Annotation of Figures from the Biomedical Imaging Literature:

A Comparative Analysis of RadLex and Other Standardized Vocabularies

Charles E. Kahn, Jr., MD, MS

Rationale and Objectives: RadLex is a standardized vocabulary developed for clinical practice, research, and education in radiology. This report sought to analyze the use of RadLex to annotate and index the captions of images from the peer-reviewed biomedical literature and to compare the number of annotations per term for RadLex and five other biomedical ontologies in a large corpus of figure captions from biomedical imaging publications.

Materials and Methods: RadLex and five other biomedical vocabularies were evaluated. A fully automated web service was used to discover the vocabularies' terms in a collection of 385,018 figure captions from 613 peer-reviewed biomedical journals. Annotations (i.e., figure-term pairs) were analyzed by vocabulary. RadLex annotations were analyzed by journal and RadLex term class.

Results: RadLex had the greatest number of annotations per term of the six vocabularies. On average, there were 10.1 RadLex annotations per figure; 380,338 figures (98.8%) were annotated with at least one RadLex term and 288,163 figures (74.8%) were annotated with six or more RadLex terms. Of 39,218 RadLex terms, 8504 (21.7%) were mapped to images in the collection, which was the highest percentage of any of the vocabularies.

Conclusions: Although comprising four to 10 times fewer terms than other vocabularies, RadLex showed excellent performance in indexing radiology-centric content. Almost all of the images in a large collection of figures from peer-reviewed biomedical journals were annotated with at least one RadLex term, and almost 75% of the images were annotated with six or more terms.

Key Words: RadLex; lexicon; ontology; vocabulary; indexing; figure captions; annotation; ARRS GoldMiner.

The RadLex radiology lexicon (www.radlex.org) is a collection of terms designed to provide a uniform vocabulary of radiology (1). It has been developed under the aegis of the Radiological Society of North America and plays an increasingly important role in radiology practice, research, and education. RadLex is being used to categorize journal articles and reviewers (2), encode radiology result information (3,4), search the content of radiology reports (5), analyze queries to web-based search engines (6), and standardize names of imaging procedures for the American College of Radiology's Dose Index Registry (7).

RadLex provides an ontology (knowledge model) that incorporates relationships between terms (8). The main relationship is the subsumption ("is-a") relation, which defines a hierarchy of subclasses. For example, the term "left lung" RadLex term identifier (RID1326) is a subclass of "lung" (RID1301), and "bronchitis" (RID34637) is a subclass of "respiratory disorder" (RID5316). In RadLex, each entity is related to a single, higher level entity, or "parent." This hierarchical representation of radiology terms allows retrieval of information more effectively.

One of the primary intended purposes of RadLex is to index resources for research and education in radiology. This study explored the extent to which RadLex could be used to annotate and index a large database of biomedical images. The study's goal was to better understand the scope and coverage of RadLex in comparison to other, more widely used biomedical ontologies.

MATERIALS AND METHODS

With permission of the American Roentgen Ray Society (ARRS), the database of the ARRS GoldMiner image search engine (goldminer.arrs.org) provided the materials for this investigation. This search engine focuses on clinically relevant images, particularly those that use medical imaging technologies (9). ARRS GoldMiner indexes images from a core set of radiology journals (*AJR American Journal of Roentgenology*, *American Journal of Neuroradiology*, *British Journal of Radiology*, *Journal of Nuclear Medicine*, *Radiology*, and *RadioGraphics*) and the European Society of Radiology's Eurorad case database. ARRS GoldMiner also includes selected images from other journals, including articles submitted to the US National

Acad Radiol 2014; 21:384–392

From the Department of Radiology, Medical College of Wisconsin, 9200 W. Wisconsin Ave., Milwaukee, WI 53226. Received September 28, 2013; accepted November 3, 2013. **Address correspondence to:** C.E.K. e-mail: kahn@mcw.edu

http://dx.doi.org/10.1016/j.acra.2013.11.007

TABLE 1. Comparison of Six Ontologies for Annotation of the ARRS GoldMiner Corpus of Figure Captions Using the Most Recent Version of Each Ontology at the NCBO BioPortal Site

