Research Resources: Curating the New Eagle-I Discovery System

DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2012, Article ID bar067, doi:10.1093/database/bar067

Original article

Research resources: curating the new eagle-i discovery system

Nicole Vasilevsky[1], Tenille Johnson[2], Karen Corday[2], Carlo Torniai[1], Matthew Brush[1], Erik Segerdell[1], Melanie Wilson[1], Chris Shaffer[1], David Robinson[1] and Melissa Haendel[1,*]

[1]Oregon Health & Science University, Library, LIB, 3181 S.W. Sam Jackson Park Rd., Portland, OR 97239-3098 and [2]Harvard Medical School, Center for Biomedical Informatics, One Kendall Square, Suite B6303, Boston, MA 02139, USA

***Corresponding author:** Tel: +503-407-5970; Fax: +503-494-3322; Email: haendel@ohsu.edu

Submitted 3 October 2011; Revised 12 December 2011; Accepted 13 December 2011

Development of biocuration processes and guidelines for new data types or projects is a challenging task. Each project finds its way toward defining annotation standards and ensuring data consistency with varying degrees of planning and different tools to support and/or report on consistency. Further, this process may be data type specific even within the context of a single project. This article describes our experiences with eagle-i, a 2-year pilot project to develop a federated network of data repositories in which unpublished, unshared or otherwise 'invisible' scientific resources could be inventoried and made accessible to the scientific community. During the course of eagle-i development, the main challenges we experienced related to the difficulty of collecting and curating data while the system and the data model were simultaneously built, and a deficiency and diversity of data management strategies in the laboratories from which the source data was obtained. We discuss our approach to biocuration and the importance of improving information management strategies to the research process, specifically with regard to the inventorying and usage of research resources. Finally, we highlight the commonalities and differences between eagle-i and similar efforts with the hope that our lessons learned will assist other biocuration endeavors.

Database URL: www.eagle-i.net

Introduction

As the volume of biological information has grown, so has a corresponding need to strengthen the practice of data collection, sharing, reuse and preservation beyond the traditional publication cycle. An enormous challenge facing science today is the management and integration of large amounts of diverse types and sources of data. One of the disciplines that has grown in response to this challenge is biocuration: biocurators manage and organize data pertaining to bioscience research from the literature, Web and other primary sources to make it accessible to the community (1,2). Well-curated data contained in repositories has proven invaluable, allowing researchers to efficiently compare and reuse data, quickly access information about their research interests, gain insight into experimental design, and discover novel connections between data from different sources (2). However, it is currently not the norm for researchers to organize their data, scientific processes, and resources in a structured way. Use of data management systems is not common in basic science laboratories. This makes resource or data sharing and reuse cumbersome. Therefore, a paradigm shift is needed where scientists perform information management throughout the research cycle, instead of it being limited to the publication phase of their projects.

Numerous databases have been developed to collect and manage bioscience data. These repositories range from gene function, such as the Gene Ontology (GO) Database (http://www.geneontology.org/), to sequence databases (http://www.ebi.ac.uk/embl/, http://www.ncbi.nlm.nih.gov/genbank/, http://www.uniprot.org/), microarrays

Database, Vol. 2012, Article ID bar067, doi:10.1093/database/bar067

(http://smd.stanford.edu/, http://www.ebi.ac.uk/arrayexpress/, https://genome.unc.edu/) and species-specific databases such as ZFIN (www.zfin.org) and WormBase (www.wormbase.org) that collect a range of data pertaining to one species. However, very few of these organizations collect information about the scientific process itself, that is, the protocols, reagents, instruments, techniques, etc. that are used during the course of experimentation. Only the most salient features of experimentation are presented within publications, and a vast amount of information exists about scientific activities and products of research that are never made available.

To faithfully reproduce a predecessor's work, one must know all the subtleties of the experiment, but laboratories vary in the degree of structure used to track both their resources and their scientific process. For example, protocols that use 'x' instrument with 'y' reagent are written down in lab notebooks, in *ad hoc* spreadsheets, on white boards, on paper towels, or not written down at all. Effective information management is a vital skill, but one that is often not being taught. Training for resource management is inadequate even at the very first stages of a scientific career. A survey of 48 undergraduate ecology programs revealed that >75% did not require students to use lab notebooks and more than half did not include any data management-related instruction in the curriculum (Carly Strasser, personal communication). Even in the context of a publication, resources are often inadequately referenced. For instance, antibodies are frequently mentioned without their corresponding catalog number. Not only do such omissions hamper the ability to reproduce an experiment, they result in missed opportunities to harvest and search data.

The lack of information that uniquely identifies the research materials has been an ongoing problem for model organism databases (1). When curating gene and protein information, the gene nomenclature is often unclear or the source organisms are not identified, preventing their inclusion into a database (3). Efforts have been made to create a consistent nomenclature for gene names, such as the International Committee on Standardized Genetic Nomenclature for Mice (4), and the HUGO Gene Nomenclature Committee (HGNC) (5). Journals such as *Nature, Science* and *Public Library of Science* (*PLoS*) now require the use of current nomenclature and accession numbers for DNA and protein sequences, but do not generally require metadata about reagents, such as antibody antigens or Entrez Gene IDs for plasmid inserts. Ultimately, much of the scientific data produced never makes it into curated databases due to an inability to uniquely identify the research resources to which it would be linked. Moreover, large numbers of research resources remain unshared due to their undocumented status or perceived lack of value.

Table 1. Participating institutions in the eagle-i Consortium

Institutions
Harvard University, Cambridge, MA
Oregon Health and Science University, Portland, OR
University of Hawaii at Manoa, Manoa, HI
Montana State University, Bozeman, MT
Dartmouth College, Hanover, NH
Morehouse School of Medicine, Atlanta, GA
Jackson State University, Jackson, MS
University of Puerto Rico, San Juan, PR
University of Alaska Fairbanks, Fairbanks, AK

The eagle-i Consortium, a collaboration between nine academic institutions (Table 1), is creating a searchable inventory of unique, rare or otherwise hard-to-find biomedical research resources in order to foster sharing and linking of resources in the larger scientific community. The nine participating institutions in the 2-year pilot project were chosen for inclusion based on their range of size, geographic location and diversity of National Center for Research Resources (NCRR)-funded institutional programs, including the Research Centers in Minority Institutions (RCMI) and IDeA Networks of Biomedical Research Excellence (INBRE) (http://grants.nih.gov/grants/guide/rfa-files/RFA-RR-09-009.html). The eagle-i project collects resource information about core facilities and services, tissue banks and cell repositories, protocols, software, animal models and organisms, human health studies, research reagents, instruments and research training opportunities. As of September 2011, the eagle-i system houses data for over 45 000 resources among the nine participating institutions (Table 2). To collect and search this resource information, an ontology-driven Data Collection Tool and Search application were designed (6). Since these tools were developed simultaneously with collection efforts, the collected data informed the way the ontology and tools were built in an iterative fashion. This iterative development cycle provided both unique benefits and challenges to the eagle-i Curation team.

