EUROPE MEDIA MONITOR

TABLE OF CONTENTS:

1 INTRODUCTION

For those who are not familiar with it, EMM is the Europe Media Monitor, a system for monitoring open source news information. EMM is developed and maintained by the Data & Text Mining Unit, in the Competences Institute of the EC Joint Research Centre (JRC).
EMM was started in 2002 as a project to support the Commission with its media monitoring activities. The main purpose of EMM is to provide monitoring of a large (but selected) set of electronic media, reducing the information flow to manageable proportions by clustering related news, categorising articles and applying Language Technology tools to derive further meta-data, such as recognising and disambiguating entities in the text, extracting quotes by and about people, applying sentiment/tonality analysis and more.

The system continuously monitors over 7.000 HTML pages and RSS feeds to find new articles published on the Internet (~250.000 articles daily). It then reads and analyses these articles and extracts information, like references to people, organisations and places in the news, extracts quotes, groups articles into categories and clusters similar articles. This last process in effect creates a view of the current biggest stories in the news in a certain language.

Highlights:

- ***New map, allowing easy visualisation of what is going on where.***
- ***New mobile apps, Category Editor Collaboration Layer and,***
- ***Lots of 'behind the scenes' developments which are making our systems more reliable tools for daily media monitoring.***

We would be very pleased to receive your feedback, and would gladly provide you with further information.
Please contact us at
emm@jrc.ec.europa.eu

For more information about the JRC please visit the following link:
https://ec.europa.eu/jrc/

INFORMATION GATHERING

2

The Europe Media Monitor is designed as a near real-time monitoring system for new publications. The system analyses publications as they flow through and continuously generates the required information products, without storing a copy of the original publication. It does not rely on (and does not have) a big information archive. Although EMM maintains an index of all retrieved material, allowing for limited historical research, the information products always refer to the original publication, mostly on the Internet.

At the core of EMM there is a chain of lightweight extensible processes each running independently and chained together using robust and reliable in-house developed web service architecture. Articles begin their flow through the processing chain as thin RSS (Really Simple Syndication) items that grow as metadata gets added at each stage of the processing chain.

EMM has been expanded with social media monitoring functionality. Currently, we are extracting the most frequent URLs, hash tags and Twits that are related to the most recent disease outbreaks, violent events and disasters.

Data collection (scraper and grabber) Highlights:

- ***Article extraction from HTML feeds without the need for custom xslts***
- ***Website scraping for unstructured sites***
- ***Access to the Internet using configurable proxy servers and user agents***
- ***Improved handling of badly formatted RSS feeds***
- ***All RSS now have an 'origin' tag***
- ***EMM has been expanded with social media monitoring functionality.***

3 INFORMATION PRESENTATION

The results of the information harvesting and processing can be accessed in a number of ways: a **NewsBrief** website (e.g. http://emm.newsbrief.eu) that allows for classical data browsing, and a full editorial and publishing system **NewsDesk** (not publicly accessible) that allows for the creation and publication of high level information products. EMM delivers emails and RSS feeds and there are (free) mobile applications for iPhone, iPad and Android tablets. Examples of current applications of the EMM technology can be found in different application domains. EMM is used in a number of traditional media monitoring applications by various EU Institutions and Agencies. **MediSys** (http://medisys.newsbrief.eu) is an instance of EMM specifically developed for internet bio-surveillance and is used by a number of Health Agencies, including the WHO. Open source intelligence for humanitarian and conflict early warning is also covered by at least 3 instances of the EMM system. **MyNews** is a web interface designed for desktop browsers, for the news items supplied by the EMM engine. It's highly customisable, since it allows each user to define his/her own specific view by selecting the topics he/she's most interested in. This is achieved – similarly to the EMM mobile apps - by allowing him/her to tune news channels focused on very specific topics.

EMM comes to you in different views:

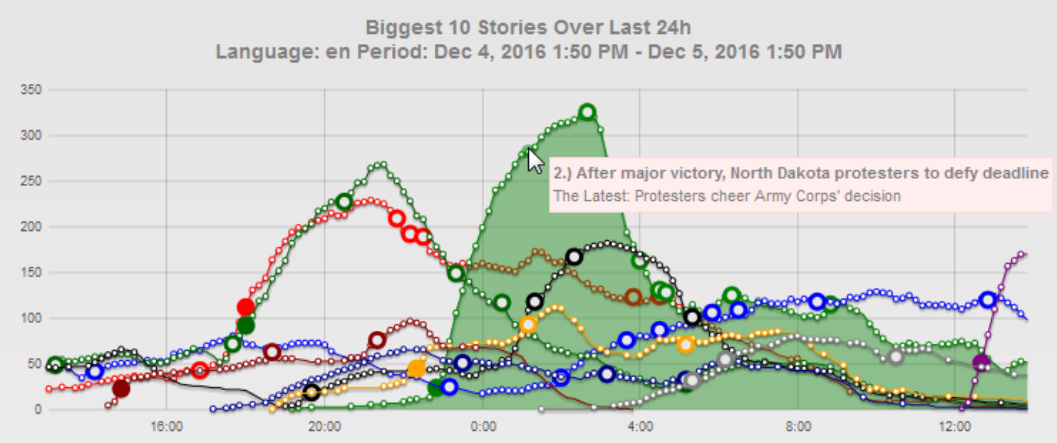
- ***NewsBrief***
- ***NewsDesk***
- ***MyNews***
- ***MediSys***
- ***EMM Mobile Apps***

Users can create as many channels as they like, and they can organise them in sets. There are many different ways they can create new channels, which increases greatly the flexibility of the tool.
At the moment, the publicly accessible instance of EMM monitors almost 20 000 RSS feeds/HTML pages from over 7000 media websites and retrieves and processes around 300 000 new news articles per day. These articles are categorized into over 2000 categories. A selected subset of these categories and the results of the clustering process can be seen on the public EMM website http://emm.newsbrief.eu.

