A System for Ontology-Based Annotation of Biomedical Data

Clement Jonquet, Mark A. Musen, and Nigam Shah

Stanford Center for Biomedical Informatics Research
Stanford University School of Medicine
Medical School Office Building, Room X-215
251 Campus Drive, Stanford, CA 94305-5479 USA
{jonquet,musen,nigam}@stanford.edu

Abstract. We present a system for ontology based annotation and indexing of biomedical data; the key functionality of this system is to provide a service that enables users to locate biomedical data resources related to particular ontology concepts. The system's indexing workflow processes the text metadata of diverse resource elements such as gene expression data sets, descriptions of radiology images, clinical-trial reports, and PubMed article abstracts to annotate and index them with concepts from appropriate ontologies. The system enables researchers to search biomedical data sources using ontology concepts. What distinguishes this work from other biomedical search tools is:(i) the use of ontology semantics to expand the initial set of annotations automatically generated by a concept recognition tool; (ii) the unique ability to use almost all publicly available biomedical ontologies in the indexing workflow; (iii) the ability to provide the user with integrated results from different biomedical resource in one place. We discuss the system architecture as well as our experiences during its prototype implementation (http://www.bioontology.org/tools.html).

Keywords: ontology-based annotation, biomedical data integration, biomedical ontologies, semantic expansion, concept recognition.

1 Introduction

The emergence of information and communication technologies has drastically changed biomedical scientific processes. Experimental data and results today are easy to share and repurpose thanks to the Web and public application programming interfaces (APIs) enabling connection to databases containing such information. As a consequence, the variety of biomedical data available in the public domain is now very diverse and ranges from genomic-level high-throughput data to molecular-imaging studies to published research articles. The paradox of such an expansion is that biomedical researchers now face the problem of extracting the specific data they need. Measures must be taken to prevent this problem from worsening as data

A. Bairoch, S. Cohen-Boulakia, and C. Froidevaux (Eds.): DILS 2008, LNBI 5109, pp. 144–152, 2008.
© Springer-Verlag Berlin Heidelberg 2008

repositories grow fast[1]. Biomedical researchers have turned to ontologies and terminologies to describe their data and turn it into structured and formalized knowledge. For instance, the Gene Ontology[2] (GO) is widely used to describe the molecular functions, cellular location and biological processes of gene products as well as integrate these descriptions across several databases.

However, most publicly available biomedical data are unstructured and rarely described with ontology concepts available in the domains. This wealth of publicly accessible biomedical data is beginning to enable cross-cutting integrative translational bioinformatics studies [1][2]. In order to develop integrative translational bioinformatics approaches to interpret these datasets, there is a strong and pressing need to be able to identify all experiments that study a particular disease. A key query dimension for such integrative studies is the sample, along with a gene or protein name. As a result, besides queries that identify all genes that have a function X – which can be reliably answered using GO – we need to conduct queries that find all samples/experiments that study a particular disease and/or the effect of an experimental agent. However, translational discoveries that could be made by mining biomedical resources are hampered because they lack standard terminologies and ontologies to describe their elements (i.e., diagnoses, diseases, samples, and experimental conditions). For example, a researcher studying the allelic variations in a gene would want to know all the pathways that are affected by that gene, the drugs whose effects could be modulated by the allelic variations in the gene, and any disease that could be caused by the gene, and the clinical trials that have studied drugs or diseases related to that gene. The knowledge needed to study such questions is available in public biomedical resources; the problem is finding that information.

