



European
Commission

Europe Media Monitor (EMM)

Joint
Research
Centre

JRC Science Hub

<https://ec.europa.eu/jrc>

Ispra: European Commission, 2018

© European Union, 2018

Reuse is authorised provided the source is acknowledged. The reuse policy of European Commission documents is regulated by Decision 2011/833/EU (OJ L 330, 14.12.2011, p. 39).

For any use or reproduction of photos or other material that is not under the EU copyright, permission must be sought directly from the copyright holders.

All images © European Union 2018

1	INTRODUCTION	04
2	INFORMATION GATHERING	05
3	INFORMATION PRESENTATION	06
3.1.	NewsBrief	07
3.2.	MyNews	09
3.3.	Mobile Devices	11
3.4.	Big Screen Map	13
3.5.	EMM Map	14
4	CUSTOMISED DOMAIN	15
4.1.	MEDISYS	15
5	EDITING TOOLS	17
5.1.	NewsDesk	17
5.2.	Channel Editor	19
5.3.	Category Editor	20
6	INFORMATION ANALYSIS	21
6.1.	Trend Impact Analysis (TIA)	21
6.2.	Media Impact Analysis (MIA)	22
7	ONGOING RESEARCH	23
7.1.	Sentiment Analysis	23
7.2.	Event Detection System	25
7.3.	Named Entity Recognition	26
7.4.	Convert Lexical Resource into Linked Data	27
7.5.	Translation System	28
7.6.	NewsExplorer	29
7.7.	EMM OSINT Suite	30

1. Introduction

FACTS AND FIGURES:

- ▶ Almost 8 000 news sites monitored
- ▶ 300 000 articles per day processed
- ▶ 70 languages
- ▶ 3 000 categories
- ▶ Running 24/7
- ▶ 25 000 visitors/day

We would be very pleased to receive your feedback, and would gladly provide you with further information. Please contact us at:

jrc-emm-support@ec.europa.eu

For more information about the Joint Research Centre (JRC) please visit the following link:

<https://ec.europa.eu/jrc>

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 **Text & Data Mining Unit**, in the **Directorate for Competences 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 metadata, 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 almost 8 000 **HTML pages** and **RSS feeds** to find new articles published on the Internet (~**300 000 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.

2. Information Gathering

HIGHLIGHTS:

- ▶ Article extraction from HTML feeds.
- ▶ Website scraping for unstructured pages
- ▶ Access to the Internet using configurable proxy servers and user agents
- ▶ Handling of badly formatted RSS feeds
- ▶ Possibility to have different flows of news articles
- ▶ Support for social media monitoring

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 are 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 Tweets that are related to the most recent disease outbreaks, violent events and disasters.

From 2018 EMM can also grab content from a live Twitter stream and from Facebook public pages.

3. Information Presentation

EMM offers a set of several applications and tools to let users access the collected news articles in many different ways: NewsBrief, NewsDesk, MyNews, MediSys and EMM Mobile Apps



NewsBrief

NewsDesk

MyNews

MediSys

EMM Mobile Apps

The results of the information harvesting and processing can be accessed in a number of ways: a **NewsBrief** website (<http://emm.newsbrief.eu>) that allows for classical data browsing; MyNews, a fully customizable news dashboard; and a full editorial and publishing system **NewsDesk**, which 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 devices.

Examples of current applications of the **EMM** technology can be found in different 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.

At the moment, the publicly accessible instance of **EMM** monitors over **23 000 RSS feeds/ HTML** pages from almost **8 000 media websites** and retrieves and processes around **300 000 new news articles** per day. These articles are categorized into over **3 000 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

3.1. NewsBrief

The primary aim of the system is to provide frequent and near real-time news updates on a selection of topics and to give an overview of top trending stories

EMM NewsBrief is a public website that provides many different views on the news published nearly real-time.

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.

