

EMM: Supporting the Analyst by Turning Multilingual Text into Structured Data

Abstract

All information-seeking professionals need to sieve through large amounts of text to retrieve the information they need so that they can stay up-to-date of developments in their field. Language Technology tools can help make the analyst's work more efficient by increasing the amount of data analysed and by speeding up the process. Software tools applied to big data may additionally provide a bird's-eye view of trends and data distributions not easily visible to the human reader. The European Commission's *Joint Research Centre* (JRC) has developed the *Europe Media Monitor* (EMM) family of applications, which aims to provide solutions for the daily media monitoring needs of a large variety of users working in diverse fields. EMM gathers and analyses hundreds of thousands of news articles every day in up to seventy languages. Due to the large scale of the effort, EMM can track topics, detect trends and act as an early warning tool. In this chapter, we present the functionality and the benefits of EMM's news analysis capacity. We also aim to make the reader aware of the potential dangers of automated large-scale media monitoring. The EMM team makes available for free a number of linguistic tools and resources that can be used by information specialists to improve their own analysis of large sets of textual data.

1. Introduction

Information is power. Organisations have always tried to be well-informed with respect to their interests and activities. This includes political organisations such as parties which want to know how the public feels about their initiatives – an intrinsic part of the democratic process. It includes companies that monitor their field and their competitors (competitive intelligence), and it includes Non-Governmental Organisations (NGOs) or public authorities such as those responsible for keeping an eye on Public Health (medical intelligence). There are of course also national law enforcement authorities that complement their internally available information with information from the internet (open source intelligence). Some of the required information can be found in the social or in the traditional media. In the past, organisations paid for manual news clipping services where human readers sieved through

the major newspapers and their selection was collated into an in-house newspaper. Some organisations continue this practice even today. The European Commission's in-house science service *Joint Research Centre* (JRC) automated this task at the beginning of this century and created its *Europe Media Monitor* (EMM). Since then, EMM has grown to monitor around 6,500 news sources in over 70 languages (status September 2016) and it offers much more functionality, including information extraction, trend detection, machine translation and the visualisation of information. Large parts of EMM are freely accessible to the public through various interfaces, including HTML for computers, apps for mobile phones and tablets, RSS for news reader software, KML for Google Maps, and more. The main functionality of EMM is summarised in Section 2. EMM is not the only automatic media monitoring system available. In Section 3, we draw a picture of the landscape of tools and services available to date.

However, can machines really substitute human intelligence? Is their output reliable? How much of the news analysis task can they take over? Do human analysts need to fear for their job? What do system users need to look out for? Section 4 addresses these questions. EMM's components were developed entirely in-house and some of these components are available for public or commercial developers of text mining software. Section 5 describes these tools and resources, and it also addresses the main theme of this book, i.e. transparency and responsibility. Section 6 summarises the main points of this chapter and gives an outlook on developments that can be expected in the near future in order to make EMM and other news analysis products even more useful.

2. The Europe Media Monitor (EMM)

EMM ( ) is an entirely automatic system of well-integrated software tools that currently visits about 6,500 different online news sources in over 70 languages and that collects and analyses a current average of about 250,000 news articles per day. A whole series of Language Technology software components then analyse these news articles, adding more and more meta-information to each article in a processing pipeline with rising complexity.

2.1 Meta-information extracted from the news

The initially unstructured text is turned into partially structured data, allowing searching of its contents like those of a database and allowing the analysis and visualisation of the information. EMM runs around the clock and many of its web pages are updated every ten minutes, letting users see at any time of the day what is currently happening around the world. Below, you find the list of information facets EMM produces for each news article:

- Time stamp (time of publication);

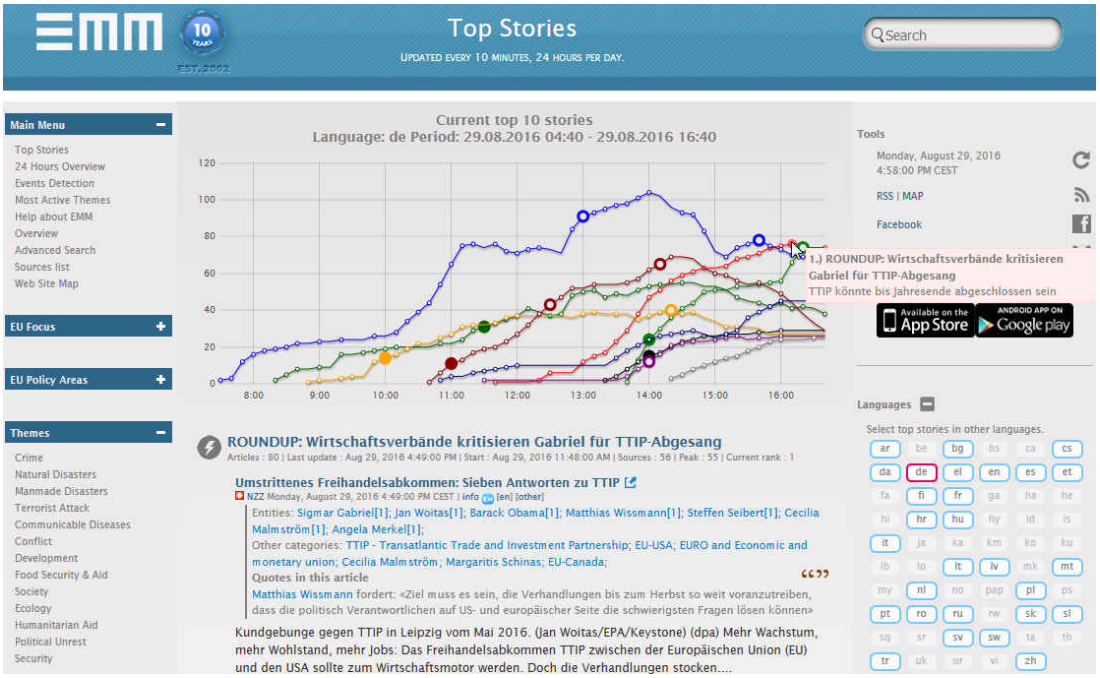

Figure 1. Entry page of EMM-NewsBrief, displaying the ten currently largest news clusters and their size development over the last hours, plus meta-information about each cluster.

- URL (internet address);
- Country where the news was published;
- Countries mentioned in the article;
- News source (e.g. newspaper name);
- Publication language;
- Information on the media type (social vs. traditional media; regional, national or international, etc.);
- News category / subject domain;
- Number of related articles (same day or in previous days);
- Related articles in other languages;
- Names of persons, organisations and locations mentioned;
- Quotations by and about people;
- Sentiment / tonality;
- Event information (who did what to whom, where and when);
- Machine translation results (translation into English only);
- Any combinations of features; average values; change over time; etc.

For ambiguous strings such as *Paris*, EMM disambiguates whether it is a person or a location name (e.g. *Paris Hilton*) and – in the case of multiple locations with the same name – it uses various heuristics to decide which one is being referred to (2006).

