Natural Language Diagnostic tool for Radiologists

Priya Deshpande, Alexander Rasin, Eli Brown, Jacob Furst, Daniela Stan Raicu
DePaul University
School of Computing
College of Computing and Digital Media
pdeshpa1@mail.depaul.edu,arasin@cdm.depaul.edu

Abstract

In Radiology, as in other medical fields, accurate diagnosis of disease is very important for an effective treatment. Teaching Files (TFs) play a vital role in the diagnosis process and are of great help to radiologists treating their patients. TFs also help students of radiology to learn more about their chosen field, and allows patients to make educated decisions about their file. Teaching file are perhaps the best reference, where doctors can learn diagnosis characteristics with the help of images, recorded discussion, references, augmenting annotations and patient history. Every hospital maintains an active collection of Teaching Files for their internal purposes, but many publicly available teaching files are available through online sources – e.g., Radiology Society of North America Medical Imaging Resource community (RSNA MIRC), MyPacs, EURORAD.These multiple public sources typically provide their own basic keyword search interface, but little else that can help doctors and patients find the relevant TFs. In this project, we integrate multiple public sources into a unified repository and propose advanced search features that will make it easier to find teaching files (and secondary sources such as textbooks or journals). Our approach supports incorporating diverse public sources with hospital's internal TF collection to provide a single search tool. We believe that such search engine should be tailored to radiologist needs, providing an understanding of natural language by understanding negation statements, substituting search term synonyms, correctly interpreting adjectives and considering the layout and type of source text. We are also integrating an image-based search that allows finding visually and structurally similar cases. Although these searches are not based on synonyms, negations. No integrated solution for teaching files are available which considers keyword based search along with negations, synonyms, adjectives etc. In our proposed work, we are going to integrate publicly available teaching file sources, journal articles related to radiology and teaching files from hospitals. The proposed system has a huge scope from text analysis as well as an image analysis. Radiologists should be able to easily contribute new cases or augment existing cases by supplying additional comments and annotating images. Our system allows radiologists to make faster and more accurate diagnosis by removing errors by the limits of human memory based on image features user can search images and relevant diagnosis information.

Keywords: Radiological Teaching Files, PACS, DICOM, Data integration

1 Introduction

In this paper we include a full overview of publicly available radiology sources, including both the advantages and limitations of these sources. We discuss our integration of several sources into our repository, our search engine design and a preliminary evaluation of

its search capability. In our system, radiologists will be able to easily contribute new cases or augment existing cases by supplying additional comments and annotating images in the single shared repository. All Teaching Files share the same overall structure but significant variations exist even within the same data sources. Teaching files can include categories such as patient history, findings, diagnosis, differential diagnosis and images related to clinical reports. Radiological journals furthermore have articles with images and text based on clinical reports, some of the data sources also have web links, power point presentations etc. Data cleaning and validation are thus an important aspects of integration when dealing with such data sources. While integrating data sources we maintained Health Insurance Portability and Accountability Act (HIPAA) constraint. We normalized all data sources and designed logical schema which support heterogeneous data sources.

Teaching Files can be accessed by a variety of users – radiologists, radiology residents, experts, clinicians, patients. Our integrated TF solution help all these users by providing a versatile search engine. Queries can vary significantly and thus relevancy of results matters in this domain. In addition to a typical search (e.g., find cases of "cardiomegaly" or "enlarged heart" condition) a user may need a teaching case (e.g., find cases that *look like* cardiomegaly but are not). Here our search engine plays a vital role by considering context of the query. The search engine applies term stemming, considers negations; e.g., "no" or "without" which are important in this domain but are not considered by other general search engines. If radiologists want to search "past history of myocardial infarction" our search engine shows the cases with history of myocardial infarction from journal dataset as well as teaching files dataset. While applying synonym substitution and negation we used Radiological Lexicon (RadLex) [1] and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [2] ontologies which provide a standardized, multilingual vocabulary of clinical terminology that is used by physicians and other health care providers for the electronic exchange of clinical health information. We also integrated synonyms from Oxford medical dictionary and working on integration of other dictionaries like Wordnet.

