School of Informatics, University of Edinburgh

Centre for Intelligent Systems and their Applications

Merging Top Level Ontologies for Scientific Knowledge Management

by

John Kingston

Informatics Research Report EDI-INF-RR-0171

School of Informatics **August 2002**
http://www.informatics.ed.ac.uk/

Merging Top Level Ontologies for Scientific Knowledge Management

John Kingston

Informatics Research Report EDI-INF-RR-0171

SCHOOL *of* INFORMATICS
Centre for Intelligent Systems and their Applications

August 2002

Proceedings of the AAAI workshop on Ontologies and the Semantic Web, AAAI-02 conference, Edmonton, Canada, 29 July 2002.

Abstract :

A group of researchers from the EPSRC-sponsored Advanced Knowledge Technologies (AKT) project had independently developed four different ontologies covering the same domain: scientific knowledge management (academics, papers, conferences, etc). A reference ontology was developed to allow communication between these four ontologies. It was proposed that the reference ontology would benefit from incorporating the top level of the OntoClean ontology (Gangemi et al, 1998) because of the meta-properties that this ontology uses to support ontology structuring decisions. A merged ontology structure is proposed and the issues that arise in the merging process are analysed and discussed.

Keywords : Ontology merging, knowledge management

Copyright © 2003 by The University of Edinburgh. All Rights Reserved

The authors and the University of Edinburgh retain the right to reproduce and publish this paper for non-commercial purposes.

Permission is granted for this report to be reproduced by others for non-commercial purposes as long as this copyright notice is reprinted in full in any reproduction. Applications to make other use of the material should be addressed in the first instance to Copyright Permissions, School of Informatics, The University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, Scotland.

Merging Top Level Ontologies for Scientific Knowledge Management

John Kingston

Artificial Intelligence Applications Institute
CISA, Div. of Informatics, University of Edinburgh
Edinburgh EH1 1HN, Scotland
J.Kingston@ed.ac.uk

Abstract

A group of researchers from the EPSRC-sponsored Advanced Knowledge Technologies (AKT) project had independently developed four different ontologies covering the same domain: scientific knowledge management (academics, papers, conferences, etc). A reference ontology was developed to allow communication between these four ontologies. It was proposed that the reference ontology would benefit from incorporating the top level of the OntoClean ontology (Gangemi et al, 1998) because of the meta-properties that this ontology uses to support ontology structuring decisions. A merged ontology structure is proposed and the issues that arise in the merging process are analysed and discussed.

Introduction

A group of researchers from the EPSRC-sponsored Advanced Knowledge Technologies (AKT) project had independently developed four different ontologies covering the same domain: scientific knowledge management (academics, papers, conferences, etc). These four ontologies – one from the Open University (OU), one from the University of Southampton, and two from the University of Edinburgh -- were collected together and the common high level concepts were identified. These concepts were as follows:

- Publications and other Documents;
- Events - especially Conferences, Seminars, Meetings and Workshops;
- Organisations, including Universities and Research Groups, and also Publishers and Funding Bodies;
- People - generally classified by their role (Academic Staff, Students, Secretaries, etc)
- Research Areas;
- Projects and other Tasks.

The researchers decided to create a "reference ontology" of this scientific knowledge management (SKM) domain that each of the ontologies could communicate to and from. The aim was to develop an ontology that could be used as a communication device between a group of researchers working in the same field, while applying techniques, tools and principles for ontology development. A workshop was held where each of the top level concepts was discussed individually, with the aim of determining its superclasses, subclasses and slots within the reference ontology. Inputs to the discussion included the definitions of concepts in the four existing ontologies; definitions of concepts in other respected ontologies (e.g. the IEEE Standard Upper Ontology); a discussion document which suggested some principles for ontological allocation, based on the proposed OntoClean approach of Guarino (Gangemi et al, 1998) and on the Cyc upper ontology (Cycorp, 2001); and the experience of the various participants. The resulting concept definitions were then built into a reference ontology.

After the reference ontology had been published internally, a discussion ensued between the participants in the workshop regarding the best top level structure of the ontology. This paper proposes a revision to the reference ontology to include the OntoClean top level ontology, in order to allow the use of OntoClean's ontology structuring principles in further development of the reference ontology. This paper therefore discusses the structure and underlying principles of OntoClean; looks at the structure of the reference ontology; then considers the issues raised in merging the two.

