Ontologies and Information Systems: the Marriage of the Century?

Domenico M. PISANELLI, Aldo GANGEMI, Geri STEVE

CNR – ISTC, Viale Marx 15, 00137, Rome, Italy

Abstract. Although is recognized that ontologies may help building better and more interoperable information systems, there is skepticism on the real impact they may have in the future. We believe that ontologies will succeed in the information system arena and no systems will ever be designed without an ontological approach. In this paper we demonstrate the effectiveness of the ontological approach by illustrating three case studies. We show how an ontological framework is able to support semantic interoperability in the domain of fishery, then we present the role of ontologies for managing clinical guidelines and finally we sketch up an ontological analysis aimed at the interoperability of genetics databases.

1. Introduction

Many people today acknowledge that ontologies may help building better and more interoperable information systems. On the other hand, many others are skeptical about the real impact that ontologies - apart from the academic world - may have on the design and maintenance of working information systems.

Our claim is that ontologies will eventually succeed in the information system arena, the "marriage" will be happy and no computerized systems in this century will ever be designed without an ontological approach.

If "no man is an island", "no system is an island" anymore: data and knowledge integration (could we say "globalization"?) are no longer an optional, but a clear necessity. In fact, the overwhelming amount of information stored in various data repositories - including those available over the web - emphasizes the relevance of knowledge integration methodologies and techniques to facilitate data sharing. The need for such integration has been already perceived for several years, but telecommunications and networking are quickly and dramatically changing the scenario.

However, the ever-increasing demand of data sharing has to rely on a solid conceptual foundation in order to give a precise semantics to the terabytes available in different databases and eventually traveling over the networks. The actual demand is not for a unique conceptualization, but for an unambiguous communication of complex and detailed concepts (possibly expressed in different languages), leaving each user free to make explicit his/her conceptualization.

Ontology is the best candidate to face these problematics. Apart from its definition in the philosophical context - where it refers to the subject of existence - *ontology* in our context is "a partial specification of a conceptualization"[1], whereas Sowa proposed the following definition influenced by Leibniz [2]:

> “The subject of ontology is the study of the categories of things that exist or may exist in some domain. The product of such a study, called an ontology, is a catalog of the types of things that are assumed to exist in a domain of interest D from the perspective of a person who uses a language L for the purpose of talking about D. [...] ”

Actually there is some disagreement on what is an ontology. Some admit informal descriptions and hierarchies, only aimed at organizing some uses of natural language; others require that an ontology be a *theory*, i.e. a formal vocabulary with axioms defined on such vocabulary, possibly with the help of some axiom schema, as in description logics (for a position see Hayes [3]).

The various opinions in the research world enrich the scientific debate and are clear symptoms of cultural vivacity. Anyhow, the aim of this paper is not to review the ontology of ontology definitions, but to emphasize their strategic role in information systems design. Rather than raising up abstract claims, our objective is to demonstrate the effectiveness of an ontological approach by illustrating three case studies.

In the next paragraph we show how an ontological framework is able to support semantic interoperability among different information systems in the domain of fishery. The third paragraph presents the role of ontologies for effective and unambiguous dissemination of clinical guidelines. The fouth one sketches up an ontological analysis aimed at the interoperability of genetics databases.

2. The Fishery Ontology Service

Specialized distributed systems are state-of-the-art of current information systems architecture. Developing specialized information resources in response to specific user needs and area of specialization has its own advantage in fulfilling the information needs of target users. However, such systems usually use different knowledge organization tools such as vocabularies, taxonomies and classification systems to manage and organize information.

Although the practice of using knowledge organization tools to support document tagging (thesaurus-based indexing) and information retrieval (thesaurus-based search) improves the functions of a particular information system, it is leading to the problem of integrating information from different sources due to lack of semantic interoperability that exists among knowledge organization tools used in different information systems.

