BIM: An Open Ontology for the Annotation of Biomedical Images

*Ahmad C. Bukhari[1], Mate Levente Nagy[2], Michael Krauthammer[2], Paolo Ciccarese[3], *Christopher J. O. Baker[1]

[1] Dept. Computer Science & Applied Statistics, University of New Brunswick, Saint John, NB, E2L 4L5, Canada

[2] Dept. Pathology & Yale Center for Medical Informatics, 300 Cedar Street, New Haven, CT 06510, USA

[3] Harvard Medical School, 25 Shattuck Street, Boston, MA 02115 USA

ABSTRACT

Biomedical images published within the scientific literature play a central role in reporting and facilitating life science discoveries. Existing ontologies and vocabularies describing biomedical images, particularly sequence images, do not provide sufficient semantic representation for image annotations generated automatically and/or semi-automatically. We present an open ontology for the annotation of biomedical images (BIM) scripted in OWL/RDF. The BIM ontology provides semantic vocabularies to describe the manually curated image annotations as well as annotations generated by online bioinformatics services using content extracted from images by the Semantic Enrichment of Biomedical Images (SEBI) system. The BIM ontology is represented in three parts; (i) image vocabularies - which holds vocabularies for the annotation of an image and/or region of interests (ROI) inside an image, as well as vocabularies to represent the pre and post processing states of an image, (ii) text entities - covers annotations from the text that are associated with an image (e.g. image captions) and provides semantic representation for NLP algorithm outputs, (iii) a provenance model - that contributes towards the maintenance of annotation versioning. To illustrate the BIM ontology's utility, we provide three annotation cases generated by BIM in conjunction with the SEBI image annotation engine.

1 INTRODUCTION

Images depicting key findings of research papers contain rich sets of information derived from a wide range of biomedical experiments. Biomedical imaging [1] employs numerous modalities such as X-Rays (CT scans), sound (ultrasound), magnetism (MRI), radioactive pharmaceuticals (nuclear medicine: SPECT, PET) or light (endoscopy, OCT) to evaluate the status of an organ or tissue. Unlike text or other non-imaging data, image data poses a number of idiosyncratic issues rendering them mainly opaque for reuse without significant manual intervention. Current practices related to the extraction of implicit knowledge provide annotations that are neither anchored with an image, nor documented in a machine-readable fashion. As a consequence images cannot be readily discovered or categorized based on their contents. In the case of biomedical images that contain some type of biological sequence data summarizing the atomic composition of biological molecules [2] a combination of optical character recognition and text extraction techniques can provide better searchability over these images such that questions like "*display of all the sequence images that show proteins from the same protein family*"- [3] could be asked, provided that annotations could be made available to a search or query engine. However, image repositories in use today restrict the features that users can search with to those described in text based image captions and predominantly encourage the syntactic keyword based search, which constitutes a significant limitation [4]. In contrast images with semantic annotation can be automatically and/or semi-automatically discovered and linked to new information. The resulting enriched images are readily reusable based on their semantic annotations and can be used in semantic search and ad-hoc data integration activities. Overall, to achieve a greater degree of reusability and interoperability over image data certain core infrastructure is required, including automated image annotation pipelines and semantic vocabularies that can anticipate and represent image related content unambiguously. Existing ontologies and vocabularies describing biomedical images, particularly sequence images, are not sufficient to fulfill the requirements mentioned above and for our use case (SEBI) [4]. This motivated us to build the BIM ontology described in this paper which was designed and modeled with the following purposes in mind: formal representation of image annotation, enhanced reusability of image related data, depiction of pre and post image processing phases, design of context aware image search engines and semantics enabled bioimaging applications.

2 THE BIM ONTOLOGY

To better understand the context where BIM is relevant we briefly describe SEBI (semantic enrichment of biomedical images). SEBI is a solution for image annotation

* To whom correspondence should be addressed: bakerc@unb.ca

that adopts a combination of technologies to comprehensively capture information associated with, and contained in, biomedical images. To achieve this SEBI utilizes information extracted from images as seed data to aggregate and harvest new annotations from heterogeneous online biomedical resources. SEBI incorporates a variety of knowledge infrastructure components and services including image feature extraction [5], semantic web data services [6], linked open data [7] and crowd annotation [8]. Together these resources make it possible to automatically and/or semi-automatically discover and semantically interlink the new information in a way that supports semantic search for images. The resulting enriched images are readily reusable based on their semantic annotations and can be used in ad-hoc data integration activities. To date the BIM ontology has been used to successfully annotate 15000 images from the Yale Image Finder [3], 85% automatically and 15% through manual crowdsourcing.