The content of eagle-i was developed as a complement to other existing resource discovery databases for publicly available resources in specific fields of study. For example, the Neuroscience Information Framework (NIF) was created to address the need for a searchable repository of publicly available neuroscience resources, though it has since expanded to numerous other non-neuroscience resources (7). The Resource Discovery System (RDS) inventories bioinformatics and service resources at Clinical and Translational Science Awards (CTSA) centers (8). All of these

Table 2. Summary of the number of resources collected at each site

Resource type	Alaska	Dartmouth	Harvard	Hawaii	Jackson State	Montana State	Morehouse	OHSU	Puerto Rico	Total
Organisms and viruses	15	14 262	1186	12 816	3	87	17	151	46	**28 583**
Instruments	225	114	1171	688	85	181	66	216	612	**3358**
Reagents	0	125	5773	184	4	173	65	233	161	**6718**
Services	66	101	984	347	42	71	52	465	110	**2238**
Software	38	47	222	65	50	43	6	150	66	**687**
Protocols	67	34	137	47	12	73	7	122	86	**585**
Core laboratories	8	20	196	36	14	15	12	37	33	**371**
Research opportunities	0	2	18	0	3	1	1	7	0	**32**
Biological specimens	0	0	0	2844	2	0	0	43	25	**2914**
Human studies	13	0	0	134	42	0	0	2	0	**191**
Total	432	14 705	9687	17 161	257	644	226	1426	1139	45 677

Electronic systems include spreadsheets, text files, MacVector and MAG-ML; non-electronic systems include lab notebooks and paper files; LIMS include Quartzy and Epic.
Note: some labs used more than one type of inventory system.

systems have overlapping and complementary purposes and resource catalogs, and have been working together on a common ontological representation of research resources and platform integration (9, http://groups.google.com/group/resource-representation-coordination).

Other recent efforts are underway to make scientific resource information semantically represented and attributable. A project called 'Beyond the PDF' was recently formed with the goal of identifying 'a set of requirements and a group of willing participants to develop a mandate, open source code and a set of deliverables to be used by scholars to accelerate data and knowledge sharing and discovery' (https://sites.google.com/site/beyondthepdf/home). While such an effort is not yet widely accepted and practiced, this is a movement towards creating and using controlled vocabularies to semantically represent research entities. A related effort, the Bioresource Research Impact Factor (BRIF), aims to quantify the impact of bioresources that are used and shared within the scientific community. Ideally, BRIF will promote and incentivize the sharing of resources by providing recognition of scientific contribution (10). The Ontology of Biomedical Investigations (OBI) (11) and The Minimum Information for Biological and Biomedical Investigations (MIBBI) Foundry (12) are working towards the goal of creating structured metadata for annotating experimental processes, results and methodologies. All of these systems have the objective of sharing resource information or primary data, making it a currency of research beyond publication and ensuring better scientific reproducibility with improved semantic identity and linking. The eagle-i system was created alongside these frameworks and complements and extends the goals these efforts aim to achieve.

A central Curation team was responsible for developing the eagle-i ontology, creating guidelines for data collection and annotation, and ensuring usability of the system by reviewing the structure and accuracy of the data in each repository. Curation at the eagle-i Consortium was rather challenging due to the short duration of the 2-year project, the geographic distribution of the members of the team, and the diversity of data types being collected. Herein, we report on our experience, address issues in biomedical resource curation in the context of eagle-i, and discuss commonalities and differences between other existing biocuration efforts to promote better curation strategies for future efforts.

Data collection

Each institution in the consortium collected information about its local research resources using three primary methods. First, an eagle-i staff scientist (known as a Resource Navigator) visited labs at each site to manually collect information about resources directly from laboratories and enter it into the eagle-i Data Collection Tool, an online data collection and curation tool. Second, selected lab staff members were authorized to enter laboratory resource information directly into the Data Collection Tool. Third, automated upload of large sets of resource data into the eagle-i repository was possible using an extract, transform and load (ETL) process.

The primary role of the Resource Navigators was to perform outreach and collect information about resources at each lab. Collection of resource-related data directly from the laboratories that house, control and best know the information allowed for capture of data about 'invisible'

Database, Vol. 2012, Article ID bar067, doi:10.1093/database/bar067

resources. This contrasts with many other types of biomedical biocuration efforts, which either curate data directly from the literature (1) or call for external submissions to submit new entries directly into a database (such as WormBase). Using Resource Navigators to collect the resource data was advantageous because it did not require waiting for external submissions and allowed information to be verified directly with the researcher. While this approach was enormously valuable in the specification of the system, it is not sustainable due to the high cost of employing dedicated staff. Therefore, eagle-i is examining strategies to obtain data via a variety of laboratory management systems in the future, so that the resource information would be fed directly into eagle-i from the output of the research labs themselves. Including lab staff members as users of the Data Collection tool is helping define requirements to this end.

The scope of work required at eagle-i differed from other databases such as ZFIN and Mouse Genome Informatics (MGI) (1). Rather than the curators performing both data entry and curation tasks, the Resource Navigation team performed data entry and a specialized Curation team handled biocuration. The Curation team further refined the data to ensure that it met eagle-i standards that were outlined in the current Curation guidelines. This workflow for the eagle-i system is summarized in Figure 1. The workflow involved both Resource Navigators and Curators and was effective for separating collection and curation tasks, as the Resource Navigators were trained scientists with experience working in laboratory settings and familiarity with the types of resources being collected, and the Curation team was more specialized in data management. As the eagle-i Consortium expands, these roles will change to accommodate data coming directly from the labs.

Many laboratories were willing to share their resources, but at the time of data collection, only 10% of the total laboratories that participated in eagle-i reported using any kind of inventory tracking system for their resources (Table 3). Of the labs that used an inventory system, spreadsheets were the most commonly used. In this case, we used an ETL process to transfer data into templates that were consistent with the ontology and upload them via an automated script. While the upload of data is more efficient than manual data entry, it was often still time consuming to prepare the data to match the specified fields. For example, if a lab had an inventory of plasmids that included only the name of the insert, it was necessary to manually add additional information such as the backbone, manufacturer, selectable marker, and the source organism and corresponding Entrez Gene ID for the insert. If researchers more systematically kept track of these resource attributes, it would be easier for authors and curators to record resources in publications and resource discovery systems such as eagle-i.

