Developing an Application Ontology for Mining Free Text Clinical Reports: The Extended Syndromic Surveillance Ontology

Mike Conway[1], John Dowling[2], and Wendy Chapman[1]

[1] University of California, San Diego, Division of Biomedical Informatics
La Jolla, California 92093, USA
http://dbmi.ucsd.edu
{mconway@ucsd.edu|wwchapman@ucsd.edu}
[2] University of Pittsburgh, Department of Biomedical Informatics
Pittsburgh, PA 15260, USA
http://www.dbmi.pitt.edu
dowling@pitt.edu

Abstract. In an increasingly globalised world, where infectious disease outbreaks can rapidly circulate through the international transport system, and the threat of bioterrorism is constant, there is a need to develop reusable resources to support early-stage disease outbreak detection. This paper presents the Extended Syndromic Surveillance Ontology (ESSO), an open source terminological ontology designed to facilitate the mining of free-text clinical documents in English to support timely disease outbreak surveillance. ESSO consists of 279 clinical concepts (Fever, Slurred Speech, Diplopia, and so on) across eight syndromes (*respiratory syndrome*, *constitutional syndrome*, and so on) and is enriched with regular expressions to support concept identification in text. The ontology is shown to have good coverage in the target domain.

Keywords: syndromic surveillance, biosurveillance, terminology, ontology, natural language processing

1 Introduction & Motivation

Effective syndromic surveillance is useful if we are to detect and contain infectious disease outbreaks at an early stage [1, 2]. The United States Centers 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."[3] That is, the focus of syndromic surveillance is the identification of disease outbreaks *before* the traditional public health apparatus of confirmatory diagnostic testing and official diagnosis can be used. Data sources for syndromic surveillance have included over the counter pharmacy sales [3], school absenteeism records [4], calls to *NHS*

[3] www.webcitation.org/5pxhlyaxX

Direct (a nurse led information and advice service in the United Kingdom) [5], and search engine queries [6].

Grouping cases into syndromes (for example, *respiratory syndrome*) rather than into specific diagnoses (for example, *pneumonia*) may provide earlier evidence of infections of public health interest, because, in their early stages, many diseases have overlapping symptoms that may not initially alarm physicians [7, 8]. Typically, clinical interactions between health workers and patients generate substantial amounts of textual data in the form of radiography reports, Emergency Room[4] reports, chief complaints and so on, which provide an obvious source of pre-diagnostic information for syndromic surveillance. However, developing methods and resources that allow public health experts to gain maximum use from these data sources has been challenging.

This paper presents an application ontology — the Extended Syndromic Surveillance Ontology (ESSO) [9] — designed to support syndromic surveillance from clinical text, building on previous work in this area, in particular the Syndromic Surveillance Ontology [10]. The remainder of the paper consists of four sections. First, we briefly review related work, before going on to describe the ontology development process. We then set forth a short evaluation section before concluding with an outline of future work.

2 Related Work

Our work has focussed on the representation of concepts (and their lexical instantiations) as they occur in clinical text (in particular Emergency Room reports). While the widely used biomedical taxonomies, for example, the Unified Medical Language System[5] (UMLS) and the Systematised Nomenclature of Medical Clinical Terms[6] (SNOMED-CT) contain many of the syndromic surveillance related terms 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, ILI-SSO, and the BioCaster ontology.

The Public Health Surveillance knowledge base (PHSkb) [11] developed by the CDC is a coding system for the communication of notifiable disease findings for public health officials in the United States. PHSkb is not suitable as a resource for *syndromic* surveillance as its focus is on *diagnosed* diseases rather than pre-diagnostic surveillance. Additionally, PHSkb is no longer under active development.

