Developing a Biosurveillance Application Ontology for Influenza-Like-Illness

Mike Conway, John Dowling and Wendy Chapman
Department of Biomedical Informatics
University of Pittsburgh
{conwaym|dowling|wec6}@pitt.edu

Abstract

Increasing biosurveillance capacity is a public health priority in both the developed and the developing world. Effective *syndromic* surveillance is especially important if we are to successfully identify and monitor disease outbreaks in their early stages. This paper describes the construction and preliminary evaluation of a syndromic surveillance orientated application ontology designed to facilitate the early identification of Influenza-Like-Illness syndrome from Emergency Room clinical reports using natural language processing.

1 Introduction and Motivation

Increasing biosurveillance capacity is a public health priority in both the developed and developing world, both for the early identification of emerging diseases and for pinpointing epidemic outbreaks (Chen et al., 2010). The 2009 Mexican flu outbreak provides an example of how an outbreak of a new disease (in this case a new variant of H1N1 influenza) can spend some weeks spreading in a community before it is recognized as a threat by public health officials.

Syndromic surveillance is vital if we are to detect outbreaks at an early stage (Henning, 2004; Wagner et al., 2006). The United States Center for Disease Control (CDC) defines syndromic surveillance as "surveillance using health-related data that precede diagnosis and signal a sufficient probability of a case or outbreak to warrant further public health response."[1] That is, the focus of syndromic surveillance is the identification of disease outbreaks *before* the traditional public health apparatus of confirmatory laboratory testing and official diagnosis can be used. Data sources for syndromic surveillance have included, over the counter pharmacy sales (Tsui et al., 2003), school absenteeism records (Lombardo et al., 2003), calls to *NHS Direct* (a nurse led information and advice service in the United Kingdom) (Cooper, 2007), and search engine queries (Eysenbach, 2006).

However, in this paper we concentrate on mining text based clinical records for outbreak data. Clinical interactions between health workers and patients generate large amounts of textual data — in the form of clinical reports, chief complaints, and so on — which provide an obvious source of pre-diagnosis information. In order to mine the information in these clinical reports we are faced with two distinct problems:

1. How should we define a syndrome of interest? That is, how are signs and symptoms mapped to syndromes?

2. Given that we have established such a set of mappings, how then do we map from the text in our clinical reports to the signs and symptoms that constitute a syndrome, given the high level of terminological variability in clinical reports.

This paper presents an application ontology that attempts to address both these issues for the domain of Influenza-Like-Illness Syndrome (ILI). The case definition for ILI, as defined by the United States Center for Disease Control is "fever greater than or equal to 100 degrees Fahrenheit

[1] www.webcitation.org/5pxhlyaxX

Proceedings of the 6th Workshop on Ontologies and Lexical Resources (Ontolex 2010), pages 58–66,
Beijing, August 2010

and either cough or sore throat."[2] In contrast to the CDC's straightforward definition, the syndrome is variously described as a cluster of symptoms and findings, including fever and cold symptoms, cough, nausea, vomiting, body aches and sore throat (Scholer, 2004). In constructing an application specific syndrome definition for this ontology, we used a data driven approach to defining ILI, generating a list of terms through an analysis of Emergency Room reports.

The remainder of the paper is divided into five parts. First, we briefly describe related work, before going on to report on the ontology development process. We then set forth an evaluation of the ontology with respect to its coverage of terms in the target domain. We go on to outline areas for future work, before finally presenting some concluding comments.

2 Related Work

In recent years there has been significant progress in interfacing lexical resources (in particular WordNet (Miller, 1995)) and upper level ontologies (like the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) (Gangemi et al., 2002) and the Suggested Upper Merged Ontology (SUMO) (Niles and Pease, 2003)). However, as our domain of interest employs a highly specialized terminology, the use of general linguistic resources like WordNet was inappropriate.

Our work has focused on the representation of ILI relevant concepts that occur in clinical reports in order to facilitate syndromic surveillance. While the widely used medical taxonomies and nomenclatures (for example Unified Medical Language System[3] and the Systematized Nomenclature of Medicine Clinical Terms[4]) contain many of the ILI relevant concepts found in clinical texts, these general resources do not have the specific relations (and lexical information) relevant to syndromic surveillance from clinical reports. Currently, there are at least four major terminological resources available that focus on the public health domain: PHSkb, SSO, and the BioCaster Ontology.

