eagle-i: biomedical research resource datasets

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Carlo Torniai[a,*], Daniela Bourges-Waldegg [b] and Scott Hoffmann[a]
[a] *Oregon Health & Science University, Portland, Oregon, USA*
[b] *Harvard University, Boston, Massachusetts, USA*

Abstract. In this paper we present the linked data sets produced by the eagle-i project. We describe the content, the features and some of the applications currently leveraging these datasets.

Keywords: eagle-i, linked data, biomedical ontologies, research resources, research profiles

1. Introduction

An important facet of biomedical research is to access 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 digital records, making the resources difficult to find, share, and reuse. The goal of eagle-i, (http://www.eagle-i.net) - a two-year project funded by National Center for Research Resources (NCRR) - is to make these "invisible" research resources more discoverable by collecting information about them and making the information available through a semantically enabled, federated search system and as linked data sets.

The eagle-i 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, retrieval and publication, a modular set of ontologies, collectively known as the eagle-i Resource Ontology (ERO) was developed. The ontology contains domain representation for research resources including organisms, instruments, protocols, constructs, antibodies, biospecimens, human studies and research opportunities.

The data generated through eagle-i's data collection tools are stored as triples in accordance with the Resource Description Framework (RDF) and made available through SPARQL endpoints (https://www.eagle-i.net/export/sparqlers/) and as linked data sets. In this paper we present the key characteristics of the datasets, provide sample SPARQL queries and discuss some of the known usages of our datasets.

2. Dataset Descriptions

Our datasets cover biomedical resources available at 25 institutions. The main resource types we collect and publish information about are:

- Biological Specimen
- Database
- Document
- Human Study
- Instrument
- Organism or Virus
- Organization
- Person
- Reagent
- Research Opportunity

[*] Corresponding author. E-mail: torniai@ohsu.edu.

- Service
- Software

The datasets can be queried through SPARQL endpoints (https://www.eagle-i.net/export/sparqlers/), can be browsed using classic and linked data browsers and are available via direct RDF download (https://www.eagle-i.net/export/rdf-download/). Full information about each dataset is available at the Data Hub (http://datahub.io/dataset?q=eagle-i). In Table 1, we present a summary of the triples in each dataset; these numbers include only the triples that are directly related to resource instances, that is (a) triples where the subject is a resource instance and (b) triples that provide minimal information (type and label predicates) for the objects in (a). We deliberately exclude from these datasets most ERO ontology triples, as they do not *per-se* represent biomedical resources. Table 2 gives an overview of the number of triples across all participating institutions devoted to each of the resource types.

A URI for a resource resolves as an HTML page when accessed by a browser. This page displays all the data about the resource and its inferred types as well as a link to the RDF download (shown in Figure 1).

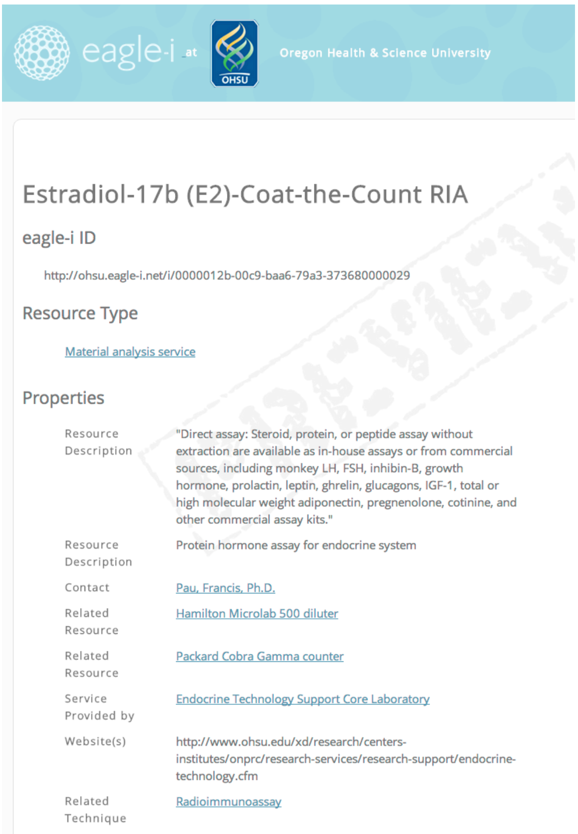
Tables 1 and 2

Number of triples published per site (Table 1) and per resource type (combined for all institutions) (Table 2) as of March 5, 2013.

