Published in final edited form as:
Proc IEEE Int Conf Big Data. 2014 October ; 2014: 790–795. doi:10.1109/BigData.2014.7004307.

Empowering Personalized Medicine with Big Data and Semantic Web Technology: Promises, Challenges, and Use Cases

Maryam Panahiazar[*,†], **Vahid Taslimitehrani**[*,†], **Ashutosh Jadhav**[*,†], and **Jyotishman Pathak**[*]

[*]Center for Science and Healthcare Delivery, Mayo Clinic, Rochester, MN, USA

[†]Ohio Center for Excellence in Knowledge-enabled Computing (kno.e.sis) College of Computer Science and Engineering, Wright State University, Dayton, OH, USA

Abstract

In healthcare, big data tools and technologies have the potential to create significant value by improving outcomes while lowering costs for each individual patient. Diagnostic images, genetic test results and biometric information are increasingly generated and stored in electronic health records presenting us with challenges in data that is by nature high volume, variety and velocity, thereby necessitating novel ways to store, manage and process big data. This presents an urgent need to develop new, scalable and expandable big data infrastructure and analytical methods that can enable healthcare providers access knowledge for the individual patient, yielding better decisions and outcomes.

In this paper, we briefly discuss the nature of big data and the role of semantic web and data analysis for generating "smart data" which offer actionable information that supports better decision for personalized medicine. In our view, the biggest challenge is to create a system that makes big data robust and smart for healthcare providers and patients that can lead to more effective clinical decision-making, improved health outcomes, and ultimately, managing the healthcare costs. We highlight some of the challenges in using big data and propose the need for a semantic data-driven environment to address them. We illustrate our vision with practical use cases, and discuss a path for empowering personalized medicine using big data and semantic web technology.

Keywords

Personalized Medicine; Big Data; Semantic Web; Smart Data; Health Care

I. INTRODUCTION

The proliferation of smart devices and exponentially decreasing cost to sequence the human genome along with growth of electronic communication via social media is generating an explosion of health related data that is specific to a given individual. It is estimated that by year 2020 healthcare data will reach 25,000 petabytes -a 50-fold increase from year 2012

[1]. While such massive amounts of streaming data provides tremendous opportunities to develop methods and applications for advanced analysis, the real value can be only realized when such information extracted from data can improve clinical decision making and patient outcomes, as well as lower healthcare costs. The McKinsey Global institute estimates that applying big data strategies to better inform decision making in U.S. healthcare could generate up to $100 billion in annual value [2].

This is particularly relevant to the practice of personalized medicine that aims to individualize the diagnosis of a disease and therapy according to the individual patient's characteristics (e.g., clinical co-morbidities and genetics), as opposed to decisions based on evidences and guidelines derived from population-based studies and clinical trials. A major emphasis of personalized medicine is to match the right drug with the right dosage to the right patient at the right time [3]. The process of personalized medicine could also be facilitated with the comparison of a new patient to patients with similar characteristics. This could lead to faster and more accurate diagnoses and consideration of therapeutic options. However, before the full potential of big data can be realized for healthcare in general, and personalized medicine in particular, several challenges related to data integration, processing and analytics, visualization and interpretation need to be addressed. Our objective is to explore technologies such as semantic web and data analysis in the context of transforming big data to smart data are applicable for personalized medicine. In particular, we present some use cases that highlight the opportunities for how big data along with semantic web and data analysis can be applied to deliver personalized medicine for improving outcomes and managing healthcare costs. The paper discusses the following:

1. The characteristics of personalized medicine and its impact in healthcare,
2. The role of big data in personalized medicine including challenges and limitations,
3. The role of technologies such as semantic web to use big data in smart way for empowering personalized medicine,
4. Use cases to illustrate some of the opportunities and potential solutions in healthcare when big data is transformed into smart data.

The subsequent sections are organized as follows: Section II reviews and discusses personalized medicine and its impact in healthcare. Section III discusses big data challenges and opportunities for healthcare. Section IV presents the role of big data in personalized medicine via use cases and highlights the role of semantic web and advanced data analysis techniques. We conclude our discussion in Section V.

