NCBO *Resource Index*: Ontology-Based Search and Mining of Biomedical Resources

Clement Jonquet, Paea LePendu, Sean M. Falconer, Adrien Coulet,
Natalya F. Noy, Mark A. Musen, and Nigam H. Shah

Stanford Center for Biomedical Informatics Research, Stanford University, US
{jonquet,plependu,sfalc,coulet,noy,musen,nigam}@stanford.edu

Abstract. The volume of publicly available data in biomedicine is constantly increasing. However, this data is stored in different formats on different platforms. Integrating this data will enable us to facilitate the pace of medical discoveries by providing scientists with a unified view of this diverse information. Under the auspices of the National Center for Biomedical Ontology, we have developed the *Resource Index*—a growing, large-scale index of more than twenty diverse biomedical resources. The resources include heterogeneous data from a variety of repositories maintained by different researchers from around the world. Furthermore, we use a set of 200 publicly available ontologies, also contributed by researchers in various domains, to annotate and to aggregate these descriptions. We use the semantics that the ontologies encode, such as different properties of classes, the class hierarchies, and the mappings between ontologies in order to improve the search experience for the *Resource Index* user. Our user interface enables scientists to search the multiple resources quickly and efficiently using domain terms, without even being aware that there is semantics under the hood.

1 The Diversity of Public Biomedical Resources

Researchers in biomedicine produce and publish an enormous amount of structured data describing everything from genomic information and pathways to drug descriptions, clinical trials, and diseases. This information is stored on many different Web sites, using idiosyncratic schemas and access mechanisms. Our goal is to enable a researcher to browse and analyze the information stored in these diverse resources. Then, for instance, a researcher studying allelic variations in a gene can find all the pathways that the gene affects, the drug effects that these variations modulate, 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 that we need in order to address such questions is available in public biomedical resources; the problem is finding that information.

The National Center for Biomedical Ontology (NCBO) maintains BioPortal, an open library of more than 200 ontologies in biomedicine [3].[1] We use the terms from these ontologies to annotate, or "tag," automatically the textual descriptions of the data that resides in diverse public resources. In our context, a biomedical *resource* is a repository of elements that may contain patient records, gene expression data, scholarly articles, and so on. A *data element* is unstructured text describing elements in the resource.

[1] http://bioportal.bioontology.org

Resource	Contains	Maintained By	Data Elements
Gene Expression Omnibus (GEO)	gene expression and molecular abundance data	National Center for Biotechnology Information	21,272
ArrayExpress (AE)	microarray data and gene-indexed expression profiles	European Bioinformatics Institute	15,190
caNanoLab (caNano)	biomedical nanotechnology research results	National Cancer Institute's cancer Nanotechnology Lab	800
Adverse Event Reporting System Data (AERS)	adverse events reported to FDA by doctors and other professionals	AersData.org	1,172,881
Clinical Trials (CT)	reports on clinical research in human volunteers	ClinicalTrials.gov	96,338
Research Crossroads (RXRD)	medical funding data	ResearchCrossroads.org	1,033,651
UniProt KB (UPKB)	protein sequence and functional information	UniProt.org	18,324
PubMed (PM)	citations and text for biomedical literature	National Library of Medicine	19,000,000

Table 1. A sample of resources that we include in the *Resource Index*. These resources are diverse and are maintained by many different groups. Please refer to `http://rest.bioontology.org/resource_index/resources/list/` for a complete listing.