* - Denotes status as member institution of original eagle-i network # - Denotes status as currently funded, CTSA eagle-i member institution

Data set	**Published triples**
University of Pennsylvania#	21866
Vanderbilt University#	19536
Oregon Health & Science University*#	76818
Harvard University*#	372288
University of Alaska Fairbanks*	15446
University of Hawai'i Manoa*	347219
Jackson State University*	10501
Montana State University*	24737
Morehouse School of Medicine*	8763
Dartmouth College*	372288
University of Puerto Rico*	35733
Clark Atlanta University	928
Charles Drew University	3940
The City College of New York, CUNY	462
Florida Agricultural and Mechanical University	2761
Howard University	2944
Hunter College, CUNY	1352
Meharry Medical College	1805
Ponce School of Medicine	2001
Texas Southern University	1474
Tuskegee University	927
Universidad Central del Caribe	5477
University of Texas at El Paso	2773
University of Texas at San Antonio	4050
Xavier University of Louisiana	9395
TOTAL	**1345484**

Resource type	**Published triples**
Biological specimens	48317
Databases	421
Documents	68890
Human studies	6491
Instruments	125696
Organisms or viruses	456107
Organizations	950930
People	180554
Reagents	177543
Research opportunities	1078
Services	62501
Software	27291



eagle-i at Oregon Health & Science University

Estradiol-17b (E2)-Coat-the-Count RIA

eagle-i ID

http://ohsu.eagle-i.net/i/0000012b-00c9-baa6-79a3-373680000029

Resource Type

Material analysis service

Properties

Resource Description	"Direct assay: Steroid, protein, or peptide assay without extraction are available as in-house assays or from commercial sources, including monkey LH, FSH, inhibin-B, growth hormone, prolactin, leptin, ghrelin, glucagons, IGF-1, total or high molecular weight adiponectin, pregnenolone, cotinine, and other commercial assay kits."
Resource Description	Protein hormone assay for endocrine system
Contact	Pau, Francis, Ph.D.
Related Resource	Hamilton Microlab 500 diluter
Related Resource	Packard Cobra Gamma counter
Service Provided by	Endocrine Technology Support Core Laboratory
Website(s)	http://www.ohsu.edu/xd/research/centers-institutes/onprc/research-services/research-support/endocrine-technology.cfm
Related Technique	Radioimmunoassay

Fig. 1. The information displayed by a browser when requesting the following resource URI [http://ohsu.eagle-i.net/i/0000012b-00c9-baa6-79a3-373680000029]. It describes a protein hormone assay for endocrine system.

3. Domain modeling

Our approach to domain modeling was driven by the following goals:

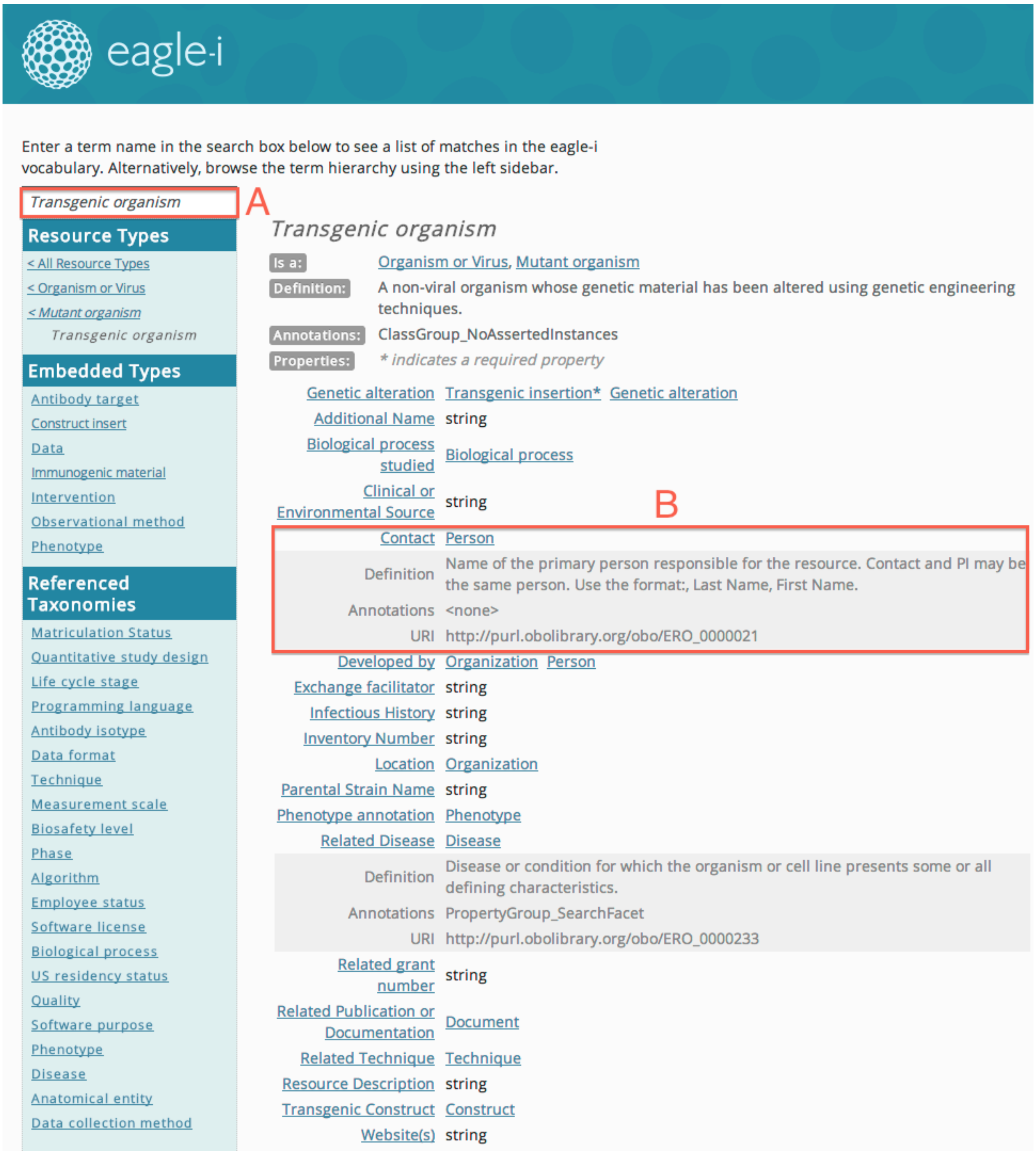
1. Reuse existing ontologies as much as possible to reduce the modeling burden and to maximize future data integration
2. Identify design patterns and ontology engineering solutions that would allow a set of ontologies to drive the eagle-i user interfaces while at the same time remaining of general use to the biomedical community at-large.

Regarding the first point, because most of our domain coverage was biomedical in nature, we referred to principles and existing ontologies within the OBO Foundry [1]. Conformance with OBO Foundry standards fixed the following design choices:

- Use of the Basic Formal Ontology [2] as the upper level ontology
- Predominant utilization of ontologies in the OBO Foundry constellation due to their quality, extensive usage and common design principles
- Application of the MIREOT principle [3] for referencing entities in external ontologies

So far as the second goal is concerned, we developed a design pattern approach to separate, within our ontology suite, the application-specific portion from the "core" content that was worth sharing with the community. Our approach, described extensively in [4], has been generalized and reused in other efforts such as the Reagent Ontology (ReO https://code.google.com/p/reagent-ontology/) and the Agent, Resource and Grant ontology (ARG https://code.google.com/p/connect-isf/). It has also led to a set of recommendations for implementing a maintenance and release pipeline using available tools and service [5].

Another key element of our ontology development process has been the coordination of efforts within the Biomedical Ontology Community. These include active collaboration and discussion with other ontology development groups (through tracker term requests, developer call participation, etc.). Although time consuming, these efforts allow easy reuse of portions of other ontologies and help achieve better data integration and interoperability.



Enter a term name in the search box below to see a list of matches in the eagle-i vocabulary. Alternatively, browse the term hierarchy using the left sidebar.

Transgenic organism

Resource Types

- < All Resource Types
- < Organism or Virus
- *< Mutant organism*
- *Transgenic organism*

Embedded Types

- Antibody target
- Construct insert
- Data
- Immunogenic material
- Intervention
- Observational method
- Phenotype

Referenced Taxonomies

- Matriculation Status
- Quantitative study design
- Life cycle stage
- Programming language
- Antibody isotype
- Data format
- Technique
- Measurement scale
- Biosafety level
- Phase
- Algorithm
- Employee status
- Software license
- Biological process
- US residency status
- Quality
- Software purpose
- Phenotype
- Disease
- Anatomical entity
- Data collection method

Transgenic organism

Is a: Organism or Virus, Mutant organism

Definition: A non-viral organism whose genetic material has been altered using genetic engineering techniques.

