Fast, Cheap and Out of Control: A Zero Curation Model for Ontology Development

Benjamin M. Good, Erin M. Tranfield, Poh C. Tan, Marlene Shehata, Gurpreet K. Singhera, John Gosselink, Elena B. Okon, and Mark D. Wilkinson

Pacific Symposium on Biocomputing 11:128-139(2006)

FAST, CHEAP AND OUT OF CONTROL: A ZERO CURATION MODEL FOR ONTOLOGY DEVELOPMENT

BENJAMIN M. GOOD, ERIN M. TRANFIELD, POH C. TAN,MARLENE SHEHATA*, GURPREET K. SINGHERA, JOHN GOSSELINK, ELENA B. OKON, MARK D. WILKINSON

*James Hogg iCAPTURE Centre for Cardiovascular and Pulmonary Research, St. Paul's Hospital, University of British Columbia, Vancouver, British Columbia V6Z1Y6, Canada. *University of Ottawa Heart Institute H355, Ottawa, K1Y-4W7, Canada*

During two days at a conference focused on circulatory and respiratory health, 68 volunteers untrained in knowledge engineering participated in an experimental knowledge capture exercise. These volunteers created a shared vocabulary of 661 terms, linking these terms to each other and to a pre-existing upper ontology by adding 245 hyponym relationships and 340 synonym relationships. While ontology-building has proved to be an expensive and labor-intensive process using most existing methodologies, the rudimentary ontology constructed in this study was composed in only two days at a cost of only 3 t-shirts, 4 coffee mugs, and one chocolate moose. The protocol used to create and evaluate this ontology involved a targeted, web-based interface. The design and implementation of this protocol is discussed along with quantitative and qualitative assessments of the constructed ontology.

Introduction

Ontologies provide the mechanism through which the "semantic web" promises to enable dramatic improvements in the management and analysis of all forms of data [1]. Already, the importance of these resources to the bio/medical sciences is made clear by the more than 1000 citations[a] of the original paper describing the Gene Ontology (GO) [2]. Because of the broad range of skills and knowledge required to create an ontology, they are generally slow and expensive to build. To illustrate, the cost of developing the GO has been estimated at upwards of $16M (Lewis, S, personal communication). This bottleneck not only slows the initial development of such systems but also makes them difficult to keep up to date as new knowledge comes available.

Conversely, projects such as DMOZ (http://dmoz.org) and BioMOBY[3][4] take a more open approach. Rather than paying curators, DMOZ lets "net citizens" build hierarchies (now utilized by Google among many others) that organize the content of the World Wide Web. BioMOBY, a web services-based interoperability framework, depends on an ontology of biological data objects

---

1064 Google Scholar citations (http://scholar.google.com) on Sept. 8, 2005

that can be extended by anyone. The successful, open, and ongoing construction of the DMOZ directories and the BioMOBY ontology hints that the power of large communities can be harnessed as a feasible alternative to centralized ontology design and curation.

We describe here a protocol meant to overcome the knowledge-acquisition bottleneck to rapidly and cheaply produce a useful ontology in the bio/medical domain. The key features of the approach are 1) the use of a web-accessible interface to facilitate collaborative ontology development and 2) the deployment of this interface at a targeted scientific conference. This paper describes the protocol and presents the results of a preliminary evaluation conducted at the 2005 Forum for Young Investigators in Circulatory and Respiratory Health (YI forum) (http://www.yiforum.ca/).

1.1. *Experimental context and target application for the YI Ontology*

The YI forum did not (outside of this study) include any research on knowledge capture or artificial intelligence. The topics covered spanned aspects of circulatory and respiratory health ranging from molecular to population-based studies, and analysis of quality of health-service provision. Attendees included molecular biologists, health service administrators, statisticians, cardio/pulmonary surgeons, and clinicians. The target task for the YI Ontology was to provide a coherent framework within which to organize the abstracts submitted to this broadly-based yet specialized conference. This framework would take the form of a simple subsumption hierarchy composed of terms associated with individual abstracts, and/or added by individual experts during the construction process. Such an ontology could be used to facilitate searches over the set of abstracts by providing legitimate, semantically-based groupings.