The different fishery information systems and portals that provide access to fishery information resources are one example of such scenario. In this paragraph we show the proposed solution to solve the problem of information integration in fishery information systems. The proposal shows how a fishery ontology that integrates the different thesauri and taxonomies in the fishery domain could help in integrating information from different sources be it for a simple one-access portal or a sophisticated web services application.

2.1 The local scenario

Fishery Ontology Service (FOS) is a key feature of the Enhanced Online Multilingual Fishery Thesaurus, a project aimed at information integration in the fishery domain. It undertakes the problem of accessing and/or integrating fishery information that is already partly accessible from dedicated portals and other web services.

The organisations involved in the project are: FAO Fisheries Department (FIGIS), ASFA Secretariat, FAO WAICENT (GIL), the oneFish service of SIFAR, and the Ontology and Conceptual Modelling Group at ISTC-CNR. The systems to be integrated are: the "reference tables" underlying the FIGIS portal [4], the ASFA online thesaurus [5], the fishery part of the AGROVOC online thesaurus [6], and the oneFish community directory [7].

The official task of the project is "to achieve better indexing and retrieval of information, and increased interaction and knowledge sharing within the fishery community". The focus is therefore on tasks (indexing, retrieval, and sharing of mainly documentary resources) that involve recognising an *internal structure* in the content of texts (documents, web sites, etc.). Within the semantic web community and the intelligent information integration research area (see for example [8] and [9]), it is becoming widely accepted that content capturing, integration, and management require the development of detailed, formal ontologies. In the following we present an outline of the FOS development and some hint of the functionalities that it carries out, thanks to the ontological approach.

2.2 Coping with heterogeneous systems: the ontological approach

An example of how formal ontologies can be relevant for fishery information services is shown by the information that someone could get if interested in "aquaculture". In fact, beyond simple keyword-based searching, searches based on tagged content or sophisticated natural-language techniques require some conceptual structuring of the linguistic content of texts.

The four systems concerned by this project provide this structure in very different ways and with different conceptual 'textures'. For example, the AGROVOC and ASFA thesauri put "aquaculture" in the context of different thesaurus hierarchies: according to AGROVOC the terms more specific than "aquaculture" are "fish culture" and "frog culture", whereas in ASFA they are "brackishwater aquaculture", "freshwater aquaculture", "marine aquaculture". Two different contexts relating respectively to species and environment point of view.

With such different interpretations of a term, we can reasonably expect different search and indexing results. Nevertheless, our approach to information integration and ontology building is not that of creating a homogeneous system in the sense of a reduced freedom of interpretation, but in the sense of navigating alternative interpretations, querying alternative systems, and conceiving alternative contexts of use.

To do this, we require a comprehensive set of ontologies that are designed in a way that admits the existence of many possible pathways among concepts under a common conceptual framework. This framework should reuse domain-independent components, be flexible enough, and be focused on the main reasoning schemas for the domain at hand.

Domain-independent, *upper* ontologies characterise all the general notions needed to talk about economics, biological species, fish production techniques; for example: *parts, agents, attribute, aggregates, activities, plans, devices, species, regions of space or time*, etc. On the other hand, the so-called *core* ontologies characterise the main conceptual habits (schemas) that fishery people actually use, namely that certain plans govern certain activities involving certain devices applied to the capturing or production of a certain fish species in certain areas of water regions, etc.

Upper and core ontologies [10,11] provide the framework to integrate in a meaningful and *intersubjective* way different views on the same domain, such as those represented by the queries that can be done to an information system.

2.3 Ontology integration and merging

Once made clear that different fishery information systems provide different views on the domain, we directly enter the paradigm of *ontology integration*, namely the integration of schemas that are arbitrary logical theories, and hence can have multiple models (as opposed to database schemas that have only one model) [12]. As a matter of fact, thesauri, topic trees and reference tables used in the systems to be integrated could be considered as *informal* schemas conceived to query semi-formal or informal databases such as texts and tagged documents.

