Using ontologies to improve semantic interoperability in health data

Harshana Liyanage
Clinical Informatics, Department of Health Care Management and Policy, University of Surrey, Guildford, UK

Paul Krause
Department of Computing, University of Surrey, Guildford, UK

Simon de Lusignan
Clinical Informatics, Department of Health Care Management and Policy, University of Surrey, Guildford, UK

ABSTRACT

The present-day health data ecosystem comprises a wide array of complex heterogeneous data sources. A wide range of clinical, health care, social and other clinically relevant information are stored in these data sources. These data exist either as structured data or as free-text. These data are generally individual person-based records, but social care data are generally case based and less formal data sources may be shared by groups. The structured data may be organised in a proprietary way or be coded using one-of-many coding, classification or terminologies that have often evolved in isolation and designed to meet the needs of the context that they have been developed. This has resulted in a wide range of semantic interoperability issues that make the integration of data held on these different systems changing. We present semantic interoperability challenges and describe a classification of these. We propose a four-step process and a toolkit for those wishing to work more ontologically, progressing from the identification and specification of concepts to validating a final ontology. The four steps are: (1) the identification and specification of data sources; (2) the conceptualisation of semantic meaning; (3) defining to what extent routine data can be used as a measure of the process or outcome of care required in a particular study or audit and (4) the formalisation and validation of the final ontology. The toolkit is an extension of a previous schema created to formalise the development of ontologies related to chronic disease management. The extensions are focused on facilitating rapid building of ontologies for time-critical research studies.

Keywords: data integration, electronic health records, interoperability, ontology, semantic
**INTRODUCTION**

Health care systems have shifted from being tightly controlled local systems to large complex systems built upon heterogeneous data sources. Classically, health data are contained within computerised medical records; one held in primary care and another in secondary care. However, the range and scope of clinical, health, and social care, and other data have multiplied. The clinical and clinically relevant information stored in these data sources exist either as structured data or as free-text. The structured data may be organised in a proprietary way or be coded using one-of-many coding, classification, or terminologies that have often evolved in isolation and designed to meet the needs of the context that they have been developed. Generally, these records are single person based though some social care records are case based (e.g., child protection) and refer to multiple members of a family. Some more informal record systems may be shared and accessed by more people. Although they are complex and heterogeneous, their scope is much more limited than the emerging use of big data in health care.

In 2008, the International Medical Informatics Association (IMIA) presented a strategic plan, ‘Towards IMIA 2015’, to develop a world-wide system approach for health care that will incorporate and integrate research, clinical care, and public health. The plan envisioned clinicians, researchers, patients, and people in general to be supported by effective informatics tools, processes, and behaviours that allow them to take informed and timely decisions in order to improve health care for all. In 2015, achieving interoperability among health systems continues to be a major challenge in health care.

We have proposed that ontological approaches help understand the semantics of chronic disease data. Rapid changes in velocity and veracity mean that hard wiring linkage is probably not a sustainable approach. This paper proposes a systematic approach for using ontologies to maximise the potential of semantic interoperability when working with complex datasets.

**INCREASING VOLUMES AND COMPLEXITY IN HEALTH DATA**

The heterogeneity of data is generally observed at three levels:

1. **semantic**: different ways of interpreting the meaning of data;
2. **syntactic**: different ways of formatting the data;
3. **structural**: different ways of storing data.

Additionally, there can be heterogeneity of the information system or platforms hosting the data. Semantic heterogeneity is often the biggest challenge as it needs to be addressed in a domain-specific manner.

Most of our present–day information needs drawing data from multiple data sources, making ‘interoperability’ a significant function in information management. Structural and syntactical interoperability is easier to achieve compared to semantic interoperability as ‘meaning’ is often contextual even within the same information domain.
Standards for Semantic Interoperability

The European Commission recommendation on cross-border interoperability of electronic health record systems defines ‘Semantic interoperability’ as the process that means ensuring that the precise meaning of exchanged information is understandable by any other system or application not initially developed for this purpose.6

In the domain of health care information technology, Health Level 7 (HL7) has been adopted as the international standard for interoperability between health information systems. The initial versions of HL7 were based on a Reference Information Model (RIM) and approached information exchange based on point-to-point information exchange. The initial versions of the RIM were too complex to implement and not consistent.7 Clinical Data Interchange Standards Consortium has also made a substantial effort to standardise trial data.8

Semantic interoperability issues are frequently documented within the health care literature. A majority of these problems are based on issues encountered during projects facilitating interaction between real-life data sources. We have attempted to organise these issues by creating a classification of semantic interoperability issues (Figure 2).

