BAOSearch: A Semantic Web Application for Biological Screening and Drug Discovery Research

Saminda Abeyruwan[4], Caty Chung[1], Nakul Datar[1], Felimon Gayanilo[1], Amar Koleti[1], Vance Lemmon[2], Christopher Mader[1], Mitsunori Ogihara[4], Deepthi Puram[1], Kunie Sakurai[1], Robin Smith[1], Uma Vempati[1], Sreeharsha Venkatapuram[1], Ubbo Visser[4], and Stephan Schürer[1,3]*

[1] Center for Computational Science, University of Miami, Florida, USA
c.mader@miami.edu,{ndatar,akoleti,ext-svenkatapuram,fgayanilo,cchung,
dpuram,ksakurai,uvempati}@med.miami.edu
[2] The Miami Project to Cure Paralysis, University of Miami Miller School of Medicine, Florida, USA
vlemmon@miami.edu
[3] Department of Molecular and Cellular Pharmacology, University of Miami Miller School of Medicine, Florida, USA
sschurer@med.miami.edu
[4] Department of Computer Science, University of Miami, Florida, USA
{visser,saminda,ogihara}@cs.miami.edu

Abstract. BAOSearch is a semantic web application for querying, browsing and downloading biological screening data relevant for drug discovery. We developed a BioAssay Ontology (BAO) in order to formalize the domain of biological screening and annotated large sets of data to make complex and diverse life science data accessible to researchers via simple queries. Our software architecture and BAO will also enable the integration with orthogonal life science databases (such as pathways and disease) and ultimately facilitate the discovery of new biomedical knowledge. BAOSearch is a multi-tier, web-based, AJAX-enabled application written primarily in Java and built following a Restful web services paradigm. The paper gives an overview of the architecture, the methods used and gives some examples of the types of queries that BAOSearch enables.

Keywords: Bioassay, ontology, drug discovery, life science, semantic search

1 Background

During the last few years small molecule biological assays performed at publicly funded screening centers have been generating very large amounts of data. The

* Senior corresponding author

largest effort is the NIH Molecular Libraries Program[5], which has the goal of developing novel chemical tools (chemical probes) to interrogate biological systems using high-throughput screening (HTS). Huge data sets generated by HTS are deposited in PubChem[6] [5]. Other public resources for small molecule screening data include ChemBank[7] or the Psychoactive Drug Screening Program K_i database[8]. In addition to data in PubChem and other public databases there are even larger data sets in pharmaceutical companies.

Our mission is to make it much easier to access, query, and analyze these diverse HTS data and thus dramatically increase their value to the chemical biology, screening and cheminformatics communities. We are also in the process of integrating and comparing various screening data sets from multiple sources. This allows researchers to compare their own data to other public data sets, for example in PubChem. A longer-term goal is to facilitate the integration of screening data with other types of life science data, such as biological pathways, disease networks, and structural biology, etc. in order to analyze HTS in the context of specific mechanisms of biological functions and to facilitate the transformation of data into knowledge (see figure 1).

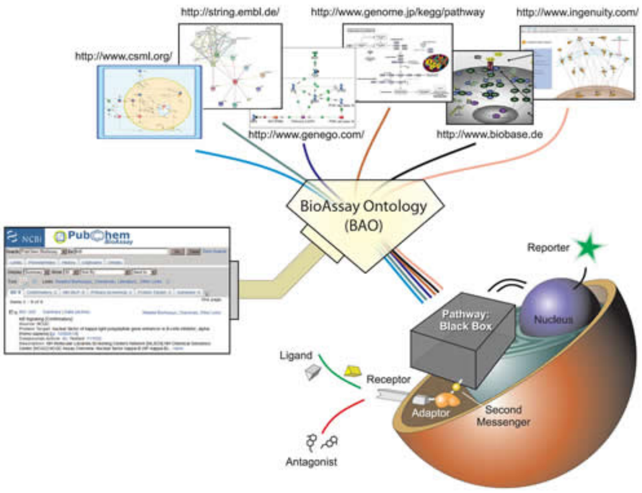


Fig. 1. Long-term goal and the importance of the central component BioAssay Ontology (BAO)

[5] http://mli.nih.gov/mli/
[6] http://pubchem.ncbi.nlm.nih.gov
[7] http://chembank.broadinstitute.org/
[8] http://pdsp.med.unc.edu/kidb.php

2 Description

The BioAssay Ontology (BAO)[9] is an extensible, knowledge-based, expressive description of biological assays (currently SHOIQ(D)). BAO defines 460 concepts of assays that are relevant to chemical biologists and drug discovery researchers. BAO also describes quantitative screening outcomes and can relate different types of outcomes on various levels. This enables the retrieval of not only the data directly specified in a search query, but also additional relevant results that a researcher is likely interested in, but may not know exists in the repository. With the description of quantitative outcomes and the many relevant categories of data for drug discovery and chemical biology, BAO makes it possible to define highly complex concepts and make them available via simple text search. Because these concepts are defined in the ontology, the obtained results will always be current with the data in the repository.

