Self-contained Patient Data in ORCA to Cope with an Evolving Vocabulary

Astrid M. van Ginneken M.D. Ph.D., Peter W. Moorman M.D. Ph.D.
Dept. of Medical Informatics, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam

Because of the benefits of standardization in healthcare data for research, decision support, and quality assessment, much research effort focuses on collection of structured patient data. Many strategies to obtain such data are based on controlled vocabularies to guide data entry in a far more flexible way than a fixed-form approach.

Medical controlled vocabularies evolve, but change is difficult to reconcile with standardization. Retrieval of data, collected with different versions of vocabularies, is not straightforward and has consequences for patient care and research. There are several strategies to cope with these problems: keep each version, keep a record of changes, or conversion of previously collected data. Each of these strategies has pros and cons regarding storage consumption, performance during patient care, and research. The approach in ORCA (Open Record for Care) is based on self-contained patient data and combines the strengths of these strategies.

INTRODUCTION

The ongoing effort in the development of controlled vocabularies, natural language processing (NLP) and structured data entry (SDE), reflect the widespread interest in more standardized collection of patient data [1-3]. The majority of clinicians still use the traditional paper record, despite the potential benefits of structured data for research, decision support, and quality assessment [4-5]. Successful introduction of SDE into practice requires a combination of education and better techniques to support data entry. All applications, supporting the collection of structured patient data face the need of adapting to the continuously evolving and expanding techniques and insights in health care. Change and standardization are difficult to reconcile. Several applications, such as Ivory, IMR-E, and ORCA (Open Record for Care) use knowledge-driven interfaces, based on controlled vocabularies to provide a flexible interface for SDE [6-10]. The problem of updating controlled vocabularies is well known by those who maintain coding schemes, such as the ICD, SNOMED, and UMLS [11-15]. Although structured patient data facilitates coding, supporting SDE is fundamentally different. Guiding data entry for the clinical narrative requires knowledge about how to describe complaints and findings, whereas coding involves classification of already available data. This paper focuses on changes to vocabularies used to support SDE. The question arises how these changes affect the use of patient data in both the routine care and research settings.

We will discuss the pros and cons of several strategies to adapt the knowledge bases underlying SDE applications. We will introduce the strategy adopted in ORCA (Open Record for Care) and discuss its strengths and weaknesses.

MODELS UNDERLYING SDE

In a previous publication, we discussed two basic models for the support of SDE: the direct model and the indirect model [16]. The term direct model denotes a direct mapping between a field on the data entry screen and an attribute in a database table. The classical example of a direct model is an interface with fixed-field forms that transparently reflect the structure of the underlying database. The term indirect model is used when an interpretation step is needed to present the patient data in a meaningful format. Most models based on instances of concepts are indirect, because the concepts (the knowledge) are needed to interpret the instances (the data). Indirect models have the advantage of being flexible, because the essence of indirection is separation between structure and content. Hence, a uniform structure can be used to store instances with a wide scope of meaning.

Direct models are rigid, but easy to query, whereas indirect models are flexible and more difficult to query. SDE for data items with fixed attributes, such as laboratory test results or drug prescriptions, can best be supported with a direct model. The indirect model has preference when data items are domain-dependent, i.e. vary highly among specialties [17].

The most typical example of domain-dependent data are the progress notes.

Adapting to change is most challenging for the highly variable domain-dependent data items. This paper addresses the problems and challenges of changing indirect models that are based on some form of controlled vocabulary for the support of SDE.

STRATEGIES FOR CHANGE

Controlled vocabularies evolve [12]. When they are used for data entry, this will have effect on the content and structure of the resulting data. Most literature about changes to controlled vocabularies pertains to classification schemes. Although coding differs from the collection of structured data at the point of care, there is overlap in the type of changes...
that may occur to the vocabulary involved. Concepts may be added or become obsolete - in which case their identifiers may be deleted or retired - , they may be merged or split, and their position in the network or hierarchy may change [11]. Questions arise about the effect of such changes on:

1. Access to patient data collected with previous versions
2. Consequences for research on data collected with different versions

We will review three possible strategies for coping with changes in a controlled vocabulary for SDE. In view of the two questions raised, we will identify their pros and cons and express these in terms of storage consumption, performance, and semantic consequences. Subsequently, we will describe the strategy in ORCA and compare that approach with the other three. In the following, we will simply speak of KB to denote the knowledge base, representing the controlled vocabulary for SDE.