Annotations: ClassGroup_NoAssertedInstances

Properties: ** indicates a required property*

Genetic alteration	Transgenic insertion* Genetic alteration
Additional Name	string
Biological process studied	Biological process
Clinical or Environmental Source	string
Contact	Person
Definition	Name of the primary person responsible for the resource. Contact and PI may be the same person. Use the format:, Last Name, First Name.
Annotations	<none>
URI	http://purl.obolibrary.org/obo/ERO_0000021
Developed by	Organization Person
Exchange facilitator	string
Infectious History	string
Inventory Number	string
Location	Organization
Parental Strain Name	string
Phenotype annotation	Phenotype
Related Disease	Disease
Definition	Disease or condition for which the organism or cell line presents some or all defining characteristics.
Annotations	PropertyGroup_SearchFacet
URI	http://purl.obolibrary.org/obo/ERO_0000233
Related grant number	string
Related Publication or Documentation	Document
Related Technique	Technique
Resource Description	string
Transgenic Construct	Construct
Website(s)	string

Fig. 2. Classes can be searched for using the autocomplete feature (A). In this case, we see information about ‘transgenic organism’. All of the properties for this class are shown. Clicking the property name displays its definition, URI and annotations as shown by feature (B). Referenced taxonomies are sets of terms used as ranges for some properties (such as the Disease taxonomy for the related disease property) while embedded types denote classes for which instances can be only created in the context of another instance. For example, a construct insert can only be created in conjunction with its containing construct.

4. Sample SPARQL queries

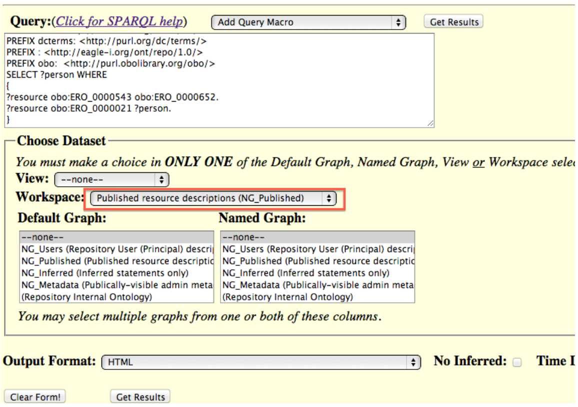
We present some sample queries that illustrate interesting usages of the eagle-i datasets. For each one, we specify the particular SPARQL endpoint used such that query results can be reproduced.

In order to better understand the classes and relationships used in the following queries we suggest referring to the eagle-i ontology browser (http://search.eagle-i.net/model/), a screen capture from which can be seen in Figure 2.

URIs of eagle-i classes and properties can also be found using Ontobee [6] (http://www.ontobee.org/browser/index.php?o=ERO) . The queries can be executed through the SPARQL interfaces (Fig.3) or passed programmatically to the endpoint.

In the following examples, we omit for brevity the declaration of the following prefixes:

```
  PREFIX rdfs:
<http://www.w3.org/2000/01/rdf-schema#>
  PREFIX obo:
<http://purl.obolibrary.org/obo/>
  PREFIX mesh:
<http://purl.bioontology.org/ontology/MSH/>
```



```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX : <http://eagle-i.org/ont/repo/1.0/>
PREFIX obo:  <http://purl.obolibrary.org/obo/>
SELECT ?person WHERE
{
?resource obo:ERO_0000543 obo:ERO_0000652.
?resource obo:ERO_0000021 ?person.
}
```

Choose Dataset

*You must make a choice in **ONLY ONE** of the Default Graph, Named Graph, View or Workspace selec*

View: --none--

Workspace: Published resource descriptions (NG_Published)

Default Graph:	Named Graph:
--none--	--none--
NG_Users (Repository User (Principal) descri	NG_Users (Repository User (Principal) descri
NG_Published (Published resource descriptic	NG_Published (Published resource descriptic
NG_Inferred (Inferred statements only)	NG_Inferred (Inferred statements only)
NG_Metadata (Publically-visible admin meta	NG_Metadata (Publically-visible admin meta
(Repository Internal Ontology)	(Repository Internal Ontology)

You may select multiple graphs from one or both of these columns.

