cultural science

# Reading News Data

Nikoleta Daskalova
*St. Clement of Ohrid University, Sofia, Bulgaria*

## Introduction

Information overflow is both inspiring and depressing. Inspiring is more or less the easy access to various information and communication resources, which in turn facilitate the exchange of ideas, knowledge and creativity. However, the impossible fathomability of infinite information space creates a feeling of depression and anxiety. Thanks to digital technology, the speed at which information is generated and distributed significantly exceeds the speed at which it can be perceived. The more the hypertext ocean of the web is filled with content, the more impossible it is to be 'tamed' by human senses.

In the ambivalent nature of information super-abundance, technological optimism and technological pessimism constantly compete. In fact, the two perspectives on the role of technology in the world of people have always been in conflict, but are now strongly intensified with the evolution and spread of the internet.[1] This text looks at a conditional technological optimism, aiming not at postulating utopian aspirations, but at illustrating how scientific and technical elites do not lose their desire to overcome depressing complexity by seeking bold optimization solutions.

The focus of this paper is on an innovative technological system for processing online news, namely the publicly available Europe Media Monitor (EMM).[2] The interest in this monitoring tool is multifaceted. In most general terms it is interesting to trace the technological solutions with which computer science specialists are trying to discipline the information flow. However, EMM is also interesting as a powerful tool for understanding reality on the basis of statistically processed news databases. This study provides examples of how EMM-enabled media content processing options may be used as a basis for further detailed analysis. Of importance are also the social and institutional intentions behind the development of such a system: what are the motives and the uses associated with such an intersection between news and intelligent software.

## The Europe media monitor system: Basic applications

EMM is an electronic media monitoring system, developed by the European Commission's Joint Research Centre (JRC). The system is designed to gather reports from news portals and to analyse the extracted information in near real time. The analysis consists of automatic processes such as classification of the news articles, generation of news summaries and clusters, named entity recognition (people, institutions), geolocation of events, etc. The EMM engine gathers about 100,000 news articles per day in approximately 50 languages (status March 2010 – Steinberger 2010) from about 2,500 hand-selected web news sources from around the world. It is a sophisticated news content management system which monitors the 'live web, i.e. the part of the web that has ever changing content'24 hours a day. [3]

The ambition to monitor the 'live web' is typical not only for the JRC. In general, EMM can be analyzed together with other popular news aggregators such as Google News, for instance. In fact, there are a few basic similarities between EMM and Google News: both engines process (gather and cluster) news out of web sources in different languages; news items are arranged into categories; headlines and top stories are selected by computer algorithms; both aggregators recognize and compile quotes; both offer news archives and visualize news timeline among other features. Or, to put it in on a broader context, such news aggregators promote a new culture of reading information: capturing and putting order of the endless news flows based on algorithmic and statistical rules. In addition, however, monitoring the 'live web' can mean more than news aggregation. A further ambition that stays behind the EMM project is not only to gather but to perform more complex analysis of the news data: the EMM project is designed to derive meta-knowledge from news stories through number of processes including multilingual and cross-lingual analysis. Such more profound cross-lingual news analysis systems are 'notable exceptions' according to the JRC experts (Steinberger 2010).[4] Indeed, a publicly visible cross-language news analysis system on a scale as large as the EMM project is hard to find – a fact that makes this project rather intriguing from a researcher's point of view.

To begin with, how is the EMM system designed? The monitoring results are specified into four public web applications: NewsBrief, NewsExplorer, MedISys and EMM-Labs.

NewsBrief provides live monitoring and breaking news detection with updates every 10 minutes. The application groups all articles related to a given topic into clusters and the largest cluster (i.e. the one with the most articles registered) is presented as the current top-ranking news story, the others listed in a descending line. Based on statistical algorithms the larger and rapidly rising size clusters are automatically classified as breaking news (Steinberger, Pouliquen & van der Goot 2009). NewsBrief is focused on short term trends detection, early alerting and up-to-date category-specific news display (ibid).

NewsExplorer uses a number of tools for a more in-depth data analysis and longer term trends identification. The application maintains an archive of daily reviews of the news on a given calendar day for 19 separate languages. Among the text mining tools applied by NewsExplorer is the detection of the most popular themes, people, organisations and locations in the news on a daily basis. There are language and country filters as well as possibilities for comparisons 'across languages and over time' (as the website slogan says).

MedISys is the sub-system for specialized medical information. The application displays only health-related articles and events based on continuous monitoring of about 250 specialist medical sites in addition to the generic EMM news. MedISys filters the information into hundreds of categories (e.g. diseases, symptoms, chemical agents, etc.) on the basis of pre-defined combinations of keywords. Additionally, the application keeps track of the statistics for each category for the identification of breaking news – that is to say public health threats.[5]

The EMM-Labs application applies interactive visualisation tools for further analysis of the data gathered by the EMM engine. Subject to visualisation are, among others, article summaries, automatically extracted social networks, real time violent events and disasters.

## Data processing Methods

Behind the so sketched 'faces' of the Europe Media Monitor lie a number of interconnected software solutions. The EMM project was started in 2001 and the first release of the system was in July 2002. However, the construction of such a system is not an endeavour which starts from scratch. The EMM matrix integrates tools previously developed within the Joint Research Centre, the main filed of research being computational linguistics.

