Towards Industrial-Strength Philosophy: How Analytical Ontology Can Help Medical Informatics

Barry Smith (IFOMIS)[1]
Werner Ceusters (Language and Computing nv)

Preprint version of paper published in
***Interdisciplinary Science Reviews*, 28 (2003), 106–111.**

Introduction

Imagine that you want to build a computerized early warning system for bioterrorism attacks.[2] For this purpose you will need real-time access to medical data from which you can draw inferences regarding day-to-day variations in the doctor's visits and hospital admissions associated with specific types of ailments in specific geographical locations. You will need to have access to data regarding statistical variations in the drugs prescribed and purchased. Such data is in fact to a large extent already available, for example in the computerized inventory systems of pharmacies, drugstores and health insurance companies.

You will immediately encounter certain problems, however, when you begin to start working with this data. For while the technology for running computerized inventories has reached an impressive state of maturity, the classification systems upon which this technology is based are the products of myriad ad hoc decisions stretching back to the early days of database design. This means that, even with regard to the limited spectrum of remedies you can

[1] The work described in this essay is supported by the Alexander von Humboldt Foundation under the auspices of its Wolfgang Paul Programme.
[2] Such a proposal is described in the *Wall Street Journal* of November 13, 2001.

purchase over the counter in drugstores, such data exists in a variety of different forms, reflecting various combinations of store- and manufacturer-generated bar-code and drug product labeling systems. When the attempt is made to bring together all of these more or less ad hoc ways of partitioning the universe of pharmaceutical products, one very quickly runs into serious problems reminiscent of the old fable of the Tower of Babel.

These problems are made all the more difficult as a result of the fact that much valuable clinical data does not reside in structured databases at all. Rather, 98% of the information that is electronically available exists in the form of digitalized documents (for example patient records, articles in local newspapers) which preserve their original natural-language form. The integration of such information requires an additional prior step: of extracting the relevant data from the text in a way that makes it capable of being integrated with structured data and also amenable to the application of automatic reasoning systems of one or other sort.

The New World of Ontology in Information Systems

Initially the problems of database integration were resolved in case by case fashion. Pairs of databases were cross-calibrated by hand, rather as if one were translating from French into Hebrew. As the numbers and complexity of database systems increased, however, the idea arose of streamlining these efforts by constructing one single benchmark taxonomy, as it were a central switchboard, into which all of the various classification systems would need to be translated only once. By serving as a lingua franca for database integration this benchmark taxonomy would ensure that all databases calibrated in its terms would be automatically compatible with each other.

Interestingly, the proposed central classification system was called by information scientists an ontology, and it was quickly recognized that work on its construction would have more than a few echoes of the metaphysics of old. For when we ask which

classifications should be used in such a benchmark taxonomy, and on the basis of what principles it should be constructed, then we are raising *philosophical* questions – and indeed many of the difficulties faced by information scientists in building an ontology turn out to be identical to problems with which philosophers have grappled since Aristotle's day. They are problems relating to universals and particulars, properties and relations, events and processes. How, in a world of continuous differences, do category boundaries arise? How can we account for the identity of an individual over time when the individual is gaining and losing parts? Are classes the mere products of human division, or do they correspond to genuine invariants on the side of the things themselves? Are classes anything more than the totality of their instances? Can one system of classifying the entities in some given domain of reality be more *correct* than another?

The underlying premise of the new information systems ontology was that it would be possible to construct a classification system so general that all databases could be reformulated in its terms. The potential advantages of ontology thus conceived are obvious. If all databases, and all the data residing in unstructured text, corpora could be made compatible in the way described, then the prospect would arise of merging all of the separately existing digital resources in such a way as to create a single knowledge base of a scale hitherto unimagined, thus fulfilling the ancient dream of a Great Encyclopedia comprehending the entirety of human knowledge. In the specific domain of medicine it would mean that the huge mass of existing digitally available resources, whether in the form of unstructured patient records or in the form of statistical data prepared by hospitals on admissions and treatment of patients, could be used as the basis for a gigantic world experiment: new forms of purely digital medical research would become possible, based on exploiting the reasoning powers and the inexhaustible memories of computers.

