Ontrez Project Report
National Center for Biomedical Ontology
November, 2007

Executive summary

Currently, genomics data and data repositories in the public domain are expanding at an explosive pace.[1] The wealth of publicly accessible data is beginning to enable cross-cutting integrative translational bioinformatics studies [1]. However, translational discoveries that could be made by mining such public resources are hampered if they lack standard terminologies to describe diagnoses, diseases, and experimental conditions. For discovery to proceed in the eras of e-science, researchers need tools to enable them to find all the data sets relevant to their area of study—spanning the biological scales from molecular studies to clinical medicine—and bridging the research modalities from high-throughput experiments to clinical trials and medical imaging. 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 data sets; the challenge is finding that information.

The key challenge common to all the needs outlined above is **to *annotate* the various resource elements consistently to identify the biomedical concepts to which they relate.** These resource elements range from experimental data sets in public repositories, to records of disease associations of gene products in mutation databases, to entries of clinical-trial descriptions. Creating ontology-based annotations from the metadata in biomedical resources and identifying diagnoses, pathological states, and experimental agents contained in those resources allows indexing of the resources, enabling end users to formulate flexible searches for biomedical data [2-6].

In the past two months, we have been developing a system that is integrated with BioPortal known as Ontrez.[2] Ontrez enables researchers to search for biomedical data (such as genomic data sets, medical images, clinical trials and published papers). Ontrez promotes translational research by enabling researchers to locate relevant biological data sets and to integrate them with clinical data to bridge the bench-to-bedside gap.

Ontrez processes the metadata-annotations of gene expression data sets, descriptions of radiology images, clinical-trial reports, as well as abstracts of Pubmed articles to annotate (or tag) them with terms from appropriate ontologies.

[1] For example, the Gene Expression Omnibus (GEO) had 369 data sets in February 2007; in the March release, the number of data sets increased to about 1500. Since its inception, GEO has been gaining data at the rate of about 300% per year.

[2] Ontrez (Ontology aware Entrez) is named after the popular search engine Entrez at NCBI for database search across 30 NCBI resources.

While BioPortal addresses the *ontological content need* of the biomedical community, Ontrez meets the *data access and data discovery need* of researchers, meeting many of the specific aims outlined in Core 2 of the NCBO grant. In fact, we envision a tight integration of BioPortal and Ontrez as we originally proposed (Figure 1). This tight integration of ontology access and ontology-based applications is essential to enable the capabilities that we present in our Ontrez use cases. In this report, we outline our methodology as well as the current status of the project.