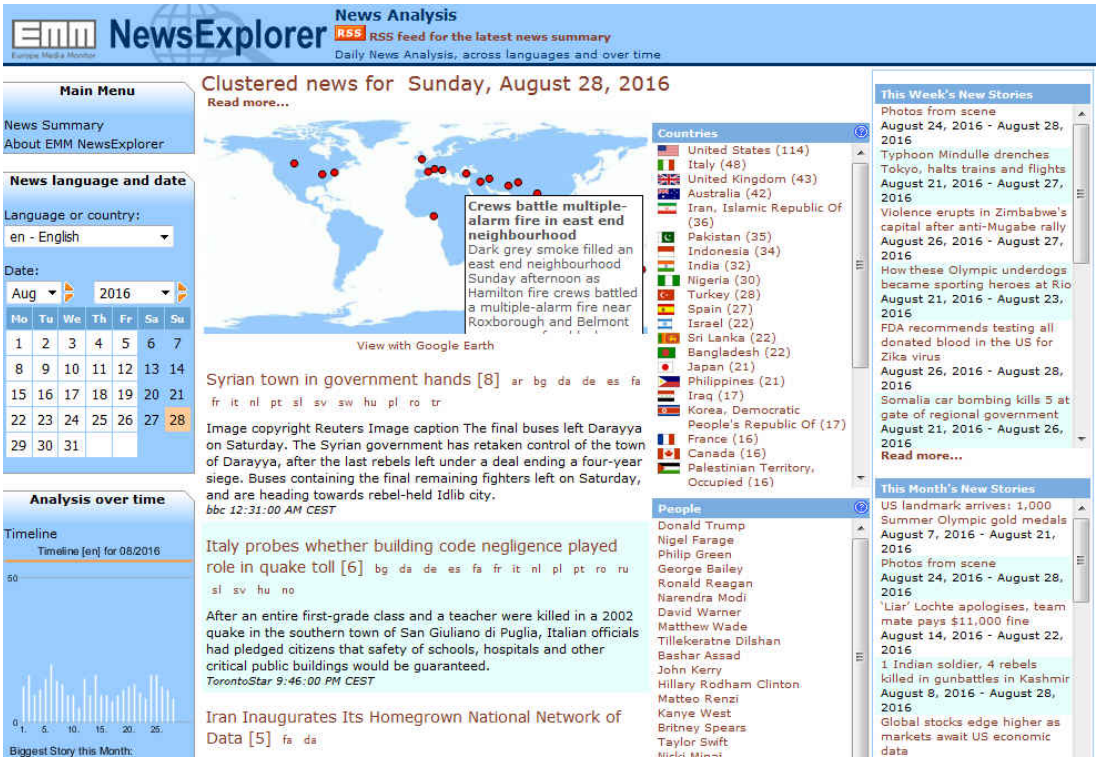

Figure 2. Entry page of EMM-NewsExplorer, displaying the largest news clusters per day, plus information on the biggest news stories this week and this month, and more.

2.2 The importance of monitoring the news in many languages

Not all information is available for all languages. While the automatic grouping of related articles and the classification into hundreds or thousands of news domains is mostly done for all languages, other text analysis tools are only available for a subset. For instance, event information (the most complex set of information aspects) is available for only eleven languages, named entities (persons, organisations and locations) as well as quotes by and about people are mostly available in at least 21 languages. Machine translation results are currently being produced for 16 languages.

Is it really necessary to monitor the news in so many different languages? Aren't all major events and facts also mentioned at least in the English language press? It may well be the case that those major events that are currently in the international focus are somehow covered by the press in many different languages. However, the viewpoints and the focus of the reporting clearly differ and large numbers of events are not reported in other languages. A study by [redacted] (2011) showed that only 51 out of 523 observed events were reported in more than one language, i.e. less than 10%. Furthermore, 350 of the 523 events (67%) were found in non-English news. This certainly confirms the observation of EMM users, who are often

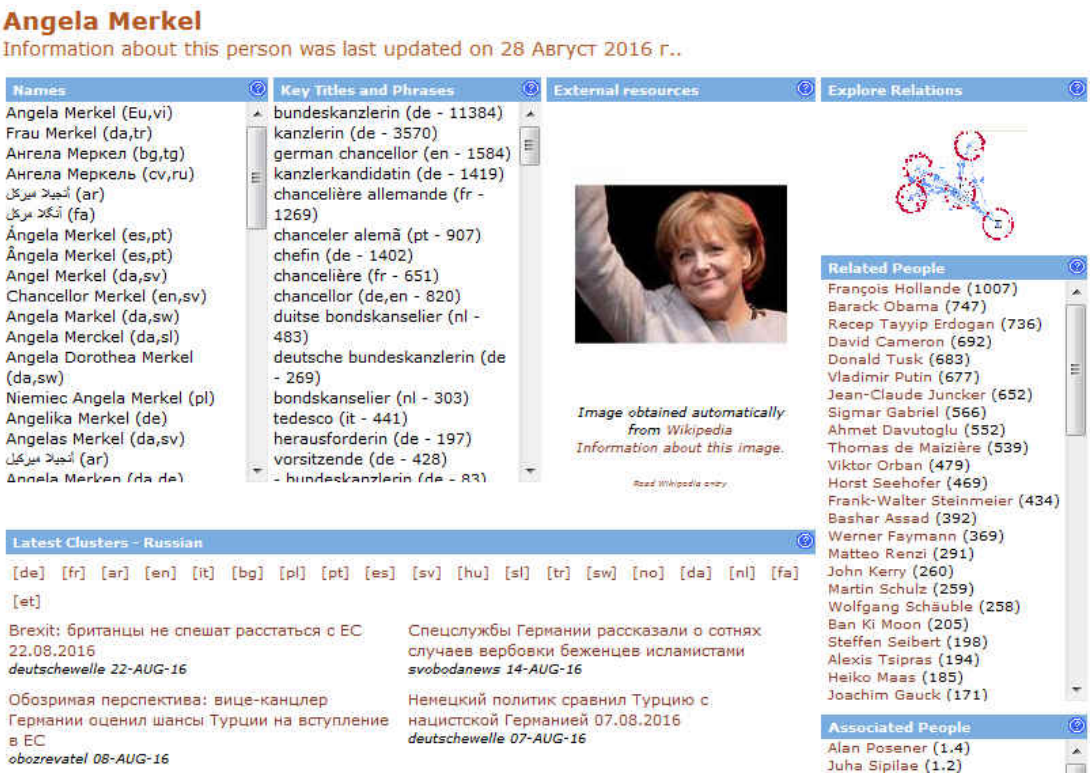

Figure 3. Entity page of EMM-NewsExplorer, showing information collected about this entity over time and in different languages, e.g. titles, name variants, quotations, etc.

interested in less-reported events such as disease outbreaks, violence, smuggling events or the mention of certain specialised organisations. When displaying all current news clusters in all EMM languages on one map, we also observe that the clusters in different languages are extremely unevenly distributed (see the top-left graph in **Figure 4**). News reporting is highly complementary across languages, both regarding facts and opinions.