In a presentation from 1998 Ralf Steinberger, the language technology project manager at the JRC and one of the leading scientists engaged in the EMM development, reports that the Joint Research Centre is in a process of building up a language engineering infrastructure. The current activities in this period are centred on automatic document indexing, automatic subject domain recognition, automatic language identification, document clustering and derivation of meta-knowledge from texts (Steinberger 1998).

In the years that followed, and from 2001 onwards related particularly to the development of the Europe Media Monitor, the initial text analysis instruments were further refined and augmented. The possibilities for multilingual text mining and cross-lingual navigation within the original news sources and the data extracted have been gradually expanded. The number of the monitored languages has been growing in accordance with the language diversity within the European Union.[6] The instruments for themes detection, summarisation of large volumes of redundant information, social network generation based on information extracted from multilingual news, and data visualization, among others, have also been improved.

To get a further understanding of how the EMM system is constructed, let us outline, in a most simplified manner, some of the main mechanisms behind the monitoring process.

The automatic categorisation and grouping of related news items into clusters is based on hierarchical algorithms. Through web crawling each monitored document is automatically keyword indexed, keywords and their relative importance being calculated with statistical formulae. Thus each text is represented by a ranked list of keywords, which, in a next step, allows for software calculation of similarities between texts. A clustering tree is built on keyword-based binary principle. For each cluster, the most typical article is identified, i.e. the one that is most similar to the cluster's centroid and its title is used as a title of the whole cluster. The clusters are then ordered by the system according to their size. As a result the users see a hierarchical representation of news themes detected from different news sources. Additionally, each cluster can be compared with previous days relevant clusters so that tracking of topics over time can be achieved (Steinberger, Pouliquen and Ignat 2005; Pouliquen et. al 2004).

Another important practice (and, as previously highlighted, advantage of the EMM project) is the cross-lingual linking among the different media sources. A useful tool in this process is the multilingual classification thesaurus Eurovoc. The thesaurus covers all official EU languages and consists of 6000 concepts hierarchically organised in a list.[7] The EMM generated clusters are mapped onto the thesaurus and the Eurovoc terms are used as descriptors/keywords for the tracing of similarities between languages. Thus cross-lingual cluster comparisons can be made.[8] The mechanisms of cross-lingual linking are based also on already made name lists of known persons and places in different languages, as well as on gazetteers, country profiles and the monolingual representations of the clusters. In this process of complex linking, the Eurovoc-component is the one with the highest impact.

Machine recognition of named entities is another important instrument integrated in EMM. The precise identification of people in the news relies on a rich knowledge base, i.e. pre-defined multilingual spelling variants. For instance, for the name of Vladimir Putin there is a collection of 100 variants, for Barack Obama – 58, David Beckham – 24. A new person name is tracked by the use of lists of first names and trigger words like 'Mr.', 'minister', 'playboy', 'actress' etc. Each new name is stored in the database together with its context. The lists of names are daily updated (Pouliquen, Tanev and Atkinson 2008; Pouliquen et al. 2005). A main problem in this respect is the identification of people having identical names. In such cases the software seeks for additional contextual information but very often the names are just merged. To solve the problem, manual human intervention can most properly identify the person in the original news text.[9] Such manual verification is actually stimulated by the EMM interface which automatically links the name to the group of articles in which it was registered. In this sense, 'human intervention' means nothing more than looking into these articles. However, this problem points to two noteworthy aspects of reading news data. First, a most accurate reading requires that the user has a minimum of preliminary knowledge about the public contexts he/she investigates.[10] Second, it raises the question of reliability of the software generated data. 'All automatic text analysis tools make mistakes, but tests have shown that the various NewsExplorer components involved mostly produce correct results', says the NewsExplorer site.[11] Such a clarification is generally indicative of the intricacies of data interpretation: on the one side, computer algorithms are designed to avoid human bias and inability to manage information overflow; on the other – there is the risk for nonhuman distortion of the picture. This implies the principal necessity of a highly sensitive use of analytic software instruments.

In addition to the EMM features, the NewsExplorer application keeps profile pages for the persons in the news. Each such sub-page contains information about: variants of the person name in different languages; frequently mentioned titles or descriptions mentioned next to the name over time; the latest news clusters in which the person has been mentioned; quotes from and about the person; associated people. The extraction of these data is supported by the statistical processing of clusters. Detection of quotations from/about a given person is achieved through algorithms based on morpho-syntactic analysis (cf. Pouliquen, Steinberger and Best 2007).

A key component of the EMM system is the structuring of social networks by the extraction of relationships between people derived automatically from news articles. There are two main methods for extraction of such relationships: through simple co-occurrence statistics or through more complex linguistic and syntactic patterns (Pouliquen, Tanev, Atkinson 2008). The second approach is useful for the detection of specific relationships such as meetings, contacts, support or criticism as well as family relationships.

The design of automated media monitoring within EMM is supplemented with other processes, including: geographical entities recognition through a world coordinating system; data visualisation through interactive figures and maps; use and integration of external tools and resources such as Wikipedia, WorldKit, GoogleEarth; real time notification of users for alarming events, etc. To sum up, EMM is constituted by a chain of algorithms combining both statistical and semantic and syntactic principles.

