31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Biomedical Data Integration – Capturing Similarities while Preserving Disparities

Stefano Bianchi[2], Anna Burla[1], Costanza Conti[3], Ariel Farkash[1], Carmel Kent[1], Yonatan Maman[1], Amnon Shabo[1] (authorship ordered alphabetically)

[1]IBM Haifa Research Labs, Haifa University, Mount Carmel, 31905, Haifa, Israel
[2] Softeco Sismat S.p.A., Via De Marini 1, WTC Tower, 16149, Genoa, Italy
[3] IMS-Istituto di Management Sanitario SRL - via Podgora, 7-20122 Milano, Italy

Abstract — **One of the challenges of healthcare data processing, analysis and warehousing is the integration of data gathered from disparate and diverse data sources. Promoting the adoption of worldwide accepted information standards along with common terminologies and the use of technologies derived from semantic web representation, is a suitable path to achieve that. To that end, the HL7 V3 Reference Information Model (RIM) [1] has been used as the underlying information model coupled with the Web Ontology Language (OWL) [2] as the semantic data integration technology. In this paper we depict a biomedical data integration process and demonstrate how it was used for integrating various data sources, containing clinical, environmental and genomic data, within Hypergenes, a European Commission funded project exploring the Essential Hypertension [3] disease model.**

I. INTRODUCTION

HEALTHCARE information systems typically contain data and knowledge related to a specific health domain with semantics unique to the originating systems [4] posing a challenge to data integration [5]. Health data warehouses are established in an attempt to accomplish such integration and support patient-centric care [6] as well as secondary use of the data such as analysis of aggregated data in the context of clinical research, public health surveillance, and systems optimization [7]. In healthcare, personalized care involves taking into account the clinical, environmental and personal genetic variations of the individual in the care processes which makes the data even more diverse [8,9]. Integration of data with diverse semantics and maintaining coherent semantics are emphasized in management of longitudinal electronic health records (EHR) or in conducting a more focused analysis based on deep understanding the data. On the semantic dimension, these requirements present conflicts that need to be addressed.

This paper describes a solution to this fundamental problem by proposing an approach of semantic data integration based on information models serving as a common language to represent health data coupled with a technology that is able to represent the data semantics. We used the HL7 v3 Reference Information Model [1] (RIM) to derive a specific data model for the integrated data, as it is a well accepted healthcare standard, while Web Ontology Language (OWL)[2] was used to build an ontology in order to converge the metadata from the disparate data sources each having a proprietary data model and terminology. The ANSI/ISO-approved RIM provides a unified 'language' to represent actions made by entities. Health data is described by associations between entities who play roles that participate in actions. For example, an organization entity plays a role of laboratory that participates in an observation action, or a person entity plays a role of a surgeon who participates in a procedure action, and so forth. Actions are related to each other through "act relationship" providing the mechanism to describe clinical statements such as "this procedure was done with the indication of that observation". The RIM includes the unique attributes of each of the entities, roles, participants and actions that are relevant to health, described in an object-oriented manner, e.g., the observation action inherits the Act class attributes and extends it, e.g., with a value attribute while the procedure action extends Act with a target site attribute, etc. The RIM is merely the underlying 'reference' model and thus is used by the various standardization groups to develop domain-specific standards such as laboratory, pharmacy, clinical documents and many others. Typically, those domains are generic in nature and can be used across the various clinical specialties. All specs have XML implementations (W3C Schemas) to enable exchange of information across networks.

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 [10]. OWL is frequently used as the framework for converging distinctive terminologies into one coherent ontology [11][12][13].

II. SEMANTIC DATA INTEGRATION

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

A. Harmonization

Integration of data coming from dissimilar data sources about the same clinical, environmental and genetic

phenomena must first undergo a process of conceptual harmonization, i.e. convergence of the sources metadata to a single and agreed-upon terminology. Take for example blood pressure measurement variables from three different cohorts of essential hypertension. Figure (1) depicts an outline of the underlying data model for three cohorts regarding their representation of blood pressure measurements taken

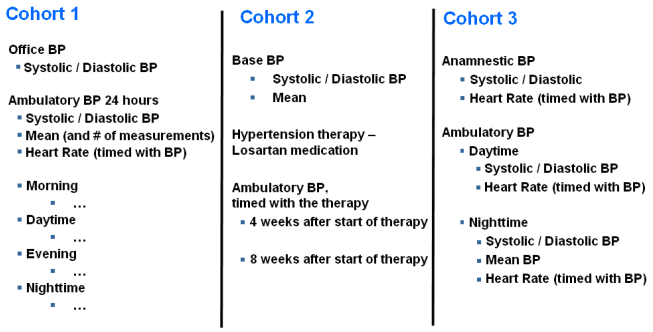

Fig. 1. Similarity and disparity in blood pressure measurement schemes

One can easily see that comparing data between the different cohorts is not a trivial task. For starters the metadata is named differently, how could 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.