II. CHARACTERISTICS OF PERSONALIZED MEDICINE AND ITS IMPACT IN THE HEALTH CARE SYSTEM

The overarching goal of "personalized medicine" is to create a framework that leverage patient EHRs and OMICS (primarily genomics) data to facilitate clinical decision-making that is predictive, personalized, preventive and participatory (P4 Medicine) [4]. The emphasis is to customize medical treatment based on the individual characteristics of

patients and nature of their diseases, diagnoses and responses to treatments. In personalized medicine, clinicians can:

1. Validate medical treatments and response to therapies. They can predict the possible side effects and detect adverse events to treatments based on genetic make up for each individual patient in comparison to other similar patients,
2. Describe better targeted therapies for individuals. They can determine - apriori - which drug(s) will work better with each individual patient, instead of adopting an empricically driven approach of trial-and-error,
3. Make better decisions on risk prediction and focus on prevention, rather than disease management. Collecting genetic information from prenatal testing can be useful to determine possible diseases in future that can either be avoided, or adequately controlled.

III. BIG DATA CHALLENGES, OPPORTUNITIES, AND POSSIBLE SOLUTIONS IN PERSONALIZED MEDICINE

Big data could have a very big impact in health care especially in personalized medicine. As reported by the Institute for Health Technology Transformation by 2013, US heath care organizations have generated 150 Exabytes of health care data [5], and this data will continue to grow in the coming months and years. This unprecedented amount of data, when meaningfully used, can provide significant insights in avoid unnecessary treatments, minimizing drug adverse events, maximizing overall safety, and eventually leading to much more effective and efficient healthcare system, and provide a path realize the objectives of personalized medicine. In the following, we explain some of the challenges for using big data in personalized medicine, and discuss the relevant use cases in the next section.

A. Variety of the data

To study personalized medicine we need to navigate and integrate clinical information (e.g. medical diagnosis, medical images, patient histories) and biological data (e.g. gene, protein sequences, functions, biological process and pathways) that have diverse formats and are generated from different and heterogeneous sources. While the last decade was dominated by challenges in handling massive amounts of data, we argue that it is now important to move focus on developing tools and techniques to make better sense of data and the use of information for knowledge discovery. Despite the general availability of datasets, interoperability is still lacking. We believe that data integration and making use of different data sources is at the core of personalized medicine. It provides physicians and researchers an integrated view across genotypes and phenotypes. Whereas some of the data are accessible in structured formats, most of the data are only available in unstructured format, such as image or text. The heterogeneity and diversity of data thus limits their accessibility and re-usability.

To address this challenge, in our research projects [6], [7] we developed an ontology-driven semantic problem solving environment to improve personalized medicine in the sense that make data and resources more “understandable” and “interpretable” for applications to

integrate, search and query. Specifically, we used semantic web technologies to make knowledge interpretable by the agent in the web to enable integration and re-usability of heterogeneous resources for knowledge discovery and prediction.

B. Quality of the data

An important aspect to ensure data quality is data standardization and terminologies with semantic mapping. This typically requires extending existing ontologies, or in many cases, development of domain-specific ontologies. One of the big challenges in ensuring data quality is understanding both syntactic and semantic differences in data sources, and how they can be harmonized. By having consensus-based common vocabularies and ontlogies, it is feasible to annotate datasets with appropriate semantics to make the resources more understandable and interpretable for applications and agents on the web [8], [7], [9].

C. Volume of the data

However, having the huge amount of data itself does not solve any problem in personalized medicine. We need to summarize or abstract data in the meaningful way to translate data to information, knowledge and finally wisdom. We still need to investigate to effectively translate large amount of data for making use of them in decision-making. To handle the huge amount of data, tools such as Hadoop systems can help us to speed up our data processing and querying.

