Tracking Media Reports on the Shiga Toxin-Producing *Escherichia coli* O104: H4 Outbreak in Germany

Jens P. Linge[1], Jas Mantero[2], Flavio Fuart[1], Jenya Belyaeva[1], Martin Atkinson[1], and Erik van der Goot[1]

[1] Joint Research Centre of the European Commission, Institute for the Protection and Security of the Citizen, Ispra, Italy
[2] European Centre for Disease Control and Prevention (ECDC), Stockholm, Sweden
{jens.linge,flavio.fuart,martin.atkinson, erik.van-der-goot}@jrc.ec.europa.eu
jas.mantero@ecdc.europa.eu
jenya.belyaeva@ext.jrc.ec.europa.eu

Abstract. In May 2011, an outbreak of enterohemorrhagic *Escherichia coli* (EHEC) occurred in northern Germany. The Shiga toxin-producing strain O104:H4 infected several thousand people, frequently leading to haemolytic uremic syndrome (HUS) and gastroenteritis (GI). First reports about the outbreak appeared in the German media on Saturday 21st of May 2011; the media attention rose to high levels in the following two weeks, with up to 2000 articles categorized per day by the automatic threat detection system MedISys (Medical Information System). In this article, we illustrate how MedISys detected the sudden increase in reporting on *E. coli* on 21st of May and how automatic analysis of the reporting provided epidemic intelligence information to follow the event. Categorization, filtering and clustering allowed identifying different aspects within the unfolding news event, analyzing general media and official sites in parallel.

Keywords: epidemic intelligence; event-based surveillance; *E. coli*; EHEC; MedISys.

1 Introduction

Internet surveillance systems have increasingly been used for early event detection and alerting of emerging public health threats [1]. The Medical Information System (MedISys, http://medisys.newsbrief.eu) is a fully automatic public health surveillance system to monitor reporting on emerging public health threats such as human and animal infectious diseases, chemical, biological, radiological and nuclear (CBRN) threats, food & feed contaminations and plant diseases [2]. The system retrieves news articles from the internet, categorizes all incoming articles according to pre-defined multilingual categories, identifies entities such as organizations, persons and locations, clusters articles and calculates statistics to detect emerging threats. Users can screen the categorized articles and display world maps highlighting event locations together with

P. Kostkova, M. Szomszor, and D. Fowler (Eds.): eHealth 2011, LNICST 91, pp. 178–185, 2012.

statistics on the reporting of health threats, countries and combinations thereof. Articles can be further filtered by language, news source, and country.

Articles are classified in a category, if they satisfy the category definition which may comprise Boolean operators, proximity operators and wild cards. Cumulative positive or negative weights can be used with an adjustable threshold. For *E. coli*, a category was introduced in MedISys in 2008 with expert input from DG SANCO and ECDC [3]. It consists of a set of patterns covering several languages. Table 1 shows the definition (as of 27 June 2011), with the weights for each pattern indicated in the second column.

MedISys monitors the volume of news per category and per country to determine sudden changes from the 14-day average number of articles in any given country-category combination (e.g. Germany - *E.coli*). If the number of articles in the last 24 hours for a country-category combination (normalized by weekday fluctuations) is significantly higher than the 14-day average, users are notified using ranking graphs on the web site or email notifications. The statistics are language-independent because of the multi-lingual patterns in the category definitions. This allows users to detect any change in a category even before the event is reported in their own language.

MedISys also clusters all news items within a time window of four hours (or more, depending on the number of recent articles) and presents the largest clusters as Top Stories.

2 Media Reports on the Outbreak

On the 19th of May, the Robert-Koch-Institute was informed about a cluster of three cases of HUS in Hamburg [4]. MedISys detected the first media report in the German newspaper Die Welt on Saturday 21st of May at 12:14 CEST (Central European Summer Time). Several other articles followed the developing story in the afternoon (see Table 2), making reference to various press releases by public health authorities in Germany, e.g. in Hamburg and Lower Saxony (in MedISys, users can distinguish between general news media and official sources). Altogether, 23 German news items triggered the MedISys *E. coli* category, of which 22 were about the outbreak (one irrelevant article was about the water quality in German lakes in the region Oberbergischer Kreis). Furthermore, there were two news reports in Farsi and one report in Mandarin; these reports were from the science pages of Iranian and Chinese newspapers and were not related to the outbreak. Due to the sudden increase in media reports on *E.coli* on Saturday (in comparison to the average value of the last 14 days), MedISys issued an automatic alert and also highlighted *E.coli* in combination with Germany to the users (in the section called ”Most active topics”).