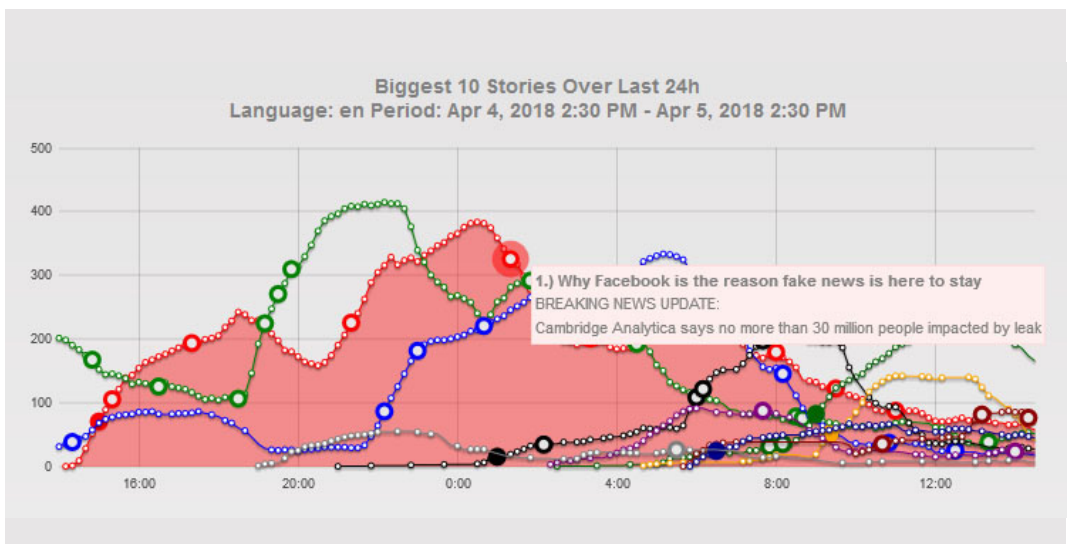
The categories visible on the public version of NewsBrief are predefined and cannot be modified by the public. On every 'category page' you can see how a particular category is defined. Most of these categories are defined by domain experts and are made available to you.

However, there are several customized versions of the NewsBrief website (most of them not publicly available) running for different clients and based on customized lists of web sources and categories. Examples are <https://nsmm.emm4u.eu>, <https://cert.europa.eu>.

On every page you can subscribe to receive the news by e-mail, or you can access the information through the RSS feed.

Feel free to include any of our RSS feeds in 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 at the moment EMM grabbed it, together with all metadata extracted by EMM.



Information presentation

3.2. MyNews

HIGHLIGHTS

- ▶ Highly and easily customisable on a 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

MyNews is a highly customisable web interface that gives access to the news items produced by the **EMM** engine.

It is user-driven, because its main focus is to offer 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.
- ▶ Multi-language 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).

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.

The screenshot displays a news channel interface. On the left, three article cards are visible, each with a title, a brief description, trigger words, and source information. On the right, a 'GEO location' map shows a world map with colored circles and numbers indicating article counts by region: North America (13), South America (15), Europe (94), Africa (2), Asia (7), and Oceania (4). The interface includes navigation icons and a footer with statistics: 142, 77, 217, 8.

Text adopted - Application of the Postal Services Directive - P8_TA-PROV(2016)0357 - Thursday, 15 September 2016 - Strasbourg - Provisional edition

A. whereas the postal market is still an area of the economy with strong prospects for growth and increasing competition, even though between 2012 and 2013 letter post services shrank by 4,85 % on average in the EU according to the European Commission Postal Statistics Database, which is in line....

Trigger words: European citizens, EU citizens

europarl_ONLINE SOURCES lang:en pub: 7 Dec 2016, 16.03

41 1 5

Investing in Europe's Youth: Questions and Answers

The recent Commission report on the implementation of the Youth Guarantee and the Youth Employment Initiative shows that the Youth Guarantee has become a reality across the EU and is yielding results. Around 9 million young people have taken up an offer under the Youth Guarantee scheme since its implementation.

Trigger words: European citizens, European citizenship

EU europa-nu_ONLINE SOURCES lang:en pub: 7 Dec 2016, 14:26

23 8

Investing in Europe's Youth: Questions and Answers

What is the initiative "Investing in Europe's Youth" about? Today the Commission has presented several key actions to support young people in Europe. This includes the launch of the European Solidarity Corps , which was announced by President Juncker in his State of the Union speech on 14 September 2016.