D. Velocity of the data

Healthcare data is continuously changing and evolving. These rapid changes in the data poses a significant challenge in creating relevant domain models on-demand to be useful for searching, browsing, and analysis of real-time content. In turn, ” this requires addressing the following issues: (1) the ability to filter, prioritize and rank the data (relevant to the domain, or use case); (2) the ability to process and ingest data quickly; and (3) the ability to cull, evolve, and hone in on relevant background knowledge” [10].

IV. USE CASES

Making personalized medicine to be successful we require accessing and processing vast volumes of structured and unstructured data about individual patients. This data includes both structural and unstructured data types. Since traditional technology is not equipped to this end, the big data and semantic web related technologies come into the play. Here we explain our platform implementing a ”Big Data” architecture and illustrate use cases that address some of the issues highlighted above.

A. Big Data platform

From an architectural perspective, the use of Hadoop in our applications as a complement to existing data systems is important: IT offers an open source technology designed to run on large numbers of connected servers to scale-up data storage and processing in a very low cost and it is proven to scale to the needs of the largest web properties in the world. The Hadoop’s architecture offers new venues for data analysis including [11]:

1. Schema on read: Users can store data in the HDFS (Hadoop data file systems) and then design their schema based on the requirements of the application.
2. Multi-use, Multi-workload data processing: Multiple users can have access to a shared data set in the same time for close to real time analysis.
3. Lower cost of storage.
4. Data warehouse workload optimization.

As Apache Hadoop has become more popular and successful in its role in enterprise data architectures, the capabilities of the platform have expanded significantly in response to enterprise requirements. For example in its early days the core components of a Hadoop system has been represented by HDFS storage and MapReduce computation system. While they are still the most important ones, many other supporting projects have been contributed to the Apache Software Foundation (ASF) by both vendors and users. These Enterprise Hadoop capabilities are aligned to the following functional areas that are a foundational requirement for any platform technology: Data Management, Data Access, Data Governance & Integration, Security and Operations [12].

The following architecture is an amalgam of Hadoop data patterns that we designed to use of Hortonworks Data Platform (HDP) in Mayo's health care systems which is shown in Figure 1. HDP is powered by Open Source Apache Hadoop. HDP provides all of the Apache Hadoop projects necessary to integrate Hadoop as part of a Modern Data Architecture [12].

Based on our architecture, we store our datasets from different resources including EHRs, Genomics, and Medical Imaging into the Hortonwork repository and then use scripting tools like Pig and Hive to clean and prepare our data. One of the applications of an implementation of this architecture at Mayo Clinic is data retrieval and cohort creation. There are many data sources available in different departments of Mayo Clinic and each one includes millions of EHRs data and creating cohort is one of the main steps in each project. Using spreadsheets for extracting records from million records of EHR data based on the cohort criteria is a time consuming and painful job. Pig is one of the big data tools that produce a sequence of MapReduce programs to run complex tasks comprised of multiple interrelated transformations. In one of our projects about the integration of different data sources such as lab results, medications, and patient demographics to predict survival score of each heart failure patients, our cohort is the patients with heart failure diagnosis event with at least one EF (Ejection Fraction) value within three months of the heart failure diagnosis date. To create our cohort, we need to extract our desirable records from the aggregation of four large datasets including one heart failure clinical trial and three EHR datasets from different Mayo's clinical systems. Using any spreadsheet based tool or even SQL to retrieve data from these datasets is almost impossible. We implemented our cohort criteria in the form of pig queries in three steps: we filter all patients with heart failure ICD9 code, then in the second step, we join the results of the first query with the patients EF records, and finally the results of the second query is be filtered based on the time intervals defined in the cohort by domain experts and clinicians. Pig translates our queries to a sequence of MapReduce jobs and the jobs are sent to the servers sequentially. Using pig to

create our cohorts is faster and easier than any other tools. Our dataset includes more than 150 million patient records that require usage of parallel querying and computation.

To compare the performance of Pig with other tools like SQL, we ran a simple test on a data file including one million rows of data and a simple operation like AVERAGE. The SQL took 18 minutes to run but Pig based alternative ran in less than two minutes on a cluster with just two nodes.