B. *Normalization*

Having crossed the hurdle of defining the metadata in comparable terms, one is still left with the challenge of deducing the meaning of the data for each metadata variable under the cohort's data model. This is due to:

- Use of different units of measurement
- Use of different classification systems
- Diversity in study protocols and in the classes of possible answers, e.g. measurement methodologies or repeated measurements

A simple example for this would be normalization of proprietary values such as internal enumerations, e.g. the data for patient gender in Cohort A may contain 0=Male, 1=Female, while in Cohort B 1=Male, 2=Female.

C. *Capturing Richness of Data*

Having similar sets of metadata represented in an agreed-upon terminology provides the basis for syntactic interoperability [14], i.e. the ability to compare two orphan items of data and reason about its level of similarity. However, biomedical data is typically complex, consisting of associations and dependencies among discrete data items as well as among common structures. As aforementioned, in the RIM, actions are related to each other through "act relationships" providing the mechanism to describe complex actions. Let's look once more at the example in (1): in Cohort 2, the Ambulatory Blood Pressure measurement is measured while the subject is treated by a medication for hypertension called Losartan. This calls for the association of the act of observing the blood pressure to the act of administering the drug so that semantics if explicitly represented in the warehouse. This information is sometimes crucial for the physician; high blood pressure measurement while under Losartan regimen has a completely different meaning then without such intervention. Another example would be when Blood Pressure and Heart Rate measurements are timed with a diagnostic procedure such as Echocardiography. In this case, Blood Pressure and Heart Rate values are not relevant for the diagnosis of follow up of hypertension itself, since measured when the patient is in a potential stressful situation, but are significant in interpreting Echocardiography findings.

Therefore, in order to capture the full richness of the data those associations should be established, preferably during the data integration process when the experts responsible for the data source can provide the implicit semantics often hidden in unstructured documentation or merely in their minds. As aforementioned, the RIM provides a unified 'language' to represent health actions such as observations, procedures and substance administrations. It facilitates the explicit representations of the rich semantics of the data. In the examples discussed above, the blood pressure and heart rate measurements are represented as RIM observations and when appropriate, these observations are associated with an echocardiography procedure or with a substance administration of Losartan.

D. *Ontology Creation*

In order to be able to compare data of different cohorts, one should first map all cohort variables to a core terminology. We used the Web Ontology Language (OWL), originating from the field of web semantic 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.

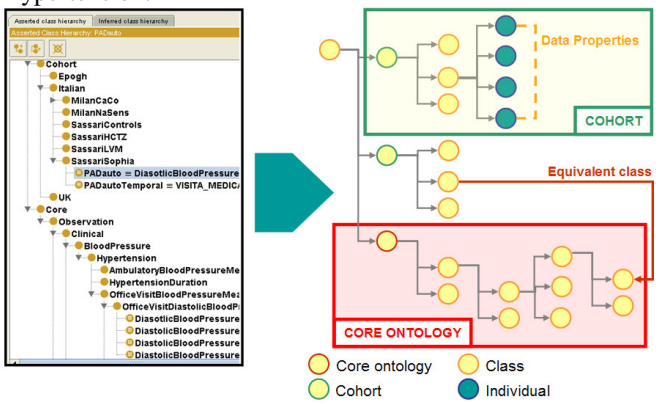

Fig. 2. Ontology schematic diagram,
left side is a screen capture of the ontology using Protege [15]

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 according to process depicted in (2). In case additional parameters are needed to refine the cohort variable definition, a cohort instance (OWL individual) is created with the needed data parameters (as Data Properties), e.g., temporal parameters. Thus, following the example depicted in (1), Cohort 2 Ambulatory Blood Pressure would need two individuals that would have the extra specification for four or eight weeks after therapy.

E. Terminology Standardization

Having a core ontology that encapsulated the needed terms to describe appropriately the target domain is mandatory for internal data comparison, i.e. of data for all cohorts in a given consortium thus conforming to the same proprietary core ontology. However, in order to be able to exchange data interoperably beyond the scope of the consortium one must rely on a standard terminology that is common and acceptable in the healthcare community. Thus, we mapped variables of the core ontology to corresponding standard terminologies. Thus, lab results, anthropomorphic measure and miscellaneous medical observations were mapped to terminologies such as LOINC[16] and SNOMED CT[17] whereas disease specification such as Ischemic heart disease and coronary thrombosis were delegated to an ICD[18] standard terminology such as: ICD8, ICD9-CM, and ICD10.