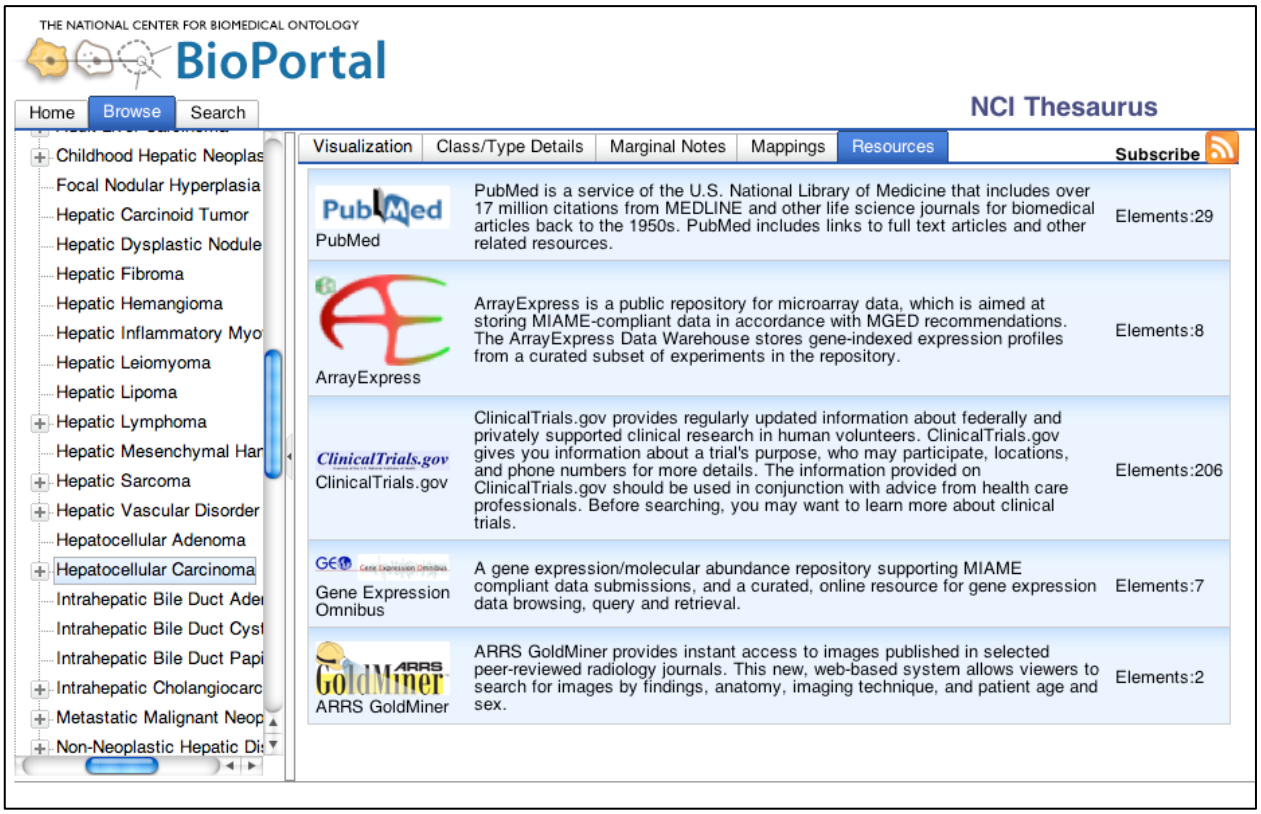

Figure 1. The integration of Ontrez into BioPortal. In this view of the NCI Thesaurus, the user can select an ontology element (in this case, *Hepatocellular carcinoma*) and see immediately a number of online resources that relate directly to that term (and the terms that is subsumes). Ontrez allows the user to link directly to the 29 articles in a small subset of PubMed that deal with Hepatocellular carcinoma and its subtypes. The user also can link to any of the eight DNA microarray datasets related to hepatocellular carcinoma in ArrayExpress or to any of the 206 clinical trials for hepatocellular carcinoma in ClinicalTrials.gov. Figure 6, on page 9, points out the various features of the interface with call outs.

Use cases enabled by Ontrez

One of the key aims for NCBO is to build resources and methods to help biomedical investigators store, view, and compare annotations of biomedical research data and to develop a query interface for the underlying data sets. Below we list the kinds of uses cases enabled by Ontrez.

1. Querying and finding data sets

Ontrez enables researchers to find biomedical datasets by expanding their searchers using ontologies. For example, if a protein expression study is labeled with a term from the NCI thesaurus (such as *pheochromocytoma*), then a researcher can query for *retroperitoneal tumors* and find data sets related to pheochromocytoma (the NCI Thesaurus will provide the knowledge that pheochromocytoma **is_a** retroperitoneal tumor). This use case is similar, in principle, to query expansion at Entrez; however, Entrez does not use ontologies, such as the NCI thesaurus or SNOMED-CT for this purpose. There are pheochromocytoma data sets in GEO, but none show up on searching for retroperitoneal neoplasms in Entrez. In Ontrez, however, a researcher could search for "retroperitoneal neoplasms" and find the relevant samples [7]. Ontrez currently enables such querying across both the Gene-Expression Ombibus (GEO) and ArrayExpress.

2. Integrating gene and protein expression data sets

Researchers who seek to identify biomarkers for stratifying cancers currently need to search for each of the different types of cancers within different databases—GEO and the Tissue Micro-Array Database (TMAD), for example. Using Ontrez, a researcher can locate all the relevant data sets across a diversity of biomedical resources; for example, the researcher could locate data sets for breast cancer across GEO and TMAD, enabling integrative studies, such as those by Quong et al., which correlate protein-expression changes with gene-expression changes [8, 9].

3. Integrative analysis of imaging and genomic data sets

Researchers recently demonstrated that imaging biomarkers of hepatocellular carcinoma can be identified by integrating gene-expression data and CT-scan imaging data. The authors correlated gene-expression profiles for hepatocellular carcinoma with features extracted from the original CT scan images [10]. Finding the data sets needed for such analyses is challenging. Discovering these data for such novel studies is readily enabled by Ontrez, which allows researchers to seamlessly search both GEO and the Goldminer repositories.

4. Searching clinical trials

Currently, Ontrez contains ontology-based annotations of all the clinical-trial descriptions in clinicaltrials.gov. In addition to indexing the trial description, we intend to augment Ontrez by processing the text describing the eligibility criteria of clinical trials from the clinicaltrials.gov resource. Once these clinical-trial eligibility criteria are annotated using ontologies, Ontrez will enable users to compute the semantic distance [11] between specific clinical trials and the medical-record data for individual patients who are candidates for those trials, using technology being developed for the AppliMatch project. (AppliMatch is an ongoing project in collaboration with Ida Sim's group, which is developing methods to identify clinical trials whose eligibility criteria are semantically close to the data in a particular patient's electronic medical record.)

Functional Specification

Figure 2 summarizes the functional specification of the Ontrez system.

