SNOMED CT Normal Form and HL7 RIM binding to normalize clinical data from cancer trials

Antonio Rico-Diez, Santiago Aso, David Perez-Rey, Raul Alonso-Calvo,
Anca Bucur, Brecht Claerhout, Victor Maojo

Abstract— **Current research in oncology, require the involvement of several institutions participating in clinical trials. Heterogeneities of data formats and models require advanced methods to achieve semantic interoperability and provide sustainable solutions. In this field, the EU funded INTEGRATE project aims to develop the basic knowledge to allow data sharing of data from post-genomic clinical trials on breast cancer. In this paper, we describe the procedure implemented in this project and the required binding between relevant terminologies such as SNOMED CT and an HL7 v3 Reference Information Model (RIM)-based data model. After following the HL7 recommendations, we also describe the main issues of this process and the proposed solution, such as concept overlapping and coverage of the domain terminology. Despite the fact that the data from this domain presents a high level of heterogeneity, the methods and solutions introduced in this paper have been successfully applied within the INTEGRATE project context. Results suggest that the level of semantic interoperability required to manage patient data in modern clinical trials on breast cancer can be achieved with the proposed methodology.**

I. INTRODUCTION

Current clinical trials require the participation of multiple institutions, largely due to novel molecular variables introduced in the process. The management of such heterogeneous information has produced new challenges in the area, related to the management of information, during the last years [1][2][3]. Semantic technologies and related research are focused, among other issues, on allowing contextualized retrieval of information instead of just performing exact or similar match queries. Due to the lack of widely adopted standards and the complexity of patient and clinical trial data, such semantic capabilities are essential to ensure the sustainability of data management in current multi-centric clinical trials. Otherwise, final users would need to know every model and data format of institutions participating in the study.

The EU funded INTEGRATE project: Driving excellence in integrative cancer research [4] aims to develop innovative infrastructures to enable data and knowledge sharing and to encourage large scale collaboration in biomedical research. Such infrastructure should provide semantic interoperability capabilities to gather multi-scale biomedical data within post-genomic clinical trials [5]. To provide a semantic interoperability layer, the first step is to use a Common Data Model (CDM) to represent the data. It should allow the use of multiple established terminologies to build a core concept dataset as the shared vocabulary. In the INTEGRATE project, the CDM developed to represent the information has been based on the HL7 Reference Information Model [6], and the core concept dataset [7] is mostly composed by SNOMED CT [8].

Antonio Rico, Santiago Aso, David Perez, Raul Alonso, Victor Maojo are with Facultad de Informatica, Biomedical Informatics Group, University Politcnica de Madrid, Campus de Montegancedo S/N, 28660 Boadilla del Monte, Madrid, Spain

Anca Bucur is with PHILIPS Research Europe, High Tech Campus 34 Eindhoven, Netherlands

Brecht Claerhout is with Custodix NV, Kortrijksesteenweg 214b3, Sint-Martens-Latem, Belgium

II. BACKGROUND

The HL7 Version 3 Reference Information Model (RIM) [6] provides an abstract model for representing health related information. The aim of HL7 RIM is to build interoperability messages to communicate among information systems, including classes with sets of attributes associated by relationships. HL7 RIM provides the representation capability required to store complex, multi-scale and heterogeneous information, such as breast cancer patient data enrolled on clinical trials. In this case, we used only a subset of the RIM for the INTEGRATE project CDM.

Besides SNOMED CT, the core concept dataset also includes some domain specific terminology concepts such as LOINC [9] for laboratory tests and HGNC [10] for gene names. SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) is a logic-based health care terminology. It provides over 310,000 concepts and over 1,300,000 relationships between them. It is organized following a structure of is_a relationships (e.g. Allergic asthma is_a Asthma), allowing the composition of new concepts by the post-coordination mechanism [8].