The challenge is to create consistent terminology labels for each element in the public resources that would allow the identification of all elements that relate to the same type at a given level of granularity. (e.g., *All carcinoma* samples versus *all Adenocarcinoma in situ of prostate* samples, where the former is at a coarser level of detail). These resource elements range from experimental data sets in repositories, to records of disease associations of gene products in mutation databases, to entries of clinical-trial descriptions, to published papers, and so on. One mechanism of achieving this objective is to map the text metadata describing the diagnoses, pathological state and experimental agents applied to a particular sample to ontology concepts allowing us to formulate refined or coarse search criteria. Creating ontology-based annotations from these resource elements metadata will enable end users to formulate flexible searches for biomedical data [3][4][5][6][7]. Therefore, the key challenge is to automatically and consistently annotate the biomedical data resource elements to identify the biomedical concepts to which they relate.

In this paper, we present a system for ontology-based annotation, which enables users to locate biomedical data related to particular ontology concepts in the BioPortal[3] ontology repository. The system's indexing workflow processes the text

[1] For example, in February 2007, the Gene Expression Omnibus (GEO) had 369 data sets; in the March 2007 release, the number of data sets increased to about 1500 and is now, in February 2008, around 2085 data sets.

[2] www.geneontology.org/

[3] www.bioontology.org/tools/portal/bioportal.html

metadata of several biomedical resource elements to annotate (or tag) them with concepts from appropriate ontologies and create an index to access these elements. As described in the following sections, the tagging is done with a concept recognition tool and the final index takes into accounts the ontology semantics that link concepts to one another (e.g., is_a relation). Our system creates an ontology-based index that can be used by existing search engines (such as Entrez, BioNavigator) to retrieve results that are complementary to the ones found with keyword based approaches. What distinguishes our system is: (i) the use of ontology semantics (ii) the ability to use almost all publicly available biomedical terminologies such as the Unified Medical Language System (UMLS) ontologies as well as Open Biomedical Ontologies, in the indexing workflow; (iii) the ability to provide the user with integrated results from different biomedical resource in one place. In the rest of the paper, Section 2 introduces the system architecture Section 3 gives an example on a GEO dataset. Section 4 presents our implemented prototype and the integration of its results in BioPortal. Section 5 concludes.

2 System Architecture

In this section we describe the system architecture consisting of different levels (Fig. 1). At the *resource level*, public biomedical resources (such as GEO and PubMed) are composed of elements that represent an abstraction for the unit of storage in theses databases. An element is identifiable and can be linked by a specific URL/URI (id), and it has a structure that defines the metadata contexts for the element (title, description, abstract, and so on). Our system retrieves[4] and downloads (through specific access tools) the element text metadata from resources, and keeps a track from both the original metadata context and element id. At the *annotation level*, the system uses a concept recognition tool called mgrep (developed by Univ. of Michigan) to annotate (or tag) resource elements with terms from a dictionary. The dictionary is constructed by including all the concept names and synonyms from a set of ontologies available at the ontology level. The annotation process is context aware, and keeps track of the context (such as title, description) from which the annotation was derived. The results are stored as annotation tables. An annotation table contains information such as "*element E was annotated with concept T in context C*".

At the *index level*, a global index combines all the annotation tables and indexes annotations according to ontology concepts. The index contains information such as: "*Concept T annotates elements E1, E2, ...*".

The system also uses relations provided at the *ontology level* to expand the annotations. This is the first step of the semantic expansion. For example, using the is_a ontology relation, for each annotation, we create additional transitive closure annotations according to the parent–child relationships subsumed by the original concept. For instance, if a resource element such as a GEO protein expression study is annotated with a concept from the ontology National Cancer Institute Thesaurus (NCIT), e.g., *pheochromocytoma*, then a researcher can query for *retroperitoneal neoplasms* and find data sets related to *pheochromocytoma*. The NCIT provides the

[4] We use public API such as Web Services or structured XML documents.