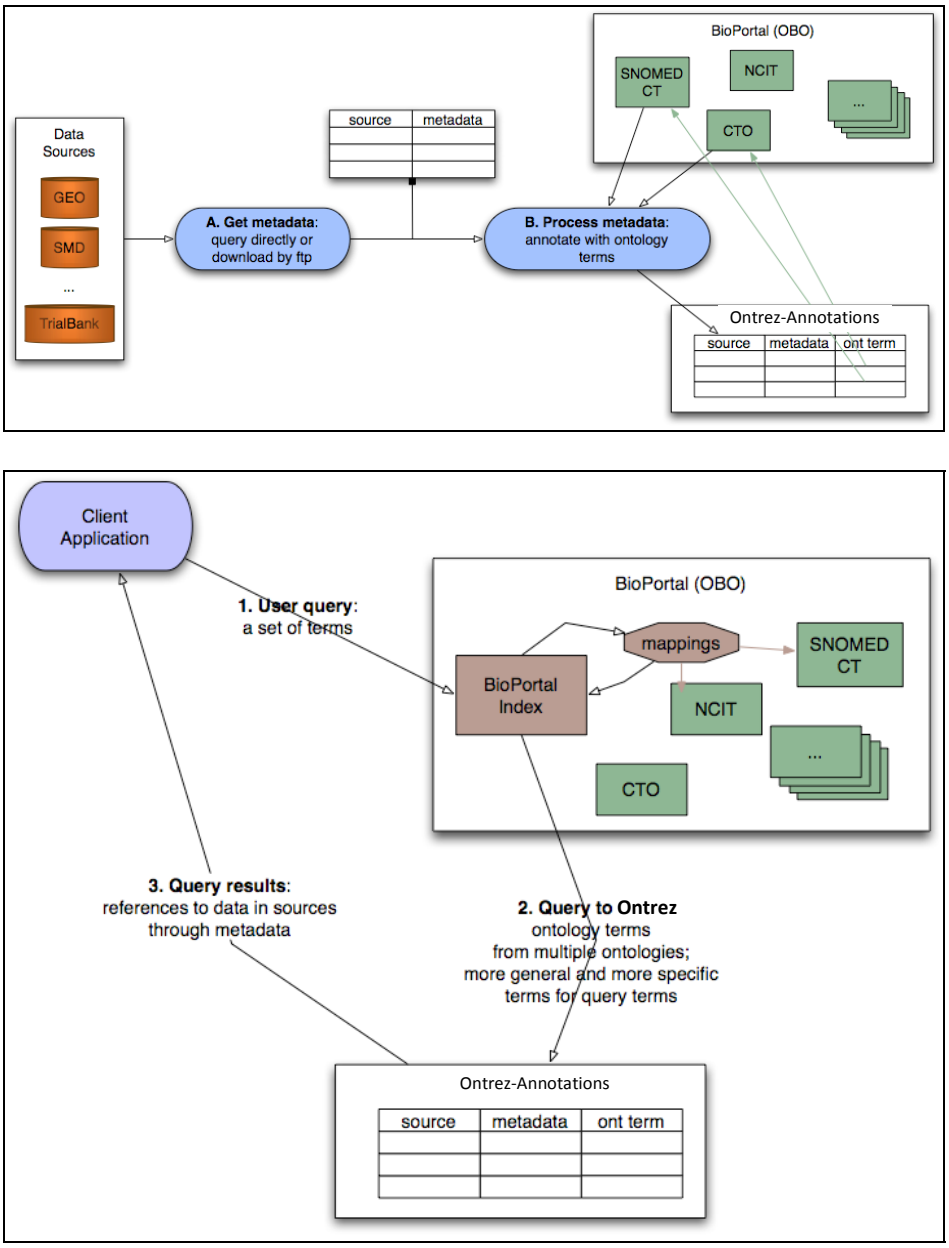

Figure 2: Ontrez functional specification. *(Top panel)* We retrieve the metadata from data resources such as GEO (A), and tag them with ontology terms using the library of ontologies in BioPortal (B). The result is stored as annotations of these data sources in Ontrez. *(Bottom panel)* User queries are formulated as a set of terms (1). We use the BioPortal index to convert the query to ontology terms. We use the subsumption relations in the ontologies and the mappings in BioPortal to expand the query. We then query the annotations of data sources in Ontrez with the expanded set of terms (2). The user receives the result (3) in terms of references to the original data sources.

The key functionality of Ontrez is to provide a service that will enable users to locate biomedical data resources related to their search for particular ontology terms.

The current Ontrez prototype has the following two main functional components: (1) an index-creation component and (2) a user-query component.

The process of index creation involves the following three steps:

1. Accessing a set of resources
2. Processing metadata annotations
3. Building the index

The query component involves the following two steps:

1. Semantic query expansion
2. Interaction and results display

Figures 3 and 4 show an example of the Ontrez system in action, processing a specific GEO element. Note that each GEO element, in this case a dataset, has a title and a description field that contain free text entered by the person creating the data set. Moreover, GEO datasets can have an additional 24 descriptors (such as agent, cell line, and species) along with their subset descriptions. We retrieve these descriptors, either by accessing the resource over the Web.[3]