Output Format: HTML

No Inferred: □

Fig. 3. The SPARQL query interface. In order to retrieve information about published resources the “Published resource description” (red box) must be selected from the Workspace dropdown menu.

4.1. *Query for researcher expertise*

The eagle-i datasets allow identification of individual expertise by leveraging the connection between resources and related techniques, diseases and instruments as well as the linkage between resources and people. As an example, the query below identifies likely experts in radioimmunoassay techniques (obo:ERO_0000652) by connecting the nodes between the technique and individuals. In this case, collecting all resources that reference the technique (obo:ERO_0000543) and returning the person(s) indicated as contact (obo:ERO_000021) for that resource.

```
SELECT DISTINCT? person WHERE
{

  ## Select Resources that have related
  ## technique radioimmunoassay

  ?resource obo:ERO_0000543 obo:ERO_0000652.

  ## Select the contact person for
  ## the resource

  ?resource obo:ERO_0000021 ?person.
}
```

If the query is executed against the OHSU endpoint (http://ohsu.eagle-i.net/sparqler/query/) it will return a contact for a set of services that involve radioimmunoassay like the one represented in Fig. 1.

4.2. *Query for animal models relevant for a particular disease*

Another interesting query is related to the identification of animal models used in the research of autoimmune diseases. The query is reported below.

```
SELECT ?resource WHERE
{
  ## Select Organism Resources

  ?resource a obo:OBI_0100026.

  ## That are model of some disease

  ?resource obo:ERO_0000233 ?disease.

  ## And the disease is an autoimmune
  ## disease

  ?disease rdfs:subClassOf mesh:D001327.
}
```

It is interesting to note that the results returned for this query, when executed against the Harvard endpoint (https://harvard.eagle-i.net/sparqler/query), include animal models related to Diabetes Mellitus, Type 1 (see for example: http://harvard.eagle-i.net/i/0000012a-25bf-7988-f5ed-943080000005) and Sjogren's Syndrome (http://harvard.eagle-i.net/i/0000012a-25bf-7988-f5ed-943080000003) because both are subsumed in the MeSH Hierarchy for Autoimmune disease (mesh:D001327).

4.3. Query for resources across datasets

For several resources in eagle-i (such as animal models) we collect information on related genes via Entrez gene IDs. This is a useful entry point for connecting non-eagle-i datasets. In the query below for instance, we probe data at the University of Puerto Rico (UPR) for resources relevant to Stony Brook investigators based on the genes they have published about. This query, when executed against the UPR endpoint (http://upr.eagle-i.net/sparqler/query/) return 4 authors related to a ErbB2 construct insert used in a particular plasmid (http://upr.eagle-i.net/i/0000012b-8e1f-e389-3bbe-1c0980000000).

```
SELECT ?entrezgeneid ?author ?resource
WHERE
{

## Query the SPARQL endpoint at Stony Brook

SERVICE
<http://link.informatics.stonybrook.edu/spar
ql/>

{
## Get the AUIs and the CUIs related to the
## entrez gene ids

?aui
<http://link.informatics.stonybrook.edu/umls
/ATN#ENTREZGENE_ID>
?entrezgeneid.
?aui rdfs:label ?label.
?aui
<http://link.informatics.stonybrook.edu/umls
/hasCUI> ?cui.

## Select the papers that have as subject
## the gene identified by the Entrez ID

?paper
<http://purl.org/dc/elements/1.1/subject>
?cui.

## Select the author of the paper

?paper
<http://vivoweb.org/ontology/core#informatio
nResourceInAuthorship> ?authorship.
?authorship
<http://vivoweb.org/ontology/core#linkedAuth
or> ?author.
?author
<http://vivoweb.org/ontology/core#hasMemberR
ole> ?membership.}

## Bind the Entrez gene IDs
## to eagle-i resources

?resource obo:ERO_0000236 ?entrezgeneid.

}
```

5. Dataset Usage

A number of groups have begun to make use of eagle-i-produced data by implementing search and visualization tools that reuse the RDF and that complement the functionality provided by eagle-i applications. The Harvard Catalyst Core Facilities Portal (http://cbmi.catalyst.harvard.edu/cores/index.html) was an early adopter on this front. 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.