INFORMATION PRESENTATION

NEWSBRIEF

4



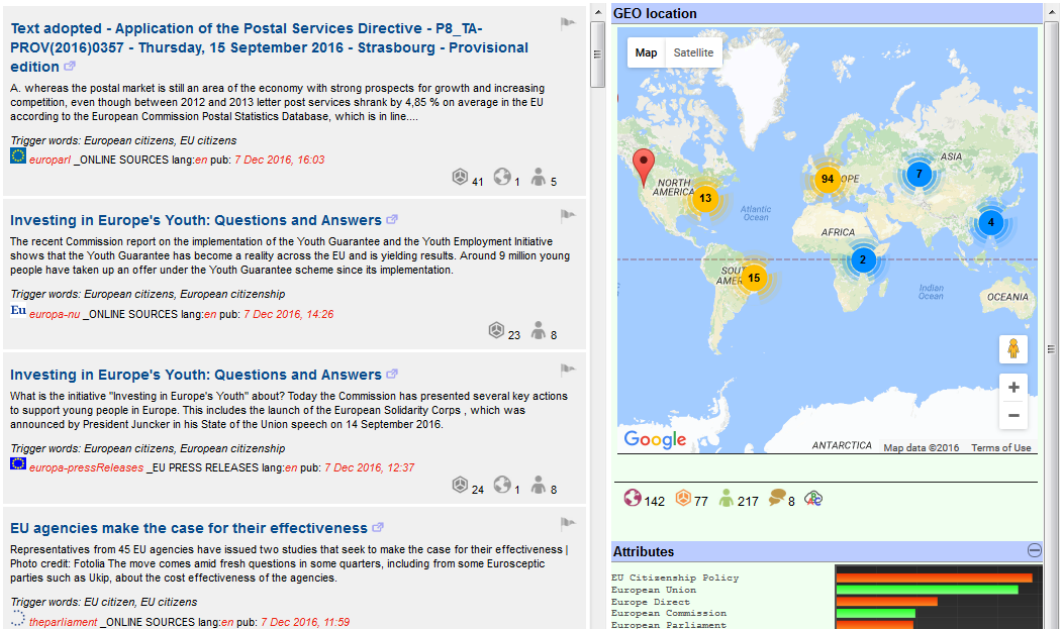
EMM NewsBrief is a public website that provides many different views on the news published right now.
The **NewsBrief** pages mostly reflect the categorisation and the topic-based clustering. The typical front page which is shown when you go to http://emm.newsbrief.eu is the result of the clustering system. Most pages accessible through the menu system reflect the result of the categorization process.
The categories are predefined and cannot be modified by the public. On every 'category page' you can click on the '?' icon to see how a particular category is defined. Most of these categories are defined by domain experts and are made available to you.
Some of the categories were defined by us, mostly when we were first developing the system and tried to generate some meaningful content.
On every page you can choose to receive the news by e-mail, or you can use the information through the RSS feed.

Feel free to include any RSS feed on your website, but please give us credit and include a link to the EMM website as a reference on your page.
You can also search EMM for information as we do maintain an index for all articles that we process. The results include a link to the original article, although in some cases it may no longer be available.

The primary aim of the system is to provide you with frequent and near real-time news updates on topics of interest to you.

INFORMATION PRESENTATION

5 MYNEWS



MyNews is a highly customisable web interface that gives access to the news items produced by the EMM engine.
It is user-driven, for its main focus is about offering you the possibility to create your own personal view by means of many different customisation options. It is based upon a "*TV metaphor*": the users, as if sitting in front of a TV with a remote control, can tune into different **channels** on the specific topics they are interested in.

There are many different types of channels users can choose from:

Category channels, associated with the EMM categories.

Country channels, associated with the source countries, or with the countries the news articles talk about.

Multilanguage top stories, associated with the EMM most active clusters in any given language.

Person or Organisation channels: associated with collections of several EMM categories and/or entities.

Search channels, associated with queries performed on text and metadata extracted from the news articles.

Channels are organised into **sets**, thus you can have many sets, each one with as many channels as you like. The structure of sets and channels is easily editable at any time, and recorded on the server for subsequent access(es).

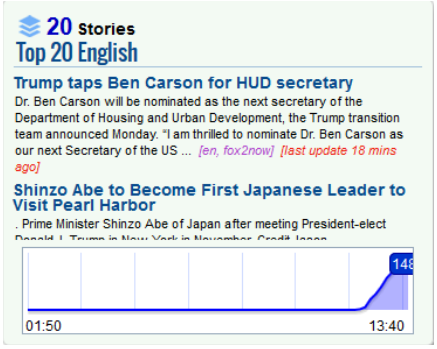
MYNEWS

When you get into a set, you see the “cover sheets” of its channels, represented by boxes. By clicking on one box you get into the details of the channel: the list of articles with the representation of all the associated metadata (categories, entities, geotags, quotes, etc.). The information is also enriched with several graphical tools: a map with the distribution of the articles, several charts, multi-language word clouds, etc.
Several refinement tools are provided: you can filter the articles based upon sources, countries, attributes (i.e.: categories and entities), languages and date/time range.

Top story channels display a list of stories, ordered by relevance – i.e. stories/clusters by topic – which are most active in that very moment or over the last 24 hours. For each of the selected languages, the twenty top stories are displayed.
Each story is listed with its main article and the representation of all the associated metadata.

Highlights:

- ***Highly and easily customizable on a per-user basis***
- ***Many different visual representations of data (charts, maps, word clouds, etc.)***
- ***Newsletters in HTML, PDF or MS-Word format, based on selection of articles.***
- ***Advanced Search channels actively catching new articles that satisfy user-defined queries***



MOBILE DEVICES

6

The EMM iPhone and Android mobile Apps provide up-to-the-minute results using Automatic Text Analysis of news articles from around the world (over 300 000 new news articles per day). Both the Apps and EMM desk system support more than 70 different languages. In-line translation to English from Arabic, Czech, Chinese, Danish, French, German, Italian, Polish, Portuguese and Swedish is supported too. The automatic story detection, groups the articles reported on the same subject, tracking the stories as they develop over time.