![Figure 1 Increase in heterogeneous data sources in health care](image1)

**Figure 1 Increase in heterogeneous data sources in health care**

**Figure 2 Classification of semantic interoperability issues**

- **Contextual (Change in meaning based on context)**
- **Linguistic (Internationalisation issues)**
- **Completeness (Different levels of completeness based on context data was captured)**
- **Meaning of data changing with time (Changes in clinical definitions)**
- **Terminological changes due to advancement of science (New concepts introduced)**
- **Categorisation (Information is categorised differently by design)**
- **Positioning errors (Incorrect spatial positioning within vocabulary structures)**
More recent developments in semantic interoperability address some of their limitations and have become more ontological in their approach. Ontologies can be used to formally represent knowledge within a domain and this enables better interoperability by allowing data to be linked at semantic level. Semantic level interaction has become a key focus of more recent versions of HL7 including the ‘Service-Aware Interoperability Framework’ and the subsequent ‘Fast Healthcare Interoperability Resources (FHIR)’ standard. They focus more on implementability in order to have a more pragmatic approach towards interoperability. FHIR in particular has been developed to support solid ontology-based analysis.

**Ontologies for Semantic Interoperability**

Ontologies can be used to build machine-interpretable semantic representations of domain knowledge. Concepts in a domain, their attributes and their inter-relationships can be expressed using statements written according to a formal logic-based specification. This allows machines to make inferences on assertions (i.e. facts as logical statements) given in a domain model (class-level assertions) or a knowledge base built on a domain model (instance-level assertions). Since ontologies can explicitly define domain semantics, they can be used to map similar data held across heterogeneous data sources.

Ontologies accelerate implementing interoperability due to the availability of robust tools and technology frameworks that promote reuse. Visual ontology development tools such as Protégé abstracts the complexities of formal representation by giving a drag-and-drop interface users to rapidly develop ontologies. They have in-built reasoners and visualisation components to facilitate accurate translation of domain information into a machine-interpretable format.

**Ontologies May Help Solve Key Challenges in Semantic Interoperability**

While semantic interoperability has made a major contribution to data utilisation between systems, it often has not been able to integrate some large heterogeneous datasets required for research. Its greatest strength lies where terms have a similar unambiguous meaning in the distributor as well as the recipient of the data. In more loosely coupled systems, the semantic meaning often differs, and as a result, it is required to have an understanding about the structural and syntactic representation of the data. Ontologies, however, provide a flexible approach to integrating data and sharing meaning and may be better able to assist in inferring meaning in complex situations. Semantic interoperability will always have a place, but has not realised benefits in all circumstances, and as health data get more complex, it becomes more challenging to make systems interoperable. Nevertheless, ontological approach enables the best possible use of data.

From a health care perspective, ontologies can be used to maximise:

- meaning that can be inferred from coded data;
- different granularities of data (of words and coding);
- the ability to cope with temporal change in definitions, clinical practice and fluctuation;
- structural (system studies, e.g. encounters, health professionals, governance and privacy).

**Table 1** Common ontology standards

<table>
<thead>
<tr>
<th>Ontology Standard</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDF (Resource Description Framework)</td>
<td>A family of World Wide Web Consortium specifications originally designed as a metadata data model</td>
</tr>
<tr>
<td>OWL (OWL Web Ontology Language)</td>
<td>A family of knowledge representation languages for authoring ontologies</td>
</tr>
<tr>
<td>SPARQL (SPARQL Protocol and RDF Query Language)</td>
<td>An RDF query language that is a semantic query language for databases, able to retrieve and manipulate data stored in RDF format</td>
</tr>
</tbody>
</table>

**Figure 3** Comparison demonstrating as to how ontologies are better suited to address semantic interoperability issues
CLINICAL ARCHETYPES – A TIME-BOUND CONCEPTUAL REPRESENTATION OF A CLINICAL CONCEPT

Clinical archetypes (e.g. openEHR) have been suggested as an alternative method for organising clinical concepts.\textsuperscript{17} They are domain-specific computable structures developed according to a reference model that enforces constraints between the concepts defined. The archetypes allow clinicians to develop clinical concepts and their relationships and visual components that allow the clinician to interact with system in a system-independent manner. While this is a robust approach for ensuring stability for clinical concept definitions, it can still be challenging to link data modelled on these definitions to non-clinical data sources. They are limited by being a representation of conceptual understanding at the time of creation and are not readily used hierarchically.

Ontologies as a Key Enabler of Linked Data

Public sector organisations and governments are moving forward to integrate information through 'Linked data', a method of linking pieces of data and information using uniform resource identifiers.\textsuperscript{18} Open data initiatives are exposing large data sets through semantic endpoints that can be queried using semantic queries to get novel insights.\textsuperscript{19} We recommend semantic enablement as a mandatory requirement for any new health information technology project to leverage the deluge of big data across the health care ecosystem. The starting point would be to use a guidance framework such as the ‘ontology toolkit for developing ontologies related to chronic disease management’ to encourage adoption at ontological approaches among clinicians who possess the domain knowledge.

