Extending the HPKB-Upper-Level Ontology: Experiences and Observations

Stuart Aitken[1]

Abstract. This paper describes our experience of extending the HPKB-upper-level ontology. Reuse by extension is key to reuse of generic upper-level ontologies, and we report on the use of structuring principles in this task. We argue that the documentation of design rationale is key to the reuse of this type of ontology, and that the HPKB-upper-level ontology would benefit from reorganisation.

1 INTRODUCTION

This paper describes an extension to the HPKB-upper-level ontology to cover *information sources* in more detail. The HPKB-upper-level ontology, which is available from the Ontolingua server [7], is intended for use in solving the "challenge problems" which have been devised as technology testbeds on the DARPA High Performance Knowledge Bases (HPKB) project [5]. AIAI are part of the HPKB programme. A major component of the first challenge problem requires searching the Web for information to answer queries about a (hypothetical) political crisis; the ability to characterise Web-based information sources in a way which identifies their ability to answer a question and the reliability of the answer is therefore important. This paper identifies concepts which would need to be represented in such an ontology, and shows how they can be implemented in Cyc.

It has been noted that there is relatively little methodological support for ontology development [1], and few reported studies on the extension of ontologies [11]. This paper describes our experience of extending an existing ontology in order to provide concrete examples of the issues and problems encountered. We then present an analysis of some of the more important issues which arise in the reuse of ontologies which define a generic upper-level conceptualisation.

Methodologies for ontology construction typically assume that a new ontology is being constructed. A middle-out approach to ontology construction has been proposed [10]. The major steps include scoping, grouping and cross-referencing concepts, producing definitions, and determining work areas. Terms in the identified work areas are then defined in middle-out fashion. It is argued that the middle-out approach avoids problems such as going into too much detail (associated with bottom-up approaches) and imposing arbitrary high-level categories (associated with top-down approaches) [10].

Project management, development-oriented, and support activities in ontology development are supported by the Methontology approach [1], which aims to specify a method for creating ontologies at a level above the language-encoding level. Ontology development includes producing a glossary of terms, and drawing diagrams such as concept classification trees and binary-relation diagrams to illustrate the connections between concepts. Terms may be drawn from other ontologies, but this is a different reuse problem to that of extending an existing ontology.

Generic ontologies which provide high-level concepts, such as event, agent, thing, and state, lack the modular structure advocated by Borst [2] and tend to have a homogeneous structure at the middle and lower levels. Terms in this type of ontology may be grouped into work areas. For example, concepts in the HPKB-upper-level ontology are grouped into 43 topical groups. These include Agents and Roles, which describe concepts and sets of relations which are central to the organisation of the ontology, as well as groups such as Emotion and Medicine which are more topic-based. In the Enterprise ontology [12], there are five work areas (all related to enterprise modelling) and, as in the HPKB-upper-level ontology, terms in each area are interrelated.

The Cyc approach to ontology development identifies a number of opposing concepts which can be used to structure the ontology. The concepts stuff-like and object-like can refer to the temporal dimension and to the nature of a substance. Events are temporally object-like, while things that exist through time, e.g. books, are temporally stuff-like as at all sub-intervals they are the same thing. However, books are object-like in nature as they cannot be subdivided and remain the same thing, unlike water, for example. We explore the use of this type of organising principle in the extension we propose in this paper. Other opposing concepts include: tangible vs. intangible, static vs. dynamic, and individual vs. collection.

Generic ontologies also differ in the degree to which they can be validated (validation is discussed further in [2]). Engineering maths and topology ontologies are capable of being validated by reference to literature in their application fields. The HPKB-upper-level, Enterprise [12] and SPAR [8, 9] ontologies do not capture knowledge in such well understood fields, therefore this form of validation is not possible. However, validation remains an important issue.

In the case study presented here, the upper ontology was already defined, and a small set of concepts were to be added. The problems we faced included the task of understanding the existing conceptualisation, but nonetheless a middle-out approach of scoping, understanding the existing ontology, then introducing intermediate level classes (i.e. classes immediately below the existing upper ontology) was productive. We describe the problems that arose in making what appeared to be 'natural' extensions to the ontology, and discuss the underlying modelling issues.

A case study of ontology extension is presented in Section 2. This is followed by a review of the modelling decisions made in the initial extension, and those implicit in the relevant section of the upper-level ontology. Section 3 also presents a revised ontology for information sources.

