User-centered semantic harmonization: A case study
Chunhua Weng a,*, John H. Gennari b, Douglas B. Fridsma a

a Department of Biomedical Informatics, University of Pittsburgh, UPMC Shadyside Cancer Pavilion Suit 301, 5150 Centre Avenue, Pittsburgh, PA 15232, USA
b Department of Medical Education and Biomedical Informatics, University of Washington, Seattle, WA, USA

Received 15 September 2006
Available online 21 March 2007

Abstract
Semantic interoperability is one of the great challenges in biomedical informatics. Methods such as ontology alignment or use of metadata neither scale nor fundamentally alleviate semantic heterogeneity among information sources. In the context of the Cancer Bio-medical Informatics Grid program, the Biomedical Research Integrated Domain Group (BRIDG) has been making an ambitious effort to harmonize existing information models for clinical research from a variety of sources and modeling agreed-upon semantics shared by the technical harmonization committee and the developers of these models. This paper provides some observations on this user-centered semantic harmonization effort and its inherent technical and social challenges. The authors also compare BRIDG with related efforts to achieve semantic interoperability in healthcare, including UMLS, InterMed, the Semantic Web, and the Ontology for Biomedical Investigations initiative. The BRIDG project demonstrates the feasibility of user-centered collaborative domain modeling as an approach to semantic harmonization, but also highlights a number of technology gaps in support of collaborative semantic harmonization that remain to be filled.

Published by Elsevier Inc.

Keywords: Semantic interoperability; Semantic harmonization; Knowledge management and engineering; caBIG™; Collaborative domain modeling; Group consensus

1. Background and motivation
Despite years of research in biomedical informatics, interoperability among healthcare information systems remain an unresolved challenge [1]. A lack of interoperability is linked to medical errors and other problems: The US Department of Health and Human Services has stated that, "Interoperability is needed for clinicians to make fact-based decisions so medical errors and redundant tests can be reduced" [2]. According to the Institute of Electrical and Electronic Engineers (IEEE), interoperability is the ability for two or more systems first, to exchange information and second, to use the exchanged information. The former can be achieved by making data structures or formats consistent across systems and is also referred to as syntax interoperability, while the latter, semantic interoperability, deals with terms denoting different concepts or concepts with mismatched scopes and uses. Semantic interoperability failures introduce severe information integration errors that are much harder to detect and resolve than syntactic interoperability problems [2].

Semantic interoperability problems in the healthcare arena can be largely attributed to the decentralized nature of multidisciplinary healthcare communities and healthcare information systems. Numerous healthcare specialties have coevolved and established their own cultures and values for a long time. Therefore, stakeholders from disparate fields often organize healthcare information in different ways from varied perspectives. At present, despite the myriad terminology services, a big problem is the lack of explicit semantics shared among multidisciplinary practitioners and stakeholders across healthcare communities. Most medical concepts in use have heterogeneous and vague def-
initions which cannot be generalized across disciplines or organizations. Consequently, the uses of healthcare information often have to be confined to local contexts [3].

Semantic interoperability has presented a challenge to a wide range of integration activities involving exchange of healthcare information, geographic information, international laws, and electric devices. As an important practical problem, it has been tackled from different angles [4–8]. Methods to achieve semantic interoperability largely fall into the following three categories: model alignment, using semantic tags or metadata, and developing shared conceptual references.

1.1. Interoperability via model alignment

The first approach, model alignment, creates mappings among models to support their semantic interoperability [9–11]. It allows multiple models to co-exist and provides only an ad hoc integration solution that does not alleviate the heterogeneity among disparate data sources. Its advantage is that sources do not have to be modified to achieve interoperability. Its drawback is that every pair of information sources needs a mapping. Therefore, this approach has poor scalability. The number of required mappings is \( C^2_K \), which is proportional to the square of the number of data sources \( N \). Tremendous computing resources are required to construct and maintain mappings, which may still have limited accuracy because mapped terms across different systems do not necessarily share the same conceptual definitions or scopes of uses.

1.2. Interoperability via metadata

The second method is to use semantic tags or metadata [12], such as the Dublin Core Metadata Initiative [13], which is one of the key technologies for building the Semantic Web. Mappings are created not directly between data sources, but either between a data source and a metadata set or between different metadata sets. Supposing there are \( N \) data sources and \( K \) metadata sets, the number of mappings to be created is \( C^2_K + N \), which is proportional to the square of the number of metadata sets \( K \) and the number of the data sources \( N \). The smaller the \( K \), the more data sources will share the metadata sets and the fewer mappings need to be created to support the interoperability. However, autonomous metadata development is often localized and introduces potential semantic interoperability problems among metadata sets. Similar problems apply to the uses of standard terminology services since regular metadata or terminology services usually carry limited contextual information and cannot easily generalize across disciplines [14].