B. Semantic Analysis of Health Search Query Log

With the growing availability of online health resources, consumers are increasingly using the Internet to seek health related information. One of the prominent ways to seek online health information is via web search engines such as Google, Bing, Yahoo!, etc. Thus in our recent study [13] in understanding consumers health information needs for cardiovascular diseases (CVD) and engaging individual patients in their treatment process, we analyzed significantly large corpus of 10 million CVD related search queries. These queries are submitted from Web search engines and directs users to the Mayo Clinic's consumer health information portal[4]. In order to understand consumer's health information needs i.e. what health topics consumer search in the context of CVD we selected 14 health categories such as Symptoms, Treatment, Food and Drugs and Medications. We categorized the CVD related search queries into the selected health categories by mapping the search queries to UMLS concepts and semantic types using UMLS MetaMap [14].

UMLS[1] incorporates variety of medical vocabularies and concepts, and maps each concept to semantic types. Therefore by using UMLS, we can attempt to infer the underlying semantics of the search query terms. For example, the search query red wine heart disease is mapped to Red wine and Heart disease UMLS concepts and their UMLS semantic types are Food and Disease and Syndrome, respectively. We utilized UMLS MetaMap tool for the mapping search queries to UMLS concepts and semantics types (data annotation). MetaMap is a tool developed at National Library of Medicine (NLM) to map text to the UMLS Metathesaurus. In order to use MetaMap tool, we need install the MetaMap server. Once the server is running, it can be queried with text input and the server returns the UMLS concepts, their semantic types, Concept Unique Identifiers (CUIs), and other details for the terms in the text. Based on the UMLS semantic types and concepts for search queries, we implemented a rule based categorization approach (with Precision: 0.8842, Recall: 0.8607 and F-Score: 0.8723) to categorize 10 million CVD related search queries.

1) Processing Challenge and Solution in Hadoop MapReduce Application—In the semantic analysis, we need to query MetaMap server for annotating each search query. The major data processing challenge in this process is that the MetaMap takes significant time to process the input queries and to return their UMLS concept mappings. For example, to process 100,000 queries on a single node with sequential processing takes around 10 hours. As MetaMap processing was very slow and the size of our dataset was fairly large (10

[1]Unified Medical Language System. Online Available: http://www.nlm.nih.gov/research/umls/

million), it would take around 40 days to finish the annotation sequentially. Therefore, we utilized Hadoop MapReduce [15] framework for speedup.

Once the search query data is submitted to the Hadoop MapReduce framework, the data is split between several mappers on 16 nodes, Figure 2. For example, given a set of Q search queries, each node process q=Q/16 search queries and each mapper process q/N search queries, where N is number of mappers. Each mapper processes multiple search queries and for each search query the mapper queries the local MetaMap server, Figure 3. Mapper output the search query with UMLS concepts and semantic types and reducers consolidated mappers output. We observed a very significant improvement in the data processing time. With MapReduce framework (16 node cluster, with each node running MetaMap server) we could reduce data processing time for 10 million search queries to around 2 days as compared to 40 days time on a single node sequential processing mode.

C. Data Annotation for Structured, Semi-structured and non Structured Resources

Massive amounts of data have been collected from structured data in databases and ontologies to semi-structured data such as XML, files or data with proprietary structure in experimental results files to unstructured text, tables and images in scientific publications and reports. A semantic representation of the data needs to be the first step toward a sustainable integration infrastructure, because a syntactic match or an ad-hoc manual integration is either error prone or not automatically reproducible [10]. In this project we looked at all pieces that required translating data sources and knowledge into the knowledge to use in clinical care with using semantic web technology such as annotation the resources with the concepts from biological and biomedical ontologies.

Semantic web technology is used in this research to make knowledge interpretable by web agents, thereby enabling the integration and re-usability of heterogeneous resources for knowledge discovery. With data annotation we tried to help machines understand unstructured and semi-structured resources such as images, text, and XML files to use them in integration with other data sources. Through tagging or annotation of data with the ontology concepts, unstructured or semi-structured data becomes standardized and understandable for agent of the web and machines to use. This annotation can be used to improve searching the annotated data, and using the annotated data to integrate with other resources. In this research we have annotated XML file, text (e.g. scientific literature, academic article), EHRs, image (e.g. radiology image, CT, molecular image, MRI), and we implemented an application to search and query the resources. More details are available in our prior works [16], [6], [17].

