SEBI: An Architecture for Biomedical Image Discovery, Interoperability and Reusability based on Semantic Enrichment

Ahmad C. Bukhari[1], Michael Krauthammer[2], Christopher J.O. Baker[1]

[1] Department of Computer Science and Applied Statistics, University of New Brunswick
Saint John, Canada
[2] Department of Pathology & Yale Center for Medical Informatics, New Haven, USA

sbukhari,bakerc@unb.ca michael.krauthammer@yale.edu

Abstract. Images depicting key findings of research papers contain rich information derived from a wide range of biomedical experiments, e.g. charts, gels, anatomical features, and protein or DNA sequence alignments. Efficient practices for accessing biomedical images are key to allowing the timely transfer of information from the research community to peer investigators and other healthcare practitioners. Searching for images of a certain type is error prone as images are still opaque to information retrieval and knowledge extraction engines due to the absence of explicit descriptions or annotation of the image contents. Moreover, traditional biomedical search engines which search image captions for relevant keywords only offer syntactic search mechanisms without regard for the exact meaning of the query. In order to resolve these challenges and to support interoperability and reusability of biomedical images, we propose a general framework for semantic enrichment of biomedical images called SEBI. SEBI utilizes the information extracted from images as seed data to harvest new annotations from heterogeneous online biomedical resources. The framework incorporates a variety of knowledge infrastructure components and services including image feature extraction, Semantic Web data services, linked open data and the crowd-sourced annotation. Together, these resources make it possible to automatically and/or semi-automatically discover and semantically interlink new information in a manner that supports semantic search for images.
Project Page: https://code.google.com/p/sebi/

Keywords: Semantic Enrichment, Image Interoperability, Semantic Image Discovery and Reuse.

1 Introduction

Due to the advancement of high throughput technologies the last two decades have produced enormous amount of data now known as “big data”. In the life sciences, spreadsheets and databases continue to be the conventional formats used to store experimental data. Some projects have proceeded further and made their data accessible via XML based web services [1], however data persisted in these formats frequently impedes integration and significantly impedes scientific discovery. More recently, researchers have adopted semantic technologies for data integration and manipulation [2, 3] where an explicit data model to describe data in an unambiguous way is used and independently generated data sets can be easily integrated under the same data model [4]. Linked Data [5] has emerged as the most adopted Semantic Web framework supporting data interoperability and reusability. Several life science and health related datasets have been transformed into linked data [6, 7].

In contrast other areas of study that are central to biomedical discovery, such as Biomedical Imaging [8], have not adopted semantic web standards. Several open access biomedical image repositories available on the internet, such as NBIA, NIH Images, and NCI Visuals Online have not published their contents semantically accessible manner. This paper presents semantic enrichment of biomedical images (SEBI) a solution which adopts a combination of Semantic Web technologies to exploit the comprehensive information associated with and contained in, biomedical images. SEBI takes the information extracted from images as seed data to aggregate and harvest new image annotations from heterogeneous biomedical resources and republishes them with semantic annotations so they are readily reusable and can be utilized in adhoc data integration activities.

2 Core Technologies in the SEBI Framework

To provide the target functionality SEBI incorporates a variety of best practice knowledge infrastructure components and services including image feature extraction, Semantic Web data services, linked open data and the crowd-annotation. Web services provide an effective medium for the use of software functionalities in a distributed environment without deploying the entire application on the client machine. SEBI relies on Bioinformatics software and services available on the web, albeit most of them have their unique access criteria and information exchange formats. To get the the full benefit out of these utilities, agile combination of these services is required and results must be available in interoperable format. Semantic Automatic Discovery and Integration (SADI) [9] is a set of best practices which allows the integration of and interoperability among resources on the Web by utilizing Semantic Web standards. SADI Services can be automatically discovered and orchestrated into complex workflows by an intelligent query client. SEBI uses SADI web services for accessing semantic image enrichment annotations from bioinformatics resources.

Linked open data [5] is a set of best practices for publishing and connecting open structured data on the web that was not previously linked. LOD is based on the standard web technologies of HTTP, and URIs, but instead of using them to serve as web pages for human readers, it extends them, using RDF, such that the information is machine readable and can be consumed automatically by machines. In SEBI the annotations received from SADI web services are transformed into linked open data along with their respective images and are kept in triplestore for further activities. SEBI uses source images from the Yale Image Finder (YIF) [10] which is one of the most widely accessed biomedical image search engines. YIF currently holds over two million biomedical images and associated metadata in its index. The data in YIF originate from open source PubMed articles under license from the NLM as XML files. SEBI acquires the YIF image data to build a semantic image repository.

3 SEBI Architecture

This section provides an overview of SEBI architecture which follows a multi-tier system design (Fig. 1) where each tier performs a number of subtasks critical to the image enrichment framework.