1.3. Interoperability via a shared conceptual reference model

The third approach, which is also the ideal solution to semantic interoperability, is to develop core ontology or a shared conceptual reference model to serve as the common ground for all systems or to define a shared metadata set [15–18]. The number of needed mappings equals the number of data sources \( N \).

The effort to construct mappings across data sources in the above three methods can be represented by a common formula: \( C^2_K + N \), where \( K \) is the number of metadata sets, and \( N \) is the total number of data sources. When \( K \) equals 1, all data sources are mapped to one metadata set provided by a shared conceptual reference model; when \( K \) equals \( N \), then all data sources require mappings for each pair of them. Therefore, construction of a shared conceptual reference model needs a minimum number of mappings to achieve semantic interoperability among information sources.

In reality, it is very difficult to construct a comprehensive conceptual reference model. Domain modeling, which is to conceptualize a domain and represent this conceptualization in computable knowledge as ontology or domain analysis models [19], is the common approach to developing shared conceptual reference models. However, most domain modeling efforts do not scale beyond individual organizations. Distributed autonomous domain modeling efforts have produced voluminous overlapping or inconsistent model resources for biomedical researchers, and contributed to the interoperability problems across healthcare information systems. A shared conceptual reference model is also required to support the interoperability of legacy systems built on existing models. Therefore, an important task in building a shared conceptual reference model is to effectively reuse extant knowledge resources from myriad domain analysis models that have been developed and have evolved constantly in a poorly-coordinated manner.

With the quickly expanding number of information and knowledge sources, researchers have been trying to find effective approaches to integrate or merge the semantics from different models. However, it is no trivial task. One panel in the 2004 WWW workshop on the semantic web for health science pointed out, “all the theoretical elements, including operations research, econometrics, systems theory, dynamical systems, and machine learning and model building, are in place today, but the practical tools for large-scale semantic unification and semantics-driven integration and interoperability are yet to be constructed” [7]. Domain consensus modeling is one of the most important challenges for domain modeling in healthcare presently [20,21].

1.4. Semantic harmonization

In this paper, we use “semantic harmonization” to refer to a domain consensus modeling approach, which includes three steps: (1) investigating semantically connected domain analysis models; (2) deriving common semantics based on the group consensus of the developers of the models; and (3) building a coherent conceptual reference
model through explicit and formal knowledge representations of the shared semantics. This approach is different from other domain modeling approaches for its consensus-based and user-centered processes. It requires community-based sharing, deriving, and integrating of common domain knowledge from distributed miscellaneous resources in a scalable manner in an open and collaborative environment [22]. It is related to an earlier concept called semantic unification [23], but highly emphasizes user-centered collaborative modeling processes, instead of reliance on automatic model integration or synthesis algorithms 

In practice, various forms of semantic harmonization for harmonizing international laws or electric device standards have existed for a long time [4–6,26]; however, the processes and methodologies have rarely been discussed and remained as vague and intricate interdisciplinary group work. Moreover, to our knowledge, there is no collaborative modeling technology that effectively supports user-centered semantic harmonization. We feel that there is a pressing need for developing effective approaches to semantic harmonization. The purposes of this paper are to share our experience with the BRIDG semantic harmonization effort, to examine the challenges and opportunities for supporting user-centered domain consensus modeling, and to inform future technology designs in support of this task.

2. Methodology of BRIDG

In the domain of cancer clinical trials, there is an increasing need for global trial data sharing and knowledge discovery. Many standards for clinical trials have been or are being developed by various clinical research standardization organizations, including the National Cancer Institute (NCI), the Food and Drug Administration (FDA), Health Level Seven (HL7), and the Clinical Data Interchange Standards Consortium (CDISC). For example, HL7 uses a Reference Information Model (RIM) to develop interchange specifications for healthcare information systems; CDISC is developing platform-independent standards to support the acquisition, exchange, submission, and archiving of clinical data for pharmaceutical companies and FDA. Meanwhile, many pharmaceutical and technology companies as well as academic researchers are developing clinical trial research knowledge models or ontologies for improving the quality of clinical trial protocols [27] or the computability of reported clinical trial data [28].

In February 2004, the Cancer Biomedical Informatics Grid (caBIG™) project was created as a voluntary virtual informatics infrastructure to develop interoperable cancer research tools [29]. The project initiated scientific collaboration among domain experts from more than 50 cancer institutes across the United States. As part of the caBIG™ efforts, the Biomedical Research Integration Domain Group (BRIDG) was formed in late 2004 to support protocol-driven clinical research through a shared domain analysis model. The mission of BRIDG is twofold: (1) to harmonize the semantics from available clinical trials information models into a shared model; and (2) to explore a methodology for user-centered semantic harmonization.

Next, we introduce the BRIDG project in terms of its desiderata for a shared reference model, modeling mechanism, major participants and source models and harmonization processes.