2.3 News display by combining extracted meta-information

EMM displays the extracted information facets next to each group of related articles so that readers receive background knowledge before even reading the text. **Figure 1** shows a snapshot of the German front page of EMM-NewsBrief. The interactive trend lines show the reporting intensity for the currently ten largest news clusters. The number of articles (y-axis) refers to the number of articles that arrived in the last four-hour period. This number is recalculated every ten minutes. The text section shows the names found in the current top cluster, consisting of 80 articles, as well as the various news categories and a quotation (reported speech). NewsBrief shows the currently most reported news from around the world. It is updated every ten minutes. **Figure 2** shows the English front page of EMM-NewsExplorer. Similarly to a newspaper, NewsExplorer is updated once a day as it summarises the ma-

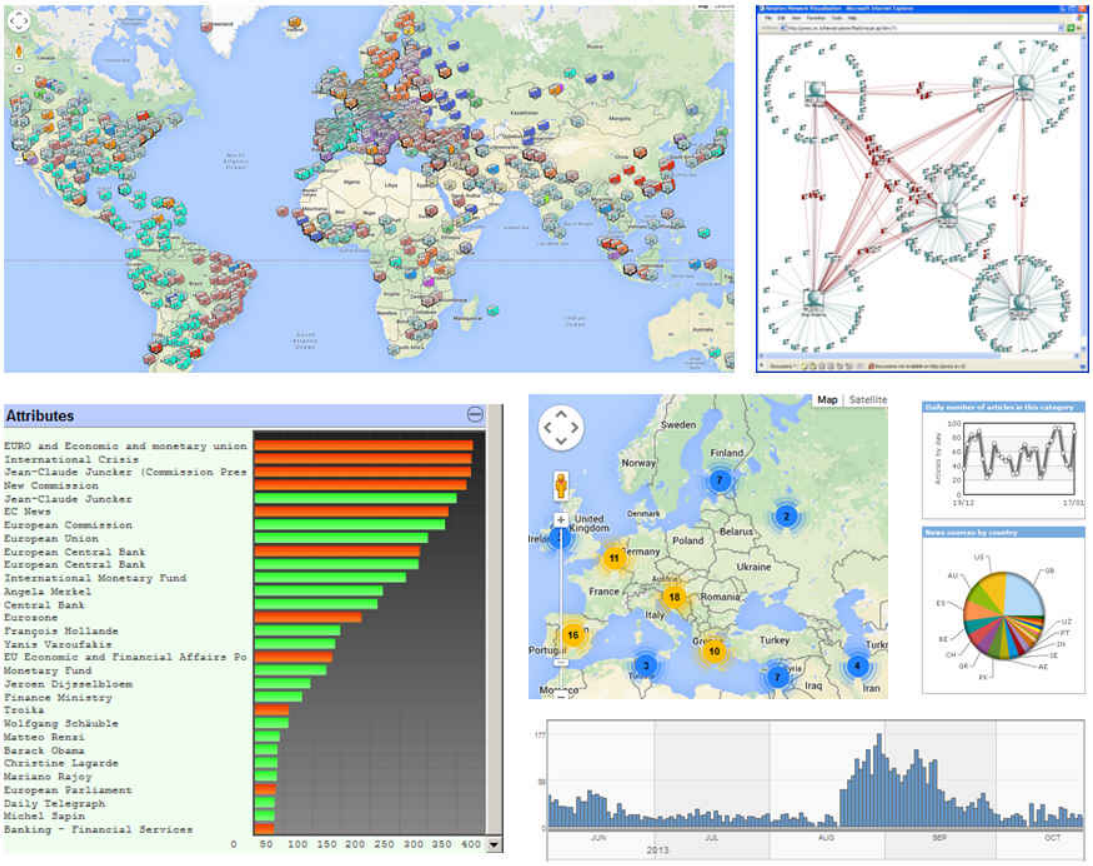

Figure 4. Various types of visualisation in EMM systems, showing geographical distributions, social networks, time lines, etc.

jor news of any given calendar day. NewsExplorer shows cross-lingual links so that users can jump to news on the same event or theme in any of the other 20 languages simply by clicking on the language link. It also shows the countries, persons and organisations most in the news that day, as well as the largest news stories this week and this month. The calendar on the left allows one to look at the news from any day since NewsExplorer's inception in 2004. Both NewsBrief and NewsExplorer contain many hyperlinks so that users can jump to the information they are most interested in. For each news cluster, there are also dedicated pages displaying all available meta-information for that cluster. For *stories*, i.e. related news clusters that are linked over several days or weeks and sometimes containing up to thousands of news articles, there are also dedicated pages showing the reporting intensity in a time line and displaying the cumulated meta-data for that story (see bottom-right graph in **Figure 4**).

What gives meta-information extracted from unstructured text so much power are the manifold ways of combining the data. For instance, in NewsExplorer, entity pages show the historically cumulated meta-information for any of over one million entities (**Figure 3**). It is possible to produce social networks of entities that fre-