[1] AIAI, University of Edinburgh, Edinburgh EH1 1HN, UK. E-Mail: stuart@aiai.ed.ac.uk

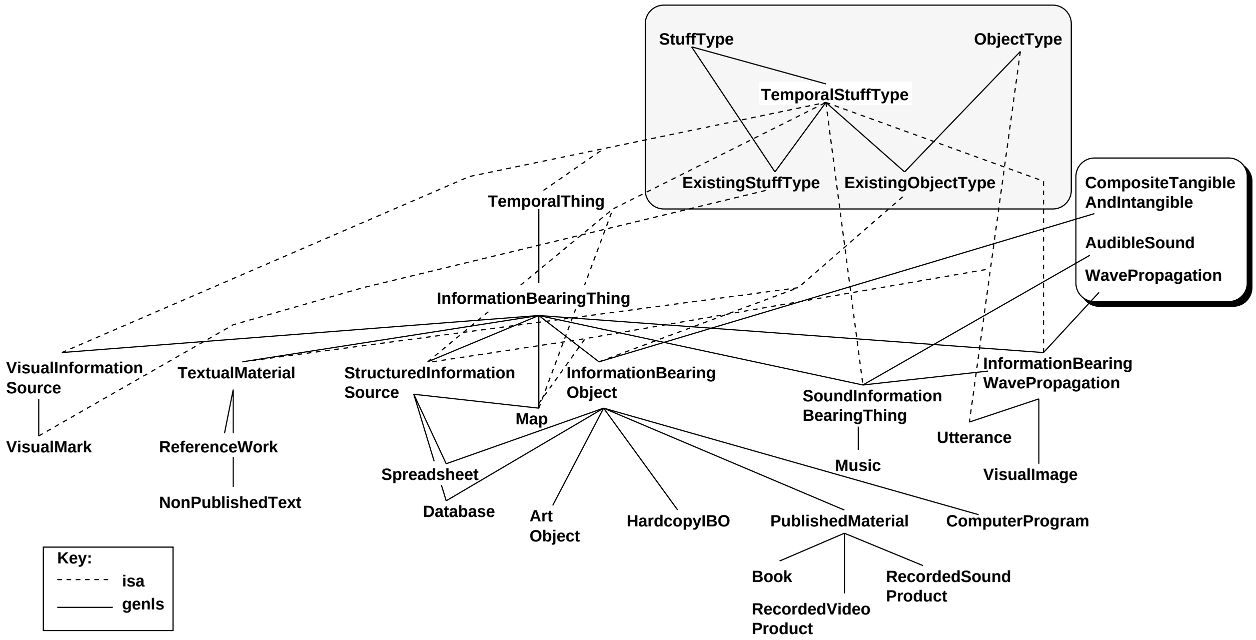

Figure 1. The existing IBT collection hierarchy

2 CASE STUDY

This section presents extensions to the Cyc BaseKB which enable explicit reasoning about the sources of information that are available to the user, or to Cyc itself. The BaseKB contains the HPKB-upper-level ontology. The domain of interest was constrained to the sources of information which were identified as being relevant to solving one of the HPKB challenge problems. Our main aims were to extend the HPKB-upper-level ontology sufficiently to cover the concepts of interest, and to gain a better understanding of the modelling issues involved.

Some types of information source are already represented in the upper-level ontology, for example, books and maps. We propose a number of new sources, and a number of intermediate-level classes which characterise new source-types. A comparable ontology of documents has been posted on the Ontolingua server [7], however, many of the classes identified there already exist in the upper-level ontology (under a very different organisation).

Acquiring information may require some actions to be taken in the world. There will be some time associated with such actions, and perhaps some risks will be involved. The BaseKB contains a model of *events* which includes events that create information-bearing objects. We have reused these existing definitions to create a model of information gathering events which is integrated into the event and temporal models, but these extensions will not be presented here.

2.1 The Domain: Information Sources

The following information sources are representative of those used in answering challenge problem (CP) questions:

- Energy Information Administration pages
- CIA factbook
- U.S. State Department Human Rights Report
- Jane's Undersea Warfare Systems (on-line)
- the Fisher Model (a listing of air capability resources)

These information sources can be characterised by capturing the type of publication: *book*, *HTML page*, *newspaper*, *model*, and *letter*; the publication media: *hardcopy* and *softcopy*; and attributes such as *authorship*, *credibility*, *language*, and *subject area*.

