pHealth 2021
B. Blobel and M. Giacomini (Eds.)

doi:10.3233/SHTI210590

Understanding the Gap Between Information Models and Realism-Based Ontologies Using the Generic Component Model

Mathias BROCHHAUSEN[a,1], Sarah J. BOST[b], Nitya SINGH[c], Christoph BROCHHAUSEN[d], and Bernd BLOBEL[e,f,g]
[a] *Dept. of Biomedical Informatics, University of Arkansas for Medical Sciences* [a], *USA*
[b] *Dept. of Health Outcomes and Biomedical Informatics, University of Florida, USA*
[c] *Emerging Pathogens Institute & Dept. of Animal Sciences, University of Florida, USA*
[d] *Institute of Pathology & Central Biobank, University and University Clinic of Regensburg, Regensburg, Germany*
[e] *Medical Faculty, University of Regensburg, Germany*
[f] *eHealth Competence Center Bavaria, Deggendorf Institute of Technology, Germany*
[g] *First Medical Faculty, Charles University of Prague, Czech Republic*

Abstract. The wide-spread use of Common Data Models and information models in biomedical informatics encourages assumptions that those models could provide the entirety of what is needed for knowledge representation purposes. Based on the lack of computable semantics in frequently used Common Data Models, there appears to be a gap between knowledge representation requirements and these models. In this use-case oriented approach, we explore how a system-theoretic, architecture-centric, ontology-based methodology can help to better understand this gap. We show how using the Generic Component Model helps to analyze the data management system in a way that allows accounting for data management procedures inside the system and knowledge representation of the real world at the same time.

Keywords. Information Models, Biomedical Ontologies, Knowledge Representation, Systems Theory, eHealth

1. Introduction

As Biomedical Informatics (BMI) is increasingly moving towards using Artificial Intelligence (AI), including Knowledge Representation and Reasoning approaches on datasets created by integrating multiple databases and datasets, BMI practice and research continues encountering problems created by the difference in requirements for information models and data integration/harmonization. Since information models are

[1] Corresponding Author, Mathias Brochhausen, Department of Biomedical Informatics, University of Arkansas for Medical Sciences, 4301 W. Markham St., Slot 782, Little Rock, AR 72205, USA; E-mail: mbrochhausen@uams.edu.

specifically designed to inform database schema development, the obstacles to data integration created in the information models are inherited in database schemata. Data about one person is frequently stored in more than one database system, with the different systems partially holding different types of data. The integration of multiple databases is crucial in order to draw meaningful inferences, e.g., about causative infectious agents, their temporal patterns, and outbreak detection, which is prime to infection prevention and public health best practices. A unique key identifier is typically associated with each entry in each database and can be used to retrieve the related information from either system. However, the same patient can be infected or sampled multiple times and at multiple locations in the course of a disease or infection, generating multiple associated entries in each system and multiple unique key identifiers, one for each database system. Thus, lack of strategic decisions on knowledge representation considerations can hinder the integration of the multiple databases due to data discrepancies. This is a major hurdle in drawing meaningful inferences from integrated data sets, especially in real-time, for making informed medical decisions. They negatively affect our ability to use and integrate data, for example when linking the database, managing the patient's demographic and follow-up patient-reported data about travel, food intake, exposure to potential carriers, etc. with the biobank samples, and their integrated database, and managing different molecular analyses related to infectiological results. Other limitations of information models are described and overcome, e.g., in Blobel et al. and Oemig and Blobel [1, 2].

Likewise, in many biobanks the entry number of a specimen is used as the reference number. Over the course of time, a tumor patient will have several entries, each receiving a new entry number. This does not only apply to specimens derived from the primary tumor, but also to those derived from secondary tumors or metastases. In the biobank, these samples must be identified and documented in the so-called sample history as follow-up samples for the first tumor. Typically, the sample history assignment is done using the patient identifier. Many systems enforce that there be only one patient identifier for each patient. If the patient has all samples taken at the same healthcare institution, this does not present a problem. However, if some of the sampling is done at another healthcare provider, tracing disease progression via the patient identifier is no longer possible because the entry number cannot be matched with the patient identifier at the location of the tissue bank. This is a common challenge especially in regional Comprehensive Cancer Centers in Germany, where patients are treated by different health care providers, and the central biobank collects samples from all the health care providers within the Cancer Center intake area. The fact that the information model enforces the existence of one and only one patient identifier hinders the integration of data about specimens curated outside of the biobank system. Thus, the requirement for a unique patient identifier, while understandable from the perspective of intra-system development, negates the fact that human beings are patients with multiple healthcare providers and, consequently, may have multiple patient identifiers.