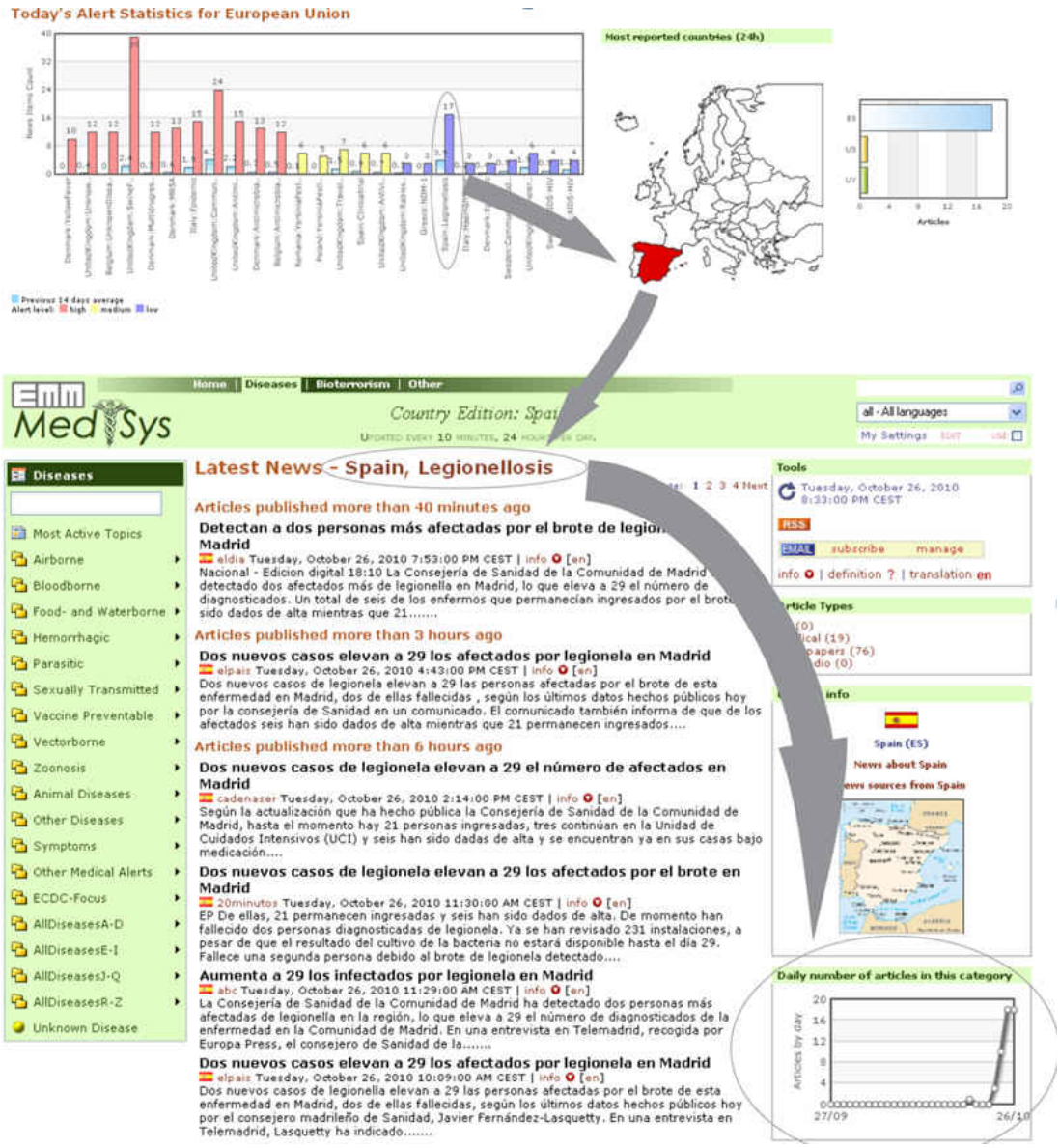

Figure 5. Early warning functionality in EMM-MedISys. The interactive alert statistics graph show today's biggest theme-country combinations and allows jumping to the related news items.

quently get mentioned together (see top right of **Figure 4**). Many more possible queries on combinations of news meta-data and their visualisation are possible, but the EMM interfaces do not allow to display all possible combinations of meta-information. They rather show the information aspects most requested by EMM's large user base.

2.4 Distributions, trends and early warning

Structured data allows various types of statistical analyses and of display (see **Figure 4**). These include simple distributions that can best be displayed as bar charts or

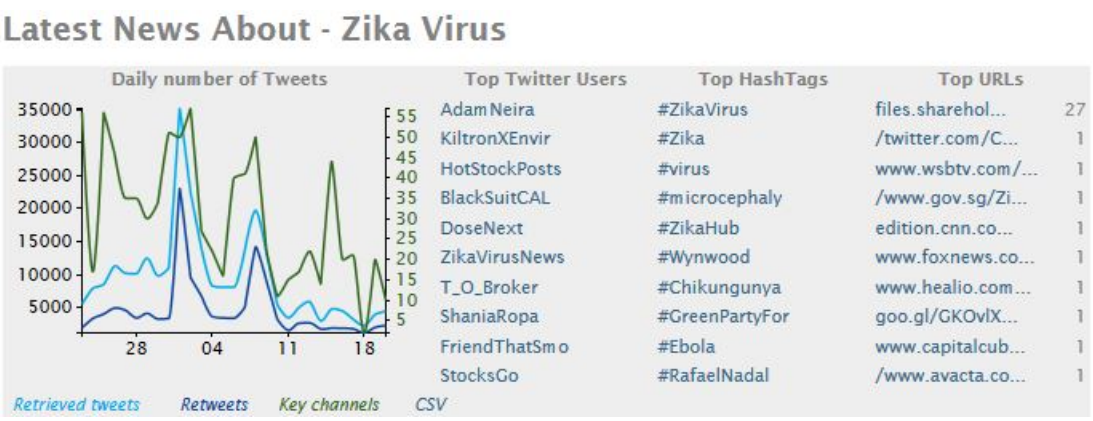

Figure 6. Display of information retrieved from Twitter on the Zika virus in EMM-MedISys: frequent users (influencers), main hashtags used, URLs mentioned in the tweets.

pie charts (e.g. distribution of categories of news in which a certain entity is mentioned) and time lines to show changes over time for any given information aspect. Networks (e.g. people frequently mentioned together) are best shown using network graphs. Anything relating to geo-locations, such as different event types being reported in different parts of the world, can be shown on a map.

A major advantage of processing large amounts of data is the fact that various statistics can be produced, displayed, and used to calculate averages and deviations. In EMM's *Medical Information System* MedISys (2016), this is used for early-warning purposes: MedISys informs users of sudden spikes regarding the number of news articles mentioning a certain health threat (e.g. TUBERCULOSIS) in combination with a certain country (see **Figure 5**). For that purpose, MedISys keeps track of the 14-day average number of articles being classified as being about a certain health threat *and* that mention a certain country (e.g. SPAIN *and* LEGIONELLOSIS). If the number of articles in this highly specific selection suddenly rises, MedISys sends out an alert. It is highly relevant that only the combination of threat and country are considered in order to also recognise *small signals*: if the same health threat is being mentioned much more regarding another country, then a rise by, say, only three articles regarding this specific country would remain unnoticed. Note that MedISys categorises news articles in all languages according to the same categories (i.e. HEALTH THREATS and COUNTRIES) so that the alert will be triggered even if the disease outbreak is mentioned in a language the users do not understand. MedISys users, which include the *European Centre for Disease Prevention and Control* (ECDC) and many other national and international Public Health monitoring organisations, focus their daily monitoring tasks on such MedISys early-warning alerts, and they additionally use MedISys to follow the development of threats already known to them. The early warning graph shown in **Figure 5** is clickable so that users will be able to view all related articles plus maps and statistics. For more information on MedISys, see (2016).

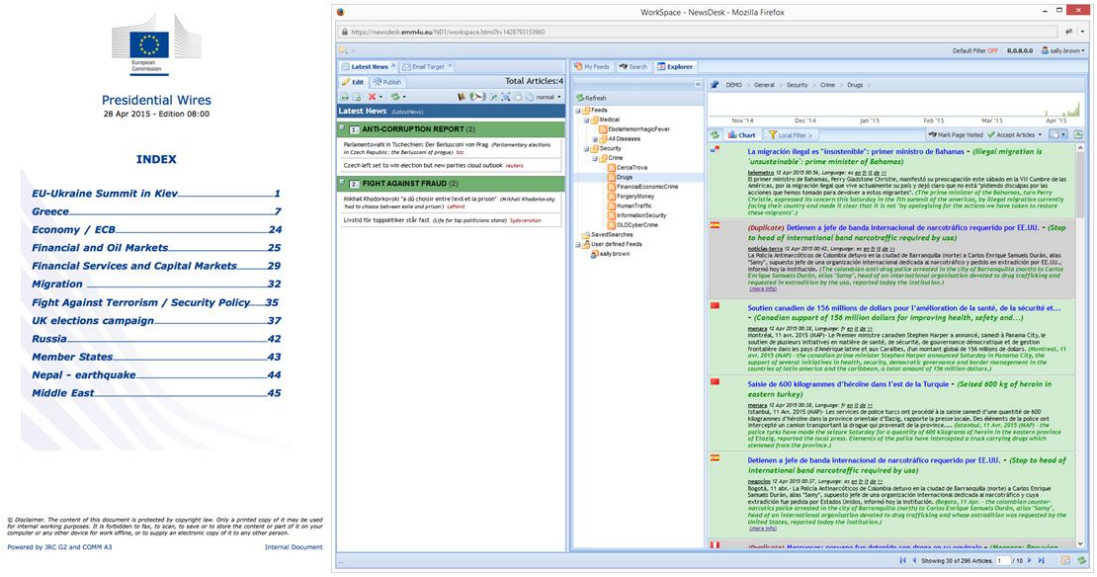

Figure 7. Interface of EMM's moderation interface *NewsDesk* (right), which allows selecting articles, defining headers and publishing a readily formatted in-house newsletter (left). The same interface can be used both for tradional media reports and for Twitter.