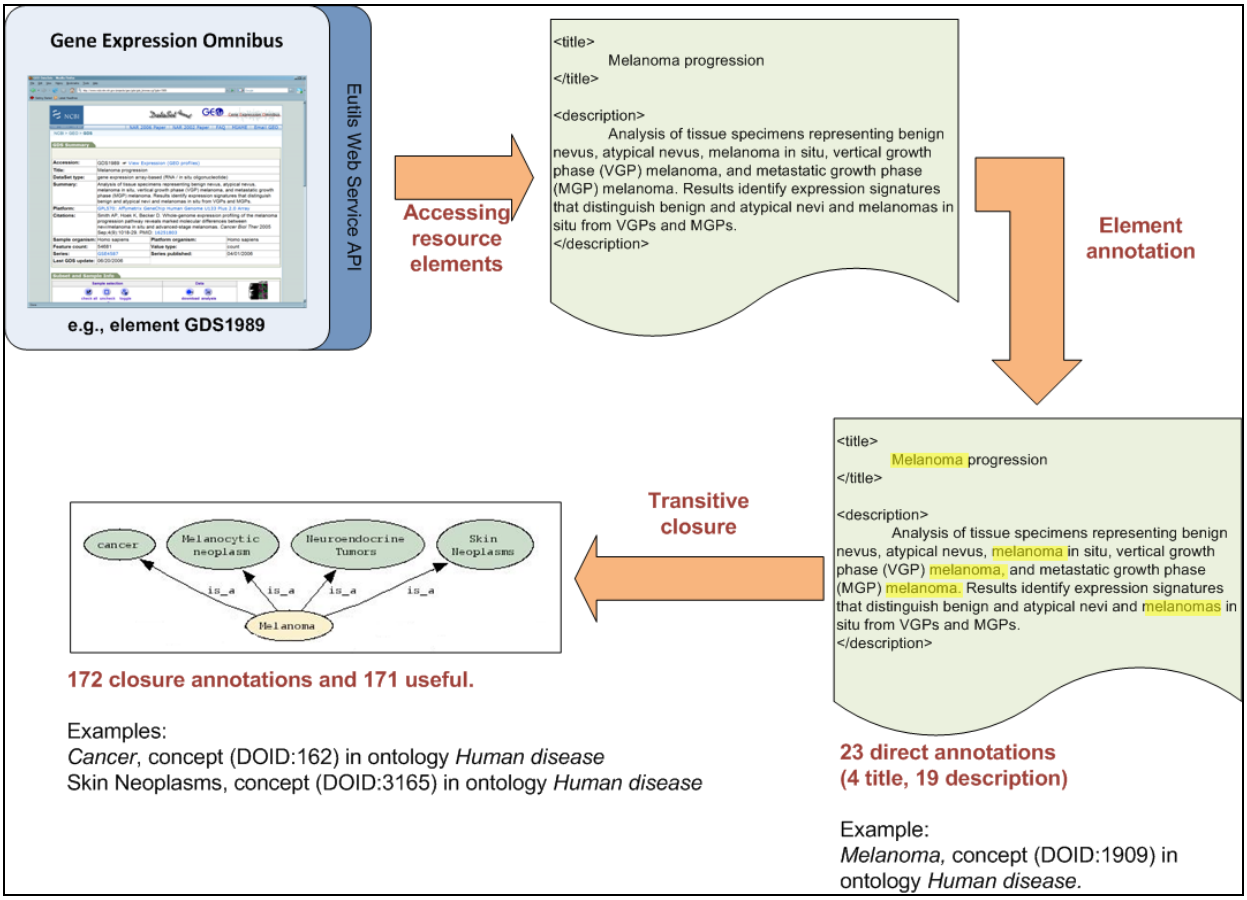

Figure 3. Indexing an element from GEO. The GEO element GDS1989 is annotated according to its title and its description, as well as by the ontology elements that comprise the transitive closure over the parent–child relationships subsumed by the direct annotations.

[3] In case of GEO we use the E-Util programs, specifically the eSearch and eSummary.

