An OGMS-based Model for Clinical Information (MCI)

Heiner Oberkampf [1,2,*], Sonja Zillner [1], Bernhard Bauer [2] and Matthias Hammon [3]

[1]Siemens AG, Corporate Technology, Munich, Germany
[2]Software Methodologies for Distributed Systems, University of Augsburg, Germany
[3]Department of Radiology, University Hospital Erlangen, Germany

ABSTRACT

More and more detailed, complex and new data about the patient's health status as well as about medical knowledge become available. The synopsis of this heterogeneous, patient-customized information is crucial for physicians to make the correct diagnosis. The problem however is that these heterogeneous data are not semantically integrated. As a result most of the available data and knowledge is often not used in their full strength in clinical decisions. Semantic integration requires annotation of clinical data with concepts or codes from established domain ontologies covering medical and clinical knowledge such as the Foundational Model of Anatomy (FMA), SNOMED CT or the International Classification of Diseases (ICD). Further, an ontologically well founded information model structuring the references to these ontologies is needed. Today's models of clinical information like the HL7 Reference Information Model however lack a well defined ontological foundation. The resulting ambiguities make it difficult to map clinical data to their schema and to reuse clinical data stored in them. In this paper we present initial work on a Model for Clinical Information (MCI) based on the Ontology of General Medical Science (OGMS) and other OBO ontologies. MCI focuses on meta-information and high-level concepts with the aim to provide a basis for data integration and knowledge exploration.

1 INTRODUCTION

More detailed, complex and new data about the patient's health status as well as about medical knowledge become available. Only this increase of available data makes individual treatments possible. The problem however is that these data are not semantically integrated. As a result most of the available data are simply not used in their full strength in clinical decisions. What we need is an integrated and standardized representation of clinical patient data reflecting the health status since this is the basis for various clinical applications like outcome analysis or other decision support systems. A standardized representation requires the use of established ontologies, vocabularies or coding systems like, e.g., the ICD[1], Logical Observation Identifier Names and Codes (LOINC)[2] or SNOMED CT[3]. In addition an information model is needed where the coded data (data with references to standardized vocabularies) is stored and structured. We identified the following requirements for a model attempting to represent clinical information using existing ontologies:

- **Integration:** data from various sources and of different format are integrated and linked
- **Standards:** data are expressed using established coding systems and terminologies
- **Interpretation:** the semantics of clinical data is consistently defined
- **Coverage:** it is possible to represent all clinical data using the model in combination with other ontologies and all clinically relevant high-level concepts are defined

Starting from these requirements we propose a Model for Clinical Information (MCI) based on the Ontology of General Medical Science (OGMS)[4]. MCI has the purpose to integrate and structure clinical data in providing concepts covering meta-information and interpretations of clinical patient data. So all basic concepts, which are needed to describe clinical information objects on the meta-level like diagnosis, findings, reports, health care provider, procedures, IDs etc. are contained in MCI. Patient data are then represented using MCI in combination with large domain ontologies and coding systems. Since high-level concepts are contained in MCI, it provides a good basis for data integration and knowledge exploration like, e.g., outcome analysis or decision support systems. MCI *structures* clinical data: for instance, MCI contains a concept like 'obo:diagnosis'[5] and sub-concepts 'mci:main diagnosis' and 'mci:secondary diagnosis', however no class for any particular disease or any ICD code. These codes are referenced. Analogously MCI contains the class 'mci:examination modality' and a subclass 'mci:patient position' but no classes for specific positions like 'standing position' or 'sitting position' since these concepts are defined in existing ontologies. For instance, the Radiological Lexicon (RadLex)[6] contains 47 subclasses of 'radlex:patient position'.

One could argue that all concepts defined by us are already available in existing ontologies like, e.g., SNOMED CT. Why another model? For us, the separation of meta-information and high-level concepts from domain knowledge seems to be a good approach (see also related work in section 4). This gives us stable core concepts while we are free to reference different ontologies depending on the data and use case.

In this paper we present initial work on MCI. Focused on structured data, such as information about diagnosis, laboratory values or different procedures, we address the above mentioned requirements. The integration of unstructured data will be addressed in future work.