2.1. Desiderata of the BRIDG model

BRIDG has defined three desiderata for developing the shared reference model: that it be (1) comprehensive; (2) consensus-based; and (3) abstraction and context neutral.

1. Comprehensive: To enable the shared model to address the needs of the broad clinical research community, BRIDG has made “being comprehensive” its important design principle. By building an open-source modeling environment, BRIDG has widely connected and engaged interested stakeholders of clinical research from academia, industry, government, standardization agencies, and patient advocates to participate in the semantic harmonization activities.

2. Consensus-based: As a user-centered project, BRIDG strongly advocates community participation and group discussions. For all source models that need to be harmonized, model developers participate in at least one harmonization session with the BRIDG harmonization committee. During such face-to-face meetings, each concept, its attributes, and its relationships to other concepts are discussed thoroughly in the group to reach agreed-upon modeling decisions.

3. Abstraction and context neutral: Earlier research on the Unified Medical Language Systems (UMLS) identified the importance of collaboration at the right level of abstraction, where involved parties would have sufficient collective experience and understanding to reach consensus [9]. In accordance with this, BRIDG tries to separate content from representations and focuses on representing the abstract meanings of the concepts shared by the clinical research communities. Subsequently, the BRIDG model only contains concept definitions, attributes, and concept relationships. By design, it is not application-oriented and excludes all implementation-specific application information.

However, the implementation of these desiderata has practical challenges. For example, when striving to be comprehensive, BRIDG has difficulty in staying focused and prioritizing relevant source models. The BRIDG model is
also too abstract to be used directly by many application-oriented users. In Section 4, we will revisit these desiderata when we describe the unresolved research issues related to the BRIDG project.

2.2. Modeling language: UML

BRIDG chose the Unified Modeling Language (UML) as the representation language for its software engineering strengths. Domain experts can easily view graphical class relationships in UML diagrams. UML models also support a model-driven-architecture so that changes in UML models can be efficiently reflected in future applications that use the models. Moreover, UML supports package importing and exporting so that collaborative modelers can work on different portions of the shared model simultaneously.

2.3. Community participation

There are three major stakeholders on the BRIDG project, which are HL7, CDISC, and NCI. Representatives from these stakeholders make up an Advisory Board and a Technical Harmonization Committee (THC), which oversee the evolution of the BRIDG model. The approximately 10-member BRIDG Advisory Board serves to identify the priorities for harmonization, and the four-member THC maintains the BRIDG model and leads all harmonization meetings. Through the open-source collaborative mode, any interested individual or parties can join the BRIDG community, learn how to work within the BRIDG project and submit relevant source models for harmonization. At the very beginning, BRIDG started with only The Study Data Tabulation Model (SDTM) contributed by CDISC, and The Clinical Trial Object Model (CTOM) contributed by NCI, but it has grown to include regular representatives and content from other organizations including NCI, CDISC, HL7, FDA, WHO, FAET (The Federal Adverse Events Taskforce), PhRMA (The Pharmaceutical Research and Manufacturers of America), and the ClinicalTrials.gov. There are also some fluid members in the BRIDG project, who participate in the harmonization activities occasionally for a particular subproject of BRIDG of interest to them.

At present, there are six active domains within the whole BRIDG modeling space, including (1) lab data modeling, (2) patient study calendar, (3) clinical trials registration, (4) adverse events, (5) SDTM, and (6) CTOM. User participation in these domains is illustrated in Fig. 1.

2.4. Group communication and workflow

BRIDG participants are widely distributed across distances and organizations. To actively engage the collaboration of a variety of parties, BRIDG employs GForge, a web-based project management and collaboration software [30], to host the project and store all discussions and the shared model at http://www.bridgproject.org. Here any project team member can download the formative shared model. Although GForge provides online communication and collaboration features, such as messaging, these have never been used. Email list services and biweekly teleconferences are set up to facilitate communications among interested BRIDG participants and to share the latest modeling activities. About fifteen to twenty people have been participating in the biweekly teleconferences on a regular basis. A face-to-face harmonization meeting also occurs.

Fig. 1. Participation in the BRIDG modeling domains as February 2007.
every month. Source model providers and any interested parties can attend those meetings.

The evolving BRIDG model is divided into two packages: the harmonized area and the staging area. All sources models are first imported into the staging area for harmonization. Then the harmonization results are stored in the harmonized area. Classes archived in the harmonized area can be reused by anyone who is interested in developing BRIDG-compliant applications.

A source model to be harmonized can be in any format, including a UML model, an excel spread sheet, a database model, or an ontology, because the harmonization is focused on the concepts and their definitions, not on the representation differences. To date, the source models that BRIDG has encountered have been database-like data dictionaries or schemas, rather than richer ontologies. These sorts of source models are re-formatted and imported into the UML modeling environment through a manual process. BRIDG modelers extract key concepts from source models and then reconstruct a UML model for these concepts and import this model as a package into the staging area of the shared BRIDG model. If the source model was an ontology with higher-order constructs such as axioms, constraints, or necessary and sufficient conditions from a description logic representation, then these would have to be ignored or transferred as text into the UML environment. However, from a practical standpoint, current applications do not typically use such constructs.