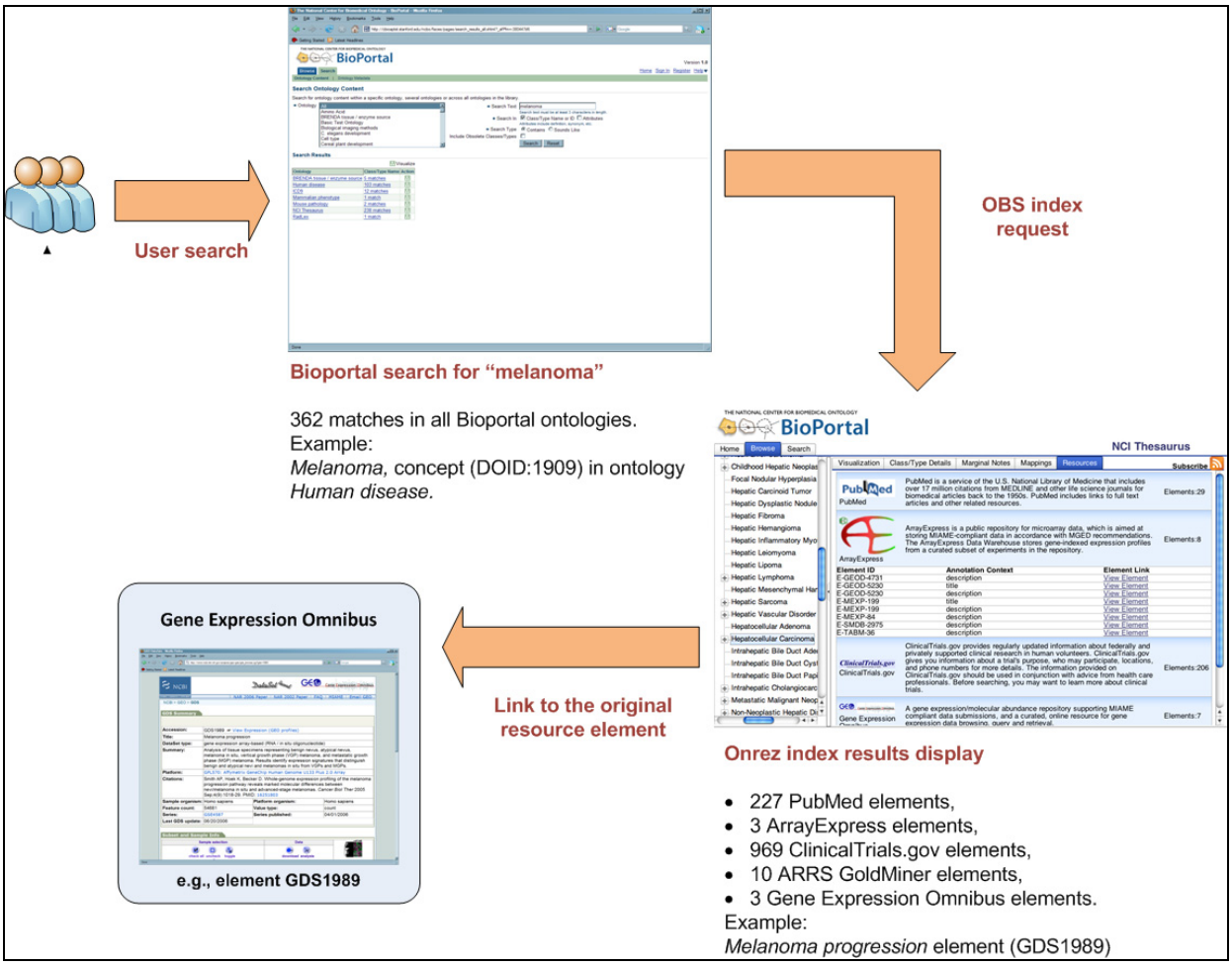

Figure 4. Performing a search of GEO using Ontrez. A user searching for "melanoma" in Bioportal is able to view the set of online data resources that have been annotated with the ontology terms related to this query. The GEO element "melanoma progression" is returned as a pertinent element for this search. (**Note**: In the current version, we have dealt only with element titles and descriptions to validate the notion of context awareness. Later, we will process the more of the metadata structure to enable a finer grained level of detail.) The display within BioPortal allows the user to view the original data set with a single click.

Design specification

Figure 5 shows the design of Ontrez, with different architectural levels that map into these five functional steps. The software components corresponding to each of these steps are described in detail in the Appendix and we summarize them here.

Components in the **Resource level** are concerned with accessing resource elements (such as experiment descriptions in GEO) both via remote access (e.g., Web services) and local access (e.g., a downloaded local copy). Components in the **Annotation level** create a dictionary of relevant ontology terms from both Bioportal and UMLS ontologies and create annotation tables by recognizing concepts in the metadata. The component performing the concept recognition is a tool called mgrep developed by The National Center for Integrative Biomedical Informatics (NCIBI). The design of the annotation level allows us to plug-in other concept recognizers as needed.

The **Index Level** computes what elements are annotated with a particular ontology term. The **Ontology level** provides the terms to use in annotation as well as performs semantic query expansion of user provided search terms.

The **User level** components integrate the annotations created by Ontrez with Bioportal's ontology search. For example a user can search for "melanoma" and get links to related resource elements (such as datasets, publications, and clinical trials).