Trigger words: European citizens, European citizenship

europa-pressReleases_EU PRESS RELEASES lang:en pub: 7 Dec 2016, 12:37

24 1 8

GEO location

Map Satellite

NORTH AMERICA 13

SOUTH AMERICA 15

AFRICA 2

EUROPE 94

ASIA 7

OCEANIA 4

Google ANTARCTICA Map data ©2016 Terms of Use

142 77 217 8

Information presentation

3.3. Mobile Devices

HIGHLIGHTS:

- ▶ The power of EMM always in your pocket
- ▶ Both for iPhone/iPad and Android devices
- ▶ Customised versions



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 the 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.

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

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 presets.



Supporting:



Android



iOS + MacOS

Information presentation

3.4. Big Screen Map

HIGHLIGHTS:

- ▶ Expanding the list of clients (JRC, Cert, Frontex, OAS, Europol, African Union, FRA, EFSA, WHO)
- ▶ Integration with EMM Finder (ability to run searches on the EMM Finder index and loop through the returned articles)
- ▶ Automatic data refresh on each loop restart
- ▶ Ability to define multiple Configuration Sets (various clients can display different data using a single instance of the application)

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.



Information presentation

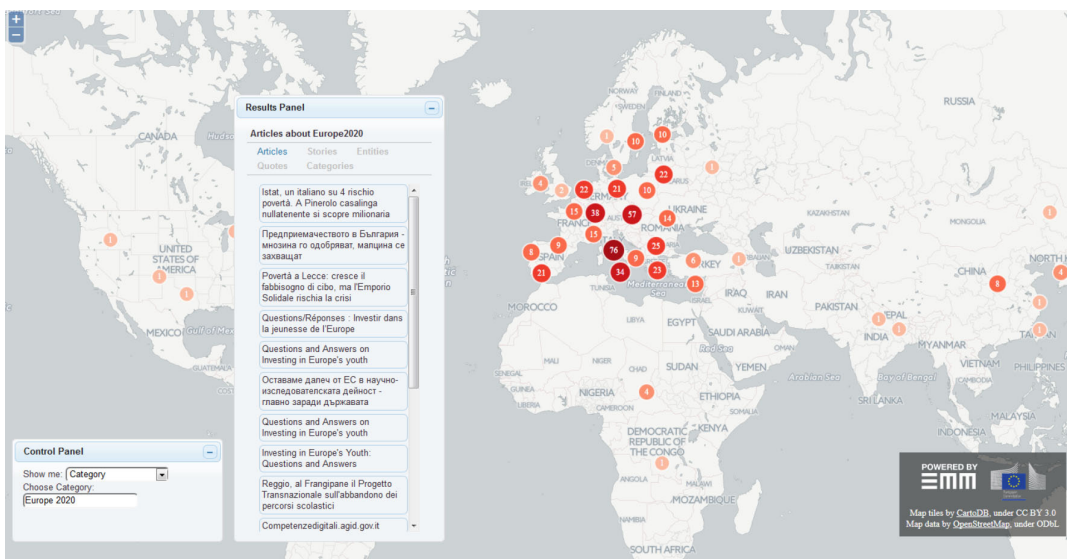
3.5. EMM Map

HIGHLIGHTS:

- ▶ Geographical distribution of the news
- ▶ Animated timelines

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 **JRC European Laboratory for Structural Assessment**, followed by **CERT**, **EEAS**, **Frontex** and **FRA**.

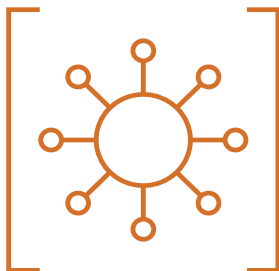


4. Customised Domain

4.1. MEDISYS

HIGHLIGHTS:

- ▶ Specifically developed for internet bio-surveillance
- ▶ Specific set of sources
- ▶ Main partners are WHO, ECDC, EFSA, EMCDDA
- ▶ Detection of disease-related information



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.

