Biositemaps Status and Plans

March 1, 2013
Stanford Biositemaps Team

Summary

This document describes the current state of the Biositemaps software and outlines possible future work. The Biositemaps Web site is up and the software is now running at Stanford. The central Biositemaps file repository is now hosted at Stanford as well. There is no longer any danger that the Biositemaps system might go down and become permanently unrecoverable.

We have identified a number of changes to the Biositemaps system that we need to make to improve its usability in the short term. In addition, we have begun exploring both what it would mean and what it would take to merge the Biositemaps and eagle-i systems. We have developed a concrete vision and proposal for integrating the two architectures into a single seamless system.

Completed Work

Biositemaps Software Redeployment

The Biositemaps software is now up and running on Stanford servers. This software includes the Biositemaps file creation and editing system, the Biositemaps Registry, and the Biositemaps Resource Discovery System (RDS) search tool. We also have moved the software to a source code control system at Stanford (https://bmir-gforge.stanford.edu/) and we have recompiled and redeployed this software. The software code base is complicated and consists of roughly 110,000 lines of code in a number of different modules. We have created an automated monitoring system for the production Biositemaps software that will notify us in real time if the software should become unavailable for any reason. This system makes calls to Biositemaps every few minutes and verifies that the system is still responsive. We list the complete set of Biositemaps components and their current disposition in Appendix 1.

The **biositemaps.org** domain name now points to the instance of the Biositemaps software at Stanford. The University of Pittsburgh instance of Biositemaps is still running, but this instance is now a backup. The **biositemaps.ncbcs.org** domain is now redirected to the canonical **biositemaps.org** domain.

Biositemaps Evaluation and Fixes

We have fixed a serious bug in the RDS system, wherein a user could not filter resource search results by resource type. The system now supports searching by resource type. The system now retrieves the available resource types from the latest version of the Biomedical Resource Ontology (BRO) in the NCBO BioPortal.

We have also identified the following usability issues in the RDS software:

- Filtering is very slow when one combines free text search with resource types
- Filtering does not update the matching resource numbers in parentheses

- The "Resource Type" facet shows only the top three levels of BRO. There is no way to directly filter for more specific resource types
- The display of multi-line items in the facet browsers (at the left hand side of RDS) is very confusing. The display should use the standard "hanging indentation" style typically seen for multi-line items in lists. This change would cause the first line to stand out to the left and allow for easier scanning.

In addition to evaluating the RDS UI, we also evaluated the code and documentation for all the Biositemaps software components. The Biositemaps API is well documented and well crafted. The RDS, on the other hand, is poorly documented and does not appear to be well written. As an example, both the RDS system and the older Biositemaps search system use the same underlying search API. Nevertheless, the RDS search is notably slower than that of the older system. While we can make incremental improvements to the RDS system, fixing all of the problems would require a substantial amount of rework and would require significant development resources.

Biositemaps and eagle-i Investigation

One possibility that we explored in our previous report on Biositemaps is the possibility of using some or all of the eagle-i infrastructure for creating and browsing Biositemaps content. The eagle-i search tool is both more user-friendly and works much better than the current Biositemaps RDS. As part of our work, we have downloaded and rebuilt the eagle-i software system and now have it running internally on a Stanford development system. eagle-i is a large system comprising more than 255,000 lines of code. We have made preliminary exploration of the source code and we believe that we have identified the places in the code where adapters could be placed to allow import and export of Biositemaps files.

We have continued thinking about how to bridge the "impedance mismatch" between the eagle-i ontology and the BRO. In order to better define the scope of the problem, we have made a list of all BRO terms that appear in any Biositemaps file. The BRO contains 485 terms but we have found that only 233 appear in any current file. Thus, it may be possible to simplify the mapping between the two systems. It may be possible to replace the "potential" problem of aligning 485 terms from BRO to eagle-i with instead the "actual problem" of mapping 233 terms. We list the terms used in Biositemaps files in Appendix 2.

The Biositemap information model also needs to be aligned with the eagle-i data model. These information models describe the properties that may appear on a resource. We reviewed an early alignment made by members of the Biositemaps and eagle-i teams from three years ago. While much of this work is still valid, additional alignment will be needed since both systems have evolved since the previous work was done.

Possible Future Work

Biositemaps Software Improvements

One path forward is simply to fix the RDS usability problems. Some of these issues such as the multi-line display are straightforward to address. Others, such as improving search performance, will take more time. A more detailed exploration of the application, and perhaps some feedback from users, may result in us adding items to this list.