The OntoClean ontology

The OntoClean ontology was originally proposed by Nicola Guarino and his researchers as a "cleaned up" version of the WordNet (Fellbaum, 1998) taxonomy. The OntoClean approach proposes that top level ontological concepts should be determined by a number of meta-properties. These meta-properties are:

- Rigidity – does this property *necessarily* hold for this individual?
- Identity – can all instances be identified by a suitable "sameness" relation?
- Dependence – can an individual be present if another individual is not fully present?

Copyright © 2002, American Association for Artificial Intelligence
(www.aaai.org). All rights reserved.

- Extensionality – an individual is extensional if everything that has the same proper parts is identical to it.
- Concreteness – an individual is concrete iff it has a physical location.
- Unity, singularity and plurality – a property carries unity if there is a common unifying relation such that all the instances are "essential wholes". Singularity implies a strong topological relation between its instances, while plurality implies a "sum" of singular wholes; if they form a whole themselves they are a *collection*, and if not they are a *plurality*.
- There is also an additional category of *types and roles:* properties can be defined as types, formal roles or material roles according to their rigidity, identity and dependence. Types and roles are "formal meta-categories based on multiple meta-properties" (Gangemi et al, 1998).

Based on these distinctions, a revised top level ontology for the WordNet system is proposed. This ontology consists of the following concepts:

- Aggregates (anti-dependent, anti-unity). There are two main subcategories: *Amounts of matter*, sometimes called groups or sets, these are extensional, because they change their identity when they change some parts; and pluralities, "mere sums of wholes which are not themselves essential wholes", which change their identity if a member is changed, but may allow a change in the parts of a member without a change in identity.
- Objects (anti-dependent, part-unity[1]). The main distinction is between physical bodies (extensional) and ordinary objects (non-extensional). There also appears to be a third category for Life Forms.
- Events (dependent, extensional). These are "temporal occurrences" which are assumed to be dependent upon their participants. All parts of an event are assumed to be essential parts; if any of the parts changed, it would be a different event.
- Features (dependent, extensional, part-unity). These are "parasitic entities" that exist insofar as their host exists. They may be relevant parts (e.g. an edge) or dependent regions (e.g. the shadow of a tree). They are typically singular as they have topological unity.
- Qualities (dependent, extensional, unity). A property such as *red, big* or *sweet* is the result of classifying a quality according to a given *conceptual space*. For example, "poor" in much of the USA is likely to equate to "rich" in parts of Africa, because the conceptual space is different. Individual qualities are assumed not to change as the conceptual space changes.
- Abstractions (anti-concrete). These are entities that do not have a physical location. They divide into Abstract Entities and Quality Spaces; the latter include conceptual spaces such as time, geometric length, and colour.

The full proposal for OntoClean's top level ontology, along with some proposed assignments of key concepts from WordNet, can be seen in Figure 1, which is taken from (Gangemi et al, 1998).

