An ontology-driven system for detecting global health events

Abstract

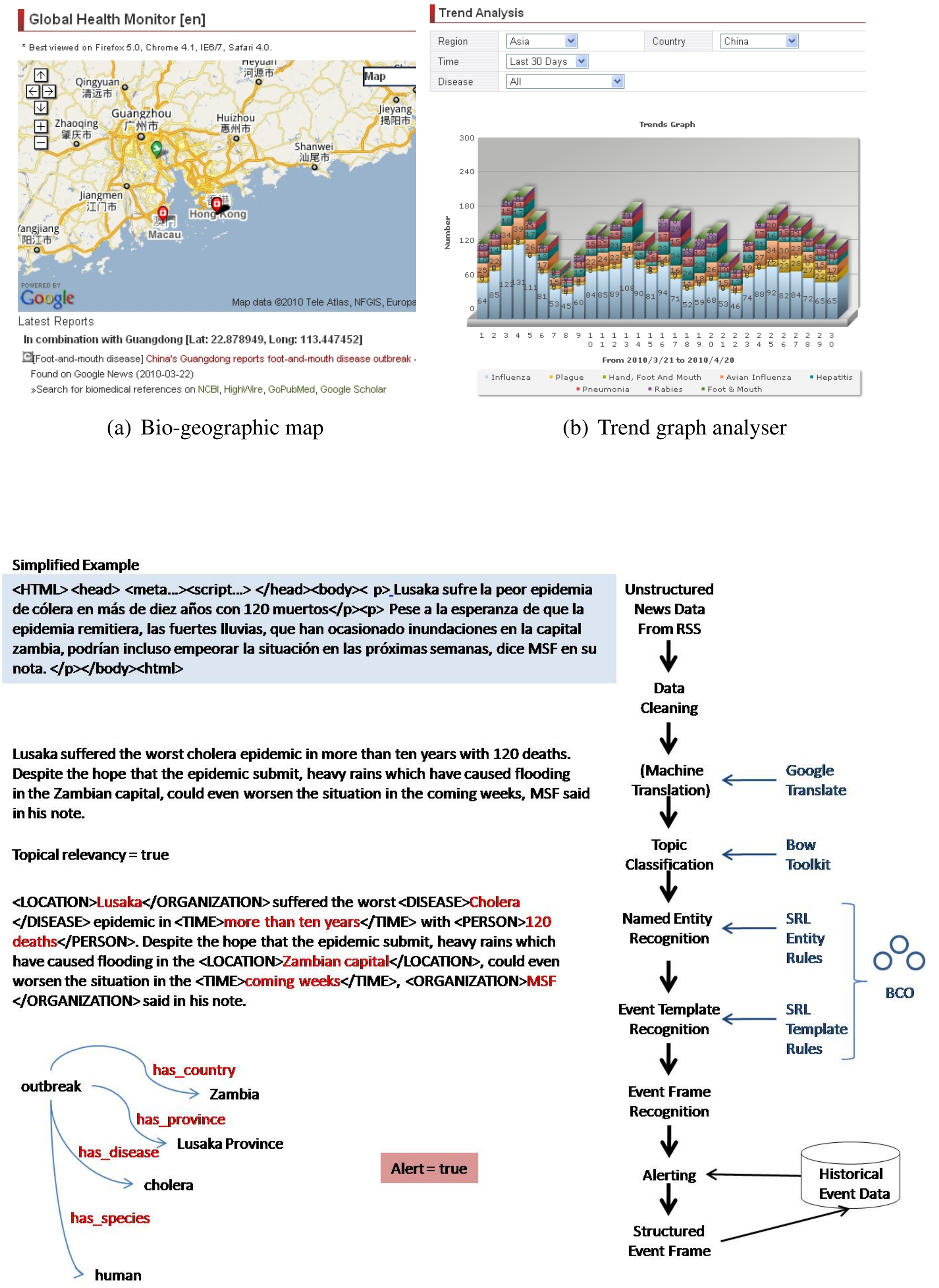
Text mining for global health surveillance is an emerging technology that is gaining increased attention from public health and governments. The lack of multilingual resources such as WordNets specifically targeted at this task have so far been a major bottleneck. This paper reports on a major upgrade to the BioCaster Web monitoring system and its freely available multilingual ontology; improving its original design and extending its coverage of diseases from 70 to 336 in 12 languages.

