MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010
© 2010 IMIA and SAHIA. All rights reserved.
doi:10.3233/978-1-60750-588-4-1164

A Model-Driven Approach for Biomedical Data Integration

David Carlson[a], Ariel Farkash[b], John T.E. Timm[c]

[a] *Contractor to the US Veterans Health Administration, Kalispell, MT, United States*
[b] *IBM Haifa Research Lab, Haifa University Campus, Mount Carmel, Haifa, Israel*
[c] *IBM Almaden Research Center, San Jose, CA, United States*

Abstract

A core challenge in biomedical data integration is to enable semantic interoperability between its various stakeholders as well as other interested parties. Promoting the adoption of worldwide accepted information standards along with common controlled terminologies is the right path to achieve this. Our paper describes a solution to this fundamental problem by proposing an approach to semantic data integration based on information models serving as a common language to represent health data coupled with technology that is able to represent the data semantics. We used the HL7 v3 Reference Information Model (RIM) [1] to derive a specific data model for the integrated data, the Web Ontology Language (OWL) [2] to build an ontology that harmonizes the metadata from the disparate data sources, the Unified Modeling Language (UML) [3] to model the data representation, and the Object Constraint Language (OCL) [4] to specify UML model constraints. To illustrate the approach, we use the Essential Hypertension Summary CDA document and related models from Hypergenes, a European Commission funded project [5] exploring the Essential Hypertension disease model.

Keywords:

CDA, Ontology, OWL, Modeling, UML, OCL

Introduction

Biomedical information repositories typically contain data related to a specific clinical domain with semantics unique to the originating systems [6]. These disparate data sources pose a challenge for data integration [7] that is paramount for improved patient-centric care [8], as well as for secondary use of the data for analysis of aggregated data in context of clinical research, public health surveillance, and decision support [9].

In this paper, we depict a complete solution to this fundamental problem by proposing an approach to semantic data integration using healthcare standard information models, ontology-based metadata harmonization, technology for creating and constraining data models, and an engine for instance generation.

The solution we present, as depicted in Figure 1, starts with a clinical domain expert creating an ontological representation of the information elements or variables of interest needed for a particular study. Based on past experiences, the clinical domain expert does not care about explicit data format, but only that certain data elements are required for further analysis.

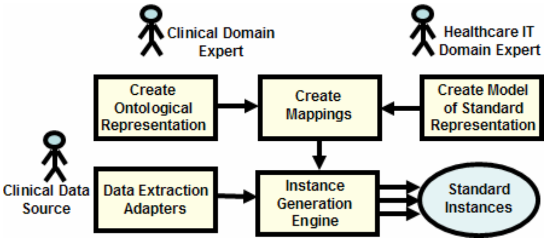

Clinical Domain Expert
Healthcare IT Domain Expert
Create Ontological Representation
Create Mappings
Create Model of Standard Representation
Clinical Data Source
Data Extraction Adapters
Instance Generation Engine
Standard Instances

Figure 1 – Solution Activity Diagram

Our approach is intended to work over multiple, heterogeneous data sources by using models based on international standards for healthcare semantics and interoperability. Using standard exchange formats, along with a set of constraints, serves to unify data into a semantically unambiguous format that makes operations on the data straightforward from a technological standpoint. The healthcare IT domain expert, familiar with healthcare data representation methods and standards, creates healthcare interoperability models. Mappings between the ontological representation and healthcare interoperability models enable the instance generation engine to produce the standard instances in the last step.

IT industry-standard modeling languages form the bridge between the clinical and healthcare IT domains and user roles required for proper integration of healthcare data. The clinical domain expert works with a “more intuitive” ontology-based approach using semantic web technologies to represent the metadata needed for harmonization, while the healthcare IT domain expert uses software modeling languages to create model-based representations of the standard format, apply constraints to this format for a domain of interest, and, in collaboration with the clinical domain expert, create mappings between the ontological representation of the variables of interest and the standard-based information models. The annotated model created by the healthcare IT domain expert at design time, is then used by the instance generation engine at runtime in order to transform the data to the standard format that conforms to the constrained model.

Background & Related Work

The HL7 v3 Reference Information Model (RIM) is used to derive consistent health information standards such as laboratory, problem and goal-oriented care, public health, and clinical research. It is an ANSI and ISO-approved standard that provides a unified health data 'language' to represent associations between entities who play roles that participate in acts. For example, a person entity plays a role of a surgeon who participates in a procedure act, and so forth. Acts may relate to other acts through "act relationships", thus providing a mechanism to describe complex actions.