## What the news data say

The wide range of technological solutions integrated into the EMM family of applications results in a wide scope of functional deployment of the system.[12] This text will try to extend the notion of EMM by discussing how the system can be used in practice and developed analytically based on the experience of the Media Monitoring Lab (MML), a civil and research project within the Media Democracy Foundation in Sofia, Bulgaria.[13] MML has been using the EMM publicly available databases to analyze Bulgarian news sites. The focus of research has been on the data generated by the NewsExplorer application.

NewsExplorer is a useful tool for media analysis mainly because of two of its functions: generating a daily record for the most common topics in the news, summarized in hierarchical clusters, and generating a list showing which people have been most mentioned in the news on a given day. The great convenience of the application is that this type of information is stored in an archive calendar which users can easily navigate. Therefore, the accumulated database shows not only what the leading news and people had been online on a given day of the month, but outlines short term and long term trends in the media agenda and media representations.

In post processing the news clusters could be subjected to an additional classification in categories. Kutseva (17 Sept. 2009) analyzes Bulgarian news sources clusters on the basis of six categories: political, accidents, international, economic, sports and miscellaneous. When comparing the results for a period for several months persistent trends and dynamics in the structure of media coverage are clearly visible.

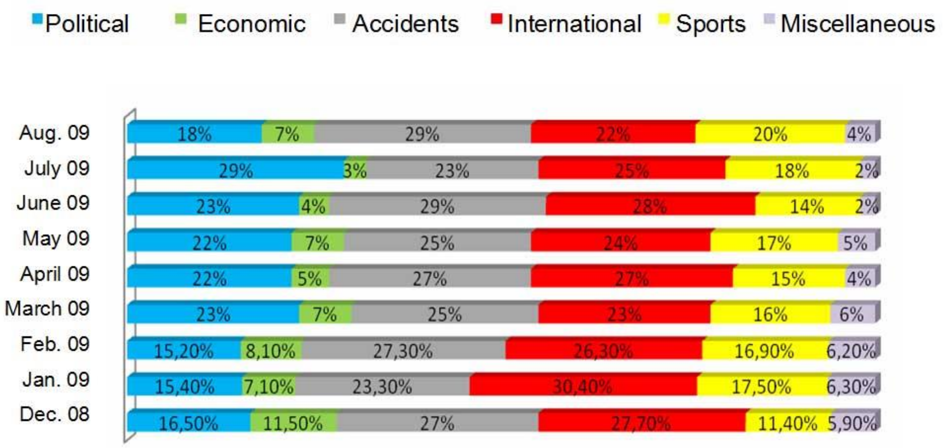

*Figure 1*. The news categories in comparison, Bulgarian news sites (Dec. 2008 – Aug. 2009). Source: Kutseva (17 Sept. 2009).

For example, data from Figure 1, processed by Kutseva, are illustrative on a macro-level of the typization of news posted on Bulgarian sites in a specific public context: before, during and immediately after elections (the European Parliament elections in June 2009 and the national elections in July 2009). On the basis of the infographic conclusions can be drawn about: the direct relationship between the change in political status quo and the displacement of news layers, reflected as a boom of political news during the elections for national parliament; the strong orientation of online news towards events from global 'hot spots' and the media taste for accidents and disasters. Such observations are respectively indicative of the illusion of co-participation and awareness, maintained by online news streams, and of the media preferences for news in the consumption of which users can project their negative emotions and fears and which do not require knowledge of the context (Kutseva 2009). The strong editorial inclination towards events in world politics is also confirmed by the secondary classification of the most frequently mentioned subjects. If we separate the names from the sphere of politics from that list, we would find consistently high presence (an average of 40% for the months from December 2008 to October 2009) of people from the global political scene (cf. Daskalova, 20 Nov. 2009). This bias towards international news is indicative of the origin of content sources — the fact is that global agencies are an important information resource for Bulgarian media and the internet greatly facilitates the supply of non-native content. Global agencies and foreign sites turn out to be a compensating content generator especially in periods of decline of political activism at the local level (e.g. around holidays like Christmas, New Year and Easter).

Post-processing and visualization of daily lists of the most frequently mentioned people in the news generated by NewsExplorer makes it possible to bring to light the seemingly invisible characteristics of the information space. For example, if we add all the names from the People column generated by the system for a period of one month, we would get an array of about 350 unique names that can be presented as a bubble chart.