The BAO enables non-experts to access knowledge that typically requires scientists from different disciplines to discover. Complex concepts that relate specific molecular targets that underlie biological function to the technologies that interrogate them can be explored. Using large sets of empirical data such as those in the BAO repository, such knowledge can be uncovered. BAO and BAOSearch, the search and query front-end, thus make up one of the first applications of semantic technology to work on large data sets to derive new knowledge in the biomedical domain. The ontology describes numerous concepts related to biological screening, including Perturbagen, Format, Meta Target, Technology, Detection, and Endpoint. Perturbagens are perturbing agents that are screened in an assay; they are mostly small molecules. Meta Target refers to the biological target, describing not just protein targets, but also pathways, biological processes or events, etc. targeted by the assay. Format describes the biological or chemical features common to each test condition in the assay and includes biochemical, cell-based, organism-based, and variations thereof. Technology describes the assay methodology, assay design, and implementation of how the perturbation of the biological system is translated into a detectable signal. Detection Method relates to the physical method and technical details to detect and record a signal. Endpoints are the final HTS results as they are usually published (such as IC50, percent inhibition, etc.). BAO has been designed to accommodate multiplexed assays. All main BAO components include multiple levels of sub-categories and specification classes, which are linked via object property relationships forming an expressive knowledge-based representation.

The current version of BAO consists of 460 OWL 2.0 classes, 36 object properties (relations), 15 data properties, and 45 individuals (not including any annotated assays). It should be noted that three major bioinformatic terminology bases: SNOMED [4], Galen [3], and GO [1] have the expressivity of EL, with additional role properties. In EL, only intersections between concepts and full existential quantification are possible. In comparison, the BAO ontology is a significant improvement in expressivity.

[9] http://bioassayontology.org/

2.1 BAOSearch

BAOSearch is an application for querying, viewing, browsing and downloading diverse high-throughput screening (HTS) for drug discovery and related life science research. We have annotated sets of assays from different sources with BAO to make complex and diverse life science data accessible to researchers via simple querying. BAOSearch is a multi-tier, web-based, AJAX-enabled application written primarily in Java and built following a Restful [2] web services paradigm.

The service-based aspect of the architecture allows the user interface (UI) to be separated from storage and manipulation of the data, and provides well-defined interfaces for UI components to access and manipulate application data. This separation of application components creates the potential of developing multiple UIs that access the same service, but which render the data differently, or run on different platforms (e.g., browsers, mobile applications). This architecture also creates an opportunity for other software applications (not only User Interfaces) to access the system to query and retrieve data.

The browser-based UI was built using JSP and JavaScript, with components from several JavaScript libraries including jQuery[10]. All data are stored in a MySQL database. SDB[11] is used as the triple-store. Other data required by the application is stored in a relational schema and are accessible using Hibernate. Figure 2 shows the high-level architecture of the BAOSearch project.

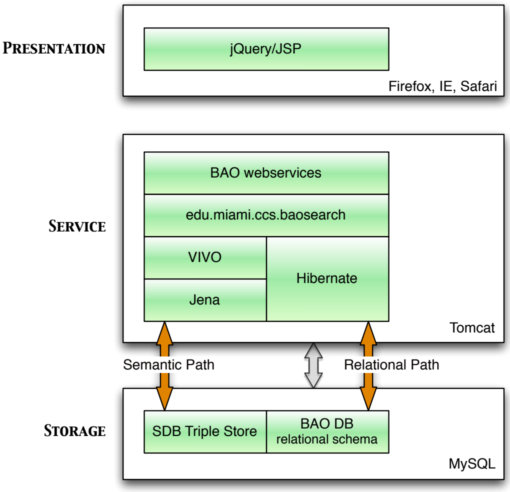


Fig. 2. High-level architecture of BAOSearch

2.2 Ontology concept visualization

BAOSearch also provides a Treemap[12] display of the ontology (display is limited to descriptions and nominals, see figure 4). This enables users to browse the ontology and retrieve individuals that display in a grid. The application middle-tier is written in Java using Jena[13] for accessing and manipulating

[10] http://jquery.com
[11] http://openjena.org/SDB
[12] http://en.wikipedia.org/wiki/Treemap
[13] http://jena.sourceforge.net

semantic data. In addition to this BAOSearch also uses components from the open source VIVO[14] project.