Biositemaps and eagle-i Integration

To evaluate the possibility of leveraging the eagle-i infrastructure for Biositemaps, we have downloaded, installed, and explored the eagle-i software. We have identified likely spots in the code that may require modification for Biositemaps compatibility. However, we would need to meet with the eagle-i software authors at Harvard to review our conclusions, brainstorm alternatives, and arrive at an implementation plan and schedule if we are to pursue the possibility of building on the eagle-i code base. We have delayed holding this meeting because of the uncertainty related to the funding of this effort. To be productive, this meeting should take place when there is clear agreement on organizational responsibilities moving forward. Our understanding is that no group has current explicit plans or funding for additional eagle-i software development. Thus we assume that any necessary software changes may need to be carried out at Stanford with guidance from the team at Harvard.

Our vision is a system in which eagle-i can import existing Biositemaps resource files and allow a user to modify them in the eagle-i editor. The files will then be available to the eagle-i search system in the same way that native eagle-i resources are. In addition to being able to import Biositemaps resource files, the system should be able to export a Biositemaps file for a given set of resources. The user can then place this file on a Web site (as occurs with the Biositemaps files today) and register this location with the biositemap registry. Such a system would provide a seamless, two-way integration of the Biositemaps and eagle-i systems. It would retain the distributed nature of the biositemaps system while allowing biositemaps users to take advantage of the eagle-i authoring and search interfaces.

In order to implement this vision, it is necessary to develop a two-way mapping between the Biomedical Resource Ontology (BRO) and the eagle-i research resource ontology (ERO). These ontologies have been developed independently over several years and, as one would expect, there are notable differences between them. In some cases there is a direct one-to-one mapping between BRO terms and eagle-i terms, sometimes with the same name, sometimes with synonyms for the same concept. For example, the term *software* is identical in both ontologies. The term *funding resource* in BRO is very close to the term *funding organization* in the eagle-i ontology. In other cases, BRO contains terms that are not in eagle-i, often because BRO contains a more granular model for some areas. For example, the term *software* in the eagle-i ontology has only two subclasses, defining specific kinds of software. In BRO, the term *software* has an entire hierarchy of dozens of classes underneath, distinguishing among many types of software. In each case, it is necessary to evaluate whether the BRO distinction is important to maintain. If so, then a corresponding concept can be added to the eagle-i ontology. If the BRO distinction is not important, then it is possible to map the BRO term to a higher-level concept in the eagle-i ontology.

Appendix 1: Biositemaps Components and Source Code

System Components

The Biositemaps software system consists of the following components:

- biositemap.org homepage
- Biositemaps editor and browser (for creating and editing Biositemaps files)
- Biositemaps Registry (for registering the location of each biositemap file)
- public file server hosting Biositemap files for others
- basic Biositemaps search interface
- Biositemaps Resource Discovery System (RDS) search tool
- Biositemaps Web services

Source Code

The Biositemaps source code is organized into the following repositories and modules.

- The original Biositemaps project is available at
  https://bmir-gforge.stanford.edu/gf/project/Biositemaps/
  and contains the following modules:

  - **BiositemapAPI** - a simple API for handling individual biositemap files (including creation, modification, querying, and saving)
  - **BiositemapEditor -** code for Biositemap Editor and Biositemap Browser using GWT
  - BiositemapWebServices
  - BioPortal-ResourceGenerator

- The Biositemaps Reference Implementation is available at:
  https://bmir-gforge.stanford.edu/gf/project/Biositemaps_ref/
  and contains the following modules:
  - **api** - Biositemaps search api and command line tools
  - **cgi** - Perl script for the Biositemaps registry
  - editor
  - **search** - basic Biositemaps search tool

- The RDS (previously called CIRWP) is available at
  https://bmir-gforge.stanford.edu/gf/project/cirwp_project/
  and contains the following modules:
  - SparqlConnection
  - cirwp_rdfloader
  - cirwp_utils
  - **cirwp_webservice** Web service that uses the Biositemaps search API to provide information (also to RDS) about the content of all the registered biositemaps
  - **cirwp_webui** - RDS user interface

Appendix 2: BRO to eagle-i Mapping Discussion

The 233 terms listed below are from the Biomedical Resource Ontology and have been used in at least one of the existing Biositemaps files. In order to move Biositemaps files into eagle-i, a mapping must exist from each of these concepts into a concept in eagle-i resource ontology (ERO). The BRO contains 485 classes, but the remaining classes are not used in any existing Biositemaps resource. Thus, it may not be essential to map all these remaining classes into their eagle-i equivalents (if any). The classes marked with "(M)" are identical to ERO classes with the same name, and mappings between these concepts already exist in BioPortal.