Keep each version of the KB
One way to deal with changes is to keep each version of the KB and to record for all patient data with which version they have been collected. One practical disadvantage of this strategy is the consumption of large amounts of storage space. A more important drawback is the run-time performance of such a solution. The necessity to retrieve different versions of KBs to view patient data collected in the past – which is more rule than exception in patient care – will substantially slow down the application.

A third problem occurs when data needs to be accessed for research purposes. Each query will have to be broken down by the user into semantically equivalent sub queries, each of which is related to one KB version. Subsequently, the results of the sub queries need to be converted into a common format to combine them for data analysis.

There is also an advantage to this strategy, regarding redefinition of concepts. One example is the NYHA class for the daily activity level of patients with heart failure. The class has four codes, ranging from I to IV. It is possible that the granularity of description changes. Perhaps in the year 2000 the NYHA class will be expressed in grades I to VIII. When the KB version at the time of data collection is available, findings can be interpreted in their appropriate context: a finding ‘NYHA class II’ would have a different meaning in 1996 than in 2000.

Keep record of changes to the KB
The second strategy involves a detailed record or log of successive changes to the KB. Regarding storage consumption, this approach has the advantage that it consumes less storage space. The problems of run-time performance, however, become even more pronounced and are not acceptable in a time-pressured environment. Instead of retrieving the appropriate version of KB, part or all of the successive changes to the original KB have to be applied to recreate the applicable version of KB. For the entry of new patient data, the problem can largely be eliminated by off-line creation of the current KB version.

For proper insight in the consequences for research, it is important to realize that a track record may be restricted to KB mutations only or be extended with semantics. Mutations only specify which terms have been added, changed, removed, or relocated. A semantic track record will specify how the meaning of a section in the old KB is represented in the new KB. In case of a pure mutation track record, the necessity of segmented queries remains. With a semantic track record, a query on the latest version of the KB can automatically generate semantically equivalent sub queries on the older versions. This advantage is restricted to the semantics expressed in the current KB version. Data involving semantically obsolete information still need to be retrieved with dedicated queries.

The advantage of interpretation of data in their original context holds, because each version of KB can be recreated. In comparison to keeping all versions of the KB, this strategy is more storage efficient, may have retrieval advantages, but it is worse in run-time performance.

Conversion of Patient Data
The third strategy involves conversion of existing patient data to conform to the successor KB for every new KB release. As a result all patient data is in one format. This approach has several advantages over the previous two. It is more storage efficient, does not cause the extra runtime burden, and does not require segmentation of research queries. Yet, each new KB release requires a repeated substantial conversion effort. Apart from the effort of writing the conversion software, there is the effort of mapping the old representation of patient data to the new one. This mapping is a semantic task with all the problems of semantic mapping in general. Well-recognized problems are deletion, merging, and splitting of concepts in successive versions. Yet, how can a concept be represented that is no longer present in the current KB? Merges and splits may result in partial matches or result in ambiguity. Restricting the allowable changes to the vocabulary may preclude a state-of-the-art knowledge base.

Although redefinition of concepts is in principle taken care of via the semantic mapping, it raises a far more fundamental question: is conversion of existing patient data in general not in conflict with the
requirement that the record be faithful and permanent [18]? Since unambiguous semantic mapping can never be guaranteed for all versions of KB to come, the validity of this strategy is at least questionable.

