Author's response to reviews

Title: The CAP cancer protocols - A case study of caCORE based data standards implementation to integrate with the Cancer Biomedical Informatics Grid

Authors:

Jonathan Tobias (jptobias@mednet.ucla.edu)
Ram Chilukuri (ram.chilukuri@semanticbits.com)
George A Komatsoulis (komatsog@mail.nih.gov)
Sambit Mohanty (mohantys2@upmc.edu)
Nicholas Sioutos (nsioutos@mail.nih.gov)
Denise B Warzel (warzel@mail.nih.gov)
Lawrence W Wright (lwright@mail.nih.gov)
Rebecca S Crowley (crowleysrs@upmc.edu)

Version: 5 Date: 7 May 2006

Author's response to reviews: see over
Dear Editor:

Thank you for providing us with the chance to respond to the excellent suggestions made by the three reviewers to this manuscript. Below we enumerate our responses and describe changes to the manuscript on a point-by-point basis.

Review 1. Dr. Simon Lin

**Suggestion 1: (minor essential)** “Page 26. CancerTNMFining was mention in the first paragraph and referred to Figure 3. However, I cannot find it in Figure 3”

Response to suggestion 1: The reference refers to a previous version of the figure and has been removed.

**Suggestion 2: (discretionary)** “It seems to me more appropriate to model the enumeration type (in figure 4) of Present/Absent more general into the thesaurus, instead of tying them to CutaneousMelanoma_Ulceration. However, it is only a modeling preference, and the authors have extensively discussed this issue in the paper.”

Response to suggestion 2: In fact the values Present and Absent are tied to the Thesaurus through their vocabulary bindings. The same value set for Present/Absent within the caDSR is reused throughout the CAP Protocol Model CDEs. Binding to the NCIT for both values is depicted in Figure 4 (upper right corner). However reuse of value sets within the caDSR is difficult depict in this already busy figure. Therefore we have elected to leave Figure 4 unchanged.

Reviewer 3: Dr. Samson Tu

**Suggestion 3: (discretionary)** “The labeling of and references to figures leave a lot to be desired.”

Response to suggestion 3. We have corrected numerous inconsistencies throughout the manuscript that represent artifacts of previous figure labeling.

**Suggestion 4: (discretionary)** “In Abstract/Methods: Define what is caCORE.”

Response to suggestion 4: We have defined the caCORE acronym in the Abstract and on page 10 in the methods and also added a one sentence description of caCORE on page 10.

**Suggestion 5: (discretionary)** “In Abstract/Result and discussion: "Using this example, we describe the process required to develop the model, the technologies and data standards on which they are based,... Clarify what "they" refers to.”

Response to suggestion 5: In the abstract (page 2) we changed “they” to “process and models”.
**Suggestion 6: (discretionary)** “p. 10 API == Application Programming Interface”

Response to suggestion 6: Corrected “programmer” to “programming”

**Suggestion 7: (discretionary)** “p. 11. Layer II - information mModels ?= information model”

Response to suggestion 7: Corrected “mModels” to “models”

**Suggestion 8: (discretionary)** “p. 21 The description of the use of 'roles' in concept definition does not capture the essence of *description logic*, where a definition gives *necessary and sufficient* conditions for an individual to be an instance of a class. Such definitions allow individuals to be classified as instances of appropriate classes and subsumption relationships to be inferred. They are not about distinguishing a concept from sibling concepts with the same parents.”

Response to suggestion 7: Our description attempts to accurately portray, in a way that will be readily understood by a general biomedical informatics audience, our particular use of description logic, which is unusual in two respects:

- All NCI Thesaurus concepts are abstract classes; it contains no instances. This is a perfectly valid use of DL, and works well for terminology services, but needs to be thought about and described somewhat differently.
- The Ontylog DL system we have used to date does not support distinguishing between "necessary" and "necessary and sufficient" conditions. This is supported in Protégé, and as part of our intended migration to Protégé we will be making this distinction, but it seemed a potential misleading or confusing digression to address that issue here.

We hope readers interested in such topics will follow the reference provided, which is focused on the DL aspect of NCI Thesaurus and the ways this use of DL differs from more traditional ones. We are reluctant to enter into these issues here, as they would be more likely to confuse than assist our readers.

**Suggestion 9: (discretionary)** “p. 25 "Alternatively, when concepts were deemed inappropriate for entry into the NCI Thesaurus, the concept was postcoordinated in the metadata using the object class or property and its qualifiers for data element concepts, or multiple value meaning codes for valid values.” Please explain further and give examples.”

Response to suggestion 9: We have added further explanation, and an example tied to figure 4 on pages 25-26.
**Suggestion 10: (discretionary)** “p. 18 "In caDSR, the use of concepts from controlled vocabularies is mandatory for object class, property, qualifiers and representation terms and optional for value meanings." Explain what's "representation terms."

Response to suggestion 10: We have added a definition of representation terms to the description of the ISO/IEC metamodel on page 17, so that the definition now precedes the reference to representation terms on page 18.