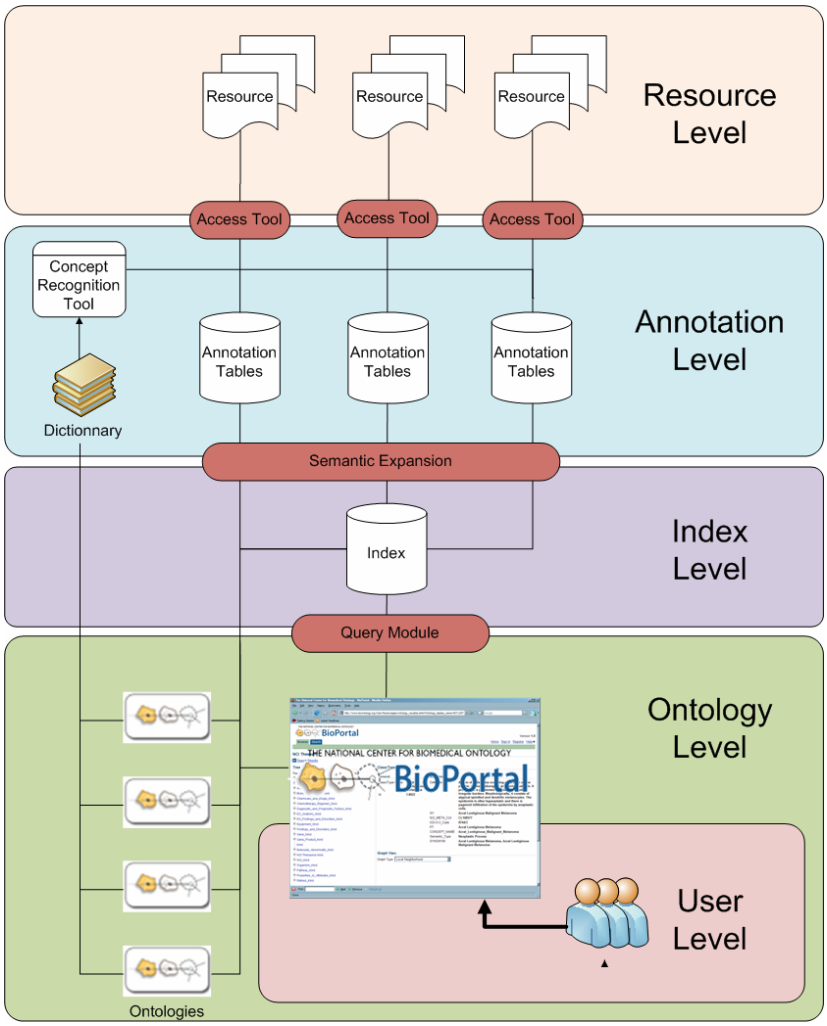

Fig. 1. The system architecture comprising of different levels. See main text for details.

knowledge that *pheochromocytoma* is_a *retroperitoneal neoplasms*. This first step is done offline because, processing the transitive closure is very time consuming – even if we use a pre-computed hierarchy – and will result in prolonged response times for the users. This use case is similar, in principle, to query expansion done by search engine like Entrez; however, Entrez does not use ontologies, therefore, there exists *pheochromocytoma* related GEO data sets, but none show up on searching for retroperitoneal neoplasms in Entrez. In our system, however, a researcher could search for *retroperitoneal neoplasms* and find the relevant samples [1].

At the *user level*, on searching for a specific ontology concept, the results provide resource elements found directly or via the step of semantic expansion. A query module performs the second step of semantic expansion i.e., expanding the user query using the knowledge ontologies provide. This module also selects and filters the appropriate annotations according to the user choices transmitted by the user interface. The semantic expansion is therefore be done both off line (e.g., such as with the is_a transitive closure) or at run time, interacting with the user and using other techniques [8], such as semantic distance [9][10]. The user receives the result in terms of references and links (URL/URI) to the original resource elements.

Remark: This architecture illustrates the generalizability of our implementation. Note the same model could be applied for domains other than biomedical informatics. The only specific components of the system are the resource access tools (which are customized for each resource) and, of course, the ontologies.