Clinical Document Architecture (CDA) [10] is a constrained subset of the RIM that specifies terminology-encoded structure and semantics for clinical documents. These documents can be serialized to XML that conforms to a published W3C XML Schema. In most applications, the general CDA structure is further constrained by a set of templates that are standardized and published in an implementation guide, such as the Continuity of Care Document (CCD) [11]. Healthcare applications that produce or consume XML instances for CDA must include the appropriate template identifiers, as specified in the implementation guide. For example, a CDA instance that includes the template identifier "2.16.840.1.113883.10.20.1.28" indicates that the instance conforms to the CCD problem observation.

Most CDA template specifications, such as CCD, are written using structured English expressions that are based on the XML schema element relationships. These conformance statements are usually implemented using Schematron rules to augment the CDA XML schema. Our work, however, includes methods and open source software tools for representing CDA documents and template constraints using the Unified Modeling Language (UML) and the Object Constraint Language (OCL). Details and examples of this approach are described in the Methods and Results sections of this paper.

The UML modeling language is dominant among IT domain users, whereas clinical domain experts often work with formal ontology definitions. The Web Ontology Language (OWL) is a semantic markup language for publishing and sharing ontologies on the World Wide Web. It is endorsed by the World Wide Web Consortium (W3C) [12]. OWL is often used as the framework for converging distinctive terminologies into a single coherent ontology; many successful examples exist in clinical research and medical informatics domains [13, 14].

There has been some prior work in both using OWL ontologies in conjunction with instance generation [15], and in using OWL to add semantic annotations to UML information models [16]. These methods are applied and extended to support ontological mapping, representation modeling, formal constraining, and instance generation in our research.

Methods

Users

The use case diagram in Figure 2 illustrates the primary activities involved in our approach and the user roles required to perform these activities.

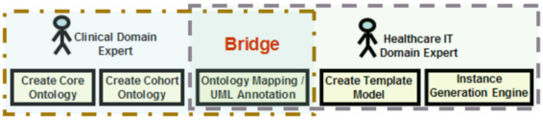


Figure 2 – Use Case Diagram

The clinical domain expert is responsible for creating the core ontology. The core ontology contains conceptual abstractions for a given clinical research domain and includes all the data elements required for secondary use by clinical researchers. The cohort ontology contains data elements and terminology specific to a data source. The cohort ontology is created by the clinical domain expert for each cohort that wishes to participate in the data integration. Using common ontology development tools such as Protégé [17], mappings are created between these cohort ontologies and the core ontology. This process is described in greater detail in the next sub section.

The healthcare IT domain expert is responsible for creating the CDA template model using a UML tool. The CDA template model contains classes, attributes, and relationships that are used to further constrain the CDA model to a particular clinical research domain. There are implicit relationships between classes in the template model and concepts in the core ontology. These relationships are made explicit by creating mappings on the CDA template model as UML annotations, providing the basis for generating the annotated template model.

The artifacts produced by these different users and the relationships between them are captured in Figure 3 below.

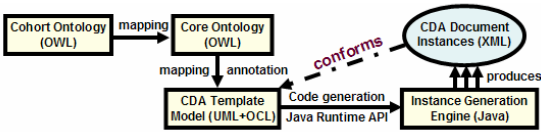


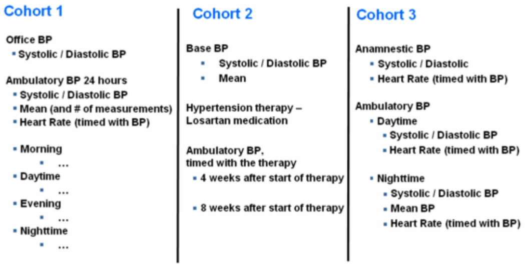
Figure 3 – Artifact Relationships

Data Integration

Healthcare data integration involves harmonization, validation, normalization, and transformation into standard structures that are accepted by the healthcare and medical research communities. Relationships between data items are often defined implicitly, e.g., in documentation or as tacit knowledge of experts. These implicit relationships must be expressed in an explicit and standard way so that analysis algorithms not aware of the implicit semantics can use them effectively.