*To whom correspondence should be addressed: heiner.oberkampf.ext@siemens.com

[1] http://www.who.int/classifications/icd/en/

[2] http://loinc.org/

[3] http://www.ihtsdo.org/snomed-ct/

[4] http://code.google.com/p/ogms/

[5] For better readability we write any concepts by prefixed annotation properties, i.e. we write 'obo:diagnosis' instead of 'obo:OGMS_0000073'.

[6] http://www.radlex.org/

2 THE MODEL

MCI provides the basis for data integration and knowledge exploration, i.e. the structural concepts for the representation of clinical data. This is at first place meta-information about the patient characteristics like diagnoses and findings (e.g. target lesions that were determined for a cancer patient) as well as provided examinations, procedures and therapies. Using MCI it should be possible to infer the changes of diagnoses and findings over time. These changes might then be analysed in the context of provided examinations and procedures e.g. in order to measure their effectiveness. This becomes possible since MCI integrates various types of information. The development is based on a dataset provided by our clinical partners covering six melanoma patients. Melanoma patients were selected since they create data in different clinical domains (diagnosis, lab values, pathology and radiology reports ...). For these patients we have information about demographics, diagnoses (ICD-10 codes), lab values (LOINC codes), procedures (OPS codes[7]), drug administration (ATC codes - Anatomical Therapeutic Chemical Classification System) as well as reports from radiology and pathology departments (partially structured free text). In the following we describe MCI in detail: the imported ontologies, the added concepts (classes and properties) and the link to terminologies.

2.1 Imports

As described above MCI is based on established upper- and mid-level ontologies, defining the basic ontological and clinical concepts (see figure 1). The main import is the Ontology for General Medical

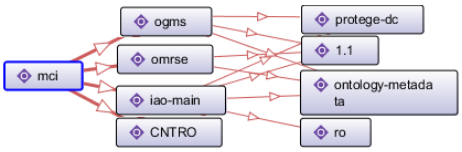

Fig. 1. MCI imports: ogms (Ontology of General Medical Science), omrse (Ontology of Medically Related Social Entities), iao-main (Information Artifact Ontology), CNTRO (Clinical Narrative Temporal Relation Ontology), 1.1 (Basic Formal Ontology), ro (OBO Relations Ontology).

Science (OGMS) which itself is based on the Basic Formal Ontology (BFO 1.1)[8]. While the BFO is completely domain independent, OGMS defines basic clinical concepts like diagnosis, clinical finding, pathological anatomical structure, sign, symptom and others (Scheuermann *et al.*, 2009). Further, MCI imports the Ontology of Medically Related Social Entities (OMRSE)[9]. OMSRE is meant to be related with OGMS and defines concepts for various clinical roles like, e.g., 'obo:patient role' or 'obo:health care provider role'. OGMS contains some concepts from the Information Artifact Ontology (IAO)[10] like 'obo:information content entity' and 'obo:data item' subclassed by OGMS with e.g. 'obo:clinical finding' and 'obo:diagnosis'. Interestingly, other concepts from IAO like 'obo:document', 'obo:report', 'obo:figure' with subclasses as well as object properties are not imported by OGMS. We decided to import almost the entire IAO ontology. The OBO Relations Ontology (RO) is indirectly imported through the import of IAO. We also imported the object properties under 'cntro:temporalRelation' from the Clinical Narrative Temporal Relation Ontology (CNTRO)[11] in order to structure clinical events.

2.2 Added Concepts

Even though the most basic terms are defined in OGMS, the representation of clinical data needs more detailed concepts. On the one hand we need to add classes and on the other hand properties for relating instance data. So far we have added about 39 classes, 11 object properties and 10 data properties. Together with the imported concepts MCI contains 310 classes, 73 object properties and 15 data properties. Additions will be made when data from other clinical departments are available.