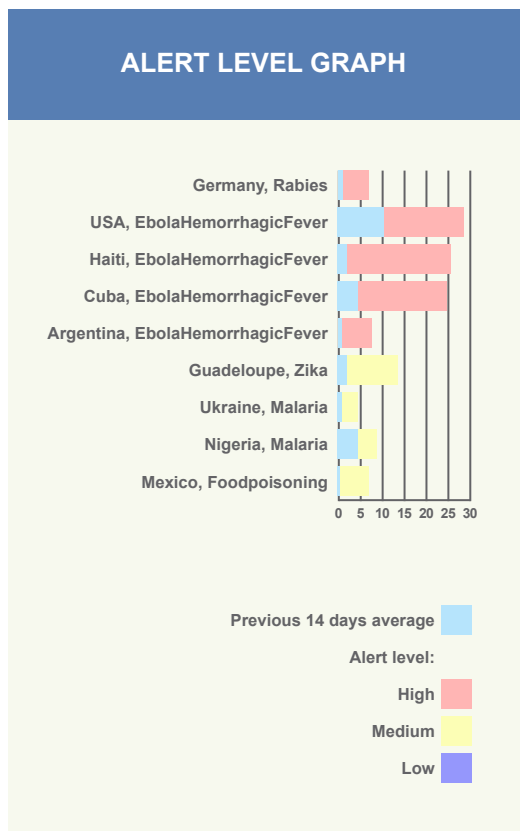
The development process is driven by our ongoing collaborations with **ECDC** on communicable diseases, **EFSA** on food safety and plant health, and **EMCDDA** on psychoactive substances. We support EU Member States in their surveillance efforts and work with international partners such as **WHO** and **G7**.

We are collaborating with **EFSA** 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**.

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 functionalities 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.



5. Editing Tools

5.1. NewsDesk

HIGHLIGHTS:

- ▶ Groupware application
- ▶ Items are tagged both by analysts and the automatic categorisation system
- ▶ Creation of reports and newsletters
- ▶ Media analysts can work together to review the news feeds and to create final products
- ▶ Social media integration

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 **Press-Review** module addresses another need common to several businesses: the aggregation of human-moderated content. The supported business process is often named **Press-Review** 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.



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 content 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 then

flow through the EMM processing chain where the categorisation system automatically adds other 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 by 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 Headquarters workgroup also publishes its own newsletters with a selection of items coming from all the countries.

NewsDesk supports highly customized templates to render the final products. EMM is maintaining now a database with more than 120 different templates.

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.

Editing Tools

5.2. Channel Editor

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)

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.

The screenshot shows the 'Channel Directory Editor' application. At the top, there are navigation links: 'Add Channel', 'Validate Channels', 'Export Channels', 'Selected Channels', 'Settings', and 'Logout'. Below this is a search bar with fields for 'Global', 'Title', 'URL', 'Description', 'Subject', 'Country', and 'Language'. A table lists several channels:

Title	URL	Description	Subject	Country	Language	State
deweekkrant	http://www.deweekkrant.nl/	deweekkrant	General News	Netherlands	Dutch	active
dewereldmorgen	http://www.dewereldmorgen.be/	De Wereld Morgen	General News	Belgium	Dutch	active
dewest-online	http://www.dewest-online.com/	De West - Suriname	General News	Suriname	Dutch	active

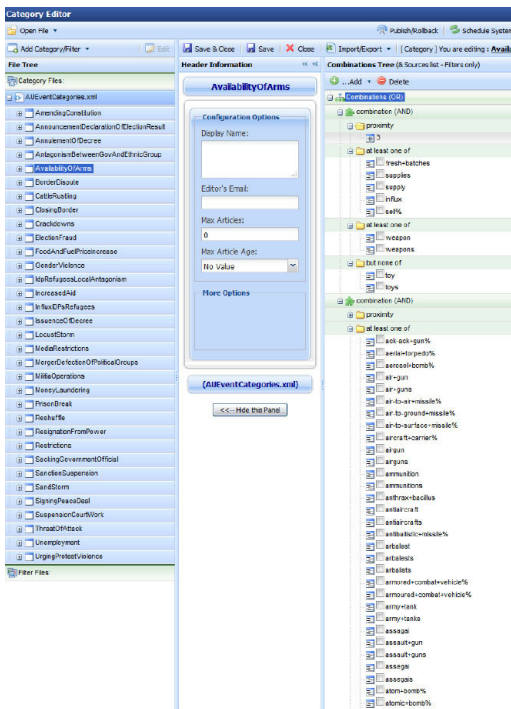
Below the table, there is a detailed form for editing a channel. The form is divided into several sections:

- Channel Title**: Format (html), URL (http://www.dewest-online.com/), Description (De West - Suriname), Subject (General News).
- General News**: Type (webnews), Country (South America), Surname (Marketing), Category (National), Language (Dutch), Update Period (Daily), Update Frequency (12), State (active).
- Feed Title**: dewest-online, dewest-online-local, dewest-online-international, dewest-online-economy.
- Feed Url**: http://www.dewest-online.com/, http://www.dewest-online.com/feed=3, http://www.dewest-online.com/feed=4, http://www.dewest-online.com/feed=5.

At the bottom, there are buttons for 'Copy to New Channel', 'Delete Channel', 'Add Channel', 'Save Channel', and 'Cancel'.

Editing Tools

5.3. Category Editor



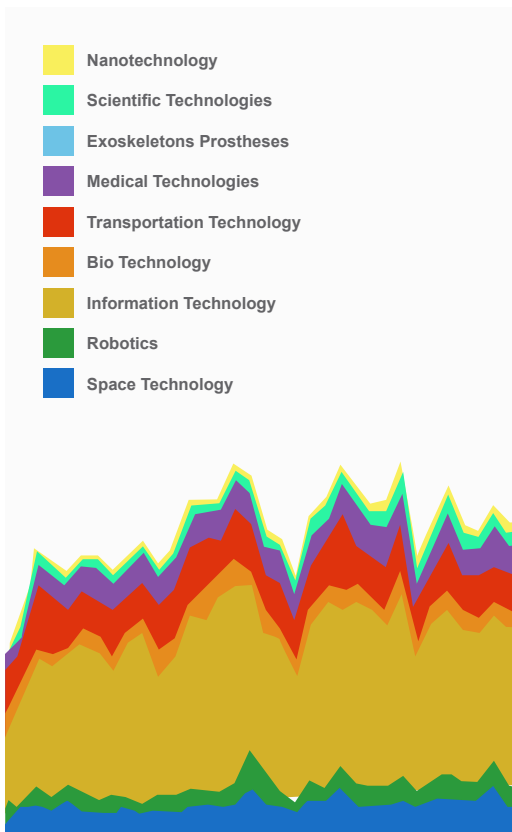
The **Category Editor** application manages the definition files used by the EMM system to categorise incoming information. The definition files are kept in a central repository and the application allows multi-user access and locking management for the repository. Through its flexible publishing mechanism,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

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.

6. Information Analysis

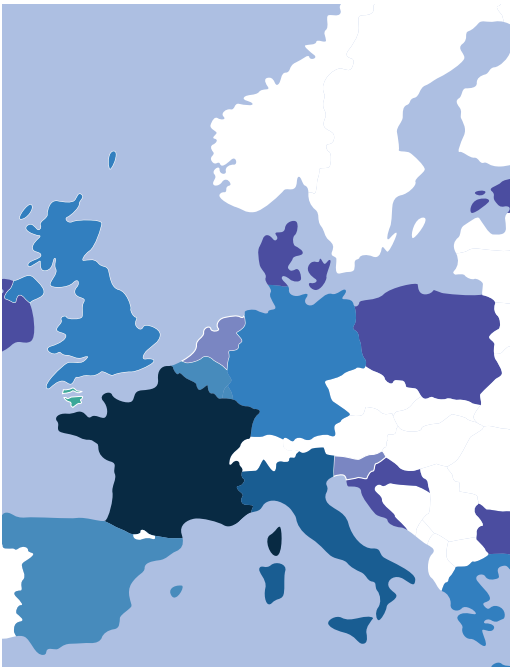
6.1. Trend Impact Analysis (TIA)



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.