There is no doubt that information models are useful and that using information models to inform database schemata is an important strategy. However, the example of integrating biobank data presented above raises the question of how the interplay between information model and domain representation can be orchestrated to ensure that both are compatible and do not lead to errors and false inferences.

In a recent paper, Brochhausen et al. [3] discussed the lack of computable semantics in medical Common Data Models and problems arising from falsely ascribing semantic capabilities to those important tools. In this paper, we explore a representational issue

that frequently occurs when using IT-oriented information models as sole representational resources. IT-oriented information models fulfill a crucial role in planning, defining, and describing the operational behavior of an IT system, such as an EHR system or a biobank information system. Due to the operational focus, the resulting representation, especially when relationships are represented, is frequently focused on data in that system alone and not on what the data represents, in our case the medical world, e.g., the patients, encounters, and prescriptions described in an EHR or the specimens, donors, and storage properties in a biobank. This problem also creates difficulties for approaches that seek to develop a knowledge representation resource (e.g. ontology-driven data management) based on an information model [4]. We stress the importance of bringing in biomedical expertise when representing the biomedical domain and the relations therein. In this paper, we discuss a use-case showing how the Generic Component Model (GCM) (Figure 1) in interplay with realism-based ontologies can be used to bolster information representation with domain specific knowledge, in this case based on clinical care and research.

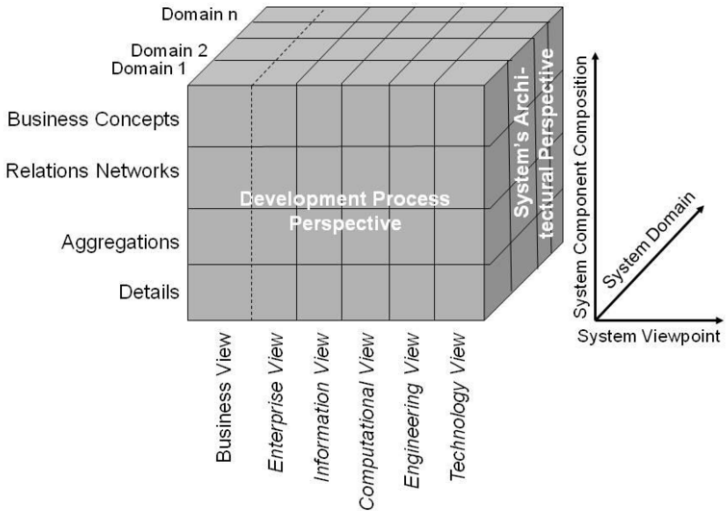

Figure 1. The Generic Component Model

The GCM is a top-level architectural model for multi-domain systems, describing the system components, their functions and interrelations structurally and behaviorally, thereby representing specific aspects (domains) by related subsystems. It is described and specified in ISO 23903 [5]. For each business case, the subsystem components and their functions and interrelations are instantiated by naming and representing them using the specific terminologies and ontologies of the domains involved in that business case. For enabling this representation of the real world system by its information technology-independent domain ontologies, the GCM specifies a Business View in addition to the five views defined by the ISO 10746 Open Distributed Processing Reference Model (RM-ODP) [6]. Furthermore, ISO 23903 introduces generic granularity levels for correctly representing and interrelating compositions/decompositions of elements. The views prescribed by the RM-ODP are Enterprise View (purpose, scope, and policies of

the system), Information View (information processing, semantics of information), Computational View (functionality of the system, functional decomposition), Engineering View (distribution of processing performed by the system), and Technology View (choice of technology for the system) [6]. Notably, a descriptive representation of the domain and its composition/decomposition is missing, which the GCM compensates by adding the Business View and granularity levels, that way correctly representing multi-domain real world systems.

The general problem of unique identifiers is not the main focus of this paper. We use longitudinal identification of tissue samples in a biobank as the leading example of how restrictions from data models are hindering integration approaches for consumption in AI methods. This is of crucial importance for the further development of precision medicine.