The Early Warning and Response System (EWRS) of the European Union [5] received a first communication by the German authorities on Sunday 22 May at 11:40. ProMED-mail covered the event in a report on Monday 23 May [6]. All other major early alerting systems (ARGUS, Biocaster, GPHIN, HealthMap, PULS) reported the event as well.

Table 1. Category definition (as of 27 June 2011) showing the patterns with their corresponding weights (using wild cards such as % for zero, one or more characters, + for whitespace/linebreak, and _ for one character). An incoming news article is selected, if the sum of the weights of all triggered patterns exceeds the threshold (which was set to 40 for this category); negative weights are used to exclude irrelevant news items.

Pattern	Weight
escherichia+coli+enfek%	20
koli+basili%	20
pałeczka+okrężnicy	20
koli+basili	20
eşerişiya+koli	20
бактерията+Ешерихия+коли	20
кишечн%+палочк%+EHEC	20
палочк%+EHEC	20
EHEC	20
enterohämorragisches+Eschericha+Coli	20
Ehec-Infekt%	20
Ehec-Keim	20
Coliba	-99
Colidiu	-99
Ešerihija+koli	20
escherichia+coli	20
ешерихи%+кол%	20
Ešerihioze	20
大腸桿菌	20
大腸埃希氏桿菌	20
大肠杆菌	20
大肠埃希氏杆菌	20
الإشـــــريكية+القولونيـــــة	20
اشريشـــــــياكلي	20
اكولـــى	20
اســـــچريچيا+كولـــي	20
ىكـــولاىا	20
enterokrwotoczn%	20
escherichia+coli+O157_H7	20
escherichia+coli+O104_H4	20
escheric%+coli%	20
e.coli+bacteri%	20
Ешерихия+коли	20
кишечн%+палочк%	20
e-coli%	20
Sukelta	20
Enterohemoraginės	20
Lazdelės	20
Žarninės	20
escherich%	20
Nakkus	20
enteroh%	20
Colibacille	20
Κολοβακτηρίδιο	20
e+coli%	20
Ešerihioze	20
Kolibakteeri	20

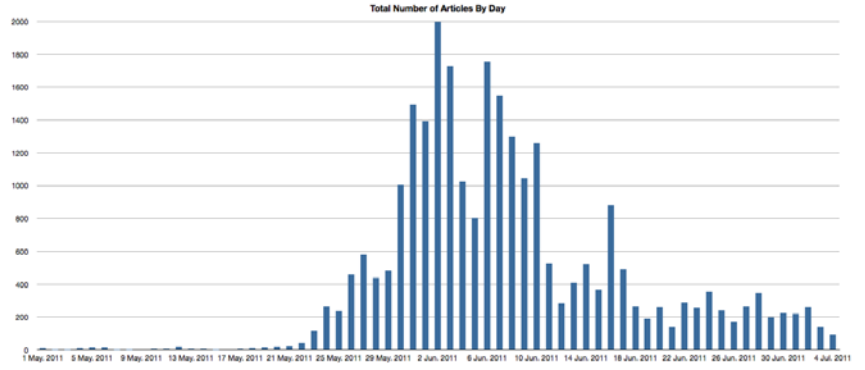

Fig. 1. Number of articles in the *E. coli* category per day (data from 1 May to 4 July 2011; all languages)

While the outbreak in northern Germany reached its peak between 21 and 23 May [4], the media reporting showed a different behaviour. As illustrated in Fig. 1, the highest number of articles per day was reached on 2 June 2011 (all languages, all countries). The data can be further filtered by language, country of publication, country mentioned in the text (using a multi-lingual category; see Fig. 2). The filter functionality clearly shows how the media attention changed geographical focus over time, following the developing situation.

In Fig. 2, we can clearly identify key aspects:

- the sudden rise on articles mentioning Germany (21-25 May), when the first cases became public,
- a peak with articles on a Swedish tourist group who got infected in Lower Saxony (26 May).
- the reporting on alleged *E. coli* contaminations in Spanish cucumber, tomatoes and salad (peak on 27 May with 107 articles),
- the reporting on the financial impact on Spanish farmers and the announcement that Spanish cucumbers had tested negative for *E. coli* (peak on 31 May with 300 articles),
- the discussion on trade restrictions in Russia for EU vegetable products (peak on 2 June),
- the announcements by German authorities that bean sprouts were the source of infection (rise in volume on 5-10 June),
- the *E.coli* cluster of cases in Bordeaux, France (peak on 16 June with 203 articles), and
- the reporting on fenugreek seeds imported from Egypt in 2009 and 2010 (from end of June on).