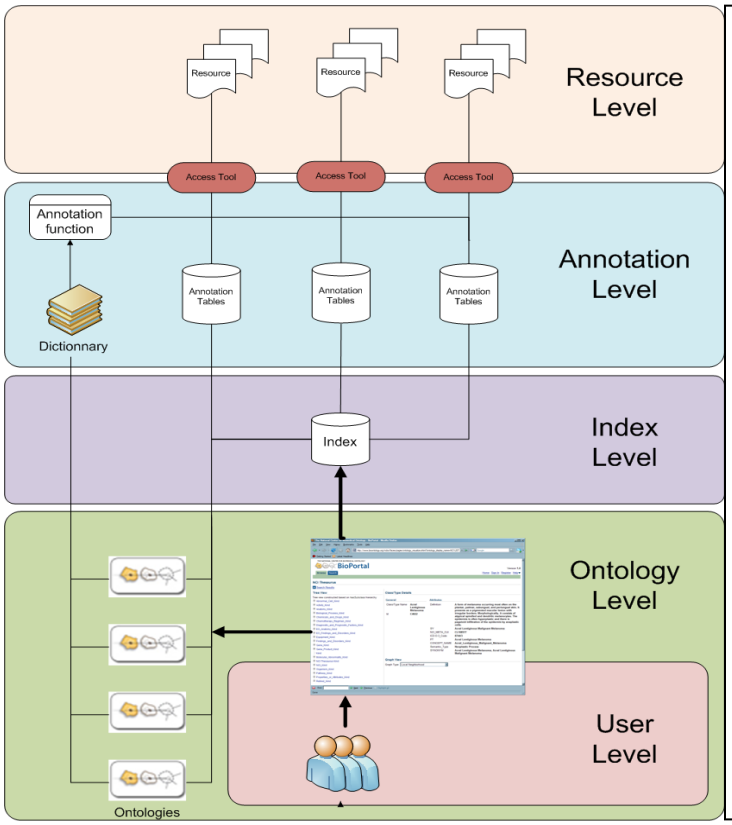

Figure 5. Ontrez design levels

The figure shows an overview of the five primary design components of Ontrez. The **Resource level** components are concerned with accessing resource elements (such as experiment descriptions in GEO) both via remote access (e.g., Web services) and local access (e.g., a downloaded local copy). The **Annotation level** components create a dictionary of relevant ontology terms and also create the annotation tables. The **Index Level** computes what elements are annotated with a particular ontology term. The **Ontology level** provides the terms used in annotation as well as semantic query expansion. The **User level** component integrates the Ontrez annotations with Bioportal's ontology search. For example a user can search for "melanoma" and get links to related resource elements (such as datasets, publications, and clinical trials).

Results and current status

In our work to develop the Ontrez prototype, we have processed (1) metadata annotations of high-throughput gene-expression data in repositories such as GEO and Array Express, (2) clinical-trial descriptions in repositories such as clinicaltrials.gov, (3) annotations of images in repositories such as Goldminer, and (4) abstracts of articles published in Pubmed. For a quick summary of the project status:

- **Figure 6** shows a snapshot of the current resources annotated in Ontrez and available for search via Bioportal.
- **Table 1** shows the current number of elements annotated from each resource that we have processed.

Our prototype uses 2,010,533 terms from 50 different ontologies (listed in the appendix) to index the resource elements shown in Table 1.

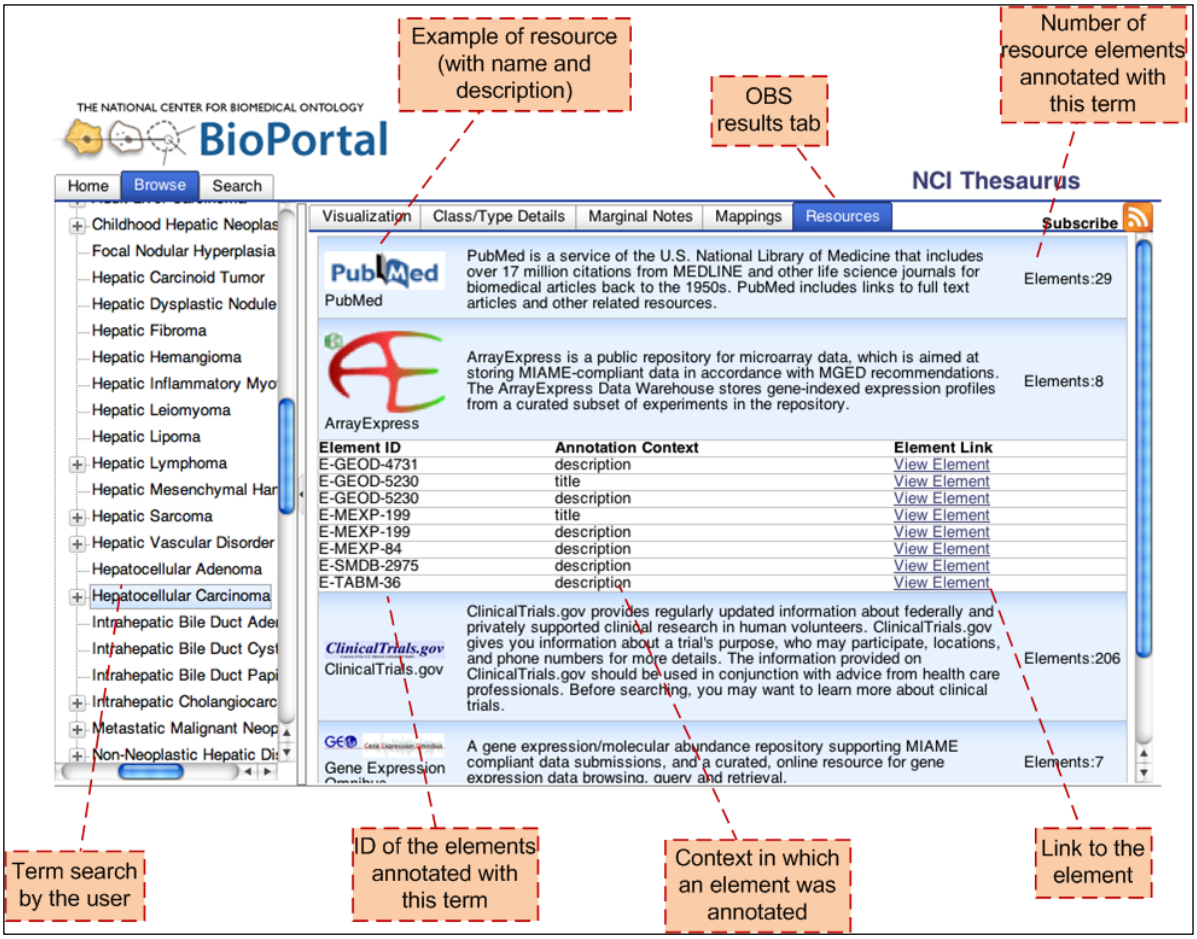

Figure 6. The figure shows a snapshot of the Ontrez user interface within BioPortal, where the user has selected the term 'Hepatocellular Carcinoma' and then navigated to the 'Resources' tab. The call outs point to and describe the different features of the interface.