Next, the technical harmonization committee prints out semantically connected concept definitions from both the sources and the shared model, juxtapose related concepts in an excel spread sheet, and distribute it to the group. This step helps modelers focus their attention on the semantically interconnected areas of related models to create mappings across models. Mappings are explicit representations of similarities and mismatches between the shared concepts (classes or attributes). During the harmonization meetings, for each concept, the chairman of the meetings reads aloud these definitions and the group discusses them and comes up with an agreed-upon definition for all. To arrive at a well-agreed-upon decision, the deliberation of a concept can take from an hour to four hours or more.

BRIDG employs stewardship to facilitate collaborative modeling. This means that only members of the technical harmonization committee can make changes to the shared model; others can only suggest changes to the technical harmonization committee. Therefore, after each harmonization meeting, a member from the technical harmonization committee applies changes to the shared model and distributes the new model to all on the shared project GForge site. On this basis, that source model is considered harmonized. The complete versioning history of the shared model is archived on the shared web site so that people can track the differences across versions. Therefore, semantic harmonization is an iterative and cumulative modeling process.

3. Result 1: BRIDG harmonization work practices

For the BRIDG project, we observed some common semantic heterogeneity across models to be harmonized and identified a set of typical semantic harmonization actions, as listed below.

3.1. Common semantic heterogeneity

Semantic heterogeneity can be either conceptualization mismatches, the way a domain is interpreted, or explication mismatches, the way a domain is represented [20]. In Table 1, we use the framework created by Visser and examples from the BRIDG project to illustrate the typical semantic mismatches across the domain models that we collected for BRIDG.

It is time-consuming and error-prone to manually categorize semantic relationships among related concepts from different domain analysis models. In the BRIDG project, this step is often the bottleneck of semantic harmonization processes.

3.2. A collection of harmonization actions

To resolve the above semantic mismatches, BRIDG has used a list of typical harmonization actions as follows:

Renaming a class: This action resolves any term mismatch. When two classes referring to the same concept have a term mismatch, one of the classes is renamed to be mapped to the other.

Redesigning an attribute (to rename or to change the type): When two attributes referring to the same concept have different data types or different terms, one of the attributes is renamed or its type is changed to be mapped to the other.

Unification of attributes for a concept: When models define different attributes for the same concept, the ultimate definition of the concept should include the unification of all possible attributes. This operator is used to resolve attribute or “concept and attribute” mismatch.

Moving attributes from a subclass to a super class: When both models have the same pair of super classes and subclasses but different attribute assignments, an attribute is moved from the subclass to the super class to make it more general.

Creating a more abstract super class for two mismatched concepts: This operator is used to resolve a concept mismatch. For example, if two models both contain the concept of “protocol”, but one refers to a document and the other refers to a series of activities, we could create a class called “protocol” as an abstract class, with “document” and “procedure” as two subclasses of it.

Creating a subsumption between two concepts: If one model defines “eligibility criteria” and the other model defines “inclusion criteria” and “exclusion criteria”, these two models represent the same concept at different abstraction levels. We can leverage “eligibility criteria” and make
it a super class of “inclusion criteria” and “exclusion criteria”.

Creating an association for two concepts: For example, “study” and “document” are two perspectives on a clinical trial protocol. For the principle investigator, a clinical trial protocol describes a study design to test a treatment, whereas for a clinical staff, a clinical trial protocol is a document to be used as a reference or guideline. Therefore, class “protocol” should have an association to both the class “study” and the class “document.”

Copying of a class: When a source model contains a distinctive class not present in the shared model, a full copy of this class is needed to insert this new concept to the shared model. Here a full copy means a copy of the class, its subclasses, and its super classes.

Deletion of an obsolete attribute: An obsolete attribute that is out of practice is removed.

Deletion of an obsolete class: An obsolete concept that contradicts the shared concepts is removed.

Model unification or merging is more of an art than a science [31]. There is no standard way to resolve each type of semantic mismatch listed in Table 1. Many of the above actions can be applied to solve each type of semantic mismatch. During a BRIDG harmonization meeting, initially several actions are recommended and reviewed, and then eventually one action is selected to revise the model. Group
discussion support is critical for reconciling diversified perspectives.

4. Result 2: unresolved research challenges for BRIDG

Over the past 18 months, the BRIDG project has revealed some challenges with regards to user-centered semantic harmonization: some are due to the gap between available technology and the goals of BRIDG, some are inherent to the task itself, and some have been caused by the to-be improved semantic harmonization methodology. Next we analyze the problems that we have encountered.