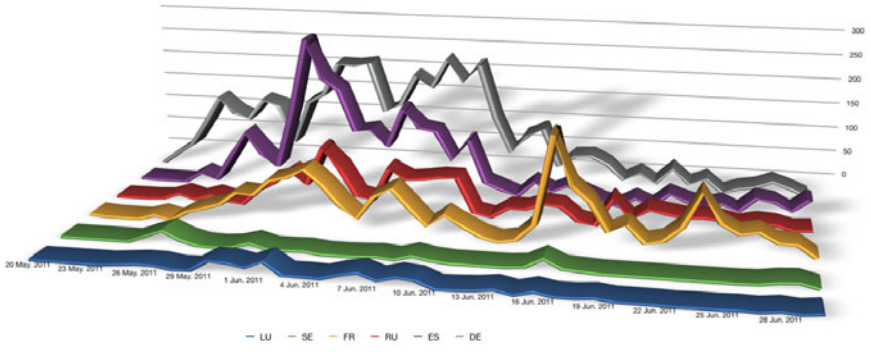

Fig. 2. Number of articles in the *E. coli* category from one of the countries Luxembourg (LU), Sweden (SE), France (FR), Russia (RU), Spain (ES) and Germany (DE)

Fig. 3 illustrates how MedISys presented statistics on the outbreak on its web site (screenshots taken on 26 May). It demonstrates how the deviation from the 14-day average alerts the users to the *E.coli*-Germany combination. All data used for the figures stem from MedISys.

In addition to the *E. coli* category, the outbreak was also visible in filters set up by EFSA [7], e.g. EFSABacteria (which contains *E. coli* as potential pathogen), EFSAEconomics and EFSACommerce in relation to Spain (impact on Spanish farmers) and Russia (trade restrictions) and EFSAFoodFeedSafety. This demonstrates that broader filters targeted at economics and commerce are able to detect changes in media reporting.

Using entity extraction, the main organizations and people mentioned in the articles can be identified by the system. As an example, we extracted a subset of articles on *E. coli* that also mentioned the Robert-Koch-Institute, Commissioner John Dalli, European Commission, ECDC, and EFSA. These entities were selected from an automatically generated list of top entities (according to number of citations). Fig. 4 summarizes the data, highlighting the following aspects:

- Most articles from the early phase mention the Robert-Koch-Institute,
- The European Commission is mentioned during the discussions on the alleged contaminations of Spanish cucumbers, tomatoes and salad; Commissioner John Dalli is cited in the peaks on 2 June and 7 June (in parallel to the European Commission),
- ECDC and EFSA are mentioned in many articles regarding the risk assessment in terms of public health at EU level; a strong peak in reporting on 30 June is due to press releases about scientific reports and risk assessments, e.g. the scientific report [8] lead to more media attention (375 articles mentioning EFSA on 5 July).

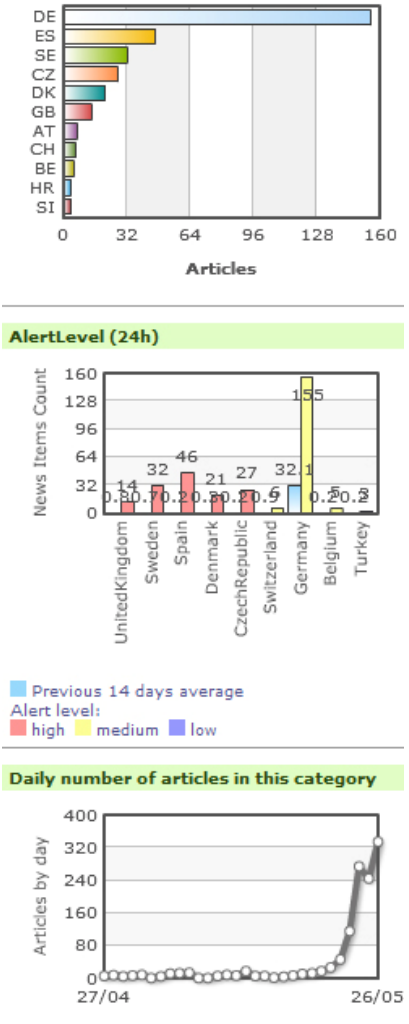

Fig. 3. Screenshot of MedISys on 26th of May (above: country distribution; centre: statistics on *E.coli*-country combinations in comparison to 14-day average values; below: daily number of articles for the *E. coli* category)