3 Example Demonstrating the Processing of a GEO Dataset

A GEO dataset represents a collection of biologically – and statistically – comparable samples processed using the same platform. We treat each GEO dataset as a resource element whose metadata we aim to process. Each GEO dataset, has a title and a summary context that contain free text metadata entered by the person creating the dataset. Consider for example the GEO dataset 'GDS1989'. This dataset is available online[5] and can be retrieved using the EUtils API.[6] GDS1989's title is: *Melanoma progression.* GDS1989's summary contains the phrase: *melanoma in situ*. Our set of ontology contains, for instance, the Human disease ontology,[7] and the concept *Melanoma* is in our system's dictionary as it is one possible term for the concept DOID:1909 in this ontology. Therefore, our concept recognition tool produces the following annotations:[8]

Element GDS1989 annotated with concept DOID:1909 in context title;
Element GDS1989 annotated with concept DOID:1909 in context summary;

The structure of the Human disease ontology shows that DOID:1909 has 36 direct or indirect parents such as for instance DOID:169, *Neuroendocrine Tumors* and DOID:4, *Disease*, therefore the transitive closure on the is_a relation generates, for instance, the following annotations:

Element GDS1989 annotated with concept DOID: 169 with closure;
Element GDS1989 annotated with concept DOID:4 with closure;

Searching for "melanoma" in BioPortal returns 109 matches[9] in the Human disease ontology including concept DOID1909. The user can access the 13 ArrayExpress

[5] www.ncbi.nlm.nih.gov/projects/geo/gds/gds_browse.cgi?gds=1989
[6] www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html
[7] http://diseaseontology.sourceforge.net/
[8] Note these two annotations involve only one annotating concept.
[9] BioPortal uses an Apache Lucene index provided by LexGrid (http://informatics.mayo.edu) to find the query related ontology concepts.

experiments, or the 673 clinical trials, or the 960 articles in PubMed and the 10 GEO datasets related to that concept.

4 Integration with NCBO BioPortal

The National Center for Biomedical Ontology (NCBO) [11] develops and maintains a Web application called BioPortal to access biomedical ontologies. This library contains a large collection of ontologies, such as GO, NCIT, International Classification of Diseases (ICD), in different formats (OBO, OWL, etc.). Users can browse and search this repository of ontologies both online and via a Web services API.

We have implemented the first prototype of the system as presented in section 2. We have written a set of Java access tools to access five resource databases. Resources processed and the numbers of annotations currently available in our system index are presented in Table 1. A public representational state transfer (REST) services API [12] is available to query the annotation index and returns XML documents describing the annotations. We have used this API to integrate the system with BioPortal as illustrated by Fig. 2.

In our prototype, we have processed: (1) high-throughput gene-expression data sets from GEO and Array Express, (2) clinical-trial descriptions from Clinicaltrials.gov, (3) captions of images from ARRS Goldminer, and (4) abstracts of articles published in PubMed. Table 1 shows both the current number of elements annotated and the number of annotations created from each resource that we have processed. Our prototype uses 48 different biomedical ontologies that give us a dictionary of 793681 unique concepts and 2130700 terms. As a result of using such a large number of terms, our system provides annotations for 99% of our subset of PubMed, and 100% of the other processed resources. The average number of annotating concepts is between 359 and 769 per element, with an average of 27% of these annotations being direct. In the current prototype, concept recognition is done using a concept recognition tool developed by National Center for Integrative Biomedical Informatics

Table 1. Number of elements annotated from each resource in the current prototype

Resource	Number of elements	Resource local size (Mb)	Number of direct annotations (mgrep results)	Total number of 'useful'[1] annotations	Average number of annotating concepts
PubMed (subset) www.ncbi.nlm.nih.gov/PubMed/	1050000	146.1	30822190	174840027	160
ArrayExpress www.ebi.ac.uk/arrayexpress/	3371	3.6	502122	1849224	525
ClinicalTrials.gov http://clinicaltrials.gov/	50303	99	16108580	48796501	824
Gene Expression Omnibus www.ncbi.nlm.nih.gov/geo/	2085	0.7	165539	772608	359
ARRS GoldMiner (subset) http://goldminer.arrs.org	1155	0.5	134229	662687	564
TOTAL	1106914	249.9	47732660	226921047	(avg)486.4