It would be expected that types of publication would be modelled taxonomically, and that media would be an attribute or a property. However, this is not the case in the existing ontology and we examine these issues in detail. The attributes identified above are modelled as would be expected (by relations) and no interesting issues arise.

The information content of an information source is a distinct entity from the information source itself and is represented by *PropositionalInformationThing* (PIT) class in the upper-level ontology. The predicate *containsInformation* relates information sources to PITs. No further treatment of this issue is required for our purposes.

2.2 Relevant Upper-Level Collections

The most relevant collection containing information sources is *InformationBearingThing* (IBT). The most relevant collections containing events are *Actions* and *InformationTransferEvents*. Some useful predicates which connect these are: *products* which can take a InformationTransferEvent and an IBT as arguments, to state that the IBT is the product of the event, and *duration* which holds of an event and the time the event lasted for. Assuming that we can represent typical examples of information gathering events, and their typical durations, these classes and predicates provide a means of representing noth objects and processes in information gathering.

2.3 Information Bearing Things

InformationBearingThings are categorised according to whether they are textual, structured, visual, or whether they are objects. An IBT may belong to several of these classes. Figure 1 shows the *genls* (subset) relations for the IBT collection. This diagram also shows the *genls* and *isa* links between IBT collections and other upper-level collections: these will be discussed in more detail later.

A number of collections of IBTs have an obvious meaning and could simply be instantiated to represent information sources in any of the HPKB challenge problem domains, for example: *Book*, *ComputerProgram*, *Database* (e.g. the Cyc BaseKB), *Map*, *RecordedSoundProduct*, *RecordedVideoProduct*, *Spreadsheet*, *Utterance*, *VisualImage*.

This list of information sources does not include all the concepts we require, for example, HTML pages are not included. In addition, this list does not make all the distinctions we might require, e.g. Books may be in paper-copy only, or also available in some electronic form.

2.4 New InformationBearingThings

The existing upper-level concepts of Book and Database are organised initially by what appears to be a concept of organisational form, i.e. textual, structured, visual, and the object/stuff-like distinction. For example, *InformationBearingObjects* is an instance of *ExistingObjectType* which means that it is a collection of spatially object-like things (i.e. things which are indivisible), but is temporally stuff-like, meaning that its instances exist through time.

The definition of ExistingObjectType begins: *"A collection of collections. Each element of each element of ExistingObjectType is temporally stufflike yet is objectlike in other ways, e.g., spatially. Any one of many timeSlices of a copy of 'Moby Dick' sitting on your shelf is still a copy of 'Moby Dick' sitting on your shelf. Most tangible objects are temporally stufflike in this fashion. That book is, of course, not spatially stufflike; spatially, it is objectlike: if we take a scalpel and slice the book into ten pieces, each piece is not a copy of 'Moby Dick'. [...]"* (Copyright 1995, 1996, 1997 Cycorp. All rights reserved.)

Not all IBTs are spatially object-like as *VisualMarks* are spatially stuff-like. All IBTS are temporally stuff-like and none are temporally object-like.

At the second level of decomposition, concepts such as 'being published' and 'in hardcopy form' are introduced as collections. Distinct types of publication are then introduced. Concepts are used to model types of information source and their properties. Noting this, we will adhere to this approach as far as possible, and will postpone criticisms of the hierarchy structure until Section 3.

Two new subcollections of Information Bearing Object (IBO) are introduced into the existing hierarchy in order to represent the new domain concepts: *SoftcopyIBO* and *Message*. Figure 2 shows the *genls* relations of the new collection hierarchy. SoftcopyIBO is introduced as a counterpart to HardcopyIBO. The natural place to locate this class is below IBO. Specifying this collection enables a distinction to be made between the electronic and paper versions of information bearing objects. Without such a collection it would not be possible to state the publication medium of IBOs such as HTML pages. The Message collection contains IBOs such as letters that are written for an identified reader. The recipients may constitute a group, in which case the IBO would be considered to be published. This collection allows a distinction to be made between letters and email, and other unpublished textual or electronic material. Messages have the spatially object-like property of IBOs in a similar way to Books. However, as they may be unpublished this collection cannot be located under Published Material and has been located directly below IBO.