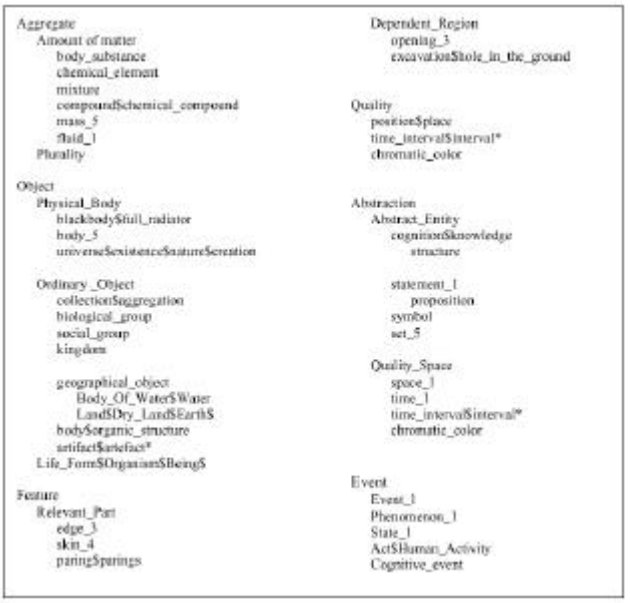

Figure 1: The proposed OntoClean top level ontology

The AKT reference ontology

Each of the six proposed top level concepts of scientific knowledge management was discussed individually at the workshop. The discussion made use of various ontology structuring principles, one of which was the principle of **re-use** – the contents of previous ontologies were re-used whenever feasible. Other principles followed included the principle of **dependence** (e.g. the existence of a Publication is dependent on the existence of a Document or other Information Bearing Object, and so Publication is placed directly beneath Information Bearing Object in the ontology[2]); the contrasting principles of **maximal detail** and **minimum commitment**; and the principle of **multi-perspective modelling** (Kingston, 2001), which was used to differentiate Technology, Research Area and Method.

[1] Every instance of P is an essential whole but there is no unifying relation common to all instances of P.

[2] This isn't an ideal way of representing dependence. This point is returned to in the discussion at the end of this paper.

An overview of the resulting ontology can be seen in Figure 2; the top level structure (Tangible Thing, Intangible Thing, Temporal Thing) is taken from the Open University's existing ontology of scientific knowledge management concepts. For more details on the structure of the AKT reference ontology, see (AKT, 2001).

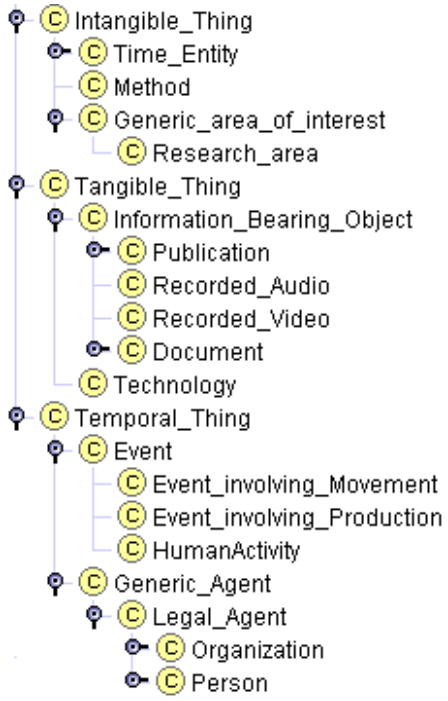

Figure 2: Selected elements of the AKT Reference Ontology of Scientific Knowledge Management

The Reference Ontology and OntoClean

The main aim of this paper is to propose a reconstruction of the AKT reference ontology so that it makes use of the top level categories from the OntoClean ontology. The expected advantage of this approach is that the meta-properties which are used to determine the top level structure of OntoClean should be of use in making principled structuring decisions throughout the ontology. It was decided that the OntoClean top level structure should be merged with the existing reference ontology rather than replacing the existing top level, and this paper discusses decisions required and difficulties encountered in carrying out this merger.

Before discussing difficulties, however, it's important to consider if there are advantages in using OntoClean's property-based approach to ontological classification. This can be illustrated by considering how the key SKM concepts could be classified in OntoClean. For example, *Documents* are extensional, they have unity, are concrete, and are anti-dependent. They are therefore best mapped to the "physical bodies" subcategory of *Objects* in OntoClean. However, these property assignments are based on a number of assumptions which are made explicit below:

- It is assumed that any document that has the same proper parts to another document is in fact the same document. We see from this that *Document* differs from *Publication* – for two publications may have exactly the same content and yet be different publications.
- If the concept *Document* carries unity, then all instances of *Document* must be "essential wholes". Essential wholes are mereologically connected, as opposed to "singular wholes" which only have a topological connection. *Documents* are thus considered to be structured text, and are differentiated from collections of unrelated paragraphs.
- If *Document* is concrete then each document has a physical location. The location may be in a filing cabinet, or on a disk. This is used to distinguish *Recorded Audio Object*, which is a sibling of *Document* and a subclass of *Information Bearing Object*, from radio waves which are a subclass of *Information Bearing Thing* but not of *Information Bearing Object*[3].
- If *Documents* are anti-dependent, then they do not require the existence of any other object in order to exist. We therefore reject the view that *Document* is a conceptual entity that only exists in the form of one or more *Publications*; instead, we consider that no *Publication* can exist unless there is an instance of *Document* to be published.