The Problems

Unfortunately, however, as experience has shown, the construction of such a single benchmark ontology proved to be a much more complex task than was originally envisaged. This ontology would have to be simple enough that it can be programmed into our computers, yet it would have to be comprehensive enough that it can allow the expression of terms derived from all competing systems of classification. The information systems community itself has responded with a series of partial ontologies, each resting on a different pragmatically motivated choice about the way an ontology should be built. Ironically, therefore, the very Tower of Babel conditions which the ontological project was initially designed to address have been recreated within ontology itself.

Ontology at IFOMIS

The Institute for Formal Ontology and Medical Information Science (http://ifomis.de) represents a new approach to solving the problems of ontology. This Institute, which was founded in 2002 in the Faculty of Medicine in the University of Leipzig, seeks a return to the original project of a common benchmark ontology. Previous efforts at ontology building have conceived this project in pragmatic terms, a project deriving its motivation from the need to solve specific problems internal to the development of computer systems or from work in closely related disciplines such as artificial intelligence or knowledge engineering research (most recently from work on the standardization of documents and processes disseminated on the Internet within the framework of the so-called Semantic Web Initiative). IFOMIS, in contrast, looks beyond the realm of software artefacts. It starts out from the idea (1) that we should attempt to get the ontology right first, before building software models, and (2) that the project of developing the needed reference ontology can profit from the theories developed by philosophers over 2500 years of ontological research.

The IFOMIS ontology is marked by the factor of realism – and by the revolutionary idea that, in constructing a database system, we should first pay careful attention to *what the world is like* to which this system is to be applied. Where existing information systems ontologies have been constructed largely by taking as their starting point existing database systems or the conceptualizations used by the practitioners within given domains, IFOMIS seeks nothing less than a comprehensive theory of the divisions and interrelations between the entities on the side of reality itself. Such a theory must be built in step by step fashion, starting from those sorts of entities with which we are most familiar and moving outward to embrace ever new categories and domains of entities as these fall within our purview. It must be constructed in such a way as to allow for revisions as the attempt to deal with new sorts of entities reveals problems in the results achieved thus far. Not least importantly, it must be in a position to do justice to the fact that the very same reality may be sliced in different ways when addressed from different perspectives.

Our principal organizing structure in this respect is captured under the heading of *granularity*. Different partitions of the same reality can be effected at different levels of granularity (as we can partition the human organism into molecules cells, organs, or bodily systems). The medical ontology of IFOMIS must for example have the resources to sustain not merely an anatomical ontology at the level of the organs within the structure of the human body, but also cell, protein, gene and molecule ontologies at successively finer resolutions. It must sustain also classifications of *processes* at different resolutions, including the chemical and biological processes taking place inside the body.

Ontology at L&C

Most recently, IFOMIS has been collaborating with the company Language and Computing nv (L&C), headquartered in Zonnegem, Belgium and with a branch also in Philadelphia, USA. L&C

includes among its customers such companies as Eclipsys, First DataBank, Merck and WebMD.

Many technology companies active in the healthcare and pharmaceutical fields offer powerful tools designed to facilitate the processing of clinical data. Such tools are however overwhelmingly based on the requirement that the data in question be *structured*, for example by adherence to one or other controlled medical vocabulary. Yet much clinical data – thus for example most patient records – exist only in unstructured form, reflecting the desire on the part of many clinicians to use natural language formulations of their own choosing in order to capture the nuances of each particular case. Though most clinicians and other healthcare workers are gradually becoming convinced of the advantages of using computers, they still prefer to retrieve data stored by others, rather than to register data themselves. There are many reasons for this, including the unavailability of systems at the point of care and the woefully incomplete integration of computers in the primary care process, so that only a fraction of the activities in relation to which clinicians would like to be able to call upon computer resources are actually supported. The issue that deserves our particular attention here is the *information structuring bottleneck*. Healthcare records, whether on paper or in computers, are preserved as an external record forming a part of an individual patient history, so that future decisions can be based appropriately on past events. *Electronic* patient record systems have obvious advantages over paper-based systems in their ability to allow for cross-patient studies and to provide support for active decision management. Both we require structured data to exist inside the machine: the data has to be represented and stored in such a way that the machine itself can manipulate it, at least for those tasks for which it is better suited than humans.