Once the core concept dataset and the CDM have been selected, the corresponding mappings are required to link core concept dataset to the CDM, i.e. terminology binding [11]. The terminology binding defines which concepts, from different vocabularies, can be stored in a given field in the model, in order to build a common representation of the information. For example, Asthma is stored in Observation.code in the RIM model, where Observation is the class and code is the attribute. Main classes and its attributes where concepts could be stored are shown at Table I.

In this process, and for the SNOMED CT vocabulary in the INTEGRATE project, the common representation of such concepts is the SNOMED CT Normal Form [12] which simplifies the binding between these concepts and the RIM as specified by HL7 recommendations [13].

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

TABLE I
CLASSES AND ATTRIBUTES FROM RIM

Class	Attribute
Entity (Device, Material, Person, ...)	Code
Procedure	Code targetSiteCode methodCode approachSiteCode
Observation	Code targetSiteCode methodCode
Substance Administration	Code targetSiteCode methodCode approachSiteCode

III. METHODS

In the process of managing data from breast cancer clinical trials, data from different and heterogeneous sources have to be integrated. It is required to provide a homogeneous interface, where all the data can be accessed. In order to match data coming from different applications and databases it is essential to represent all data using a canonical or normal form. Using the SNOMED CT normal form standardizes the storage and retrieval of pre and post-coordinated expressions from clinical records in the model. The normalization of a given SNOMED CT concept also simplifies the implementation of the HL7 recommendations. Concepts are transformed into a normal form that can be used to bind data to the data model according to HL7 guidelines to achieve such binding, but it also presents some open issues. The original terminologies are annotated with the associated RIM classes and stored using an ontology representation language (Ontology Web Language - OWL). This information is afterwards used to expand and execute queries in the integrated repository.

A. SNOMED CT Normal Form

A normal form is a view that can be generated for any valid expression by applying a set of logical transformation rules [12]. A concept may be labeled either primitive or fully-defined. A concept is primitive when its roles and parents do not fully express its meaning. Primitive concepts do not have the unique relationships needed to distinguish them from their parent or sibling concepts. Fully defined concepts can be differentiated from their parent and sibling by the meaning of their relationships. One of the main characteristics of SNOMED CT data in normal form is that all the concepts from the resulting expression will be primitive. There are two distinct normal forms.

- The **long normal form** states all the attributes that can be inferred from concepts referenced by the expression.
- The **short normal form** enables more efficient retrieval testing due to fewer conditions, as the attributes do not include those from the closest primitive. However, there is no loss of specificity because any candidate that fulfills the conditions of the short normal form inevitably fulfills the conditions of the long normal form.

During the process described in this work, the use of the short normal form has been selected as it maintains all the information needed in a shorter form facilitating the binding of SNOMED CT with the HL7 model.
An expression in SNOMED CT short normal form can be basically divided in two parts:

- **Focus concepts:** The closest primitive super-types of the concept to normalize.
- **Refinement:** The defining relationships of the concept. These relationships can be grouped into zero, one or more role groups. These role groups were created to clarify expression by allowing to group relationships in a more readable form. However, the presence of those role groups is not required in the process of the terminology binding; only the information of the relationships is required.

The refinement is formed by a series of attribute-value pairs where attribute is the type of relationship and value must be another normalized expression (nested expression).

B. HL7 Terminology Binding to SNOMED CT

The current domain vocabularies of the INTEGRATE project, that are being used for semantic representation of concepts, are SNOMED CT and LOINC. Both terminologies are enriched with information from HL7 recommendations, gathered into a common OWL file, and finally stored into a semantic repository (i.e. SESAME [14]). A script to generate the OWL is already provided by the SNOMED CT owners but a few modifications were necessary in order to facilitate the retrieval of the storage information in HL7 and to include the LOINC terminology. The most important modification is the inclusion of the class and attribute from the RIM model where each concept should be stored. That is, the information provided by the terminology binding process. Another transformation introduced by this process is the inclusion of the LOINC terminology in the resulting OWL. The main relationships of each concept are stored with the concept itself, thus a new declaration of relationships is needed. Due to the nature of the LOINC vocabulary, all the concepts can be established to be Observation.code in the terminology binding process.
The process of calculating the class and attribute of each concept uses the default version of the SNOMED CT OWL to get the required information from the terminology. After that information has been obtained, the modified OWL can be created. The existing overlaps indicated by HL7 recommendations, must be solved as well as the semantic contradictions between an expressive SNOMED CT vocabulary and the HL7 RIM. The information that a SNOMED CT concept can express and how it is stored, according to the HL7 RIM model, necessarily requires a linkage method between objects that each part contains.
Once the normal form is obtained, and therefore the information implicit in the concept can be retrieved, a binding process is needed to store each piece of information keeping the semantic properties of SNOMED CT. There are cases where a concept is related to more than a single class