Information Analysis

6.2. Media Impact Analysis (MIA)



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.

7. Ongoing Research

7.1. Sentiment Analysis

HIGHLIGHTS:

- ▶ 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
- ▶ Use of indirect expressions of emotions in the context of the fight against disinformation

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 three 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:

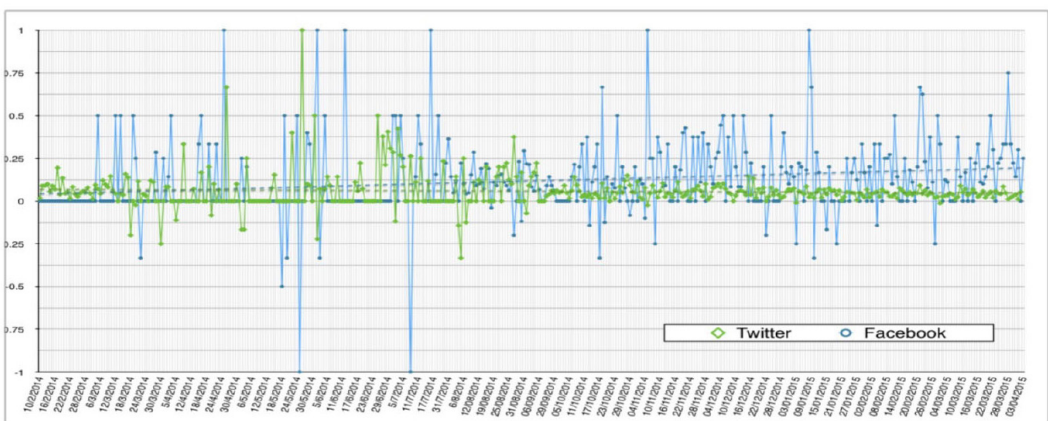
1. The first one computes the overall tonality of a news article, based on dictionaries of sentiment-bearing words with an associated polarity and intensity score (positive, highly positive, negative, highly negative).

This variant of the **Sentiment Analysis** system is implemented in the main **EMM** application and has also been used in a pilot 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.

- ▶ Below is an example of a sentiment graph obtained by computing the sentiment polarity of Facebook and Twitter messages about Information Technology topics.
- ▶ The research in this area envisages the extension and improvement of the existing systems, as well as their eventual merging.



7.2. Event Detection System

HIGHLIGHTS:

- ▶ Large range of event types
- ▶ Highly multilingual
- ▶ Supports both real-time monitoring and long term analysis
- ▶ Integrates algorithms for fine-grained event location and time detection

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 Crises. It uses a lightweight semantic approach and Natural Language Processing techniques to produce frame-like structured metadata.

New event types are added and event template structure customized according to information need of the system users, by deploying Machine Learning techniques.

The system currently processes news in 10 languages: English, French, Spanish, Italian, Portuguese, Russian, Bulgarian, Turkish, Romanian and Czech.

It provides a map interface for global/local situation monitoring and also access to historical datasets of fully automatic and human moderated events for quantitative analysis.

Based on finite-state parsing of locative and temporal expressions, events location and time are used to place the event 'bubbles' in a map interface and to assign them exact time stamps.

Ongoing Research

7.3. Named Entity Recognition

HIGHLIGHTS:

- ▶ Automatic detection of links between different forms of the same entity
- ▶ Automatic cross-lingual lexicon extension
- ▶ Automatic rule creation

Named Entity (NE) Recognition consists of extracting from text, expressions referring to entities like person names, organization names, etc.

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.

In the context of the **NE Recognition**, statistical extensions consist of developing hybrid methods combining statistical and rule-based approaches in order to improve the **Recognition** 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 Recognition** based on how the person names and organisation names are found in the news articles.

es protestas pacíficas contra el presidente **Bashar el Asad**.

fr la Lybie conquise pour la Syrie de **Bachar Al-Assad**.

de syrischen Führung unter **Präsident Baschar al-Assad** sei.

it contro il regime del **presidente Bashar al Assad** in Siria.