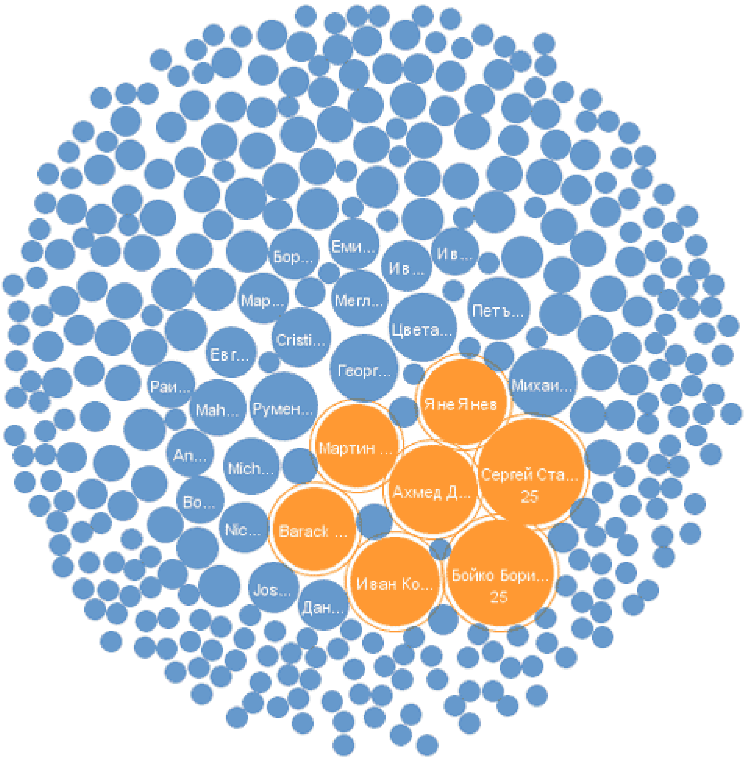

*Figure 2.* The most popular names in Bulgarian online media in terms of frequency of presence in the news, June 2009. Bubble size represents the number of days in which the given name is among the most frequently mentioned subjects. Source: Daskalova (17 July 2009).[14]

Figure 2 illustrates the total set of unique names that were registered in the People section of NewsExplorer for the Bulgarian language news sources for a period of one month, namely 1-30 June 2009. The data was gathered by manual accumulation of the lists of names generated in the People section for every day of the month. The author (Daskalova, 17 July 2009) uses the bubble chart technique in particular because it clearly illustrates sets consisting of dozens to hundreds of items as it is in the given case. As seen, Figure 2 is made up mostly of small bubbles, i.e. names registered in the entry for just one day of the month, which actually explains the great number of subjects. However, it is noteworthy that a considerable portion of those people are regularly present in the news. In the foreground we see the core consisting of statistically most significant players — persons that can be called the leading role actors in news stories for the month. Analyzing who they are, how their popularity in the news varies over the months, and looking at the original media and event contexts where the names appear are all analytical interpretation techniques to deconstruct statistical peaks and to reveal the qualitative meanings behind the quantitative accumulation. Of course, this applies both to the analysis of subjects in the news and on the analysis of news clusters. All this is in favour of a reasoned and data validated expertise of mediatized publicity.

Among other possible (and practiced in the MML) approaches to reading databases is the grouping of most common subjects present in the news on the basis of various criteria. This means going beyond the mere counting of data and adding recognizable features for each subject. Generated units can be grouped into categories (Figure 3), and categories can be subjected to more detailed sections: for example, the grouping of names around public entities (political parties and institutions) allows for making more detailed observations of representations of political news in the area. It becomes possible to answer questions like: what are the trends in news interest to the government and the opposition, how intense is the media coverage of individual political leaders, how much is the public presence of a political party multifaceted and how is it concentrated in the figure of the party leader, in what type of events the representatives of a certain political force accumulate greater media interest, etc.

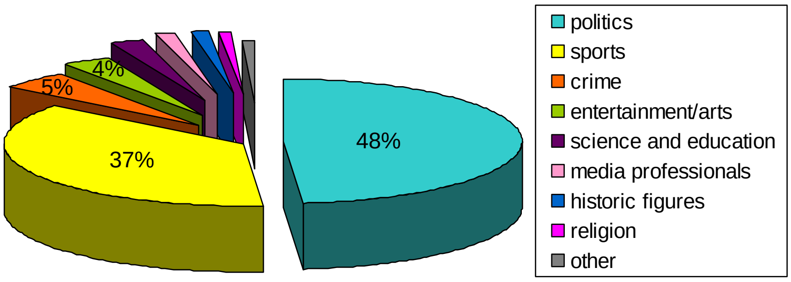

*Figure 3.* Most popular persons in the news in categories, Bulgarian news sites, August 2009. Source: Daskalova (17 Sept. 2009).

Among the useful features of NewsExplorer is the registration of quotations related to actors in the news. This allows for the exploration of the rhetorical strategies of public figures. Of importance for the media analysis is mostly what and who the leading media characters (i.e. the biggest bubbles in Figure 2) speak about. Precisely because they are the most intensively present in the media, they are central authors of messages in the public sphere and together with the media they set the tone of public discourse and public agenda. In this context, the phenomena that MML diagnosed for the elections in 2009 in Bulgaria included: replacement of political debate and institutional positions by personal verbal attacks between political opponents; mutual influence between the strategies of public speaking of the central election enemies; domination of the yellow tones in the political messages; populist uses of the EU for the purposes of the election campaign, etc.

To get a better idea of the nature of the conclusions reached through analytical reading of the databases in NewsExplorer, we can add other central conclusions regarding the Bulgarian news sites, including: solidification of political scandals as the norm to attract media attention; media conformity to power; correspondence of media interest to voter behaviour of society; a direct link between power shifts and media coverage of political actors; interweaving between criminal and political news stories; clear elections-oriented media placement of the leading political parties, etc. All these conclusions are based on extremely precise monitoring and deconstruction of statistical variables.