4.1. A gap between a reference model and applications

Designed to be a domain reference, the BRIDG model has been developed at a high level and to be context-neutral. Aimed at being generic and reusable for different application contexts, the BRIDG model has nevertheless been criticized for providing little application development support and for being disconnected from realistic application models. “How can we use the harmonized BRIDG model?” has been a recurring question frequently posted in teleconferences and harmonization meetings. Participants often first show enthusiasm in this effort and agreed with its importance but give up due to difficulties connecting the model to real-world uses.

In fact, this challenge is not unique to BRIDG, but inevitable for any reference model [32]. Reference ontologies are broad and deep, such as the comprehensive and abstract BRIDG model, while application models are narrow and concrete, and often focused on particular user needs. As pointed out by Brinkley et al. based on their experience with the Foundational Model of Anatomy, reference ontologies are too large and detailed to be used “out-of-the-box” in applications, even when developers are aware of them and would like to use them [32]. How we can make a reference model usable in practice is still an open research challenge because changes in the evolving reference models need to be systematically reflected in the applications of reference models. Looking back, we could have incorporated a broad range of application needs into consideration during the BRIDG harmonization process, instead of postponing this process until the completion of the BRIDG model. Timely consideration of application scenarios could have served as a formative evaluation tool for the abstract BRIDG model.

4.2. UML for semantics representation

Despite its desirable software engineering features listed in Section 2.2, UML is not a satisfying knowledge representation language for constructing a shared domain reference. UML diagrams are full of associations, which represent different relationships between classes, such as “whole and part”, “actor and action”, “cause and consequence”, and many others. These different types of associations created a lot of ambiguity in a UML model.

Moreover, UML is not a formal knowledge representation language for defining coherent class relationships and constraints. In UML diagrams, relations of a super class are not automatically inherited to subclasses; therefore, modelers can purposefully or unintentionally overwrite a super class when defining attribute and class-relationships for subclasses, which often create inconsistency in a UML model. Furthermore, UML does not well support knowledge reuse and attribute definitions cannot be easily shared across classes. For example, in the BRIDG model, many classes require the same attribute “ID”. Therefore is no mechanism for sharing the definition of the same attribute among different modelers. Sometimes “ID” is defined as a string, and other times “ID” is defined as an integer, which causes much redundancy and inconsistency.

Alternative domain modeling tools to UML include Protégé [33] and OWL [34], which provide more formal and coherent modeling mechanisms. However, these languages are necessarily further from the underlying database technology used by most applications. In addition, the default user interface for Protégé or Protégé/OWL requires a fairly high level of modeling sophistication, whereas our domain modelers were much more comfortable with the style of graphical modeling user-interfaces provided by UML tools. Therefore, we realize that no single existing domain modeling tool satisfies user needs perfectly, especially for the mixed BRIDG modeling group. A combination of multiple tools or a tool that integrates features from multiple tools may work for such collaborative modeling tasks in the future.

4.3. Knowledge provenance

Semantic harmonization essentially distills and integrates domain semantics from existing knowledge sources. It is a modeling process based on but beyond semantics merging. To make a shared reference model convincing to assorted communities in a domain, we need to provide source meta-information or a description of the origins of the shared semantics, which is also called knowledge provenance [35]. In addition, the harmonization result is often structurally different from all the sources. Knowledge provenance is necessary to support evidence-based modeling by making explicit the origins of the semantics in the harmonization result.

Earlier in the BRIDG project, the old versions of the BRIDG model did not provide sufficient knowledge provenance information so that some users could not tell how their source models related to the shared reference model. We learned this lesson from user feedback and inserted mappings to the original source models for each class and attribute in the current version of the BRIDG model. These mappings effectively provided preliminary provenance information; however, we saw a great need for a sys-
tematic approach to tracking and maintaining links between a harmonization result and the origins of the semantics.

4.4. Model alignment and merging

Semantic harmonization relies on the merging of semantically connected domain models, which is “the process of finding commonalities between two different ontologies A and B and deriving a new ontology C that facilitates interoperability between computer systems that are based on the ontologies A and B” [36]. During semantic harmonization, a key task is to compare various source models and unify those semantically connected areas. It is tedious to compare thousands of concepts in related models manually. In the BRIDG project, we did not use an automatic model alignment tool and we relied on manual concept comparisons, which was very time-consuming and error-prone. Although it may be that we could have used a tool such as PROMPT [10] for this step, it seemed as though the overhead cost for converting models into Protégé and learning PROMPT was too high. Most participants requested some assistance for model alignment such as receiving recommendations of merging options.

4.5. Handling divergence

Semantic harmonization is essentially a dynamic knowledge engineering process. The shared reference model and source models inevitably coevolved. On one hand, the shared reference model needs to be ready to assimilate any knowledge updates from source models to maintain knowledge provenance. On the other hand, the changes in the shared model are requested to be quickly communicated to model developers who use the shared reference model as a foundation for domain modeling. It is important to effectively support two-way synchronization, or divergence handling for both source models and the shared reference model. Version control has been studied in software engineering for a long time; however, divergence handling for co-evolving models is still a practical challenge.