In order to benefit from the ontology integration framework, we must transform informal schemas into *formal* ones. In other words, thesauri and other terminology management resources must be transformed into (formal) ontologies. To perform this task, we apply the techniques of three methodologies: OntoClean [11], ONIONS [13], and OnTopic [14] which are described in detail in the literature.

OntoClean consists of principles for building and using upper ontologies for core and domain ontology analysis, revision, and development. The OntoClean methodology is based on highly general ontological notions drawn from philosophical ontology, especially from what is now called “analytic metaphysics”, and it is general enough to be used in any ontology effort, independently of a particular domain.

As claimed by the authors: “The OntoClean methodology provides a formal, consistent and straightforward way to explain some of the most common misunderstandings in conceptual modeling regarding the taxonomic or subsumption relation.”[11]. In fact, those taxonomies feeding the first steps of an ontological analysis very often confuse subsumption with instantiation (e.g.: “John” is an istance of “Human” whereas the class “Humans” is subsumed by the class “Mammals”). Not to forget that another frequent feature of bad taxonomies is the systematic confusion between “part_of” and “is_a” (subsumption) relationships (e.g.: “engine” part_of “car”; “car” is_a “vehicle”).

ONIONS (“ONtological Integration Of Naïve Sources”) is a set of methods for enhancing the informal data of terminological resources to the status of formal ontological data types. OnTopic is about creating dependencies between topic hierarchies and ontologies. It contains methods for deriving the elements of an ontology that describe a given topic, and methods to build 'active' topics that are defined according to the dependency of any individual, concept, or relation in an ontology.

3. Ontologies and Clinical Guidelines

Guidelines for clinical practice are being introduced in an extensive way in more and more different fields of medicine [15,16]. They have the goal of indicating the most appropriate decisional and procedural behavior optimizing health outcomes, costs and clinical decisions.

Guidelines can be expressed in a textual way as recommendations or in a more formal and rigid way as protocols or flow diagrams. In different contexts they can be either a loose indication for a preferred set of choices or they can be considered a normative set of rules.

Clinical practice guidelines are seen as a tool for improving the quality and cost-efficiency of care in an increasingly complex health care delivery environment. It has been proved that adherence to plans may reduce cost of care up to 25% [17].

However the overwhelming number of guidelines available makes it difficult to select the right one. Just to give an idea of the figures, it is reported that there are 855 different guidelines for British GPs ranging from a single page to small booklets of more than 15 pages [18].

Computerization may increase the effectiveness of both the information retrieval of guidelines and the delivery of guideline-based care. In an optimal scenario they are integrated with the information systems operational at the point of care. The full potentialities of computerized systems can be exploited in such an environment where different processes are executed in parallel on several patients. In this context such systems must be able to retrieve the updated situation of every patient, as well as to give an overall report on the ward, freeing the physicians to concentrate more on clinical decisions. Keeping track of the parallel activities performed, they should avoid unnecessary duplication of tasks and prevent possible omissions.

3.1 The unambiguous representation of guidelines

Several research projects deal with the computer representation and implementation of guidelines. The scenario is evolving from stand-alone workstations to telematics applications that - utilizing e.g. the Internet - not only support the use of guidelines, but also enable their development and dissemination.

Such a knowledge sharing requires the definition of formal models for guidelines representation. The models should have a clear semantics in order to avoid ambiguities.

The role of ontologies is that of making explicit the conceptualizations behind a model. The definition of ontologies - i.e. the formal description of the entities to which a model makes a commitment and of the relations holding among the entities - is the groundwork for making a standard model acceptable and sharable. An ontology library is not normative, but allows an inter-subjective, explicit and formal agreement on the semantics of the primitives of a model, by referring to more generic primitives (generic theories).

We believe that such an approach can facilitate the standardization process by allowing an explicit mapping in a formal ontology of the concepts represented in the heterogeneous models proposed so far.