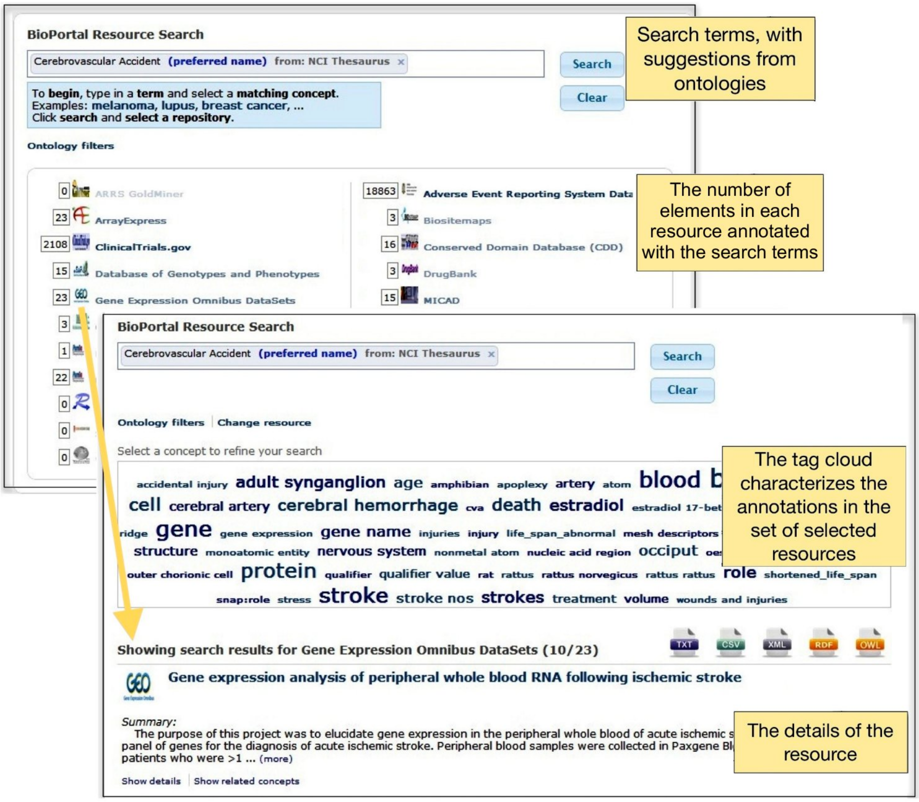
An *annotation*—a central component—is a link from an ontology term to a data element, indicating that the data element refers to the term. We then use these annotations to "bring together" the data elements from these resources.

The *Resource Index* contains a growing set of data that are maintained originally by a variety of different institutions (Table 1). The user interface is available at `http://bioportal.bioontology.org/resource_index_ui`. We currently index 22 resources using a 1.5Tb MySQL database, containing 16.4 billion annotations, 2.4 million ontology terms, and 3.5 million data elements.

2 Use Case Scenarios

We will describe the functionality of the *Resource Index* through two use case scenarios.

Scenario 1: A researcher studying the causes and treatments for strokes in humans is interested in learning more about the genetic basis of the response to such conditions by searching the literature. She already knows that some related conditions such as stroke, transient ishaemic attack, and cerebral bleeding fall under the general category of cerebrovascular accidents (Figure 1). Therefore, she starts by entering "cerebr-" and immediately gets feedback in the form of suggested terms from various ontologies. She selects and initiates a search for *Cerebrovascular Accident* from the NCI Thesaurus. She notices a number of hits from several resources and drills down to read more about the data from both the Gene Expression Omnibus (GEO) and the Database of Genotypes and Phenotypes (dbGaP) resources. Interestingly, although both of these datasets are maintained at the National Center for Biotechnology Information (NCBI), NCBI does not link them together.



BioPortal Resource Search

Cerebrovascular Accident (preferred name) from: NCI Thesaurus ×

To **begin**, type in a **term** and select a **matching concept**.
Examples: melanoma, lupus, breast cancer, ...
Click **search** and **select a repository**.

Ontology filters

0	ARRS GoldMiner	18863	Adverse Event Reporting System Data
23	ArrayExpress	3	Biositemaps
2108	ClinicalTrials.gov	16	Conserved Domain Database (CDD)
15	Database of Genotypes and Phenotypes	3	DrugBank
23	Gene Expression Omnibus DataSets	15	MICAD
3			
1			
22			
0			
0			
0			

Search terms, with suggestions from ontologies

The number of elements in each resource annotated with the search terms

BioPortal Resource Search

Cerebrovascular Accident (preferred name) from: NCI Thesaurus ×

Select a concept to refine your search

The tag cloud characterizes the annotations in the set of selected resources

Showing search results for Gene Expression Omnibus DataSets (10/23)