Ongoing Research

7.4.

Convert Lexical Resource into Linked Data

HIGHLIGHTS:

- ▶ Large list of names and spelling variants
- ▶ Available for download from the EU Open Data portal

The JRC-Names resource is a highly multi-lingual 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-КлодЮнкер, 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.

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.

Details and download:

[https://data.europa.eu/euodp/en/data/dataset/jrc-
emm-jrc-names](https://data.europa.eu/euodp/en/data/dataset/jrc-emm-jrc-names)



Ongoing Research

7.5. Translation System

HIGHLIGHTS:

- ▶ 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 translations to the Statistical Machine 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.

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

7.6. NewsExplorer

HIGHLIGHTS:

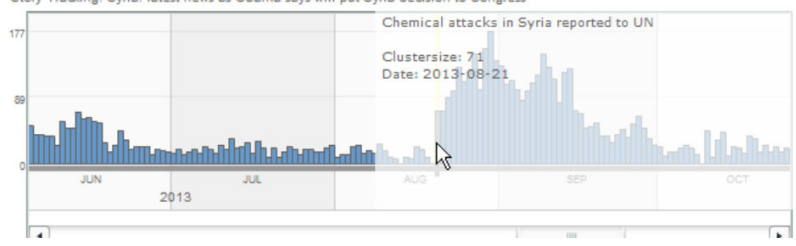
- ▶ Analysis over time and across languages

NewsExplorer (<http://emm.newsexplorer.eu>) 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 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**.

Syria: latest news as Obama says will put Syria decision to Congress

Story information Stories consist of time-linked news clusters with overlapping keywords. Keywords: Syria, Lebanon, Turkey / Bashar Assad, Arab League / syrian, al, regime, opposition, government, damascus Importance: 31860 articles in 1099 clusters Start date: Thursday, October 20, 2011 End date: Tuesday, April 1, 2014	Related People Bashar Assad (11109) Kofi Annan (2358) Ban Ki Moon (1625) Barack Obama (1625) Sergei Lavrov (1376) John Kerry (1192) Lakhdar Brahimi (1017)
Timeline Story Tracking: Syria: latest news as Obama says will put Syria decision to Congress 	Associated People Bashar Assad (135.6) Kofi Annan (37.0) Ban Ki Moon (21.9) Sergei Lavrov (20.3) Barack Obama (20.1) Other Names Arab League (8092) UN Security Council (7762) United Nations (5548) Hezbollah (3662) Human Rights Watch (3079) European Union (2770) Free Syrian Army (2392) White House (1633) NATO (1585)

Ongoing Research

7.7. EMM OSINT Suite

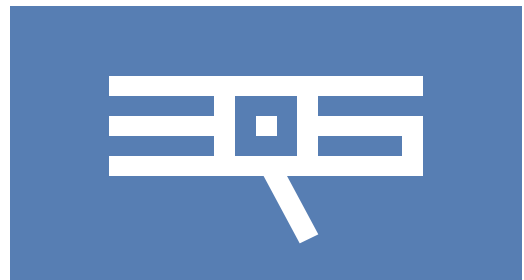
HIGHLIGHTS:

- ▶ **Category Matching** – a new module based on EMM Core technology has been added to the application to allow the end-user define categories and tag the documents in the workspace of the tool
- ▶ **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**

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, organisations, geolocations, phone numbers and custom types defined by the end user.

Analysis and Reporting views are provided to visualise the data and export them into third-party tools.



GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: <http://europa.eu/contact>

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696, or
- by electronic mail via: <http://europa.eu/contact>

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: <http://europa.eu>

EU publications

You can download or order free and priced EU publications from EU Bookshop at: <http://bookshop.europa.eu>. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see <http://europa.eu/contact>).

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.



EU Science Hub
ec.europa.eu/jrc



@EU_ScienceHub



EU Science Hub - Joint Research Centre



Joint Research Centre



EU Science Hub