Without appropriate technology support, the BRIDG team uses a stewardship mechanism to handle divergences. Only the steward of the BRIDG model, a technical harmonization member, can revise the BRIDG model. All other revision suggestions based on group discussions are reviewed by this person and applied to the shared model. This person also releases and distributes the new versions of the BRIDG model. A big problem with this method is, as mentioned above, that the steward can be a bottleneck to the harmonization process and severely interfere with the group’s collaboration efficiency. In addition, we lack an efficient mechanism to show the changes to participants before and after harmonization. Model change representation, an extension to the classic problem of “software code change representation” in software engineering, has been a big technical challenge for the BRIDG harmonization effort.

4.6. Consensus-based modeling in open-source communities

According to the feedback provided by the BRIDG participants, group discussions and consensus-based modeling have been helpful from the collaboration perspective. It showed each participant’s commitment to making the model generic and their respect for the needs of other communities. However, consensus-based modeling has proven technically impossible for a distributed open-source modeling community such as BRIDG without sophisticated technology support. It was impossible to organize large semantic harmonization face-to-face meetings or to achieve a consensus among this large distributed community through teleconferences or emails. Through monthly face-to-face harmonization meetings, BRIDG only achieved the “small group consensus” in its harmonization effort, by which every concept included in the shared reference model has the definitions agreed-upon by each source model contributor and the technical harmonization committee. This was a compromise. From the BRIDG experience, we can see the value of supporting collaborative modeling and negotiation as well as a substantial technology gap in achieving these when a big group of participants is involved.

4.7. Tradeoffs between being comprehensive and focused

BRIDG strived for comprehensiveness, but again, because of the large group involved with many diverse perspectives, it suffers from the challenge for focus control in its modeling process. To make the harmonization process efficient, it is important to stay focused. There are two important principles for focus control: (1) identifying the shared concepts across all domain models; (2) strictly following the top-down principle and moving from abstract to context-specific concepts. However, focus control in practice is hard. Related research questions include “How can we efficiently identify semantic connections across models?” and “How can we efficiently recognize a semantically relevant source information model?”

In the BRIDG project, many interested parties in this volunteer project wanted their source models to be harmonized into the shared conceptual reference. Interests of modeling subgroups on the BRIDG project include adverse events, study calendar, lab specimen, and best modeling practice, etc. Therefore, the domain modeling requests from various perspectives expanded fast and the complexity of the project management increased quickly. BRIDG needed to know how these requests overlap and how an individual’s interest is relevant to the final shared semantics. Overall, while focus control is critical for semantic harmonization from the project management perspective, it
involves challenges in the areas of relevance measurement and harmonization tasks prioritizing.

4.8. Group discussions and rationale capture

The goal of increased interoperability between communities will not be achieved through further formalization and abstraction. Rather, negotiation within, and especially between communities, is indispensable [37]. In the BRIDG project, modelers frequently use “sticky notes” or annotations in UML diagrams to document the sources of a concept or to make suggestions about how the model or a particular class or attribute should be harmonized. During a harmonization process, a concept could go through several statuses: to be reviewed, to be approved, to be modified, to be incorporated or discarded, etc. Valuable knowledge provenance information or harmonization design rationales were embedded in these notes and shared within the group. Since semantic harmonization involves group collaboration among multidisciplinary domain experts, rational modeling practices are crucial to support systematic modeling. Poor management of design rationales conveyed in these notes may have caused unnecessary recurring group discussions and sometimes created negative social impact on the group work, especially for those whose opinions are not adopted.

Given the above challenges and the lack of adequate technology support in the BRIDG project, we relied on human practice guidelines to regulate harmonization activities. For instance, we encouraged all modelers to develop use cases before modeling to find a common focus shared by the group. Moreover, we provided detailed instructions about handling divergence, such as how to get the latest version of the shared model, how to synchronize the local version with the evolving shared model, and so on. However, there was a lack of formal mechanisms to enforce these guidelines; therefore, variations of harmonization practices were common.

Overall, the BRIDG project did not resolve the above research challenges, such as choosing appropriate representation mechanisms for reference models, balancing being comprehensive and staying focused and supporting group consensus within the large open-source community. We also identified several technology gaps such as with model alignment and merging, handling divergence, support of consensus-modeling in distributed open-source communities, and support for knowledge provenance and design rationale capture. Such knowledge could provide implications for the design of future collaborative modeling or semantic harmonization technology. For example, in prior work, we established that annotations can effectively capture “design rationale” and support progress tracking, group discussion, and group activity coordination during collaborative writing processes [38–40]. Therefore, it is possible to develop a web-based model annotation system to support semantic harmonization online.