All apps support automatic detection of people & organisations and produce views of what was said by and about people or organisations.

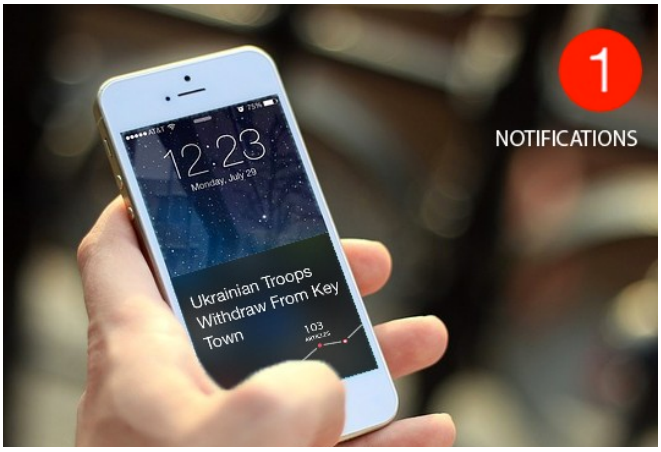
Highlights:

- ***Android phone version released***
- ***Customized version for CERT-EU released***
- ***Other customization for some clients will be released soon***

INFORMATION PRESENTATION

MOBILE DEVICES

Real-time Notifications



Real-time alerts allow custom notifications based on changes in the specific data set the user has defined. When a logical threshold is activated the system displays a notification directly on the user's mobile device.

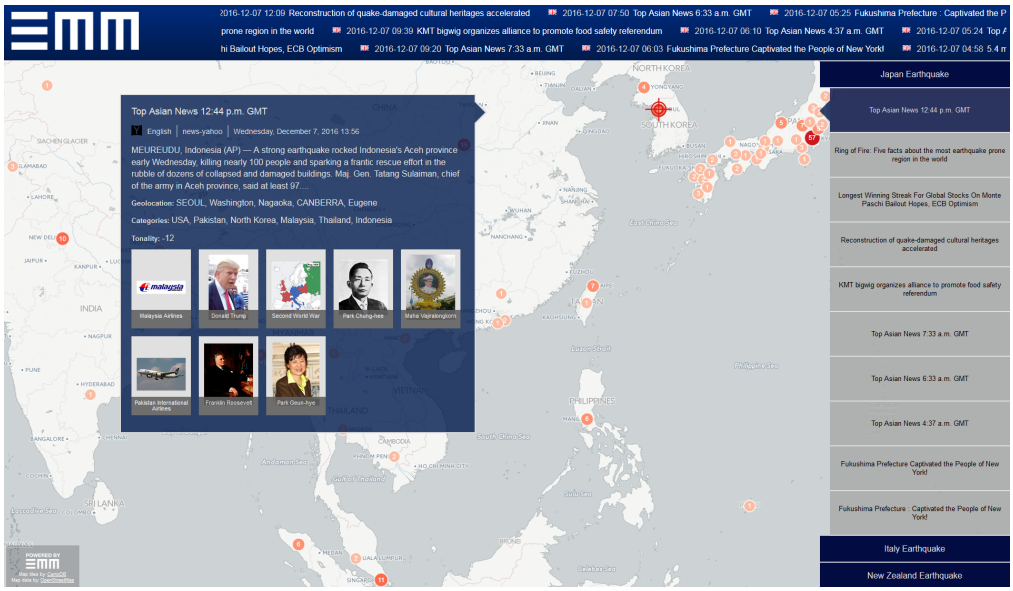
By merging our notifications with the system's core notification we alert the user only when it is appropriate. For example, notification will wait silently when the user is asleep and will schedule the notifications to be presented a few minutes after the user has started using the device. This is being done without any user intervention or pre-settings.

Supporting Android, iOS, PC, Linux & MacOS

INFORMATION PRESENTATION

7

BIG SCREEN MAP



The Big Screen Map is an application that automatically loops through the latest news from the EMM system. It is designed to run on large-format screens. The application is fully configurable providing the ability to select the languages and the categories to be displayed.

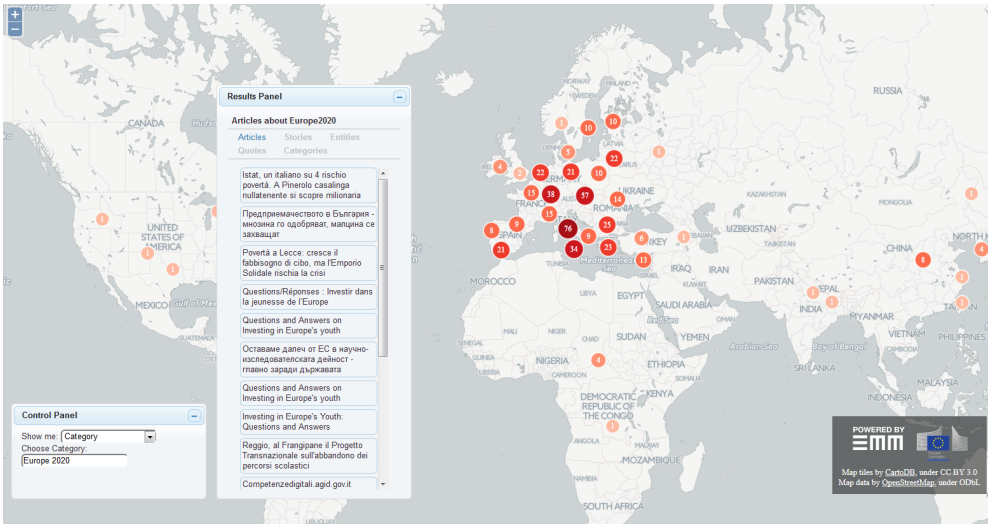
New developments in the application are:

- ***Expanding the list of clients (JRC, APIIPA, Cert, Frontex, OAS, Europol, African Union, FRA)***
- ***Integration with Finder (ability to run any finder query and loop through the returned articles)***
- ***App updated to work with version 2 of the emmApp API***
- ***Introduction of Configuration Sets (multiple clients can display different data using a single instance of the application)***

INFORMATION PRESENTATION

EMM MAP

8



The EMM Map is another useful and popular application that shows the geographical distribution of the news in the EMM system. The news can be displayed by top stories, 24 hour stories, country, category and entity. Timelines can be displayed for stories.