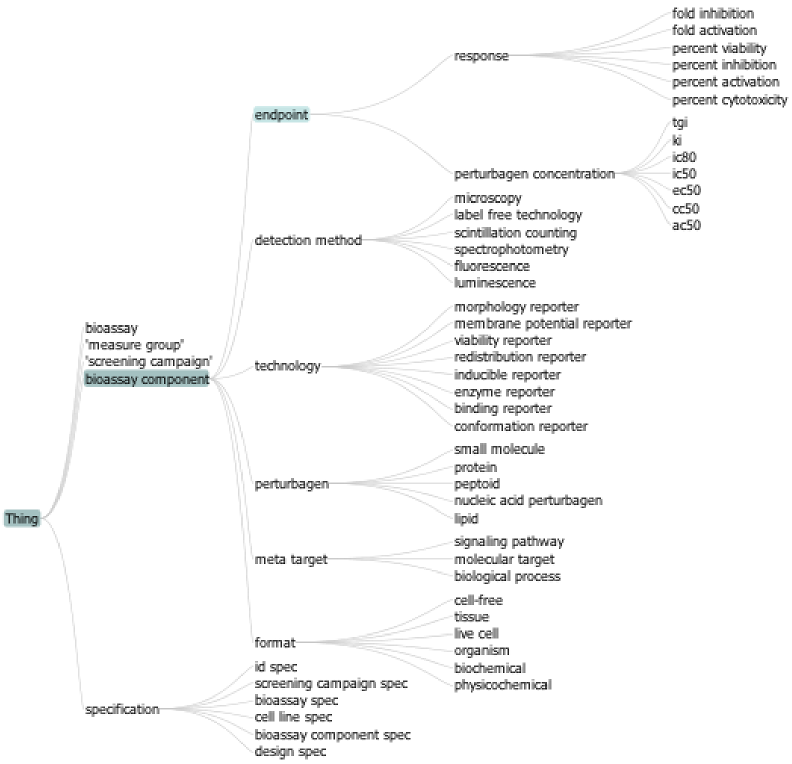
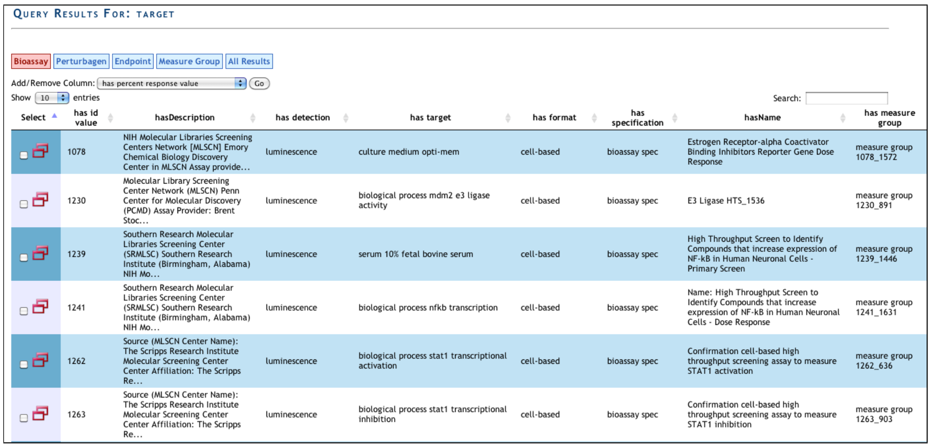


Fig. 3. Treemap view on parts of BAO

2.3 Search interface and grid display

The primary search interface is a simple text-based search box, which gives users the ability to enter sets of search terms and see results in a gridded (*spread sheet-like*) display that is categorized by concepts from the ontology. Major concepts are displayed above the grid. Clicking on a concept (e.g. target) shows relevant search results within that category. The columns of the grid represent the individual targets (e.g., "has target") and relations (e.g., "has endpoint").