[2] www.webcitation.org/5q22KTcHx

[3] www.nlm.nih.gov/research/umls/

[4] www.ihtsdo.org/snomed-ct/

2.1 PHSkb

The Public Health Surveillance knowledge base PHSkb (Doyle et al., 2005) developed by the CDC is a coding system for the communication of notifiable disease[5] findings for public health professionals at the state and federal level in the United States. There are however several difficulties in using the PHSkb directly in an NLP orientated syndromic surveillance context:

1. Syndromic surveillance requires that syndromes and signs are adequately represented. The PHSkb emphasizes *diagnosed* diseases. That is, the PHSKb is focused on post diagnosis reporting, when laboratory tests have been conducted and the presence of a disease is confirmed. This approach is not suitable for syndromic surveillance where we seek to identify clusters of symptoms and signs *before* a diagnosis.

2. PHSkb is no longer under active development.

2.2 SSO

The Syndromic Surveillance Ontology (SSO) (Okhmatovskaia et al., 2009) was developed to address a pressing problem for system developers and public health officials. How can we integrate outbreak information when every site uses different syndrome definitions? For instance, if State **X** defines *sore throat* as part of ILI, yet State **Y** does not, syndromic surveillance results from each state will not be directly comparable. When we apply this example to the wider national scene, with federal regional and provincial public health agencies attempting to share data with each other, and international agencies, we can see the scale of the problem to be addressed.

In order to manage this data sharing problem, a working group of eighteen researchers, representing ten functional syndromic surveillance systems in the United States (for example, Boston Public Health Department and the US Department of Defense) convened to develop standard

[5] A notifiable disease is a disease (or by extension, condition) that must, by law, be reported to the authorities for monitoring purposes. In the United States, examples of notifiable diseases are: Shigellosis, Anthrax and HIV infection.

definitions for four syndromes of interest (*respiratory*, *gastro-intestinal*, *constitutional* and *ILI*)[6] and constructed an OWL ontology based on these definitions. While the SSO is a useful starting points, there are several reasons why — on its own — it is insufficient for clinical report processing:

1. SSO is centered on *chief complaints*. Chief complaints (or "presenting complaints") are phrases that briefly describe a patient's presenting condition on first contact with a medical facility. They usually describe symptoms, refrain from diagnostic speculation and employ frequent abbreviations and misspellings (for example "vom + naus" for "vomiting and nausea"). Clinical texts — the focus of attention in this paper — are full length documents, normally using correct spellings (even if they are somewhat "telegraphic" in style). Furthermore, clinical reports frequently list physical findings (that is, physical signs elicited by the physician, like, for instance reflex tests) which are not present in symptom orientated chief complaints.

2. The range of syndromes represented in SSO is limited to four. Although we are starting out with ILI, we have plans (and data) to extend our resource to four new syndromes (see Section 5 for details of further work).

3. The most distinctive feature of the SSO is that the knowledge engineering process was conducted in a face-to-face committee context. Currently, there is no process in place to extend the SSO to new syndromes, symptoms or domains.

2.3 BioCaster Ontology

The BioCaster application ontology was built to facilitate text mining of news articles for disease outbreaks in several different Pacific Rim languages (including English, Japanese, Thai and Vietnamese) (Collier et al., 2006). However, the ontology, as it stands, is not suitable for supporting text mining clinical reports, for the following reasons:

1. The BioCaster ontology concentrates on the types of concepts found in published news outlets for a general (that is, non medical) readership. The level of conceptual granularity and degree of terminological sophistication is not always directly applicable to that found in documents produced by health professionals.

2. The BioCaster ontology, while it does represent syndromes (for example, constitutional and hemorrhagic syndromes) and symptoms, does not represent physical findings, as these are beyond its scope.

In addition to the application ontologies described above, the Infectious Disease Ontology provides an Influenza component (and indeed wide coverage of many diseases relevant to syndromic surveillance). In Section 5 we describe plans to link to other ontologies.