The Syndromic Surveillance Ontology (SSO) [10] was developed to provide a set of common syndrome definitions for public health professionals in order to facilitate data sharing. A working group of eighteen researchers, representing ten syndromic surveillance systems in the United States convened to develop

[4] Also known as *Casualty Departments* or *Accident & Emergency Departments*

[5] `www.nlm.nih.gov/research/umls`

[6] `www.ihtsdo.org/snomed-ct`

standard definitions for four syndromes of interest [12] (*respiratory*, *gastrointestinal*, *influenza-like-illness* and *constitutional*) and constructed an OWL[7] ontology based on these definitions. While the SSO provides a useful starting point, there are two main reasons why — on its own — it is insufficient for clinical report processing: First, SSO is centred on *chief complaints*. Chief complaints (or "presenting complaints" in British English) 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 that describe not only symptoms, but patient history and diagnoses. Second, the number of syndromes in SSO is limited to four, whereas comprehensive syndromic surveillance requires the representation of further syndromes (for example, *hemorrhagic syndrome* and *neurological syndrome*).

The Influenza-Like-Illness Syndromic Ontology (ILI-SSO) [13] is an extension of the SSO designed to supplement the limited consensus definitions found in the SSO, with the goal of providing a general NLP-oriented terminological resource for identifying Influenza-Like-Illness syndrome in clinical texts. The ILI-SSO is subsumed by the current work.

The BioCaster application ontology was built to facilitate text mining of news articles for disease outbreaks in several different Pacific Rim languages (Japanese, Thai, Vietnamese, Simplified Chinese, and so on) in addition to English [14]. It is used to power a real time, multi-lingual, publicly accessible online biosurveillance text mining system[8] that classifies news stories of epidemiological interest and populates a Google Map with geographically coded new cases. However, as the BioCaster system concentrates on *news reports*, representing the concepts, relations and lexical instantiations found in *clinical reports* is beyond the scope of the BioCaster ontology.

In addition to the application ontologies described above, the Infectious Disease Ontology[9] provides coverage of symptoms and diagnoses relevant to syndromic surveillance.

3 Developing the Ontology

Work began with the construction of a term list by author JD (a board certified infectious disease physician with thirty years of experience in clinical practice). The term identification process involved the domain expert reading multiple clinical reports, searching through textbooks and utilising professional knowledge. Terms were then consolidated into a list of concepts. Next, the concept list was compared to the Syndromic Surveillance Ontology, and concepts from the SSO reused where available. ESSO consists of 279 concepts (compared to 94 in SSO)

[7] The Web Ontology Language (OWL) is a World Wide Web Consortium standard for representing ontologies: `http://www.w3.org/TR/owl-ref/`

[8] `http://born.nii.ac.jp`

[9] `http://infectiousdiseaseontology.org`

spread across eight syndromes important to syndromic surveillance (see Table 1 for a list of syndromes and example concepts).

Table 1. ESSO Syndromes and Example Concepts

Syndrome	**No. Concepts***	**Example concepts**
Rash	33	Hives, Itching, Sores
Hemorrhagic	21	Hemoptysis, Melena, Epistaxis
Botulism	16	Botulism, BellsPalsy, SlurredSpeech
Neurological	52	Coma, Confusion, Headache
Constitutional	40	Fever†, Lethargy, Myalgia
InfluenzaLikeIllness	55	Fever†, Chill, Malaise
Respiratory‡	84	Plague, Rales, QFever
Gastrointestinal‡	30	AbdominalPain, Nausea, Rotavirus

* Number of concepts in each syndromic category

† Note that the SKOS data model allows "polyhierarchies" (for example, the concept Fever has `skos:broader` syndrome *InfluenzaLikeIllness* and *Constitutional*)

‡ *Respiratory* and *Gastrointestinal* syndromes are subdivided into **specific** and **sensitive** syndromes

The ontology is encoded in SKOS (Simple Knowledge Organisation System[10], a World Wide Web Consortium data standard for encoding thesauri and terminologies), with the syndromic hierarchical backbone of the ontology represented using `skos:narrower` and `skos:broader` (see Figure 1 for a screenshot of the Fever concept within the Protégé editor). Note that the Extended SSO subsumes all the concepts and relations present in the SSO, with all SSO concepts and relations reorganised to conform with the SKOS standard.