The Extended Ontology Toolkit for Chronic Disease Management

The toolkit for supporting the development of ontologies related to chronic disease management was developed as an outcome of a consensus process which took place in a forum at the Medical Informatics Europe (2012) conference.\textsuperscript{20} A key objective of developing the toolkit was to overcome problems associated with the semantics of datasets originating from heterogeneous data sources.

This toolkit suggested a four-step approach for developing ontologies:
1. Identification and specification of data sources;
2. Conceptualisation of semantic meaning;
3. How available routine data can be used as a measure of the process or outcome of care;
4. Formalisation and validation of the final ontology. It recommends tools that can be used for engineering ontologies and can be extended for building ontologies in other areas of health care.

![Figure 4 The extended ontology toolkit for chronic disease management](http://informatics.bmj.com/)

<table>
<thead>
<tr>
<th>Process</th>
<th>Tools/Methods</th>
<th>Ontologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Identification and specification of data sources</td>
<td>DOLCE</td>
<td>Generic health concept ontology</td>
</tr>
<tr>
<td>2. Conceptualisation of semantic meaning</td>
<td>OntoClean</td>
<td>Ontology mappings to terminologies</td>
</tr>
<tr>
<td></td>
<td>OWL/RDF</td>
<td>- Read2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- CTV3</td>
</tr>
<tr>
<td></td>
<td>SPARQL</td>
<td>- SNOMED-CT</td>
</tr>
<tr>
<td></td>
<td>SPARQL</td>
<td>- ICD 10</td>
</tr>
<tr>
<td></td>
<td>PROTEGE</td>
<td>- MeSH Terms</td>
</tr>
<tr>
<td>3. Formalisation</td>
<td></td>
<td>Data source ontology</td>
</tr>
<tr>
<td>4. Validation</td>
<td></td>
<td>Study requirement ontology</td>
</tr>
</tbody>
</table>
Since the initial development of the toolkit, we have utilised it in a number of studies dealing with routine health data. Most feedback received from clinicians involved was related to the steep learning curve for adopting standard tools that limit using them within the limited timeframe available for research studies. As a result, we extended the toolkit to facilitate rapid development of ontologies that focused on data-centric studies. The elements developed as the extension included a generic health concept ontology and a set of mappings of the concepts to chapters in frequently used controlled vocabularies. In addition, the toolkit included a ‘data source ontology’ to conceptualise metadata associated with the heterogeneous data sources incorporated into study. Finally, a study-requirement ontology was introduced to semantically represent the salient features of a research study.

DISCUSSION

Semantic interoperability is appropriate in situations where the data structure is known and where there is transferrable meaning. However, as we work with more and more data sources that are loosely defined, the standards established for semantic interoperability are increasingly difficult to use.

Ontologies allow better use of data in situations involving complex heterogeneous data sources. Furthermore, the existing stack of ontology tools facilitates a more enhanced user experience for achieving semantic interoperability. Clinical practitioners are often reluctant to adopt ontology development tools due to the steep learning curve associated with them. While the initial version of the ontology toolkit was focused on a structured method of developing ontologies for chronic disease management, the extension was more focused on having some pre-built ontological components that will reduce the learning curve and reduce the ontology development time in real-world data-centric studies.

COMPARISON WITH THE LITERATURE

Intermediate Processors of Health Information (IPHI) have been suggested as a mechanism for facilitating interoperability while enforcing privacy, ethical and data quality constraints. Such mediating entities could be used to provide rapid access to health data to interested stakeholders. Automated negotiation of data sharing among trusted stakeholders would be highly desirable for accelerating information governance approval processes and IPHI could be key to achieve such goals. In order to realise the maximum potential of IPHIs, we need to consider using ontologies to facilitate the mapping between health information producer and consumer.

There are several emerging models of interoperability that have been proposed within the health care sector in England. The National Information Board of the Department of Health is working towards realising the Care Act 2014 (a part of the government digital strategy), which includes integration across care services. The BCS has recently published interoperability guidance for health and care networks to enhance existing methods of data sharing across organisations. This guidance reflects the need for a broader use of ontologies.

CONCLUSIONS

There is a clear trend of adopting ontologies for enabling semantic interoperability by key stakeholders that facilitate health information exchange. The most important value offered by ontologies in this context is the ability to allow technology agnostic methods of communicating the meaning of similar concepts used within the domain. As we move towards achieving more comprehensive levels of semantic interoperability, ontologies have proved to be more dynamic than other methods used.

Nevertheless, there are some limitations of ontologies as they are scaled to represent large complex information domains. Building large ontologies can be time consuming and can require considerable amount of input from domain experts. These methodological issues need to be resolved in order to have main stream adoption that will demonstrate the collective effect of using ontologies for better semantic integration within the health care ecosystem.

Informaticians looking to work with large health datasets and looking to work with big data need to extend their capability to deliver semantically interoperable systems by arming themselves with an ontological toolkit that will boost adoption of ontologies and encourage participation of domain experts.