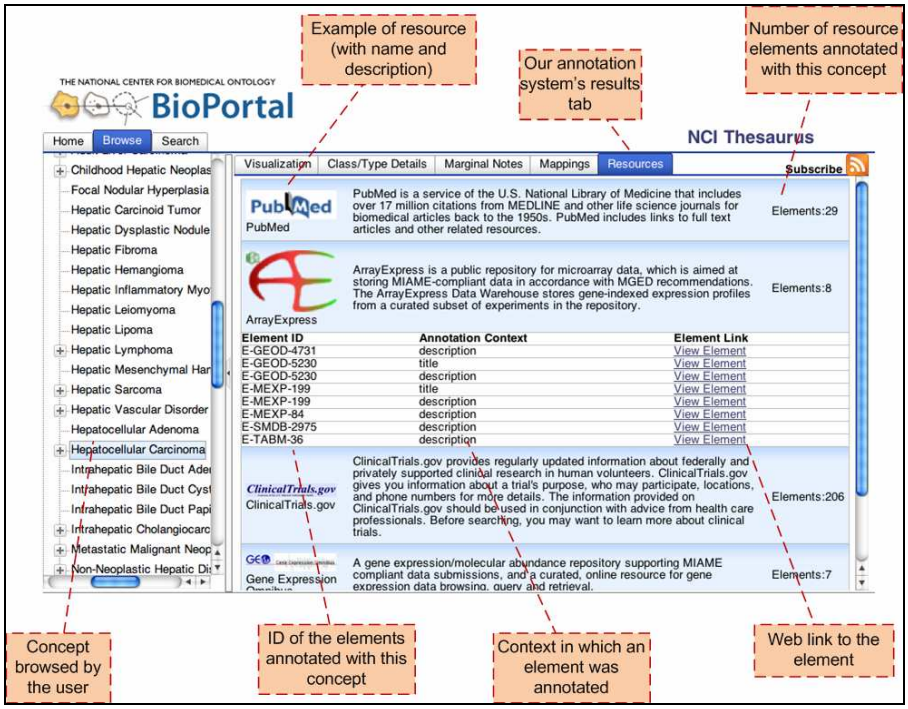

Fig. 2. User interface within BioPortal. In this view, a user browsing the NCIT in BioPortal, can select an ontology concept (in this case, *Hepatocellular carcinoma*) and see immediately the numbers of online resource elements that relate directly to that concept (and the concepts that it subsumes). The interface allows the user to directly access the original elements that are associated with *Hepatocellular carcinoma* for each of the indexed resources.

(NCIBI) called *mgrep*. [10]We rely on this tool which reported a very high degree of accuracy (over 95%) in recognizing disease names [13]. The prototype design of the annotation level is such that we can plug-in other concept recognizers. The prototype is available online http://alpha.bioontology.org/.

5 Conclusion

In this paper, we have described the prototype implementation of an ontology-based annotation system. The system's objective is to annotate (offline) a large number of biomedical resources and to provide an index up to date of annotated resources elements. We use ontologies (and not simply terminologies) both for annotation as

[10] We have conducted a comparative evaluation of this tool with the gold standard in the biomedical community, MetaMap [14]. It has a higher precision in recognizing concepts, and it is more scalable as well as open to outside dictionary (not tied to the UMLS structure as MetaMap is.).

well as semantic expansion of the annotations. The NCBO hosts one of the largest library of biomedical ontologies and our system allows a user to search for various biomedical data related to a specific ontology concepts in one place; greatly enhancing the value of the ontology repository. Our system can process text metadata of gene-expression data sets, descriptions of radiology images, clinical-trial reports, as well as abstracts of PubMed articles to annotate them automatically with concepts from appropriate ontologies. It promotes biomedical translational research by enabling users to locate relevant biological data sets and to integrate them with clinical data to bridge the bench-to-bedside gap.