In addition to the standard thesaurus apparatus of preferred labels, alternative labels and hidden labels provided by SKOS, in order to facilitate "off the shelf" concept recognition, for each concept we include both regular expressions and links to external vocabularies. Table 2 provides a description of SKOS data relations for the concept Fever, while Figure 2 shows a simplified graph representation of the same concept.

The ontology is freely available under an open source licence.[11]

4 Evaluation

In recent years, significant research effort has focussed on evaluation methods for ontologies and terminologies [15, 16], yet no single "best practice" approach to ontology evaluation has emerged. We have adopted a "triangulation" strategy to audit the ESSO, concentrating on *coverage* (does the ontology contain

[10] `http://www.w3.org/2004/02/skos/`

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

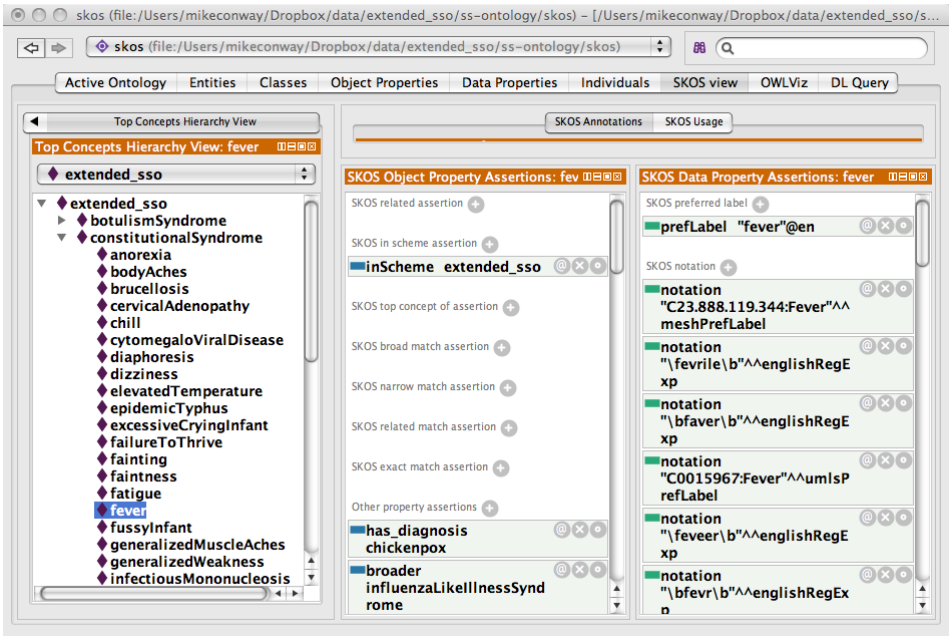

Fig. 1. Example of FEVER concept within the Protégé 4 Editor (SKOS-plugin)