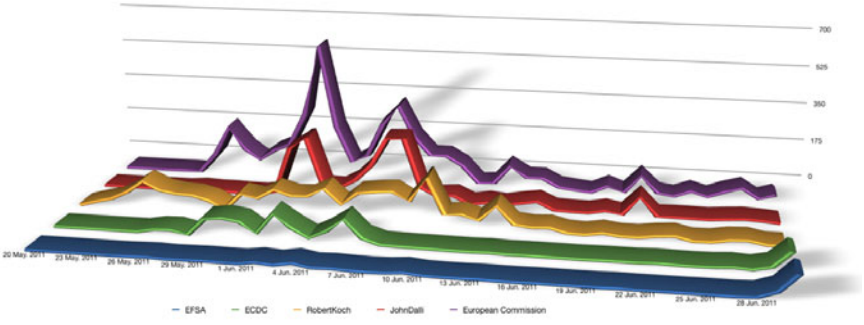

Fig. 4. Number of articles in the *E.coli* category that also mention one of the entities Robert Koch Institute, Commissioner John Dalli, European Commission, ECDC and EFSA

3 Conclusions

Media monitoring is a well-established technique for Epidemic Intelligence [9]. In the case of the *E.coli* outbreak, the early media reports were identified quickly and accurately; an alert was published on Saturday 21st of May via MedISys. Obviously, the authorities in Germany were informed earlier via the established indicator-based systems in Germany, and the EWRS message published on Sunday 22 notified all EU member states about the outbreak. Thus, media monitoring was less used for early alerting, but rather to identify key aspects of the developing story (which were further disseminated, e.g. in reports for the European Commission). Since MedISys covers both general media and official sites, users were able to analyze their reports in parallel.

Acknowledgments. We thank Prof Dr Nikolaos Stilianakis and Dr Agnès Rortais for fruitful discussions and comments on the manuscript.

References

1. Linge, J.P., Steinberger, R., Weber, T.P., Yangarber, R., van der Goot, E., Al-Khudhairy, D., Stilianakis, N.I.: Internet surveillance systems for early alerting and health threats. Euro. Surveill. 14(13), 19162 (2009)
2. Linge, J.P., Belyaeva, J., Steinberger, R., Gemo, M., Fuart, F., Al-Khudhairy, D., Bucci, S., Yangarber, R., van der Goot, E.: MedISys: Medical Information System. In: Asimakopoulou, E., Bessis, N. (eds.) Advanced ICTs for Disaster Management and Threat Detection: Collaborative and Distributed Frameworks, pp. 131–142. IGI Global (2010)
3. Mantero, J., Belyaeva, J., Linge, J.P.: How to maximise event-based surveillance web-systems: the example of ECDC/JRC collaboration to improve the performance of MedISys. JRC Scientific and Technical Reports (EUR collection). Publications Office of the European Union (2011) ISBN 978-92-79-19755-0. ISSN 1018-5593, doi: 10.2788/69804

4. Frank, C., Werber, D., Cramer, J.P., Askar, M., Faber, M., Heiden, M.A., Bernard, H., Fruth, A., Prager, R., Spode, A., Wadl, M., Zoufaly, A., Jordan, S., Stark, K., Krause, G.: The HUS Investigation Team: Epidemic Profile of Shiga-Toxin-Producing Escherichia coli O104:H4 Outbreak in Germany - Preliminary Report. N Engl. J. Med., June 22 (2011)
5. Guglielmetti, P., Coulombier, D., Thinus, G., Van Loock, F., Schreck, S.: The Early Warning and Response System for communicable diseases in the EU: an overview from 1999 to 2005. Euro. Surveill. 11(12), 666 (2006), http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=666
6. ProMED-mail: E. coli VTEC - Germany: RFI. ProMED-mail 2011, May 23: 20110523.1566, http://www.promedmail.org (accessed July 3, 2011)
7. Linge, J.P., Belyaeva, J.: Customization of MedISys for the monitoring of food and feed hazards (2011), http://www.efsa.europa.eu/en/supporting/pub/140e.htm
8. EFSA: Tracing seeds, in particular fenugreek (Trigonella foenum-graecum) seeds, in relation to the Shiga toxin-producing E. coli (STEC) O104:H4 2011 Outbreaks in Germany and France. EFSA Report. Question No EFSA-Q-2011-00817 (2011), http://www.efsa.europa.eu/en/supporting/pub/176e.htm (issued on July 05, 2011)
9. Paquet, C., Coulombier, D., Kaiser, R., Ciotti, M.: Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. Euro. Surveill. 11, 665 (2006), http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=665