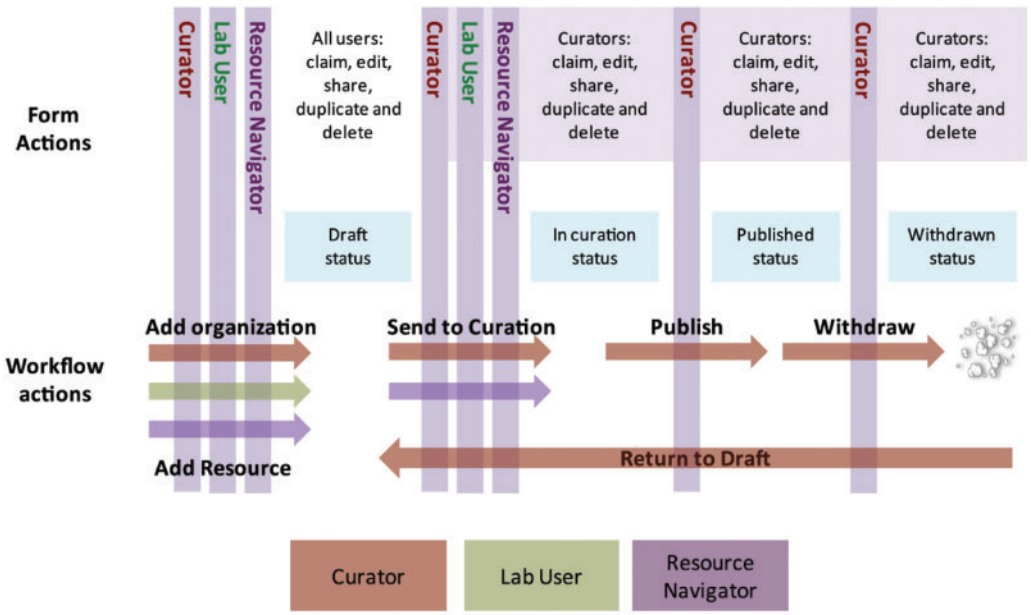

Figure 1. Workflow of the eagle-i team. The role of the Resource Navigators is to collect and add data to the system, such as organizations or resources. All users (Curators, Lab Users and Resource Navigators) can enter data into the Data Collection tool in draft state. To edit a record, it must be 'claimed' by the user and then 'shared' after editing. Curators and Resource Navigators can send resources to curation. Data 'in curation' is managed by the Curation team and subsequently published, where it is visible in the Search interface. After a record is published, a Curator can withdraw, duplicate or delete the record, or return the record to draft for further editing.

Table 3. Summary of laboratories that use a lab inventory system and type of system used at each institution in the eagle-i Consortium

Institution	Type of inventory system					Total number of labs	Percentage of labs with inventory systems (%)
	Electronic	Non-electronic	LIMS	Database	Unspecified		
Alaska	6		1	1		15	47
Dartmouth					1	57	2
Harvard	3		1	1	2	206	3
Hawaii	23			2	2	105	26
JSU	6	1			2	20	40
Montana	1			1	1	77	4
MSM	3	1	1		1	19	26
OHSU	9		2	1	2	75	15
UPR						103	0
Total	51	2	5	6	11	677	10

While there is an apparent need for scientists to use controlled vocabularies and share data in a consistent, structured format, motivating them to do so is a challenge (13). It has been suggested that the best way to motivate researchers to share data is by mandates from funding agencies and journal publishers (13). Both the NSF and NIH now require data management plans for new grant proposals, but the implementation has not been enforced. Perhaps if easy-to-use tools were available, this would increase researcher compliance. Some have argued that mandates are not the answer; it is autonomy, mastery, and purpose that will drive researchers to comply (14). According to Michael Lesk, Chair of the Department of Library and Information Science at Rutgers University, 'we need ways to reach out to individual researchers and provide simple methods for their participation (15)'. We need to demonstrate why they should care about referencing research resources as semantic entities: reasoning across data (for example, identifying artifacts or experimental bias), locating research resources (making science faster and less expensive) and interoperability across projects (promoting synthetic science and developing new hypotheses).

One mechanism to facilitate researchers' understanding of the importance of specifically referencing research entities is to provide training in information and data management as part of the curriculum for new scientists (e.g. undergraduate and graduate students). Similarly, if tools existed that were part of the normal scientific workflow that enabled semantic tagging of information, it would be much simpler to comply with data sharing standards. Commercial Laboratory Information Systems (LIMS) exist to aid laboratories with inventorying their resources, but are often very expensive or beyond the needs for most academic labs. Systems like BioData (www.biodata.com) or Quartzy (www.quartzy.com) are specifically designed for academic labs and integration of these systems into a scientists' workflow could facilitate both the effort towards use of shared terminologies and data and resource sharing. Such systems could ideally feed resource repositories and auto-populate manuscripts and grant reports with little effort on part of the researchers. As more systems such as these and eagle-i are developed, it is hoped that researchers will see the utility and apply more rigorous data management and dissemination processes to their workflow.

Ontology-driven data collection and search tools

The eagle-i framework includes a Data Collection Tool and a Search application that are both ontology-driven, providing structured vocabulary and enabling logical connections between data items. Further description of this technology has been previously published (6,9). The data storage technology is a Resource Description Framework (RDF) triple-store, which allows 'many-to-many' relationships between resources and data (16). Other biomedical databases, such as UniProt (http://expasy3.isb-sib.ch/~ejain/rdf/), also use RDF technologies. Tools using RDF can create subsumptive hierarchies of both the resource types and the properties that are used to relate one type to another (17). For example, a DNA sequencer *is_a* sequencer and a sequencer *is_a* instrument. Use of RDF can enable enhanced search capability for the end users. The use of an ontology for eagle-i facilitates query by inference and interoperability with other data and databases.

The earliest eagle-i data modeling was based on questions and search scenarios collected from domain experts.

Database, Vol. 2012, Article ID bar067, doi:10.1093/database/bar067

Requirements generated through these scenarios also drove initial data collection efforts; the structure and type of resource data gathered by Resource Navigators heavily influenced subsequent modeling. We then leveraged external ontologies and vocabularies in the building of the eagle-i ontology in OWL (Web Ontology Language) (6), including those of the OBO Foundry (http://www.obofoundry.org/), the Ontology for Biomedical Investigations (OBI) (http://obi-ontology.org/page/Main_Page), NIF (http://neurolex.org/wiki/Main_Page), Biomedical Resource Ontology (BRO) (http://bioportal.bioontology.org/ontologies/1104) and Medical Subject Headings (MeSH) (http://www.ncbi.nlm.nih.gov/mesh). However, many of the resource types being collected by eagle-i did not have representation in preexisting ontologies. As a result, we used the requirements from eagle-i to drive our ontology development and to provide feedback to existing ontologies. The resulting eagle-i resource ontology (ERO) is a combination of classes already existing in external biomedical ontologies and taxonomies, and classes we created in order to be able to represent the information about resources inventoried by eagle (6).