New developments include the possibility to set a default configuration when the application starts.
The first production installation of the application was done at the European Laboratory for Structural Assessment Unit, JRC, followed by CERT, EEAS, Frontex and FRA.

CUSTOMISED DOMAIN

9

MEDISYS

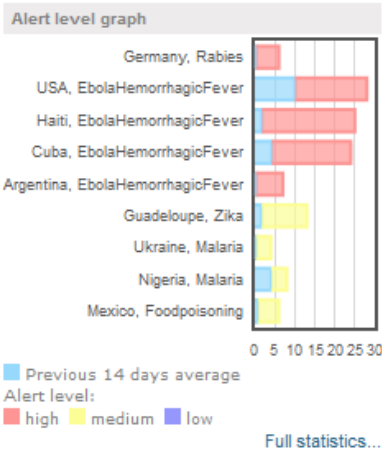
MediSys is an instance of EMM specifically developed for internet bio-surveillance and is used by a number of Health Agencies, including ECDC, EFSA and WHO. A system for the detection of disease-related information published on Twitter was deployed as part of the MediSys website.

The development process is driven by our ongoing collaborations with ECDC on communicable diseases, EFSA on food safety and plant health and EMCDDA on psychoactive substances. We support EU member states in their surveillance efforts and work with international partners such as WHO and G7.
In 2016, we have seen an intense media interest in the Zika virus epidemic in the Americas. The spread of the virus and reports on birth defects were monitored in news media and Twitter.

We are collaborating with EFSA, university of Lleida and Institut d'Investigació de la Generalitat de Catalunya (IRTA) on monitoring plant health threats. More than 150 categories on bacteria, fungi, insects, mollusks, nematodes, oomycetes and viruses that pose a threat to plant health have been added to MedISys. An entire ontology of 350 plant health threats has been developed.

In support of EU member states, we have set up accounts for analysts in Italy for monitoring public health events in Italy during mass gathering events. The NewsDesk tool is used for producing newsletters and sending them to all stakeholders.

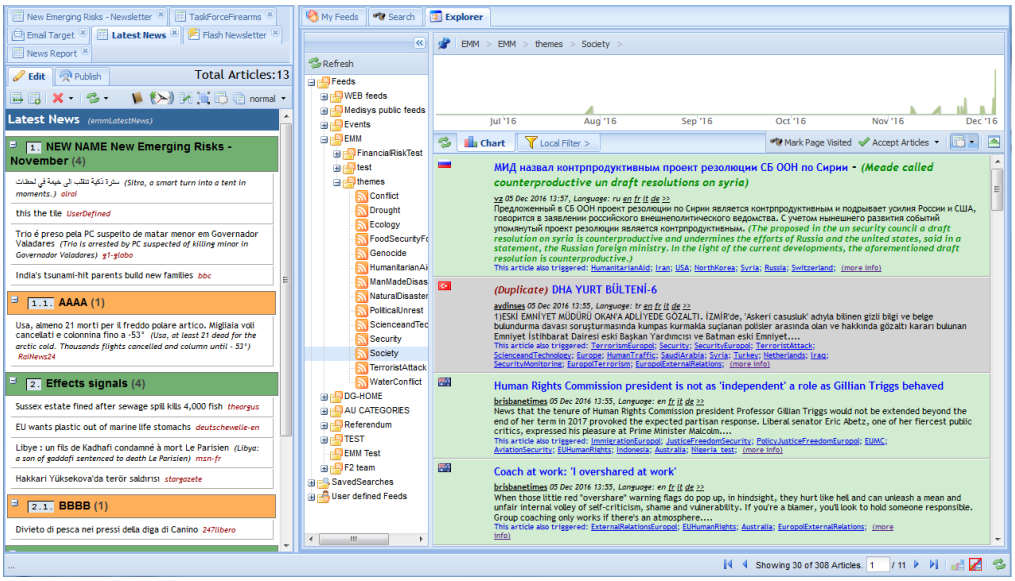
MedISys is also used as processing chain for the WHO HDRAS portal with functionality for commenting and risk assessments in user groups. A similar portal is routinely used by the G7 countries within the GHSAG EAR project. We are involved in the development of the future WHO EIOS portal which will allow various user groups to analyze, assess and comment on news jointly.



EDITING TOOLS

NEWSDESK

10



NewsDesk is a groupware application that allows a community of users organised in workgroups to create reports and newsletters by selecting news items coming from registered sources as well as manually uploaded documents. NewsDesk offers a wide range of tools and features to ease the process of collecting, searching and filtering news items.
Although the Open Source monitoring remains the EMM core business (~300.000 automatically analyzed news per day), the newly developed “PressReview” module addresses another need common to several businesses: the aggregation of human-moderated content. The supported business process is often named “PressReview” because one of the main products is the Daily Press Review report: a daily selection of most representative news in the press from different countries.