In Section2 we discuss RadLex and SNOMED ontologies. In Section4 we discussed our current results to show how our system accurately handle synonym, negation and adjectives. We expect to reduce radiologists frustration by providing them with integrated teaching files solution. Otherwise while doing diagnosis radiologists refers different sources and its very difficult for them to memories which source provides which information. Our system allows radiologists to make faster and more accurate diagnosis by removing errors by the limits of human memory.

2 Related work

Our literature survey is based on articles from Journal of Radiology, Radiographics, Digital Imaging, IEEE and other established medical publication venues. We reviewed papers that discussed the need for data integration of radiological sources or reports. There are many papers that argue the need for Big Data utilization and disparate source integration to better serve the medical field, which greatly inspired us to proceed with building IRIS engine. Ron Gutmark [3] argued for building a system that reduces errors in radiological images using teaching file database. Easy-to-use computer teaching files are useful for training physicians,

serve as a reference tool for experienced physicians and help them to improve diagnostic accuracy. Author of [4] discussed how critical radiologic images are for diagnosis, teaching needs and research. They were particularly interested in using case-based radiologic teaching files for radiology teaching. Their proposed architecture meant to be integrated with existing medical image databases (featured by MIRC interoperability), but it is not publicly available. Availability of a large and diverse set of clinical cases need the integration of profiles published by Integration Healthcare Enterprise (IHE) [5].

Having a repository of pathology-proven cases in a dashboard also has the potential to enhance and encourage the formation of accurate teaching files, as well as educational publications in the form of case series or case of the day submissions.[6] As the use of positron emission tomography computed tomography (PET-CT) has increased rapidly, there is a need to retrieve relevant medical images that can assist an image interpretation. Building a database which may provide integrated repository with images to predict diagnosis accurately [7]

Larger clinical reference datasets that are relevant to a larger number of patients may help critical diagnosis. Data integration or centralized open data repository required for clinical data, patient history, physical exam findings, laboratory data, imaging data which can help analyzing diseases more precisely and accurately. Authors form the paper discussed about the how big data analysis could be helpful radiology [8]. From this survey we can conclude that in radiology there is need to integrate clinical reports and images and generate database which may help radiologists to diagnose disease.

3 Proposed System

In this section we focus on datasets that we have successfully integrated and planning to integrate into our data repository. We also overview the methods behind our proposed system; we expect several more sources to be integrated in the near future. Study shows how the existing resources are useful in radiology field as well as what are the limitations of these resources. On different websites teaching files information is available. The following is the list of sources and repositories we have evaluated in determining what databases are currently available to doctors. We intend to integrate these sources as we have integrated RSNA MIRC, MyPacs, RadLex and others.

RSNA MIRC:

Radiology Society of North America Medical Imaging Resource Community. RSNA MIRC is a large repository with teaching files including the information about the history of patient, diagnosis, differential diagnosis, findings, discussion as well as external references (journal articles). Radiological terms are highlighted and linked to RadLex browser (see discussion about RadLex below).

Strengths: Over two and a half thousand teaching files available. Text search that includes body, title, author, abstract and keywords of the TF. Radlex terms are automatically linked to Radlex browser.

Weaknesses: Search is done verbatim with no processing to interpret the goals (e.g,. synonyms, negation). A variety of specialized search fields (e.g., anatomy, age, image modality) built into the UI but not yet implemented. No image-based search is possible.

Mypacs.net:

Publicly available teaching file resource, where radiologists can create, modify and upload teaching files.
Strengths: A total 32,396 cases are available with 202,986 images. User can search records based on anatomy, pathology, modality, age, gender, etc.
Weaknesses: Search engine has no consideration of synonyms, negation and adjectives. No image-based search is available.

RadLex:

Radiology Lexicon term browser. RadLex.org is an ontological system that provides a comprehensive lexicon vocabulary for radiologists. RadLex browser was developed by RSNA and includes 75505 defined terms.