The eagle-i ontology was used to generate and pre-populate fields in the Data Collection Tool in order to create consistent annotations about the resources (Figure 2). Ontology classes are linked to an external glossary, which displays the definitions of the terms to assist in proper usage. A free text 'resource description' field was available to annotate potentially valuable or unique features about the resource that were not captured in other fields. Combined free text and ontology-driven fields are common among other databases, notably the *Saccharomyces cerevisiae* and *Mus musculus* databases, Saccharomyces Genome Database (SGD) (http://www.yeastgenome.org/) and MGI (http://www.informatics.jax.org/) (18,19). Using a combination of controlled vocabularies and free text allowed extensive information to be captured for each resource and enhanced search by providing multiple axes of classification. The capture of free text descriptions also allowed for collection of new requirements for data representation.

The Search application is intended to allow researchers to find and enable reuse of the resources that were collected in the eagle-i repositories. Researchers can search for matches on both ontology classes and instance labels. For example, DNA sequencing is a technique and c-Myc is the label of a plasmid insert. Users can use the ontology to filter based on organization and resource subtypes. Researchers then may contact the owner of the resource through a link on the results page.

The eagle-i ontology also includes properties that can link data to outside resources such as Entrez Gene ID or PubMed ID (PMID). These attributes are valuable, as a name alone is insufficient for identification. Adding Entrez Gene IDs for plasmid inserts, antibody antigens, transgenic organisms, etc. allows the user to link to gene information at NCBI and therefore many other data types. In addition to external resources, the Data Collection Tool allows linking between resource records. For instance, an instrument record can be linked to a technique, a contact person, manufacturer or a related publication or other documentation in the repository. This is important within the context of semantic searching, as it allows users to locate resources using different entry points, pathways, search terms and methods.

During the initial development of the eagle-i ontology, there were many as yet unrecorded synonyms for eagle-i classes, which limited the capability of the system. While synonyms were available from existing ontologies, such as OBI, new terms or even imported terms did not always have every known synonym. To address this issue, additional synonyms were obtained from domain experts and large vocabularies such as Systematized Nomenclature of Medicine—Clinical Terms (SNOMED) and MeSH. These synonyms were added to the underlying eagle-i ontology so results would be returned when users searched for those particular terms or their synonyms. For example, if a user searched for a reagent that was used in '*in situ* hybridization' and entered the abbreviation 'ISH', only one result would be returned based on text matching. Because of the inclusion of the synonym in the ontology, 135 additional reagent results are returned. It is important to note that these synonyms were added to the eagle-i ontology consistent with the orthogonal set of OBO Foundry ontologies.

Data curation

The eagle-i Curation team was responsible for ensuring that the data collected was within scope for inclusion, that it was annotated correctly, and that logical connections between the resource types were correctly associated. A summary of the number of resources inventoried at each site is given in Table 2. It is common for curation work to be performed by domain experts in the field who have the knowledge and experience to interpret the data itself. This necessary expertise must be balanced, however, with an understanding of the information science practices and technical realities central to modern curation work: archiving, indexing, data modeling, ontology development and usage, etc. The eagle-i Curation team brought varied educational and professional backgrounds to the project, which included expertise in biological sciences, Semantic Web technologies, library and information science and medical informatics. The central Curation team was located at OHSU and Harvard. This diversity was needed to build

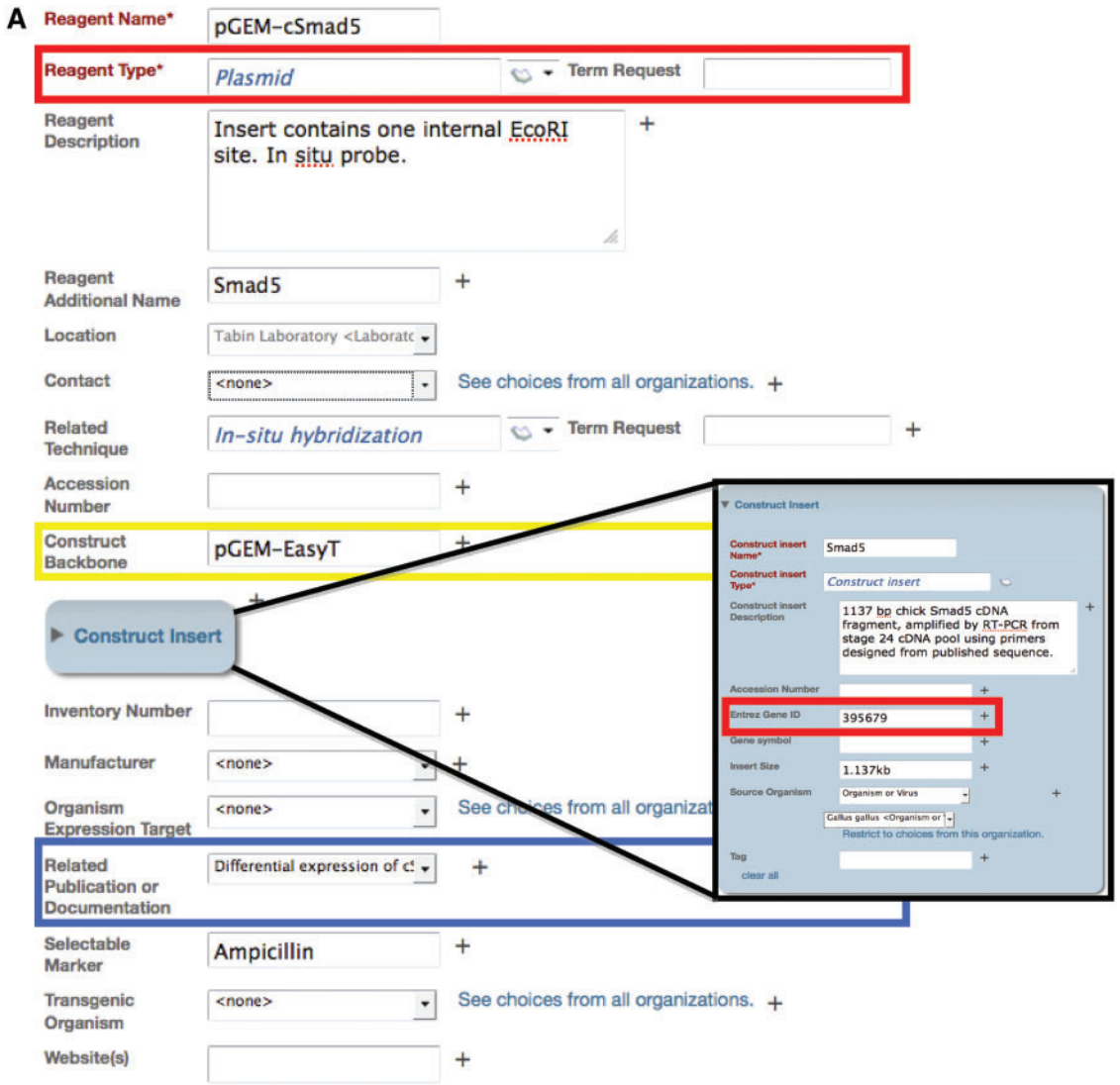

Figure 2. Example of an annotation form in the Data Collection Tool for the plasmid reagent type. (**A**) The Data Collection Tool contains annotation fields that are auto-populated using the ontology (red box) and free text (yellow box). Fields in the Data Collection Tool can also link records to other records in the repository, such as related publications or documentation (blue box). Users can request new terms be added to the ontology using the Term Request field. Inset: Construct insert is an embedded class in the plasmid form and contains information that corresponds to other databases, such as Entrez Gene ID. (**B**) The search result upon searching for this specified plasmid. Only the fields that are filled out in the data tool are displayed in the search interface. Search results can be returned for this plasmid by searching on any of the fields that are annotated for this record. Text that is colored blue links to other records in the search interface. Hovering over the 'i' icons displays the ontological definition of the term, as in the example of the technique, *in situ* hybridization.

out the system. However, a staff of this size would not be required for on-going curation for nine sites after the initial pilot stage.