NEWSDESK

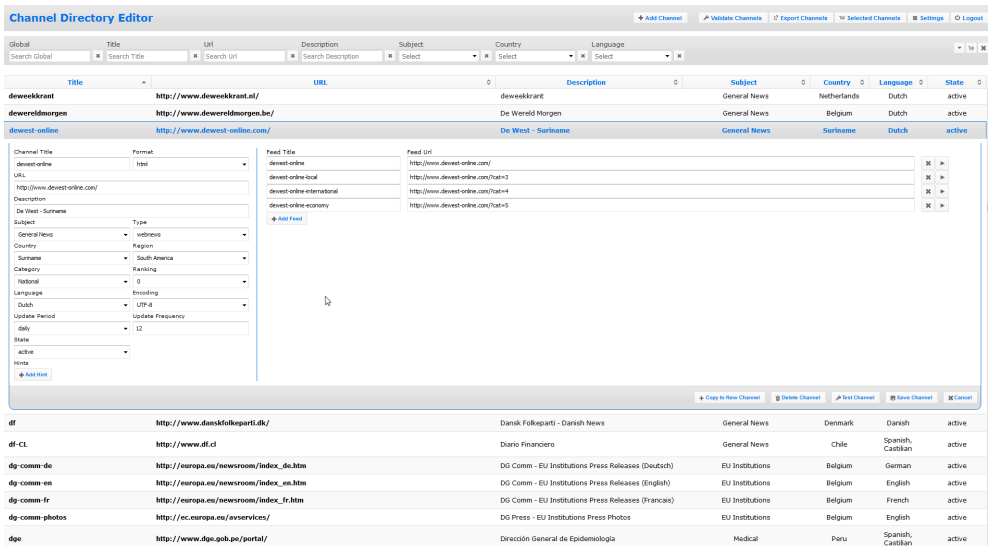
The IT infrastructure needed to support the press review workflow implements a distributed system that allows several groups of analysts to cooperate in the collection, tagging, and publishing of regional content. A central workgroup oversees the activities of the others and prepares products that aggregate contents from all the countries. The above use case has been recently implemented within the European Parliament Media Monitor Platform (EPMM).
The EP press review involves different actors across all the 28 EU countries. In each country there is an EP Information Office, responsible for coordinating the collection of news items for that country. The cuttings are manually uploaded every day by a contractor company (one for each country) and uploaded via the Document Upload module of NewsDesk. Each country has a dedicated workgroup in NewsDesk. All the country workgroups are orchestrated and supervised by the Headquarter workgroup located at EP premises.
The news items **Document Upload** module lets the users define meta-information about the item, like title and description in two languages, the publication date, whether the EP is mentioned in the title of the news item, the type of the uploaded item, the source, and so forth. Probably one of the most interesting features is the possibility for the user to manually assign one or more categories to the uploaded item. Every uploaded item will flow then through the EMM processing chain where the categorisation system automatically adds additional categories to the item (based on alert and filter specification).

So at the end of their journey through the press review system all the news items are tagged both by analysts and the automatic categorisation system.
Once all the cuttings have been uploaded in NewsDesk, each workgroup makes a final selection of the most significant items for that specific country by adding them to one or more newsletters/reports. At the end of the selection process each newsletter can still be edited and further refined. After the editing step, the final product - the newsletter in HTML, PDF and DOCX formats - is generated and sent to the subscribers. In the meanwhile, the Headquarter workgroup also publishes its own newsletters with a selection of items coming from all the countries.
All workgroups can also access the System View module of NewsDesk to retrieve statistics of uploaded and published news items in order to perform the accounting process.

Highlights:

- ***Items are tagged both by analysts and the automatic categorisation system***
- ***Creation of reports and newsletters***

11 CHANNEL EDITOR



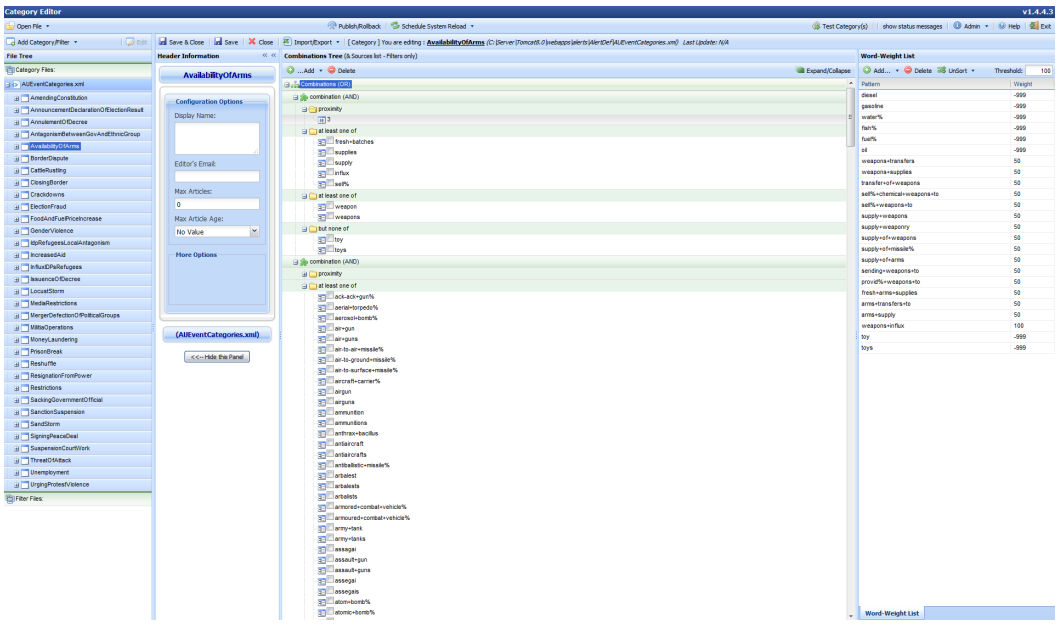
The Channel Editor application allows complete management of the sources monitored by the EMM system. Sources can be easily filtered with the advanced search functionality. Also, the application features a source validation mechanism to ensure articles can be properly read by the system. The flexible export options allow different sets of sources to be published to various processing chains or to be saved in xml and xlsx format.