In our ontology guidelines are distinguished in "paper guidelines" and "web guidelines". Some common concepts - like "author" - pertain to both of them, whereas "URL" and "last-checked" are peculiar of the web guidelines. They are also categorized in five different kinds, as defined in the Guideline Interchange Format standard (GLIF) [19]: "guideline for care of clinical condition", "screening and prevention", "diagnosis and prediagnosis management of patients", "indications for use of surgical procedures", "appropriate use of specific technologies and tests". Such classification is furtherly specialized by us: for example a "guideline for care of clinical condition" may be a "therapy assessment", a "pharmacologic therapy" or a "disease management".

As far as the formal representation of guidelines is concerned, our ontology integrates some of the most relevant modeling efforts so far produced: notably PROforma [20], EON [21], Asbru [22] and GLIF [19]. It is also an evolution of a model previously defined in the context of the SMART system [23]. A "guideline" is a kind of "plan" which is a method of a "procedure", and it is represented by a "flowchart" (for more details see [24,25]).

The concept of "flowchart" pertains to the "diagrams" ontology. It is defined as a set of nodes and edges like an ordinary graph with some restrictions. Every flowchart has a first and a last node, only four kinds of nodes are allowed: single nodes, branching nodes, synch nodes and cycle nodes. Moreover the flowchart ontology allows for recursion, i.e. a node may be expanded into a flowchart.

3.2 The ontology of planning

The ontology of clinical guidelines accounts for the structural part of a guideline, but no semantics is attached to it. The semantics of the actions involved pertains to the *planning* ontology, where simple nodes represent elementary actions and branching nodes enquiries and decisions. The recursion allowed in the flowchart domain, where a node of a flowchart may be expanded into a flowchart, is isomorph to the planning ontology, where an elementary action may be refined into a plan.

We believe that in our model it is appropriate to capture the distinction between the structural part of a guideline, represented by the flowchart, and its semantics, represented by the plan. A third level is that of the procedure, i.e. what is actually performed.

This ontology integrates some of the most relevant work in the guideline modeling field. It is GLIF-compliant, i.e. each concept defined in GLIF is represented in it (e.g. the "synch" node after parallelization of activities). It takes into account the ProForma task ontology which categorizes tasks into: actions, enquiries and decisions and allows recursive definition of them (a plan is made of tasks, a task may be a plan).

It has been proven that the introduction of guidelines can significantly decrease the costs of care and therefore they are a "hot topic" in the agenda of health care professionals. Guidelines are mushrooming and computers can help in retrieving them and can give assistance during their execution.

However such a widespread diffusion poses new problems, not only in terms of credibility and acceptability, but also concerning non-ambiguity in knowledge dissemination. Formal models with a clear semantics should be defined in order to represent guidelines and facilitate their diffusion.

4. Ontology for the Interoperability of Genetics Databases

We are witnessing the unification of biology, since both biochemists and genetists now recognize a single universe of genes and proteins, and such unification is made possible also by the ever-increasing availability of the sequences of entire genomes.

Such an availability should rely on solid conceptual foundations in order to give a precise semantics to the data present in the different genome databases and accessible over the networks.

The conceptualization task - which is the groundwork for solid foundations - is not an easy one to be achieved, since a deep analysis of the structure and the concepts of terminologies is needed. Such analyses can be performed by adopting an *ontological* approach for representing terminology systems and for integrating them in a set of ontologies.

We performed an ontological analysis of the Metathesaurus™ [26], a terminology data-bank developed in the context of the Unified Medical Language System (UMLS) project by the U.S. National Library of Medicine [27,28]. It collects millions of terms belonging to the most important nomenclatures and terminologies defined in the United States and in other countries too. About 700,000 preferred terms, named "concepts", have been chosen as representative of a set of synonyms and lexical variants in different languages. It is probably the largest repository of terminological knowledge in medicine.

Recently we extended ontolological analysis in the genetics field.

As a case study, we investigated on the molecular function ontology defined by the Gene Ontology Consortium (http://www.geneontology.org) [29]. We implemented a wrapper translating from the XML ontology definition into LOOM, a formalism suitable for automatic classification [30].