2.5 Social media analysis

Monitoring the social media is growing increasingly popular because it gives a better feel for what the population thinks without having to resort to opinion polls. Sites collecting posts on product or movie reviews and the more generic Twitter application are monitored regularly. In EMM, Twitter is currently being mined for several purposes, including: (1) For some threats, Twitter is searched for additional information, such as the most relevant hashtags/keywords, for the most active users and for links to photos or videos (see **Figure 6**). (2) The non-public *Citizens & Science* project has the aim to gauge the attitude of the population towards scientific and technical developments and to measure attitude and reporting intensity across countries. (3) EU institutions follow Twitter for posts on a small number of main subjects (e.g. the *Brexit*) and include their analysis results in the twice-daily newsletters (see Section 2.7). The idea is to complement the information present in news articles with information provided by the general public (2013). Social media posts are widely considered to be a rich complement to traditional media monitoring in the field of emergency and crisis management ( ).

2.6 Users of the Europe Media Monitor EMM

The EMM systems were developed by the European Commission's *Joint Research Centre* (JRC), whose mandate is to give scientific-technical support to the EU insti-

tutions, to EU agencies, to national authorities of the European Union's member states, as well as to EU partner countries and partner organisations. The latter includes the African Union, the Organisation of American States and many United Nations sub-organisations. Additionally, several thousand anonymous internet users access the public web pages every day. A small number of users have their own EMM installation, the majority access EMM output via the internet on the JRC's servers. EMM offers several hundred generic news categories. Many users additionally have their own customised categories. These are not publicly accessible as they usually overlap partially with other categories.

2.7 EMM moderation tool NewsDesk

EMM pre-processes the media data for the end users by categorising the news, by extracting meta-data and by establishing links. Many organisations have teams who use this as input for their own work of digesting the information and of turning it into in-house newsletters. They use the *NewsDesk* groupware application (see **Figure 7**) which allows the production of structured reports by creating sections, dragging and dropping individual articles into them, changing titles, etc. NewsDesk then uses customised templates that allow producing readily formatted newsletters at the push of a button. Analysts that are scanning the news feeds can send notifications to selected recipients by e-mail and SMS, and they can push a selection of news to update other systems like corporate web sites.

NewsDesk also provides support for a decentralised moderated information collection: Teams in the various member states manually select relevant local documents or field reports that can optionally include enclosures such as audio, video, or printed press excerpts. The initial manually assigned metadata is then enriched with the results of EMM's automated analysis and aggregated at headquarters to provide analysts with a more complete global view. Several media monitoring products (e.g. *Daily Press Reviews*) are generated and distributed in the headquarters and the member states.

From the data collected during the daily media monitoring activities, analysts can prepare periodic or ad-hoc media analysis reports: They start from a search in NewsDesk to identify the target dataset of news (period to be covered, countries, sources, main topics, etc.) and, with the *Media Analysis* module, they can review all the meta-data of the news articles and eventually enrich them with new sets of tags specific to each analysis exercise. Analysts might want to capture a wide range of information facets including: whether specific entities (people or organisations) are explicitly mentioned in the title; the type of news article (editorial, interview,…); the media used to publish that piece of news (such as TV/radio broadcast, printed press, on-line news, etc.); the media tier; tonality of the whole article and the perceived sentiment expressed specifically towards the mentioned entity; whether the entities are the primary topic of the article or they are just mentioned; what role the

Figure 8. Left: MyNews interface showing articles, meta-information and distributions on the current news selection. Right: EMM's mobile applications for iPad, iPhone and Android phones.

author gives to the entities (e.g. decision makers,…); and who are the reported speakers, if any. The moderated dataset is used to produce tables or graphs to be included in final reports.

2.8 Mobile applications and customised EMM views

Since 2013, EMM has also been available as an app for mobile devices such as telephones and tablets.[1] The mobile applications for Apple and Android are configurable (see the right side of **Figure 8**). For instance, each user can customise their own starting page, they can select the most important categories and information aspects, and they can select the languages they want to see. Upon popular demand, this configurability was then also introduced to EMM's desktop applications. This new EMM interface, which is called *MyNews* (see the left of **Figure 8**), keeps user preferences synchronised with the corresponding mobile app. It requires a login and is currently only available inside the EU institutions.

2.9 Event scenario template filling – Who did what to whom, where and when

Having an automatic system that collects the news, identifies the user-relevant article and displays additional automatically extracted meta-information together with each article is useful, but it still leaves a lot of work to the end user. For event types that may be dangerous to people (e.g. natural disasters, accidents, outbreaks of contagious diseases, etc.), the EMM team developed software that extracts more precise information from each article (2008). This includes the event type,

[1] EMM apps can be downloaded from http://emm.newsbrief.eu, from Google Play and iTunes.

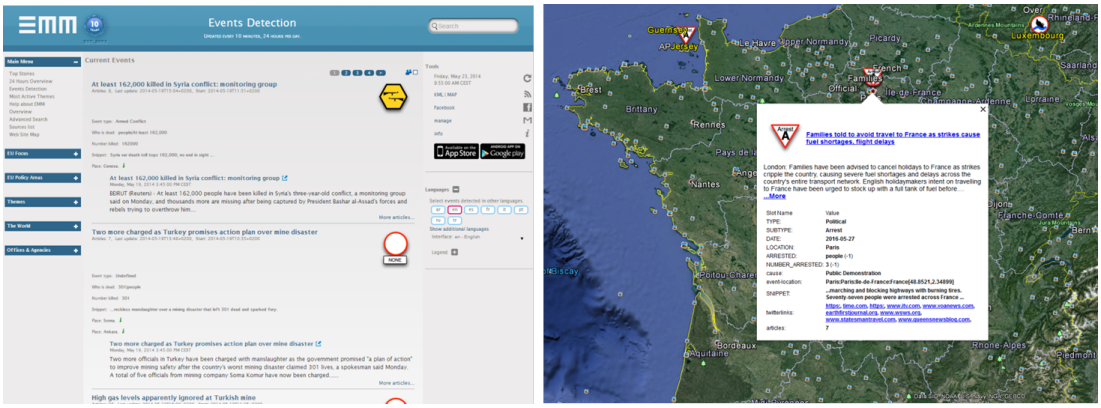

Figure 9. EMM's event recognition system, which presents event information in structured text format or on interactive maps.

information regarding the victims (dead, hospitalised, etc.), the perpetrators (who did it), the weapons used, the disease agent, the location, and more. The tool identifies events in eleven languages and displays the results in table format or on interactive maps (see **Figure 9**). Apart from the early-warning functionality when applied to live news feeds, a benefit of the software would be that it generates structured data that can be fed into a database so that long-term statistics and trends can be produced. However, due to the complexity of the tool and the resulting error rate, human moderation is in practice required.