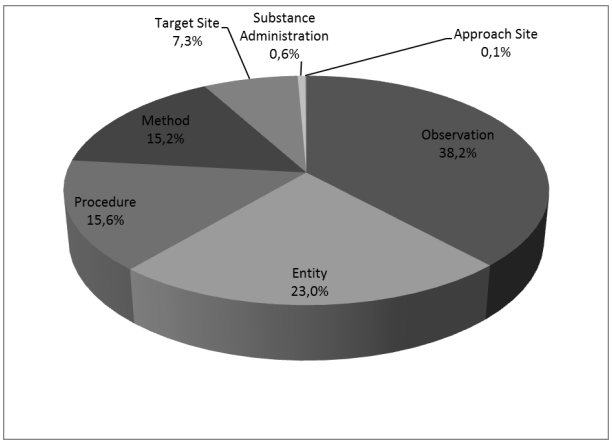

Fig. 1. Binding of SNOMED CT concepts with HL7 RIM classes

from the RIM model. This issue is called overlap and the results of the analysis carried out by this work show how this overlapping can be solved. In addition, SNOMED CT use many types of relationships, the most common are the is_a relationships that establish a hierarchy between the SNOMED CT concepts. But there are also other types such as Finding Site, Method, Associate Morphology, Part of, etc. Some of these relationships are considered by the HL7 recommendations, although there are others that are not included. When relationships not present at the HL7 recommendations appear in a normal form, a location in the RIM model is not automatically set and must be established through an analysis of the relationships and its occurrences in the SNOMED CT terminology.

IV. RESULTS

By using the previous methods in the process,we carried out an analysis of SNOMED CT and HL7 RIM binding to achieve semantic interoperability on breast cancer clinical trials. Some of the most important issues found in this analysis, were the overlapping during the mapping of SNOMED CT Concepts into the RIM model. The proportion of concepts for each RIM class can be observed on Fig. 1. Note that this graph is just indicative in the proportion of concepts for each class. It shows the percentage of concepts for each main RIM class, including overlapping concepts between the different classes and attributes found (for example the methods are the same for Procedures and Observations).

A. Overlapping of SNOMED CT binding to HL7 RIM

The study of the SNOMED CT Terminology and its binding with the RIM model showed that a set of concepts could be stored in multiple places in the model. These overlaps have been also mentioned in the Normal Form to HL7 section. An analysis revealed around 20K overlaps, the origin of most of these overlaps can be seen on Table II. On the other hand 247683 concepts don't have any overlap.

In fact, every overlap found is produced due to some of the parents of the concept. This parents may belong to an HL7 RIM class but a supperior parent (and it's offspring) belongs to another. This creates a double possibility in the offspring of the fist concept. For example, Evaluation Procedure and its entire offspring are Observation in the RIM model. However, Procedure, which is parent of Evaluation Procedure, belongs to the Procedure class from the RIM. The adopted solution was to select the more specific classes from the sub-tree so the overlapping is resolved as an Observation. Overlaps between Procedure and SubstanceAdministration are analogous, but in this case, the sub-tree is Administration of substance which is mapped to SubstanceAdministration despite this SNOMED CT concept and all its offspring are a sub-tree of Procedure.