1.2. *Motivation and novelty of conference-based knowledge capture*

Research in natural language processing and machine learning is yielding significant progress in the automatic extraction of knowledge from unstructured documents and databases [5][6]; however these technologies remain highly error-prone and, to our knowledge, no widely used public ontology in the life sciences has ever been built without explicit, extensive expert curation. Thus, given the costs of curation, it would be preferable to identify methodologies that facilitate extraction of machine-usable knowledge directly from those who possess it. In order to achieve this, several preliminary steps seem necessary:

1. Domain experts need to be identified
2. These experts need to be convinced to share their knowledge.

3. These experts must then be presented with an interface capable of capturing their specific knowledge.

Scientific conferences seem to provide a situation uniquely suited to inexpensive, rapid, specialized knowledge capture because the first two of these requirements are already met by virtue of the setting; experts are identified based on their attendance and, at least in principle, they attend with the intention of sharing knowledge. Clearly, the principle challenge lies in generation of an interface that facilitates extremely rapid knowledge acquisition from expert volunteers.

2. Interface Design

The architecture chosen for this project borrows techniques from a new class of knowledge acquisition systems that attempt to harness the power of the Internet to rapidly create large knowledge bases. Projects in this domain are premised on the assumption that, by distributing the burden of knowledge representation over a large number of people simultaneously, the knowledge acquisition bottleneck can be avoided [7][8][9][10]. Two active projects in this domain are Open Mind Common Sense[10], and Learner2[11][12] . Both of these efforts focus on gathering "common sense" knowledge from the general public with the aim of producing knowledge-based systems with human-like capabilities in domains such as natural language understanding and machine translation.

These large, open, Internet-based projects are premised on the idea that there is little or no opportunity for explicit training of volunteers, and in principle no *strong* motivation to participate. This is similarly true of the conference participants engaged in this study, and thus based on these similarities, the interface developed for this knowledge capture experiment was modeled after the template-based interface of the Learner2 knowledge acquisition platform (http://learner.isi.edu).

Learner2 follows two basic design patterns:

1. Establish a system that allows the knowledge engineer to passively control knowledge base *structure*, while allowing its *content* to be determined entirely by the subject matter experts.
2. Use a web-enabled, template-based interface that allows all volunteers to contribute to the same knowledge base simultaneously and synergistically in real-time.

The "iCAPTURer" knowledge acquisition system presented here applies and adapts these principles to the task of knowledge capture in the conference setting.

2.1. *Specific challenges faced in the conference domain*

The iCAPTURer experiment faced unique challenges by virtue of its expert target-audience. Learner2 is designed to capture "common sense" knowledge, and operates by generating generic, user-agnostic fill-in-the-blank templates. For example, in order to collect statements about objects and their typical uses, a volunteer might be presented with "A [blank] *is typically used* to smash something" and asked to fill in the blank. In order to capture specific, expert knowledge however, it is necessary to adapt the contents of these templates to target each volunteer's specific domain of expertise. The following section details our adaptation of the Learner2 approach to meet this challenge.

3. Methods - Introducing the iCAPTURer

3.1. *Preprocessing*

Prior to the conference, terms and phrases were automatically extracted from each abstract using the TermExtractor tool from the TextToOnto ontology engineering workbench [5]. The TermExtractor was tuned to select multi-word terms using the "C-value" method [13]. This process produced a corpus of terms and phrases linked directly to the abstracts. This corpus provided the first raw material for the construction of the ontology and provided a mechanism to match the contents of the templates to the volunteer's area of expertise.

In addition, the nascent ontology was seeded with a concept hierarchy taken from the Unified Medical Language System Semantic Network (UMLSsn; http://www.nlm.nih.gov/research/umls/). The UMLSsn was selected as the "upper ontology" in order to provide a common semantic framework within which to anchor the knowledge capture process [14].

3.2. *Priming the knowledge acquisition templates - term selection*

Two priming models were employed to ensure that relevant knowledge was captured and that expert volunteers were presented with templates primed with concepts familiar to them. After logging into the system, the volunteer first makes a choice between priming the system with a keyword entered as free text, or priming the system through selection of a specific abstract (preferably their own).

In the abstract-driven model, the term to be evaluated is randomly selected from the pre-processed auto-extracted terms associated with the selected abstract. In this way, the expert is preferentially asked about terms from an abstract that they are presumptively familiar with, though there is nothing stopping them from selecting abstracts at random.

In the keyword-driven model, the system first checks the knowledge base for partial matches to the keyed-in term, and if found, selects one at random. If no matches are found the term is added to the knowledge base and is considered meaningful.