The need for structured *data representation and storage* is undeniable and well understood. The mistake has however been made of inferring that this implies a need for structured *data entry*. As a consequence, structuring is imposed at the level of the data capture modalities such as rigorous data entry forms, point and click

interfaces, structured menus built out of check-boxes and the like. There is also structuring at the level of content by using coding and classification systems or controlled vocabularies. But is it really necessary to require the structuring be done by the user, given that most users do not like the constraints imposed by such structured data entry, and only some few are willing to accept these constraints for the sake of the benefits which they bring when retrieving information? That is, only some few are willing to accept the burden of structured data entry as the price to be paid for powerful information retrieval. But is this burden affordable, let alone justifiable, given the fact that, as already noted, many clinicians share the view that faithful recording of patient data can be achieved only by using natural language?

L&C has, in light of the above, developed an innovative approach to the processing of clinical data based on the use of sophisticated natural language processing technologies combined with what may be the world's most comprehensive terminology-ontology of the medical domain together with powerful formal systems for managing medical terminologies.

IFOMIS and L&C

IFOMIS and L&C work together on complementary objectives. L&C is faced with the constant need to update and refine its ontology resources. However most ontology developers with whom it might collaborate employ weak formalisms, such as description logics, which are fine for toy problems – or for the construction of ontologies of simplified worlds – but which do not match the requirements of the extremely large ontologies which are characteristic of the biomedical domain. What is missing in these approaches is the type of critical methodology characteristic of analytical philosophy as a tool for observing and understanding complex phenomena. IFOMIS takes ontology development seriously in precisely the sense required by L&C in its work on large ontologies, and it is not afraid to deviate from the mainstream

of ontology development where this is shown to be required by a careful analysis of the problems ontologists face. Equally, LinkBase® (L&C's medical ontology) and LinkFactory® (L&C's ontology management system) are invaluable to the work of IFOMIS, not least as bootstrapping devices to enable IFOMIS achieve its goals more quickly. More specifically, LinkBase allows IFOMIS to identify the important issues in healthcare ontology development and to find outstanding problems and difficulties. At the same time IFOMIS contributes its critical resources to L&C in ways designed to improve the foundations of the LinkBase ontology and to extend and amend it where necessary. LinkFactory allows IFOMIS to develop ontologies in a professional, industry-proven environment. The ultimate goal is to show that applying the methodology and ontological realism of the IFOMIS framework to the systems developed by L&C will lead to measurable improvements in efficiency and reliability.

The IFOMIS ontology will improve LinkBase's current formal definitions of the concepts and binary relations between concepts which are its building blocks by adding more standardization of a sort which employs a clean ontological theory. This will serve to refine and consolidate the structure of LinkBase itself, while at the same time giving it extra reasoning power and opening up the possibility of developing new sorts of algorithms in the future. These improvements will draw on the application of philosophical rigour already as a result of the fact that the language used will be that of first-order logic, the language in which the IFOMIS ontology is defined. The formal rigour of the IFOMIS framework is then imported into LinkBase on the meta-level: few changes are made to the elements themselves, but rather their place in an IFOMIS domain ontology is "tagged".

Why Healthcare?

The domain of medicine has been selected by IFOMIS for application purposes not only in light of its intrinsic significance but also because of the ontological challenges which it presents.

First, medicine calls for an ontology which can allow the simultaneous application of distinct perspectives (of, for example, doctor and patient, of pharmacologist and geneticist) to one and the same reality. Medicine is a domain which can sustain classifications reflecting causally relevant distinctions at more than one level of granularity.

Second, medicine is a domain which has the potential to show how a good ontology can yield demonstrable benefits in human welfare. If clinicians are truly of the view that patient data should be recorded as unstructured text, then the ideal situation would be one in which they can enter information in the preferred natural language form, but in such a way that we can have the machine analyse and structure this input automatically. This calls for advanced facilities for natural language understanding, and this in turn requires powerful ontologies.