Gene expression analysis of peripheral whole blood RNA following ischemic stroke

Summary:
The purpose of this project was to elucidate gene expression in the peripheral whole blood of acute ischemic s
panel of genes for the diagnosis of acute ischemic stroke. Peripheral blood samples were collected in Paxgene Bl
patients who were >1 ... (more)

The details of the resource

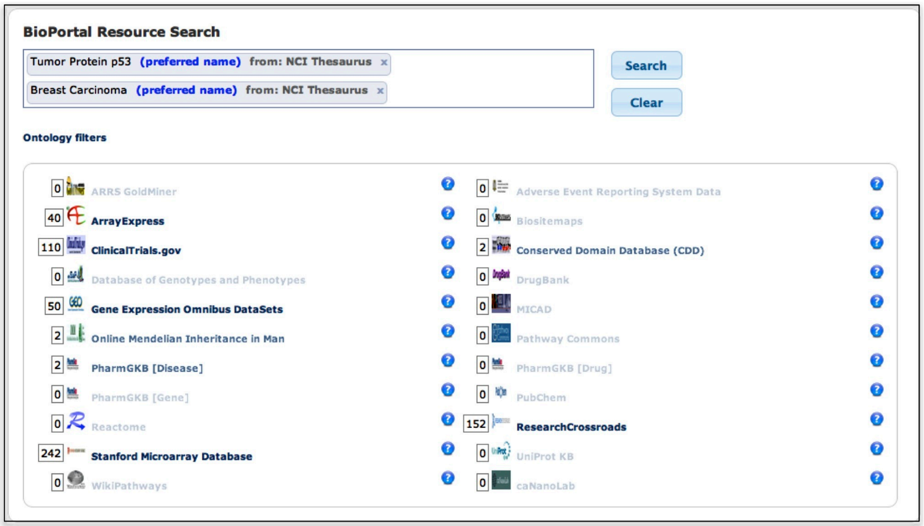
Fig. 1. Searching the *Resource Index*. The user searches for resources on “cerebrovascular accidents” and finds gene-expression data from rats and humans that are relevant to different types of cerebrovascular accidents, such as stroke and cerebral bleeding.

She focuses on GEO: the tag cloud emphasizes other terms that are ranked highly in these 23 elements. Thus, she can get an idea of what these datasets are about. She selects “Stroke” in the tag cloud and gets to the datasets that are tagged with “Cerebrovascular Accident,” “Stroke,” and “Treatment.” A similar series of steps on dbGaP leads her to two elements tagged with “Cerebrovascular Accident,” “Stroke,” and “Physiology.”

As a result of her search, she has quickly located gene-expression data (from rats) that is connected to genotype-phenotype data (from humans). In rats, researchers studied the gene-expression level response to both stroke and to drugs used to treat stroke. In humans, researchers studied genotypes that predispose humans to stroke and affect the physiology of the outcome. Thus, the user has satisfied her goal: she learned more about the genetic basis for strokes, some treatments, and outcomes.

Scenario 2: Suppose the user is interested in the role of tumor protein p53 in breast cancer. He can search the *Resource Index* for “Tumor Protein p53” AND “Breast Carcinoma” (Figure 2). The search results summarize resources annotated with both terms.

The user can see that there is relevant data linking p53 to breast cancer in such resources as AE, ClinicalTrials.org, GEO, Stanford Microarray Database, AERS, Re-



BioPortal Resource Search

Tumor Protein p53 (preferred name) from: NCI Thesaurus ×

Breast Carcinoma (preferred name) from: NCI Thesaurus ×

Ontology filters

0	ARRS GoldMiner	0	Adverse Event Reporting System Data
40	ArrayExpress	0	Biositemaps
110	ClinicalTrials.gov	2	Conserved Domain Database (CDD)
0	Database of Genotypes and Phenotypes	0	DrugBank
50	Gene Expression Omnibus DataSets	0	MICAD
2	Online Mendelian Inheritance in Man	0	Pathway Commons
2	PharmGKB [Disease]	0	PharmGKB [Drug]
0	PharmGKB [Gene]	0	PubChem
0	Reactome	152	ResearchCrossroads
242	Stanford Microarray Database	0	UniProt KB
0	WikiPathways	0	caNanoLab

Fig. 2. The search for resources that contain both “Tumor Protein p53” AND “Breast Carcinoma.”

search Crossroads, and others. He can access the data within each resource quickly and navigate between resources easily.