Highlights:

- ***Ability to import a channel directory into the application***
- ***Global source validation (produces an xls report for all sources)***
- ***Audit functionality (track which users worked on a specific source)***
- ***Integration with Scraper/ Grabber logs (being able to monitor from within the application the output and health of a source)***

CATEGORY EDITOR

12

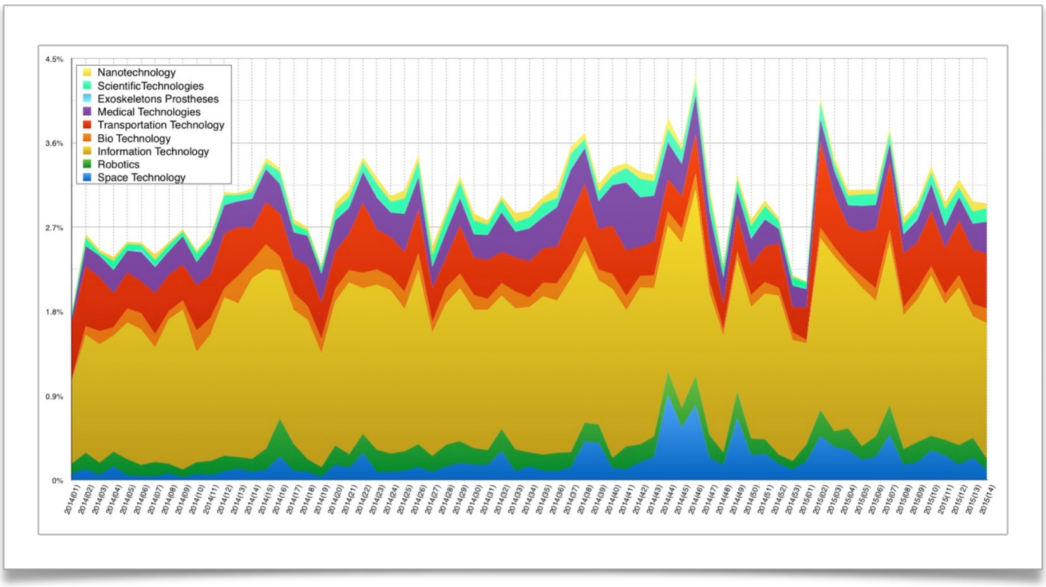


The Category Editor application manages the definition files used by the EMM system to categorise incoming information. The definition files are kept in a central repository and the application allows multi-user access and locking management for the repository. Through its flexible publishing mechanism, the application can use a single repository to serve multiple processing chains.

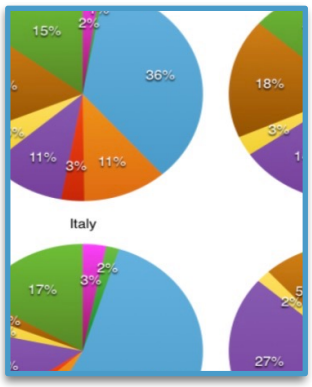
Category Editor Collaboration Layer

The Category Editor Collaboration Layer is a completely new concept of the most powerful EMM tool, the Category Editor. The Collaboration layer allows users from different organisations to work together on the **category definition**. Each organisation retains complete control over its own category repository while at the same time being able to engage with other partners in order to produce improved definitions that will yield better data. A deep integration with versioning software allows for easy merging, rollback and comparison between category definitions.
A notification system that sends messages to users when a new version is added to the collaboration layer is also developed.

13 TREND IMPACT ANALYSIS (TIA)

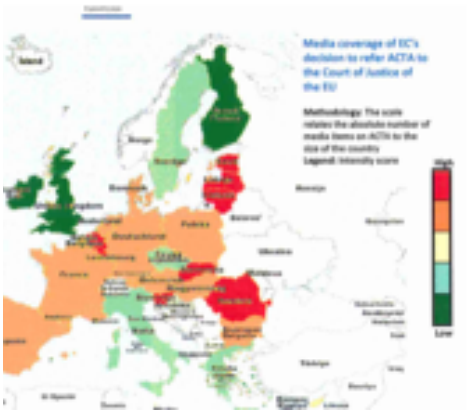


Trend Impact Analysis (TIA) is a new tool that allows users to explore trends in reporting. They can analyse articles collected by EMM in multiple dimensions and interactively, from multiple perspectives. This includes a broad range of information resulting from EMM's automatic analysis. The supported dimensions are time (day, month, year, epoch), topic of article, language of the article, country of publisher and reach of publisher (International, National, Regional or Local). These dimensions can be combined with advanced queries to result in any kind of aggregated data analysis. TIA provides a wizard that guides the users through the selection of the visualisation media (Charts and Maps), the configuration of the dimensions to visualise and finally the selection and filtering of the data to fill these charts or export the results for other analyses. This wizard also supports the predefined configuration and live chart types and/or maps that can then be displayed in dashboards.



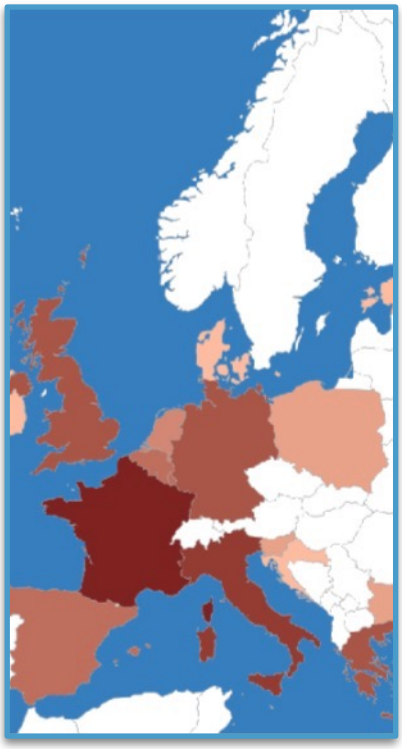
MEDIA IMPACT ANALYSIS (MIA)