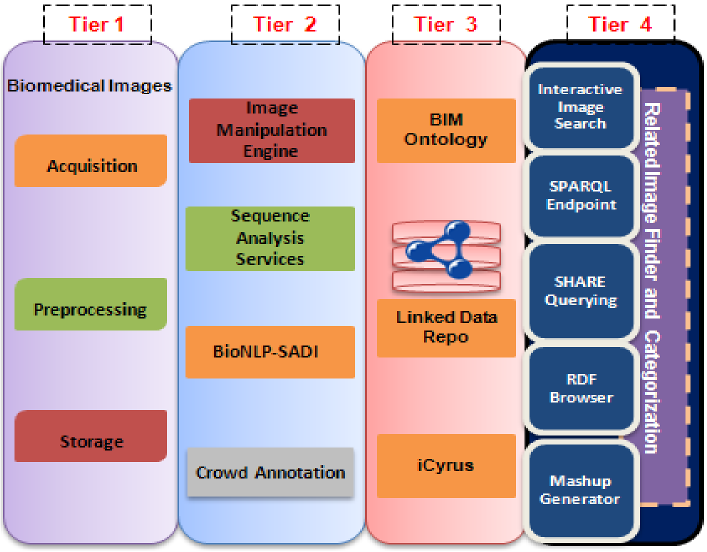

Fig. 1. The Graphical Architecture of SEBI

Tier 1 - Image Data Manipulation: Tier 1 establishes connections with YIF and PubMed concurrently to crosscheck the image metadata. Consequently, the IDM tier resolves data redundancy and populates any missing image metadata. All the relevant metadata (e.g. Title, authors, Pubmed IDs, and year for each article) are extracted and stored in MySQL for a quick lookup, and in a triplestore as parallel storage.

Tier 2 - Image Annotation Generation through SADI: Tier 2 comprises of four sub-modules each of which is responsible for generating a certain type of a sequence image annotation. The image manipulation engine (IME) in tier 2 works as a standalone component that analyses images and classifies them to an image category. The IME algorithm functions both at global and local levels. At the global level, it works to classify sequence images as distinct from the other image categories such as radiographs, gels, or microscopy images [9]. The local function annotates and classifies sequence images into DNA and protein sequence images. Subsequently IME invokes optical character recognition services that process images, applying quality enhancement filters such Grayscale, Gaussian and Laplace filters followed by optical character extraction. Pseudo code of IME algorithm can be viewed in [11].

Biological Sequence Analysis Services: The IME module outputs are used as service inputs to generate sequence image annotations. The SADI sequence analysis service module is designed to retrieve annotations for biological sequences from various biological sequence analysis tools such as *HMMER, BLAST, Pfam, ProSite,* and *GO*.

BioNLP Annotation Generation Module: The *BioNLP* annotation module [2] extracts named entities, such as drug, protein, and lipid names found in the body of a research document where the image originates and/or from the image captions. The *BioNLP* annotation module further normalizes the entities to canonical names defined in online resources e.g. PDB[1] and DrugBank[2] and publishes them in RDF to annotate the images.

Crowd-Annotation Generation Module: Semi-automatic annotation, where automatic annotation is not feasible due to noisy input, is made possible through the introduction of a crowd annotation [12] technique. All images which fail to produce new annotations in the IME and BioNLP modules are automatically ingested. Salient features of the crowd annotation module are that it allows a user to annotate, delete, or update annotations. Users can keep annotations private or share them other registered users.

Tier 3 - The Semantic Enrichment Chamber: Tier 3 is an enrichment chamber holding the Biomedical Image ontology (BIM ontology) [11] that represents the required semantic vocabularies to RDFize annotations generated at tier 2. The BIM ontology provides vocabularies for the sequence image annotation, provenance and crowd annotation modules. The provenance module imports the classes and subclasses from the PAV ontology [13] and interfaces with a text annotation module for text segments. Fig. 2 illustrates the semantic enrichment of an image. All RDFized annotations and image metadata are stored in a dedicated triplestore called iCyrus[3].

Tier 4 - Client Interaction: Tier 4 is a portal through which a user can retrieve images based on semantic annotations generated in tier 3. There are multiple entry points for different users and use cases. For non technical users we provide interactive search based on faceted browsing initiated by keyword input. In addition, a back end *related-image* algorithm (full details of this algorithm will be disclosed in a forthcom-

[1] www.rcsb.org

[2] www.drugbank.ca

[3] cbakerlab.unbsj.ca:8080/icyrus/index1.jsp

ing manuscript) calculates the similarity of images based on their annotations and can link users both to semantically similar images and their corresponding image source documents. Images can also be accessed directly through the *developer friendly* SEBI API providing support for the creation of dynamic image content mashups. Moreover, SEBI's SPARQL endpoint is compatible with semantic federated query clients such as SHARE [14] and Hydra[4] that make use of SPARQL queries for data integration. A complete guide to the use of the API usage is available here[5].