3.3. *Term evaluation*

After the volunteer chooses an abstract or enters a keyword, they are presented with the term-evaluation page. This page presents them with a term and requests them to decide if it is "meaningful", "not-meaningful", or if they do not understand it ("X is a meaningful term or phrase {True, False, I don't know}"). If they are unable to make a judgment on the term, another term is presented and the process repeats. If they indicate that the term is not valid, then the term's "truth value" is decremented in the knowledge base and another term is presented for judgment. Only terms above a set truth value are presented. This allows for rapid pruning of invalid entries from the active knowledge base without any permanent corpus loss. Approximately 50% of the terms extracted using text mining were judged nonsensical, hence this pruning was a critical step in the development of the ontology. If a term is rated as "meaningful", its truth value is raised and the term is considered selected.

3.4. *Relation acquisition*

Once a valid concept is selected, the system directs the volunteer to attach relations to the concept that will determine its position in the ontology. Two types of relation were targeted in this study, synonymy (same as) and hyponymy (is a).

To capture synonyms, a simple fill-in-the-blank template was presented. For example, if the term "muscle" was selected as valid, the volunteer would then be invited to enter synonyms through a template like: *The term or phrase [blank] means the same thing as "muscle"*.

A different format was used for capturing the hyponym relation. The hyponym template asks the volunteer to select a parent-term from a pre-existing hierarchical vocabulary (initially seeded with the UMLSsn) rather than letting them type one in freely. This approach was selected with the goal of producing a sensible taxonomic structure. During the knowledge capture process, terms

added to this hierarchy became new classes that future terms could be classified under, thus allowing the ontology to grow in depth and complexity.

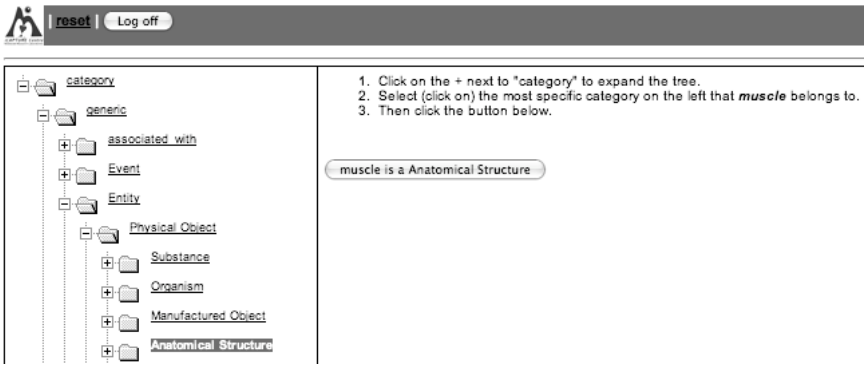

Figure 1: Hyponym collection. "Muscle" is being placed as a child of "Anatomical Structure".

As each task is completed, the volunteer is returned to a task selection screen and the completed task's button is removed. When each of the tasks are completed for the select term, another term is selected and the process repeats.

3.5. *Volunteer recruitment and reward*

To assist in volunteer-recruitment, conference attendees were motivated by a 5 minute introductory speech at the welcome reception, by flyers included in the conference handouts, and by the promise of mystery prizes for the most prolific contributors. Points were awarded to the user for each piece of knowledge added to the system. A simple user management system allowed the users to create accounts, log out, and log back in again while keeping track of their cumulative score throughout all sessions. Anonymous logins were also allowed.

4. Observations

In this preliminary study, qualitative observation of volunteer response to the system was a primary objective. As such, the enthusiastic response the project received from the organizers and the participants in the conference was encouraging, and the willingness of the volunteers to spend significant amounts of time entering their knowledge was unanticipated. From conversations with the participants, it became clear that the competitive aspect of the methodology was often their primary motivation, and this was especially true for the most prolific contributors who indicated a clear "determination to win". Some volunteers also indicated a simple enjoyment in playing this "intellectual game".

Another important observation was that the tree-based interface used to capture the hyponym relation (see figure 1) was not readily understood by the majority of participants. This interface required the user to understand relatively arbitrary symbols and to click multiple times in order to find the correct parent for the term under consideration. In contrast, the interface used in the later qualitative evaluation (discussed in section 6) required just a single click for each evaluation, resulting in no confusion or negative comments and *more than 11,000 collected assertions in just three days* from a similar number and composition of volunteers.