2.2.1 Added Classes OGMS contains the concept 'diagnosis' but we additionally need sub-concepts 'main diagnosis' and 'secondary diagnosis' since this distinction is often made in clinical practice. Further we added subclasses of 'clinical finding' as shown in figure 2. Important for the evaluation of finding data is the

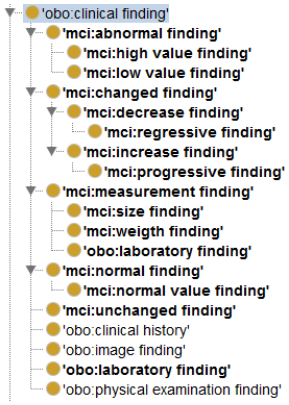

Fig. 2. Added subclasses under 'obo:clinical finding' (in bold).

distinction between 'normal findings' and 'abnormal findings'. For instance, lab values are tagged as being 'low', 'normal' or 'high' for some given patient. Similarly, a lymph node presenting certain characteristics, such as a short axis diameter larger than 1 cm (or 1.5 cm, depending on its location) is categorized by radiologists as 'abnormal' while lymph nodes not showing these changes are classified as 'normal'. An abnormal anatomical structure might be pathological or non-pathological: under 'snap:material_entity' we defined 'mci:anatomical structure' with subclasses 'mci:non-pathological anatomical structure' and 'obo:pathological anatomical structure'. Important for evaluation of the health status is the detection of changes of findings throughout the body. As an example lymph nodes might be progressive or regressive in size, body temperature might increase or decrease and the synopsis of this information can be crucial for the physician to make the correct diagnosis. Further,

[7] The German 'Operationen und Prozedurenschlüssel' is a coding system for procedures.

[8] http://www.ifomis.org/bfo/1.1

[9] http://code.google.com/p/omrse/

[10] http://code.google.com/p/information-artifact-ontology/

[11] http://informatics.mayo.edu/CNTRO/

we group findings that can be measured such as laboratory values, size findings and body weight findings. Additionally, we defined 'mci:examination modality' (under 'obo:data item') with subclasses 'mci:imaging modality' and 'mci:patient position'.

2.2.2 Added Object Properties The import of RO, IAO and CNTRO brings us 62 object properties - we added some more for the relation between instance data (see figure 3). For instance, we

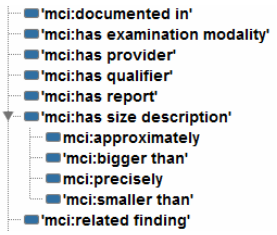

Fig. 3. Added object properties.

have the 'mci:has provider' property which relates some 'obo:data item' or 'obo:health care process' with a health care provider (or more precisely an entity which 'obo:is bearer of' some 'obo:health care provider role' as shown in figure 4). The direct relation of the patient to the respective diagnosis is implicitly given through the named graph mechanism (described below). The 'mci:documented

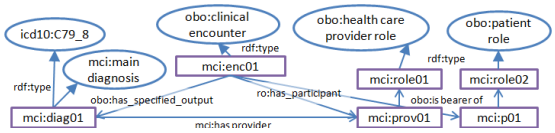

Fig. 4. The patient's main diagnosis related to a health care provider.

in' is used to link information to the source, where it was extracted from (e.g. a finding might be documented in a report or an image). The 'mci:size description' properties are used for the description of size findings such as lymph node enlargements. Typical statements like 'all lymph nodes in thorax under 1 cm', or 'axillary lymph node 1.4 cm' can be represented using these relations (see figure 6).

2.2.3 Added Data Properties Data properties are used for temporal information and finding descriptions. Even though we imported some of the object properties from CNTRO for the temporal structuring of events (see section 2.1), we do not import the CNTRO object properties for start and end dates since we found it more convenient to define them as data properties. On the one hand we cannot use the standard inference mechanisms of CNTRO for event sequencing as described in (Tao *et al.*, 2012), but on the other hand it is a lot easier to write SPARQL queries with restrictions or computations on date values. The inference of relations like 'cntro:before' or 'cntro:during' based on these data properties is done with by SPARQL UPDATE queries. Relations to identifiers are defined as subclasses of 'skos:notation'[12]. All added data properties are illustrated in figure 5.

[12] http://www.w3.org/2004/02/skos/

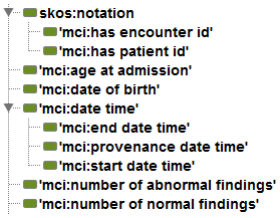

Fig. 5. Added data properties.