The new subcollections of IBO allow HTMLPages to be defined as published material in softcopy. *PlainHTMLPages* are essentially textual, and hence this is a specialisation of HTMLPage. *Letters* and

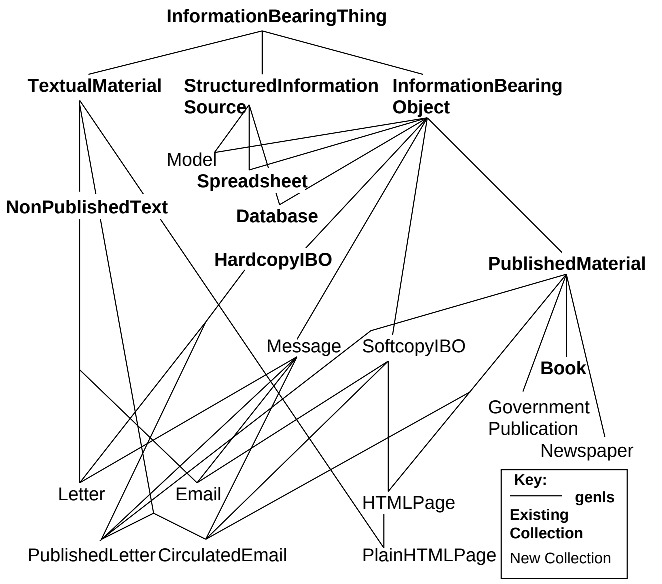

Figure 2. The extended IBT collection hierarchy

Email are IBOs which are textual, and have an identified recipient. They differ as to their publication medium. Email that is circulated, and letters that are published become published textual material, as opposed to unpublished text.

The concepts of published material and non-published text must be mutually exclusive. As a result we are led to define different classes for *Letters* and *PublishedLetters* where otherwise we might say that only an attribute of the object (the letter) changes on publication. This situation arises from the extensions which were introduced on largely intuitive grounds, and from the existing hierarchy structure where published material and non-published text are concepts which hold of types of publication. One remedy is to retract the `(genls Letter NonPublishedText)` assertion and allow instances of letter to be non-published text or published material as appropriate. However, the more general issues of how to structure the ontology to make it more understandable, and more amenable to extension need to be addressed.

3 A CRITIQUE OF THE IBT ONTOLOGY

The foregoing discussion suggests that it might be beneficial to review the structure of the existing ontology of information sources, prior to extending it. The HPKB-upper-level ontology is not composed from modules and it does not appear feasible to introduce this type of organisation. However, the structuring principles and the rationale for the design of the IBT ontology can be examined and clarified, and some validation given for the concepts used.

The principles underlying the structure of the existing hierarchy are not clear. At the first level of decomposition of IBT there are seven classes. Three of these (InformationBearingObject, SoundInformationBearingThing and InformationBearingWavePropagation) generalise to classes outside of IBT, see the shadowed box in Figure 1, and so have distinguishing features. Of these three classes, one is a subclass of another - indicating some redundancy in the *genls* definitions. Of the four remaining classes, Map and StructuredInformationSource have the same *isa* links and Map is a subset of StructuredInformationSource.