Figure 4 Shows the annotation of images with Kino [6]. Kino is an integrated tools to annotate unstructured and semi structures resource. More details about this project could be found in our previous papers [16], [6], [17]. Figure 5 shows the Image Maker as a tools implemented by author for annotating images and add meta data to each selected part of the images. More detail about this tools could be found in [18].

The use cases above highlight our efforts and progress till date to address some of the challenges in realizing the potential of smart data for personalized medicine. In particular,

we posit the need for creating a robust, scalable and flexible semantic-driven big data infrastructure that will enable higher data quality, web-scale integration and advanced analytics for improving health outcomes and lowering costs.

V. CONCLUSION

Big data is already starting to demonstrate its economic and clinical value in the field of personalized medicine. However, to realize its full potential, we posit that “smart data” is a requirement to enable down-stream analysis and extraction of meaningful information. This will, in our opinion, enable large-scale data science using techniques such as inductive reasoning and topic modeling, exploratory data analysis will allow discovery of data patterns that, unlike traditional statistically driven methods which are hypotheses based, will be independent of a specific hypotheses.

Beside technology and infrastructure challenges for using smart data to enable personalized medicine, there are several other challenges that were not discussed in this paper. In particular, it is becoming increasingly clear that to leverage big data in a smart way, healthcare organizations and policy makers alike need a fundamental shift in their decision making, and embrace a ‘brave new world’ that promotes data sharing with appropriate security and privacy protections, new policy guidelines for collaborative national and international data science efforts, and strategic funding and investment in training data scientists all working jointly toward the goal of delivering personalized high-quality care at lower costs.

References

1. Roski J, Bo-Linn GW, Andrews Ta. Creating value in health care through big data: opportunities and policy implications. Health affairs (Project Hope). Jul.2014 33(7):1115–22. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25006136. [PubMed: 25006136]
2. Datta, M. How Big Data Will Lower Costs and Advance Personalized Medicine. GEN Exclusives; 2013. Tech Rep
3. Jain, KK. Textbook of Personalized Medicine. Springer; 2009.
4. Casey O, Tarczy-Hornoch P. Personolized medicine: challenges and opportunities for translational bioinformatics. 2013; 10(5):453–462.
5. Teli, N. Big Data: A Catalyst for Personalized Medicine. 2014. [Online]. Available: http://healthcare-executive-insight.advanceweb.com/Features/Articles/Big-Data-A-Catalyst-for-Personalized-Medicine.aspx
6. Ranabahu, A.; Parikh, PP.; Panahiazar, M.; Sheth, AP.; Logan-Klumpler, F. Kino: A Generic Document Management System for Biologists Using SA-REST and Faceted Searc; 2011 IEEE Fifth International Conference on Semantic Computing; Sep. 2011; p. 205-208.
7. Panahiazar M, Sheth AP, Ranabahu A, Vos Ra, Leebens-Mack J. Advancing data reuse in phyloinformatics using an ontology-driven Semantic Web approach. BMC medical genomics. Jan. 2013 6(Suppl 3):S5. Suppl 3. [PubMed: 24565381]
8. Panahiazar, M.; Ranabahu, A.; Taslimi, V.; Yalamanchili, H.; Stoltzfus, A.; Leebens-Mack, J.; Sheth, AP. PhylOnt : A Domain-Specific Ontology for Phylogeny Analysis; IEEE International Conference on Bioinformatics and Biomedicine; 2012;
9. Pathak J, Kiefer RC, Bielinski SJ, Chute CG. Applying semantic web technologies for phenome-wide scan using an electronic health record linked Biobank. Journal of biomedical semantics. Jan. 2012 3(1):10. [PubMed: 23244446]