1 Introduction

The number of countries who can sustain teams of experts for global monitoring of human/animal health is limited by scarce national budgets. Whilst some countries have advanced sensor networks, the world remains at risk from the health impacts of infectious diseases and environmental accidents. As seen by the recent A(H5N1), A(H1N1) and SARS outbreaks, a problem in one part of the world can be rapidly exported, leading to global hardship.

The World Health Organization (WHO) estimates that in the future, between 2 to 7.4 million people could be at risk worldwide from a highly contageous avian flu virus that spreads rapidly through the international air travel network (WHO, 2005). Pandemics of novel pathogens have the capacity to overwhelm healthcare systems, leading to widespread morbidity, mortality and socio-economic disruption (Cox et al., 2003). Furthermore, outbreaks of livestock diseases, such as foot-and-mouth disease or equine influenza can have a devastating impact on industry, commerce and numan health e.g. (Blake et al., 2003). The challenge is to enhance vigilance and control the emergence of diseases. Whilst human analysis remains essential to spot complex relationships, automated analysis has a key role to play in filtering the vast volume of data in real time and highlighting unusual trends using reliable predictor indicators.

BioCaster [*citation withheld*] is a Web 2.0 monitoring station for the early detection of infectious disease events. The system exploits a high-throughput semantic processing pipeline, converting unstructured news texts to structured records, alerting events based on time-series analysis and then sharing this information with users via geolocating maps (Fig. 1(a)), graphs (Fig. 1(b)) and alerts. Underlying the system is a publicly available multilingual application ontology. Launched in 2006 [*citation witheld*], the BioCaster Ontology (BCO) has been downloaded by over 70 academic and in-



Global Health Monitor [en]

* Best viewed on Firefox 5.0, Chrome 4.1, IE6/7, Safari 4.0.

Latest Reports

In combination with Guangdong [Lat: 22.878949, Long: 113.447452]

[Foot-and-mouth disease] China's Guangdong reports foot-and-mouth disease outbreak -
Found on Google News (2010-03-22)
»Search for biomedical references on NCBI, HighWire, GoPubMed, Google Scholar

(a) Bio-geographic map

Trend Analysis

Region	Asia	Country	China
Time	Last 30 Days		
Disease	All		

(b) Trend graph analyser

Simplified Example

<HTML> <head> <meta...><script...> </head><body>< p>_Lusaka sufre la peor epidemia de cólera en más de diez años con 120 muertos</p><p> Pese a la esperanza de que la epidemia remitiera, las fuertes lluvias, que han ocasionado inundaciones en la capital zambia, podrían incluso empeorar la situación en las próximas semanas, dice MSF en su nota. </p></body><html>

Lusaka suffered the worst cholera epidemic in more than ten years with 120 deaths. Despite the hope that the epidemic submit, heavy rains which have caused flooding in the Zambian capital, could even worsen the situation in the coming weeks, MSF said in his note.

Topical relevancy = true

<LOCATION>Lusaka</ORGANIZATION> suffered the worst <DISEASE>Cholera </DISEASE> epidemic in <TIME>more than ten years</TIME> with <PERSON>120 deaths</PERSON>. Despite the hope that the epidemic submit, heavy rains which have caused flooding in the <LOCATION>Zambian capital</LOCATION>, could even worsen the situation in the <TIME>coming weeks</TIME>, <ORGANIZATION>MSF </ORGANIZATION> said in his note.