THE CHALLENGE

It is obvious from the above mentioned three strategies which challenges remain:
- Preferably one version of KB operational (for storage and runtime efficiency)
- No conversion of patient data
- Solution for redefinition of concepts

Strategy in ORCA

In ORCA, data entry and retrieval are supported with only one KB version and patient data, collected with previous versions, can be retrieved and displayed without the need for data conversion. The essence of the strategy in ORCA is threefold:

1. In the knowledge base for support of SDE there is a thesaurus of concepts (the lexicon) and a concept network in the form of a directed graph.
2. No concepts may be deleted from the thesaurus.
3. The instances representing the actual data can be interpreted with the thesaurus of concepts alone.

We will briefly review those aspects of the ORCA knowledge model that are relevant for its strategy to tackle the challenges. In the discussion, we will address the consequences of this strategy for storage consumption, runtime performance for access to old patient data, research with patient data, and redefinition of concepts.

The knowledge model of ORCA. In ORCA, there is a thesaurus of concepts, in which each concept has a unique identifier, and a name. The thesaurus contains all concepts that have ever been defined. In other words, concepts are never deleted from the thesaurus. When a term becomes obsolete, but the underlying concept remains valid, the new name is represented in the thesaurus with an incremented version number and a time-stamp. In this way, changes to concept names over time are properly reflected.

The description knowledge is further represented by a separate directed graph of concepts. When concepts become obsolete, they can be deleted from the network. Such concepts will no longer be available for future data entry.

The ORCA concept network is based on conceptual graphs [19]. As opposed to the purpose of most conceptual graph applications, the description knowledge in ORCA is not designed for inferencing, but defines how the predefined medical concepts can be combined into meaningful descriptions. The concept graph is a special application of conceptual graphs [20]. The children of each concept represent its descriptors. The connecting relationships do only drive the behavior of the user interface for SDE. The ORCA knowledge has 7 different relationships.

Has-Specialization (HS) denotes class-subclass. Has-Feature (HF) represents a feature of the parent. Has-Value (HV) indicates the need for a numeric description. Refers-To (RT) is used to describe the same child from more than one context. HF and HV relationships are inherited by class-subclass children. During data entry, the user selects concepts while he traverses the concept network. Depending on the sequence of selections some concepts need to be excluded as descriptive options. For example, after selecting findings stomach, the location of an ulcer can no longer include other intestinal parts than the stomach. The remaining relationships enable context-sensitive presentation of descriptive options.

Findings are represented by instances of the selected concepts in the form of a tree. The order of the instances in the tree reflects the order in which the corresponding concepts are present in the current version of the concept network. The relationships are not stored in the patient data as these only contain information for user interface behavior. For example, [ulcer] – HF – [location] – HS – [cardia] in the KB is represented in the patient data by the sequence ulcer – location – cardia. Instance attributes will represent absence or presence and cardinality.

A path or subtree of instances, starting from the top node of the tree, reflects a (composite) finding in its context: the sequence endoscopy – stomach – ulcer represents a stomach ulcer, whereas the sequence endoscopy – duodenum – ulcer reflects a duodenum ulcer. The instances are linked to the related concepts in the thesaurus. Data entered with SDE is displayed as a tree or in a very simple text format.

![Knowledge Network, Thesaurus, Patient Data Diagram](image)

Figure 1. Example of a new network and old patient data. For simplicity, letters represent concept names and the numbers the versions of these names. *Italics* denote instances. Concept B1 is no longer present in the network and F2 represents a previous concept name for F3. Patient data can always be correctly interpreted with the thesaurus, irrespective of the changes to the knowledge network.
DISCUSSION

Storage consumption
Storage consumption by the ORCA KB is limited to the thesaurus and the current version of concept network. Although the thesaurus grows over time with new concepts and versions of terms, it is a relatively simple table that requires much less storage space than successive versions of the entire evolving and expanding concept network.

Performance and access to old patient data
In ORCA, patient data have a self-contained format: the necessary context for interpretation is conveyed in the thesaurus and the structure of the patient data trees. The version of the concept network at the time of data entry is not needed for interpretation. Some analogy can be made with laboratory test results that are stored with the reference values at the time of the test. Each value is stored with its context for interpretation. Without the need for conversion, this strategy adheres to the requirement of permanence and faithfulness of the CPR.