2.10 Open Source Intelligence Tools (OSINT)

Some EMM customers have the requirement that they can analyse their own documents or documents that are of other types than the media. The EMM team has therefore developed a desktop application called OSINT ([redacted] 2010) that allows ingesting locally stored text collections or harvesting relevant texts from the internet via a web search interface. The documents are downloaded and converted from HTML, PDF or Microsoft-Office formats into a machine-readable text format. Users can then apply software that recognises entities and relations between them. The entity types include person and organisation names, email addresses, URLs, VAT numbers, vocabulary from user-provided term lists, and more. The results can be displayed in graphs or in table format (see **Figure 10**) and they can be saved for future reference.

3. Other media monitoring services and tools

EMM is not the only system that allows monitoring the media. A web search for the term *media monitoring* yields an ample amount of companies offering services to

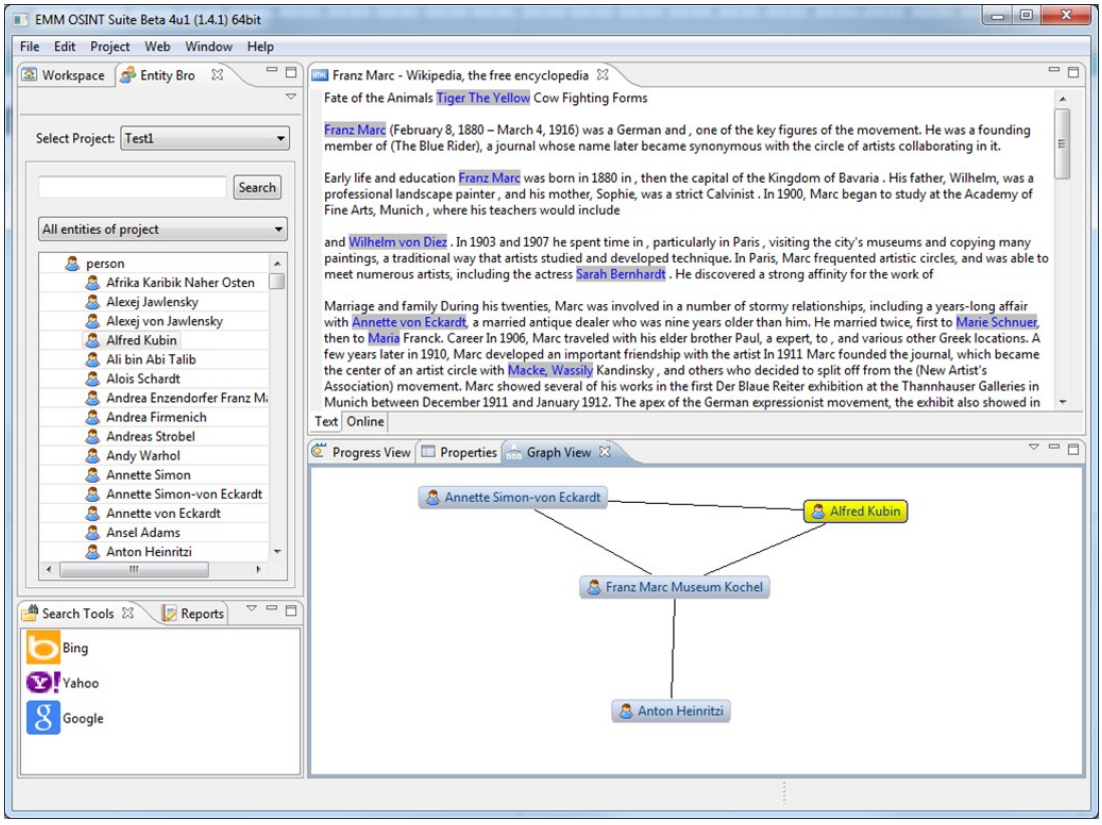

Figure 10. Data presentation in the Open Source Intelligence Suite OSINT.

monitor the printed and also the social media.[2] While the technical details of such commercial services are often not disclosed, the websites seem to reveal that these companies mostly collect large volumes of online news (and partially also Social Media posts), that they allow users to formulate queries consisting of Boolean search word combinations, that they deliver the resulting text collections and produce statistics showing media volume and changes over time. The overwhelming majority of services are offered for English language news, but there are some that gather news in large amounts of different languages. As the query formulation is effectively done by the client and functionality mainly consists of filtering those news items that contain the query words, offering a multilingual service is not a challenge. A noteworthy exception seems to be SiloBreaker[3], which offers the monitoring of the news and other text sources in twelve languages and which addition-

[2] To mention just a few companies offering media monitoring services or solutions: LexisNexis (lexisnexis.com), L'Eco della Stampa (Ecostampa.it), Kantar Media (kantarmedia.com), Cision (Cision.com), Visual Box (visualbox.it), Cyberalert (cyberalert.com), Africa News Monitoring (africanewsmonitoring.com), Infojuice (infojuice.eu), Mention (mention.com), Meltwater (meltwater.com), Selpress (selpress.com), and more. All URLs mentioned here were last visited in the week 19-23.09.2016

[3] See www.silobreaker.com.

ally extracts entities, produces time lines and visualises entity relations. SiloBreaker seems to use Machine Translation to translate texts into English and to then apply English language text analysis software tools to the results. According to the *native language hypothesis* ([redacted] 2004) – and also in our own experience – text analysis results are better when applied in the native language as information gets lost during the translation process. Especially names frequently get distorted during automatic translation ([redacted] 2012). However, it is difficult to get a clear idea of what happens behind the scenes in these commercial systems as scientific publications are rarely available.

In comparison, our own system EMM has over one thousand ready-made multilingual categories. EMM extracts information from the news articles (entities, quotations, events, social networks). EMM puts order into the results by clustering related news items and it links related news over time and across languages. On the other hand, EMM does not serve the commercial public so that users cannot normally create their own categories.[4]

Search engines also often provide multilingual news search functionality. *Google News*[5] categorises the news articles and it groups related items so that users can directly access different viewpoints about the same event. *Bing News*[6] and *Yahoo News*[7] offer various news categories, but there is no evidence that they group related articles. With all three services and with EMM, users can customise languages, news sources and subject areas they want to see. The web version of EMM does not allow users to personalise the interface by selecting their own set of categories, but EMM's public mobile applications do and that functionality is also available in the EC-internal EMM system called *MyNews* (see Section 2.8). Unlike EMM, the three search engines do display photos, but they do not display extracted meta-information such as entities mentioned, quotations extracted and other categories to which an article also belongs. They also do not allow the display of cross-lingual links between related news and they do not display timelines like the ones in **Figures 1, 4, 5 and 6**.[8]

There are a number of more exploratory and/or academic projects dedicated to media monitoring and news analysis. In 2015, IBM Watson presented its English language news analysis system called … *News Explorer* ([redacted]

[4] The spin-off company OSVision (www.osint.com) has licensed EMM tools and may serve such clients.

[5] Available at news.google.com. See also https://support.google.com/news/answer/106259.

[6] Available at www.bing.com.

[7] Available at www.yahoo.com/news.

[8] Google News occasionally shows small numbers of hyperlinked person, organisation or place names next to the news item and – for small numbers of English language news – there is also a short-term time line.