3 Constructing the Ontology

Work began with the identification of ILI terms from clinical reports by author JD (a board-certified infectious disease physician with thirty years experience of clinical practice) supported by an informatician [author MC]. The term identification process involved the project's domain expert reading multiple reports,[7] searching through appropriate textbooks, and utilizing professional knowledge. After a provisional list of ILI concepts had been identified, we compared our list to the list of ILI concepts generated by the SSO ILI component (see Section 2.2) and attempted to reuse SSO concepts where possible. The resulting ILI concept list consisted of 40 clinical concepts taken from SSO and 15 new concepts. Clinical concepts were divided into three classes: Disease (15 concepts), Finding (21 concepts) and Symptom (19 concepts). Figure 1 shows the clinical

[6] A demonstration chief complaint classifier based on SSO is available at:
`http://onto-classifier.dbmi.pitt.edu/onto_classify.html`

[7] De-identified (that is, anonymized) clinical reports were obtained through partnership with the University of Pittsburgh Medical Center.

concepts covered. As part of our knowledge engineering effort, we identified concepts and associated relations for several different syndromes which we plan to add to our ontology at a later date.[8]

Early on in the project development process, we took the decision to design our ontology in such a way as to maintain consistency with the BioCaster ontology. We adopted the BioCaster ontology as a model for three reasons:

1. A considerable knowledge engineering effort has been invested in BioCaster since 2006, and both the domain (biosurveillance) and application area (text mining) are congruent to our own.

2. The BioCaster ontology has proven utility in its domain (biosurveillance from news texts) for driving NLP systems.

3. We plan to import BioCaster terms and relations, and thus settled on a structure that facilitated this goal.

The BioCaster ontology (inspired by the structure of EuroWordNet[9]) uses *root terms* as interlingual pivots for the multiple languages represented in the ontology.[10] One consequence of following this structure is that all clinical concepts are *instances*.[11] Additionally, all specified relations are relations between instances.

Relations relevant to the syndromic surveillance domain generally were identified by our physician in conjunction with an informatician (MC). Although some of these relations (like `is_bioterrorismDisease`) are less relevant to ILI syndrome, they were retained in order to maintain consistency with planned future work. Additionally, we have added links to other terminological resources (for example, UMLS and Snomed-CT)

[8]Note that finer granularity was used in the initial knowledge acquisition efforts (for example, we distinguished *sign* from *physical finding*).

[9]`http://www.illc.uva.nl/EuroWordNet/`

[10]Note that we are using *root term* instead of the equivalent EuroWordNet term *Inter Lingual Index*.

[11]Note that from a formal ontology perspective, concepts are instantiated in text. For example, "Patient **X** presents with *nausea* and *high fever*" instantiates the concepts **nausea** and **high fever**.

Lexical resources and regular expressions are a vital component of our project, as the ontology has been built with the public health audience in mind (in practice, state or city public health IT personnel). These users have typically had limited exposure to NLP pipelines, named entity recognizers, and so on. They require an (almost) "off the shelf" product that can easily be plugged into existing systems for text analysis.

The ontology currently includes 484 English keywords and 453 English regular expression. The core classes and relations were developed in Protege-OWL, and the populated ontology is generated from data stored in a spreadsheet (using a Perl script). Version control was managed using Subversion, and the ontology is available from a public access Google code site.[12] Figure 2 provides a simplified example of relations for the clinical concept instance *fever*.

4 Evaluation

In recent years, significant research effort has centered on the evaluation of ontologies and ontology-like lexical resources, with a smorgasbord of techniques available (Zhu et al., 2009; Brank et al., 2005). Yet no single evaluation method has achieved "best practice" status for all contexts. As our ontology is an application ontology designed to facilitate NLP in a highly constrained domain (that is, text analysis and information extraction from clinical reports) the notion of *coverage* is vital. There are two distinct questions here:

1. Can we map between the various textual instantiations of ILI concepts clinical reports and our ontology concepts? That is, are the NLP resources available in the ontology (keywords, regular expressions) adequate for the mapping task?

2. Do we have the right ILI concepts in our ontology? That is, do we adequately represent all the ILI concepts that occur in clinical reports?