2.2.4 Link to Terminologies The representation of heterogeneous clinical data using terminologies might be realized through links from instance data (individuals) to terminologies (mainly classes). These links can be established either using the rdf:type or an annotation property. If we have a class in the terminology exactly matching our need such as the relation to ICD codes, then rdf:type is a good solution for the link and reasoning mechanisms can be easily applied. If a statement is more complex and there is no single corresponding concept in the terminology, we need to post-coordinate concepts from the terminology, e.g. for further location specifications (right, dorsal ...). For this purpose we defined annotation properties like, e.g., 'mci:has qualifier'. Here it would not make sense to use the rdf:type property. The current version of MCI has links to ICD-10, OPS[13], ATC[14], LOINC, the FMA and RadLex.

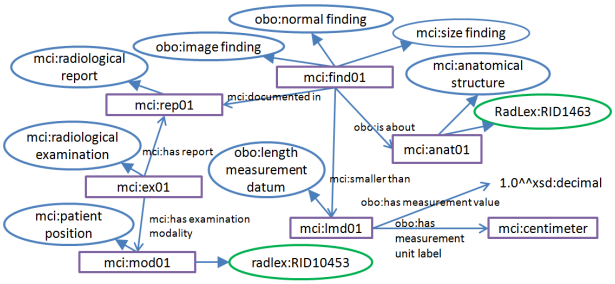

Fig. 6. Representation of a size finding and examination modalities (RID10453 = 'standing position', RID1463 = 'lymph node of thorax').

2.3 Data Sets and Named Graphs

For clarity and better query performance we separate the triples using different datasets for MCI, the instance data and the referenced ontologies. Additionally, the separation allows us to have different reasoning levels for the different datasets. MCI is held with OWL-reasoning while the patient data and the referenced ontologies without any reasoning. Further, we use named graphs in order to group patient data triples for the context of clinical encounters. The separation of triples belonging to different clinical encounters is necessary since e.g. some clinical department might realize different roles within the context of different encounters (e.g. admission/discharge role). Similarly, the 'mci:age at admission' makes sense only within the context of some clinical encounter.

[13] We transformed the the OPS hierarchy from in XML to RDF using XSLT.

[14] We transformed the hierarchy of the German version of the Anatomical Therapeutic Chemical Classification System from Excel to RDF.

3 EVALUATION

As an initial step we transformed the provided data of six melanoma patients from an i2b2 database into MCI and stored the RDF triples in a Jena Fuseki triple store. All structured and coded data, i.e. demographics, diagnosis (ICD-10), procedures (OPS), administered drugs (ATC), laboratory values (LOINC), health care providers were successfully imported. MCI integrates structured data and defines their semantics through the ontological basis (OGMS). Standardized established vocabularies are used to represent the data. Using federated SPARQL queries it is possible to combine data of MCI with knowledge contained in other ontologies. Data can be retrieved on different granularity: e.g. one can get the number of abnormal findings within some time interval or one could retrieve newly appeared high value findings. With respect to coverage we restricted this initial work to structured data – the representation of unstructured data is future work. Clearly we need a more thorough evaluation based on a larger set of patients with different diagnoses, to test performance and coverage of MCI.

4 RELATED WORK

MCI is based on the OGMS ontology, which is based on the BFO1.1. There are other ontologies besides OGMS, extending the BFO with clinical terms like, e.g. the Computer-based Patient Record (CPR)[15] or the Translational Medicine Ontology (TMO) (Luciano *et al.*, 2011). Even though CPR defines a lot of important core concepts of the clinical domain (which inspired the creation of MCI), it was easier for us to base MCI on OGMS due to its clarity and illustrative descriptions of concepts (Scheuermann *et al.*, 2009). TMO is more focused on studies and the representation of molecular and genomic data and lacks classes for a detailed representation of clinical findings other than measurements. Due to the available patient data we had, TMO was somehow not the right ontology to start with. As we extend our model to molecular and genomic data an alignment to TMO seems to be appropriate.

Other OGMS-based ontologies like, e.g., the Oral Health and Disease Ontology (OHD) or the Ontology for Newborn Screening, Follow-up and Translational Research (ONSTR) are very domain specific, containing e.g. findings like 'onstr:positive newborn dried blood spot screening test finding' MCI would only reference.