5. Quantitative Results

5.1. *Volunteer contributions*

During the 2 active days of the conference, 68 participants out of approximately 500 attendees contributed to the YI Ontology. Predominantly, volunteers contributed their knowledge during breaks between talks and during poster sessions at a booth with computer terminals set up for the purpose; however several participated from Internet connections in their hotel rooms. The quantity of contributions from the different participants was highly non-uniform, with a single volunteer contributing 12% of the total knowledge added to the system.

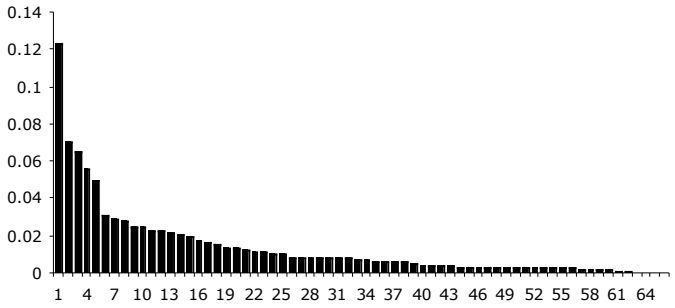

Figure 2 : Distribution of participant contributions. The X axis denotes the participant number, the Y-axis the fraction of the knowledge base contributed by that individual.

5.2. *Composition of the YI Ontology*

The pre-processing text mining step yielded 6371 distinct terms associated with the 213 abstracts processed. These auto-extracted terms were not added to the ontology until they had been judged meaningful by one of the volunteers via the term-evaluation template. 464 auto-extracted terms were evaluated by the conference volunteers. Of these, 232 were judged meaningful and 232 were

judged not meaningful. In addition, the 429 terms entered directly by volunteers (in the keyword initialization) were all considered to be meaningful. Thus in total the potential corpus for the ontology consisted of 661 validated terms.

Table 1. Captured Terms

	text-extracted	judged meaningful	judged not meaningful	Added directly	Total meaningful
Count	6371	232	232	429	661

5.3. *Relationships in the YI Ontology*

Of the 661 concepts, 207 were assigned parents in the UMLSsn rooted taxonomy. Of these, 131 concepts came from the auto-extracted set and 76 came from the directly entered set. As terms could be linked to different parents, 38 additional parental relationships were assigned to terms within this set, bringing the total number of hyponym relations assigned up to 245. 219 of the accepted terms were associated with at least one synonym, with many linked to multiple synonyms.

Table 2. Hyponyms

Total number of categories (including the UMLSsn)	469
Total categories added -at the YI forum	207
Added categories created from auto-extracted terms	131
Added categories created from terms added as keywords	76

Table 3. Synonyms

Total distinct targets (number of distinct synonyms entered)	340
Total distinct sources (number of terms annotated with a synonym)	219
Sources from auto-extracted terms	153
Sources from terms added as keywords	66

6. Quality Assessment

The evaluation of the YI Ontology was conducted in similar fashion to the initial knowledge capture experiment. Following the conference, the 68 participants in the conference study and approximately 250 researchers at the James Hogg, iCAPTURE Centre for Cardiovascular and Pulmonary Research were sent an email requesting their participation in the evaluation of the YI ontology. The email invited them to log on to a website and answer some questions in exchange for possible prizes. 65 people responded to the request. Upon logging into the website, the evaluators were presented with templates that presented a term, a hyponym relation, or a synonym relation from the YI

Ontology. They were then asked to make a judgment about the accuracy of the term or relation. For synonyms and hyponyms, they were asked to state whether the relationship was a "universal truth", "true sometimes", "nonsense", or "outside their expertise". For terms, they were asked whether the term was a "sensible concept", "nonsense", or "outside their expertise". After making their selection, another term or relation from the YI ontology that they had not already evaluated was presented and the process repeated.

Again, participants were provided motivation through a contest based on the total number of evaluations that they made (regardless of what the votes were and including equal points for indicating "I don't know"). Participation in the evaluation was excellent, with 5 responders evaluating every term and every relation in the ontology. During the three days of the evaluation, 11,545 votes were received, with 6060 on the terms, 2208 on the hyponyms, and 3277 on the synonyms. 93% of the terms, 54% of the synonyms and 49% of the hyponyms enjoyed more positive than negative votes overall.