TABLE II
OVERLAPPINGS OF SNOMED CT CONCEPTS MAPPED TO HL7 CLASSES

HL7 Classes Overlap	# SNOMED CT concepts
Procedure-Observation	17.463
Procedure-Subs.Admin	2.029
TOTAL	19.492

B. SNOMED CT Normal Form storage on HL7-based CDM

The SNOMED CT normal form may also produce additional overlappings when mapping to HL7. The focus concept of the normal form in some cases belongs to a different branch of SNOMED CT that correspond to another HL7 class. The focus concept is frequently considered to decide the HL7 class, however if the focus concepts belongs to a different HL7 class this may be misleading. To solve this issue, it was decided to use the original concept (not normalized) to set the HL7 RIM class, and in case it is not defined, then use the focus concept to set it. There are also concepts that are linked by recommendations to the HL7 RIM-based model, but are usually involved in a relationship on the normal form whose type establishes the attribute in the model. In that case, to obtain the class of the HL7 RIM, the focus concept should be used. For example:

```
Sodium Measurement
   Focus Concept = Procedure
   Relationships =
    Component = Sodium
    Method = Measurement-action
```

In this example, Measurement action does not belong to any RIM class but the relationship type (method) sets that its attribute is methodCode. To obtain the corresponding class, the original concept (Sodium Measurement), and the corresponding RIM class (Observation) have to be considered. With those pieces of information, Measurement action can be linked to Observation.methodCode in the given example. Issues related to relationships of SNOMED CT that are not included within the HL7 recommendations, and consequently do not have an HL7 RIM class associated, have been also analyzed. In order to store all the information that contains those relationships, the sixty different attributes present at

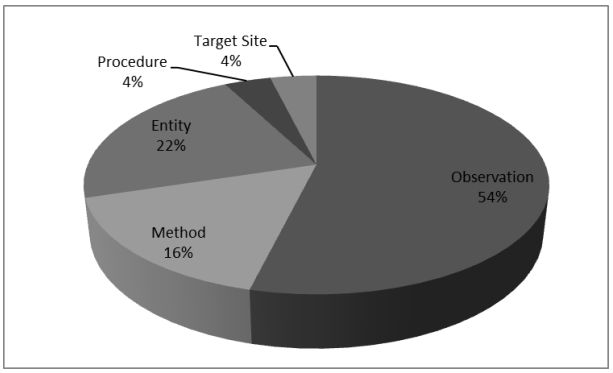

Fig. 2. Binding of the SNOMED CT subset of concepts used in clinal data from the INTEGRATE project with HL7 RIM classes

the refinements have to be also allocated within the CDM and stored in a database following that CDM.

C. Coverage of Breast Cancer Patient Data

With the proposed approach of HL7 RIM-based CDM and the core dataset containing SNOMED CT and LOINC terminologies, data available within the INTEGRATE project about breast cancer patients was coded and analyzed. The format of the data in its origin was HL7 v2 and v3 messages which were loaded into the CDM through an ETL process. The data includes around 2,500 acts performed to patients on clinical trials. The distribution of the data applying the proposed methods is shown on Fig. 2. By applying the normalization described above and the terminology binding to each piece of data from clinical trials at the INTEGRATE project framework, the proposed approach stores into the HL7 RIM-based CDM, the corresponding SNOMED CT and LOINC information as presented on Fig. 2. Every act concept is therefore expanded; the corresponding implicit information extracted and finally stored into the CDM. Such process allows the applications of the project, and research institutions to retrieve information not just by an act concept, but contextualized with all the information inherent to the specific concept.

V. CONCLUSIONS

In this paper we have described and analyzed the mechanisms proposed to supply a terminology binding between SNOMED CT and other terminologies as LOINC with the HL7 RIM model. The majority of the conflicts are related to overlapping introduced by HL7 RIM and SNOMED CT structures, and the lack of mappings for certain concepts and relationships. This work proposed a solution for each issue and it has been tested during the development of the INTEGRATE EU project in order to provide semantic interoperability for clinical trials in breast cancer context. There are still some challenges and limitations to address in order to achieve the main objective at the level of the terminology binding mainly related with the mapping of the relationships present in the normal form into the HL7-based CDM but the results of applying this methods until now suggests that this work may be done with an analysis of those relationships.