**Suggestion 11: (discretionary)** “p. 26. The alleged Figure 3 showing CancerTNMFinding seems to be missing.”

Response to suggestion 11: The reference refers to a previous version of the figure and has been removed.

**Suggestion 12: (discretionary)** “p. 27 The alleged Figure 3 showing subclasses of CutaneousMelanoma seems to be missing.”

Response to suggestion 12: The reference is incorrect and has been corrected to Figure 4. Additionally we edited the sentence associated with the figure reference to make it clear that we are referring to relationships in the ontology that could be used to determine valid aggregations of the hypothetical model being discussed.

**Suggestion 13: (discretionary)** “p. 27-28 "The relationship between BreastSurgicalPathologySpecimen and SurgicalPathologySpecimen exists only in the metadata" What semantics does the metadata post-coordination have? Without a logical model, what meaning can one ascribe to such post-coordination. The same problem pertains to discussion on pp. 34-36.”

Response to suggestion 13: The semantics of the metadata are simply the semantics of ISO/IEC 11179. Specifically post-coordination in the metadata currently takes the form of ordinal lists in which the object class or property is annotated with a primary concept that conveys the most essential meaning relevant to the real world object or characteristic, and the remaining semantic content is considered as modifying this essential concept. Remaining concepts become property qualifiers or object class qualifiers. We have added a sentence on page 28 indicating that this is the main constraint to the semantics. The resulting meaning is truly implicit and based largely on linguistic characteristics. This issue is addressed extensively in the discussion.

**Suggestion 14: (discretionary)** “p. 13 Layer IV - controlled vocabularies and ontologies I am surprised that, in the methodology, the use of controlled vocabularies does not appear to extend to allowed values of attributes (value sets). The discussion section talks about future extension where permissible values to have the same semantic annotation as classes and attributes. It seems to me that without such controlled vocabulary for permissible values, semantic interoperability is not possible. Perhaps I am missing something. If so, the authors should explain more clearly.”
Response to suggestion 14: The reviewer is not correct. The methodology does allow binding of permissible values to the formal semantics of the description logic. However, at the time that this paper was written, the tooling did not easily support annotation of permissible values, and thus the process was manual as explained in the manuscript. We have modified the manuscript to clarify this issue:

**Suggestion 15: (discretionary)** “The paper can be significantly enhanced by a comparison of the approaches to semantic interoperability taken by caBIG and by HL7 version 3. HL7 derive its claim on interoperability through the use of standardized datatypes, shared reference information model, and binding of permissible attribute values to controlled terminologies. The information model describes here appears to be radically different from that of HL7. The data elements appear to correspond to HL7 Observations, where data element concept corresponds to the "code" attribute of Observation. In HL7, the code attribute have values taken from controlled terminologies which may use description logic to post-coordinate concepts. It seems that this framework does not have the drawbacks identified in this paper (post-coordination of metadata, arbitrary distinction between property and qualifier, and lack of structure in qualifiers). Modeling problems are still there, but at least the modeling is done in a terminological language whose compositional rules are better defined (e.g., use of DL, as is done in NCI thesaurus). Perhaps the semantic annotation and HL7 approaches are complementary in the sense that the annotations create mappings between a given data schema and a canonical schema (and creating new data element concepts as necessary), whereas HL7 RIM is very much a top-down centrally harmonized methodology.

Response to suggestion 15: The authors agree that the relationship of the described methodology provides an interesting contrast to HL7 RIM and that both methods may be complimentary. However the comparison of this approach to HL7 RIM is almost certainly deserving of an entire manuscript itself. Other modeling efforts within caBIG (e.g. BRIDG) have used elements of HL7 RIM modeling practices, and take a more top-down approach, but still employ the annotation scheme developed by caBIG. Thus we prefer to defer this discussion to future manuscripts where the comparison can be made within the context of a completed project.

**Suggestion 16: (discretionary)** “This reviewer is uncomfortable with the idea that the metadata semantic annotation may express relationships that contradicts the source ontology. Sure researchers need to have concepts and relationships that inconsistent with the source ontology. However, a better approach seems to be a methodology of subsetting and extending the source ontology, rather than living (dangerously) with contradictions which may invalidate your inferences.”

Response to suggestion 16: The authors share the view that there is potential danger in permitting the semantic metadata to potentially conflict with the ontology. However, the constraints of almost all data modeling are such that this conflict is always inevitable. Data models are by their very nature contextualized, and therefore often extend the formal semantics in often unpredictable ways. Although the central
central canonical ontology can be modified to exist as a set of locally extended ontologies – the issue of conflict both between the extensions and also between the ontology and the metadata remains. The flexibility provided by allowing the metadata to express contextualized relationships is of such practical importance, that we believe that the dangers of this approach are worth enduring. What is needed to ameliorate these dangers is a methodology for uncovering such inconsistencies between metadata and ontology in order to appropriately harmonize the models and the ontology. We have elected not to further expand this section because we believe this topic is adequately covered in our discussion.