Inspired by Grigonyte et al. (2010), we attempted to address these two related issues using

[12]`http://code.google.com/p/ss-ontology`

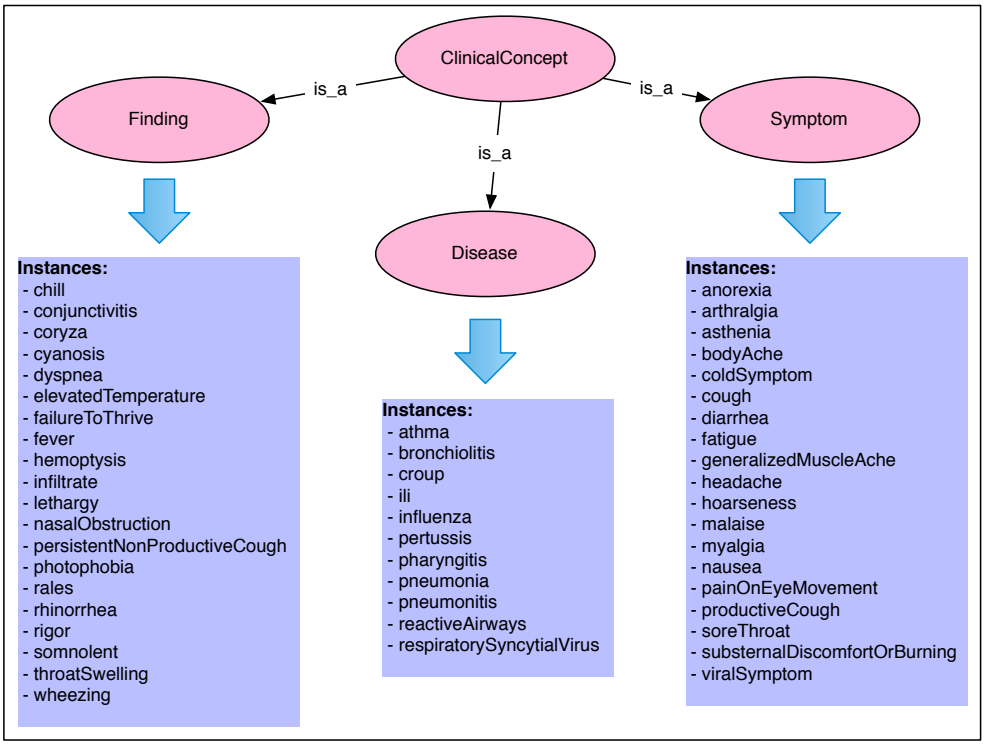

Figure 1: Clinical concepts.

techniques derived from terminology extraction and corpus linguistics. Our method consisted of assembling a corpus of twenty Emergency Room clinical reports which had been flagged by experts (not the current authors) as relevant to ILI. Note that these articles were not used in the initial knowledge engineering phase of the project. We then identified the "best" twenty five terms from these clinical reports using two tools, *Termine* and *KWExT*.

1. ***Termine*** (Frantzi et al., 2000) is a term extraction tool hosted by Manchester University's National Centre for Text Mining which can be accessed via web services.[13] It uses a method based on linguistic preprocessing and statistical methods. We extracted 231 terms from our twenty ILI documents (using *Termine*'s default configuration). Then we identified the twenty-five highest ranked *disease*, *finding* and *symptom* terms (that is, discarding terms like "hospital visit" and "chief complaint").

2. ***KWExT*** (Keyword Extraction Tool) (Conway, 2010) is a Linux based statistical keyword extraction tool.[14] We used *KWExT* to extract 1536 unigrams, bigrams and trigrams using the log-likelihood method (Dunning, 1993). The log-likelihood method is designed to identify n-grams that occur with the most frequency compared to some reference corpus. We used the FLOB corpus,[15] a one million multi-genre corpus consisting of American English from the early 1990s as our reference corpus. We ranked all n-grams according to their statistical significance and then manually identified the twenty-five highest ranked *disease*, *finding* and *symptom* terms.