New distributed projects should not underestimate the utility and cost of face-to-face time when it comes to data annotation and quality assurance (QA) training. Frequent and planned communication is key, especially with regard to new guidelines or functionality, to ensure that the data is consistently and accurately annotated. In this technological age, communication via email, teleconference, and videoconference is convenient and economical. However, there is still benefit to in-person interactions. For example, the technical lead on the Biodiversity Heritage Library spends 50% of his work time traveling to meet with project partners in order to ensure that data standards are met and communicated (Chris Freeland, personal communication). In the context of the eagle-i pilot project, we leveraged in-person training and the aforementioned

Database, Vol. 2012, Article ID bar067, doi:10.1093/database/bar067

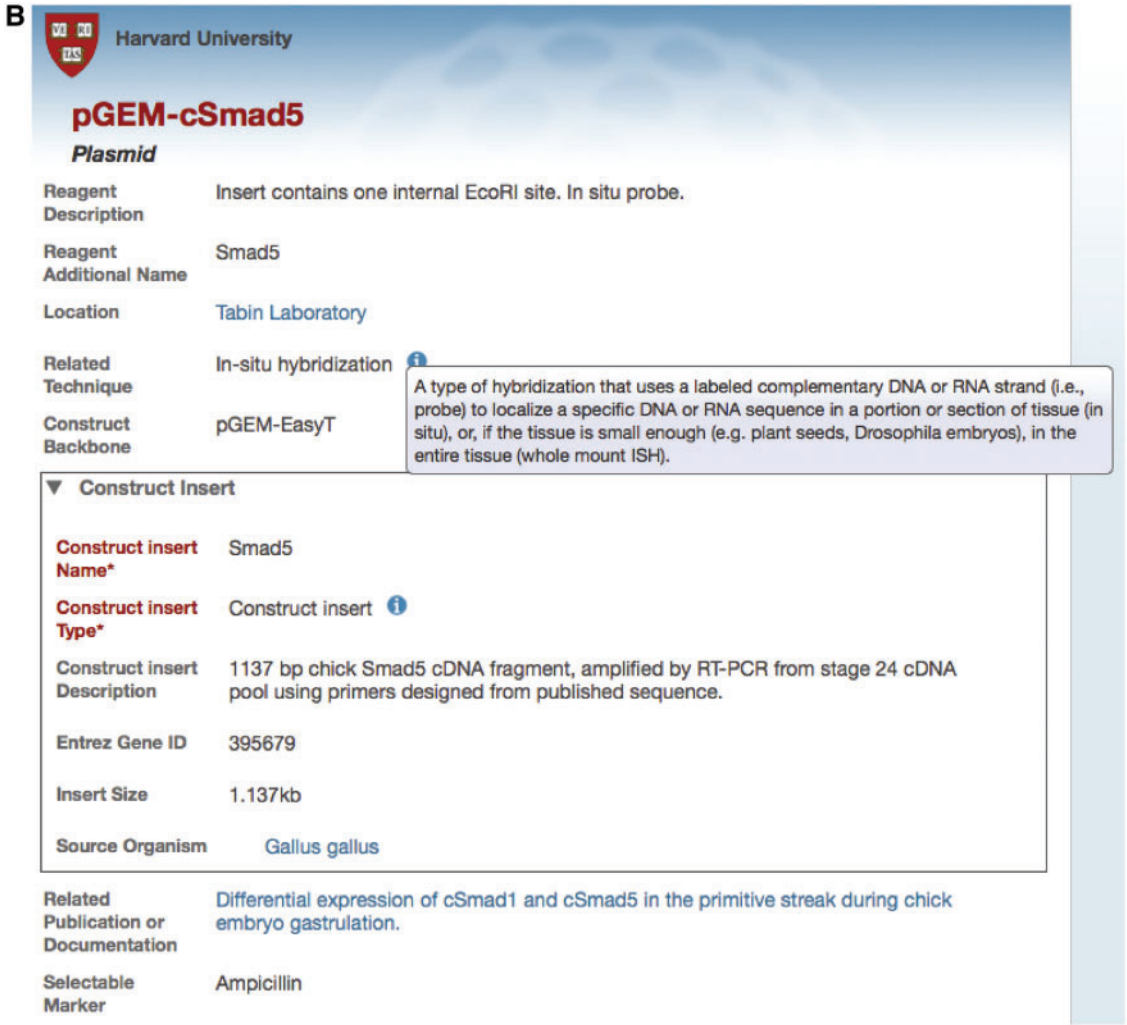

Figure 2. Continued.

methods to facilitate communication and quality data entry in the eagle-i system.

The development of the eagle-i curation workflow and guidelines was an iterative process. The earliest guidelines were compiled in a document but not closely adhered to, leading to inconsistent or poorly annotated data. When the Data Collection tool came online, the guidelines were integrated as online help for greater visibility. The online help was intended to facilitate both data entry and curation; greater awareness of the standards enhanced data consistency. The current online guidelines for eagle-i data entry and curation are available here: http://bit.ly/eicurationguidelines.

The initial data collection efforts emphasized quantity and diversity above all else. This was necessary since a significant amount of data was needed to define the parameters of the ontology and of the system itself. Stricter guidelines were later developed regarding the type of resources and data annotations that were to be included. Quality, not quantity, became the overriding factor in determining which data to incorporate and guidelines were enhanced to help the Resource Navigators and Curators determine what type of data were suitable for inclusion in eagle-i. At this time, the Curators needed to ensure that an adequate amount of information was included in each record and that the information was consistently annotated. Within each record, there were both required fields and highly desired fields. Highly desired fields were not required in order to allow flexibility, yet they helped guide users to the types of data that would make the record more meaningful from a search perspective. Decision trees illustrating the resource properties provided guidance for proper annotation of each resource type (Figure 3).