There are other approaches realizing the representation of clinical patient data through a combination of some standardized information model with terminologies. The mostly used information models are the HL7 Reference Information Model (RIM)[16] and the OpenEHR Entry Model[17] (Beale and Heard, 2007). In (Markwell *et al.*, 2008) the combination of HL7 and OpenEHR with SNOMED CT are compared. In (Schulz *et al.*, 2010) and (Karlsson *et al.*, 2011) the relationship of information models and ontologies is analyzed. In (Rector *et al.*, 2006) a methodology for the binding of information models with ontologies is presented and demonstrated for the binding of HL7 RIM with SNOMED CT. Since the HL7 RIM lacks ontological consistency (Smith *et al.*, 2013) one can see our work as an attempt to create an information model on the basis of well defined established upper-level ontologies and thus with clear semantic interpretation of the high level concepts.

[15] http://code.google.com/p/cpr-ontology/

[16] http://www.hl7.org/implement/standards/rim.cfm

[17] http://www.openehr.org/

With respect to data integration from various sources the i2b2 platform[18] provides a good basis. With i2b2 all heterogeneous clinical data related to a patient are available at one point and structured data can be mapped to standardized coding systems. The problem however is that since i2b2 data is not in RDF format the data cannot be easily combined with knowledge contained in ontologies like SNOMED CT. Additionally, the i2b2 data schema is too generic to represent more complex data such as findings descriptions.

5 CONCLUSION

We presented initial work on an OGMS-based Model for Clinical Information. Extensions will be made as more clinical data becomes available. Further we need to find a solution for encoding clinical data at a very granular level in such a way that more detailed clinical questions can be answered. Since omics data is becoming more important in clinical practice and especially in the context of personalized medicine, we will extend MCI to capture omics data as well. For this purpose we plan to align MCI to TMO.

ACKNOWLEDGEMENTS

This research has been supported in part by the DI4CS Project, which is funded by the German Federal Ministry of Economics and Technology under grant number 01MQ07016. The responsibility for this publication lies with the authors.

REFERENCES

Beale, T. and Heard, S. (2007). An ontology-based model of clinical information. *Studies in health technology and informatics*, (Pt 1), 760–4.

Karlsson, D., Berzell, M., and Schulz, S. (2011). Information Models and Ontologies for Representing the Electronic Health Record. In *ICBO2011*, pages 6–10, Buffalo, NY, USA.

Luciano, J. S., Andersson, B., Batchelor, C., Bodenreider, O., Clark, T., Denney, C. K., Domarew, C., Gambet, T., Harland, L., Jentzsch, A., Kashyap, V., Kos, P., Kozlovsky, J., Lebo, T., Marshall, S. M., McCusker, J. P., McGuinness, D. L., Ogbuji, C., Pichler, E., Powers, R. L., Prud'hommeaux, E., Samwald, M., Schriml, L., Tonellato, P. J., Whetzel, P. L., Zhao, J., Stephens, S., and Dumontier, M. (2011). The Translational Medicine Ontology and Knowledge Base: driving personalized medicine by bridging the gap between bench and bedside. *Journal of biomedical semantics*, **2**(Suppl 2), S1.

Markwell, D., Sato, L., and Cheetham, E. (2008). Representing clinical information using SNOMED Clinical Terms with different structural information models. In K. Spackman and R. Cornet, editors, *Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008)*, pages 72–79.

Rector, A., Qamar, R., and Marley, T. (2006). Binding Ontologies & Coding systems to Electronic Health Records and Messages. In *KR-MED 2006*, pages 11–19.

Scheuermann, R. H., Ceusters, W., and Smith, B. (2009). Toward an Ontological Treatment of Disease and Diagnosis. In *2009 AMIA Summit on Translational Bioinformatics*.

Schulz, S., Schober, D., Daniel, C., and Jaulent, M.-C. (2010). Bridging the semantics gap between terminologies, ontologies, and information models. *Studies in health technology and informatics*, **160**(Pt 2), 1000–4.

Smith, B., Vizenor, L., and Ceusters, W. (2013). Human Action in the Healthcare Domain: A Critical Analysis of HL7's Reference Information Model. In *Johanssonian Investigations. Essays in Honour of Ingvar Johansson on His Seventieth Birthday*, pages 554–573. Ontos Verlag.

Tao, C., Wongsuphasawat, K., Clark, K., Plaisant, C., Shneiderman, B., and Chute, C. G. (2012). Towards Event Sequence Representation , Reasoning and Visualization for EHR Data. In *2nd ACM SIGHIT International Health Informatics Symposium*, pages 801–805, Miami, Florida, USA.

[18] https://www.i2b2.org/