Journals:

Following are the data sources and journals linked from teaching file resources.
MIRC: RSNA Radiographics, Radiology
Radlex: BIR publications -British Institute of Radiology, AJNR
MyPacs.net : MedScape
Gamuts: European Journal of Cardio Thoracic Surgery, AJR (American Journal of Roentgenology),Radiographics, Radiology

CTisus:

Huge repository with radiological images, quizzes and CT protocols
Strengths: 237,814 images available. Video files are a good additional source for learning.
Weaknesses: Supplemental case information such as diagnosis, history of the patient and differential diagnosis are not available. No image-based search.

Medscape:

Latest medical news and information about drugs and diseases are available for radiology students and physicians.
Strengths: Articles focused on different anatomical structures, including experts viewpoints and guidance.
Weaknesses: No search engine is available. There are no teaching files that could provide radiologists with images, patient history, differential diagnosis and other valuable information.

Content-based queries on the casimage database with the IRMA framework [9]:

An integration of a multimedia teaching and reference database in a PACS environment.
Strengths: Integration of PACS with other types of image formats. Data editing and annotation made possible through client-server framework.
Weaknesses: Not publicly available (we were unable to determine if it is used internally). Database includes only 8,723 images. No concept of context- or image-based search.

RADTF [10]:

Teaching file solution, which is compatible with RadLex. Differential diagnosis, quiz modes are available.
Strengths: Natural language processing used to ingest radiologic reports. Search engine uses RadLex anatomy concept terms, stemming, ranking of results based on detected negation, hedge, or uncertainty expressions.
Weaknesses: Not publicly available. The link provided in the article is no longer active, the system appears to be defunct.

infoRAD Vendor-Neutral Case Input into a Server-based Digital Teaching File System RADICS[11]:

The RadICS server could handle CT, MR, computed radiography, and digital radiography images adhering to the DICOM format.
Strengths: Efficiently harvesting images from DICOMcompliant PACS with minimal radiologist work flow disruption.
Weaknesses: Not available publicly. Unable to determine if the system is still used or available to anyone.

Biomedical Image Metadata Manager (BIMM) [12]:

BIMM system provides retrieval of similar images using semantic features of metadata.
Strengths: Based on imaging observation, 2D regions of interest (ROIs) stored as metadata. Authors discussed content-based image retrieval capabilities.
Weaknesses: Not available publicly. The link provided in paper is defunct, unable to find any traces of the system in recent years.

The caBIG™Annotation and Image Markup Project [13]:

This project developed a mechanism for modeling, capture and serializing image annotation, which is readable by human as well as machine.
Strengths: Project allows for streamlined image annotation.
Weaknesses: Not available publicly, software is not currently supported.

Radiology Teacher [4]:

Web-based teaching file development and distribution program.
Strengths: Allow authors to create, edit, and delete cases and images with descriptions and annotations. Quiz mechanism is provided for the radiologists. Image annotation feature integrates an interface to the Medical Illustrator software.
Weaknesses: Anyone can changes Teaching Files. Some features claimed in the paper (e.g., image annotation) is not available. Database only contains 321 cases.

Yottalook:

A radiologist-targeted search engine ("powered by Google Custom Search"). Searches a variety of sources such as radiopaedia.org, Americal Journal of Radiology, University of Michigan Medical School, MyPacs. Strengths: search engine with relevance or date ranking, provides users with ability to choose category of search (e.g., CT, Ultrasound, .edu).

Weaknesses: The data sources are not integrated as one – users who chose a search category (e.g., X-Ray) are then redirected to original source in their specific format (e.g., a webpage or PowerPoint file).

Radiopaedia:

open-edit radiology resource.
Strengths: 25,640 cases and 10,409 articles are available to help radiologists in diagnosis.
Weaknesses: No search engine support for teaching files. No additional categories (e.g,. DDX, findings) provided.

Gamuts:

Comprehensive lists of imaging differential diagnoses
Strengths: Disease states linked to symptoms, disease names and causes. Images linked to goldminer radiology search engine.
Weaknesses: No search engine available.