Another interesting reuse of eagle-i RDF data is the CoreSearch service at Oregon Health & Science University (http://www.ohsu.edu/research/coresearch/).

Leveraging the Plumage tool (http://ctsiatucsf.github.com/plumage/), developed by the Clinical & Translational Science Institute at the University of California San Francisco, CoreSearch allows for visualization and search of OHSU core laboratories, their service offerings and their instruments by converting the RDF data accessed through the eagle-i SPARQL endpoints to static HTML pages that can be optimally indexed by Google and other search engines. Eagle-i datasets are also used in the context of CTSAConnect project (http://www.ctsaconnect.org/) to link clinicians to basic researchers through publications and research resources.

6. Updating maintenance and scalability

The creation of the eagle-i datasets over the three years of the project was possible as a result of the dedicated work of resource navigators (Ph.D. level scientists contacting laboratories and collecting resource information) and curators responsible for the data entry and quality control of collected data. After the grant ends, each institution in the network will be in charge of maintaining and updating their local datasets with the help of detailed guidelines and tools (https://open.med.harvard.edu/display/eaglei/Training). The eagle-i soft stackware, which is available as open source, will continue to be maintained and enhanced by the development team at Harvard with contributions from the open source community. The

Harvard team will also operate central components that tie the eagle-i network together (node registry, central search, global instances repository). The eagle-i ontology has been and will remain an open source community resource (https://code.google.com/p/eagle-i/) that is updated through tracker requests. Each new eagle-i software release incorporates the latest release of the eagle-i Resource Ontology. To assure data alignment in accordance with these regular ontology changes, the developers concurrently release data migration scripts to update each of the triple stores. Other sustainability efforts for data collection are related to the integration of the eagle-i backend with laboratory inventory management systems such as iLab (http://www.ilabsolutions.com/) and billing systems such as Vanderbilt CORES (http://www.mc.vanderbilt.edu/root/vumc.php?site=CORES).

7. Discussion

The eagle-i linked data is a suite of different datasets: one for each institution participating to the network. This choice was motivated by a desire to allow each institution to control their own search and data entry applications as well as to assign their own URIs for their respective published instance data.

The lack of a single SPARQL query interface to search over all of the eagle-i datasets at once, but is easily overcome using programmatic access. As an internal response, for curation purposes we have developed a simple web application that allows eagle-i curators to select multiple SPARQL endpoints and issue queries against them in bulk.

Another characteristic of our datasets is related to the usage of numeric URIs for most of the classes and properties. This choice was driven by the decision to adhere to the OBO Foundry Principles. From the perspective of ontology development, it makes sense to have the semantics of a particular resource be conveyed by its textual and logical definition rather than by a human readable URI or `rdf:label`. This is convenient to avoid misuse of entities when, for example, a label of a particular entity changes. This makes writing SPARQL queries less straightforward but we have found that good documentation of the ontology through the ontology browser and Ontobee are of great help for our end users.

Another problem we had to face while creating the dataset was related to particular “instances” that did not belong to any institution but were supposed to be “global” (i.e. used form each institution). Examples of this kind of instances are Organizations, Manufactures or any kind of resource that are not tied to particular institutions in our network. For these kinds of resources we use a particular name space and we store them in a dedicated repository.

Acknowledgment

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

[1] B. Smith, et al., The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, Nat Biotechnol, 25 (11) (2007), 1251-5.

[2] P. Grenon and B. Smith, SNAP and SPAN: Towards Dynamic Spatial Ontology, Spatial Cognition & Computation: An Interdisciplinary Journal 4 (2004), 69-104.

[3] M. Courtot, F. Gibson and A. Lister, MIREOT: the Minimum Information to Reference an External Ontology Term, ICBO, 2009.

[4] C. Torniai, M. Brush, N. Vasilevsky, et al., Developing an application ontology for biomedical resource annotation and retrieval: challenges and lessons learned, ICBO, 2011.

[5] C. Torniai, M. Brush and M. Haendel, A pipeline for biomedical ontology maintenance and release, Blog post, http://blog.carlotorniai.net/388/, Accessed March 05 2013.

[6] Z. Xiang, C. Mungall, A. Ruttenberg and Y. He, Ontobee: A Linked Data Server and Browser for Ontology Terms, ICBO, 2011, pp. 279-281.