BRO:Aggregate_Human_Data
BRO:Algorithm (M)
BRO:Animal_Models_Facility
BRO:Application_Programming_Interface
BRO:Assessment_Material_Resource
BRO:Atlas_Generation
BRO:Backup_Archives_and_Disaster_Recovery
BRO:Biobank_Facility
BRO:Biomaterial_Manufacture
BRO:Biomaterial_Supply_Resource
BRO:Biomedical_Supply_Resource
BRO:Biosafety_Level_Facility
BRO:CT_Scanner (M)
BRO:Calculation_of_Solvent_Accessible_Area
BRO:Calendar_Schedule_and_Resource_Management
BRO:Cardiovascular_Model
BRO:Cell_Biology_Facility
BRO:Cell_Culture_Facility
BRO:Cell_Model
BRO:Certificate_Program
BRO:Clinical_Care_Data
BRO:Clinical_Data
BRO:Clinical_Research_Data
BRO:Clinical_Trial_Data
BRO:Communication_Interface
BRO:Communication_and_Collaborative_Work
BRO:Computational_Geometry
BRO:Computational_Model
BRO:Computational_Service
BRO:Contact_Modeling
BRO:Continuum_Method_Simulation
BRO:Controllers
BRO:Cortical_Modeling
BRO:Data_Acquisition_Software
BRO:Data_Analysis_Software
BRO:Data_Annotation
BRO:Data_Computation_Service
BRO:Data_Exploration
BRO:Data_Mining_and_Inference
BRO:Data_Processing_Software
BRO:Data_Repository
BRO:Data_Resource
BRO:Data_Service
BRO:Data_Storage (M)
BRO:Data_Storage_Repository
BRO:Data_Storage_Service (M)
BRO:Data_Transfer_and_Communication
BRO:Data_Transforms
BRO:Data_Visualization (M)
BRO:Database (M)
BRO:Database_Software
BRO:Delimited_Table
BRO:Dissemination_Vehicle

BRO:Metabolomics_Facility
BRO:Microarray_Scanner
BRO:Microscope (M)
BRO:Microscopy_Facility
BRO:Model_Fitting_Algorithm
BRO:Modeling_and_Simulation
BRO:Modular_Component
BRO:Molecular_Biology_Facility
BRO:Molecular_Dynamics
BRO:Molecular_Force_Field_Calculator
BRO:Molecular_Interaction
BRO:Molecular_Model
BRO:Molecular_Modeling_and_Classification
BRO:Molecular_Trajectory_Data
BRO:Molecular_Visualization
BRO:Molecular_and_Cellular_Data
BRO:Monte_Carlo_Simulation
BRO:Multibody_Dynamics
BRO:NMR_Facility
BRO:Narrative_Resource
BRO:Natural_Language_Processing
BRO:Network_Characterization
BRO:Network_Interaction_Model
BRO:Network_Model
BRO:Network_and_Communication
BRO:Neuromuscular_Model
BRO:Next_Generation_Sequencer (M)
BRO:Numerical_Calculation_of_Electrostatic_Potential
BRO:Numerical_Integrator
BRO:Numerical_Method
BRO:Numerical_Model
BRO:Ontology
BRO:Ontology_Development_and_Management
BRO:Ontology_Visualization
BRO:Optimizer
BRO:Outreach_Program
BRO:PCR_Facility
BRO:PCR_Instrument (M)
BRO:PET_Scanner (M)
BRO:Paper
BRO:Pathology_Laboratory_Screening
BRO:Pattern_Inference_Algorithm
BRO:Pattern_Recognition
BRO:Pattern_and_Motif_Inference
BRO:People_Resource
BRO:Pharmacokinetics_Pharmacodynamics_Expertise
BRO:Phenotypic_Measurement
BRO:Phlebotomy_Facility
BRO:Physico-Chemical_Model
BRO:Physiology_Facility
BRO:Physioloigcal_Model
BRO:Pipeline_Manager
BRO:Portal