The ontological analysis put in evidence the necessity of refining some assumptions made by the Gene Ontology developers. For example, metonymy is often used, since both enzymes and their functions are used in the same taxonomy.

5. Conclusions

Heterogeneity of information in data bases schemata or in other semi-formal information repositories is due essentially to the different conceptualizations, the different intended meanings of the terms which constitute the information in the repository. Such inherent polysemy of terminological information is reflected by widespread polysemous phenomena within existing terminologies.

Usually terminological sources shows the following features:

- *Lack of axioms*: for example, ICD10 shows nude taxonomies, without axioms or even a natural language gloss.
- *Semantic imprecision* (cycles, relation range violation, etc.): for example, the semantic network used as the top-level of the UMLS Metathesaurus includes a set of templates for its taxonomy, but the semantics of such templates is not defined at all: after careful analysis, the best that we could do is considering UMLS templates as default axioms.
- *Ontological opaqueness* (lack of motivation for choosing a certain predicate, or lack of reference to an explicit, axiomatized generic ontology, or at least to a generic informal theory): for example, systems in which concepts and relations in the top-level are non-axiomatized and undocumented: they may appear to have been chosen with disregard of formal ontology: possibly no trace of mereological, topological, localistic, dependence notions is retrievable.
- *Linguistic awkwardness in naming policy*: for example, systems in which purely formal architecture considerations originate a lot of redundancy and cryptic relation and concept names.

Ontology integration may act as a reference activity for information integration architectures and standardization work. Our experience has proved that the ontologies produced by means of the ONIONS methodology support:

- *Formal upgrading* of terminology systems: term classification and definitions are now available in a common, expressive formal language;
- *Conceptual explicitness* of terminology systems: (local) term definitions are now available, even though the source does not include them explicitly;
- *Conceptual upgrading* of terminology systems: term classification and definitions are translated so that they can be included in an ontology library which has a subset constituted of adequate generic ontologies;
- *Ontological comparability*, since pre-existing ontology libraries pertaining to different fields are largely employed.

In conclusion, we point out the following important features of ontologies:

- Semantic explicitness.
- An explicit taxonomy.
- Explicit linkage to concepts and relations from generic theories.
- Absence of polysemy within a given formal context.
- Modularity of contexts.
- Some minimal axiomatization to detail the difference among sibling concepts.
- A good naming policy.
- Rich documentation.

Our research aims at showing that integration of heterogeneous information systems can take advantage from the framework of formal ontology. We recognize that, in a software engineering perspective, many drawbacks are still present in ontology-based information systems.

Tools for efficiently supporting an ontology-based system design are not reliable enough for being employed in industrial practice. However we do believe that in a near future better engineered instruments will be available and the advantages of ontology-based design and integration of information systems will be tangible.

To this aim, the ontology design can profit also from the peculiar *Lyee* approach [31]. Even an ontology, in its essence, is a piece of software, and Lyee is a powerful and innovative software engineering methodology and tool for implementing software. It has been shown that such a tool allows an implementation faster than by using traditional methods.

Exploring the potentially fruitful synergies between the "classic" ontological analysis method and the revolutionary Lyee approach, will be a prominent issue in our future research.

References

[1] Guarino N (ed.), *Formal Ontology in Information Systems*, Amsterdam, IOS-Press, 1998.

[2] Sowa J, communication to the *ontology-std* mailing list, 1997.

[3] Hayes P, note on the meaning of “ontology”, http://ksl-web.stanford.edu/email-archives/srkb.messages/647.html.

[4] http://www.fao.org/fi

[5] http://www4.fao.org/asfa

[6] http://www.fao.org/agrovoc

[7] http://www.onefish.org

[8] http://www.ontoweb.org

[9] http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-6/web-agent/www/i3.html

[10] Gangemi A, Guarino N, Masolo C, Oltramari A.: Understanding Top-Level Ontological Distinctions, in: H. Stuckenschmidt (ed), *Proceedings of the IJCAI 2001 Workshop on Ontologies and Information Sharing.*

[11] Guarino N, Welty Ch. Evaluating Ontological Decisions with Ontoclean. Communications of the ACM, 2002, vol.45 (2): 61-65.