(c) BioCaster processes

Figure 1: (a)BioCaster's bio-geographic map for a suspected foot-and-mouth outbreak on 22nd March 2010 with links to the multilingual ontology, NCBI, HighWire, GoPubMed and Google Scholar; (b) The trends analyser showing aggregated document counts for health events in China between 13th March and 12th April 2010; (c) The system's pipeline of processes with example semantic markup.

dustrial groups worldwide. This paper reports on a major upgrade to the system and the ontology - expanding the number of languages from 6 to 12, redefining key relations and extending coverage in the number of diseases from 70 to 336, including many veterinary diseases.

2 Background

As the world becomes more interconnected and urbanized and animal production becomes increasingly intensive, the speed with which epidemics spread becomes faster, adding to pressure on biomedical experts and governments to make quick decisions. Traditional validation methods such as field investigations or laboratory analysis are the mainstay of public health but can require days or weeks to issue reports. The World Wide Web with its economical and real time delivery of information represents a new modality in health surveillance (Wagner and Johnson, 2006) and has been shown to be an effective source by the World Health Organization (WHO) when Public Health Canada's GPHIN system detected the SARS outbreak in southern China from news reports during November 2002. The recent A(H1N1) 'swine flu' pandemic highlighted the trend towards agencies using unvalidated sources. The technological basis for such systems can be found in statistical classification approaches and light weight ontological reasoning. For example, Google Flu Trends (Ginsberg et al., 2009) is a system that depends almost entirely on automatic statistical classification of user queries; MedISys-PULS (Yangarber et al., 2008), HealthMap (Freifeld et al., 2008) and BioCaster use a mixture of statistical and ontological classification; and GPHIN (Mawudeku and Blench, 2006) and Argus (Wilson, 2007) rely on a mixture of ontological classification and manual analysis.

Compared to other similar systems BioCaster is characterized by its richly featured and publicly downloadable ontology and emphasizes critical evaluation of its text mining modules. Empirical results have included: topic classification, named entity recognition, formal concept analysis and event recognition [*citations withheld*]. In the absence of a community gold standard, task performance was assessed on the best available human standard - the ProMED-mail network (Madoff and Woodall, 2005), achieving F-score of 0.63 on 14 disease-country pairs over a 365-day period.

Despite initial skepticism within the public health community, health surveillance systems based on NLP-supported human analysis of media reports are becoming firmly established in Europe, North America and Japan as sources of health information available to governments and the public (Hartley et al., 2010). Whilst there is no substitute for trained human analysts, automated filtering has helped experts save time by allowing them to sift quickly through massive volumes of media data. It has also enabled them to supplement traditional sources with a broader base of information.

In comparison with other areas of biomedical NLP such as the clinical and genetics' domains, a relative lack of building block resources may have hindered the wider participation of NLP groups in public health applications. It is hoped that the provision of common resources like the BCO can help encourage further development and benchmarking.

3 Method

BioCaster performs analysis of over 9000 news articles per day using the NPACI Rocks cluster middleware (http://www.rockcsclusters.org) on a platform of 48 3.0GHz Xeon cores. Data is ingested 24/7 into a semantic processing pipeline in a short 1 hour cycle from over 1700 public domain RSS feeds such as Google news, the European Media Monitor and ProMED-mail. Since 2009, news has also being gathered under contract from a commercial news aggregation company, providing access to over 80,000 sources across the world's languages.

The new 2010 version of BioCaster uses machine translation into English (eleven languages) to source news stories related to currently occurring infectious and environmental disease outbreaks in humans, animals and plants.

Access to the site is freely available but login registration applies to some functions such as email alerts. Processing is totally automatic, but we have the potential within the login system to enable human moderated alerts which broadcast to Twitter (http://twitter.com/biocaster) and RSS.

Below we describe in detail two key aspects of the system that have been significantly upgraded: the BCO and the event detection system.