Reviewer 2: Dr. John Sinard

Suggestion 17: (major compulsory) “My biggest problem with the manuscript, and honestly why it has taken me so long to review it, is that I have read it multiple times, trying to understand what the “message” is. This manuscript does not propose any hypothesis which is subsequently tested or evaluated, but rather describes the arduous process of adapting a CAP cancer protocol into a caBIG framework. Although separate Methods, Results, and Discussion sections are provided, most of the quite lengthy manuscript is actually methods. This is not necessarily a bad thing in the Informatics domain, since “process” constitutes a lot of what informatics is, but it does leave me wondering why one is bothering to do this if it is so difficult unless it provides some advantages. Possible uses and advantages to this approach are suggested in the discussion, but actual use of the resulting data model is never described, nor is there any data supporting the claim that this method of modeling the protocols has any real advantages over the existing protocols, or some other data-element persistent implementation which is not burdened by the requirement that every component be an element in the NCI Thesaurus.”

Response to suggestion 17: We appreciate the concerns of the reviewer. With regard to the message, we have added a sentence to page 9 to further drive home the point that the method is intended to support semantic interoperability. This message is tightly woven into almost all section of this paper. As the reviewer suggests the paper is indeed intended to be read as a Methods paper describing a new technical methodology of significant importance to the development of semantic interoperability in biomedicine. The intent of the paper is to provide a case study that can be studied by others who choose to use the methodology to implement data standards. This purpose was reinforced by Reviewer 1 in his review of the manuscript. Because this is a methods paper – traditional quantitative results and analysis are simply not applicable. If the confusion stems from the required BMC headings (Methods, Results, Discussion) – the authors would be happy to alter these headings to more appropriate ones provided by the editor or reviewer.

Suggestion 18: (major compulsory) “In addition, the manuscript seems to have had some trouble defining its audience. The introduction is well written and introduces the caBIG project, and some of its components, to the uninitiated, making it appropriate for even non-informatics people. The manuscript then dives pretty deep into ontologies,
vocabularies, and modeling, and is sufficiently esoteric to be appreciated only by a rather different audience. Personally, I feel that this work would benefit from selecting one of the two audiences. If the former, more diagrams and a toned down description of the process using vernacular terminology would be better received. If the latter, then much of the introductory material could be eliminated (and I would definitely not be an appropriate reviewer!)

Response to suggestion 18: The intent of the paper is clearly spelled out in the Introduction and Background. The intended audience is informaticians – as opposed to non-informatics readers. The technical detail presented is essential in a case study, because others who wish to use this methodology must be able to reproduce it. We present background on caBIG and the domain because it is important to understanding the context in which the modeling was performed. For example, it is important to describe the social structures of caBIG so that readers will understand that a cross-cutting workspace (Vocabularies and CDEs) is working with a domain workspace (Tissue Banking and Pathology Tools) on this project.

**Suggestion 19: (major compulsory) “caCORE is never really defined”**

Response to suggestion 19: We have defined the caCORE acronym in the Abstract and on page 10 in the methods and also added a one sentence description of caCORE on page 10.

**Suggestion 20: (major compulsory) “The legend for Figure 2 seems to refer to what is actually figure 3, and the legend for figure 3 seems to refer to what is in figure 2.”**

Response to suggestion 20: The figure legends and all references have been corrected.

**Suggestion 21: (major compulsory) “The text in page 26 refers to figure 3 two times, but the first reference seems to be talking about something which is not in figure 3 at all, or any other figure – most likely a figure which has been subsequently deleted. Same for the reference on page 30.”**

Response to suggestion 21: Because the figure legends have been switched for Figures 2 and 3, the figure references in the text are now correct and require no further revision.

**Suggestion 22: (major compulsory) “Finally, figure 3, as well as figure 2, is labeled “Fig 2”.”**

Response to suggestion 22: The figures have been corrected.

**Suggestion 23: (major compulsory) “Page 11: Layer II – information mModels”**

Response to suggestion 23: Corrected “mModels” to “models”
**Suggestion 24: (major compulsory)** Page 12 “…it has NSC Number is ‘007’”

Response to suggestion 24: Corrected by removing word “is”

**Suggestion 25: (major compulsory)** page 42 “…Cerner/DHTI to …”

Response to suggestion 25: Modified to “Cerner DHTI”

**Suggestion 26: (major compulsory)** Page 28 refers to five CDEs being created for CutaneousMelanomaNegativeSurgicalMargin, but only two are listed in the figure.

Response to suggestion 26: The text at issue is as follows:

“For example, CutaneousMelanomaNegativeSurgicalMargin inherits three attributes from the general class SurgicalMargin, and extends this class with two additional attributes (Figure 3). Three CDEs will be created for Surgical Margin, and five will be created for CutaneousMelanomaNegativeSurgicalMargin.”

The text is correct as it stands. As indicated, the additional two attributes extend the parent class with three attributes. Thus the child class will generate 5 CDE’s – three of which are derived from inheritance relationships. That is why only two attributes are indicated in the figure.