[12] Calvanese D, De Giacomo G, Lenzerini M.: A Framework for Ontology Integration. *Proceedings of 2001 Int. Semantic Web Working Symposium (SWWS 2001).*

[13] Gangemi A, Pisanelli DM, Steve G.: An Overview of the ONIONS Project: Applying Ontologies to the Integration of Medical Terminologies. *Data and Knowledge Engineering*, 1999, vol.31, pp. 183-220 (1999)

[14] Gangemi A, Pisanelli DM, Steve G.: The OnTopic Methodology for Supporting Active Catalogues with Formal Ontologies. ISTC-CNR-OCMG Internal Report iii-01 (2001)

[15] Woolf SH. Practice Guidelines, a New Reality in Medicine: Impact on Patient Care. *Arch Intern Med*, 1993; 153: 2646-2655.

[16] Grimshaw JM, Russell IT. Effect of Clinical Guidelines on Medical Practice: a Systematic Review of Rigorous Evaluations. *Lancet*, 1993; 342: 1317-1322.

[17] Clayton PD, Hripcsak G. Decision support in healthcare. *Int. J. of Bio-Medical Computing*, 1995; 39: 59-66.

[18] Hibble A., Kanka D., Penheon D., Pooles F: Guidelines in general practice: The new Tower of Babel? *British Medical Journal*, 1998; 317: 862-3.

[19] Ohno-Machado L et al. The Guideline Interchange Format, *Journal of American Medical Informatics Association*, 1998; 6.

[20] Fox J, Johns N, Rahmanzadeh A. Disseminating Medical Knowledge: The PROforma Approach *Artificial Intelligence in Medicine*, 1998; 14.

[21] Musen MA, Tu SW, Das AK, Shahar Y. EON: A component-based approach to automation of protocol-directed therapy. *JAMIA*, 1996; 3.

[22] Shahar Y, Miksch S, Johnson P. The Asgaard project, *Artificial Intelligence in Medicine*, 1998; 14.

[23] Pisanelli DM, Consorti F, Merialdo P. SMART: A System Supporting Medical Activities in Real Time, *Proc. Medical Informatics Europe*, 1997.

[24] Pisanelli DM, Gangemi A, Steve G, "Towards a Standard for Guideline Representation: an Ontological Approach", *Journal of American Medical Informatics Association*, vol.6 S4, pp. 906-910, 1999.

[25] Pisanelli DM, Gangemi A, Steve G, "The Role of Ontologies for an Effective and Unambiguous Dissemination of Clinical Guidelines", in R Dieng, O Corby (eds.), "Knowledge Engineering and Knowledge Management. Methods, Models, and Tools", Berlin, Springer-Verlag, pp. 129-139, 2000.

[26] Humphreys BL, Lindberg DA, "The Unified Medical Language System Project", *Proceedings of MEDINFO 92*, Amsterdam, Elsevier, 1992.

[27] Pisanelli DM, Gangemi A, Steve G, "An Ontological Analysis of the UMLS Metathesaurus", *JAMIA*, vol. 5 S4, pp. 810-814, 1998.

[28] Pisanelli DM, Gangemi A, Steve G, "A Medical Ontology Library that Integrates the UMLS Metathesaurus™", *Lecture Notes in Artificial Intelligence* 1620, Berlin, Springer Verlag, pp. 239-248, 1999.

[29] The Gene Ontology Consortium, "Gene Ontology: tool for the unification of biology", *Nature Genetics*, vol.25, pp.25-29, 2000.

[30] MacGregor RM, "A Description Classifier for the Predicate Calculus" *Proceedings of AAAI 94, Conference,* 1994.

[31] Negoro F, "Principle of Lyee software", *Proceedings of 2000 International Conference on Information Society in 21st Century (IS2000), pp.441-446,* 2000.