Harmonization

Integration of data from dissimilar data sources must first undergo a process of conceptual harmonization, i.e. convergence of the sources metadata to a single and agreed-upon terminology. For example, blood pressure measurements from three different cohorts of essential hypertension are outlined in Figure 4. This outline depicts the underlying data model for the blood pressure measurements taken by the three cohorts.



Cohort 1

Office BP
- Systolic / Diastolic BP

Ambulatory BP 24 hours
- Systolic / Diastolic BP
- Mean (and # of measurements)
- Heart Rate (timed with BP)

- Morning
  - …
- Daytime
  - …
- Evening
  - …
- Nighttime
  - …

Cohort 2

Base BP
- Systolic / Diastolic BP
- Mean

Hypertension therapy – Losartan medication

Ambulatory BP, timed with the therapy
- 4 weeks after start of therapy
- 8 weeks after start of therapy

Cohort 3

Anamnestic BP
- Systolic / Diastolic
- Heart Rate (timed with BP)

Ambulatory BP
- Daytime
  - Systolic / Diastolic BP
  - Heart Rate (timed with BP)
- Nighttime
  - Systolic / Diastolic BP
  - Mean BP
  - Heart Rate (timed with BP)

Figure 4 – Various blood pressure measurement schemes

Comparing data between the different cohorts is not a trivial task. The metadata is named differently, so how can one deduce that: Cohort 1 "Office BP", Cohort 2 "Base BP", and Cohort 3 "Anamnestic BP" all refer to the same conceptual data? Furthermore, looking at Ambulatory Blood Pressure findings one can see that Cohort 1 temporal divisions are to "Morning, Daytime, Evening, and Nighttime", whereas in Cohort 3 we find "Daytime and Nighttime" only; Cohort 2 blood pressure observations relate to four and eight weeks after start of therapy, thus completely incomparable to the above data.

Using OWL, we leveraged technology used for semantic web representation, to map all cohort variables to a core ontology able to represent the base conceptual terms for the target domain, e.g. Essential Hypertension. A schematic diagram is shown in Figure 5.

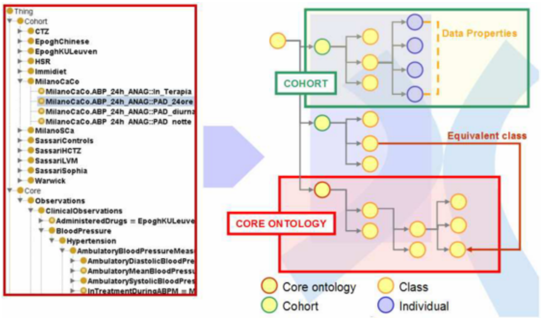

Thing
Cohort
CTZ
EpoghChinese
EpoghKULeuven
HSR
Immidiet
MilanoCaCo
MilanoCaCo.ABP_24h_ANAG::In_Terapia
MilanoCaCo.ABP_24h_ANAG::PAD_24ore
MilanoCaCo.ABP_24h_ANAG::PAD_diurna
MilanoCaCo.ABP_24h_ANAG::PAD_notte
MilanoSCa
SassariControls
SassariHCTZ
SassariLVM
SassariSophia
Warwick
Core
Observations
ClinicalObservations
AdministeredDrugs = EpoghKULeuve
BloodPressure
Hypertension
AmbulatoryBloodPressureMeas
AmbulatoryDiastolicBloodPre
AmbulatoryMeanBloodPressu
AmbulatorySystolicBloodPres
InTreatmentDuringABPM = M
COHORT
Data Properties
Equivalent class
CORE ONTOLOGY
Core ontology
Class
Cohort
Individual

Figure 5- Ontology schematic diagram, left side is a screen capture of ontology using Protégé

The process starts by creating a cohort class (OWL class) for each metadata variable, thus each cohort contains a flat list of cohort classes. We then map each cohort variable in accordance with harmonization effort to a core ontology class by specifying an equivalent class relationship. In case of n:1 mapping, cohort instances (OWL individuals) are created, allowing the class to maintain 1:1 mapping, and additional parameters (as Data Properties) are added to capture the instance disparities. Thus, following the example shown in Figure 4, Cohort 2 Ambulatory Blood Pressure would contain one class with two individuals, having a temporal parameter to specify for four or eight weeks after therapy.

Normalization & Validation

Having crossed the hurdle of defining metadata in comparable terms, one is left with the challenges of deducing and validating data values for each metadata variable under the cohort's data model, as well as normalizing values in correspondence to harmonized standard units. This task is a complex one due to differences in units of measurement, classifications, and diversity of protocols. We do not elaborate here on these efforts.