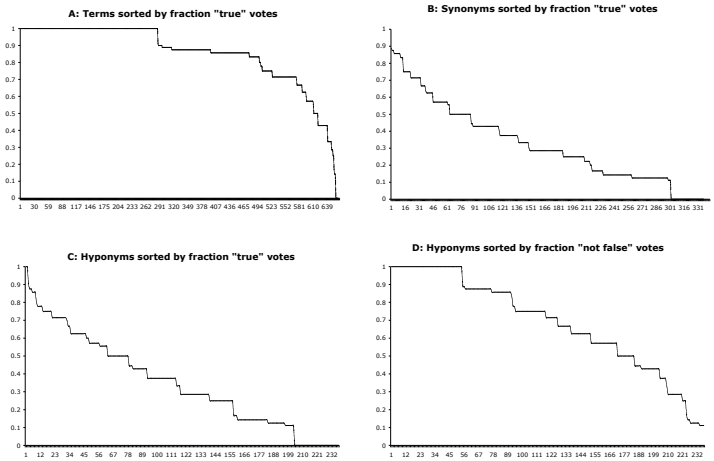

Figure 3 : The positive consensus agreement for captured terms (A), synonyms (B), and hyponyms (C). For A, B and C, the y-axis indicates the fraction of the votes for "universal truth". This value is used to sort the assertions indicated on the X-axis. The y-axis on D indicates the level of positive consensus for the hyponyms if the "true sometimes" votes are counted with the "universal truth" votes indicating a "not-false" category.

Figures 3a, 3b, and 3c display plots of the fraction of "true" votes received for each term, synonym and hyponym in the ontology. These curves illustrate strong positive consensus for the large majority of captured terms, but considerable disagreement regarding the quality of the captured synonyms and hyponyms. To some extent this may have been caused by the exclusion of the "sometimes" category from the term evaluations, but even when the

"sometimes" votes are merged with the "true" votes, there are still considerably fewer positive votes for the hyponyms and synonyms and less agreement among the voters. This is illustrated for the hyponyms in Figure 3d.

Table 4. Examples of assertions and associated votes

Assertion	% positive	% sometimes	% negative
Term: "*wild type*"	100	NA	0
Term: "*epinephrine e*"	50	NA	50
Term: "*blablala*"	0	NA	100
Hyponym: "*asthma* is a *disease*"	100	0	0
Hyponym: "*factor xiia* is a *coagulation factor*"	50	50	0
Hyponym: "*stem cells* are a kind of *transmission electron microscopy*"	0	11	89
Synonym: "*positive arrhythmia* is the same as *abnormal pacing of the heart*"	89	11	0
Synonym: "*lps treatment* is the same as *lipopolysaccaharide treatment"*	50	37.5	12.5
Synonym: "*Cd34* is the same as *aneurysm*"	0	14	86

Table 4 gives some examples of the contents of the YI ontology. These examples illustrate that the voting process successfully identified high quality components that should be kept, low quality components that should be discarded, and questionable components in need of refinement. These assessments could be used to improve the overall quality of the ontology through immediate pruning of the obviously erroneous components and by guiding future knowledge capture sessions meant to clarify those components lacking a strong positive or negative consensus.

Summary

Between April 29th and April 30th 2005, 661 terms, 207 hyponym relations, and 340 synonym relations were collected from 68 volunteers at the CIHR National Research Forum for Young Investigators in Circulatory and Respiratory Health. In a subsequent community evaluation, 93% of the terms, 54% of the synonyms and 49% of the hyponyms enjoyed more positive than negative votes overall. The rudimentary ontology constructed from these terms and relationships was composed at a cost of the 4 t-shirts, 3 coffee mugs, and one chocolate moose that were awarded as prizes to thank the volunteers.

Discussion

This work addresses the key bottleneck in the construction of semantic web resources for the life sciences. Ontology construction to date has proven to be extremely, possibly impractically, expensive given the wide number of expert

knowledge domains that must be captured in detail. Thus, it is critical that a rapid, accurate, inexpensive, facile, and enjoyable approach to knowledge capture be created and ubiquitously deployed within the life science research community. To achieve this, a paradigm shift in knowledge capture methodologies is required. The open, parallel, decentralized, synergistic protocol presented in this study represents a significant deviation from the centralized, highly curatorial model employed in the development of all of the major bio/medical ontologies produced to date.

