eagle-i: ontology-driven federated search and data entry tools for discovering biomedical research resources

Melissa Haendel[1,*] Carlo Torniai[1], Nicole Vasilevsky[1], Scott Hoffmann[1],
and Daniela Bourges-Waldegg[2]
[1] Oregon Health & Science University, Portland, Oregon, USA
[2] Harvard Medical School, Boston, Massachusetts, USA

ABSTRACT

In this paper we present the eagle-i system architecture, which was designed for data entry and storage of semantically enabled and ontology-driven data about biomedical research resources. We describe the content, the features and some of the applications currently leveraging eagle-i datasets.

1 INTRODUCTION

An important facet of biomedical research is access to the right tools and research resources needed to answer specific biological questions. Biomedical resources are generated, purchased, and used during the course of research. Information about such resources is often sequestered in lab notebooks or a lab's files, making the resources difficult to find, share, and reuse. The goal of the eagle-i Network (http://www.eagle-i.net) is to make these "invisible" research resources more discoverable by collecting information about them and making the information available in a semantically enabled, federated search system.

2 EAGLE-I SYSTEM FEATURES

The eagle-i system architecture is composed of four main components: an underlying ontology; data collection tools comprising an ETL toolkit and a web-based, ontology driven Semantic Web Entry and Editing Tool (SWEET); institutional triple-store repositories; and a central web-based search application.

To support structured data collection and information retrieval, the underlying data model utilizes a modular set of ontologies, collectively known as the eagle-i Resource Ontology (ERO). The ontology contains domain representation for research resources such as organisms, instruments, constructs, antibodies, biospecimens, human studies and research opportunities. An application-specific ontology layer containing annotations to the core ontology drives the application functionality (Torniai *et al.*, 2011). Requests for ontology additions and modifications are collected from the community via a googlecode issue tracker (http://code.google.com/p/eagle-i/). At regular intervals (coinciding with new software and ontology releases), the requested changes are compiled, tested, and released to the public.

The eagle-i system relies on rich data describing the research resources. The eagle-i team developed the SWEET, a graphical front-end to an eagle-i repository through which resource information is entered and curated. The SWEET features online forms with four types of fields (see **Fig. 1.**).

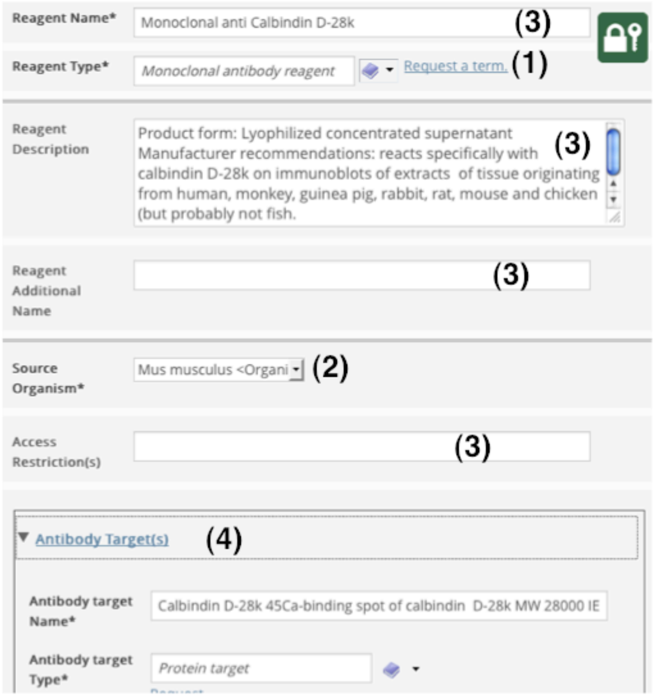
(1) A **dropdown of hierarchical terms that are derived from specific ontology branches** (representing specific resource subtypes or referenced taxonomies such as diseases or anatomy partonomy)
(2) A **drop-down menu for resources** (other instances) that exist in the repository
(3) **Free text fields**
(4) **Fields for describing related resources** called embedded instances

Field names are representative of object and datatype properties. Annotation properties drive preferred labels and tool tips for annotation consistency. In order to avoid the computational expense of dealing with the upper-level terms and axioms of the ontology, and to provide an end-user-friendly view of the ontology, a data model is generated. This data model is constructed with only the classes and properties necessary to represent data in the data store. The simplicity of the typical SWEET user's data entry experience belies the complexity of the ontology integration and data storage processes happening behind-the-scenes.