3.1 Ontology

3.1.1 Aim

The BioCaster Ontology aims:

- To describe the terms and relations necessary to detect and risk assess public health events in the grey literature;
- To bridge the gap between (multilingual) grey literature and existing standards in biomedicine;
- To mediate integration of content across languages;
- To be freely available.

The central knowledge source for BioCaster is the multilingual ontology containing domain terms such as diseases, agents, symptoms, syndromes and species as well as domain sensitive relations such as a diseases causing symptoms or agents affecting particular host species. This allows the text mining system to have a basic understanding of the key concepts and relationships within the domain to fill in gaps not mentioned explicitly in the news reports. To the best of our knowledge the BCO is unique as an application ontology, providing freely available multilingual support to system developers interested in outbreak surveillance in the language of the open media.

The BCO however has little to say outside of its application domain, e.g. in disease-gene interaction or for supporting automatic diagnosis. As discussed in Grey Cowell and Smith, (2010), there are many other resources available that have the potential to support applications for infectious disease analysis including controlled vocabularies and ontologies such as the the Unified Medical Language System (UMLS) (Lindberg et al., 1993), International Classification of Diseases (ICD-10) (WHO, 2004), SNOMED CT (Stearns et al., 2001), Medical Subject Headings (MeSH) (Lipscomb, 2000) and the Infectious Disease Ontology (IDO) (Grey Cowell and Smith, 2010). In [*citation withheld*] we discussed how BCO compared to such ontologies so we will focus from now on the implication of the extensions.

3.1.2 Scope

The new version of the BCO now covers 12 languages including all the United Nation's official languages: Arabic (968 terms), English (4113), French (1281), Indonesian (1081), Japanese (2077), Korean (1176), Malaysian (1001), Russian (1187), Spanish (1171), Thai (1485), Vietnamese (1297) and Chinese (1142). Currently news in all 12 languages are available via the Web portal but news in additional languages such as German, Italian and Dutch are being added using machine translation.

3.1.3 Design

Like EuroWordNet (Vossen, 1998), on which it is loosely based, the BCO adopts a thesaurus-like structure with synonym sets linking together terms across languages with similar meaning. Synonym sets are referred to using *root terms*. Root terms themselves are fully defined instances and provide a bridge to external classification schemes and nomenclatures such as ICD10, MeSH, SNOMED CT and Wikipedia. The central backbone taxonomy is deliberately shallow and taken from the ISO's Suggested Upper Merged Ontology (Niles and Pease, 2001). To maintain consistency and computability we kept a single inheritance structure throughout. 18 core domain concepts corresponding to named entities in the text mining system such as DISEASE and SYMPTOM were the results of analysis using a formal theory (Guarino and Welty, 2000).

We have endeavoured to construct human definitions for root terms along Aristotelean principles by specifying the difference to the parent. For example in the case of *Eastern encephalitis virus*:

> *Eastern equine encephalitis virus is a species of virus that belongs to the genus Alphavirus of the family Togaviridae (order unassigned) of the group IV ((+)ssRNA) that possesses a positive single stranded RNA genome. It is the etiological agent of the eastern equine encephalitis.*

We are conscious though that terms used in the definitions still require more rigorous control to be considered useful for machine reasoning. To aid both human and machine analysis root terms are linked by a rich relational structure reflecting domain sensitive relations such as *causes(virus,disease)*, *has_symptom(disease, symptom)*, *has_associated_syndrome(disease, syndrome)*, *has_reservoir(virus, organism)*.

In such a large undertaking, the order of work was critical. We proceeded by collecting a list of notifiable diseases from national health agencies and then grouped the diseases according to perceived relevance to the International Health Regulations 2005 (Lawrence and Gostin, 2004). In this way we covered approximately 200 diseases, and then explored freely available resources and the biomedical literature to find academic and layman's terminology to describe their agents, affected hosts, vector species, symptoms, etc. We then expanded the coverage to less well known human diseases, zoonotic diseases and diseases caused by toxic substances such as sarin, hydrogen sulfide, sulfur dioxide and ethylene. At regular stages we checked and validated terms against those appearing in the news media.