Table 1. The current number of elements annotated from each resource that we have processed.

Resource	Number of elements	Resource file size (Kb)	Number of direct annotations	Number of closure annotations	Total number of 'useful'[4] annotations
PubMed subset	10164	13461	187686	681973	857459
ArrayExpress	2751	2880	143134	484758	619133
ClinicalTrials.gov	43918	8379	1206939	6792430	5217115
Gene Expression Omnibus	546	163	16494	100984	116234
ARRS GoldMiner	1155	494	53082	290935	340915
TOTAL	**58534**	**25377**	**1607335**	**8351080**	**7150856**

Ontrez implements ontology-based search, which is an area of active research in the computer-science community [3, 4]. There are potential problems that we may encounter in scaling up the prototype to production use. However, we believe that these problems are addressable in the amount of time that we have left in the project.

[4] To prevent redundant annotation entries, we do not add those terms in the closure annotations for which the element was already annotated directly.

The issues that we have identified are:

1. **Success of processing the metadata depends on the ability to identify accurately concepts such as diseases, drugs, anatomical parts, and so on in the free-text annotation.**
   - In our preliminary studies, we have very encouraging results for identifying disease names in GEO text annotations. We rely on the mgrep tool developed by NCIBI, which has a very high degree of accuracy (over 95%) in recognizing disease names. We have not yet conducted an evaluation of the precision and recall of mgrep in recognizing concepts of different types, such as diseases, drugs, and anatomical parts, however. We are aware that NCIBI is developing entity recognizers for tagging entities such as cytobands and microsatellite markers, and we hope to utilize these new tools in our future work.
   - Relationships are very hard to identify in free text. Therefore, we will focus initially on the entities that are relevant to the data source (e.g., disease names for GEO), rather than on the relations.
2. **Success of processing the metadata-annotations depends on the ability to access the public data resources.** Currently, we have processed GEO, ArrayExpress, PubMed, Clinicaltrials.gov, and a subset of Goldminer. We have access to the Stanford Tissue Micro-Array Database (which will soon be available publicly) as well as to the public data sets in the Stanford Microarray Database (SMD). We will decide on the inclusion of new resources for indexing based on their value to the community and will ask for feedback from the Bioportal User Group when prioritizing the resources to include in our work.
3. **The quality of the search results depends on the quality of query expansion that we can achieve.** In our preliminary results, we have been able to identify relevant data sets across GEO and TMAD with approximately 77% accuracy [6]. We note published results from Moskovitch et al. [3], which demonstrate that concept-based searches outperform keyword queries and that search results improve monotonically as additional concepts are included in the search.

Summary

Our preliminary results demonstrate that the metadata available in public biomedical data can be processed automatically by computer programs, and that the resulting annotations can enable several diverse use cases. Ontrez can process metadata-annotations of gene-expression data sets, descriptions of radiology images, clinical-trial reports, as well as abstracts of PubMed articles to annotate them automatically with terms from appropriate ontologies. Ontrez enables researchers to search biomedical data (such as genomic data sets, images, clinical trials, and published papers) in an ontology-driven manner from within BioPortal. We believe that, as we expand Ontrez 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.

References

[1] Butte, A.J. and I.S. Kohane, *Creation and implications of a phenome-genome network.* Nat Biotechnol, 2006. 24(1): p. 55-62.

[2] Spasic, I., et al., *Text mining and ontologies in biomedicine: making sense of raw text.* Brief Bioinform, 2005. 6(3): p. 239-51.

[3] Moskovitch, R., et al., *A Comparative Evaluation of Full-text, Concept-based, and Context-sensitive Search.* J Am Med Inform Assoc, 2007. 14(2): p. 164-174.

[4] Sneiderman, C.A., et al., *Knowledge-based Methods to Help Clinicians Find Answers in MEDLINE.* J Am Med Inform Assoc, 2007. 14(6): p. 772-780.

[5] Shah, N.H., et al. *Ontology-based Annotation and Query of Tissue Microarray Data*. in *AMIA Annual Symposium*. 2006. Washington, DC.

[6] Shah, N.H., et al., *Ontology-driven Indexing of Public datasets for Translational Bioinformatics*, in *AMIA 2008 STB Submission*. 2007: Stanford.

[7] Butte, A.J. and R. Chen, *Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics.* AMIA Annu Symp Proc, 2006: p. 106-10.