2. Methods

Applying the methods to the use-case at hand first requires considering the role of the information models and the database schemata according to the views of the GCM. Both belong to the Information View, since their primary focus is to guide information processing within the system. Hence, they do not provide real world knowledge, such as the fact that one person can have more than one patient ID and that sampling of tumor progression is frequently done by different healthcare providers. Notably, those aspects lie outside each individual EHR system and are ill-fitted to be represented within the Information View. However, from our database example, it is clear that not accounting for those aspects of the Business View may lead to errors in the system.

Our approach is to use representation of data in the Resource Description Framework (RDF) along with realism-based ontologies. Smith & Ceusters [7] propose a realism-based ontology development as the general methodology to inform knowledge representation. Their approach is based on experience with multiple implementations in biomedical informatics, which have influenced the evolution of their methodology from the beginning [8-11]. Following this approach, we use realism-based ontologies to represent the Business View. Brochhausen and Blobel [12] have outlined a strategy to use the Generical Component Model (GCM) to assist the representation of relations in realism-based biomedical ontologies (Figure 2).

RDF [13] is a Semantic Web Standard that allows the representation of information in a machine-interpretable way. For each entity, RDF provides a unique identifier. RDF data can be annotated and used along domain descriptions provided in the Web Ontology Language [14]. We are using RDF representations and OWL ontologies following the principles described by Smith and Ceusters [7].

3. Results

Figure 2 shows the result of using an RDF representation of the data in our leading example. This representation allows for the representation of multiple entry numbers referring to specimens from a single donor. The different classes are modeled based on best practice of realism-based ontology developed as described in [7].

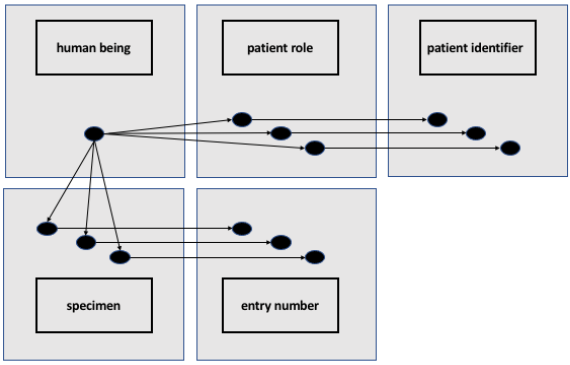

Figure 2. Representation of RDF individuals and OWL classes representing a human being having multiple patient roles, corresponding to multiple patient IDs and multiple specimens derived from that human being corresponding to multiple entry numbers.

4. Discussion

Our example shows that an information model restricting each patient to one and only one patient identifier is not sufficient to fulfill the GCM Information View, which provides the semantics. From a real world perspective, we do know that one person may be identified by multiple patient IDs throughout their life. The use of those patient IDs might overlap temporally, if the person sees multiple healthcare providers over the same period of time. In addition to the information model, a realism-based representation is needed that specifies that a person can be a patient at multiple healthcare providers and may have multiple patient IDs associated with them.

In the current stage, we have applied the GCM and the basic tenets of realism-based ontology development to one use-case. The use-case presented is an extremely common use-case in managing specimens from multiple organizations or biobanks. In addition, the successful deployment of ISO 23903 for integrating different domains and knowledge spaces including their specific models has been demonstrated for the integration of HL7 privacy and security specifications [15] in ISO 13606 EHR communication [16], the harmonization of concepts from ISO 12967 [17] and ISO 13940 [18] or the mapping of open EHR (ISO 13606) archetypes [19], ISO 13972 clinical models [20] and HL7 FHIR resources [21].

The next step is to develop computational ways to leverage the GCM system analysis and its results immediately in developing information models, database schemata, and ontologies. In addition, we plan to explore using the GCM to transition from information models to ontologies and *vice versa*.

Traditionally, technical standards have been focused on a specific domain or technology, while healthcare is multi-disciplinary by nature and therefore in need of the GCM multi-domain architecture, impacting many standards from HL7 [22] or ISO/TC215 [23]. With the turn to more complex approaches such as IoT or Smart Cities, etc., also other domains and their standards developing organizations such as OMG [24], IEEE 70xx projects [25] or ISO/IEC JTC1 [26] adopt ISO 23903.