Exclusively leveraging the user community for contribution to the database may have been logistically simpler

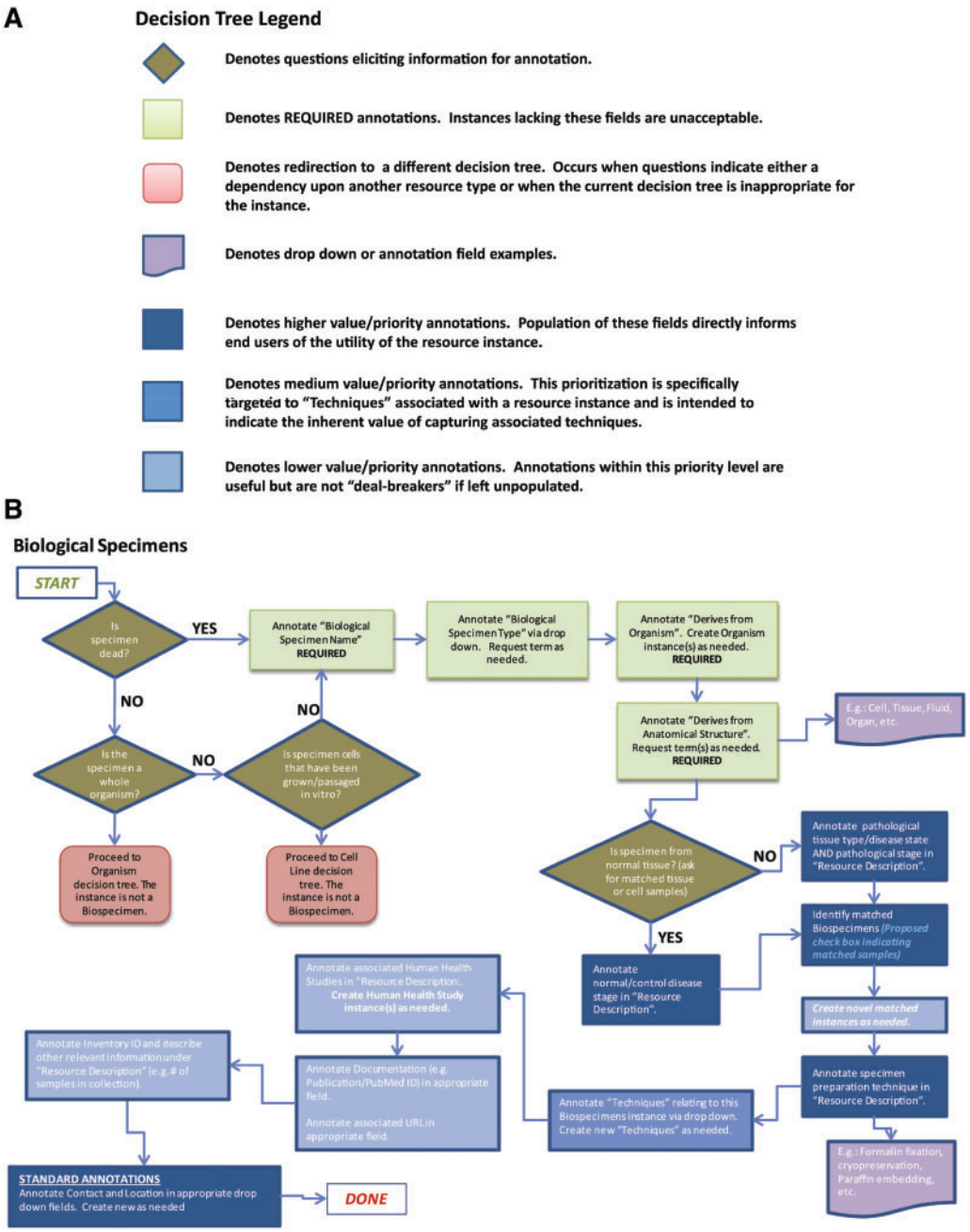

Figure 3. Decision trees were used to assist with data collection and curation. (**A**) Decision tree legend. (**B**) The decision tree for biological specimens. Required and highly desired fields are indicated by green and blue colors, respectively. Each resource type had 2–3 required fields, between 4–8 highly desired fields and the other fields were considered optional or applied only to specific subtypes of resources.

than employing Resource Navigators, but would have presented quality issues due to the distinct lack of structured terminology within the community. Similar database projects, such The Immune Epitope Database and Analysis Resource (IEDB) (www.immuneepitope.org), have reported that inconsistent usage of terminology by the community led to reproduction of work and inconsistently annotated resources (20). Relying upon contributions from the community could also potentially lead to not only increased inconsistency but also fewer contributions in general. The Neuron Registry has experienced this as the number of new articles contributed has markedly declined over time (http://pons.neurocommons.org/page/Neuron_registry). While some select lab users were trained to perform

Database, Vol. 2012, Article ID bar067, doi:10.1093/database/bar067

data entry into the eagle-i system, the majority of data was collected by Resource Navigators to produce better and more voluminous data. The use of Resource Navigators resulted in the substantive information about real resources that was required in order to build the ontology and user interfaces.

It has frequently been reported that the 'build it and they will come' attitude with regards to building data repositories has not been entirely successful (13). However, some databases have had success with 'crowdsourcing' such as EcoliWiki (http://www.ecoliwiki.net) and The Arabidopsis Information Resource (TAIR) (http://www.arabidopsis.org), the latter of which is now focusing on community curation due to funding cuts. Now that the eagle-i pilot system is in place, the goal will be to determine how to best tailor data entry and incentivize scientists to contribute data. A potential approach could involve collaboration between laboratory inventory tracking systems such as BioData (www.biodata.com/) or Quartzy (www.quartzy.com/), where researchers would use such systems for structured inventory management, facilitating transfer into a system such as eagle-i for research sharing. We need a flexible and extensible system to be able to integrate data from a wide variety of sources and/or domain dependent tools. Future efforts are underway to pursue such collaborations.

Quality assurance

For any biological database, there is always a need to keep up with the changing internal and external landscape as we strive to meet end user needs. In the case of eagle-i, the requirements for data annotation evolved as the data were collected, the ontology updated, and the guidelines enhanced. This created legacy data that met our initial data quality guidelines but required updating in order to meet the new standards. To ensure data quality and currency, a variety of methods were used, including manual revision, application of metrics, and building tools for automating processes. Some of the issues that were routinely addressed included poor naming of resources, misapplication of ontology classes and properties, inconsistent usage of terminology in free text fields, insufficient annotation for resource types, lack of links between records, and incorrect, outdated, or broken links on the records. Issues such as these were encountered both through routine curation and through bulk QA efforts.