10. Krishnaprasad, T.; Sheth, AP. Semantics for Big Data. Semantics-Empowered Approaches to Big Data Processing for Physical-Cyber-Social Applications. AAAI Technical Report FS-13-04
11. White, T. Hadoop: The definitive guide. O'Reilly Media, Inc; 2012.
12. Hortonwork. A Modern Data Architecture with Apache Hadoop, The Journey to a Data Lake. Hortonworks; 2014. Tech Rep
13. Jadhav, AS.; State, W. Online Information Searching for Cardiovascular Diseases : An Analysis of Mayo Clinic Search Query Logs. 2014.
14. Aronson, R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program; Proceedings/AMIA … Annual Symposium AMIA Symposium; Jan. 2001; p. 17-21.
15. Dean, J.; Ghemawat, S. MapReduce : Simplified Data Processing on Large Clusters. Google,Inc.; 2004. Tech Rep
16. Panahiazar, M.; Leebens-Mack, J.; Ranabahu, A.; Sheth, AP. Using semantic technology for Phylogeny; AMIA.Annual Symposium proceedings, TBI; 2012. p. 175no. iEvoBio
17. Panahiazar, M.; Ranabahu, A.; Taslimi, V.; Yalamanchili, H.; Stoltzfus, A.; Leebens-Mack, J.; Sheth, AP. PhylOnt : A Domain-Specic Ontology for Phylogeny Analysis; IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2012; 2012; p. 106-116.
18. Panahiazar M. My Health: A Multidimensional Approach for Personalized Medicine Empowered with Translational Medicine. 2014 TBI, 2014 Joint Summits on Translational Science. 2014

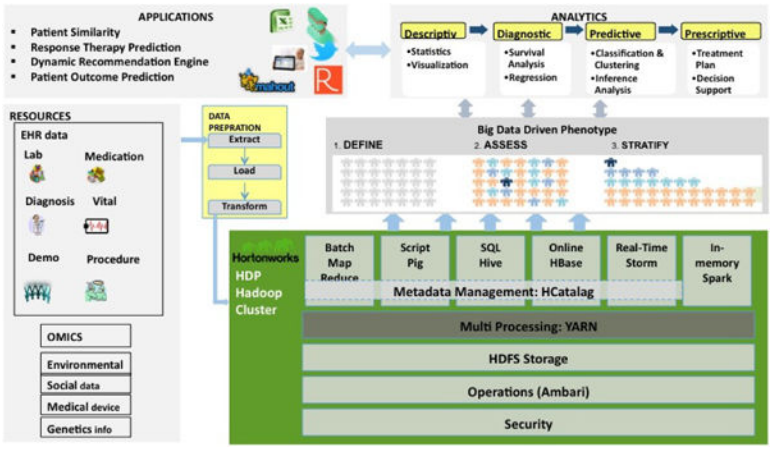

Figure 1.
Mayo's Big Data Architecture

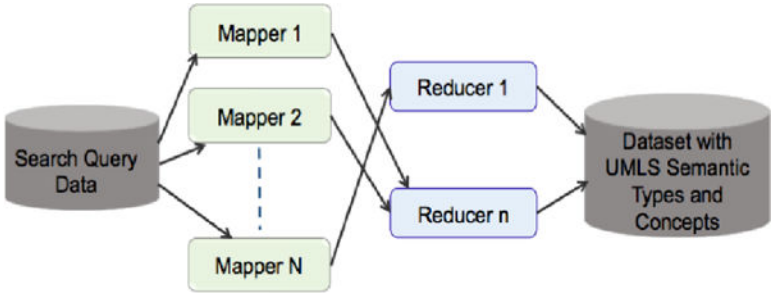

Figure 2.
Hadoop MapReduce framework with 16-node cluster for processing of the search queries with UMLS MetaMap.