Term lists derived using the *Termine* and *KWExT* tools are presented in Tables 1 and 2 respectively. For both tables, column two ("Term") details each of the twenty-five "best" terms (with respect to each term recognition algorithm) ex-

[13] www.nactem.ac.uk/software/termine/

[14] http://code.google.com/p/kwext/

[15] www.webcitation.org/5q1aKtnf3

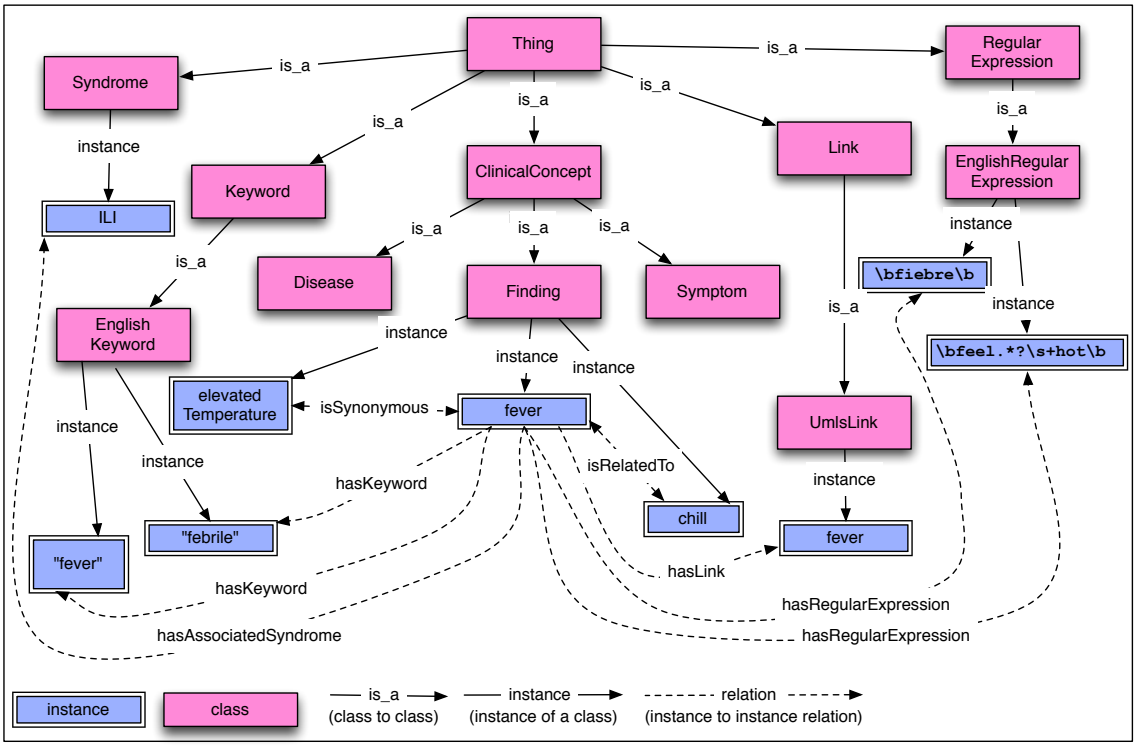

Figure 2: Example of clinical concept "fever" and its important relations (note the diagram is simplified).

tracted from our twenty document ILI corpus. Column three ("Concept") specifies the concept in our ontology to which the term maps (that is, the lexical resources in the ontology — keywords and regular expressions — can map the term in column two to the clinical concept in column three). For instance the extracted term *slight crackles* can be mapped to the clinical concept RALE using the keyword "crackles." Note that "-" in column three indicates that no mapping was possible. Underlined terms are those that *should* be mapped to concepts in the ontology, but currently are not (additional concepts and keywords will be added in the next iteration of the ontology).

There are two ways that mappings can fail here (mirroring the two questions posed at the beginning of this section). "Shortness of breath" *should* map to the concept DYSPNEA, but there is no keyword or regular expression that can bridge between text and concept. For the terms "edema" and "lymphadenopathy" however, no suitable candidate concept exists in the ontology.