2015).[9] It collects news items, extracts names of entities and 'topics' from them and provides various ways of searching and of visualising the data. A timeline, for instance, shows the number of articles that contain the user-provided search words and a relationship graph displays all related entities and articles mentioning a certain entity, thus facilitating the discovery of direct or indirect links between entities.
The *NewsReader* (to appear; 2016) is an EU-funded research project running from 2013 until 2016.[10] It aims at analysing large volumes of historical Dutch, English, Italian and Spanish news by extracting events based on entities and their relationships and by storing the results in a structured database. Important aspects are the completion of event descriptions, the deduplication of the highly redundant news information, the storage of the event information in RDF format (triplets) and retrievability. The NewsReader team has focused their efforts on six specific subjects (e.g. automotive industry, criminal networks).[11] The software is freely available for download.
Similarly, the large-scale GDELT project[12] (2013) aims at producing a historical event database by analysing news archives and newly incoming articles in "over 100 languages". For 65 languages, part of the ingested documents gets automatically translated and then analysed by an English language event extraction system (note our reservations towards the analysis of automatically translated text in Section 3). The system can be queried online and even downloaded. For an assessment and an evaluation of the system, see . (2013).
The *BBC News Labs*[13] explore various ways of analysing the media flow automatically, based on an automatic enrichment of news data. The objective is to give journalists faster access to information (hence saving time and money), giving access to media data not currently well-indexed, to allow more transparency across the BBC's news processes, etc. They are involved in a range of different projects with academia.[14] For instance, the *Summa*[15] Project aims to detect trends and the evolution of story lines across many languages. Unfortunately, it is hard to find details or publications on their work.
The projects mentioned in this section include some of the major efforts, but the list is by no means exhaustive. The abundance of work shows the interest in automated content analysis and media monitoring and the expected benefits for advanced analysis of world events. The *Europe Media Monitor* was one of the first multilingual and openly accessible systems, with EMM-NewsBrief coming online in 2012 and EMM-NewsBrief in 2004.

---

[9] Available at https://news-explorer.mybluemix.net/.
[10] See http://www.newsreader-project.eu
[11] See the brochure at http://www.newsreader-project.eu/files/2012/12/NWR_Brochure_2015.pdf.
[12] See http://www.gdeltproject.org.
[13] See http://bbcnewslabs.co.uk and the tab named 'Projects'.
[14] See http://bbcnewslabs.co.uk/categories/academic-research/.
[15] See http://bbcnewslabs.co.uk/projects/summa/.

4. Text analysis quality and usage warnings

Analysing text automatically and converting it into disambiguated structured data is not a trivial matter. The question therefore arises: How good and reliable is the computational analysis? It goes without saying that automatic systems will always make mistakes, but it is not easy to quantify the error rate because so many different components are involved and not all tasks are equally difficult. The EMM team has published formal scientific evaluation results for almost all of its components. We will provide a few generic answers below and we refer to (2009) and the publications mentioned therein for details.

4.1 Reliability of the automatically extracted information

The best way to get a feel for the analysis quality is to look at the openly accessible EMM online applications[16] with a critical eye. Website visitors will see that most EMM-generated information is correct, but there will also be some obvious errors: Regarding news classification, articles will only be grouped into a certain news category if they contain the relevant words and word combinations, but some articles will nevertheless not pertain to that class. This can happen, for instance, when the news talks about a movie or a song about that subject instead of about a real event. The reader can find a formal evaluation in . (2008). Regarding the automatic grouping (clustering) of related news articles, the reader will occasionally find some articles that do not pertain to the same news story, and it may also happen that there are two clusters about the same story, but clustering mostly works very well. Regarding entity recognition, most frequently found person names will be correctly identified, but occasionally, other words may be marked as being a person (e.g. *Dunya News*) because the words are written in uppercase and because the context of that name misleads the computer program. For a formal evaluation, see (2009). Regarding geo-location recognition, places are occasionally wrong because many locations are homographic with common words (e.g., there are locations called *Split, And, For, Bush*, etc.). Sometimes, the system picks a location from the text that is circumstantial, instead of identifying the name of the place where the event happened. See (2006) for more details. Regarding machine translation, due to the complexity of the task, translation results should be expected to be good enough to give the reader an idea of the contents and the relevance of the text. However, they cannot be expected to be grammatically correct. (2012) showed that EMM's statistical machine translation system produces state-of-the-art performance, but it is less good than translations produced by the major developers of translation software. Regarding quotation extraction, the results are almost always correct (2007). Extracted event information, however, frequently contains partial errors because it is composed of

[16] See http://emm.newsbrief.eu, http://medisys.newsbrief.eu and http://emm.newsexplorer.eu.

so many different information aspects: event type, location, actor, victim, weapons/means/disease, etc. (2008).
Regarding trends and the related early-warning functionality, the situation is slightly different because trend detection additionally gets influenced by the intensity of reporting, which is not controlled by EMM. It is a media fact that reporting intensity is not necessarily directly related to the objective importance of the event. Furthermore, the selection of the news sources monitored may not always be perfectly balanced, which would also lead to a less-than-objective result. EMM aims to monitor all major news sources of a country, but it also tries to collect articles from around the world in English and other widely-spoken languages. Some news sources are added on the request of EMM users. It is not always possible to verify whether the news sources are balanced regarding political or other biases.

4.2 Caveat – pitfalls to avoid

Media monitoring is not identical to reality monitoring. Media reporting is biased as it is influenced by the political and the geographical bias of the news sources. As we have seen in the previous section, automatic text analysis additionally is error-prone. It is useful to keep these facts in mind to make best use of automatic content analysis systems. Especially systems showing aggregated results such as trends or early warnings should have the following features in order to be useful and to avoid that users are mislead: The users must be able to verify the data that led the system to its conclusion, i.e. they should offer a drill-down functionality; and they should show the list of news sources included in the analysis. System evaluation results should be available measuring the accuracy of the analysis quality. These results should have been produced on the same type of data as the data on which it is applied. For instance, if a sentiment analysis system was trained and evaluated on subjective social media data, but it will be applied to seemingly more objective news data, we can expect that the performance will be considerably lower. In summary, automated systems should be accompanied by transparent evaluation results, they should offer a drill-down functionality so that users can explore and judge the original data, and users should be aware of the intrinsic bias of such systems, as well as of the expected error rate.

4.3 Automated versus human content analysis

Automated content analysis does work, to a certain extent, and we know for certain that EMM users find it useful. Computer programs can sieve through huge volumes of text, making it easy to get a fuller picture of what is going on and detecting trends and distributions that are not easily visible to the human eye. However, this does not mean that human analysts will no longer be needed. To the contrary: the crucial steps of verifying the data, of selecting the most relevant pieces of infor-

mation and of drawing conclusions are best performed by people. Human analysts do not even have less work than without the existence of computers because available data volumes keep growing and there is a clear benefit to digesting a lot of it. The best results are achieved by exploiting the best capabilities of both people and computers: the speed and consistency of large data processing by machines and the intelligent digestion by human analysts.

5. Transparency and responsibility

How does the *Europe Media Monitor* software fit the theme of this book? What initiatives are there to take responsibility in order to increase transparency? EMM was developed for the European Institutions, but it was decided to make most functionality openly and freely accessible to the public. EMM is a meta-news site. It gives an organised and unbiased overview of information found on many different news sources, including across languages and countries. It is useful for readers who want to look beyond their favourite newspaper, who want to get informed of what other sources say and how the same news story is presented in other countries. We believe that understanding the focus and the viewpoint of neighbouring countries – and even of far-away countries that may have highly different views – is a good step towards higher levels of democracy.