So by applying the property-based classification approach of OntoClean to Documents, we have been forced to decide that Documents differ from Publications (a distinction that was not made clear in some of the four original ontologies of SKM); that Publications depend on the existence of Documents and not vice versa; that Documents are structured; and that while information need not have a physical location, Documents must have a physical location, whether they exist on paper or in some other form. It is this requirement to make precise definitions of concepts that is the main advantage of applying the OntoClean approach to ontology structuring.

The remainder of this section describes the proposed mapping of the other key SKM concepts to the OntoClean top level, and briefly discusses decisions that had to be made in order to assign ontology structuring properties:

[3] with acknowledgements to the Cyc upper ontology (Cycorp, 2001) from which the definitions of *Recorded Audio Object*, *Information Bearing Thing* and *Information Bearing Object* are taken.

- **Events** are explicitly represented at the top level of OntoClean, and are considered to be extensional and dependent. In order to define events as extensional, however, it's necessary to take a strong view of what constitutes a composite event: if even one sub-event changes, then it is considered to be a different composite event. This seems to fly in the face of intuition in terms of conferences, for a conference is an event, and yet if a conference cancels one talk, it is not usually considered to be a different conference. It is, however, a different event in the eyes of OntoClean.
- **Organisations** are anti-extensional (because they can change some of their parts without changing their identity), and have part-unity. We can safely assume that organisations are anti-dependent -- we can probably all think of committees that have continued to exist despite their *raison d'être* no longer existing -- and so *Organisations* can be considered to be ordinary objects in OntoClean. In fact, the subcategory of *SocialGroup* seems to be the most appropriate superclass for *Organisation*.
- **People** have unity, are concrete, and are anti-dependent; they should therefore be objects in OntoClean's hierarchy. OntoClean treats *Life Forms* as a special case of *Objects*; presumably this is because they are neither physical bodies (extensional) nor ordinary objects (anti-extensional)[1]. However, the four original ontologies have subclasses of *Person* that include *Academic*, *Student* and *Manager*. These are **roles**; and since they are anti-rigid, supply no identity criterion, are dependent (on the continued existence of the activity they perform, if nothing else), and also carry unity, they are most properly classified as *material roles*. Future discussion must therefore bear in mind this dual classification of People as both Objects (i.e. types) and roles.
- **Research areas** are anti-concrete -- they have no physical location, as opposed to technologies, which do -- but do not appear to be classified in a conceptual space (so they are not Qualities). They are therefore classified as *Abstract Entities*.
- **Tasks** are dependent on their performer and, if we accept a strong view that a task that changes any of its subtasks is a different task, they are extensional. It therefore seems reasonable to consider them to be isomorphic with *HumanActivity* in OntoClean, which is a subclass of Event.

[1] Perhaps this is because persons are normally considered extensional, but in the domain of transplant surgery they are clearly anti-extensional.

Merging the OntoClean top level into the AKT Reference Ontology

If two ontologies are to be merged at the top level, there are three possible approaches: a new top level can be created which contains all the concepts from the two original top levels, or one of the classifications can remain as the top level and the other can be introduced at a lower level. It was decided that it was undesirable to try to create a new top level that included all the top level concepts from both the current ontology and from OntoClean, since these concepts were created according to differing principles. Instead, the OU's existing tripartite top level classification was retained as the top level, and the OntoClean top level is introduced at the second level of the ontology.

Mapping OntoClean concepts to the OU top level

The mapping between OntoClean and the OU top level ontology appeared relatively straightforward for Events (classified as Temporal Things), Abstractions and Qualities (classified as Intangible Things) and Objects and Aggregates (classified as Tangible Things). However, the classification of Features raised some difficulty. OntoClean's definition of features is that they are "parasitic entities" that exist insofar as their host exists, and that they are essential wholes and singular entities. However, these features are subdivided into "relevant parts" of their host (a bump or an edge) and "dependent regions" (a hole in piece of cheese; the shadow of a tree). Both relevant parts and dependent regions are "parasitic entities", but relevant parts are tangible and non-temporal, while holes and (possibly) shadows appear to be intangible, and shadows may also be considered to be temporal.