5 Further Work

While the current ontology covers only ILI, we have firm plans to extend the current work along several different dimensions:

- Developing new relations, to include modeling DISEASE → SYMPTOM, and DISEASE → FINDING relations (for example TONSILLITIS **`hasSymptom`** SORE THROAT).
- Extend the application ontology beyond ILI to several other syndromes of interest to the biosurveillance community. These include:
  - *Rash Syndrome*
  - *Hemorrhagic Syndrome*
  - *Botulic Syndrome*
  - *Neurological Syndrome*
- Currently, we have links to UMLS (and also Snomed-CT and BioCaster). We intend to extend our coverage to the MeSH vocabulary (to facilitate mining PubMed) and also the Infectious Disease Ontology.

	Term	Concept
1	abdominal pain	-
2	chest pain	-
3	urinary tract infection	-
4	sore throat	SORE THROAT
5	renal disease	-
6	runny nose	CORYZA
7	body ache	MYALGIA
8	respiratory distress	PNEUMONIA
9	neck stiffness	-
10	yellow sputum	-
11	mild dementia	-
12	copd	-
13	viral syndrome	VIRAL SYN.
14	influenza	INFLUENZA
15	febrile illness	FEVER
16	lung problem	-
17	atrial fibrillation	-
18	severe copd	-
19	mild cough	COUGH
20	asthmatic bronchitis	BRONCHIOLITIS
21	coronary disease	-
22	dry cough	COUGH
23	neck pain	-
24	bronchial pneumonia	PNEUMONIA
25	slight crackles	RALE

Table 1: Terms generated using the *Termine* tool

	Term	Concept
1	cough	COUGH
2	fever	FEVER
3	pain	-
4	shortness of breath	-
5	vomiting	-
6	influenza	INFLUENZA
7	pneumonia	PNEUMONIA
8	diarrhea	DIARRHEA
9	nausea	NAUSEA
10	chills	CHILL
11	abdominal pain	-
12	chest pain	-
13	edema	-
14	cyanosis	CYANOSIS
15	lymphadenopathy	-
16	dysuria	-
17	dementia	-
18	urinary tract inf	-
19	sore throat	SORE THROAT
20	wheezing	WHEEZING
21	rhonchi	-
22	bronchitis	BRONCHIOLITIS
23	hypertension	-
24	tachycardia	-
25	respiratory distress	PNEUMONIA

Table 2: Terms generated using the *KWExT* tool

- Currently evaluation strategies have concentrated on *coverage*. We plan to extend our auditing to encompass both *intrinsic* evaluation (for example, have our relations evaluated by external health professionals using some variant of the "laddering" technique (Bright et al., 2009)) and *extrinsic* evaluation (for example, plugging the application ontology into an NLP pipeline for Named Entity Recognition and evaluating its performance in comparison to other techniques).

In addition to these ontology development and evaluation goals, we intend to use the ontology as a "gold standard" against which to evaluate automatic term recognition and taxonomy construction techniques for the syndromic surveillance domain. Further, we seek to integrate the resulting ontology with the BioCaster ontology allowing the potential for limited interlingual processing in priority languages (in the United States, Spanish).

Currently we are considering two ontology integration strategies. First, using the existing mappings we have created between the ILI ontology and BioCaster to access multi-lingual information (using OWL datatype properties). Second, fully integrating — that is, *merging* — the two ontologies and creating object property relations between them.

For example (using strategy 1), we could move from the string "flu" in a clinical report (identified by the `\bflu\b` regular expression) to the ILI ontology concept `ili:influenza`. In turn, `ili:influenza` could be linked (using a datatype property) to the BioCaster root term `biocaster:DISEASE_378` (which has the label "Influenza (Human)."). From the BioCaster root term, we can — for example — generate the translation "Gripe (Humano)" (Spanish).

6 Conclusion

The ILI application ontology developed from the need for knowledge resources for the text mining of clinical documents (specifically, Emergency Room clinical reports). Our initial evaluation indicates that we have good coverage of our domain, although we plan to incrementally work on improving any gaps in coverage through a process of active and regular updating. We have described our future plans to extend the ontology to new syndromes in order to provide a general commu-

nity resource to facilitate data sharing and integration in the NLP based syndromic surveillance domain. Finally, we actively solicit feedback on the design, scope and accuracy of the ontology.

