Semantic representation of public health indicators: framework and a proposed practical application

A. Nagy, G. Surján, G. Héja

Introduction

Public health indicators are important sources of information in the everyday practice of healthcare administration and decision-making. The amount of data involved means a problem for those who need to access public health indicators regularly and efficiently. The scope of this work is to establish an ontological framework in which public health indicator data can be represented in order to make it semantically searchable. We also propose a data warehouse application that enables the user to access public health data. We used the indicator set defined by the ECHI (European Community Health Indicators) project as a data source of manageable size for our ontology. The idea of semantic representation of public health indicators is a largely unexplored one, while previous work has been done [1].

The problem

Utilizing public health indicators is affected by several problems. Some of these include poor quality of data, both in terms of availability and reliability. This is at least partly attributable to substantial differences in the structure of the health care system in different countries, differences in methodologies and indicator definitions.

Another serious source of problems, the one we address in this work, is the vast amount of data involved. In Hungary, ESKI (National Institute for Strategic Health Research) is responsible for maintaining a data warehouse consisting of more than 2000 individual health indicators. In Europe, the WHO “Health for all” database consists of about 600 indicators.

Browsing a data volume of this magnitude requires some navigational aid. Two currently utilized methods are free-text search and taxonomies. They both have their obvious drawbacks. Free-text search strongly depends on wording and may miss hits if they use a different phrase for the same entity; moreover, searching by relation is also impossible.

The major problem with taxonomies is that they are largely arbitrary and lack semantic background. A particular method of constructing taxonomies may provide useful results in one context, but another problem may require a completely orthogonal approach, one that is probably missing from the taxonomy in question.

An important realization that we had to make when describing health indicators is that we described *information objects* instead of real-world phenomena. However, at a later stage, these information objects are linked to those real world entities they convey information about.

Another point that we had to make clear was the concept of health indicator. What is an individual health indicator? It is a *numeric description* of a *collection of entities* that is relevant to health care. Health indicators have *dimensions*: they talk about a population that is formed by people of different age, gender, health condition, etc. Further dimensions are the location, and the time at which or during which indicator data was collected. Of these dimensions, however, only two are ubiquitous: the place and the time at (or during) which the indicator data was collected. All other dimensions optional: they apply to some indicators but not for others. For this, we regard a single indicator as a collection of numerical data from which one can select a particular numeric description by supplying nothing more but a geographical area and a time.

Before setting out to plan and implement an ontology, we have to consider the potential use for it. The system we have in mind is a semantic search facility that enables the user to search and access health indicators from many different sources, and also make it possible to compare indicator definitions to help find differences in methodologies. Potential users of this facility would be doctors, epidemiologists and health care decision makers.

We imagine that the most useful type of query to end users would be a special kind of query that finds relevant indicators along a collection of relationships that we consider “relevant”. This way, the end user doesn't have to pay attention to details of the ontology, but still can benefit from a semantically represented set of indicators.

Methods and sources

Ontology building is done using Protégé 3.3 in OWL 1.0. Using a reasoner is important both during the ontology building phase, to be able to pinpoint inconsistencies as soon as they are introduced, and also in the resulting application, as we intend to use the reasoner to answer queries. So far both Pellet and FaCT++ are being considered; the final choice will be a matter of performance evaulation.

We use Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [2] as a top-level ontology. DOLCE is philosophically well-founded and relatively easy to use.

The indicator set that we first chose to represent is a subset of the upcoming European Community Health Indicators project (ECHI) called “Short List Section 1” that consists of 40 smaller groups of indicators. It was chosen as the first set to be represented because it has a convenient size and it incorporates data from multiple sources, thus providing an opportunity to explore the benefits of semantic search in managing indicator data from multiple sources.

Results

Work done so far mostly went into the establishment of a framework that can represent health indicators both in a semantically correct and a reasonably efficient way. This task, while resulted in a fairly simple ontology, was nontrivial.

A substantial part of the work is the mechanism that we used implement meaning-independent search in the ontology. This provides a way to handle user requests without a well-defined meaning, for example a query like “indicators related to smoking”. To achieve this, we grouped some properties under a common superproperty, and marked this superproperty as transitive. The name of this superproperty is *is-related-to*. When the reasoner searches for indicators satisfying a restriction on *is-related-to*, the reasoner walks along properties of all indicators that are subproperties of *is-related-to*. As it is a transitive property, searching doesn't stop at the first level, but instead recursively progresses among the properties of the individuals found in the first turn and so on, until all “related” individuals have been examined or the given individual has been found.

At this time, ontology building is in progress, and the system is already able to answer queries on the *is-related-to* property. As new properties are introduced, the set of subproperties for *is-related-to* have to be carefully chosen and thoroughly tested for correctness so that all indicators get included in the result that we *intuitively consider relevant*, but the number of irrelevant indicators that get selected should be near zero.

Discussion

Difficulties other than those mentioned above rise from the different ways we may regard health indicators and the collections or populations that they express information on. We have considered three different approaches to this problem.

The first, most simplistic approach is to represent a population as a class by itself, and to link age and other restrictions to it by properties in the definition. This means, however, that the indicators have to be represented as classes instead of individuals, because in OWL-DL, the fillers of an individual's property cannot be classes, only instances. When such an ontology is classified using a reasoner, strange and uninterpretable conclusions arise. One such conclusion would be translated to plain text as “Every number of births to mothers aged 20-25 is a number of all births.” Such a conclusion is definitely false. This situation could have been resolved by either “reinterpreting” the semantics of *is-a* for the class of indicators. We decided to drop this approach, as it is almost sure that even if it would result in a usable application, it would later cause serious problems in interoperability with other ontologies.

The second approach is to regard a population or collection that a health indicator describes as a set with one or more criteria that control what elements are chosen into the set. We can also describe a set indirectly by describing these criteria. This approach ensures the highest level of semantic correctness, but appears to be cumbersome to use.

This is why we chose a third approach. Some indicators by their nature appear to describe a fictive “average member” or “generic member” of a collection. This “generic member” concept is essentially a placeholder for any entity that is a member of the collection in question. With this new concept, many indicators are conveniently expressed without resorting to talking about criteria, as simply as one would speak about a real person. The price to pay for this simplicity is the somewhat obscure meaning of a generic entity.

Conclusion

It has been proven that the concept of representing public health indicators is viable. The first objective for further efforts is to complete the description of ECHI Short List. After that, we will plan and implement a web service or offline program to perform semantic search on this on this ontology. A first step in this second phase would be to perform queries still within the ontology editor to assess exactness and overall performance of different reasoners. After that, a user interface would be designed to allow users to form queries. The final step would be the planning and implementation of the facility to translate queries from the user interface into OWL expressions.

References

[1] Surján G. et al.: A pilot ontological model of public health indicators (Comp. Biol. Med 2006. 36:802-16)

[2] Gangemi A. et al.: Sweetening ontologies with DOLCE p. 166 in Proceedings of Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web : 13th International Conference, EKAW 2002