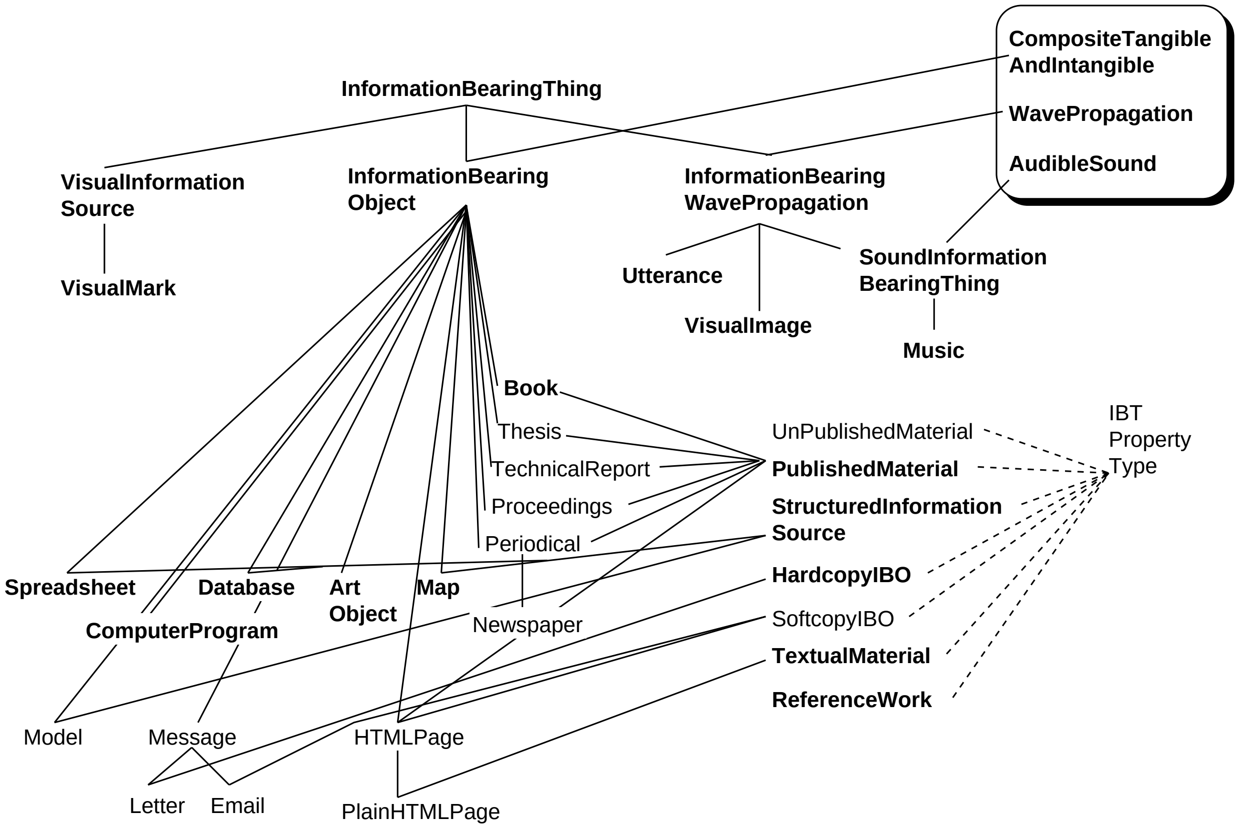

Figure 3. A proposal for restructuring the IBT hierarchy

TextualMaterial and InformationBearingObject (IBO) are of type ExistingObjectType. This in turn generalises to both TemporalStuffType and ObjectType. Map and StructuredInformationSource are stated to be of types TemporalStuffType and ObjectType - types which they would appear to belong to if they had also been defined to be of ExistingObjectType. In summary, the classes Map, StructuredInformationSource, and TextualMaterial appear to make no distinctions which are not made in InformationBearingObject. With the exception of Map, the subclasses of StructuredInformationSource generalise to IBO, indicating that there should probably be a subset relation between them. One subclass of TextualMaterial generalises to IBO, while the others do not. It is hard to explain why these subclasses do not have the IBO properties.

The VisualInformationSource class is distinguished by its lack of properties: it is not an ExistingObjectType and does not generalise to any class outside IBT. This class is only stated to be of TemporalStuffType. The definitions allow the subclass VisualMark to be existing-stuff-like in nature (note that ExistingStuffType and ExistingObjectType form a partition). In general, IBTs are not spatially stuff-like as (presumably) dividing an IBT into pieces will also divide the information encoded there.

In conclusion, there appear to be three distinct subclasses of IBT: InformationBearingWavePropagation, InformationBearingObject, and VisualInformationSource.

We suggest a reorganisation of the IBT heirarchy. The three distinct subclasses of IBT identified above form the first level of decomposition. Publication kinds are subclasses of InformationBearingObject. The kinds of IBOs can be reorganised as a taxonomy, and properties can be defined as concepts. The proposed revision of the IBT hierarchy is shown in Figure 3. Concepts may be an inherent part of the definition of a publication type, e.g. books must be published, and email must be softcopy. Exclusive properties such as being published, or not, should not be assigned to certain publication types if that property may change, e.g. subclasses of Message cannot be unpublished by definition (as noted above).

Figure 3 includes the publication types of the original IBT ontology, plus those introduced above. Additional classes are taken from the Documents ontology [7], namely: Thesis, TechnicalReport, Proceedings, and Periodical (subclasses of these are not illustrated for reasons of clarity). The opportunity of reorganising the IBT hierarchy allows the structure of that ontology to be adopted in part.

The collection *IBTPropertyType* is introduced as a collection of collections of IBTs. The members of this class are collections which define properties such as publication media, published or unpublished, textual and structured. The names of these properties in original ontology have retained where possible.