14



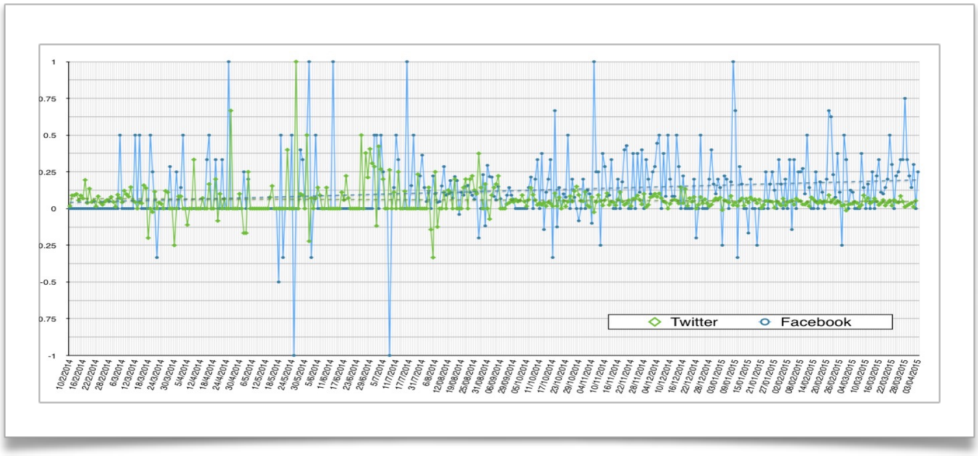
The Media Impact Analysis Tool - MIA supports the typical media impact process which is:
"A combination of procedures, methods and tools by which an event or subject may be judged as to its effects on the population due to media coverage via a report"

MIA is a tool aimed at supporting Media Analysts. It's main function is to allow them to manually tag articles from EMM. The tags can vary depending on the analysis campaign. MIA allows analysts to select the items from EMM using EMM's new Finder application. It then allows them to tag the articles and finally export the resulting data into Excel or as Maps.



ONGOING RESEARCH

15 SENTIMENT ANALYSIS



Sentiment Analysis is the field in Text Mining that deals with the automatic discovery and classification of opinions and sentiments from written text. In general, opinions and sentiments are classified according to their “polarity”, into positive, negative or neutral.
Our in-house sentiment/tonality system analyses - detects and classifies - opinions and sentiments expressed in traditional and social media texts.
Currently, we are working with 3 variants of the sentiment/tonality system that have been implemented and are in use (or ready for use, in the case of the Named Entity one), which we are constantly upgrading by extending to new languages or improving the methods underlying the polarity classification:

- ***Above is an example of a sentiment graph obtained computing the sentiment polarity of Facebook and Twitter messages about IT topics.***

- ***The research in this area envisages the extension and improvement of the existing systems as well as their eventual merging.***

SENTIMENT ANALYSIS

1 - The first one computes the overall tonality of a news article, based on dictionaries of sentiment-bearing words with an associated polarity and intensity score (positive, high positive, negative, high negative). This variant of the Sentiment Analysis system is implemented in the main EMM application and has also been used in a recent project aiming at determining the general public's opinion in Science and Technology topics (Citizen and Science project - CAS).

2 - The second variant computes the sentiment expressed about a Named Entity (person, organization), based on dictionaries of sentiment-bearing words and on a set of heuristics (for negation, proximity, etc.).

3 - The third system classifies the sentiment expressed in a Social Media text (tweet or Facebook post/ comment) based on a hybrid approach - using a supervised machine learning algorithm and dictionaries of sentiment-bearing words to abstract some of the features used for learning. Currently, this variant of the system has been implemented and is working in the context of the "Citizens and Science" project, PUBSY references JRC96113 and JRC96546.

Additional directions for research in which we are currently working are:

- ***Classification of indirectly expressed sentiment/emotion - useful for the case of news, where sentiment is not obviously expressed, but triggered in the readers through the use of specific journalistic techniques;***
- ***Extension and expansion of a knowledge base on concepts and situations that trigger certain affective reactions.***

ONGOING RESEARCH

EVENT DETECTION SYSTEM

16

The live **Automated Event Extraction (AEE)** system automatically determines, for each article cluster, the type, date, exact location, number and type of victims, perpetrators and more of the main reported event, ranging over a large list of event types in the domains of Conflict, Man Made Disasters, Natural Disasters and Humanitarian Crisis. It uses a lightweight semantic approach and NLP techniques to produce frame-like structured metadata.

Recent developments include:

- ***New event types have been deployed: Deportation/ Extradition, Small Arms Trafficking, Human Trafficking, Prison Break, Influx of IDP/ Refugees, etc.***
- ***Three additional languages have been added to the Event Detection System: Bulgarian, Turkish and Romanian. We are also developing event extraction resources for the Czech language.***
- ***A new event location detection algorithm has been deployed. It implements a weighting schema based on a locative expression parsing grammar and on structured event information, such as victims, perpetrators and type-defining keywords. We expect to base the event 'bubble location' on the map interface on this new slot after completing a thorough performance evaluation.***

ONGOING RESEARCH

17 NAMED ENTITY GUESSER

Acronym and multi-word entity extraction

Multi-word entities, such as organisation names, are frequently written in many different ways (e.g. European Commission, European Union Commission, EC, …).
Acronym and multi-word entity extraction consist in automatically detecting, from news articles, links between short forms (acronyms) and long form (multi-word entities) of the same entity.
A cross-lingual acronym and multiword entity extraction system has been developed and evaluated. It should be integrated in the EMM chain in the coming months.

Named Entity Guesser - statistical extension

In the context of the NE Guesser, statistical extensions consist of developing hybrid methods combining statistical and rule-based approaches in order to improve the Guesser output. It includes experiments on automatic cross-lingual lexicon extension and experiments on automatic rule creation.
Automatic cross lingual lexicon extension harmonises the lexical resources we have for different languages.
For instance, by extending lexicons of the less-covered languages based on lexicons we have for the well-covered languages.Automatic rule creation generates new rules for the NE Guesser based on how the person names and organisation names are found in the news articles.