During data entry, only the current version of the concept network is needed. Older KB versions or reconstruction via the track record are not necessary.

Research with patient data
This is a more challenging topic as the question arises how patient data can be accessed, which data tree structure is no longer compatible with the current version of the KB.

Specialization of concepts can be achieved in ORCA by adding HS children to a concept. A search for the old concept will then also include the more specialized data. When specific concepts are merged, the query will only be complete when it includes both the specific concepts on the merged concept.

The semantics of the self-contained patient data are relatively insensitive to the relocation of concepts in the KB. The following example may illustrate this. Imagine that data representing a ‘stomach ulcer’ have been stored in the following ways, according to the order of concepts in successive versions of the KB:

Endoscopy  ulcer  location  stomach
Endoscopy  stomach  findings  ulcer
Endoscopy  findings  stomach  ulcer

When the researcher wishes to retrieve all patients with a stomach ulcer, the essence of the query is that ulcer and stomach are present in the same path. One cannot simply search for ulcer and stomach separately because the query would also retrieve patients with for example ‘a duodenum ulcer’ and ‘a stomach erosion’. The strategy for querying must be based on what we call membership of path. When ulcer and stomach are nodes in the same path, all three examples will produce a positive query result. Negations can be included in the query by setting criteria for the instance attributes. The membership of path strategy is a semantic query strategy that will include a variety of representations with the same meaning. Since all concepts are present in the thesaurus, query building is not dependent on the current version of KB. The first step in a membership of path query produces all matching path types in the data. In the second step, the user will have to select which paths should be included in the data query.

Redefinition of concepts
How does the ORCA model deal with redefinition of concepts? With redefinition of concepts we mean that the level of detail of description has changed, but the semantics of the concept remain the same. In case of the new definition of NYHA class with eight grades, the new concept name will be ‘NYHA class 2000’. Furthermore, grades V – VIII will be added as new descriptors.

What happens if a clinician wants to retrieve patients with NYHA class grade II? Since ‘NYHA class’ and ‘NYHA class 2000’ are names for the same concept all instances with grade II will be retrieved. However, the meaning of grade II recorded in 1997 is not the same as grade II recorded in 2001. A hybrid retrieval result is unlikely as ORCA will show the two existing data paths in which the instances of ‘NYHA class – grade II’ and ‘NYHA class 2000–grade II’ are present, see query 1 in Fig 2.

Figure 2. Matching paths for queries on NYHA class grades II and IV. Those without a cross will be included in the data query.

The different version names indicate that the two names reflect a different definition. As shown in Fig 2, the user will choose to use the path ‘NYHA class – grade II’. Another query on grade IV will produce two more paths and give him the option to add the path ‘NYHA class 2000 – grade IV’ to his data query. For analysis purposes, a conversion of the query results to one format will be needed, but this does not affect patient data in the record.
In routine patient care the performance of the ORCA strategy is independent of the extensiveness or number of changes to the KB, because only the current version and the thesaurus are necessary. In a research setting, relocations and redefinitions of concepts will produce more matching path types. Apart from path type selection, the resulting query will only be more time-consuming for the database. Obsolete concepts, however, require awareness of the user and active browsing of the thesaurus to include them in the query.

The work described in [11-13] focuses on the conceptual reliability of research queries, using established coding vocabularies over the boundaries of updates. The challenge here is to have semantic upgrades of the mapping between the coding vocabulary and the vocabulary on which the own data repository is based. The use of the ORCA vocabulary for SDE makes good runtime performance essential. The studies mentioned are in fact complementary to the ORCA strategy and become important when the intended semantic mapping between the ORCA vocabulary and GALEN is developed.

CONCLUSION

The strategy of self-contained patient data, as adopted in ORCA, combines the advantages of the other three strategies, except for the descriptors of obsolete concepts. At this moment, retrieval for research purposes is under development. More research is needed to investigate how robust the semantic retrieval of the data trees is. Are there structures that cannot be adequately retrieved via the membership of path strategy? The already discussed advantages of the ORCA strategy certainly justify further research in this direction.

References