The definition of spatial regions in ontologies has been an issue for a number of researchers. Ontologies of aircraft and their movement have to define regions of airspace as concepts (Grant, 2000); ontological definition of land areas have been a key feature in the DARPA-sponsored HPKB, SUO and RKF projects; and the correct definition of holes is a subject that has been debated in much detail (see for example (Varzi, 1996)). There is general agreement that spatial regions are entities, but the question of whether they are tangible entities or intangible entities is rarely addressed; it seems to be assumed that their tangibility depends on their contents, unless there are compelling reasons otherwise. So a land region is generally assumed to be tangible, while a region of airspace is usually considered to be intangible. Alternatives to this view

involve taking a strict definition of a region as an area or a volume, without reference to its contents – a region therefore only exists in a (spatial) conceptual space, and becomes an Abstraction in OntoClean's terms; or to take an even stricter definition of tangibility, by stating that any place that a human can place their hand – which includes any geographical location – is tangible. A similar strict-definition approach may be applied to the question of whether shadows are temporal: while shadows often move (and are thus spatially temporal), this is always due to the movement of either the light source or the occluding object, and if a strict view is taken that the "shadow region" of an object is illuminated by a defined light source which does not move relative to the object, then the shadow region is not temporal.

In short, the merging of OntoClean's Features into the OU's top level ontology can be accomplished if a strict view is taken of one or more definitions and/or the class of Features is broken into two. These requirements might be taken to imply that there is a weakness in one or both ontologies; but it is perhaps closer to the truth to say that the concept of spatial regions is one of a group of concepts that *any* ontological structure struggles to cope with. A full discussion of these "difficult" concepts is beyond the scope of this paper, but a working heuristic is that any concept for which individual instances can be defined at or between arbitrary point(s) in a continuous conceptual space will be difficult to classify ontologically. Examples of such concepts include colours (in an ontology of art objects built for Interpol, the ontology used to describe the colours of paintings was entirely different to the ontology used to describe the colours of ceramics (Wielinga, personal communication)), and financial value (which is a function of two conceptual spaces – supply and demand). Spatial regions are especially difficult, because they can be defined between two arbitrary points on each of two or three conceptual spaces, depending whether the spatial region is 2D or 3D.

Mapping SKM concepts to the merged ontology

Once the top level ontologies are merged, it is necessary to see if the definitions of SKM concepts in the two ontologies are consistent. For Documents, Events, Research Areas and Tasks, no problems arise. However, Organisations and People are considered to be Agents in the AKT reference ontology, while OntoClean defines them to be subcategories of Object. Agents are considered to be TemporalThings, but the merged ontology considers OntoClean's Objects to be tangible but not temporal things.

For People, the answer to this dilemma lies in the dual classification noted earlier; people as types (individual human Life Forms) are indeed non-temporal (unless measured on a scale of decades and centuries), but people's roles have no meaning unless those performing the roles act as an agent in some way. So People need to be classified both as Objects (specifically, Life Forms) and as Agents. For Organisations, it's a similar story: a University or a Funding Body may indeed be a Social Group, but it also acts as an agent by carrying out activities and constraining or encouraging the activities of other agents. Organisations therefore need a dual classification similar to that given to People.

To express this using OntoClean's meta-properties, the OntoClean top level classification is largely concerned with classifying *types*; any object that is also capable of performing a *role* can claim a secondary classification as an agent of some kind. The questions of whether the class of Agents needs a detailed substructure to encompass all its roles, and of whether the OntoClean top level should be modified to include Agent as a class, are left as questions for future work.

The proposed revision of the reference ontology, incorporating the OntoClean top level concepts, is shown in Figure 3.