3 MATERIALS AND METHODS

BIM ontology has been created to provide the standardized semantic representation of the annotations generated to describe a biomedical image by SEBI. BIM can further be used for annotating the associated text references by a machine or human. In order to collect the relevant terms, relationships / properties for sequence related images, we reviewed literature mentioning sequence analysis algorithms [9] such as BLAST, HMMER, Prosite, and the conserved domain database. A total number of 23 papers published from 2006 to 2015 were selected from different journals. We focused on actual depictions and discussion of sequence alignment outputs, rather than the algorithms, to distill the typical terms, concept and relations used. In order to accumulate terminologies associated with non-sequence image types such as: X-Rays, ultrasound, MRI, radioactive pharmaceuticals endoscopy, we selected a random sample set of papers from the Journal of *Bioimaging* and applied the *SNOMED-CT*[1] *and BioNLP web services* [10] to expedite the knowledge elicitation process. The *SNOMED-CT* and NLP web services provided the exact annotation location (e.g. start and stop annotation word) wherever a term existed in the paper. Manual evaluation of the outputs extracted from papers was performed, whenever relevant terms were found they were categorized and documented. While modeling the BIM ontology, a number of ontologies relating to annotation and biomedical imaging were also consulted and where appropriate, classes and properties were reused.

Table 1 depicts the ontologies, prefixes and namespaces of the existing ontologies that have been employed in the modeling of BIM ontology. We have reused the vocabularies defined in Annotation Ontology (AO) [11] to model the biological concepts mentioned in an image caption. AO is an open-source ontology for annotating the scientific documents on the web. In AO, all the annotations are regarded as resources and fall under the instance category of the *Annotation class*. Each annotation has some *hasTopic*, context predicates and object class. Objects can be a particular entity such as protein or chemical name, a disease, or reified fact, while the context refers to a certain text segment inside the sentence (see Fig.3). This simple reference model makes it possible to integrate the extracted information semantically. The provenance of annotations is modeled with Provenance, Authoring and Versioning (PAV) ontology [12] e.g. predicates such as *createdBy*, *createdOn* describe the annotation creator and date of creation. PAV provides the terminologies for tracing provenance of the digital entities that have been published on the web and then accessed, transformed and consumed. To cover high-level scientific research concepts, terms from the *Semanticscience Integrated Ontology* (SIO) were imported [13]. SIO provides a simple, integrated ontology of types and relations to describe objects, processes and their attributes. SIO behaves as an upper level ontology and supplies many high-level biomedical concepts. To represent the structural information of a biological sequence semantically, we incorporated a number of classes and relationships from Sequence Ontology (SO) [14] ontology such as *transcript, primary-transcript, intron*, *mRNA, insertion sequence.*

Table 1. Well-known vocabularies utilized in BIM modeling

Ontology/Vocabulary	*Prefix*	*Namespace*
Annotation Ontology	AO	http://purl.org/ao/
Provenance Authoring & Versioning Ontology	PAV	http://purl.org/pav/
SemanticScience Integrated Ontology	SIO	http://semanticscience.org/ontology/sio.owl
Sequence Ontology	SO	http://purl.obolibrary.org/obo/so.owl
Friend Of A Friend	FOAF	http://xmlns.com/foaf/0.1/
SIOC Ontology	SIOC	http://rdfs.org/sioc/ns#
SKOS ontology	SKOS	http://www.w3.org/2004/02/skos/core
Exif Ontology	exif	http://www.kanzaki.com/ns/exif#
Time Ontology	TIME	http://www.w3.org/TR/owl-time/
Semantic DICOM Ontology	DICOM	http://purl.bioontology.org/ontology/SEDI
DBpedia Ontology	DBpedia	http://dbpedia.org/ontology/

The Exif[2] ontology [15] mainly describes the Exif format of picture data semantically, and provides useful vocabularies supporting the pre-processing and usage of Exif images. In BIM ontology, we used the Exif terminologies to define image orientation and size using *Ex-*

if:Orientation, Exif:ImageWidth, Exif:ImageHeight and corresponding vocabularies to represent the stages of image processing e.g *Exif:WhiteBalance.* DICOM (Digital Imaging and Communications in Medicine) [16] is a standard to represent the medical image information worldwide. Most of the available medical images modalities follow the DICOM standards to capture, store and disseminate the medical image information. However, the DO (DICOM Ontology) [17] serves the purpose of integrating and explicitly representing the concepts and relationships of DICOM in machine readable and human understandable format. In BIM ontology, we imported DO classes to represent the information associated with radiology images and to represent image capturing detail semantically. The FOAF [18] vocabulary describes people, their relations with other people, and objects that are related to a person-to-person connection.