III. INSTANCE GENERATION

Once metadata is harmonized, normalized and standardized, data was verified to be adequate to its corresponding metadata under the given terminology. Finally, proprietary data is ready to be transformed to standard instances.

A. Data Model

The RIM is only the underlying meta-model for expressing domain-specific data. As aforementioned, various standardization groups develop generic standards such as laboratory, pharmacy and clinical documents. Those RIM-derived standards have XML implementations (W3C XML Schemas) to capture the complex semantics in the dominant structure for content representation[19]. XML instances that follow these schemas are adjusted to the jargon of the domain. For example a Clinical Document Architecture (CDA) instance will use terms such as: section, consumable and entry, whereas these terms are mapped to Act [classCode=DOCSECT], Participation [typeCode=CSM], and ActRelationship [typeCode=COMP] correspondingly, under the RIM meta-model.

B. Data Representation and Constraints

Prior to instance generation a two step constraining process must be performed. The first step entails selecting the most appropriate RIM-based standard to encapsulate the data. The second step involves additional constraining that should be applied to those generic standards.

1) Domain Standards selection: Existing standards are selected to be the basis of the data model. For example, in our research at Hypergenes, the standards Clinical Document Architecture (CDA) [20], Genetic Variation (GV) [21] and Family History (FH) [22] were used to best capture the clinical, environment, and genomic data involved.

*2) Templates development***:** A template is an expression of a set of constraints on the RIM or a RIM derived model that is used to apply additional constraints to an instance of data which is expressed in terms of some other Static Model. Templates are used to further define and refine these existing models to specify a narrower and more focused scope [23]. The standards selected in the first step are further constrained to create templates of the standards which reflect the specific structure to fit the target domain, e.g., Essential Hypertension. This step includes merely constraining of the selected standards so that instances are always valid against the generic standard but also comply with the templates [24]. As stated above, we use the RIM in order to capture the richness of the data. Nevertheless, richness may lead to diversity, thus we use templates as a means to facilitate semantic interoperability among interested parties by narrowing down the large number of compositional expressions allowed by the RIM to a nailed-down structure for each piece of data. By supplying a "closed template", i.e. a strict template specification constraining each datum to a specific location within the XML, we enable users and utilities, such as decision support tools, to comprehend unequivocally the integrated biomedical data.

IV. HYPERGENES USE CASE SUMMARY

A. Overview

Hypergenes is a European Commission funded project that aims at building a method to dissect complex genetic traits using essential hypertension as a disease model [3]. Most of the common-complex, chronic diseases, that have a high prevalence in our populations, arise through interaction between genetic, environmental and life-style factors. To understand the composite origin of these diseases, there is a need to know the path from genotype to phenotype. To that effect, the Hypergenes consortium includes major data sources of genomic, clinical and environmental data, each having its proprietary data set and data model.

B. Data Description

Hypergenes consortium contains data from twelve data sources regarding 4000 individuals (hypertensive and normotensive subjects). Genomic data included information on one million tag single nucleotide polymorphism (tag-SNPs) [25] obtained by an DNA microarray [26] created by fully robotized Illumina Bead Station 500 GX [27], for array based high-throughput SNPs genotyping. Clinical and environmental data collected from the disparate data sources varied in data model and terminology. Each cohort contained between 30 and 500 variables plagued with proprietary data enumerations, different languages, duplicity, partial similarity with implicit or no specification of

relationship, parameterization, and miscellaneous internal inconsistencies that had to be resolved.

C. Data Integration Process

The harmonization process involved consulting with scientific experts in order to elucidate exact intention in each data set. First, metadata was discussed to identify the list of variables and their meaning, variable associations, and parameterization. The next iterations were aimed at deducing the meaning of the data assigned to each variable, i.e. the units used, translation of terms, and disambiguation of proprietary enumerations, e.g., 15mm vs. 1.5cm, si/no vs. yes/no and true/false, and 0=Male/1=Female vs. 1=Male/2=Female and M=Male/F=Female, etc.

The core ontology taxonomical structure was built based on data analysis preliminary results and the macro-classes of intermediate phenotypes and environmental risk factors defined for Essential Hypertension. The mapping process used the core ontology as a reference mapping for the variables in each cohort. During the mapping a spreadsheet was created associating semantically equivalent variables of different cohorts to the core ontology. Among the most significant issues encountered in this task were:

1) Cohorts variables that need their original context through 'bound variables', e.g., dates, times, types of instruments used at diagnosis, etc.