3 Implementation Details

We use the public REST services that NCBO BioPortal provides to create a *dictionary* of terms to use for direct annotations. Currently, the dictionary contains 4.1 million entries, which are based on the 2.4 million classes plus any synonyms that are specified in the ontology using SKOS properties or other user-specified properties.

To access the information in the remote resources, we build a custom *wrapper* for each resource with the help of a subject matter expert on the resource. The wrapper extracts the text fields describing the data elements within a resource and records the context. We use the context later in scoring the annotations: For example, we may give annotations appearing in the title a higher weight based on an expert’s recommendation for that resource. In some cases, resources already tag elements with ontology terms, so the wrapper directly extracts the curated annotation and applies an appropriate weight.

After we have the term dictionary and the set of textual descriptions for the data elements, we create *direct annotations* from the text of a resource element using an off-the-shelf concept recognition tool, which in our case is MGREP [7]. Then we use subclass relations to traverse ontology hierarchies to create new, *expanded annotations*: If a data element has a direct annotation with a particular class, we add an annotation with every superclass of that class to that data element. We preserve the information on the distance between the class in the direct annotation and the parent, using this distance as one of our scoring components.

We have conducted a comparative evaluation of two concept recognizers—the University of Michigan’s MGREP and the National Library of Medicine’s MetaMap—and found that MGREP has clear advantages in large-scale service oriented applications [5]. The precision varies depending on the text in each resource and type of entity being recognized: from 93% for recognizing biological processes in descriptions of gene expression experiments to 60% in Clinical Trials, or from 88% for recognizing disease terms in descriptions of gene expression experiments to 23% for Medline abstracts [4, 6]. The average precision is approximately 85%, average recall is 78%. In a separate evaluation, in which we used a manually curated gold-standard for diabetes data in the GEO resource, we found that the expanded annotations increases recall by as much as 27% (from 34% to 43.2%).

When the user searches the *Resource Index*, we use the BioPortal search service for autocomplete. We then use the *Resource Index* database to expand the search results using the precomputed class hierarchy and ontology mappings. We rank the results based on several factors: (1) the field in which the annotation appears; (2) whether the annotation corresponds to the preferred name or a synonym; and (3) whether the annotation is a direct annotation or it was inferred by traversing the hierarchy or by using ontology mappings.

All of the *Resource Index* data is also programmatically accessible to developers through Web services, which they can use to mine the data for knowledge discovery (`http://www.bioontology.org/wiki/index.php/NCBO_REST_services`).

4 The Use of Semantics in the *Resource Index*

The *Resource Index* leverages the semantics in the ontologies in several different ways:

Synonyms: Many biomedical ontologies specify not only labels (preferred names) but also synonyms for the class names, which we use during annotation. For example, a keyword search of caNanoLab resource with “adriamycin” would normally obtain no results. However, because the ontologies we use have defined “doxorubicin” as a synonym for “adriamycin,” the *Resource Index* retrieves all caNanoLab elements annotated with the term “doxorubicin.” We assign slightly lower scores when using synonyms.

Auto-complete: As the user types a term into the search box, they receive immediate feedback giving both preferred names and synonyms for matching classes from different ontologies.

Hierarchies: We use subclass relations to traverse ontology hierarchies to create additional, expanded annotations, which improves the recall of search on general terms. For example, a keyword search in GEO with “retroperitoneal neoplasm,” will obtain no results. However, the same search via the *Resource Index* retrieves data annotated with its child term in NCI Thesaurus, “pheochromocytoma.” We give lower scores to annotations that are increasingly distant from a direct annotation.