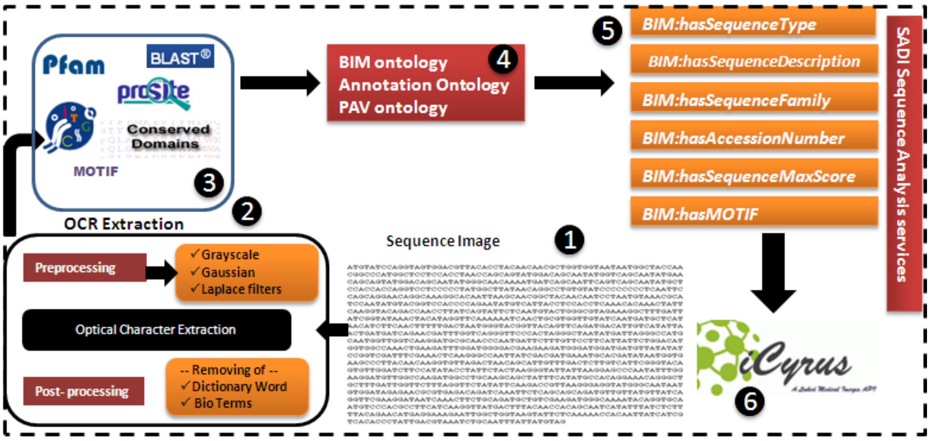

Fig. 2. A Graphical View of Semantic Enrichment of a Sequence Image

Implementation: A modular prototype of the framework is being developed. iCyrus and BioNLP-SADI[6] sub-modules of the system have been released. iCyrus supports technical and non-technical users in developing applications that incorporate biological image data. Unit level testing, prior to a full performance evaluation has been initiated. Currently the *related-image* algorithm is being optimized to achieve the high precision.

4 Conclusions

This paper introduces a framework designed to facilitate biological image discovery and reuse. Core to the framework is the authoring of image annotations that describe specific features of the image content. These annotations are generated using semantic web services following a series of image classification and image pre-processing steps including optical character recognition of biological sequence information in images. Annotations are published in RDF and persisted in triples. The resulting semantically enriched images are readily reusable and can be employed in ad-hoc data integration projects. To augment the discovery of biomedical images we propose the use of an

[4] http://ipsnp.wikidot.com/hydra
[5] https://code.google.com/p/icyrus/wiki/AccessiCyrusThroughJena
[6] http://cbakerlab.unbsj.ca:8080/bionlp-sadi-web-demo/

algorithm to compute image relatedness based on semantic annotations. In addition to proposing a novel use of semantic annotation permitting image discovery and reuse, this research work seeks to lay a foundation for a new paradigm in information retrieval whereby targeted access to the scientific knowledge is mediated primarily through image search, discovery of related images and linking to source publications describing scientific investigations. Instead of seeking scientific literature by entering keywords, users will first query for an image, confirm its relevance to their goals and, based on its relatedness to other images, find other source publications where the related image was first published. This approach to document retrieval and knowledge discovery may in future provide a valuable alternative or addition to current best practices.

References

1. Bhagat, Jiten et al.: BioCatalogue: a universal catalogue of web services for the life sciences. Nucleic acids research. 689-694 (2010).
2. Bukhari, A. C., Klein, A., Baker, C.: Towards Interoperable BioNLP Semantic Web Services Using SADI Framework. Data Integration in the Life Sciences. 69-80 (2013).
3. Nelson, E. K et al.: LabKey Server: an open source platform for scientific data integration, analysis and collaboration. BMC bioinformatics 12.1 (2011).
4. Shadbolt, N., Wendy, H., Berners-Lee, T.: The semantic web revisited. Intelligent Systems, IEEE Vol. 21(3). 96-101 (2006).
5. Bizer, C., Tom, H., Berners-Lee. T. : Linked data-the story so far." International journal on semantic web and information systems. Vol 5(3). 1-22 (2009).
6. Callahan, A. et al.: Bio2RDF release 2: Improved coverage, interoperability and provenance of life science linked data. The Semantic Web: Semantics and Big Data. 200-212 (2013).
7. Bukhari, A.C., Baker, C.: The Canadian health census as Linked Open Data: towards policy making in public health. Data Integration in the Life Sciences. (2013)
8. Webb, A., George, C. K.: Introduction to biomedical imaging. Medical Physics. Vol. 30(8) 2267-2267 (2003)
9. Wilkinson, M., Vandervalk, B., McCarthy, L.: The Semantic Automated Discovery and Integration (SADI) Web service Design-Pattern, API and Reference Implementation. Journal of Biomedical Semantics 2(1), 5–23 (2011).
10. Songhua, X., McCusker, J., Michael K.: Yale Image Finder (YIF): a new search engine for retrieving biomedical images. Bioinformatics Vol 24(17). 1968-1970 (2008).
11. Semantic Enrichment of Biomedical Image, https://code.google.com/p/sebi/
12. Dijkshoorn, C. et al.: Personalization in crowd-driven annotation for cultural heritage collections. UMAP Workshops. (2012)
13. Ciccarese, P. et al.: PAV ontology: provenance, authoring and versioning. Journal of biomedical semantics. Vol 4(1) (2013)
14. Wilkinson, M.D. et al.: SADI, SHARE, and the in silico scientific method. BMC bioinformatics 11.Suppl 12 (2010)