As we expanded the number of diseases to include animals we found a major structural reorganization was needed to support animal symptoms. For example, a high temperature in humans would not be the same as one in bovids. This prompted us in the new version to group diseases and symptoms around major animal familes and related groups, e.g. *high temperature (human)* and *high temperature (bovine)*.

A second issue that we encountered was the need to restructure the hierarchy under *OrganicObject* which was divided between *MicroOrganism* (i.e. infecting agents) and *Animal* (affected hosts).

The *MicroOrganism* class contained bacterium, helminth, protozoan, fungus and virus; however, this posed a tremendous problem, since some organisms such as worms and mites (e.g. scabies) also infect humans. This prompted us to restructure the organic classes using the traditional Linnean taxonomy as a guideline, although this is probably not free from errors (e.g. virus is probably not an organism).

3.2 Event alerting system

Figure 1(c) shows a schematic of the modular design used by the BioCaster text mining system. Following on from machine translation and topic classification is named entity recognition and template recognition which we describe in more detail below. The final structured event frames include slot values normalized to ontology root terms for disease, pathogen (virus or bacterium), country and province. Additionally we also identify 15 aspects of public health events critical to risk assessment such as: spread across international borders, hospital worker infection, accidental or deliberate release, food contamination and vaccine contamination.

Latitude and longitude of events down to the province level are found in two ways: using the Google API up to a limit of 15000 lookups per day, and then using lookup on BCO taxonomy of 5000 country and province names derived from open sources such as Wikipedia.

Each hour events are automatically alerted to a Web portal page by comparing daily aggregated event counts against historical norms (Collier, 2010). Login users can also sign up to receive emails on specific topics. A topic would normally specify a disease or syndrome, a country or region and a specific risk condition.

In order to extract knowledge from documents, BioCaster maintains a collection of rule patterns in a regular expression language that converts surface expressions into structured information. For example the surface phrase "man exposes airline passengers to measles" would be converted into the three templates "**species(human); disease(measles); international_travel(true)**". Writing patterns to produce such templates can be very time consuming and so the BioCaster project has developed its own light weight rule language - called the Simple Rule Language (SRL) and a pattern building interface for maintaining the rule base (McCrae et al., 2009). Both are freely available to the research community under an open source license. Currently BioCaster uses approximately 4000 SRL rules to identify events of interest in English. Using SRL allows us to quickly adapt the system to newly emerging terminology such as the 11+ designations given to A(H1N1) during the first stages of the 2009 pandemic.

The SRL rulebook for BioCaster can recognize a range of entities related to the task of disease surveillance such as bacteria, chemicals, diseases, countries, provinces, cities and major airports. Many of these classes are recognized using terms imported from the BCO. The rule book also contains specialised thesauri to recognize subclasses of entities such as locations of habitation, eateries and medical service centres. Verb lists are maintained for lexical classes such as detection, mutation, investigation, causation, contamination, culling, blaming, and spreading.

Some examples of SRL rules for named entity recognition are shown in Table 1 and described below:

Rule D3 in the rulebook tags phrases like 'mystery illness' or 'unknown killer bug' by matching on strings contained within two wordlists, @undiagnosed and @disease, separated by up to one word.

```
D3: :- name(disease){ list(@undiagnosed) words(,1) list(@disease) }
S2: :- name(symptom) { list(@severity) list(@symptom)}
CF1: contaminated_food("true") :- "caused" "by" list(@contaminate_verbs_past)
list(@injested_material)
SP4: species("animal") :- name(animal,A) words(,3) list(@cull_verbs_past)
```

Table 1: Examples of SRL rules for named entity and template recognition. Template rules contain a label, a head and a body, where the head specifies the template pattern to be output if the body expression matches. The body can contain word lists, literals, and wild cards. Various conditions can be placed on each of these such as orthographic matching.