ONGOING RESEARCH

CONVERT NAME RESOURCE IN LINKED DATA

18

The **JRC-Names** resource is a highly multilingual named entity resource for person and organisation names. JRC-Names consists of large lists of names and their many spelling variants (up to hundreds for a single person), including across scripts (Latin, Greek, Arabic, Cyrillic, Japanese, Chinese, etc.). For example, the spellings Jean-Claude Juncker, Jean Cloud Junker, Jean-Claude Juencker, Жан-Клод Юнкер, جان كلود جونكر, Ζαν Κλοντ Γιούνκερ, 让-克洛德•容克, and many others have all been identified as referring to the 12th President of the European Commission.

容

JRC-Names has been available for download since September 2011 as a text file. The new linked data edition, accessible through the European Union's Open Data Portal, offers more structured and machine-readable data. It also contains more information compared to the previously released resource, including: titles and function names that have been historically found next to the person mentions; information about the time period during which name variants and their titles were found; various frequency counts; as well as links to other linked datasets such as DBPedia.

ك

Д

19 TRANSLATION-SYSTEM

The **EMM Translation System** is a real-time machine translation system that translates live news articles from 17 languages (Arabic, Czech, Danish, Dutch, Farsi, Finnish, French, German, Greek, Italian, Latvian, Lithuanian, Polish, Portuguese, Russian, Spanish and Swedish) into English.
The translation service allows our users to get an idea of the main content of an article and to determine whether a news item is relevant for their field of interest.
Due to the large number of news articles and languages, the EMM Translation System was optimised for processing high volumes of text and tries to avoid language-dependent tools such as syntactic parsers and morphological analysers.
It is a phrase-based statistical machine translation (SMT) system (based on Moses), for which we have trained language and translation models.

Customised Features:

- ***The EMM translation system focuses on translating news gathered through the Europe Media Monitor.***
- ***It is optimised for the news domain in order to increase translation quality and speed.***
- ***The system uses different translation models for titles and content.***
- ***Geolocations and named entities detected previously are used for suggesting English***

The EMM Translation System is made of the connection module (a Java servlet which connects the translation module to the EMM news processing pipeline) and Moses servers located on different machines.
Translated articles (titles and descriptions) are available in the EMM family of applications.

ONGOING RESEARCH

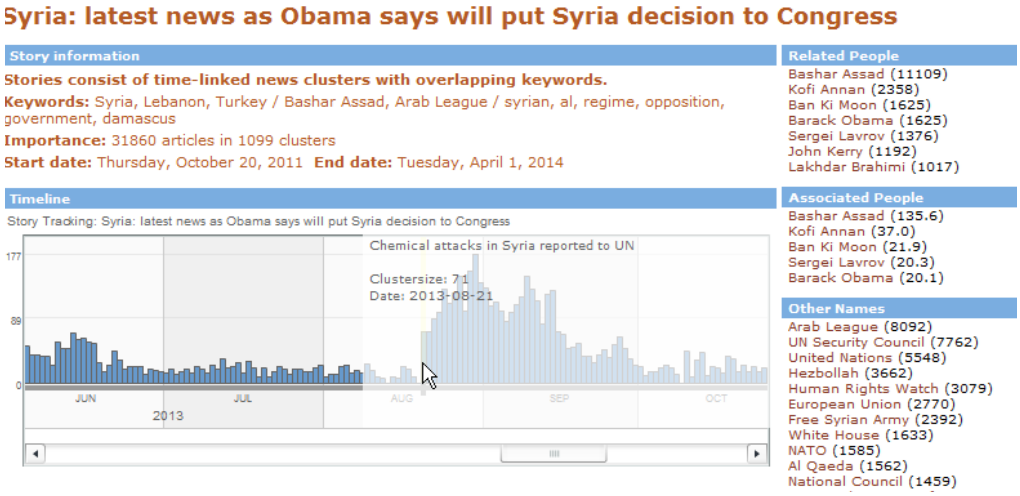
NEWS EXPLORER: 20

Analysis over time and across languages

NewsExplorer links related news across 21 languages (including Arabic and Russian), allowing users to discover possible differences in viewpoints and in the intensity of media reporting in different countries.

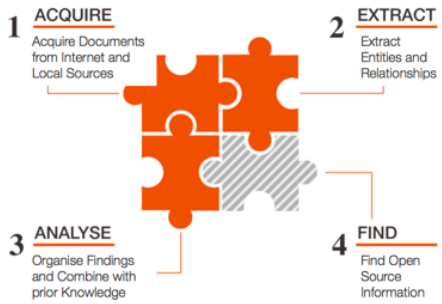
The system also supports users in exploring the news over longer periods of time. Timelines – for stories that catch the media's attention over weeks or even months – are interactive so that users can read up what happened in the past, including across languages. With the calendar function, EMM readers can check what happened on any given day in the past since 2004.

Efforts are currently under way to integrate the historical and the cross-lingual news cluster linking into NewsBrief, MyNews and MedISys.



21 EMM OSINT SUITE

The **Open Source Intelligence Suite (OSINT)** is a desktop software application based on EMM technology which helps to find, acquire and analyse data from the Internet and local sources. Designed for analysts in law enforcement authorities, it is used in other authorities, such as customs and tax authorities as well.



The EMM OSINT Suite comprises a set of tools to support the core processes of intelligence gathering from open sources. Documents in multiple file formats can be acquired from the public Internet as well as from local sources and stored in a user workspace. A built-in entity extraction module identifies persons, organizations, geolocations, phone numbers and custom types defined by the end user. Analysis and Reporting views are provided to visualise the data and export it into third party tools.

In the latest version 2.3 the following features were added:

- ***Category Matching – a new module based on EMM Core technology has been added to the application to allow the end user define categories to tag the documents in the workspace of the tool.***
- ***Improved Custom Entity Support – the module allows adding custom entity types based on user-defined patterns to the system.***
- ***Full-text index of documents in the workspace.***
- ***Software Updates – the deployed tool can be updated automatically.***
- ***Support for Linux desktop computers.***

CONTACT US

European Commission
Joint Research Centre
Directorate I. Competences
Unit I.3 - Text and Data Mining
Via E. Fermi, 2749 - TP 460
I-21027 Ispra (VA)/Italy

emm@jrc.ec.europa.eu