2) Mapping across cohorts that requires one-to-many or many-to-many associations. For similar but non-comparable semantics, either changes in the core ontology or conversion rules were introduced.

This effort iteratively refined the OWL ontology described in section II D.

Creating an Essential Hypertension specific template for instance generation involved the refinement of HL7v3 CDA [20] for representing the clinical and environmental data and Genetic Variation [21] for representing clinical-genomics data with raw genomic data encapsulated in HL7-compliant instances. In order to specify the constraints specific to Essential Hypertension, we needed to create a HL7v3 Template [23]. Since there is not yet an approved representation of a template, we temporarily created a 'template instance', i.e. an all-encompassing instance that includes an XML structure skeleton for each datum required. A JAVA based engine was built that takes as input the OWL ontology, the template instance, and data from a cohort and generates a valid XML instance that follows the domain standard schema while adhering to the constraints specified by the template.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's Seventh Framework Program FP7/2007-2013 under grant agreement no 201550.

REFERENCES

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

[2] Web Ontology Language, [Online]. Available: http://www.w3.org/TR/owl-features/

[3] Hypergenes FP7 European Commission Project, [Online]. Available: http://www.Hypergenes.eu/

[4] Veli N. Stroetmann et al. "Semantic Interoperability for Better Health and Safer Healthcare". SemanticHEALTH Project Report. January 2009. Published by the European Commission, [Online]. Available: http://ec.europa.eu/information_society/ehealth.

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

[6] J. D. Gold, M. J. Ball. (2007). "The Health Record Banking imperative: A conceptual model" IBM Systems Journal, Vol 46, No 1.

[7] BJ Bock, CT Dolan, GC Miller, WF Fitter. "The Data Warehouse as a Foundation for Population-Based Reference Intervals". American Journal of Clinical Pathology, 2003, 120:662-670.

[8] Ruano G: Quo vadis personalized medicine? Personal Med (2004) 1(1):1-7.

[9] Davis RL, Khoury MJ: The journey to personalized medicine. Personal Med (2005) 2(1):1-4.

[10] Smith, Michael K.; Chris Welty, Deborah L. McGuinness (2004-02-10). [Online]. Available: http://www.w3.org/TR/owl-guide/.

[11] Dee McGonigle, Kathleen Mastrian, "Nursing Informatics and the foundations of knowledge". p.97 [Online]. Available: http://nursing.jbpub.com/informatics

[12] Stefan Schultz, Martin Boeker, Holger Stenzhorn, "How Granularity Issues Concern Biomedical Ontology Integration". MIE (2008), p. 863.

[13] Christine Golbreich, Songmao Zhang, Olivier Bodenreider, "The foundational model of anatomy in OWL: Experience and perspectives", Web Semantics: Science, Services and Agents on the World Wide Web, Volume 4 , Issue 3 (September 2006), Pages 181-195

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

[15] Protege, a free, open source ontology editor and knowledge-base framework. [Online], available: http://protege.stanford.edu/

[16] Logical Observation Identifiers Names and Codes (LOINC), [Online], Available: http://loinc.org/

[17] SNOMED Clinical Terms (SNOMED CT), [Online], available: http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

[18] International Classification of Diseases (ICD), [Online], available: http://www.who.int/classifications/icd/en/

[19] Shabo, S. Rabinovici-Cohen, P. Vortman, "Revolutionary impact of XML on biomedical information interoperability". IBM Systems Journal, Volume 45, Number 2, 2006 [Online]. Available: http://www.research.ibm.com/journal/sj/452/shabo.html

[20] HL7v3 Clincal Document Architecture, Release 2. [Online], http://www.hl7.org/v3ballot/html/infrastructure/cda/cda.htm

[21] HL7v3 Clinical Genomics, Genetic Variation. [Online], Available : http://www.hl7.org/v3ballot/html/domains/uvcg/uvcg.htm

[22] HL7v3 Clinical Genomics, Pedigree Topic, Family History. [Online], http://www.hl7.org/v3ballot/html/domains/uvcg/uvcg_Pedigree.htm#POCG_DO000000UV-Pedigree-ic

[23] Template project on the HL7 ballot site, [Online], Available: http://www.hl7.org/v3ballot/html/infrastructure/templates/templates.htm

[24] Li J, Lincoln MJ. Model-driven Clinical Document Development Framework. AMIA Annu Symp Proc. 2007 Oct 11:1031.

[25] International Hapmap Project. [Online], Available: http://www.hapmap.org/whatishapmap.html

[26] DNA microarray from Wikipedia. [Online], Available: http://en.wikipedia.org/wiki/DNA_microarray

[27] Illumina Bead Array Reader [Online], Available : http://www.illumina.com/pages.ilmn?ID=29