We also leveraged the *DBpedia* ontology [19], a multi-domain ontology that is mainly designed to cover the Wikipedia infoboxes. In version 3.2, there are roughly 359 classes and 1775 properties, which cover a vast range of common and life science concepts. In contrast, the Dublin core Metadata [20] vocabulary was used to represent general meta-data attributes for documents such as titles, authors, subjects, descriptions, date, type, and format. Core concepts from time and relationship ontologies were imported to describe concepts relating to time units (e.g. minutes, seconds) and relations between objects. The Semantically-Interlinked Online Communities (SIOC pronounced as "shock") [21] is a domain ontology, which perfectly defines and interlinks all the online communities' concepts such as posts, comments, and users. Similarly, the Simple Knowledge Organization System (SKOS) [22] is a generalized model written in RDF for sharing and interlinking organizational knowledge with semantic description. We reused the terms *SKOS:prefLabel, SKOS:Concept, SIOC:Item and SIOC:userAccount* from SIOC and SKOS ontologies. To assemble the BIM ontology model, we used the *Protégé*, editor [23]. However, to efficiently manage and utilize the BIM vocabularies, an ontology-publishing server called *UNBvps (http://cbakerlab.unbsj.ca/unbvps/)* was set up.

The server provided a range of control functions, including management of provenance, versioning of the source vocabularies, and delete/update functions. We enhanced the Neologism plugin [24] on our server to reduce the time spent developing and publishing vocabularies with conventional ontology authoring techniques i.e. using *Protege* and internet publishing. To identify the appropriate semantic mappings between existing ontologies and BIM ontology, a Java program that suggests the possible mappings was created. The program extracted the tables and column names, storing them as variables and invoked a *WordNet*[3] web service that lexically compared each variable with the ontology entities to find possible matches. The overall goal was to provide candidate matches for subsequent curation; a comprehensive benchmarking of the algorithm's performance was not derived. A cursory evaluation of the derived mappings showed three types of results; (i) mappings that fully met our requirements, which suggested predicates such as *hasPubMedID* and *hasPMCID* in the *FRBR-aligned Bibliographic Ontology* [25] *(FaBio)*; (ii) mappings that were insufficiently defined, like the image Feature property that existed in BioPortal; and (iii) mappings with hosted resources that did not appear trustworthy.

4. USE CASES

This section demonstrates the BIM ontology modeling with three different use cases.

4.1 Use case 1: Automatic sequence image annotation

To perform enrichment of a biological sequence image with semantic annotations, a cluster of SADI web services [26] was developed. When the SEBI platform sends a request to semantically annotate an image, a number of web services are invoked serially. The image extraction and analysis service takes the image and applies the image processing filters to improve the image contrast and to improve the image resolution. Subsequently, the OCR extraction web service receives a processed image and applies an algorithm to extract the optical characters from the image. BIM ontology supplies the necessary vocabularies to express the pre and post image processing stages such as: *BIM:hasImageResolution* and *BIM:ImageFilters* used to semantically represent features that have been used to process an image. Subsequently the OCR extraction web service pulls out the sequence (optical characters) from an image while BIM ontology represents that sequence string as *BIM: SequenceBlock.* Later the extracted sequence string has been passed to the sequence analysis web services to generate annotations on a sequence image. The SADI sequence analysis service module has been designed to retrieve annotations for biological sequences from various biological sequence analysis tools such as *HMMER, BLAST, Pfam, ProSite, and GO*. Fig. 1 displays the semantic modeling provided by the BIM ontology to enrich a sequence image with semantic annotations. The annotations harvested by the sequence analysis services (by exposing sequence analysis software as web services) provide useful information about a sequence image. The newly generated annotation further underpins the image similarity module of SEBI that accurately fetches the relevant/similar sequence images from the scientific literature. To preserve the provenance of an image and annotations curated on an image, BIM ontology reuses the vocabularies provided by the PAV ontology as displayed in Fig.1. The terms such as: *pav:createdBy* and *pav:createdOn* have been recruited to represent the web service and the annotation creation date respectively. However, the terms such as