Of great importance is also the question of what remains outside the statistical peaks, i.e. what the statistically invisible tells us. Or in other words, what type of events become current breaking news and which events remain in the periphery or even out of news media interest. One particularly symptomatic example refers to the news arrays from January 2009. At the time, two important events happened in Bulgaria: a gas crisis as a result of the conflict between Ukraine and Russia on the one hand, and large-scale civil protests outside parliament, on the other. 'Clean quantitative comparison of information about the development of the gas crisis (80 clusters, with more than 1000 information units) and the protests (reflected in only 9 clusters with a total of 47 information units) suggests that the media prefer to cover the stories of political elites, and not those of citizens' (Kutseva 2009).

In conclusion we can summarize that the applicability of EMM as a primary analytical resource for assessing the media phenomena and processes is extremely high mainly because the system formats the media agenda in [co-]measurable quantities. Moreover, using the model of the mechanisms for analysis proposed here, structural distinctions in media sources within a country and comparisons with foreign-language media arrays could be made. The analysis can be extended in many directions based on the great number of technological solutions offered by EMM. On the other hand, the very media representations constructed through the statistical and semantic tools of EMM within its web applications give significant signs of key developments in society. As it is discussed in the following section, producing indications to alert the public about important global events is among the main reasons for establishing such an electronic monitoring system.

## The Europe media monitor: Institutional context

The acquaintance with the technological parameters and analytical potential of EMM eventually leads us to the question: what is the social background that triggers the development of a computer instrument as powerful as EMM?

The origin of the idea of creating a robust media monitoring platform refers to the DG Communication of the European Commission. For its needs and the needs of the Commission, the Directorate was required to make daily summary of news from a wide range of European and world media. And since this was a task that required too much time and effort, DG took steps to facilitate the process. For this purpose, DG Communication contacted the Joint Research Centre (JRC) and the two sub-structures of the European Commission began to collaborate on the establishment of an electronic system for media monitoring. The Institute for the Protection and Security of the Citizen (IPSC) at JRC undertook the technical implementation of the project. Following its launch in 2002, EMM began to function as an extremely useful tool among EU institutions. Over the years, the platform has been constantly expanding — both in terms of scope (linguistic and geographic) and in terms of areas of use of the software.

The technological support originally sought by DG Communications does not cover all motives that inspired the development and continuous improvement of this intelligent system. Simplifying and accelerating the process of handling large databases are functional purposes that fit very well in the macro-objectives of the institutional environment in which the software product and underlying technologies have been developed. It is the Institute for Protection and Security of the Citizen at JRC that is actively developing methods within the scientific scope of computational linguistics, on the basis of which

EMM tools are deployed. As it is well known, data processing techniques are traditional priority in the operational activities of national and supra-national organizations in the field of security. Significantly, such institutions are usually supported by innovative technology research cores. The clearest example is the internet itself — its creation was the result of the work of the Advanced Research Projects Agency at the Ministry of Defense of the United States. In this sense the analogy with the Joint Research Centre of the European Commission is inevitable. Indeed, the mission of the JRC is historically linked with the issue of security: the Centre was established back in 1957 under the Treaty establishing the European Atomic Energy Community (EURATOM). Since then the activities of the JRC have been inextricably linked with EURATOM's objectives to promote nuclear protection and security for European citizens and since 2000 the Centre has focused its research in areas such as statistics, macroeconomic modeling, financial econometrics and sensitivity analysis, social multi-criteria evaluation and knowledge assessment.[15]

The development of Europe Media Monitor in an institutional environment devoted to the problem of security inevitably affects the internal structural features of the system. For example, news posted on NewsBrief are categorized into thematic accents, including: crime, natural disasters, manmade disasters, terrorist attack, communicable diseases, conflict, development, food security & aid, society, ecology, humanitarian aid, political unrest, security. Special emphasis is placed also on the algorithmic mechanisms to detect events related to security for the early discovery of threats and dangerous developments.[16] There is also the option for users to subscribe to immediate or daily alert updates of registered security threats. That is to say the system automatically detects threat-related keywords in news articles and sends alerts via e-mail. The system also provides geo-mapping of conflicts and terrorist attacks as well as visualization of the alert level of countries in combining booth quality (recognition of keywords, security related themes and quotes) and quantity (statistical measurement of occurrence in the news) aspects.

In addition, as previously mentioned, one of EMM's applications focuses entirely on the issue of health security. In the MedISys sub-system, the process of identifying events from data arrays has been further tailored to healthcare. This type of monitoring is considered of great significance for the monitoring of diseases, epidemics and especially for the timely prevention of hazardous situations (Yangbarber et al. 2007). The idea of developing the application has come from the European Commission's Health and Consumers Directorate General (SANCO).

EMM capabilities are actively used by the European Anti-Fraud Office (OLAF) and the majority of the services of the European Commission. The JRC serves not only the EU institutions and government organizations in the Member States, but also a broader network of institutions outside the EU (e.g. in the USA, Canada, China), as well as international organizations (including various United Nations and Pan-African sub-organizations) (Steinberger 2010).