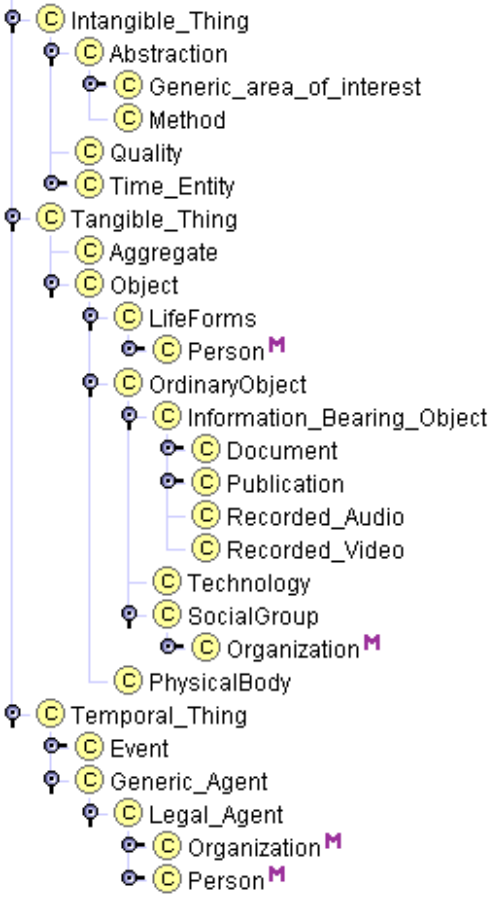

Figure 3: The proposed revision of the AKT reference ontology incorporating OntoClean top level concepts

Discussion

The application of the OntoClean ontology and its meta-properties to scientific knowledge management has required decisions and clarification of assumptions that have assisted in accurate ontological modelling of this domain. A proposal was made to merge the OntoClean top level with an existing "reference ontology" of SKM, in order to provide a well-justified basis for making further ontological decisions. The proposed merger has raised a number of issues regarding ontological classification. These include:

- When is it helpful to take a "strict" view of a parameter? When is a strict definition too strict?
- How should ontologies deal with concepts where individuals can be defined between arbitrary points on a continuum, such as colour, financial value, or spatial regions?
- How should ontologies handle relationships between concepts that are definitional but aren't truly taxonomic? This arises in the context of Documents and Publications in the AKT reference ontology: it was noted that any publication was dependent on the existence of an information bearing object, but the only way this could be expressed in the ontology was to create a class-subclass link between Publication and InformationBearingObject. Is there a better way of express such dependencies?
- Are types and roles sufficiently orthogonal that they require separate ontologies?

These questions are not only important for the domain of scientific knowledge management, but for all ontology researchers.

Acknowledgements

This work was supported under the Advanced Knowledge Technologies (AKT) Interdisciplinary Research Collaboration (IRC), which is sponsored by the UK Engineering and Physical Sciences Research Council under grant number GR/N15764/01. The AKT IRC comprises the Universities of Aberdeen, Edinburgh, Sheffield, Southampton and the Open University. The EPSRC and the Universities comprising the AKT IRC are authorised to reproduce and distribute reprints for their purposes notwithstanding any copyright annotation hereon. The views and conclusions contained herein are those of the author and should not be interpreted as necessarily representing official policies or endorsements, either express or implied, of the EPSRC or any other member of the AKT IRC.

References

AKT, 2001. *Reference Ontologies Version 1.* Advanced Knowledge Technologies Project http://kmi.open.ac.uk/projects/akt/ref-onto

Cycorp, 2001. *Cyc Upper Ontology*. http://www.cyc.com/cyc-2-1/index.html

Fellbaum C. (ed), 1998. *WordNet: An Electronic Lexical Database*. Boston, Mass: MIT Press. See also "WordNet: A lexical database for the English language", http://www.cogsci.princeton.edu/~wn/.

Gangemi A.; Guarino N.; and Otramari A., 2001. *Conceptual Analysis of Lexical Taxonomies: The Case of WordNet Top-Level*. In Proceedings of FOIS 01, Ogunquit, Maine, October 17-19 2001. Also available as a technical report of LADSEB-CNR, Padova, http://www.ladseb.pd.cnr.it/infor/ontology/Papers/OntologyPapers.html.

Grant T.J., *Identifying Planning Applications from Domain Analysis*. 19th workshop of the UK Planning and Scheduling SIG, Open University, 14-15 Dec 2000.

Kingston J.K.C., 2002. *Multi-Perspective Modelling: A Framework for Re-usable Knowledge*. Forthcoming.

Varzi A.C. *Reasoning about Space: The Hole Story*. Logic and Logical Philosophy, 4, 1996, pp. 3-39.