BIM:hasSequenceType, BIM:hasMutationResidue, BIM:hasConservedResidue, BIM:hasMOTIF, BIM:hasProteinInteractionSite explicitly define the outputs of sequence analysis software. All terms relating to sequence analysis have been defined for the first-time in BIM ontology, as we did not find their accurate representation in any ontologies available online. Additionally, we can utilize *time:Instant* to capture the hours, minutes and seconds for *createdOn*.

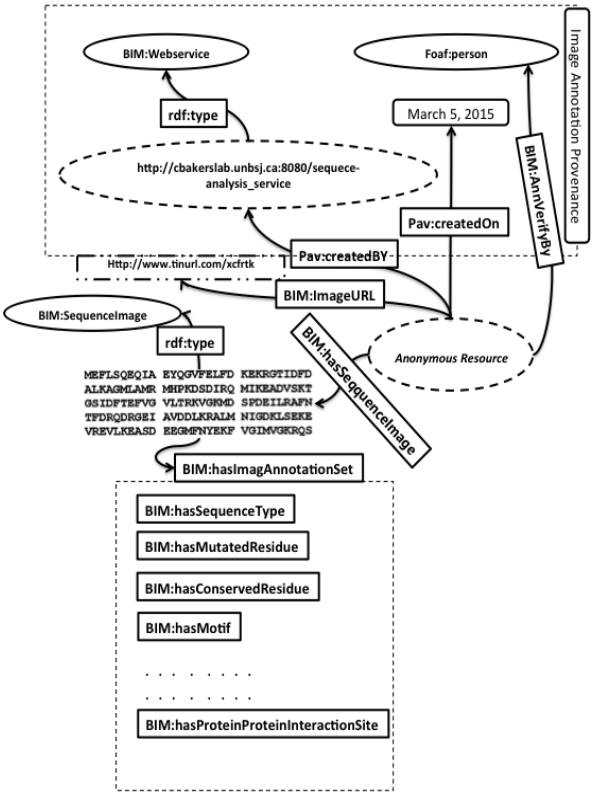

Fig. 1 BIM Model of Automatic Sequence Annotation by a Web Service

4.2 Use Case 2: Crowd-based semi automatic annotation

Semi-automatic annotation, where automatic annotation is not feasible due to poor quality input images, is made possible through the introduction of a crowd annotation technique. All images that fail to produce new annotations through web services are forwarded to the crowd annotation module of SEBI. Salient features of the crowd annotation module are as follows: Users can annotate, delete, or update annotations, maintain private annotations or share them with other legitimate users. BIM provides vocabularies through which a user can maintain image provenance, for instance it documents the author (human or machine) that has curated an annotation and the location *(xy-coordinates)* inside an image. Moreover, the crowd annotation module provides a utility through which a user can select and annotate a portion within an image. To support such activities BIM ontology supplies the crowd annotation module with *BIM:CTScan*, *BIM: hasSomeLesion*, and *BIM: polygonCoordinates* to semantically express the intra image annotation and the position of the annotation inside an image. *BIM: Resolution* class has further subclasses in *BIM: Width sameAS Exif:ImageWidth* and *BIM: Height sameAS Exif:Imageheight.* The *BIM:AnnotationRevision* class facilitates a user to track the legacy annotation made on an image along with information on the creator/software agent.