An important feature of EMM is that inside the Commission Intranet there exists a slightly more elaborate version of EMM that processes next to the public sources also a number of commercial wire providers. Another advantage of the Intranet version of EMM is that users can change the category definitions — an option which is not accessible on the public website.[17] But even in its current incomplete form, which is accessible to the public, Europe Media Monitor enjoys great popularity among internet users. The public web pages are visited by an average of 30,000 anonymous users per day (Steinberger 2010).

## Conclusion: Intelligent information processing and the limits of knowledge

Should we summarize in one sentence the philosophy behind the Europe Media Monitor system, it could be said that it is a tool for improving knowledge about the surrounding world. This is the leitmotif of the research team that developed EMM. This is also the general idea that synthesizes the institutional intentions of the European Commission to create a 24-hour monitoring system that alerts should a threat is detected in various places around the world. Of course, technological control of information flows is based on purely practical reasons: 'the increasing amount of data on the Web needs to be processed in order to help users who are looking for specific information. Therefore, summarization systems are becoming more and more useful because they provide shorter versions of texts, avoiding users wasting their time' (Balahur et al. 2009). This is what technological optimism is all about — about an effort to overcome the burden of information overflow: 'applying techniques from text mining, automated machine learning and statistical analysis can help to reduce this overload of information' (Atkinson & Van der Goot 2009).

In short, knowledge optimization is achieved through intelligently modelled electronic monitoring of global media resources available on the internet. This process implies at least minimal trust in the media as carriers of knowledge on one hand. On the other, the reliability of resulting final factual accuracy is determined by the very nature of the monitoring tool. Despite the principal possibility for a low percentage of software mistakes, the sheer scale of processed data in conjunction with the wide diversity in the origin of information resources (linguistic and media) is a prerequisite for factual accuracy. Moreover, the JRC provides qualitative development of the analyzed information resources. For example, one of the current priorities of the research team is 'adding blogs as a new text type to the current news monitoring system' (Steinberger, Pouliquen & van der Goot 2009).[18] The perspective of knowledge management includes ever more complete automation of media analysis; another current priority of the JRC is 'adding opinion mining functionality to the existing information extraction components' (ibid). The reasoning behind such a step once again lies in the logic of seeking to improve awareness: 'Information complementarity not only applies to contents, but also to opinions: by considering points of view from around the world, readers will get a less biased, and more balanced, view on world events' (Steinberger 2010). This unshakable argument perfectly fits in the well-known rhetoric of democratic functioning public sphere. Moreover, the successful automation of this kind of analytical processes will turn EMM into an even more powerful tool for media analysis and will reveal new opportunities for secondary interpretation by pundits of trends in media coverage and in society in general.

Optimization of knowledge about the surrounding world is actually new knowledge. The idea behind a monitoring system such as EMM aims to go beyond the explicitly expressed, to detect and reveal hidden information in the thousands of scattered sources, to make primary invisible knowledge visible. Or, as Wired magazine wrote about EMM, 'That universe of information contains early warnings about everything from natural disasters to political unrest — if you can read the data' (Rogers 2008).

Expanding knowledge frontiers through an intelligent software system such as EMM raises the question about the limits of use of next generation technology. Ultimately, through intelligent machine-readable data observed subjects and objects become much more transparent to those that control the monitoring tool. For example, the EMM-Labs application is designed to track even family and intimate relationships between individuals. The question here is what the purpose of collecting information of such a personal nature is — would they necessarily be purposes related to improving the security of citizens?

We are witnessing diverse examples of uses of the knowledge generated by smart technology. On the one hand, the greater part of the EMM system is freely available for public use. On the other, there is the example of the Media Monitoring Lab which analytically uses the free resources of EMM to generate secondary detailed expertise in its civil ambitions to fill the lack of expertise in the media field in Bulgaria. A third example is the 'We Feel Fine' project based on a data collection engine that automatically scours the internet every ten minutes, harvesting human feelings from a large number of blogs.[19] The authors of the project define it as 'an artwork authored by everyone'.[20] A fourth type of strategy is offered by Newstin, a news and metadata technology innovations company, which offers metadata and media analysis products targeted at clients in the media industry, government organizations and multinational corporations.[21]

In all cases the key question remains – who will control the meta-knowledge of the future? 'Software, based on some of the principles of artificial intelligence, will be able to know much more about us than we know about ourselves. It will be able to derive new meaning and information by comparing content, which, taken separately, do not contain the same information. This effectively means great expansion of knowledge, but also of control. Both information flows and human behaviour will become much more transparent and filled with meaning long remained unsolved. Questions about the use of this knowledge will undoubtedly be fundamental' (Spassov 2008).

The answer could be sought only in a public debate far outside the purposes of the study presented here. However, it is important to emphasize that such a discussion requires a high degree of commitment on the part of the largest generator of knowledge: science. If nothing else, this means a need for more intense dialogue between the humanities and social sciences, on the one hand, and computer science and software engineering, on the other. Positive example in this respect is the hybridization of various fields of science as in projects such as computational social science (Lazer et. al), digital humanities (Raben 2007), and cultural science (Hartley 2009). After all, it is not only a question of reading data but also of how we read reality.