Transformation to Intermediate Data Representation

Data is first extracted via a suitable adapter from data source proprietary formats, such as an excel file or MySQL database, and copied into a generic *data container*. The data container is conceptually a map where the *key* is a cohort variable and the *value* is the matching value. Additional inference is performed by the instance generation engine receiving both the data container and the ontology as input.

Capturing Data Semantics

Having similar sets of metadata represented in an agreed-upon terminology provides the basis for semantic interoperability [18]. Biomedical data is typically complex, consisting of associations and dependencies between discrete data items as well as between common structures. Consider the example in Figure 4. In Cohort 2, the Ambulatory Blood Pressure is measured while the subject is treated by a medication called Losartan. Associating the act of observing the blood pressure and the act of administering the drug will explicitly represent the semantics. This relationship is crucial to physician as the significance of high blood pressure while under a Losartan regimen is different than under other circumstances. To capture the full context of the data, these kinds of associations should be established during the data integration process when the experts responsible for the data source can provide the implicit semantics often hidden in unstructured documentation or in their minds.

As described in the background, the HL7 v3 RIM provides a unified 'language' to represent such relationships and context. CDA, as a RIM-derived domain specific standard, facilitates the explicit representation of the rich semantics of healthcare data. Referring back to the examples discussed above, the blood pressure measurements are represented as CDA observations and, when appropriate, these observations are associated with a substance administration of Losartan.

CDA Model

The CDA UML model was created as an implementation model that is primarily based on two artifacts: (1) the CDA Refined Message Information Model (R-MIM) from HL7 and

(2) the CDA XML Schema. This implementation model was developed to support the existing code generation and serialization mechanisms present in the Eclipse Modeling Framework (EMF) [19]. The model was imported into an EMF model and ultimately transformed into a set of Java classes. The Java classes in conjunction with a set of additional utility classes make up the base runtime API that can be used to produce, consume and validate instances of CDA.

Template Modeling &Annotation

The template model is a domain-specific model that constrains the CDA model. Classes in a template model extend those in the CDA model. Constraints are modeled using directed associations, property redefinitions, and OCL expressions. The CDA Profile for UML is used to capture additional metadata needed during model transformation and at runtime. Annotations on template model elements including UML classes and properties are used to describe all core ontology variables and their possible parameterizations, each appearing at a unique location in the template model. Annotations are used to map between the core ontology and the CDA template model. After an annotated template model has been created, it is transformed into an implementation model which leads to the generation of a domain-specific API for constructing and validating instances.

Instance Generation Engine

The instance generation engine takes a data container that contains data values corresponding to variables in the cohort ontology as input and produces CDA document instances that conform to the template model. Using the ontology mappings, which were specified by the clinical domain expert at design-time, it resolves each variable in the data container to a corresponding variable in the core ontology. Annotations from the template model are then used to map core ontology variables to unique paths in the output tree and store data values in the leaves of the tree. Values that were specified as default or fixed in the template model such as template identifiers and coded attributes are also generated automatically. Additionally, we support a registry of CDA templates that enables instance validation.

Results & Discussion

In the frame of Hypergenes, an FP7 European Commission funded project exploring the Essential Hypertension disease model, we had to deal with 18 historical cohort data sources with diverse clinical and environmental data. We chose HL7 v3 RIM meta-model and data types for data representation and CDA as our data model. Additionally we needed to apply a template to constrain CDA to a document specialized for describing an Essential Hypertension Summary document (EH-CDA). Needless to say it was a perfect opportunity to put theory to test. In this section we will describe how the technology was used as well as illustrate a concrete example based on work done for Hypergenes project.

Essential Hypertension Ontology

Hypergenes project assimilated clinical data from 18 cohort data sources. The harmonization process involved consulting with scientific experts in order to elucidate exact intention in each data element. The metadata was discussed at length in order to identify the list of variables, their meaning, variable associations, value ranges, and additional parameterization. The core ontology taxonomical structure was built based on data analysis of preliminary results and the macro-classes of intermediate phenotypes and environmental risk factors defined for Essential Hypertension. The core ontology was used as a reference for mapping the variables in each of the cohorts.

Essential Hypertension Template Model

Once metadata was fully accounted for, we created a template model that constrains CDA to Essential Hypertension report. Figure 6 depicts an excerpt pertaining to Blood Pressure Finding; the full model comprising of several hundred templates.