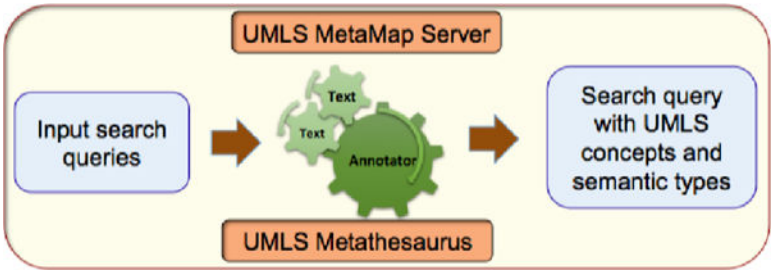

Figure 3.
Functional overview a mappers task: annotation of the search queries with UMLS concepts and semantic types using MetaMap tool.

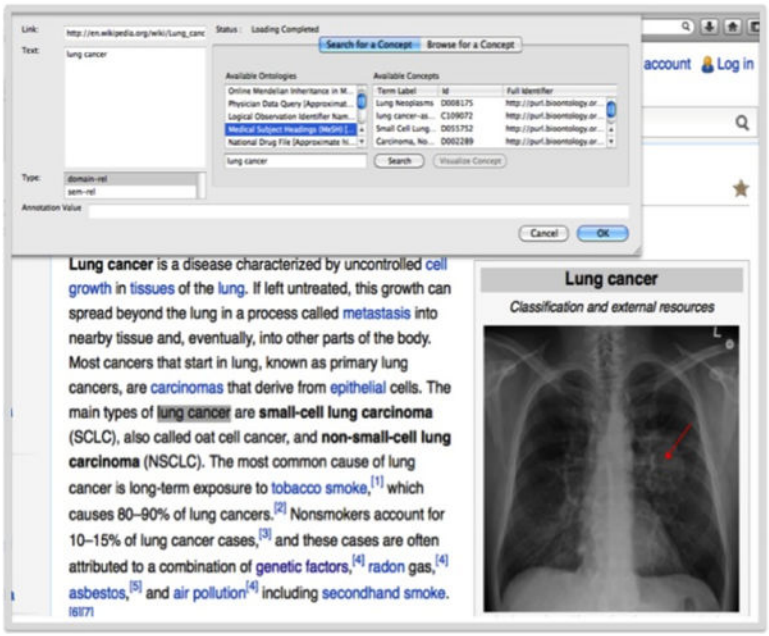

Figure 4.
Annotation of Images with the Concept from Ontology

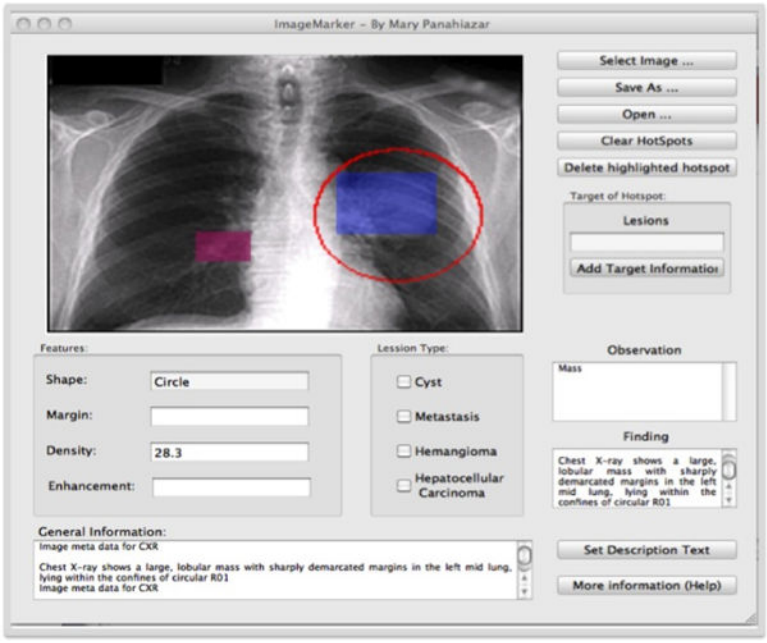

Figure 5.
Image Maker is a tools Implemented for Annotating Images to add Meta Data to the Images