GoldMiner:

Helps user to search images and articles from peer-reviewed biomedical journals. Uses National Library of Medicine(part of NIH) to discover medical concepts in figure captions, and retrieve relevant images.
Strengths: Recognizes abbreviations, synonyms, and types of diseases.
Weaknesses: Search results can depend on the presence of specific words in the figure captions.

EURORAD (European Society of Radiology):

A peer-reviewed educational tool based on teaching cases. Teaching files with clinical history, image findings, discussion, final diagnosis and differential diagnosis lists available.
Strengths: 6691 cases and user can search based on anatomical structure.Support for 2 languages is available.
Weaknesses: Search results do not consider negations and synonyms. No image-based search.
Table 1 shows comparative study of different data sources. Our goal is to integrate all of the available public sources and let users retrieve accurate results by augmenting search with synonyms and correctly interpreting negation and adjectives. In our database system we captured the data from these publicly available sources and cleaned the data to a normalized schema before loading it. In the Results section we present a comparative analysis using our system (IRIS) comparing it to RSNA and MIRC search engines and applying Google search to the original websites. We used RadLex synonym terms dictionary and Oxford medical dictionary for keyword synonym substitution. Figure 1 shows the workflow of our IRIS engine with web-based user interface. Here we used "Cardiomegaly" as a keyword for our search. Our search engine retrieves results based on "Cardiomegaly" as well "Enlargement of Heart" which is an automatically detected synonym. Based on both of these terms, IRIS search engine retrieves and shows TF images along with associated text.

	Type of queries (negations, adjectives, etc.)	NLP capabilities					
Search Engine	**Keyword-based search**	**synonyms**	**morphological forms**	**Relationships between terms**	**Corrected forms of spelling errors**	**Relevance feedback**	**Other Functionality**
RadTF	YES	NO	YES	YES(NA)	NO	NO	
Render	YES	NO	NO	YES(NA)	NO	YES	
GoldMiner	YES	NO	NO	NO	YES	YES	
Yottalook	YES	YES	NO	YES	YES	YES: Ranking Query Results	Survey added at the end of document
Google	YES	NO	NO	NO	YES	YES	
MIRC	YES	NO	NO	NO	NO	NO	
MyPacs	YES	NO	NO	NO	NO	YES	
Gamuts	NO	NO	NO	NO	NO	NO	Disease states linked to symptoms, diseases and causes. Images linked to goldminer
CTisus	YES	NO	NO	NO	Prompt User To Check Spelling	NO	
Casimage	NA	NO	NA	NA	NO	NO	
Project with Osirix	NA	NA	NA	NA	NO	NO	
RadICS	NA	NO	NO	NA	NO	NA	
BIMM	YES	NO	NO	NA	NO	NO	
CaBIG	YES	NO	NO	NO	NO	NO	
Radiology Teacher	YES	NO	NO	NO	NO	NO	
Medscape	YES	NO	NO	NO	YES	NO	

Table 1: Data Sources comparative study.

4 Results

We illustrate advantages of our approach using a few simple queries given to us by our radiologist collaborators that compare state-of-the-art tools and our integrated database with the same content. We have used MIRC, MyPACs and RSNA RadioGraphics and Radiology journals as a third integration data source. MIRC is one of the best known available public sources with about two thousand public teaching files. MyPACs is another widely used public source that contains over 30 thousand files, but our proof-of-concept uses two thousand teaching files for consistency with MIRC. The data used here has gone through initial cleaning and normalization (e.g., removal of blank and duplicate entries). Our IRIS database improves results along two dimensions: search queries produce more relevant results (per source) and results from multiple data sources are merged into single easy-to-query source.
A.Single-term search and negation:

Figure 2 summarizes the comparative performance of different searches. With single term