## Notes

1. On the one hand, the Internet is considered a technological miracle that will transform the world in new and better way. On the other hand, the utopian vision of the super-potential of the new communication tool is contrasted to the belief that computers have a dehumanizing destructive influence. For an overview of these two opposing theses, see Castells (2004: 354–5), Wellman (2004).

2. See http://emm.jrc.it/ or http://press.jrc.it/overview.html.

3. See http://emm.newsbrief.eu/overview.html.

4. In this context, Steinberger (2010) compares EMM only to the Newstin system (http://www.newstin.com/). Newstin offers real-time, semantic, multi-language and cross-language document categorization and web mining; the system news database comprises 120 million documents and 6 billion metadata items updated from over 160,000 weighted global sources in 11 languages (http://www.newstin.com/en_US/product). Based on these characteristics, the Newstin founders promote their products as 'superior to Google News' (http://www.newstin.com/).

5. See the MedISys overview, http://emm.newsbrief.eu/overview.html.

6. In addition, languages spoken outside the EU family are also monitored by the EMM engine.

7. The Eurovoc thesaurus has been developed since 1982 by the European Parliament in collaboration with the EC's Publications Office. It contains terms relevant to the activities of the European Parliament in 22 EU languages, as well as in Croatian and Serbian. Eurovoc can be accessed on http://europa.eu/eurovoc.

8. The process is described in Pouliquen, Steinberger and Ignat (2003).

9. In my experience with the EMM NewsExplorer application I have come across this problem as well as across software errors in the identification of places, cities, boulevards, etc. as persons. The percentage of such errors does not exceed 4% of the People section names accumulated on a monthly basis.

10. For example, if I know nothing about contemporary public developments in Italy, it would be extremely hard for me to read the data derived from Italian news sources.

11. See http://emm.newsexplorer.eu/NewsExplorer/readme.html.

12. An interesting fact in JRC's work is that due to the small number of computational linguists engaged in the research, the team 'always had to use minimalistic methods and try to achieve with them as much as possible' (Steinberger 2010).

13. See http://www.fmd.bg/ [in Bulgarian]. The MML project was launched in 2008 by the Media Democracy Foundation with the financial support of the Trust for Civil Society in Central and Eastern Europe. After its start, the MML has been additionally supported by the Media Program South East Europe of Konrad Adenauer Foundation. The project is aimed at developing independent monitoring of the media in Bulgaria. Its particular focus is on the analysis of media and politics relationships. The scope of observation covers print, broadcast and online media. Since December 2008 the Lab has been critically monitoring the media developments in the country by generating a large number of studies, data and expert opinions. The MML has introduced a number of innovative tools for media monitoring including the instrumental use of the EMM NewsExplorer application. All monitoring results have been popularized on news conferences and public discussions and are all available on the website of the Media Democracy Foundation.

14. The bubble chart was generated though the Many Eyes website. Many Eyes is an experiment brought by IBM Rersearch and IBM Cognos Software group, and aimed at democratizing visualization (Many Eyes 2007). It is a data visualizations website which gives users the possibility to upload data sets, to create, share and discuss visualizations.

15. See http://ec.europa.eu/dgs/jrc/index.cfm?id=3840&lang=en.

16. For more information about the event monitoring technology in the area of security within the EMM see Atkinson et al. (2008).

17. The author would like to thank Erik van der Goot and Ralf Stainberger from the Joint Research Centre for the information on the differences in in-house and public use of EMM, as well as for providing useful reference for research from 'behind the scenes' of EMM (personal e-mail correspondence, January 2009).

18. Studies on detection and sentiment analysis of attitudes expressed in blog posts are already under way (Balahur et al. 2009).

19. See http://wefeelfine.org/methodology.html; Kamvar and Harris (2009).

20. See http://wefeelfine.org/mission.html.

21. See http://www.newstin.com/en_US and Luft (2008).

## References

Atkinson, M. & E. Van der Goot (2009) 'Near Real Time Information Mining in Multilingual News'. In: *Proceedings of the 18th International World Wide Web Conference (WWW'2009)*, pp. 1153-1154. Madrid, 20-24 April 2009.

Atkinson et al. (2008) 'Online Monitoring of Security-Related Events'. In: *Proceedings of the 22nd International Conference on Computational Linguistics (CoLing'200*8). Manchester, UK, 18-22 August 2008, http://langtech.jrc.it/Documents/2008_08_Coling_demo-paper-events_CD.pdf.

Balahur, A. et. al (2009) 'Opinion Mining on Newspaper Quotations'. In: *Proceedings of the workshop 'Intelligent Analysis and Processing of Web News Content' (IAPWNC)*, held at the 2009 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 523-526. Milano, Italy, 15.09.2009.

Castells, M. (2004) *The Rise of the Network Society*. Sofia: Lik.

Daskalova, N. (20 November 2009) I am. Bulgarian Online Media in the Europe Media Monitor, 1 September – 15 November 2009. Sofia: Foundation Media Democracy, http://www.fmd.bg/?p=4714.