The positive consequences of this approach are that 1) knowledge can be captured directly from domain experts with no additional training, 2) a far larger number and diversity of experts can be recruited than would ever be feasible in a centralized effort and 3) because the approach involves no paid curators, the overall cost of ontology development is very low.

The negative aspect of the approach is that the knowledge collected is "dirty", requiring subsequent cleaning to achieve high quality. Future versions of the iCAPTURer software will attempt to improve on the quality of the captured knowledge by integrating the evaluation phase directly with the knowledge capture phase. In this "active learning" approach, the questions will be tuned on-the-fly to direct knowledge capture efforts to areas of uncertainty or contention within the developing ontology and to quickly weed out assertions that are clearly false. The present study describes just one step of such a multi-step process, with obvious opportunities for immediate improvement in the next iteration based on the knowledge gathered during the evaluation.

In comparison to existing methodologies, which tend to separate the biologists from the ontologists, the iCAPTURer approach demonstrates dramatic improvements in terms of cost and speed. If future work confirms that this approach can also produce high quality ontologies, the emergence of a global semantic web for the life sciences may occur much sooner than expected.

Acknowledgments

Funding provided by Genome Canada, Genome British Columbia, Genome Prairie, the Canadian Institute for Health Research, and the Michael Smith Foundation. Thanks to Yolanda Gil and in particular to Timothy Chklovzki for important contributions during the design and conception of the iCAPTURer. Thanks also to the organizers of the YI Forum, in particular Ivan Berkowitz and Bruce McManus. We would also like to thank all of the volunteer knowledge engineers without whom this work would simply not be possible.

References

1. T. Berners-Lee, J. Hendler, and O. Lassila. "The Semantic Web". *Scientific American*, **284**:5 pp 34-43, May (2001)
2. M. Ashburner *et al*, "Gene Ontology: Tool for the Unification of Biology". *Nature Genetics*. **25**:1 pp 25–29 (2000)
3. M.D.Wilkinson, M. Links, "BioMOBY: an open-source biological web services proposal". *Briefings In Bioinformatics* 3:4. pp 331-344 (2002)
4. M.D. Wilkinson, H. Schoof, R. Ernst, D. Haase. "BioMOBY successfully integrates distributed heterogeneous bioinformatics web services. The PlaNet exemplar case", *Plant Physol* **138**, pp 1-13 (2005)
5. A. Maedche, S. Staab, "Ontology learning". In S. Staab and R. Studer, editors, *Handbook on Ontologies*, pp 173-189 (2004)
6. P. Cimiano, A. Hotho, S. Staab, "Clustering Concept Hierarchies from Text", in *Proceedings of 4th International Conference on Language Resources and Evaluation* (2004)
7. T. Chklovski, "Using Analogy to Acquire Commonsense Knowledge from Human Contributors", PhD. thesis, MIT Artificial Intelligence Laboratory technical report AITR-2003-002 (2003)
8. T. Chklovski. "LEARNER: A System for Acquiring Commonsense Knowledge by Analogy", in *Proceedings of Second International Conference on Knowledge Capture*. (2003)
9. M. Richardson, P. Domingos "Building Large Knowledge Bases by Mass Collaboration", In *Proceedings of the International Conference on Knowledge Capture* (2003)
10. P. Singh, T. Lin, E.T. Mueller, G. Lim, T. Perkins, W. L. Zhu, "Open Mind Common Sense: Knowledge Acquisition from the General Public", *Lecture Notes in Computer Science*, **2519**, pp 1223–1237 (2002)
11. T. Chklovski, "Designing Interfaces for Guided Collection of Knowledge about Everyday Objects from Volunteers", in *Proceedings of 2005 Conference on Intelligent User Interfaces* (2005)
12. T. Chklovski, Y. Gil, "Towards Managing Knowledge Collection from Volunteer Contributors", in *Proceedings of 2005 AAAI Spring Symposium on Knowledge Collection from Volunteer Contributors* (2005)
13. K.T. Frantzi, S. Ananiadou, J. Tsujii, "The c-value/nc-value method of automatic recognition for multi-word terms", *Lecture Notes in Computer Science*, **1513**, pp 585-600. (1998)
14. Niles, A. Pease, "Towards a Standard Upper Ontology", in *Proceedings of the international conference on Formal Ontology in Information Systems* (2001)