[14] http://www.vivoweb.org



Query Results For: target

Add/Remove Column: has percent response value

Show 10 entries

Search:

Select	has id value	hasDescription	has detection	has target	has format	has specification	hasName	has measure group
□	1078	NIH Molecular Libraries Screening Centers Network [MLSCN] Emory Chemical Biology Discovery Center in MLSCN Assay provide...	luminescence	culture medium opti-mem	cell-based	bioassay spec	Estrogen Receptor-alpha Coactivator Binding Inhibitors Reporter Gene Dose Response	measure group 1078_1572
□	1230	Molecular Library Screening Center Network (MLSCN) Penn Center for Molecular Discovery (PCMD) Assay Provider: Brent Stoc...	luminescence	biological process mdm2 e3 ligase activity	cell-based	bioassay spec	E3 Ligase HTS_1536	measure group 1230_891
□	1239	Southern Research Molecular Libraries Screening Center (SRMLSC) Southern Research Institute (Birmingham, Alabama) NIH Mo...	luminescence	serum 10% fetal bovine serum	cell-based	bioassay spec	High Throughput Screen to Identify Compounds that increase expression of NF-kB in Human Neuronal Cells - Primary Screen	measure group 1239_1446
□	1241	Southern Research Molecular Libraries Screening Center (SRMLSC) Southern Research Institute (Birmingham, Alabama) NIH Mo...	luminescence	biological process nfkb transcription	cell-based	bioassay spec	Name: High Throughput Screen to Identify Compounds that increase expression of NF-kB in Human Neuronal Cells - Dose Response	measure group 1241_1631
□	1262	Source (MLSCN Center Name): The Scripps Research Institute Molecular Screening Center Center Affiliation: The Scripps Re...	luminescence	biological process stat1 transcriptional activation	cell-based	bioassay spec	Confirmation cell-based high throughput screening assay to measure STAT1 activation	measure group 1262_636
□	1263	Source (MLSCN Center Name): The Scripps Research Institute Molecular Screening Center Center Affiliation: The Scripps Re...	luminescence	biological process stat1 transcriptional inhibition	cell-based	bioassay spec	Confirmation cell-based high throughput screening assay to measure STAT1 inhibition	measure group 1263_903

Fig. 4. Part of the grid display as one of the results of BAOSearch

2.4 Examples

We show three selected examples that the BAOSearch is able to answer with the integration of our SPARQL interface. However, we will elaborate only one example in details due to space limitations.

Example 1: Show all compounds from assays with an inhibitory mode of action that show a percentage response of 50% or greater at ≤10 μM screening concentration. This example relates to a common query for compounds with an IC50 value of less than a certain cutoff (here ≤10 μM). Such a query should also return results of differently named IC50 endpoints (e.g. AC50), but which a user may not know exist. A user querying the database may also be interested in returning other relevant endpoints, such as IC80 values ≤10 μM (if it existed in the repository) or other result types such as potent inhibitors screened at less than the IC50 concentration. With the semantic definition of IC50 in our ontology, we can achieve both.

Example 2: All assays with compounds that have a mode of action *activation* and show a percentage response of $\geq$ 50% at ≤10 μM screening concentration.

In addition to assays with compounds that have an endpoint activation of 50 % at <10 μM, this query also returns assays with an EC50 or an AC50 (if the mode of action is activation) value of <10 μM. This example also illustrates one of the constructive reasoning mechanisms of the BAO ontology. In the ontology *activation* was defined as *equivalent* to *stimulation* (among other equivalent classes, e.g. agonist). As the reasoning system returns results that satisfy the original query and the inferred query, searching *'activation'* returns exactly the same results as querying for *'stimulation'* independent from the specific term used to describe the pharmacological action.

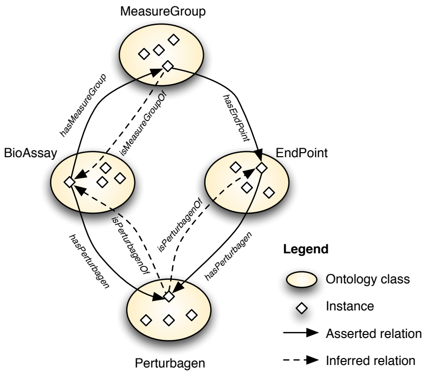


Fig. 5. Relationships between BioAssay, EndPoint, and Perturbagen in our BAO ontology.

Example 3: With this example, we illustrate a specific case concerning three concepts: endpoint, bioassay, and perturbagen. Figure 5 shows the relevant relationships between these concepts[15] (there are more in the ontology). Of particular interest was the relation *'has perturbagen'* that holds between endpoint and perturbagen as well as bioassay and perturbagen. The ontology specifies that this property has an *inverse* relationship with *'is perturbagen of'*. Thus, we use this inference in order to retrieve eligible instances (individuals).