Rule S2 allows severity indicators such as 'severe' or 'acute' to modify a list of known symptoms in order to identify symptom entities.

Rule CF1 is an example of a template rule. If the body of the rule matches by picking out expressions such as 'was caused by tainted juice', this triggers the head to output an alert for contaminated food.

Rule SP4 identifies the victim species as 'animal' in contexts like '250 geese were destroyed'.

The rulebook also supports more complex inferences such as the home country of national public health organizations.

Since BioCaster does not employ systematic manual checking of its reports, it uses a number of heuristic filters to increase specificity (the proportion of correctly identified negatives) for reports that appear on the public Web portal pages. For example, reports with no identified disease and country are rejected. Since these heuristics may reduce sensitivity they are not applied to news that appears on the user login portal pages.

4 Results and Discussion

Version 3 of the ontology represents a significant expansion in the coverage of diseases, symptoms and pathogens on version 2. Table 2 summarizes the number of root terms for diseases classified by animal familes.

The thesaurus like structure of the BCO is compatible in many respects to the Simple Knowledge Organization System (SKOS) (Miles et al., 2005). In order to extend exchange and re-use we have produced a SKOS version of the BCO which is available from the BCO site. We have also converted the BCO terms into 12 SRL rule books (1 for each language) for entity tagging. These too are freely available from the BCO site.

As the ontology expands we will consider adopting a more detailed typing of diseases such as *hasInfectingPart* to indicate the organ affected or *hasProtectionMethod* to indicate broad classes of methods used to prevent or treat a condition. The typology of diseases could also be extended in a more fine grained manner to logically group conditions, e.g. *West Nile virus encephalitis*, *Powassan encephalitis* and the *Japanese B encephalitis* could be connected through a *hasType* relation on *encephalitis*.

5 Conclusion

Text mining for global health surveillance is an emerging technology that is gaining increased attention from public health and governments. Multilingual resources specifically targeted at this task have so far been very rare. We hope that the release of version 3 can be used to support

Species	N	Example
Avian	22	Fowl pox
Bee	6	Chalk brood disease
Bovine	24	Bluetongue
Canine	4	Blastomycosis (Canine)
Caprine	14	Contagious agalactia
Cervine	2	Chronic wasting disease
Equine	17	Strangles
Feline	4	Feline AIDS
Fish	2	Viral hemorr hagic septicemia
Human	216	Scarlet fever
Lagomorph	2	Myxomatosis
Non-human primate	16	Sylvan yellow fever
Other	2	Crayfish plague
Rodent	8	Colorado tick fever (Rodent)
Swine	12	Swine erysipelas

Table 2: Major disease groups organized by affected animal family. N represents the number of root terms.

a range of applications such as text classification, cross language search, machine translation, query expansion and so on.

The BCO has been constructed to provide core vocabulary and knowledge support to the BioCaster project but it has also been influential in the construction of other public health oriented application ontologies such as the Syndromic Surveillance Ontology (Okhamatovskaia et al., 2009). The BCO is freely available from [site withheld] under a Ceative Commons license.

References

Blake, A., M. T. Sinclair, and G. Sugiyarto. 2003. Quantifying the impact of foot and mouth disease on tourism and the UK economy. *Tourism Economics*, 9(4):449–465.

Collier, N. 2010. What‘s unusual in online disease outbreak news? *Biomedical Semantics*, 1(1), March. doi:10.1186/2041-1480-1-2.

Cox, N., S. Temblyn, and T. Tam. 2003. Influenza pandemic planning. *Vaccine*, 21(16):1801–1803.

Freifeld, C., K. Mandl, B. Reis, and J. Brownstein. 2008. Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports. *J. American Medical Informatics Association*, 15:150–157.