Another big step towards more transparency and democracy was the EU decision – many years ago – to give free access to EU documents via the internet,[17] allowing citizens to search and retrieve EU legislation and other official data from the EU's Official Journal, and more. EU decision takers also became aware that other public data can be highly useful for businesses so they decided – already in 2003 – to encourage that data from EU institutions and from EU member states should be made freely accessible to the public, including for commercial purposes.[18] In 2014, the *EU Open Data Portal* was launched, from where many types of public data can be downloaded for free.[19]

In order to build the *Europe Media Monitor*, the EMM team needed to develop or acquire large volumes of multilingual text data and language technology software tools. Starting in 2006, the team released highly multilingual data and tools in order to support commercial and scientific Research & Development teams in their endeavour to build multilingual Language Technology solutions ([redacted]

[17] See, e.g. https://europa.eu/european-union/documents-publications/.

[18] See Directive 2003/98/EC of the European Parliament and of the Council on the re-use of public sector information (http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32003L0098:EN:NOT), as well as Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission Documents (http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:330:0039:0042:EN:PDF).

[19] Accessible at https://data.europa.eu/.

2014).[20] The intention was to speed up developments in the field of automated text analysis and especially for cross-lingual applications such as machine translation and cross-lingual information access. The availability of such software is expected to help citizens communicate and to help intra-EU trade grow.

6. Conclusion and future developments

Automated media monitoring is a well-established practice inside the European Institutions. Up to 300,000 online news articles per day in over seventy languages are being processed by the *Europe Media Monitor* EMM. EMM groups related articles and links them over time and across languages; it classifies the news into main categories; filters news according to specific user interests; extracts information; produces statistics and visualises the results and trends. Selected information from social media streams complement the news data: Information is extracted from Twitter and links to main images and videos retrieved from the Tweets are offered to the user.

The hundreds of institutions using EMM daily and the thousands of anonymous online users show that the system is perceived to be useful. However, people need to know the limitations of automated systems and be aware of the possible pitfalls. All automatic text analysis applications will make mistakes and the results will very much depend on the input data used. Any automatically generated statistics, trends and early-warning messages should not be taken for granted; they should be looked at with great care. For that purpose, it is important that the software allows drilling down, i.e. look at the data underlying the analysis. Is the selection of sources biased? Are the individual analysis steps of an acceptable quality? Is the system transparent? If these conditions are satisfied, the machines' capacity to process large volumes of data combined with the human analysts' capability to draw conclusions is extremely powerful.

While – to our knowledge – EMM is one of the world's largest news monitoring system, considering the amount of languages covered and the number of linguistic analysis tools being integrated, it could do much more. EMM extracts huge amounts of meta-data, allowing in principle a large variety of aggregated views. However, EMM only shows some of the many possible views while further interesting data combinations are not available and there is no interface allowing users to make their own meta-data queries on the data. Building a data representation and an interface that would allow users to do this is a major task that yet needs to be tackled. Time series analyses would be particularly insightful. How have the major media themes developed over time, and what are the differences across countries. Which countries are the trendsetters that dominate the international media landscape? What are the main subject domains and the main concepts mentioned in the

[20] The highly multilingual resources and software tools can be downloaded from https://ec.europa.eu/jrc/en/language-technologies.

context of certain persons and organisations? Who are the other entities mentioned in the same context, and how are these changing over time? This type of analysis is not currently available in EMM. Linking almost any data extracted from the news over time and looking at changes and trends will be extremely interesting, especially when comparing this trend data across different countries. The future is bright, but there is lots to do.

Having the responsibility to create more transparency? Yes. The focus of this book is on creating more transparency in private organisations, but we are convinced that good information from the media on fields of public interest is generally beneficial to the society. Having a more pluralistic view on matters is informative, and especially reading about the viewpoint of other countries on the same themes. It will probably generate more trust, but in any case deeper knowledge and understanding.

References

(2013). Detecting Event-Related Links and Sentiments from Social Media Texts. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 25-30.

(2015). Social network relationship mapping. US patent, Publication number 8977979 B2, https://www.google.com/patents/US8977979.

(2010). Desktop text mining for law enforcement. Intelligence and Security Informatics (ISI), Conference Proceedings, pp. 138-140.

EBS Surveillance Work Group (2016). Event-Based Surveillance During EXPO Milan 2015: Rationale, Tools, Procedures, and Initial Results. Health Security, Volume 14, Number 3, 2016 Mary Ann Liebert, Inc.

(2004). Language-specific Models in Multilingual Topic Tracking. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 402-409.

(2013). GDELT: Global Data on Events, Location and Ton, 1979-2012. ISA Annual Convention, Vol. 2. No. 4.

(2011). Online news event extraction for global crisis surveillance. In: Nguyen N.T. (ed.): Transactions on Computational Collective Intelligence V, Springer LNCS series 6910, pp. 182-212. Springer, Heidelberg.

2011). Exploring the usefulness of cross-lingual information fusion for refining real-time news event extraction. Proceedings of the 8th International Conference Recent Advances in Natural Language Processing (RANLP'2011), pp. 210-217. Hissar, Bulgaria, 12-14 September 2011.

(2006). Geocoding multilingual texts: Recognition, Disambiguation and Visualisation. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), pp. 53-58. Genoa, Italy, 24-26 May 2006.

(2007). Automatic detection of quotations in multilingual news. Proceedings of the International Conference *Recent Advances in Natural Language Processing* (RANLP'2007), pp. 487-492. Borovets, Bulgaria, 27-29 September 2007.

(2016). Building Event-Centric Knowledge Graphs from News. Journal of Web Semantics. ISSN: 1570-8268. 37-38, 132-151.

(2009). Cross-lingual Named Entity Recognition. In: (eds.): Named Entities - Recognition, Classification and Use, Benjamins Current Topics, Volume 19, pp. 137-164. John Benjamins Publishing Company. ISBN 978-90-272-8922 3.

(2012). A survey of methods to ease the development of highly multilingual text mining applications. Language Resources and Evaluation, June 2012, Volume 46, Issue 2, pp 155–176.

(2009). An Introduction to the Europe Media Monitor Family of Applications. In: (eds.): Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR'2009), pp. 1-8. Boston, USA. 23 July 2009.

(2008). Text Mining from the Web for Medical Intelligence. In: (eds.): Mining Massive Data Sets for Security. pp. 295-310. IOS Press, Amsterdam, The Netherlands.

(2014). An overview of the European Union's highly multilingual parallel corpora. Language Resources and Evaluation Journal (LRE).

(2012). The use of social media within the global disaster alert and coordination system (GDACS). Proceedings of the 21st International Conference on World Wide Web, pp. 703-706.

(2008). Real-time news event extraction for global crisis monitoring. International Conference on Application of Natural Language to Information Systems. Springer, Berlin Heidelberg.

(2012). ONTS: "OPTIMA" News Translation System. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 25–30, Avignon, France, April 23 - 27 2012.

(to appear). Newsreader: how semantic web helps natural language processing helps semantic web. Special issue Knowledge-based systems, Elsevier. ISSN: 0950-7051.

(2013). Comparing GDELT and ICEWS Event Data. Analysis 21, pp. 267-297.