The potential that is represented by the idea of a truly integrated system of electronic patient records requires not only adequate resources for dealing with clinical language; it requires also robust database classification systems and structured medical terminologies that enjoy a high degree of representational adequacy to the domain of medical phenomena themselves. The database and terminology systems associated with existing clinical data entry paradigms satisfy neither of these requirements – and here, too, the methodology of realist ontology propounded by IFOMIS can be of help.

The pharmaceutical industry is well aware of the importance of effective knowledge and information management. Bringing a new drug to the market is a multi-stage process that typically takes between 7 and 15 years. Huge amounts of information have to be

gathered, analysed and communicated along the way. Between 100 and 1000 people intervene somewhere in the development life cycle for any given drug, whether in feasibility studies, planning, clinical trial monitoring, medical writing, regulatory affairs, post-marketing surveillance or pharmacovigilance. Tens of thousands of documents are generated and have to be analysed. The IFOMIS medical ontology will thus need to comprehend, too, the various types of entities involved in these complex processes and most importantly in those complex processes we call clinical trials. A clinical trial is a controlled experiment in which the effectiveness of a given therapy is measured in systematic fashion in relation to preselected groups of patients. The IFOMIS medical domain ontology must thus be expressive enough to represent the structures involved in the standard types of trials. It should comprehend classification systems for therapies, patient populations and outcomes. It should help in the development of standards not only for the representation of trial data but also for the preparation of clinical protocols and of the guidelines which specify procedures for diagnosis and treatment.

The First Industrial-Strength Philosophy

LinkBase is a medical domain ontology developed by L&C and designed to support machine understanding of medical texts in a way that allows external standardized medical terminologies and ontologies to be used also for this purpose, in spite of the fact that the latter were not themselves designed for text understanding. This task turns out to be staggeringly complex, since the major terminology systems were constructed also without any basis in a prior rigorous ontology. They are often internally inconsistent and suffer from other logical defects, including classificatory gaps, cycles and terminological ambiguity. Most importantly, they manifest stark incompatibilities amongst themselves. LinkBase constitutes an all-embracing “container” ontology marked by strong logical coherence and conceived in such a way that external medical terminologies can be safely be mapped into it. But there are

problems: medical phenomena can be highly complex, and the language used to describe them can manifest subtle nuances of context, in ways which resist coherent treatment within a simple logical framework.

For millennia, when human beings have encountered problems in understanding reality, they have turned to philosophers for solutions. Now, when we encounter problems in understanding how to *represent* and *reason about* reality, we must do the same. The cause of the aforementioned ambiguities and inconsistencies was precisely the lack of a unified framework for understanding many of the basic formal relationships that structure reality (of object to process, of universal to particular, of part to whole, of function and execution, and so forth). The IFOMIS ontology provides a coherent, unified understanding of these relationships. Its implementation as a top-level or "backbone" ontology for LinkBase will thus not only provide a more robust framework for the clarification of existing ambiguities and discrepancies in and between ontologies, but also provide a template for future revision and augmentation of those ontologies. The implementation of a philosophically sound top-level ontology will thus serve the urgent needs of successful integration of existing terminology systems as well as serving as a useful guide for future algorithm development, not only in medicine but also, in principle, in a variety of other domains in which robust terminologies and efficient and reliable natural language understanding are pressing needs.

Literature on Applied Ontology

Guarino, Nicola 1995 "Formal Ontology, Conceptual Analysis and Knowledge Representation", *International Journal of Human-Computer Studies*, 43, 625-640.

Guarino, Nicola (ed.) 1998 *Formal Ontology in Information Systems*, Amsterdam, Berlin, Oxford: IOS Press. Tokyo, Washington, DC: IOS Press (Frontiers in Artificial Intelligence and Applications), 1998.

Johansson, Ingvar 1989 *Ontological Investigations. An Inquiry into the Categories of Nature, Man and Society*, New York and London: Routledge.

Koepsell, David R. 2000 *The Ontology of Cyberspace: Law, Philosophy, and the Future of Intellectual Property*, Chicago: Open Court.