Ginsberg, J., M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014.

Grey Cowell, L. and B. Smith. 2010. Infectious disease informatics. In Sintchenko, V., editor, *Infectious Disease Informatics*, pages 373–395. Springer New York.

Guarino, N. and C. Welty. 2000. A formal ontology of properties. In Dieng, R. and O. Corby, editors, *EKAW-2000: Proc. 12th Int. Conf. on Knowledge Engineering and Knowledge Management*, pages 97–112.

Hartley, D., N. Nelson, R. Walters, R. Arthur, R. Yangarber, L. Madoff, J. Linge, A. Mawudeku, N. Collier, J. Brownstein, G. Thinus, and N. Lightfoot. 2010. The landscape of international biosurveillance. *Emerging Health Threats J.*, 3(e3), January. doi:10.1093/bioinformatics/btn534.

Lawrence, O. and J. Gostin. 2004. International infectious disease law - revision of the World Health Organization's international health regulations. *J. American Medical Informatics Association*, 291(21):2623–2627.

Lindberg, Donald A.B., L. Humphreys, Betsy, and T. McCray, Alexa. 1993. The unified medical language system. *Methods of Information in Medicine*, 32:281–291.

Lipscomb, C. 2000. Medical subject headings (MeSH). *Bulletin of the Medical Library Assocation*, 88:256–266.

Madoff, Lawrence C. and John P. Woodall. 2005. The internet and the global monitoring of emerging diseases: Lessons from the first 10 years of promed-mail. *Archives of Medical Research*, 36(6):724 – 730. Infectious Diseases: Revisiting Past Problems and Addressing Future Challenges.

Mawudeku, A. and M. Blench. 2006. Global public health intelligence network (gphin). In *Proc. 7th Int. Conf. of the Association for Machine Translation in the Americas, Cambridge, MA, USA*, August 8–12.

McCrae, J., M. Conway, and N. Collier. 2009. Simple rule language editor. Google code project, September. Available from: http://code.google.com/p/srl-editor/.

Miles, A., B. Matthews, and M. Wilson. 2005. SKOS Core: Simple knowledge organization for the web. In *Proc. Int. Conf. on Dublin Core and Metadata Applications, Madrid, Spain*, 12–15 September.

Niles, I. and A. Pease. 2001. Towards a standard upper ontology. In Welty, C. and B. Smith, editors, *2nd Int. Conf. on Formal Ontology in Information Systems FOIS-2001, Maine, USA*, October 17–19.

Okhamatovskaia, A., W. Chapman, N. Collier, J. Espino, and D. Buckeridge. 2009. SSO: The syndromic surveillance ontology. In *Proc. Int. Soc. for Disease Surveillance, Miami, USA*, December 3–4.

Stearns, M. Q., C. Price, K. A. Spackman, and A. Y. Wang. 2001. SNOMED clinical terms: overview of the development process and project status. In *Proc. American Medical Informatics Association (AMIA) Symposium*, pages 662–666.

Vossen, P. 1998. Introduction to EuroWordNet. *Computers and the Humanities*, 32:73–89.

Wagner, M. and H. Johnson. 2006. The internet as sentinel. In Wagner, M. et al., editor, *The Handbook of Biosurveillance*, pages 375–385. Academic Press.

WHO. 2004. *ICD-10, International Statistical Classification of Diseases and Related Health Problems, Tenth Revision*. World Health Organization, December.

WHO. 2005. Avian influenza: assessing the pandemic threat. Technical Report WHO/CDS/2005.29, January.

Wilson, J. 2007. Argus: a global detection and tracking system for biological events. *Advances in Disease Surveillance*, 4.

Yangarber, R., P. von Etter, and R. Steinberger. 2008. Content collection and analysis in the domain of epidemiology. In *Proc. Int. Workshop on Describing Medical Web Resources (DRMED 2008), Gotenburg, Sweden*, May 27th.