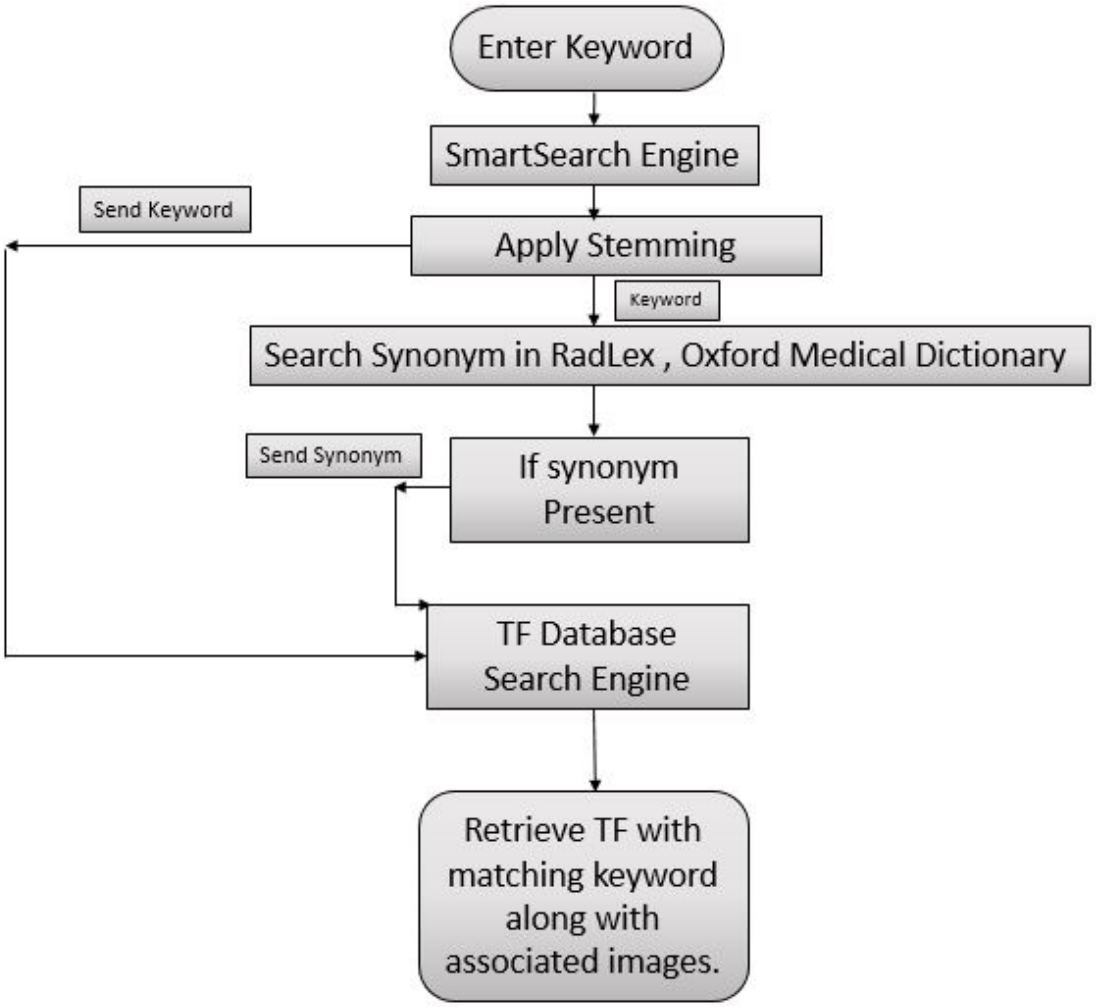

Figure 1: IRIS Engine Workflow.

Cardiomegaly search MIRC search engine results in 56 results, MyPacs shows 2 results with 2k teaching files. Our IRIS engine retrieves 56 cases in MIRC, as search engine also search for "Enlargement of heart" which is synonym for "Cardiomegaly". For MyPacs we found 2 additional cases, i.e., 4 cases in total. We also integrated RSNA journal as one of the data source in our search engine. RSNA have search engine to retrieve the articles from the database. It retrieves the results based on entered phrases, words etc. Filter can be applied to research article, case report. Articles can be retrieved by relevancy or date. Free articles are highlighted as Free. For cardiomegaly as a keyword RSNA journal searches 519 results. Our search engine could also find 130 articles from RSNA. We also used Google Site search as one of the comparative tool to compare the results. Google search shows 24 and 72 results for MIRC and MyPacs site search. Our next search was based negation. We tried searching No Cardiomegaly. Oddly no search engine applies the concept of negation while retrieving results. MIRC and MyPacs shows same teaching files as with cardiomegaly. Our search engine smartly replaces No Cardiomegaly with Normal Heart and retrieves 10 results