Methods can be extended to provide semantic interoperability not only to oncology-related clinical trials, but other areas that require similar solutions. The gathering of complex, highly heterogeneous and multi-scale data from post-genomic cancer patients is one of the main existing challenges to ensure the success of clinical trials in modern oncology. However, this challenge can be solved by the terminology binding of SNOMED CT and HL7 RIM explained in this paper.

VI. ACKNOWLEDGMENTS

This work has been partially funded by European Commission within the INTEGRATE project (FP7-2009-IST-270253).

REFERENCES

[1] Perez-Rey D, Maojo V, Garcia-Remesal M, Alonso-Calvo R, Billhardt H, Martin-Sanchez F, Sousa A. ONTOFUSION: Ontology-based integration of genomic and clinical databases. Computers in Biology and Medi-cine Volume 36, Issues 78, JulyAugust 2006; 712730

[2] Alonso-Calvo R, Maojo V, Billhardt H, Martin-Sanchez F, Garcia-Remesal M, Perez-Rey D. An agent- and ontology-based system for integrating public gene, protein, and dis-ease databases. Journal of Biomedical Informatics, 2007 Feb; 40(1):17-29.

[3] M. Tsiknakis, D. Kafetzopoulos, G. Potamias, C. Marias, A. Analyti, and A. Manganas. Developing European bio-medical grid on cancer: The ACGT integrated project. Proc. HealthGrid 2006 Conf., Valencia, Spain, Stud Health Technol. Inf., vol. 120, pp. 24758, Jun. 68.

[4] Fp7-integrate.eu [homepage on the internet]. Driving excellence in integrative cancer research [updated 23 November 2012; cited 25 November 2012]. Available from: http://www.fp7-integrate.eu/index.php/project

[5] Paraiso S, Perez D, Alonso R, Claerhout B, Schepper K, Hennebert P, Lhaut J, Leeuwen J, Bucur A. Semantic in-teroperability solution for multicentric breast cancer trials at the INTEGRATE EU Project. In Proceedings of the 6th International conference in Health Informatics 2013.

[6] Beeler G, Case J, Curry J, Hueber A, Mckenzie L, Schadow G, Shakir AM. HL7 Reference Information Model. 2003.

[7] Bucur A, Leeuwen J, Claerhout B, Schepper K, Perez D, Alonso R. Evaluation of the Semantics of Eligibility Crite-ria of Clinical Trials in Oncology based on relevant Medi-cal Ontologies. Proceedings of the 2012 IEEE 12th Inter-national Conference on Bioinformatics & Bioengineering (BIBE). 413-421.

[8] Donnelly K, SNOMED-CT: The Advanced Terminology and Coding System for eHealth. Medical and care compu-netics 3 2006: 279-290.

[9] McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, Forrey A, Mercer K, DeMoor G, Hook J, Williams W, Case J, Maloney P. LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update. Clinical Chemistry. 2003; 49 No. 4: 624-633.

[10] Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA. Genenames.org: The HGNC resources in 2011. Nucleic Acids Research. 2011; 39: 514-519.

[11] Tim Benson. Principles of Health Interoperability HL7 and SNOMED. London: Springer-Verlag; 2010.

[12] College of American Pathologists. SNOMED Clinical Terms Guide. Transforming Expressions to Normal Forms. August 2006. www.cap.org/apps/docs/snomed/documents/transfor-mations_to_normal_forms.pdf

[13] Cheetham E, H. Dolin R, Markwell D, Curry J, Gabriel D, Hausam R, Knight B, Rector A, Spackman K, Townend I. Using SNOMED CT in HL7 v3 Implementation Guide, Release 1.5. 2008.

[14] Broekstra J, Kampman A, Harmelen F. Sesame: A Gener-ic Architecture for Storing and Querying RDF and RDF Schema. The Semantic Web. 2002; 2342: 54-68.