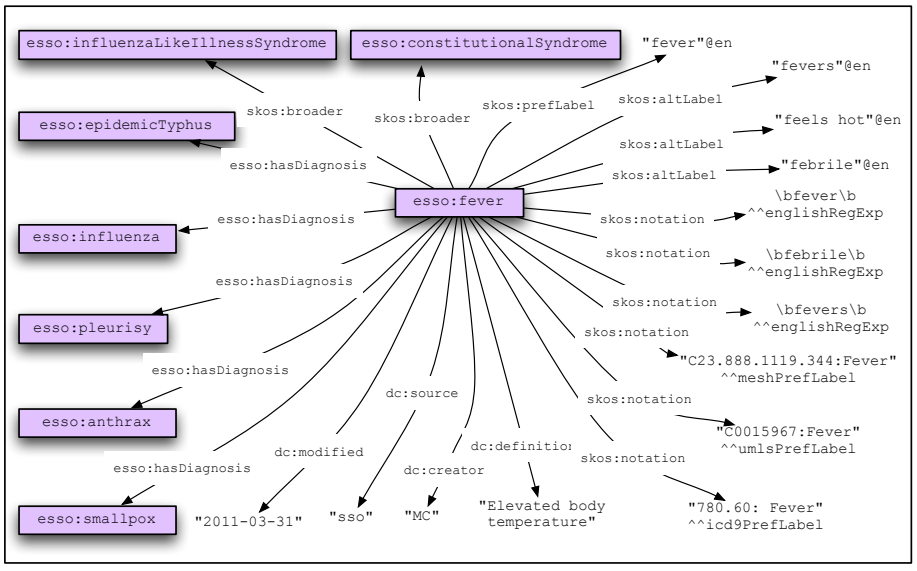

Fig. 2. Extended SSO Relations for the Concept FEVER

Table 2. Selected Relations for the Extended SSO Concept FEVER

Relation	Example
`skos:inScheme`[a]	FEVER `inScheme` EXTENDEDSSO
`skos:broader`[b]	FEVER `broader` CONSTITUTIONALSYNDROME
`skos:prefLabel`	FEVER `prefLabel` "fever"
`skos:altLabel`	FEVER `altLabel` "febrile"
`skos:notation^^umlsPrefLabel`[c]	FEVER `umlsPrefLabel` "C0015967"
`skos:notation^^meshPrefLabel`	FEVER `meshPrefLabel` "C23.888.119.344"
`skos:notation^^englishRegExp`	FEVER `englishRegExp` "\bfev\b"
`esso:has_diagnosis`	FEVER `hasDiagnosis` CHICKENPOX
`esso:dataCategory`[d]	FEVER `dataCategory` "sign"
`dc:creator`[e]	FEVER `creator` "MC"
`dc:source`	FEVER `source` "sso"
`dc:created`	FEVER `created` "2011-03-31"
`dc:modified`	FEVER `modified` "2011-03-31"
`dc:definition`	FEVER `definition` "Elevated body temperature"

[a] The `skos:inScheme` relation places a SKOS concept in a named Knowledge Organisation System

[b] `skos:broader` is read as "has broader category"

[c] `skos:notation` provides a mechanism for creating links to external vocabularies

[d] Clinical concept types are: *diagnosis*, *syndrome*, *sign*, *chest radiography*, and *bioterrorism disease*

[e] "`dc`" (Dublin Core) is a widely used metadata standard that can be used to augment SKOS with editorial information

the concepts we need for syndromic surveillance?), *relation quality* (are the relations in the ontology correct?) and *classification accuracy* (how well do the terms and regular expressions in ESSO perform at classifying clinical texts?). Currently, we have completed preliminary evaluation of ESSO's *coverage* of the target domain using a technique derived from terminology extraction and corpus linguistics [17]. First, we extracted terms from 300 Emergency Room reports[12] using the `TerMine`[13] term extraction tool [18]. We then went on to examine the twenty most statistically significant terms generated by `TerMine` (filtering out terms not relevant to the infectious disease domain) and found that only two of the `TerMine`-generated terms were not represented in ESSO — the two terms were "acute distress" and "apparent distress" — indicating that our domain coverage is adequate. Examples of significant terms extracted by `TerMine` which are contained in ESSO include "chest pain", "sore throat", "night sweat", and "vaginal bleeding."

[12] Deidentified Emergency Room reports were sourced from the University of Pittsburgh Medical Center.

[13] `TerMine` uses a combination of linguistic and statistical techniques to identify all terms in a document set, and then ranks these extracted terms according to their "termness". A web accessible version of the tool is hosted at: `http://www.nactem.ac.uk/software/termine/`

5 Conclusion

In conclusion, we have presented the Extended Syndromic Surveillance Ontology, an open source terminological resource designed to facilitate English language clinical text mining for syndromic surveillance. Our next task is to extend our preliminary evaluation to assessing relation quality and classification accuracy, with the medium term goal of using the ESSO as a gold standard against which we can evaluate new synonym extraction algorithms.