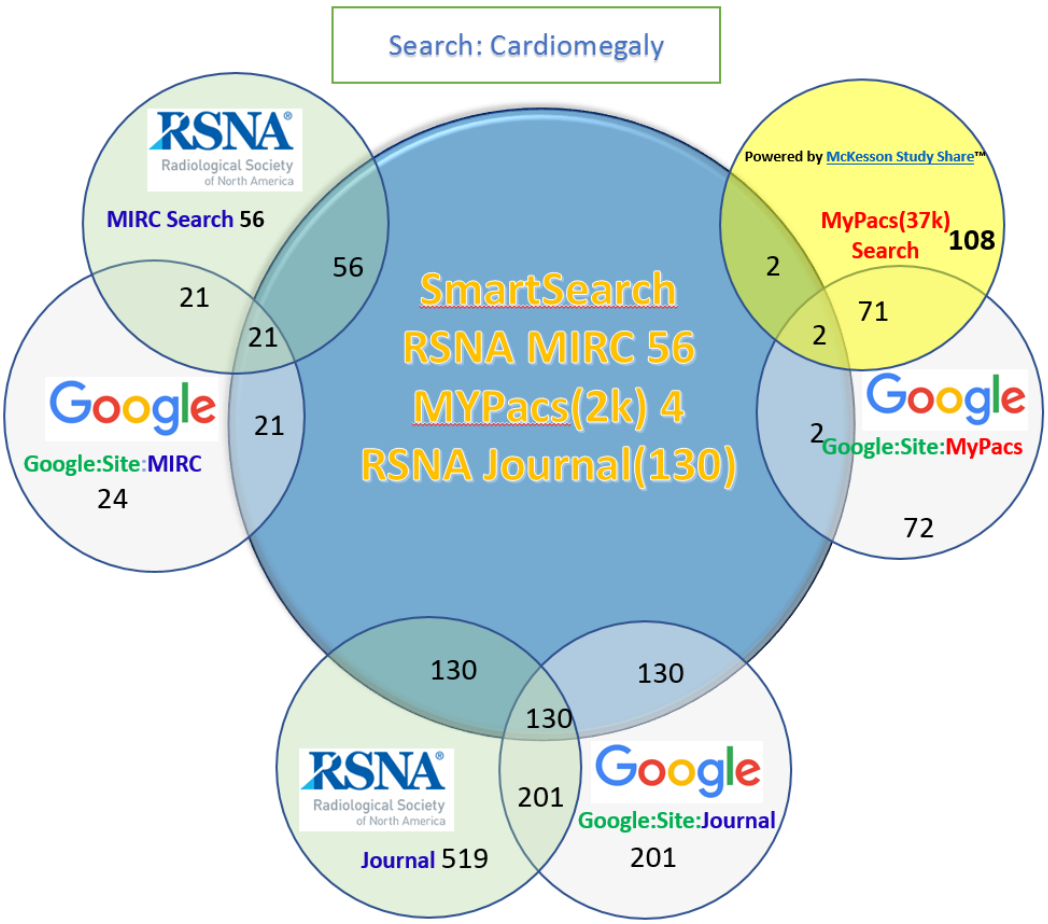

Figure 2: Search results "Cardiomegaly"

with MIRC data source and 8 with MyPacs data source. Using a dictionary built using RadLex and Oxford dictionary our prototype automatically replaced No Cardiomegaly with Normal Heart. Some teaching file sources (e.g., MIRC) use RadLex-based annotations; we automatically augment the rest of the imported data in our database.

B.Synonym search:

We tried synonym search on all these teaching file resources. For irregularly shaped keyword MIRC and MyPacs shows zero results; Google shows 4 results; while our search engine shows 3 results. Using synonym substitution as "abnormally shaped" search engine shows more 5 results. Here the results are not only counted as how many results search engine retrieves , but more about relevancy of results. Our system also results in 197 journals for abnormal shape keyword based search.

