist
aljazeera-en 19/02/2005 20:19

Rome Demonstrators March for Hostage
guardian 19/02/2005 16:52

People

- Giuliana Sgrena
- Florence Aubenas
- Silvio Berlusconi
- Romano Prodi
- Enzo Baldoni
- Susilo Bambang Yudhoyono
- Saddam Hussein
- Simona Torretta



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 one year of news analysis, information about 125,000 persons was collected, with up to 60 variants for the same name.



Iyad Allawi
Iyad Alawi
Ajad Allawi
Ιγιάντ Αλάουι
Ijad Alawi
Illyad Allaoui
Iyad Alauí
Ayat Allavi
ایاد علاوي
Eyad Allawi
Iyad al-Allawi
伊亚德·阿拉维

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, between 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 language 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>.

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*. The thesaurus (<http://europa.eu.int/celex/eurovoc>) 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.

Ψήφισμα σχετικά με τα ανθρώπινα δικαιώματα στην Αιθιοπία
 Το Ευρωπαϊκό Κοινοβούλιο,
 - έχοντας υπόψη το ψήφισμά του της 18ης Ιουλίου 1996 σχετικά με τα ανθρώπινα δικαιώματα στην Αιθιοπία ((EE C 261 της 9.9.1996, σελ. 166)).
 Α. έχοντας υπόψη ότι ο σεβασμός των ανθρωπίνων δικαιωμάτων, οι δημοκρατικές αρχές και το κράτος δικαίου αποτελούν ουσιώδη στοιχεία της αναθεωρημένης Ε' Σύμβασης του Λομέ και ότι το σύνταγμα της Αιθιοπίας επίσης περιλαμβάνει το σεβασμό των ανθρωπίνων δικαιωμάτων,
 Β. λαμβάνοντας υπόψη την εν εξελίξει διαδικασία

Desc ID	Descriptor text
04310305	political violence(10)
123602	human rights(10)
7221060704	Eritrea(9)
7221060702	Ethiopia(9)
043604010301	democratization(8)
723119	EC countries(7)
081602	peacekeeping(7)

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 News Explorer*.

Document Profile

Title: Seizure of plutonium at Munich airport (E-3083/95)

Author: Martin Schulz (PSE)

Text Language(s): English

Source: http://crnfn.com/digital/jam/wires/9903/13/plutonium_eu.html

Related Documents: 12 ([click here to view](#))

Retrieval Date: 03.05.1998

Creation Date: 27.03.1998

Text Length: 287 words

Keywords (Occurrence Frequency)

TUI (3), Commission (7), Karlsruhe (3), seizure (6), OJ (2), plutonium (3), suitcase (3), German (4), material (4)

Descriptors

plutonium, import, illicit trade, Federal Republic of Germany, EAEC Joint Research Centre, airport, fraud

Names

Organisations: Commission, European Institute for Transuranium Materials (TUI), Joint Research Centre, PSE

People: Martin Schulz, Mrs. Breyer, Mr. Papoutsis

Geographical References

Germany (11)
 German (4), Karlsruhe (3), Munich (2), Germany (1), Federal Republic of Germany (1)

Europe (1)
 European (1)

Dates

10/01/1996

17/08/1995

04/09/1995

10/08/1994

Customs Tariff (TARIC) Product Groups

plutonium (3) (2844205100010)

suitcase (3) (392321000010)

Document Summary

E-3083/95 by Martin Schulz (PSE) Seizure of plutonium at Munich airport
 In the summer of 1994 a suitcase containing plutonium illegally imported into Germany was seized in sensational circumstances at Munich airport in the Federal Republic of Germany. The Commission (Euratom safeguards directorate) was alerted by the German authorities in the early afternoon of 10 August, 1994, that some material might be seized.

Contact us for more information

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

Ralf.Steinberger@jrc.it

Language Technology

Support to External Security Unit
 Institute for the Protection and Security of the Citizen

Joint Research Centre / TP 267
 I - 21020 Ispra (VA), Italy
 Tel: +39 - 0332 786271
 Fax: +39 - 0332 785584



Language Technology

Multilingual Text Analysis



Find information
 in large collections
 of multilingual text