5. Related work

There is a wealth of related work that aims at achieving system interoperability as a general concept. An important example of such work is the semantic web [41], which aims at making web resources and services interoperable. In addition to this general approach, we also compare our work to three more specific examples of interoperability projects within medical informatics: the UMLS, the Inter-Med, and the OBI project.

5.1. Interoperability and the semantic web

The Semantic Web has a number of different aspects and definitions; here, we focus on those aspects that deal with resolving differences among the terminologies used by web resources. In particular, for two web resources to interoperate, they must share at least some terms and concepts; in our terms, the web resources must use a common reference ontology to avoid semantic mismatches [32]. As should be clear, the idea of reference ontologies for the semantic web is the same idea as the shared conceptual reference model used by BRIDG and that we espouse here for semantic harmonization.

However, there are significant differences. First, the semantic web has embraced the use of OWL as its knowledge representation language: OWL is built up from RDF, and is an expressive language that is well-suited for certain types of inference. In theory, OWL has a number of advantages over a simpler language such as UML. However, in practice the difference between models created in UML (for BRIDG) and OWL (for the semantic web) are often not so great, because much practical domain modeling work is in creating very simple concept and relationship definitions, without using the richer knowledge representation capabilities of OWL.

Rather than focusing on knowledge representational choices, we emphasize the hard work of face-to-face negotiations and the need for improved technology support for this critical step in semantic harmonization. Yet early formulations of the semantic web seemed to brush over this step and assume that reference ontologies would be easily created and that semantic differences might be resolved automatically, perhaps through classification (the inference method for OWL). However, more recent views of the semantic web have recognized the hard work involved in reference ontology development and maintenance [41].

Finally, the semantic web is usually viewed as existing across a broad range of web resources. Thus, the semantic web envisions a network of related reference ontologies, perhaps building on each other in a principled manner. In contrast, our focus is much narrower: for BRIDG, all stakeholders are interested in interoperability around systems that work with clinical trial protocols. Thus, we have proposed building a single shared conceptual reference model, rather than a network of models.
5.2. UMLS

Within healthcare, The Unified Medical Language System (UMLS) is currently the most widely used Interlingua. A core component of the UMLS is the metathesaurus, which provides a unifying paradigm that integrates machine-readable knowledge from a variety of sources including patient record systems, bibliographic databases, factual databases, expert systems, etc. The metathesaurus keeps a copy for all the knowledge sources and reserves the names, meanings, hierarchical contexts, attributes, and inter-term relationships in the sources, and also creates bridging relationships among related concepts across knowledge sources. As with the semantic web and our own work, one goal of UMLS is to provide support for connecting terms across multiple sources.

UMLS is different from BRIDG in multiple ways. First, the BRIDG model unifies various aspects of all the concepts in the clinical research domain and creates a shared generic representation for each concept, while the UMLS model retains various representations of important biomedical concepts from all sources and creates only post hoc mappings among them. Since the local usage of separately built source models is respected, there is no mechanism to ensure the cross-framework consistency. Thus UMLS terminologies have different formalisms and degrees of completeness, as well as uncoordinated updating policies.

Second, the construction of the BRIDG model relies on consensus among domain experts through discussions of deep semantics for each concept, while the UMLS model is produced by automated processing of machine-readable versions of its source vocabularies, followed by human review and editing by subject experts [21].

Third, BRIDG is a truly collaborative project in that every concept in the model is based on group consensus among all stakeholders present in harmonization meetings, while it is less clear if a true collaboration has been developed among the various groups who maintain the component vocabularies from which the UMLS is constructed [9].

Finally, UMLS accommodates dynamic links to all knowledge sources and the mappings among them can be automatically updated regularly, while BRIDG hard-codes representations of concepts from all the knowledge sources and does not provide tools that facilitate dynamic knowledge updates. Overall, BRIDG and UMLS have different philosophies: BRIDG aims at a shared generic standard for knowledge sharing through conceptual unification, while UMLS is about facilitating conceptual exchange across systems through automatic model alignment.

5.3. InterMed

Prior to BRIDG, InterMed was an important collaborative informatics initiative that developed a common model for guidelines (and to a lesser extent, clinical trial protocols) known as the Guideline Interchange Format (GLIF) [42]. Researchers on this project first hoped to make GLIF an Interlingua in the clinical guideline world, thereby supporting guideline interoperability. However, they soon discovered that true interoperation among the four major guideline representation methodologies, including Medical Logic Modules (MLM), GEODE, MBTA, and EON, was impossible due to their incompatible functionality support.

Thus, InterMed developed a generic model that “would capture a large common subset of functionality shared by the different models and that would facilitate its adaptation to local settings and its integration with other systems such as electronic medical records and order entry systems” [42].