C. Adjective-partitioned search:

Our next search was based on different types of pulmonary edema. Ultimate results should include cases related to pulmonary edema. MIRC search for severe pulmonary edema, moderate pulmonary edema and mild pulmonary edema shows 19, 12, 17 results respectively with significant amount of overlap. Our search engine results in 23 cases with results relevant to pulmonary edema. Out of these 23 cases, 2 teaching files use mild, 6 teaching files use word

severe and rest 15 teaching files we interpret as moderate. Google produces 9 for MIRC and 210 for MyPacs. Using classification algorithms in machine learning we can improve our search results more accurate and precise.Using NLP and classification model, results can be improved

5 Conclusion

Our proposed system is an integrated database repository for radiology teaching files with text and images. Which allows radiologists to make faster, more confident and accurate diagnoses by removing the innate error caused by the limits of human memory. Search radiology reports interpreting negation, checking for synonyms, and considering adjectives and correlation between search terms. We are going to integrate data sources we mentioned in related work.Based on extensive discussions with experienced (25+ years) radiologists, IRIS will be a great improvement of existing tools – currently radiologists use internal TFs with limited search capability.

References

[1] R. S. of North America (RSNA), Radlex ontology, `http://www.radlex.org/` (2017).

[2] S. I. I. H. T. S. D. Organization, Snomedct ontology, `http://www.snomed.org/` (2017).

[3] R. Gutmark, M. J. Halsted, L. Perry, G. Gold, Use of computer databases to reduce radiograph reading errors, Journal of the American College of Radiology 4 (1) (2007) 65–68.

[4] R. Talanow, Radiology teacher: a free, internet-based radiology teaching file server, Journal of the American College of Radiology 6 (12) (2009) 871–875.

[5] M. Dos-Santos, A. Fujino, Interactive radiology teaching file system: the development of a mirc-compliant and user-centered e-learning resource, in: Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, IEEE, 2012, pp. 5871–5874.

[6] L. R. Margolies, G. Pandey, E. R. Horowitz, D. S. Mendelson, Breast imaging in the era of big data: structured reporting and data mining, American Journal of Roentgenology 206 (2) (2016) 259–264.

[7] K. H. Hwang, H. Lee, G. Koh, D. Willrett, D. L. Rubin, Building and querying rdf/owl database of semantically annotated nuclear medicine images, Journal of Digital Imaging (2016) 1–7.

[8] A. P. Kansagra, J. Y. John-Paul, A. R. Chatterjee, L. Lenchik, D. S. Chow, A. B. Prater, J. Yeh, A. M. Doshi, C. M. Hawkins, M. E. Heilbrun, et al., Big data and the future of radiology informatics, Academic radiology 23 (1) (2016) 30–42.

[9] C. Thies, M. O. Güld, B. Fischer, T. M. Lehmann, Content-based queries on the casimage database within the irma framework, in: Workshop of the Cross-Language Evaluation Forum for European Languages, Springer, 2004, pp. 781–792.

[10] B. H. Do, A. Wu, S. Biswal, A. Kamaya, D. L. Rubin, Informatics in radiology: Radtf: A semantic search–enabled, natural language processor–generated radiology teaching file 1, Radiographics 30 (7) (2010) 2039–2048.

[11] A. W. Kamauu, S. L. DuVall, R. J. Robison, A. P. Liimatta, R. H. Wiggins III, D. E. Avrin, Vendor-neutral case input into a server-based digital teaching file system 1, Radiographics 26 (6) (2006) 1877–1885.

[12] D. Korenblum, D. Rubin, S. Napel, C. Rodriguez, C. Beaulieu, Managing biomedical image metadata for search and retrieval of similar images, Journal of digital imaging 24 (4) (2011) 739–748.

[13] D. S. Channin, P. Mongkolwat, V. Kleper, K. Sepukar, D. L. Rubin, The cabig annotation and image markup project, Journal of digital imaging 23 (2) (2010) 217–225.