The design rationale of the proposed hierarchy has been stated, and examples given. Further, validation of the ontology is made possible by stating the origin of the terms used. By these means we believe we have increased the possibility that others could make additional extensions to the IBT ontology, and do so in a systematic way.

4 DISCUSSION

The experience of reusing the HPKB-upper-level ontology has highlighted a number of issues. The opposing concepts stuff-like vs. object-like are applicable to the spatial and temporal descriptions of information bearing things. However, all IBTs are temporally stuff-like and the majority of classes are spatially object-like. The stuff-object dimension is not useful as an organising principle at the level of the ontology we considered. This is not to dispute its use at higher levels, or in other regions of the ontology.

It was apparent that concepts were used to model both properties and kinds of information sources within the same hierarchy. An attempt to extend this ontology led to modelling errors. Further examination of the ontology revealed some redundancies in the *genls* and *isa* structure. These are not significant in operational terms, but further complicate the task of understanding the structure and design rationale of the ontology.

We have considered generic ontologies which formalise consensus, or common-sense knowledge, i.e. knowledge which is not drawn from established fields of study, and which cannot be formally verified. We conclude that statements of principles, or guidelines, of ontology design, along with the provision of information which allows some validation of the ontology are particularly important for the reuse and extension of this type of generic ontology.

ACKNOWLEDGEMENTS

This work is sponsored by the Defense Advanced Research Projects Agency (DARPA) under grant number F30602-97-1-0203. The U.S. Government is authorised to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation hereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing official policies or endorsements, either express or implied, of DARPA, Rome Laboratory or the U.S. Government.

REFERENCES

[1] Blázquez, M., Fernández, M., Garcia-Pinar, J.M., and Gómez-Pérez, A. Building Ontologies at the Knowledge Level using the Ontology Design Environment. *Proceedings of KAW'98 Eleventh Workshop on Knowledge Acquisition, Modeling and Management*, Banff, 1998. URL: `http://ksi.cpsc.ucalgary.ca/KAW/KAW98/blazquez/`

[2] Borst, P., Akkermans, H., Top, J. Engineering Ontologies. *International Journal of Human-Computer Studies*, Vol. 46, No. 2/3, pp. 365–406.

[3] Gruber, T.R. and Olsen, G.R. An Ontology for Engineering Mathematics. *Fourth International Conference on Principles of Knowledge Representation and Reasoning*, Doyle, J., Torasso, P., and Sandewall, E. (Eds.), Morgan Kaufmann, 1994. URL: `http://www-ksl.stanford.edu/knowledge-sharing/ontologies/html/engineering-math.text.html`

[4] van Heijst, G., Schreiber, A.T., and Wielinga, B.J. Using Explicit Ontologies in KBS Development. *International Journal of Human-Computer Studies*, Vol. 46, No. 2/3, pp. 183–292.

[5] HPKB Information about the HPKB program can be found at URL: `http://www.teknowledge.com/HPKB/`

[6] Lenat, D. *Leveraging CYC for HPKB Intermediate-level Knowledge and Efficient Reasoning* URL: `http://www.cyc.com/hpkb/proposal-summary-hpkb.html`

[7] Ontolingua Ontology Server has URL: `http://www-ksl-svc.stanford.edu`

[8] Shared Planning and Activity Representation (SPAR) URL: `http://www.aiai.ed.ac.uk/~arpi/spar`

[9] Tate, A. Roots of SPAR. to appear in the *Knowledge Engineering Review, Special Issue on Putting Ontologies to Use*, (eds.) Tate, A., and Uschold, M.F., 1998.

[10] Uschold, M. and Gruninger, M. Ontologies: principles, methods and applications. *Knowledge Engineering Review*, Vol. 11:2, 1996, pp. 93–136.

[11] Uschold, M., Clark, P., Healy, M., Williamson, K., and Woods., S. An Experiment in Ontology Reuse. *Proceedings of KAW'98 Eleventh Workshop on Knowledge Acquisition, Modeling and Management*, Banff, 1998. URL: `http://ksi.cpsc.ucalgary.ca/KAW/KAW98/uschold/`

[12] Uschold, M., King, M., Moralee, S., and Zorgios, Y. The Enterprise Ontology. to appear in the *Knowledge Engineering Review, Special Issue on Putting Ontologies to Use*, (eds.) Tate, A., and Uschold, M.F., 1998. Also available from AIAI as AIAI-TR-195, URL:`http://www.aiai.ed.ac.uk/~entprise/`