To support quality assurance, we developed procedures, workflows, and a tool to compare and analyze the data. Since we had chosen triple-store technology, SPARQL Protocol and RDF Query Language (SPARQL) queries (http://www.w3.org/TR/rdf-sparql-query/) were used; SPARQL is a mechanism to query for specific sets of triple statements in an RDF triple store (16). We developed a simple web application to perform precompiled SPARQL queries on the data at any snapshot in time. The query results displayed the exact instances that required updating, which were updated manually or via more efficient and less error-prone bulk curation scripts. The bulk curation tool was particularly useful for bulk migration of data from one field or type to another, and for applying changes to multiple resources simultaneously. Other systems have automated features to determine if and when curators need to perform QA. For instance, the myGrid project has developed an automated monitoring service that checks for issues in the myExperiment system (http://www.myexperiment.org/) (21). In the eagle-i system, the frequency of changes to the data model often necessitated updates to the data.

One way data quality was evaluated was by a quantitative assessment of various data characteristics. For instance, a cutoff was defined for the minimum number of filled properties for each resource type. We performed analysis on instrument records by comparing the number of fields filled out at two different time points, before and after a quality assurance effort. The instrument analysis showed up to a 4-fold increase in the number of annotations for most of the repositories after enhancing the quality of the data (Figure 4). This method proved to be effective because instruments are one of the least complex resource types and there is little variability between the properties used to describe instrument instances. However, use of quantitative field metrics did not necessarily facilitate quality assurance of more complex biological resources, such as biological specimens, reagents, organisms, and viruses. For example, organism resource properties included many fields related to transgenic organisms, which are not

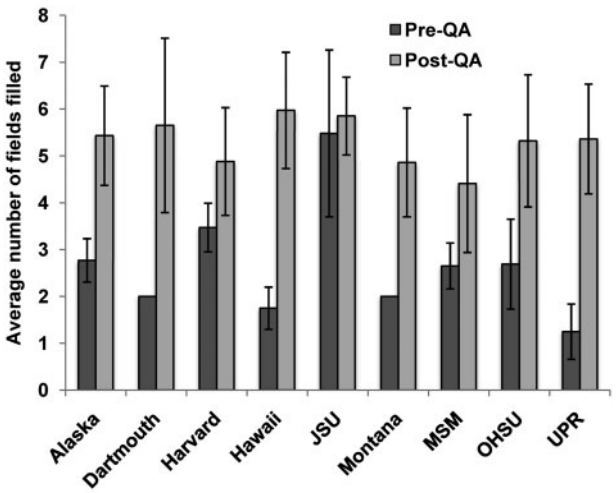

Figure 4. Average number of fields recorded for instruments before and after a QA effort. There was a 1.1–4.3-fold increase in the average number of filled fields after the QA effort. Error bars indicate standard deviation.

applicable to wild-type organisms. Similarly, plasmids may or may not have an insert, so a plasmid that is used as a control vector would be lacking data in the 'construct insert' fields. Therefore, using metrics for the number of annotated fields were not always informative, necessitating manual spot checks to be performed to ensure data quality.

While the metrics informed the data quality for certain annotations, additional information was required to ensure annotation quality. Instruments that had the same label and type were considered inappropriately annotated; for instance, the instrument label, 'Flow cytometer' and the instrument type, 'Flow cytometer'. Such resources were identified by *ad hoc* SPARQL queries and more specific labels were given for the instrument name to improve the overall quality of the record, e.g. replacing the name 'Flow cytometer' with 'BD FACSAria I flow cytometer'. This was also important for search purposes, enabling the user to search on different specific terms, as well as returning detailed and useful information about the resource.

The ability to track the provenance of changes is important to establish consistent practices between curators and ensure the accuracy and currency of the data. Tracking changes also allows establishment of metrics over time to further determine strategies for quality assurance. The pilot eagle-i system only recorded the last modifying user and record creator; detailed notes regarding significant changes (such as changing the resource type) were added to a comments field that was not visible in the search results. Ideally, these types of changes will be recorded directly in the repository in the future. Any new system developer should consider this important facet when developing a curation interface.

eagle-i usability

With all of the aforementioned development, data collection, biocuration, etc., if the users cannot find what they are looking for, then the eagle-i system will not have succeeded in meeting its goal to make research resources accessible. A beta version of the eagle-i system was released to the nine participating institutions prior to the conclusion of the pilot stage of the project. Researchers at the nine sites were invited to use the system and asked to provide feedback based on their experience using the search tool. The survey of 259 eagle-i users demonstrated that eagle-i addresses a real need and functions as intended. Ninety one percent of users were 'satisfied' or 'very satisfied' overall with the eagle-i Search. Seventy one percent felt that eagle-i would be particularly useful to them if scaled to institutions nationally. The vast majority of respondents (97%) learned about resources they did not know were available. Sixty three percent said they would be 'very likely' or 'likely' to contribute their lab's resources to eagle-i. The majority of respondents (83%) said that they 'would return' or 'would probably return' to the Search application. The Search user interface, search/browse functions, auto suggest, and filtering by resource location were all rated as 'excellent' or 'good' by most respondents. In addition, we have received excellent suggestions from two usability studies and an expert review of eagle-i; we plan to incorporate these suggestions in ongoing development.

Lessons learned

Our biocuration experience within the eagle-i project was unique due to the ontology-driven technology, the diverse and geographically distributed Curation team, the separation between data collection and curation, and the wide spectrum of resources annotated. Nevertheless, the following lessons learned can be generalized to many other biocuration projects.

Balancing the data you need with the data you can get

Most researchers are not adept at creating structured data and in almost all systems there exists the need to balance data provided by end users with the added quality and structure that can be provided by biocuration. In the case of eagle-i, we employed Resource Navigators to obtain and enter the data and built an ontology-based curation interface to serve high quality data to our Search interface. However, we are now addressing the issue of how to better obtain more and higher quality data with less effort. This may result in modifications to our Data Collection tool for ease of use by researchers directly, enhancement of ETL methods, and interoperability with the improving landscape of basic research LIMS. Therefore, a clear understanding of your user community's capability and incentives to provide high quality data should drive decisions about what types of tools, interfaces and curation will be required. Additionally, the ability to adapt the system to address different needs and maintain the data over time is important. Finally, developers of any system can have an eye towards longer term instruction of data providers in capturing better and more structured data.

Documentation and quality assurance are iterative

Establishing defined QA processes and metrics *a priori* to monitor the quality of the data is critical. These processes should be reviewed and updated regularly following changes to the ontologies and system functionality and as part of routine QA efforts. In addition to a pre-determined QA checklist, random spot-checks and manual QA is also valuable, as new changes to the system can be missed. Deciding on how to address legacy data ahead of time as the system evolves is also important; for example, will data

annotated with less expressivity be updated when more fields are added? To promote data quality, clear standards and guidelines should be developed, and the target end user for documentation should be established. For the eagle-i project, this meant producing documentation that could be used not only by the current Curation and Resource Navigation teams, but also by future institutional participants who may not have extensive training in data management and resource annotation. We also developed curation-specific documentation for our QA processes. Therefore, some documentation can serve multiple needs and by doing so ensure consistency, but there may also be the need for task or user-specific documentation. Tailoring the documentation facilitates higher quality data.