In this example we queried for all perturbagens that have a percentage response of ≥50 % in at least three assays. The SPARQL query was as follows:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX bao: <http://www.bioassayontology.org/bao#>
# results
SELECT ?pert
WHERE {
{ ?pert rdf:type bao:BAO_0000021 .
    ?pert bao:BAO_0000361 ?assay .
    ?assay bao:BAO_0000209 ?measureGroup .
    ?measureGroup bao:BAO_0000208 ?endpoint .
    ?endpoint bao:BAO_0000195 ?percentResponseValue .
} UNION {
        ?pert rdf:type bao:BAO_0000021 .
        ?pert bao:BAO_0000361 ?assay .
        ?assay bao:BAO_0000209 ?measureGroup .
        ?measureGroup bao:BAO_0000208 ?endpoint .
        ?endpoint bao:BAO_0000337 ?percentResponse .
        ?percentResponse bao:BAO_0000195 ?percentResponseValue .
    }
    FILTER (?percentResponseValue >= 50)
}
GROUP BY ?pert
HAVING (count(distinct ?assay) >= 3)
```

In this query, we used the inferred relation *'is perturbagen of'*, which points to either an endpoint or a bioassay. The query separately checked for bioassay

[15] The concept 'measure group' exists to accommodate multiplexed assays; it is not used in this example.

instances and endpoint instances. This syntax allows for the expression of the notion of *’at least’* in a simple way. Specifically, we use the syntactic extensions available in ARQ ‘ SPARQL[16] implementation. The ’GROUP BY’ extended clause groups the unique ?pert result set (?pert is a variable here) in a row-by-row basis. The ’HAVING’ clause applies the filter ’count(distinct ?assay))’ to the result set after grouping. The results of the query were as follows. First, we queried for the compound and obtained:

```
(1) (?pert=<bao#individual_BAO_0000021_646704>)
```

We then use this result (bao:individual_BAO_0000021_646704)[17] for the next query:

```
SELECT ?assay ?percentResponseValue
WHERE {
{     bao:individual_BAO_0000021_646704 bao:BAO_0000361 ?assay .
    ?assay bao:BAO_0000209 ?mg .
    ?mg bao:BAO_0000208 ?endpoint .
    bao:individual_BAO_0000021_646704 bao:BAO_0000361 ?endpoint .
    ?endpoint bao:BAO_0000195 ?percentResponseValue .
} UNION {
        bao:individual_BAO_0000021_646704 bao:BAO_0000361 ?assay
        ?assay bao:BAO_0000209 ?mg .
        bao:individual_BAO_0000021_646704 bao:BAO_0000361 ?endpoint .
        ?endpoint bao:BAO_0000337 ?percentResponse .
        ?percentResponse bao:BAO_0000195 ?percentResponseValue .
    }
    FILTER (?rv >= 50)
}
```

Here are the final results:

```
(1) (?assay=<bao#individual_BAO_0000015_1262>)
    (?percentResponseValue="116.84"^^xsd:float)
(2) (?assay=<bao#individual_BAO_0000015_1306>)
    (?percentResponseValue="106.48"^^xsd:float)
(3) (?assay=<bao#individual_BAO_0000015_1316>)
    (?percentResponseValue="99.42"^^xsd:float)
```

In the Example 3 query, bioassay, endpoint, and response value could easily be further specified, e.g. using BAO concepts meta target or technology. This allows the construction of complex queries in a simple manner. Thus the *inverse* relationship *’is perturbagen of’* allows for directly querying of compounds that may act via an artifactual mechanism (e.g. active in many assays using a particular technology) or that may be promiscuous for a certain target class.

As BAO includes concepts for targets, technologies, detection, etc (see above), perturbagen subclasses of interest can be directly defined in the ontology using the same approach; e.g. compounds that are promiscuously active in luciferase reporter gene assays. The individuals that are members of such

[16] http://jena.sourceforge.net/ARQ/group-by.htm

[17] All results are individuals with a working URI. URIs are abbreviated due to space limitations; e.g. the complete URI to the first result is http://www.bioassayontology.org/bao#individual_BAO_0000021 _646704.

a class are automatically inferred using the current curated assays (with their BAO annotations).

These three examples illustrate some of the features that can be used in complex search queries with an underlying DL-based ontology. Other features such as role hierarchies, quantifiers, nominals etc. were also used in our ontology.

3 Summary

We have developed an ontology for the purpose of analyzing biological assay and screening data with semantic information. 300 PubChem assays were curated and 194 were loaded in the ontology. The ontology was published in its first version (0.9) and is available at http://bioassayontology.org. This is the first ontology to describe this domain, and certainly the first time that bioassay and HTS data have been represented using expressive description logic. There are numerous advantages to this approach; most importantly it opens new functionality for querying and analyzing HTS data sets and the potential for discovering knowledge that is not explicitly stated by inference.

Using large sets of empirical data such as those in the BAO repository, such knowledge can be uncovered. The current repository will grow to well over a billion records that are available in the triple store. Because the results are subject to reasoning, the system has hit an upper bound limit on the number of triples that can be handled by a reasoner. At the moment the system is capable of operating on a scale of millions of triples with reasoning. This project will also provide a foundation of real (life science) data to improve reasoning algorithms and develop novel solutions to efficiently operate on very large data sets.

We are currently in the process of creating a web portal with an easy-to-use querying interface that incorporates this functionality. A user will be able to query data from PubChem (and other databases) using BAO terminology and collect groups of results for further analysis. It will also allow end users to formulate their own queries via a graphical user interface. Future developments will include an annotation tool for domain experts that will aid in the curation process and the incorporation of additional data sources.

In the future BAO will also enable integration with orthogonal life science databases such as biological pathways, diseases or adverse drug reactions and ultimately facilitate the discovery of new biomedical knowledge

Acknowledgement

This project is funded by the NIH under the grant number NHGRI (1RC2HG00566801). The authors thank Mark Southern for contributions to the project. We also acknowledge resources of the Center for Computational Science at the University of Miami. Vance Lemmon holds the Walter G. Ross Distinguished Chair in Developmental Neuroscience.

References

1. Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hil, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.
2. R.T. Fielding. *Architectural styles and the design of network-based software architectures.* PhD thesis, Citeseer, 2000.
3. J. Rogers and A. Rector. Galen's model of parts and wholes: experience and comparisons. *Proc AMIA Symp*, pages 714–718, 2000.
4. K. A. Spackman, K. E. Campbell, and R. A. Cote. Snomed rt: a reference terminology for health care. *Proc AMIA Annu Fall Symp*, pages 640–644, 1997.
5. Y. Wang, E. Bolton, S. Dracheva, K. Karapetyan, B. A. Shoemaker, T. O. Suzek, J. Wang, J. Xiao, J. Zhang, and S. H. Bryant. An overview of the pubchem bioassay resource. *Nucleic Acids Res*, 38(Database issue):D255–66, 2010.

A Minimal requirements

1. The application has to be an end-user application, i.e. an application that provides a practical value to general Web users or, if this is not the case, at least to domain experts.
   GIVEN

2. The information sources used should be under diverse ownership or control should be heterogeneous (syntactically, structurally, and semantically), and should contain substantial quantities of real world data (i.e. not toy examples).
   all GIVEN

3. The meaning of data has to play a central role. Meaning must be represented using Semantic Web technologies. Data must be manipulated/processed in interesting ways to derive useful information and this semantic information processing has to play a central role in achieving things that alternative technologies cannot do as well, or at all;
   OWL 2.0 ontology, 460 classes, please see description for details

B Additional Desirable Features

In addition to the above minimum requirements, we note other desirable features that will be used as criteria to evaluate submissions.

- The application provides an attractive and functional Web interface (for human users)
  YES

- The application should be scalable (in terms of the amount of data used and in terms of distributed components working together). Ideally, the application should use all data that is currently published on the Semantic Web.
  YES

- Rigorous evaluations have taken place that demonstrate the benefits of semantic technologies, or validate the results obtained.
  Currently ongoing with project team and domain experts, planned in near future (Q1-2/2011) with end-users

- Novelty, in applying semantic technology to a domain or task that have not been considered before
  First ontology for bioassays, big impact potential

- Functionality is different from or goes beyond pure information retrieval
  Curation, annotation of future bio assay experiments, potential statistical learning. The application has clear commercial potential and/or large existing user base.

- Contextual information is used for ratings or rankings
  Not yet

- Multimedia documents are used in some way
  No

- There is a use of dynamic data (e.g. workflows), perhaps in combination with static information
  Static ontology is used for curation workflow. New knowledge is then added to static ontology.

- The results should be as accurate as possible (e.g. use a ranking of results according to context)
  Not yet

There is support for multiple languages and accessibility on a range of devices
Not at this time