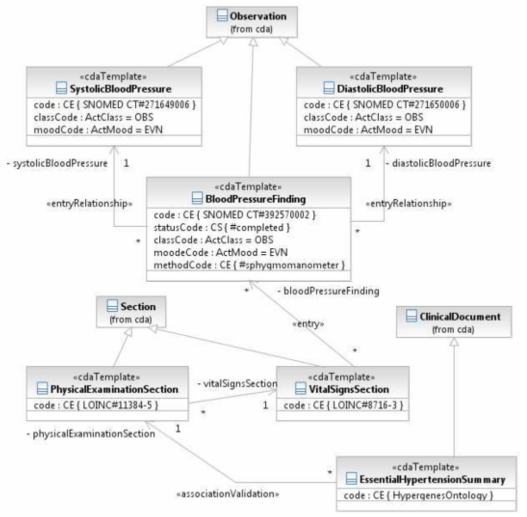
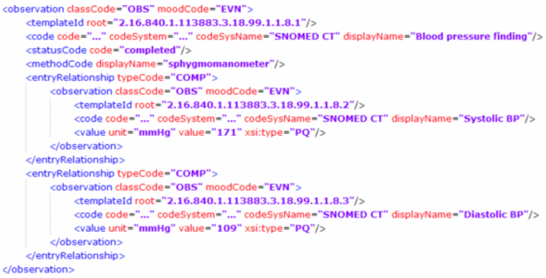

Observation
(from cda)
«cdaTemplate»
SystolicBloodPressure
code : CE { SNOMED CT#271649006 }
classCode : ActClass = OBS
moodCode : ActMood = EVN
«cdaTemplate»
DiastolicBloodPressure
code : CE { SNOMED CT#271650006 }
classCode : ActClass = OBS
moodCode : ActMood = EVN
- systolicBloodPressure
1
1
- diastolicBloodPressure
«entryRelationship»
«cdaTemplate»
BloodPressureFinding
code : CE { SNOMED CT#392570002 }
statusCode : CS { #completed }
classCode : ActClass = OBS
moodeCode : ActMood = EVN
methodCode : CE { #sphygmomanometer }
«entryRelationship»
*
*
*
- bloodPressureFinding
Section
(from cda)
«entry»
ClinicalDocument
(from cda)
*
«cdaTemplate»
PhysicalExaminationSection
code : CE { LOINC#11384-5 }
- vitalSignsSection
«cdaTemplate»
VitalSignsSection
code : CE { LOINC#8716-3 }
*
1
- physicalExaminationSection
1
*
«cdaTemplate»
EssentialHypertensionSummary
code : CE { HypergenesOntology }
«associationValidation»

Figure 6 – Blood Pressure Observation in template model

The *BloodPressureFinding* class in the EH-CDA template model *extends* the *Observation* class from the CDA model. The template identifier was specified in a property of the <<cdaTemplate>> stereotype. Additionally, the code attribute was used to capture metadata about the specific code in SNOMED-CT. This gives the template precise semantics from a clinical perspective. The directed associations in the diagram (e.g. *VitalSignsSection* to *BloodPressureReading*) were converted into equivalent OCL constraints during the model-to-model transformation.

Instance Generation for Essential Hypertension

The model in Figure 6 was used to generate a runtime API that enabled the creation of the CDA XML instance of a Blood Pressure finding depicted in Figure 7 below.



```
<observation classCode="OBS" moodCode="EVN">
    <templateId root="2.16.840.1.113883.3.18.99.1.1.8.1"/>
    <code code="..." codeSystem="..." codeSysName="SNOMED CT" displayName="Blood pressure finding"/>
    <statusCode code="completed"/>
    <methodCode displayName="sphygmomanometer"/>
    <entryRelationship typeCode="COMP">
        <observation classCode="OBS" moodCode="EVN">
            <templateId root="2.16.840.1.113883.3.18.99.1.1.8.2"/>
            <code code="..." codeSystem="..." codeSysName="SNOMED CT" displayName="Systolic BP"/>
            <value unit="mmHg" value="171" xsi:type="PQ"/>
        </observation>
    </entryRelationship>
    <entryRelationship typeCode="COMP">
        <observation classCode="OBS" moodCode="EVN">
            <templateId root="2.16.840.1.113883.3.18.99.1.1.8.3"/>
            <code code="..." codeSystem="..." codeSysName="SNOMED CT" displayName="Diastolic BP"/>
            <value unit="mmHg" value="109" xsi:type="PQ"/>
        </observation>
    </entryRelationship>
</observation>
```