BRO:Document_Retrieval
BRO:Dynamic_Model
BRO:Education
BRO:Electron_Microscope (M)
BRO:Experimental_Measurement
BRO:Exploratory_Data_Analysis
BRO:Fabrication_Facility
BRO:Facility_Core
BRO:Faraday_Cage (M)
BRO:Feature_Analysis
BRO:Federal_Funding_Resource
BRO:Finite_Element_Model
BRO:Flow_Cytometry_Cell_Sorting_Facility
BRO:Fourier_Transform
BRO:Funding_Resource
BRO:Gene_Expression
BRO:Genomic_Phenotypic_Analysis
BRO:Genomics_Facility
BRO:Grant_Preparation_Expertise
BRO:Graph_Analysis
BRO:Graph_Viewer
BRO:Graph_Viewers
BRO:Graphical_Integration
BRO:Grid_Computing (M)
BRO:Homology_Modeling
BRO:Human_Studies_Compliance_Resource
BRO:Image (M)
BRO:Image_Algorithm
BRO:Image_Processing
BRO:Imaging
BRO:Imaging_Facility
BRO:Immune_Monitoring_Facility
BRO:Immunohistochemistry_Facility
BRO:Individual_Human_Data
BRO:Industry_Partnership_Expertise
BRO:Inference_From_Data
BRO:Information_Resource
BRO:Information_Retrieval
BRO:Instrument (M)
BRO:Instrument_Manufacture
BRO:Integration_and_Interoperability_Tool
BRO:Interaction_Modeling
BRO:Interaction_Network
BRO:Interactive_Network_Analysis
BRO:Interactive_Tool
BRO:Interactive_Web-Based_Tool
BRO:Journal_Article (M)
BRO:Knowledge_Environment
BRO:Knowledge_Extraction
BRO:Knowledge_Mining_and_Capturing
BRO:Knowledgebase
BRO:Laboratory_Supply_Resource
BRO:Light_Microscope
BRO:Linear_Algebra_Tool
BRO:MRI_Scanner
BRO:Manifold_Viewers
BRO:Mass_Spectra_Identification
BRO:Mass_Spectrometer (M)
BRO:Material_Resource
BRO:Material_Service
BRO:Mechanical_Simulation
BRO:Medical_Device (M)
BRO:Mesh_Model
BRO:Metabolism_Facility

BRO:Prediction_of_Side-Chain_Conformations
BRO:Processing_Pipeline
BRO:Protein-Protein_Interaction
BRO:Protein_Expression
BRO:Protein_Interaction_Modeling
BRO:Protein_Model
BRO:Proteomics_Facility
BRO:Protocol_Development_Expertise
BRO:Psychometrics_Expertise
BRO:Publication
BRO:RNA_Model
BRO:Radioisotopes_Facility
BRO:Reagent_Manufacture
BRO:Reagent_Resource
BRO:Registration
BRO:Regulatory_Compliance_Resource
BRO:Regulatory_Policy_Resource
BRO:Regulatory_Signaling_Network_Reconstruction
BRO:Relational_Database
BRO:Research_Animals_Facility
BRO:Resource
BRO:Robotic_Arrayer
BRO:Searching_Sorting_and_Indexing
BRO:Segmentation
BRO:Seminar_Series
BRO:Sequence_Annotation
BRO:Sequence_Similarity_Searching
BRO:Sequence_Visualization
BRO:Service_Resource
BRO:Shape_Analysis
BRO:Signal_Processing
BRO:Signaling_Network_Reconstruction
BRO:Software (M)
BRO:Software_Development_Tool
BRO:Software_Distribution
BRO:Software_Documentation
BRO:Source_Code
BRO:Standalone_Application
BRO:Statistical_Algorithm
BRO:Statistical_Analysis
BRO:Statistical_Package
BRO:Structural_Model
BRO:Structure-Based_Protein_Classification
BRO:Structured_Data
BRO:Structured_Knowledge_Resource
BRO:Support
BRO:Technical_Support
BRO:Three_D_Image
BRO:Three_Dimensional_Data
BRO:Time_Series_Analysis
BRO:Tissue_Organ_Facility
BRO:Toolkit
BRO:Toxicology_Expertise
BRO:Training_Resource
BRO:Training_Service (M)
BRO:Two_D_Image
BRO:Visualization (M)
BRO:Wavelet_Transform
BRO:Web_Service
BRO:Website
BRO:X-Ray_Crystallography_Facility
BRO:XML_Data
BRO:XML_Database