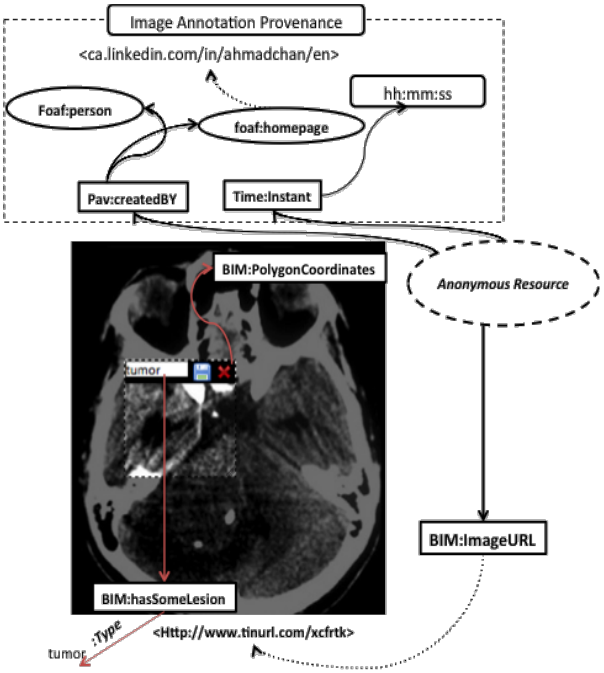

Fig. 2 BIM Crowd-Sourced Modeling of a Biomedical Image

4.3 Use Case 3: Text associated with an image

In SEBI, the BioNLP annotation module extracts named entities, such as drug names, diseases, chemicals, proteins, lipids or GO terms found in the captions or in the descriptions of a biomedical image in a paper. The BioNLP annotation module further normalizes the entities to canonical names defined in online resources e.g. PDB and DrugBank and publishes them in RDF to annotate the images. The BIM ontology incorporates the Annotation Ontology and PAV ontology vocabularies to semantically annotate the concepts and relationships. Fig. 3 explains the BIM ontology modeling on the caption of an image where a drug is

[1]http://ihtsdo.org/snomed-ct/
[2]http://www.kanzaki.com/ns/exif
[3]http://wordnet.princeton.edu/
[4]http://www.rcsb.org/pdb/home/home.do
[5]http://www.drugbank.ca/

mentioned.

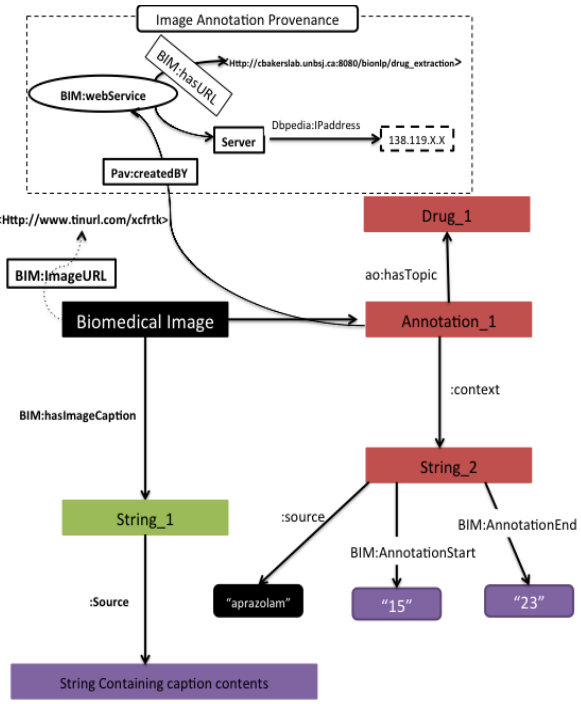

Fig. 3 BIM Ontology Modeling for Image Associated Text

CONCLUSIONS

This paper introduces Biomedical Image ontology (BIM) that supports the publication of annotations on, features identified within a biomedical image generated by SEBI tools. BIM was created to address the dearth of appropriate ontologies and appropriately integrated semantic metadata targeted to annotating diverse biomedical images, particularly images depicting biological sequences. BIM supports both the creation of machine generated and human curated annotations which can be reused in multiple knowledge discovery tasks or resources. These include; image mashups, linked image data, semantic image search and the computing of image similarity, which along with provenance annotations indicating an image's source publication permits the linking of publications containing related images. The SEBI framework is designed to facilitate all these goals.