Koepsell, David R. (ed.) 1999 *Proceedings of the Buffalo Symposium on Applied Ontology in the Social Sciences* (*The American Journal of Economics and Sociology*, 58: 2).

Schulze-Kremer, Steffen 1997 "Adding Semantics to Genome Databases: Towards an Ontology for Molecular Biology", in *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, T. Gaasterland, et al. (eds.), Halkidiki, Greece.

Searle, John R. 1995 *The Construction of Social Reality*, New York: Free Press, 1995.

Simons, Peter M. and Dement, Charles W. 1996 "Aspects of the Mereology of Artifacts", in: Roberto Poli and Peter Simons, ed., *Formal Ontology*. Dordrecht: Kluwer, 1996, 255-276.

Smith, Barry (in press) "Ontology", in Luciano Floridi (ed.), *Blackwell Guide to the Philosophy of Computing and Information*, Oxford: Blackwell.

Smith, Barry and David M. Mark 2001 "Geographic Categories: An Ontological Investigation", *International Journal of Geographic Information Science*, 15: 7, 591–612.

Smith, Barry and Varzi, Achille C. 2002 "Surrounding Space: The Ontology of Organism-Environment Relations", *Theory in Biosciences*, 121, 139–162

Smith, Barry and Zaibert, Leonardo 2001 " The Metaphysics of Real Estate", *Topoi*, 20: 2.

Welty, Christopher and Smith, Barry (eds.), *Formal Ontology and Information Systems*, New York: ACM Press.

Literature on medical ontology and natural language understanding

Ceusters, W., Buekens, F., De Moor, G., Waagmeester, A. 1998 "The distinction between linguistic and conceptual semantics in medical terminology and its implications for NLP-based knowledge acquisition", *Methods of Information in Medicine*, 37, 327-33.

Ceusters, W., Cimino, J., Rector, A. 1997 "Medical Language and Terminologies", in: M. Sosa-Iudicissa, N. Oliveri, C. A. Gamboa and J. Roberts (eds.), *Internet, Telematics and Health*, 197-203, Amsterdam: IOS Press.

Ceusters, W., Lovis, C., Rector, A., Baud, R. 1996 "Natural language processing tools for the computerised patient record: present and future", in P. Waegemann (ed.), *Toward an Electronic Health Record Europe '96, Proceedings*, 294-300.

Ceusters, W. and Smith, B. "Ontology and Medical Terminology: why Descriptions Logics are not enough",. TEPR 2003 (in press).

Ceusters, W., Spyns, P. and De Moor, G. 1998 “From Syntactic-Semantic Tagging to Knowledge Discovery in Medical Texts”, *International Journal of Medical Informatics* 52, 149-157.

Ceusters, W. 1999 “Language Engineering Tools for Healthcare Telematics”, in: *Proceedings of the Third European Conference on Electronic Healthcare Records, Eurorec’99, Seville (Spain), 6-7 May 1999*, 135-139.

Flett, A., Casella dos Santos, M., Ceusters, W. 2002 “Some Ontology Engineering Processes and their Supporting Technologies”, in: A. Gomez-Perez and V. R. Benjamins (eds.), *Ontologies and the Semantic Web*, EKAW2002, Berlin: Springer, 154-165.

Hahn, U., Schulz, S., Romacker, M. 1999 “Part-Whole Reasoning: A Case Study in Medical Ontology Engineering”, *IEEE Intelligent Systems and Their Applications*, 14: 5, 59-67.

Rector, A.L., Zanstra, P., Solomon, D., Rogers, J., Baud, R., Ceusters, W., Claassen, W., Kirby, J., Rodrigues, J. M., Rossi-Mori A., van der Haring, J., Wagner, J. 1998 “Reconciling Users’ Needs and Formal Requirements: Issues in Developing a Reusable Ontology for Medicine”, *IEEE Transactions on Information Technology in Biomedicine*, 4, 229 - 242.

Tange, H. J., Hasman, A., de Vries Robbe, P.F., Schouten, H. C. 1997 “Medical Narratives in Electronic Medical Records”, *International Journal of Medical Informatics*, 46: 1, 7-29.