We believe that as we expand the system with additional ontologies and process additional biomedical resources, we will serve an even wider user population, broadening the reach and impact of the NCBO in enabling translational research.

Acknowledgements

This work is supported by the National Center for Biomedical Computing (NCBC) National Institute of Health roadmap initiative; NIH grant U54 HG004028. We also acknowledge assistance of Manhong Dai and Fan Meng at University of Michigan as well as Chuck Kahn for the access to the Goldminer resource.

References

[1] Butte, A.J., Kohane, I.: Creation and implications of a phenome-genome network. Nature Biotechnology 24(1), 55–62 (2006)

[2] Butte, A., Chen, R.: Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. In: American Medical Informatics Association Annual Symposium, AMIA 2006, Washington DC, USA, p. 106 (2006)

[3] Spasic, I., et al.: Text mining and ontologies in biomedicine: making sense of raw text. Brief Bioinformatics 6(3), 239 (2005)

[4] Moskovitch, R., Martins, S.B., Behiri, E., Weiss, A., Shahar, Y.: A Comparative Evaluation of Full-text, Concept-based, and Context-sensitive Search. American Medical Informatics Association 14(2), 164–174 (2007)

[5] Sneiderman, C.A., et al.: Knowledge-based Methods to Help Clinicians Find Answers in Medline. American Medical Informatics Association 14(6), 772–780 (2007)

[6] Shah, N.H., Rubin, D.L., Supekar, K.S., Musen, M.A.: Ontology-based Annotation and Query of Tissue Microarray Data. In: American Medical Informatics Association Annual Symposium, AMIA 2006, Washington DC, USA, pp. 709–713 (2006)

[7] Khelif, K., Dieng-Kuntz, R., Barbry, P.: An ontology-based approach to support text mining and information retrieval in the biological domain. Universal Computer Science 13(12), 1881–1907 (2007), Special Issue on Ontologies and their Applications

[8] Bhogal, J., Macfarlane, A., Smith, P.: A review of ontology based query expansion. Information Processing and Management 43, 866–886 (2007)

[9] Lee, W.J., Raschid, L., Srinivasan, P., Shah, N., Rubin, D., Noy, N.: Using Annotations from Controlled Vocabularies to Find Meaningful Associations. In: Cohen-Boulakia, S., Tannen, V. (eds.) DILS 2007. LNCS (LNBI), vol. 4544, pp. 264–279. Springer, Heidelberg (2007)

[10] Caviedesa, J.E., Cimino, J.J.: Towards the development of a conceptual distance metric for the UMLS. Biomedical Informatics 37(2), 77–85 (2004)
[11] Ashburner, M., Sim, I., Hute, C.G., Solbrig, H., Storey, M.A., Smith, B., Day-Richter, J., Noy, N.F., Musen, M.A.: National Center for Biomedical Ontology: Advancing Biomedicine through Structured Organization of Scientific Knowledge. OMICS A Journal of Integrative Biology 10(2), 185–198 (2006)
[12] Fielding, R.T., Taylor, R.N.: Principled Design of the Modern Web Architecture. ACM Transactions on Internet Technology 2(2), 115–150 (2002)
[13] Xuan, W., Dai, M., Mirel, B., Athey, B., Watson, S., Meng, F.: Interactive Medline Search Engine Utilizing Biomedical Concepts and Data Integration. In: BioLINK SIG: Linking Literature, Information and Knowledge for Biology, Vienna, Austria (July 2007)
[14] Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: American Medical Informatics Association Annual Symposium, AMIA 2001, Washington DC, USA, pp. 17–21 (2001)