[8] Quong, J.N., et al. *Correlation of protein and gene expression for the stratification of breast cancer patients*. in *Breast Cancer Symposium*. 2007. San Francisco, California: American Society of Clinical Oncology.

[9] Basik, M., S. Mousses, and J. Trent, *Integration of genomic technologies for accelerated cancer drug development.* Biotechniques, 2003. 35(3): p. 580-2, 584, 586 passim.

[10] Segal, E., et al., *Decoding global gene expression programs in liver cancer by noninvasive imaging.* Nat Biotechnol, 2007. 25(6): p. 675-80.

[11] Caviedes, J.E. and J.J. Cimino, *Towards the development of a conceptual distance metric for the UMLS.* J Biomed Inform, 2004. 37(2): p. 77-85.

Appendix

Description of Ontrez design levels

Figure 5 showed the design of Ontrez, with different architectural levels that provide the five functional steps for annotating resources. The software components corresponding to each of these steps were summarized in the main text and are described in detail here.

The resource level – Accessing resources

The ***resource level*** comprises biomedical data sets, repositories, and databases, and provides a Web interface to query those resources. Resources are composed of elements (such as GEO data sets) 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. An element is associated with structured metadata for the element (title, description, abstract, and so on). In order to annotate these resource elements, Ontrez either can download a local copy of the resource elements or can access them directly online (preferred) through a public interface, such as Web services. Processing online resources and keeping the set of annotated elements up to date is done by a resource-specific access tool.

Annotation level – Processing the metadata-annotations

The ***annotation level*** processes the metadata corresponding to each element from an online data resource to assign ontology terms to that element; Ontrez stores these assignments in annotation tables. The annotation of elements with ontology terms is performed using a dictionary that provides the set of terms to identify in the metadata. The annotation process is context aware, meaning that we keep track of the context (such as title, description) in each element's metadata where the annotation was derived. The resulting annotation table associates each resource-element identifier with terms belonging to the ontologies in BioPortal. The annotation table contains information such as "Element E was annotated with term T in context C".

Note: The actual concept recognition step in a particular context (the title or description) is done using the **mgrep** tool developed by The National Center for Integrative Biomedical Informatics (NCIBI). The design of the annotation level components is such that we can plug-in other concept recognizers.

Index level – Building the index

At the ***Index level***, a global index is constructed that indexes annotations according to dictionary terms from the ontologies in BioPortal. The index contains information such as: "term T annotates elements E1, E2 … in resource R1". We also perform a transitive closure. That is, for each annotation found in the annotation table, we assign additional terms to annotate that element based on the parent–child relationships subsumed by the original term.

Ontology level – Semantic query expansion

The ***Ontology level*** provides access to the ontology content used for creating the dictionary and for computing transitive closure. The ontologies are also used for semantic query expansion. For example, when a user issues a query with a term such as "skin neoplasm", ontologies may be used to expand the query by adding more specific terms to the query (e.g., specializations of skin neoplasm, such as malignant melanoma), as well as by adding terms from other ontologies that are "semantically close" to the query terms or explicitly mapped to them. We have an active research project to evaluate alternative semantic-distance measures that we may use in Ontrez for query expansion.

User level – Interaction and results display

At the ***User level***, Ontrez provides a user who is searching for a specific ontology term (e.g., Hepatocellular carcinoma) with resource elements that have are found directly in the Ontrez index or via the semantic query expansion. The search service is provided directly through the existing Bioportal user interface. Ontrez results are displayed inside Bioportal and a Web link is available to each of the online data elements (see Figure 6).

List of Ontologies used in Indexing

Bioportal Ontologies	UMLS Ontologies
Biological imaging methods	SNOMED Clinical Terms, 2007_01_31
Dictyostelium discoideum anatomy	RxNorm Vocabulary, 07AA_070503F

Human developmental anatomy, timed version	Micromedex DRUGDEX, 2007_04_02
Basic Test Ontology	National Drug File - Reference Terminology, 2004_01
Human disease	Medical Subject Headings, 2005_2005_01_17
Human developmental anatomy, abstract version	National Cancer Institute Thesaurus, 2006_03D
Cereal plant development	
Plant environmental conditions	
Plant growth and developmental stage	
Amino Acid	
PATO	
Cell type	
RadLex	
Common Anatomy Reference Ontology	
Physico-chemical methods and properties	
Mouse adult gross anatomy	
Mouse gross anatomy and development	
Subcellular Anatomy Ontology (SAO)	
Pathway ontology	
Protein-protein interaction	
Xenopus anatomy and development	
Maize gross anatomy	
Drosophila gross anatomy	
Evidence codes	
FlyBase Controlled Vocabulary	
Drosophila development	
Mosquito gross anatomy	
Cereal plant trait	
Mouse pathology	
Environment Ontology	
Cereal plant gross anatomy	
Proteomics data and process provenance	
Sequence Ontology	
Fungal gross anatomy	
BRENDA tissue / enzyme source	
Zebrafish anatomy and development	
Medaka fish anatomy and development	
Mammalian phenotype	
Molecule role (INOH Protein name ontology)	
Protein modification	
Galen	
Multiple alignment	
C. elegans development	
Physico-chemical process	