REFERENCES

1. Perkel, J. M. (2013). *Life Science Technologies: Mass Spec Imaging: From Bench to Bedside*. Science, 340(6136), 1119-1121
2. Webb, A., & Kagadis, G. C. (2003). *Introduction to biomedical imaging*. Hoboken: Wiley.
3. Xu, S., McCusker, J., & Krauthammer, M. (2008). *Yale Image Finder (YIF): a new search engine for retrieving biomedical images.* Bioinformatics, 24(17), 1968-1970.
4. Bukhari, A. C., Krauthammer, M., & Baker, C. J. SEBI: *An Architecture for Biomedical Image Discovery, Interoperability and Reusability based on Semantic Enrichment.*
5. Corke, P. (2011). *Image Feature Extraction*. In Robotics, Vision and Control (pp. 335-379). Springer Berlin Heidelberg.
6. Fensel, *D., Facca, F. M., Simperl, E., & Toma, I. (2011). Semantic web services. Springer* Science & Business Media.
7. Bizer, C., Heath, T., & Berners-Lee, T. (2009). *Linked data-the story so far.*
8. Dijkshoorn, C., Oosterman, J., Aroyo, L., & Houben, G. J. (2012, July). *Personalization in crowd-driven annotation for cultural heritage collections.* In4th International Workshop on Personalized Access to Cultural Heritage PATCH 2012, Montreal, Canada, July 16-20, 2012.
9. Li, H., & Homer, N. (2010). *A survey of sequence alignment algorithms for next-generation sequencing.* Briefings in bioinformatics, 11(5), 473-483.
10. Bukhari, A. C., Klein, A., & Baker, C. J. (2013, January). *Towards Interoperable BioNLP Semantic Web Services Using the SADI Framework.* In Data Integration in the Life Sciences (pp. 69-80). Springer Berlin Heidelberg.
11. Ciccarese, P., Ocana, M., Garcia-Castro, L. J., Das, S., & Clark, T. (2011). *An open annotation ontology for science on web 3.0. J. Biomedical Semantics*,2(S-2), S4.
12. Ciccarese, P., Soiland-Reyes, S., Belhajjame, K., Gray, A. J., Goble, C. A., & Clark, T. (2013). *PAV ontology: provenance, authoring and versioning. J. Biomedical Semantics,* 4, 37.
13. Dumontier, M., Baker, C. J., Baran, J., Callahan, A., Chepelev, L. L., Cruz-Toledo, J., ... & Hoehndorf, R. (2014). *The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery*. J. Biomedical Semantics, 5, 14.
14. Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., & Ashburner, M. (2005). *The Sequence Ontology: a tool for the unification of genome annotations*. Genome biology, 6(5), R44.
15. Alvarez, P. (2004). *Using extended file information (EXIF) file headers in digital evidence analysis*. International Journal of Digital Evidence, 2(3), 1-5.
16. Mildenberger, P., Eichelberg, M., & Martin, E. (2002). *Introduction to the DICOM standard. European radiology*, 12(4), 920-927.
17. Kahn Jr, C. E., Langlotz, C. P., Channin, D. S., & Rubin, D. L. (2011). *Informatics in Radiology: An Information Model of the DICOM Standard 1.*Radiographics, 31(1), 295-304.
18. Golbeck, J., & Rothstein, M. (2008, July). *Linking Social Networks on the Web with FOAF: A Semantic Web Case Study*. In AAAI (Vol. 8, pp. 1138-1143).
19. Töpper, G., Knuth, M., & Sack, H. (2012, September). *Dbpedia ontology enrichment for inconsistency detection.* In Proceedings of the 8th International Conference on Semantic Systems (pp. 33-40). ACM.
20. Dublin Core Metadata Initiative. (2012). Dublin core metadata element set, version 1.1.
21. Breslin, J. G., Decker, S., Harth, A., & Bojars, U. (2006). *SIOC: an approach to connect web-based communities*. International Journal of Web Based Communities, 2(2), 133-142.
22. Miles, A., & Pérez-Agüera, J. R. (2007). *Skos: Simple knowledge organisation for the web. Cataloging & Classification Quarterly*, 43(3-4), 69-83.
23. Rubin, D. L., Noy, N. F., & Musen, M. A. (2007). *Protege: a tool for managing and using terminology in radiology applications*. Journal of Digital Imaging, 20(1), 34-46.
24. Basca, Cosmin et al. (2008). Neologism: Easy Vocabulary Publishing.
25. Shotton, D., & Peroni, S. (2011). *FaBiO: FRBR Aligned Bibliographic Ontology*.
26. Wilkinson, M. D., Vandervalk, B. P., & McCarthy, E. L. (2011). *The Semantic Automated Discovery and Integration (SADI) Web service Design-Pattern, API and Reference Implementation*. J. Biomedical Semantics, 2(8).

Availability: The BIM ontology, version 1.0 is scripted in OWL/RDF is available on Bioportal and can be accessed at: http://bioportal.bioontology.org/ontologies/BIM