Figure 7 – EH-CDA Blood Pressure finding Observation

As the hypertension model contained hundreds of templates to model we used the Jena API, Eclipse UML2 API, and models for CDA, data types, and vocabulary from the MDHT project to programmatically generate the template model from a minimal complete, conforming instance [20]. The template model was decorated with annotations that map variable names from the core ontology to relative paths in the instance. We used a depth first traversal of the template model to convert these relative paths into a map of variable to absolute paths. Given an incoming record (i.e. data container), variables were extracted and used to look up the absolute path which was in turn used to construct the corresponding path of objects in the instance. We followed this approach for 11,472 records coming from 4,000 patients deriving from 18 historical cohorts of Hypergenes project. Each record contained up to 1500 unique data elements or variables that mapped to the same number of paths in the output instance.

Conclusion

In this paper we discussed a model-driven approach for integrating biomedical data using three complementary technologies. We used semantic technology in the form of an ontology definition language (namely OWL) to describe data elements of interest for a particular clinical research domain. We discussed the use of XML-based healthcare interoperability standards for clinical data and the role they play in semantic interoperability across multiple data sources. Finally, we discussed the use of UML to bridge the gap between the clinical domain expert and the healthcare IT domain expert and to facilitate the generation of a runtime that produces conforming instances. We validated our approach by using it to integrate clinical data in the EU-funded Hypergenes project.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Program FP7/2007-2013 under grant agreement n° 201550.

References

[1] HL7 Reference Information Model, Health Level Seven, http://www.hl7.org/v3ballot/html/infrastucture/rim/rim.htm

[2] OWL, http://www.w3.org/TR/owl-features/

[3] UML at OMG, http://www.omg.org/spec/UML/2.0/

[4] Object Constraining Language specification, http://www.omg.org/technology/documents/formal/ocl.htm

[5] Hypergenes FP7 EC Project, http://www.Hypergenes.eu/

[6] Stroetmann V. et al. “Semantic Interoperability for Better Health and Safer Healthcare”. SemanticHEALTH Project Report, 2009. Published by the European Commission, http://ec.europa.eu/information_society/ehealth

[7] Heiler S: Semantic interoperability. ACM Computing Surveys 27(2):pp271-273, 1995.

[8] Gold JD, Ball MJ. “The Health Record Banking imperative”, IBM Systems Journal, Vol 46, No 1, 2007

[9] Bock BJ. et al. “The Data Warehouse as a Foundation for Population-Based Reference Intervals”. American Journal of Clinical Pathology, 120; pp662-670, 2003.

[10] Dolin RH. et al, “HL7 Clinical Document Architecture, Release 2”, JAMIA 2006;13:pp30-39

[11] Ferranti et al, “The Clinical Document Architecture and the Continuity of Care Record: A Critical Analysis”, JAMIA 2006; 13:pp245-252.

[12] Smith MK, Welty C, McGuinness DL. http://www.w3.org/TR/owl-guide/, 2004

[13] Schultz S., Boeker M., Stenzhorn H., “How Granularity Issues Concern Biomedical Ontology Integration“. MIE, p. 863, 2008.

[14] Golbreich C., Zhang S., Bodenreider O., “The foundational model of anatomy in OWL: Experience and perspectives”, Web Semantics: Science, Services and Agents on World Wide Web, Vol 4, Issue 3:pp181-195, 2006.

[15] Farkash A. et al. “Biomedical Data Integration – Capturing Similarities While Preserving Disparities”, proceeding of IEEE EMBC 2009.

[16] Carlson D, “Semantic Models for XML Schema with UML Tooling,” proceeding of SWESE 2006.

[17] Protégé, a free, open source ontology editor and knowledge-base framework. http://protege.stanford.edu/

[18] Heiler S: Semantic interoperability. ACM Computing Surveys 27(2):271-273, 1995.

[19] Eclipse Modeling Framework (EMF), http://www.eclipse.org/modeling/emf/

[20] Farkash et al. “Facilitating the Creation of Semantic Health Information Models from XML Contents”, CSHALS 2010

Address for correspondence

Ariel Farkash
IT for Healthcare & Life Sciences
IBM Haifa Research Lab.
E-mail: arielf@il.ibm.com