Mappings: BioPortal stores mappings between classes in different ontologies. We use these mappings to expand the set of annotations as well. For example, a search with the concept “treatment” from MeSH retrieves the elements annotated with “therapeutic procedure” in SNOMED-CT because there is a mapping between these two concepts.

5 Challenges and Future Plans

We are currently working on expanding the *Resource Index* to include more resources. Our goal is to index up to 100 public resources, including PubMed, which provides access to all research articles in biomedicine. Naturally, we have already encountered scalability challenges, with our original workflow taking too long to process each resource. We have analyzed the metrics on ontologies in order to re-structure the database; this restructuring has enabled us to reduce the processing time for one of our larger datasets from one week to one hour [2]. With this type of performance, we can now annotate extremely large datasets such as PubMed, which contain over 19 million records. We have already indexed the last five years of it (20%).

One limiting factor in increasing the number of resources that we index is the need to develop custom access tools for most resources. However, most resource access tools follow the same principles, so we have built templates that enable our collaborators to build them easily and quickly to process their own datasets and to include them in the *Resource Index*.

The evolution of ontologies and resources is another key challenge that we must address. There are three types of updates that we must deal with: (1) new data elements to resources; (2) new versions of ontologies used for indexing; (3) new ontologies. We have used ontology metrics to optimize the index update [2] for the latter two cases. Given the scale of the index, however, some elements of evolution still remain challenging and we are currently addressing them.

6 Conclusion

The *Resource Index* enables domain experts to search heterogeneous, independently developed resources. While we use ontologies and semantics “under the hood” to improve the quality of the results and to simplify the user interaction, the users are not aware of this complexity. They use a simple search-box interface and can drill down on the specific resources that contain their terms of interest or any other relevant terms.

Acknowledgments: This work was supported in part by the National Center for Biomedical Ontology, under roadmap-initiative grant U54 HG004028 from the National Institutes of Health.

References

1. A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *AMIA Annual Symposium*, pages 17–21, Washington, DC, USA, 2001.
2. P. LePendu, N. F. Noy, C. Jonquet, P. R. Alexander, N. H. Shah, and M. A. Musen. Optimize first, buy later: Analyzing metrics to ramp-up very large knowledge bases. In *9th International Semantic Web Conference (ISWC 2010)*, Shanghai, China, 2010. Springer.
3. N. F. Noy, et. al. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.*, 37((Web server)):170–173, 2009.
4. I. N. Sarkar. Leveraging Biomedical Ontologies and Annotation Services to Organize Microbiome Data from Mammalian Hosts. In *AMIA Annual Symposium,*, Washington DC, 2010.
5. N. H. Shah, N. Bhatia, C. Jonquet, D. L. Rubin, A. P. Chiang, and M. A. Musen. Comparison of concept recognizers for building the Open Biomedical Annotator. *BMC Bioinformatics*, 10(9:S14), September 2009.
6. J. S. Simon N. Twigger, Joey Geiger. Using the NCBO Web Services for Concept Recognition and Ontology Annotation of Expression Datasets. In *Workshop on Semantic Web Applications and Tools for Life Sciences, SWAT4LS’09*, Amsterdam, the Netherlands, 2009.
7. W. Xuan, M. Dai, B. Mirel, B. Athey, S. J. Watson, and F. Meng. Interactive Medline Search Engine Utilizing Biomedical Concepts and Data Integration. In *BioLINK: Linking Literature, Information and Knowledge for Biology, SIG, ISMB’08*, pages 55–58, Vienna, Austria, 2007.

Appendix: Compliance with Semantic Web Challenge Requirements

Compliance with the Semantic Web Challenge minimal requirements

✓ *The application has to be an end-user application.*

The *Resource Index* has an easy-to-use interface geared towards biomedical researchers: `http://bioportal.bioontology.org/resource_index_ui`. The researchers do not even need to be aware that semantic technologies are driving the user interface, and can use it through a simple keyword-search mechanism.