5. Conclusion

Based on these preliminary results, we conclude that the GMC can inform development of information models and knowledge representation resources, thus, filling a relevant gap in AI usage for biomedical data.

References

[1] Blobel B, Ruotsalainen P, Oemig F. Why Interoperability at Data Level Is Not Sufficient for Enabling pHealth? Stud Health Technol Inform. 2020;273:3–19.
[2] Oemig F, Blobel B. (2010) Harmonizing the Semantics of Technical Terms by the Generic Component Model. Stud Health Technol Inform. 2010;155:115–121.
[3] Brochhausen M, Bona J, Blobel B. The role of axiomatically-rich ontologies in transforming medical data to knowledge. Stud Health Technol Inform. 2018;249:38–49.
[4] Zhang Y-F, Tian Y, Zhou T-S, Araki K, Li J-S. Integrating HL7 RIM and ontology for unified knowledge and data representation in clinical decision support systems. Comput Methods Programs Biomed. 2016 Jan 1;123:94–108.
[5] International Organization for Standardization. ISO 23903:2021 Health informatics - Interoperability and integration reference architecture - Model and framework (retrieved from: https://www.iso.org/standard/77337.html)
[6] International Organization for Standardization. ISO/IEC 10746 Information technology - Open distributed processing.
[7] Smith B, Ceusters W. Ontological realism: A methodology for coordinated evolution of scientific ontologies. Appl Ontol. 2010 Nov 15;5(3–4):139–88.
[8] Hogan WR, Hanna J, Joseph E, Brochhausen M. Towards a Consistent and Scientifically Accurate Drug Ontology. CEUR Workshop Proc. 2013;1060:68–73.
[9] Ceusters WM, Spackman KA, Smith B. Would SNOMED CT benefit from Realism-Based Ontology Evolution? AMIA Annu Symp Proc. 2007;2007:105–9.
[10] Lin Y, Xiang Z, He Y. Brucellosis Ontology (IDOBRU) as an extension of the Infectious Disease Ontology. J Biomed Semant. 2011 Oct 31;2:9.
[11] Grenon P, Smith B, Goldberg L. Biodynamic ontology: applying BFO in the biomedical domain. Stud Health Technol Inform. 2004;102:20–38.
[12] Brochhausen M, Blobel B. Architectural approach for providing relations in biomedical terminologies and ontologies. Stud Health Technol Inform. 2011;169:739–43.
[13] Resource Description Framework (RDF) Primer, https://www.w3.org/TR/rdf11-concepts/ (retrieved 09/05/2021).
[14] Web Ontology Language (OWL) Primer, https://www.w3.org/TR/owl2-primer/ (retrieved 09/05/2021).
[15] HL7 International. HL7 Version 3 Domain Analysis Model: Composite Security and Privacy, Release 1
[16] International Organisation for Standardisation. ISO 13606:2019 Health informatics – Electronic health record communication. ISO: Geneva; 1998.
[17] International Organisation for Standardisation. ISO 12967:2020 Health informatics – Service architecture (HISA). ISO: Geneva; 2020.
[18] International Organisation for Standardisation. ISO 13940:2015 Health informatics – System of concepts to support continuity of care. ISO: Geneva; 2015.
[19] openEHR Foundation. www.openEHR.org (last accessed on 10 December 2015).
[20] International Organisation for Standardisation. ISO 13972:2015 Health informatics - Detailed clinical models – Characteristics and processes. ISO: Geneva; 2015. (Currently under revision to ISO 13972:2020 Health informatics - Clinical information models - Characteristics, structure and requirements.)
[21] HL7 International. HL7 Fast Healthcare Interoperability Resources (FHIR). http://www.hl7.org/fhir/
[22] HL7 International. http://www.hl7.org
[23] International Organization for Standardization. TC 215 Health informatics. https://www.iso.org/committee/54960.html
[24] Object Management Group. *https://www.omg.org*
[25] IEEE 70xx Standards on Ethics in Autonomous and Intelligent systems Standards. https://ethicsinaction.ieee.org/p7000/
[26] ISO/IEC JTC1 Information Technology. https://www.iso.org/isoiec-jtc-1.html