Acknowledgments

This project was partially funded by Grant Number 3-R01-LM009427-02 (NLM) from the United States National Institute of Health.

References

Brank, J., Grobelnik, M., and Mladenić, D. (2005). A Survey of Ontology Evaluation Techniques. In *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005)*, pages 166–170.

Bright, T., Furuya, E., Kuperman, G., and Bakken, S. (2009). Laddering as a Technique for Ontology Evaluation. In *American Medical Informatics Symposium (AMIA 2009)*.

Chen, H., Zeng, D., and Dang, Y. (2010). *Infectious Disease Informatics: Syndromic Surveillance for Public Health and Bio-Defense*. Springer, New York.

Collier, N., Shigematsu, M., Dien, D., Berrero, R., Takeuchi, K., and Kawtrakul, A. (2006). A Multilingual Ontology for Infectious Disease Surveillance: Rationale, Design and Challenges. *Language Resources and Evaluation*, 40(3):405–413.

Conway, M. (2010). Mining a Corpus of Biographical Texts Using Keywords. *Literary and Linguistic Computing*, 25(1):23–35.

Cooper, D. (2007). *Disease Surveillance: A Public Health Informatics Approach*, chapter Case Study: Use of Tele-health Data for Syndromic Surveillance in England and Wales, pages 335–365. Wiley, New York.

Doyle, T., Ma, H., Groseclose, S., and Hopkins, R. (2005). PHSkb: A Knowledgebase to Support Notifiable Disease Surveillance. *BMC Med Inform Decis Mak*, 5:27.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

Eysenbach, G. (2006). Infodemiology: Tracking Flu-Related Searches on the Web for Syndromic Surveillance. In *American Medical Informatics Association Annual Symposium Proceedings (AMIA 2006)*, pages 244–248.

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic Recognition for Multi-word Terms. *International Journal of Digital Libraries*, 3(2):117–132.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., and Schneider, L. (2002). Sweetening Ontologies with DOLCE. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, pages 166–181.

Grigonyte, G., Brochhausen, M., Martin, L., Tsiknakis, M., and Haller, J. (2010). Evaluating Ontologies with NLP-Based Terminologies - A Case Study on ACGT and its Master Ontology. In *Formal Ontology in Information Systems: Proceedings of the Sixth International Conference (FOIS 2010)*, pages 331–344.

Henning, K. (2004). What is Syndromic Surveillance? *MMWR Morb Mortal Wkly Rep*, 53 Suppl:5–11.

Lombardo, J., Burkom, H., Elbert, E., Magruder, S., Lewis, S. H., Loschen, W., Sari, J., Sniegoski, C., Wojcik, R., and Pavlin, J. (2003). A Systems Overview of the Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE II). *J Urban Health*, 80(2 Suppl 1):32–42.

Miller, G. (1995). WordNet: A Lexical Database for English. *Communications of the Association for Computing Machinary*, 38(11):39–41.

Niles, I. and Pease, A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03)*, pages 23–26.

Okhmatovskaia, A., Chapman, W., Collier, N., Espino, J., and Buckeridge, D. (2009). SSO: The Syndromic Surveillance Ontology. In *Proceedings of the International Society for Disease Surveillance*.

Scholer, M. (2004). Development of a Syndrome Definition for Influenza-Like-Illness. In *Proceedings of American Public Health Association Meeting (APHA 2004)*.

Tsui, F., Espino, J., Dato, V., Gesteland, P., Hutman, J., and Wagner, M. (2003). Technical Description of RODS: a Real-Time Public Health Surveillance System. *J Am Med Inform Assoc*, 10(5):399–408.

Wagner, M., Gresham, L., and Dato, V. (2006). *Handbook of Biosurveillance*, chapter Case Detection, Outbreak Detection, and Outbreak Characterization, pages 27–50. Elsevier Academic Press.

Zhu, X., Fan, J.-W., Baorto, D., Weng, C., and Cimino, J. (2009). A Review of Auditing Methods Applied to the Content of Controlled Biomedical Terminologies. *Journal of Biomedical Informatics*, 42(3):413 – 425.