				Annotated Terms		Annotated Figures		Annotations		
Ontology	Version	Release Date	No. of Terms	Number	Percent	Number	Percent	Number	Per Term	Per Figure
FMA	3.1	March 3, 2010	83,281	5398	6.5	324,376	84.2	1,288,568	15.5	3.3
ICD-10-CM	2011_01	January 1, 2011	91,590	1635	1.8	84,987	22.1	104,095	1.1	0.3
LOINC	236	June 1, 2011	171,399	7683	4.5	380,834	98.9	5,008,536	29.2	13.0
MeSH	2012	September 9, 2011	229,698	15,792	6.9	381,978	99.2	3,097,452	13.5	8.0
RadLex	3.8	February 19, 2013	39,218	8504	21.7	380,338	98.8	3,871,573	98.7	10.1
SNOMED CT	2011_07_31	July 31, 2011	395,036	41,371	10.5	384,492	99.9	11,588,578	29.3	30.1
Total			1,010,222	80,383	8.0	385,018		24,958,802	24.7	64.8

FMA, Foundational Model of Anatomy; ICD-10-CM, International Classification of Diseases, Version 10, Clinical Modification; LOINC, Logical Observation Identifier Names and Codes; MeSH, Medical Subject Headings; SNOMED CT, Systematized Nomenclature of Medicine–Clinical Terms.

The Annotated Terms column shows the number of terms from each ontology that appeared in the annotations. The Annotated Figures column shows the number of figures captions from the collection that were annotated.

Library of Medicine's open-access PubMed Central collection. All of the journals are indexed by PubMed and provide the images as part of articles that are made freely available through the web, usually after a subscriber-only period of 6 to 24 months. As of May 2013, ARRS GoldMiner contained information about 385,018 figures from 613 peer-reviewed biomedical journals. Available information included each figure's source, caption, and the title of article in which the figure appeared. For the purposes of the current study, the textual information for each figure consisted of the figure's caption text concatenated to the title of the article in which the figure appeared.

To provide a point of comparison for the performance of RadLex, five other widely used biomedical ontologies were explored. The Foundational Model of Anatomy (FMA) is an anatomical reference vocabulary (10), part of which has been incorporated into RadLex. The International Classification of Diseases, 10th version, Clinical Modification, includes standardized terms and codes for symptoms, signs, abnormal findings, and diseases; it is based on the medical classification list developed under the auspices of the World Health Organization, modified for clinical use in the United States (11). The Logical Observation Identifiers Names and Codes (LOINC) lexicon provides a standard for identifying medical laboratory procedures and observations (12,13). The Medical Subject Headings (MeSH) vocabulary is used by the US National Library of Medicine to index the biomedical literature (14). The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) offers a standard to encode the meanings of health information and supports effective recording of clinical data (15). The National Center for Biomedical Ontology (NCBO) BioPortal (16,17) provided current versions of all six ontologies (Table 1). Except for RadLex, all of the ontologies are part of the U.S. National Library of Medicine's Unified Medical Language System (18).

The NCBO Annotator (19) was used to discover terms from these six ontologies in the article title and figure caption text for each entry in the ARRS GoldMiner database. This process, called "annotation," produced a list of terms associated with each figure. This fully automated web service identified the appearance of ontology terms in a block of text, which was passed to the server through an application programming interface using Simple Object Access Protocol and Representational State Transfer architecture. The NCBO Annotator web service presents the input text to a concept recognition tool along with a dictionary that consists of all concept names from the specified ontologies and other string forms, such as synonyms or labels that syntactically identify the concepts. The Annotator recognizes concepts using string matching on the dictionary to produce a set of direct annotations. For terms that have synonyms or abbreviations, the preferred term is used in the output. The current work did not use the Annotator's semantic expansion components, which allow the system to use "is-a" relation transitivity, semantic distance, and/or cross-ontology mapping to expand the set of annotations to include parent concepts or related concepts from other ontologies. The appearance of a term was counted only once for each figure in the database, even if that term or a synonym appeared more than once in the figure's caption text or article title. The number of annotations per figure was tallied to ascertain the extent to which each ontology covered the content of figure captions. Conversely, the number of annotations per term was assessed as well. For RadLex terms, the possibility of a power-law relationship between the number of annotations per term and the frequency of such terms was explored.

To better understand the use of RadLex terms, they were divided into major classes. A major RadLex class was defined as a term whose parent was the top-level term "RadLex entity" (RID1). For example, "anatomical entity" (RID3) *is-a* "RadLex entity" (RID1) and hence was defined here as