Tool and technology choices

Considering end users is essential to the development of applications and evaluation of technology choices, particularly with regard to system and data management requirements. The development of a single interface for curators and end users, such as that developed by eagle-i, has advantages and disadvantages. One advantage is that refining a single interface as the requirements are gathered from all users is simpler and less expensive. On the other hand, it is seldom the case that a single tool will meet all user needs. The eagle-i Data Collection Tool was intentionally built to capture the complex, multifaceted semantic relationships in the data; however, this tool is less intuitive for lab users. Some databases, such as TAIR, primarily use prepared spreadsheet templates for user-based contributions to address this issue. Certainly, one should consider differential functionality or development of different tools for different users.

Tradeoffs always exist with respect to technology choices and prioritization of resources. For example, the triple-store used in the eagle-i pilot system limited our ability to have advanced provenance features. Upon considering technology choices when building a curation tool, using a tool that can track different levels of provenance and has granular reporting capabilities is important for the following reasons: first, one can never know in advance all the aspects of the data that will be relevant to report on. Second, being able to have detailed provenance information for each record takes less curator time than keeping detailed notes and is more reliable because it does not rely on manual entry. Third, such capabilities further enable assessment of curator consistency, in particular when training new curators or after data model or system changes. In summary, one should evaluate technology limits and roadblocks with respect to requirements in the curation process and data lifetime, both at the inception of the project and at regular intervals thereafter.

Conclusion

The main goals of the eagle-i pilot project were to collect a diverse range of resources, model the ontology and Data Collection tool to accommodate this diversity of resource types, and to develop a user-friendly search interface to query the eagle-i repositories. Due to the geographic distribution and uniqueness of the nine sites participating in the project, a wide range of resources were collected, from common reagents such as antibodies and plasmids, to more rare and unusual resources such as submarines and domesticated caribou breeding colonies. The eagle-i Curation team was the bridge between the scientific knowledge base and the platforms that provide access to this knowledge. They were responsible for facilitating a cycle in which information is both provided and consumed. The development of the tools with an underlying ontology enabled structured representation of resources, semantic linking and advanced search capabilities. Finally, the Search application was intended to connect researchers with invisible or rarely shared resources, and to promote resource discovery and reuse; preliminary feedback indicates it may be successful in that regard. However, more work needs to be done on the overall system to improve usability and efficiency. Beyond these technical challenges, the eagle-i project has been an interesting social study on scientists and how they manage information about the entities used within the process of research. eagle-i and projects like it can help address these issues by providing a platform that encourages participation by rewarding researchers with access to valuable resources.

Acknowledgements

We should like to thank Daniela Bourges-Waldegg and Ted Bashor for their work in developing the Data Collection Tool and the Search Applications, Jackie Wirz for her help with edits, discussion and preparation and formatting the figures and Scott Hoffman for his help with formatting.

Funding

Funding for open access charge: The National Institutes of Health and the American Recovery and Reinvestment Act (1U24RR029825-01).

Conflict of interest. None declared.

References

1. Hirschman,J., Berardini,T.Z., Drabkin,H.J. *et al*. (2010) A MOD(ern) perspective on literature curation. *Mol. Genet. Genomics*, **283**, 415–425.
2. Howe,D., Costanzo,M., Fey,P. *et al*. (2008) Big data: the future of biocuration. *Nature*, **455**, 47–50.
3. Haendel,M. (2009) amplify the impact of your research: ensure that your data can be integrated into the electronic data stream. In: *Sixth Zebrafish Conference*, Rome, Italy.
4. Maltais,L.J., Blake,J.A., Chu,T. *et al*. (2002) Rules and guidelines for mouse gene, allele, and mutation nomenclature: a condensed version. *Genomics*, **79**, 471–474.
5. Bruford,E.A., Lush,M.J., Wright,M.W. *et al*. (2008) The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res.*, **36**(Database issue), D445–D448.
6. Torniai,C., Brush,M., Vasilevsky,N. *et al*. (2011) Developing an application ontology for biomedical resource annotation and retrieval: challenges and lessons learned. In: *Proceedings in International Conference on Biomedical Ontology*, Buffalo, NY, pp. 101–108.
7. Gupta,A., Bug,W., Marenco,L. *et al*. (2008) Federated access to heterogeneous information resources in the Neuroscience Information Framework (NIF). *Neuroinformatics*, **6**, 205–217.
8. Tenenbaum,J.D., Whetzel,P.L., Anderson,K. *et al*. (2011) The Biomedical Resource Ontology (BRO) to enable resource discovery in clinical and translational research. *J. Biomed. Inform.*, **44**, 137–145.
9. Borromeo,C., Whetzel,P., Haendel,M. *et al*. (2011) A linked open data approach to interoperability between biomedical resource inventories. *Summit on Translational Bioinformatics*, San Francisco, CA.
10. Cambon-Thomsen,A., Thorisson,G.A., Mabile,L. *et al*. (2011) The role of a bioresource research impact factor as an incentive to share human bioresources. *Nat. Genet.*, **43**, 503–504.
11. Peters,B. and The OBI Consortium (2009). Ontology for biomedical investigations. *Nature Precedings*, http://precedings.nature.com/documents/3623/version/1 (17 August 2011, date last accessed).
12. Taylor,C.F., Field,D., Sansone,S.A. *et al*. (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotech.*, **26**, 889–896.
13. Nelson,B. (2009) Data sharing: empty archives. *Nature*, **461**, 160–163.
14. Pink,DH. (2011) *Drive: The Surprising Truth About What Motivates Us*, Reprint. New York: Riverhead Trade.
15. Lesk,M. (2011) Encouraging Scientific Data Use [Internet]. The Fourth Paradigm a nature network blog, http://blogs.nature.com/fourthparadigm/2011/02/ (7 February 2011, date last accessed).
16. Bodenreider,O. and Stevens,R. (2006) Bio-ontologies: current trends and future directions. *Brief Bioinform.*, **7**, 256–274.
17. Bug,W.J., Ascoli,G.A., Grethe,J.S. *et al*. (2008) The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics*, **6**, 175–194.
18. Costanzo,M.C., Skrzypek,M.S., Nash,R. *et al*. (2009) New mutant phenotype data curation system in the Saccharomyces Genome Database. *Database*, **2009**, doi:10.1093/database/bap001.
19. Hancock,J.M., Adams,N.C., Aidinis,V. *et al*. (2007) Mouse Phenotype Database Integration Consortium: integration [corrected] of mouse phenome data resources. *Mamm. Genome*, **18**, 157–163.
20. Peters,B., Sidney,J., Bourne,P. *et al*. (2005) The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol*, **3**, e91.
21. Goble,C., Stevens,R., Hull,D. *et al*. (2008) Data curation + process curation=data integration + science. *Brief. Bioinform.*, **9**, 506–517.