Daskalova, N. (17 September 2009) In New Roles. Bulgarian Online Media in the Europe Media Monitor, July-August 2009. Sofia: Foundation Media Democracy, http://www.fmd.bg/?p=4323.

Daskalova, N. (17 July 2009) The King of Pop. Bulgarian Online Media in the Europe Media Monitor, June 2009. Sofia: Foundation Media Democracy, http://www.fmd.bg/?p=3627.

Hartley, J. (2009) 'From Cultural Studies to Cultural Science'. *Cultural Science*, 2(1), http://cultural-science.org/journal/index.php/culturalscience/article/view/19/68.

Kamvar, S. and J. Harris (2009) *We Feel Fine: An Almanac of Human Emotion*. New York: Scribner.

Kutseva, G. (2009) 'The themes in EMM'. *Culture Weekly*, No 13 (2540), 3 April 2009.

Kutseva, G. (17 Sept. 2009) Extreme Makeover. Bulgarian Online Media in the Europe Media Monitor, July-August 2009. Sofia: Foundation Media Democracy, http://www.fmd.bg/?p=4363.

Lazer, D. et. Al (2009) 'Computational Social Science'. *Science*, 323(5915): 721-723.

Luft, O. (2008) 'Innovations in  Journalism – Newstin'. *Journalism.co.uk*, http://blogs.journalism.co.uk/editors/2008/02/11/innovations-in-journalism-newstin/.

Many Eyes (2007) 'Democratizing Visualization', http://manyeyes.alphaworks.ibm.com/blog/2007/01/31/democratizing-visualization/.

Pouliquen, B., H. Tanev and M. Atkinson (2008) 'Extracting and Learning Social Networks out of Multilingual News'. In: *Proceedings of the social networks and application tools workshop (SocNet-08)* pp. 13-16. Skalica, Slovakia, 19-21 September 2008, http://langtech.jrc.it/Documents/SocNetMultilingualNewsVF2.pdf.

Pouliquen B., R. Steinberger and C. Best (2007) 'Automatic Detection of Quotations in Multilingual News'. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP'2007),* pp. 487-492. Borovets, Bulgaria, 27-29.09.2007, http://langtech.jrc.it/Documents/0709_RANLP_Quotation-detection_BP-RS-CB_final.pdf.

Pouliquen, B. et al. (2005) *Multilingual person name recognition and transliteration. Journal CORELA - Cognition, Représentation, Langage. Numéros spéciaux, Le traitement lexicographique des noms propres,* http://edel.univ-poitiers.fr/corela/document.php?id=490.

Pouliquen et. al (2004) 'Multilingual and cross-lingual news topic tracking'. In: *Proceedings of the 20th International Conference on Computational Linguistics (CoLing'2004),* Vol. II, pages 959-965. Geneva, Switzerland, 23-27 August 2004.

Pouliquen, B., R. Steinberger and C. Ignat. (2003) 'Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus'. In: *Proceedings of the Workshop Ontologies and Information Extraction at the Summer School The Semantic Web and Language Technology – Its Potential and Practicalities (EUROLAN'2003).* Bucharest, Romania, 28 July – 8 August 2003, http://langtech.jrc.it/Documents/EuroLan-03_Pouliquen-Steinberger-et-al.pdf.

Raben, J. (2007) 'Introducing Issues in Humanities Computing'. *Digital Humanities Quarterly,* 1(1), http://digitalhumanities.org/dhq/vol/1/1/000008/000008.html.

Rogers, A. (2008) 'Tracking the News: A Smarter Way to Predict Riots and Wars'. *Wired.* http://www.wired.com/science/discoveries/magazine/16-07/pb_news.

Spassov, O. (2008) *Web 3.0: the future of the internet or the end of the romantic side of the net* [manuscript].

Steinberger, R. (2010) 'Challenges and Methods for Multilingual Text Mining'. In: *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010).* Valletta, Malta, 19-21 May 2010.

Steinberger, R., B. Pouliquen & E. van der Goot (2009) 'An Introduction to the Europe Media Monitor Family of Applications'. In: Fredric Gey, Noriko Kando & Jussi Karlgren (eds.): *Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR'2009),* pp. 1-8. Boston, USA. 23 July 2009.

Steinberger, R., B. Pouliquen and C. Ignat (2005) 'Navigating multilingual news collections using automatically extracted information'. *Journal of Computing and Information Technology* - CIT 13, 2005, 4, 257-264, http://cit.zesoi.fer.hr/downloadPaper.php?paper=767.

Steinberger, R. (1998) *Language Engineering for UCLAF,* http://langtech.jrc.it/Documents/Slides-981023_Steinberger_LE-Presentation-UCLAF.pdf.

Wellman, B. (2004) 'The Three Ages of Internet Studies: Ten, Five and Zero Years Ago'. *New Media and Society*, 6 (1): 108–114.

Yangbarber et al. (2007) 'Combining Information about Epidemic Threats from Multiple Sources'. In: Proceedings of the Workshop Multi-source Multilingual Information Extraction and Summarization (MMIES'2007) held at RANLP'2007, pp. 41-48. Borovets, Bulgaria, 26 September 2007, http://langtech.jrc.it/Documents/0709_RANLP-MMIES_Yangarber-et-al.pdf.