BRIDG and InterMed are alike in many ways in that they both pursue the same mode of sharing, which is to encourage the community to adopt a shared aggregated standard. This view is different than the UMLS or the semantic web approach which both focus on linking or mapping different vocabularies or ontologies together. GLIF was also developed via a consensus-based multi-institutional collaboration and contained most of the important features that are needed in clinical guidelines [42]. In addition, the BRIDG project verified some of the findings from InterMed, e.g., (1) “standardization works most smoothly if focused on well-defined common components”; (2) “using existing standards as a starting point, while aiding in establishing credibility and consensus, does not always meet the modeling requirements”; (3) face-to-face meetings are crucial precursors to the effective use of distance communication technologies; and (4) abstraction is important for building a shared standard [42].

The BRIDG project is built on the major success factors or design principles of other projects, such as open-source collaboration, utilization of a variety of source models, and concept abstraction. However, it does not receive funding for tool development as was the case for both the UMLS and the InterMed project; therefore, it faces great challenges in continuing costly face-to-face meetings for consensus-based domain modeling that involve a large interested community. After two years of harmonization experiences, we realize that the project efficiency would have been improved had we had the resources to develop some model alignment and collaborative semantic unification tools. So far all these tasks remain tedious manual processes.

5.4. The OBO foundry and OBI project

In recent years, domain-agreed upon standards have been recognized as important for achieving semantic interoperability, especially by the ontology research community. The Open Biomedical Ontology Foundry and the Ontology for Biomedical Investigations (OBI) are two examples of community-based, domain-specific standardization efforts at the knowledge level.

The OBO Foundry [43] is a collaborative experiment, involving a group of ontology developers who have agreed in advance to the adoption of a set of principles specifying
best practices in ontology development. These principles are designed to foster interoperability of ontologies within the broader OBO framework, and also to ensure a gradual improvement of quality and formal rigor in ontologies, in ways designed to meet the increasing needs of data and information integration in the biomedical domain.

The Ontology for Biomedical Investigations (OBI) project is developing an integrated ontology for the description of biological and medical experiments and investigations by leveraging broad international research communities. The purpose of this ontology is to support the consistent annotation of biomedical investigations, regardless of the particular field of study. This project was formerly called the Functional Genomics Investigation Ontology (FuGO) project [44]. To date, the OBI project contains seventeen ontology development groups including a “clinical trials ontology” group.

These two efforts, like BRIDG, emphasize community feedback and community convergence on a single reference ontology. The OBI project is focused on metadata development for annotation purposes, while the BRIDG project is focused on developing a shared domain analysis model and supporting model-driven application development. Moreover, as we mentioned in Section 2.4, the BRIDG project does not include many source models as strictly defined ontologies, but more information models. To some degree, BRIDG strives for agreed-upon content as a first step toward interoperability, while OBI aims for both agreed-upon content and a formal representation (in ontology). Thus, the BRIDG work is complementary to these more formal ontology efforts—as future work, we could take the consensus that BRIDG achieved around content, and explore how well or easily this content could be expressed in a formal ontology, one that might be defined by an OBO or OBI-style effort.

6. Conclusion

There is little prior empirical knowledge about semantic harmonization. In this paper, we bridge this knowledge gap by reflecting on our experiences with the BRIDG project and summarize some unresolved challenges for user-centered semantic harmonization processes.

At present, domain modeling is convenient for almost any organization, but there is little support for community-based modeling that effectively utilizes existing domain knowledge resources. BRIDG made a first step to explore this open ground. The BRIDG project successfully established collaboration across different clinical research communities in academia, industry, government, and organized them to work together toward a shared ambitious goal. Instead of building yet another new standard in an isolated research mode, BRIDG explored an open-source approach for supporting community-based domain analysis and knowledge synthesis and contributed first-hand experience about this methodology, its inherent challenges, and potential pitfalls of modeling practice in building a domain reference model.

With the soon-to-be completed pilot phase of the caBIG program, it is still too early to see the fruits of this modeling effort, but we want to bring the attention of our research community to a few technology gaps in support of collaborative modeling and domain knowledge synthesis. As we describe in Section 4, some of these gaps include a need for better modeling tools (4.2), methods for managing knowledge provenance (4.3), technologies for model alignment (4.4), tools for divergence handling and model change representation (4.7), technologies for very large group collaborations (4.6), and tools for annotation and design rationale capture (4.8). Semantic harmonization is a challenging research area. Several intertwined unresolved research issues from software engineering, computer-supported collaborative work, particularly collaborative modeling, and semantic knowledge representation and engineering all come into play. For collaborative domain-consensus modeling to scale, future technology support will be indispensable and should be urgently investigated.

Acknowledgments

This work has been funded through the caBIG™ project by the National Cancer Institute. The authors express their great appreciation to the two reviewers for their constructive and insightful review comments. The manuscript has benefited tremendously from the reviews. The authors also thank Roger Day, Heather Piwowar, and Diane Paul for their helpful comments.

References


[34] OWL. Available from: <http://www.w3.org/TR/owl-features/>.