References

1. Henning, K.: What is Syndromic Surveillance? MMWR Morb Mortal Wkly Rep 53 Suppl, 5–11 (2004)
2. Wagner, M., Gresham, L., Dato, V.: Case Detection, Outbreak Detection, and Outbreak Characterization. In: Wagner, M., Moore, A., Aryel, R. (eds.) Handbook of Biosurveillance, pp. 27–50. Elsevier Academic Press (2006)
3. Tsui, F., Espino, J., Dato, V., Gesteland, P., Hutman, J., Wagner, M.: Technical Description of RODS: A Real-time Public Health Surveillance System. J Am Med Inform Assoc 10(5), 399–408 (2003)
4. Lombardo, J., Burkom, H., Elbert, E., Magruder, S., Lewis, S.H., Loschen, W., Sari, J., Sniegoski, C., Wojcik, R., Pavlin, J.: 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 (2003)
5. Cooper, D.: Case Study: Use of Tele-health Data for Syndromic Surveillance in England and Wales. In: Lombardo, J., Buckeridge, D. (eds.) Disease Surveillance: A Public Health Informatics Approach pp. 335–365. Wiley, New York (2007)
6. Eysenbach, G.: Infodemiology: Tracking Flu-Related Searches on the Web for Syndromic Surveillance. In: American Medical Informatics Association Annual Symposium Proceedings (AMIA 2006). pp. 244–248 (2006)
7. Centers for Disease Control: Recognition of Illness Associated with the Intentional Release of a Biologic Agent. MMWR Morb Mortal Wkly Rep 50(41), 893–7 (2001)
8. Kuehnert, M.J., Doyle, T.J., Hill, H.A., Bridges, C.B., Jernigan, J.A., Dull, P.M., Reissman, D.B., Ashford, D.A., Jernigan, D.B.: Clinical Features that Discriminate Inhalation Anthrax from Other Acute Respiratory Illnesses. Clin Infect Dis 36(3), 328–36 (2003)
9. Conway, M., Dowling, J., Tsui, R., Chapman, W.: Developing an Application Ontology for Mining Clinical Reports: The Extended Syndromic Surveillance Ontology. In: International Society for Disease Surveillance. Abstract (2010)
10. Okhmatovskaia, A., Chapman, W., Collier, N., Espino, J., Buckeridge, D.: SSO: The Syndromic Surveillance Ontology. In: Proceedings of the International Society for Disease Surveillance (2009)
11. Doyle, T., Ma, H., Groseclose, S., Hopkins, R.: PHSkb: A Knowledgebase to Support Notifiable Disease Surveillance. BMC Med Inform Decis Mak 5, 27 (2005)
12. Chapman, W., Dowling, J., Baer, A., Buckeridge, D., Cochrane, D., Conway, M., Elkin, P., Espino, J., Gunn, J., Hales, C., Hutwagner, L., Keller, M., Larson, C., Noe, R., Okhmoatovskaia, A., Olson, K., Paladini, M., Scholer, M., Sniegoski, C., Thompson, D., Lober, B.: Developing Syndrome Definitions Based on Consensus and Current Use. Journal of the American Medical Informatics Association 17, 595–601 (2010)

13. Conway, M., Dowling, J., Chapman, W.: Developing a Biosurveillance Application Ontology for Influenza-Like-Illness. In: Proceedings of the 6th Workshop on Ontologies and Lexical Resources. pp. 58–66. Coling 2010 Organizing Committee, Beijing, China (2010)
14. Collier, N., Matsuda Goodwin, R., McCrae, J., Doan, S., Kawazoe, A., Conway, M., Kawtrakul, A., Takeuchi, K., Dien, D.: An Ontology-Driven System for Detecting Global Health Events. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). pp. 215–222. Coling 2010 Organizing Committee, Beijing, China (2010)
15. Zhu, X., Fan, J.W., Baorto, D., Weng, C., Cimino, J.: A Review of Auditing Methods Applied to the Content of Controlled Biomedical Terminologies. Journal of Biomedical Informatics 42(3), 413 – 425 (2009)
16. Brank, J., Grobelnik, M., Mladenić, D.: A Survey of Ontology Evaluation Techniques. In: Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005). pp. 166–170 (2005)
17. Grigonyte, G., Brochhausen, M., Martin, L., Tsiknakis, M., Haller, J.: 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). pp. 331–344 (2010)
18. Frantzi, K., Ananiadou, S., Mima, H.: Automatic Recognition for Multi-word Terms. International Journal of Digital Libraries 3(2), 117–132 (2000)