✓ *The information sources used should be under diverse ownership or control.*

The 22 resources in the *Resource Index* are all under different ownership and control, including resources from different countries. None of the resources are under the control of NCBO itself. The ontologies in BioPortal that we use for annotations also come from many different sources, none of which are under NCBO control. Refer to `http://bioportal.bioontology.org/ontologies` and `http://rest.bioontology.org/resource_index/resources/list/` for the list of ontologies and resources.

✓ *The information sources used should be heterogeneous.*

Data resources in the *Resource Index* are always in different format (often XML) and offer different means of access (mostly Web services).

✓ *The information sources used should contain substantial quantities of real world data.*

All data in the *Resource Index* are real world data produced in the biomedical domain. The system indexes 3.5 million elements (for a total size of 3.7Gb of text data) across 22 resources by using 16.4 billion annotations. The complete database size with indexes is around 1.5Tb.

✓ *The meaning of data has to play a central role. Meaning must be represented using Semantic Web technologies.*

Section 4 highlights the semantic technologies that we used. We use ontologies as sources of terms for entity recognition, extracting class descriptions (preferred names and synonyms). We use the class hierarchies and ontology mappings to expand the search results.

✓ *Data must be manipulated/processed in interesting ways to derive useful information. This semantic information processing has to play a central role in achieving things that alternative technologies cannot do as well, or at all.*

Section 2 illustrates the use cases and value added of our system. Examples of possible new queries not possible otherwise are also given.

Compliance with the Semantic Web Challenge additional features

✓ *The application provides an attractive and functional Web interface (for human users).*

The *Resource Index* interface uses a simple search box and auto-complete for user input. The view changes interactively. There is a tag cloud to highlight important terms. The user can quickly get to the details for the element of interest.

✓ *The application should be scalable.*

The *Resource Index* has 1.5Tb of data, including 16.4 billion annotations that use 2.4 million classes from 200 ontologies. It provides access to 22 distributed, independently developed, and heterogeneous resources. We continue to add resources, after optimizing the *Resource Index* [2].

✓ *Rigorous evaluations have taken place that demonstrate the benefits of semantic technologies, or validate the results obtained.*

We have evaluated the accuracy of the indexing workflow: average precision is 85%, recall is 78%. We have conducted a comparative evaluation of MGREP with MetaMap [1, 5], the gold-standard in biomedicine. Different Annotator use cases confirmed similar results [6, 4].

✓ *Novelty in applying semantic technology to a domain or task.*

Ontology-based indexing is not new in biomedicine, however it is usually restricted to indexing a specific resource with a specific ontology (vertical approach). We adopt a horizontal approach, accessing annotations for many important resources and a very high number of ontologies.

✓ *Functionality is different from or goes beyond pure information retrieval.*

Use cases and examples presented in section 2 show how the *Resource Index* search goes beyond traditional information retrieval, by using the semantics that is encoded in the ontologies.

✓ *The application has clear commercial potential and/or large existing user base.*

The Healthline search engine—adopted by Yahoo!—uses similar semantic technologies for searching health-related information and generates profit. Unlike Heathline, the ontologies and data for the *Resource Index* are publicly available. When we apply our workflow to electronic medical records we expect transformative, revenue-generating commercial applications as well.

✓ *Contextual information is used for ratings or rankings.*

We use the context of where the annotated terms appear in the original resource (title, description, etc.) in scoring annotation results.

Multimedia documents are used in some way.

We index only textual data for the moment. However, we can extend the approach to other kinds of documents (i.e., image, sounds) by changing the tools that we use for entity recognition.

There is a use of dynamic data, perhaps in combination with static information.

There is no use of dynamic data such as workflow or data streams.

✓ *The results should be as accurate as possible (e.g. ranking of results according to context).*

We rank the annotations based on several parameters: the context in the original data element, the distance between the ontology terms, and the use of mappings.

There is support for multiple languages and accessibility on a range of devices.

We currently use an entity-recognizer for English. However, as BioPortal now contains ontologies in multiple languages, we can start using entity recognizers for other languages.