The data generated through SWEET are stored as triples in accordance with the Resource Description Framework (RDF) model and made available through SPARQL endpoints (https://www.eagle-i.net/export/sparqlers/). The data are available for consumption as Linked Data for those interested in representing their resources in a highly searchable and readily edited format. The SWEET offers an access point for easy URI assignment as well as a portal to a permanent home for RDF data for many different types of resources.

---

[*] To whom correspondence should be addressed: haendel@ohsu.edu



Reagent Name*	Monoclonal anti Calbindin D-28k
Reagent Type*	*Monoclonal antibody reagent*
Reagent Description	Product form: Lyophilized concentrated supernatant Manufacturer recommendations: reacts specifically with calbindin D-28k on immunoblots of extracts of tissue originating from human, monkey, guinea pig, rabbit, rat, mouse and chicken (but probably not fish.
Reagent Additional Name	
Source Organism*	Mus musculus <Organi
Access Restriction(s)	
Antibody target Name*	Calbindin D-28k 45Ca-binding spot of calbindin D-28k MW 28000 IE
Antibody target Type*	*Protein target*

Fig. 1. A sample form used to enter an antibody resource in the SWEET. Fields marked with **(1)** represent dropdowns of term hierarchies from specific ontology branches. The field annotated **(2)** allows links to existing resource instances (in this case, organisms) in the repository. **(3)** denotes free text fields and **(4)** a collection of fields containing information about a linked, embedded resource instance. Embedded instances must be tied to a parent resource (antibody target to monoclonal antibody reagent, for example).

Beginning with nine sites, the eagle-i Network has grown to include more than 25 institutions. Each of these institutions installs and maintains their own SWEET instance along with the RDF data it produces. The eagle-i Search Tool draws on the data present in each of the federated institutional data repositories and the shared node. Like the SWEET, it also makes use of a slim version of ERO and its application ontology in the data model. Auto-complete functionality allows users to match their text search string to ontology terms (or their synonyms) or to the resource instances and return relevant results. Faceted browsing lets the user filter search results by resource type or subtype.

3 BROADER USE FOR EAGLE-I GENERATED DATA

A number of groups have begun to make use of eagle-i-produced data by implementing search and visualization tools that reuse the RDF. An early adopter and user of the eagle-i data was the Harvard Catalyst Core Facilities Portal (http://cbmi.catalyst.harvard.edu/cores/index.html). This portal component generates HTML pages from the core facilities data stored at Harvard University's eagle-i repository. The automated production of these pages ensure that the service offerings as well as the contact information for each core are standardized, centralized and current, as they are maintained through the eagle-i SWEET. A similar reuse of eagle-i RDF data using the Plumage tool at Oregon Health & Science University allows for compilation, visualization and enhanced search of specifically core laboratories and their service offerings (http://www.ohsu.edu/research/coresearch/). The Plumage tool, developed by the Clinical & Translational Science Institute at the University of California, San Francisco (http://ctsiatucsf.github.com/plumage/), converts the RDF data to static HTML pages that can be optimally indexed by Google and other search engines.

4 CHALLENGES

Efforts on the development of the eagle-i software suite and ERO continue. Evolution of these integrated pieces is driven largely by the goal of improved user experience. Challenges being addressed include:

- **Limiting variation in user-entered data:** Simplification of the user interface, ontology data model, and improved help documentation
- **Allowing ontology flexibility by institution:** Planned implementation of site-specific ontologies and process to centrally monitor proposed terms

As work in these areas proceeds, eagle-i and ERO provide an effective open-source framework for the collection, storage, and publication of research resources.

ACKNOWLEDGEMENTS

This work has been funded from the National Center for Research Resources (NCRR), part of the National Institutes of Health (NIH). Grant Number #U24 RR 029825.

REFERENCES

Torniai C., Brush M., Vasilevsky N., Segerdell E., Haendel M., Shaffer C., Johnson T., Corday K. (2011). Developing an application ontology for biomedical resource annotation and retrieval: Challenges and lessons learned, International Conference of Biomedical Ontologies (ICBO).