Ontology Mapping: A Way Out of the Medical Tower of Babel?

(Summary of Invited Talk)

Frank van Harmelen

Vrije Universiteit Amsterdam, Dept. of Artificial Intelligence,
De Boelelaan 1081a, 1081HV Amsterdam, Netherlands
frank.van.harmelen@cs.vu.nl

The Problem

Integration of different information sources has been a problem that has been challenging (or perhaps better: plaguing) Computer Science throughout the decades. As soon as we had two computers, we wanted to exchange information between them, and as soon as we had two databases, we wanted to link them together.

Fortunately, Computer Science has made much progress on different levels:

Physical interoperability between systems has been all but solved: with the advent of hardware standards such as Ethernet, and with protocols such as TCP/IP and HTTP, we can nowadays walk into somebody's house or office, and successfully plug our computer into the network, giving instant world-wide physical connectivity.

Physical connectivity is not sufficient. We must also agree on the *syntactic form* of the messages we will exchange. Again, much progress has been made in recent years, with open standards such HTML and XML.

Of course, even syntactic interoperability is not enough. We need not only agree on the form of the messages we exchange, but also no the meaning of these messages. This problem of *semantic interoperability* is still wide open, despite its importance in many application areas, and despite decades of work by different disciplines within Computer Science.

It is clear that the problem of semantic interoperability is also plaguing Medical Informatics. Terminological confusion is plaguing the interoperability of data sources, and is hindering automatic support for document searches. [10] provides a study of synonymy and homonymy problems on gene-names in Medline. They established that genes have on the average 2-3 different names; cross-thesaurus homonymy is often up to 30%; and almost half of the of acronyms used to denote human genes also have another meaning in Medline entirely unrelated to human genes[1]. The conclusion of a report [8] by the same research group states: *"Information extraction and literature mining efforts will be strongly affected by this ambiguity, and solving this problem is essential for these research fields."*

[1] My favorite example is PSA, (Prostate Specific Antigen), which also stands for Pot Smokers of America.

S. Miksch et al. (Eds.): AIME 2005, LNAI 3581, pp. 3–6, 2005.

This problems is by no means unique to genomics. Perhaps the oldest taxonomy around is the Linneaus "Systema Naturae" [5]. The modern-day on-line version of this system[2] [1] lists 411 homonyms, with the same name used for birds as well as insects, and insect as well as fish.

Multiple Solutions

The problem of semantic interoperability has been the subject of research in different fields over many decades. Different variants of the problem received names such as "record linkage" (dating back to Newcombe's work on linking patient records [6], and surveyed in [11]), schema integration [7], and more recently ontology mapping (see [2] and [4] for recent surveys).

An important development in this historical progression is the move towards ever richer structure: the original record linkage problem was defined on simple strings that were names of record-fields; the schema-integration problem already had the full relational model as input; while ontology mapping problems are defined on full hierarchical models plus rich axiomatisations. Each step in this progress has all the solutions of the previous steps to its disposal (since each later model subsumes the earlier ones), plus new methods that can exploit the richer structures of the objects to be aligned.

Current approaches to ontology mapping deploy a whole host of different methods, coming from very different areas. We distinguish linguistic, statistical, structural and logical methods.

Linguistic Methods are directly rooted in the original record linkage work all the way back to the early 60's. They try to exploit the linguistic labels attached to the concepts in source and target ontology in order to discover potential matches. This can be as simple as basic stemming techniques or calculating Hamming distances, or can use specialised domain knowledge. An example of this would be that the difference between *Diabetes Melitus type I* and *Diabetes Melitus type II* are not an innocent difference to be removed by a standard stemming algorithm.

Statistical Methods typically use *instance data* to determine correspondences between concepts: if there is a significant statistical correlation between the instances of a source-concept and a target-concept, there is reason to belief that these concepts are strongly related (by either a subsumption relation, or perhaps even an equivalence relation). These approaches of course rely on the availability of a sufficiently large corpus of instances that are classified in both the source and the target ontology.

Structural Methods exploit the graph-structure of the source and target ontologies, and try to determine similarities between these structures, often in coordination with some of the other methods: if a source- and target-concept have

[2] `http://www.taxonomicon.net/`

similar linguistic labels, then dissimilarity of their graph-neighbourhoods can be used to detect homonym problems where purely linguistic methods would falsely declare a potential mapping.

Logical Methods are perhaps most specific to mapping *ontologies* (instead of mapping record-fields or database-schemata). After all, in the time-honoured phrase of [3], ontologies are "*formal specifications* of a shared conceptualisation" (my emphasis), and it makes sense to exploit this formalisation of both source and target structures. A particularly interesting approach is to use a third ontology as background knowledge when mapping between a source and a target ontology: if relations can be established between source (resp. target) ontology and different parts of the background knowledge, then this induces a relation between source and target ontologies. A serious limitation to this approach is that many practical ontologies are rather at the semantically lightweight end of Uschold's spectrum [9], and thus don't carry much logical formalism with them.

Where Are We Now?

Undoubtedly, the problem of semantic integration is one of the key problems facing Computer Science today. Despite many years of work, this old problem is still open, and has actually acquired a new urgency now that other integration barriers (physical, syntactic) have been largely removed.

Given the difficulty of the problem, and the amount of work already spent on it, it seems unlikely that the problem of ontology mapping will yield to a single solution. Instead, this seems more the kind of problem where many different partial solutions are needed.

Currently, our toolbox of such partial solutions is already quite well stocked, and is still rapidly growing. However, a theory of which combination of partial solutions to apply in which circumstances is still entirely lacking.

References

1. S.J. Brands. Systema naturae 2000. website, 1989–2005. `http://www.taxonomicon.net/`.
2. J. Euzenat (coord). State of the art on ontology alignment. Technical Report D2.2.3, Knowledge Web, 2004. http://knowledgeweb.semanticweb.org.
3. T.R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199–200, 1993.
4. Y. Kalfoglou and M. Schorlemmer. Ontology mapping: the state of the art. *Knowledge Engineering Review*, 18(1):1–31, 2003.
5. C. Linneaus. *Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis.*. Holmiae (Laurentii Salvii), editio decima, reformata edition, 1758.
6. H.B. Newcombe, J.M. Kennedy, S.J Axford, and A.P. James. Automatic linkage of vital records. *Science*, 130:954–959, 1959.
7. Erhard Rahm and Philip A. Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal: Very Large Data Bases*, 10(4):334–350, 2001.

8. M. Schuemie and J. Kors. Assessment of homonym problem for gene symbols. Technical report, Erasmus University Medical Center Rotterdam, 2004. Report D4.4 for the ORIEL project.
9. M. Uschold and M. Gruninger. Ontologies: principles, methods, and applications. *Knowledge Engineering Review*, 11(2):93–155, 1996.
10. M. Weeber, R.J.A. Schijvenaars, E.M. van Mulligen, B. Mons, R. Jelier, C.C. van der Eijk, and J.A. Kors. Ambiguity of human gene symbols in locuslink and medline: Creating an inventory and a disambiguation test collection. In *Proceedings of the American Medical Informatics Association (AMIA 2003)*, pages 704–708, 2003.
11. W. Winkler. The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Bureau of the Census, Washington, DC, 1999.