TraitBank: Practical semantics for organism attribute data

Cynthia S. Parr[a,c], Katja Schulz[§a], Jennifer Hammock[a], Nathan Wilson[bd], Patrick Leary[be], Jeremy Rice[b], Robert J. Corrigan Jr.[a]

[a] National Museum of Natural History, Smithsonian Institution, Washington, D. C., USA

[b] *Marine Biological Laboratory, Woods Hole, Massachusetts, USA*

[c] *Current address: United States Department of Agriculture, Agricultural Research Service, Beltsville, Maryland, USA*

[d] *Current address: MassChallenge, Boston, Massachusetts, USA*

[e] *Current address: California Academy of Sciences, San Francisco, California, USA*

[§]*Corresponding author, schulzk@si.edu*

Abstract. Encyclopedia of Life (EOL) has developed a new repository for organism attribute (trait) data called TraitBank (http://eol.org/traitbank). TraitBank aggregates, manages and serves attribute data for organisms across the tree of life, including life history characteristics, habitats, distributions, ecological relationships and other data types. We describe how TraitBank ingests and manages these data in a way that leverages EOL's existing infrastructure and semantic annotations to facilitate reasoning across the TraitBank corpus and interoperability with other resources. We also discuss TraitBank's impact on users and collaborators and the challenges and benefits of our lightweight, scalable approach to the integration of biodiversity data.

Keywords: biodiversity, ontologies, Semantic Web, traits, ecology, evolution, taxonomy, aggregation

1. Introduction

While human knowledge of life on Earth is vast, there is no easy way to query all the information accummulated in hundreds of years of biodiversity research and documentation. Even simple questions like "which plants have yellow flowers?" or "what do sharks eat?" are impossible to answer with confidence.

Biologists have captured and managed information about morphology, behavior, life history, and ecological interactions in many different ways. Most of this information survives in the form of free text or data tables in published papers, if it survives at all [19]. Lately communities have started to annotate those papers [3], extract information from text [27, 39], and build special-purpose databases of trait data, for example, TRY for plants [23] and SeaLifeBase[1] for marine organisms. In addition, modern researchers are more likely to archive and share data sets associated with their published studies in open data repositories such as Dryad [40] and PANGAEA[2]. While this is a critical development, there is still little standardization in how biologists talk about the characteristics of organisms, how they describe the context of their observations, and how they document the methods with which the data were collected. This means that the information in many data sets is not easily discovered, integrated or repurposed.

This lack of data standards impedes progress in the ecological, conservation, and phylogenetic research communities, who need effective ways to quickly discover and consume data in the coming era of data-intensive science [e.g., 16, 17, 18]. For example, marine environmental modelers need high-quality inputs about large numbers of species in order to understand current and historical distributions of species; how these distributions are impacted by environmental changes such as climate change, overharvesting, or invasive species; how biological communities function to provide ecosystem services; and what could happen to these services under future scenarios that change the composition of these

[1] http://sealifebase.org

[2] http://www.pangaea.de

[2] http://www.pangaea.de

communities. Such large-scale data have also been identified by DIVERSITAS[3] and the Global Earth Observation Biodiversity Observation Network (GEO BON)[4] as likely to be required by the Intergovernmental science-policy Platform on Biodiversity and Ecosystem Services (IPBES) [34]. Aggregating and standardizing these data, making them freely re-usable, and providing discovery mechanisms for them could facilitate rapid analyses for investigators interested in these urgent problems.

This paper describes TraitBank®, a system designed by the Encyclopedia of Life (EOL) to acquire, organize and serve biodiversity attribute data on a global scale across the entire tree of life – currently estimated at nearly two million species.[5] It describes our approaches to semantics, details TraitBank's implementation, and evaluates the system with respect to implications for interoperability and impact on community and provider processes.

2. Approach

TraitBank mobilizes data from diverse sources including biodiversity databases (e.g., Global Biodiversity Information Facility (GBIF)[6], Global Biotic Interactions[7], Ocean Biogeographic Information System (OBIS)[8], Paleobiology Database[9]), literature repositories (e.g., Dryad [40], Ecological Archives[10], PANGAEA[11]), natural history collections, and citizen science projects. Legacy or previously unpublished data are also represented, and some data are derived from text mining projects [27, 39]. Access to the data is free and open. While some data sets are released under the Attribution Creative Commons License[12], most TraitBank data can be used and redistributed without copyright restrictions.

In addition to traditional "trait data" like *body size*, *flower color*, and *onset of fertility*, TraitBank also features structured attributes like the *number of sequences in GenBank*, *type specimen repository*, and *human population density within the geographic range of a taxon*. TraitBank data include individual measurements (e.g., the wood density of a particular tree) as well as statistics (e.g., the mean body mass from a particular sample). In addition, there are facts derived from the literature (e.g., blue whales are known to prey on krill or dandelions have yellow flowers).

TraitBank leverages EOL's existing network of content partners and Content Creation Community [38] and employs the EOL relational database frameworks (providing advanced taxonomic names resolution) in combination with existing data standards and domain ontologies. Rather than developing a comprehensive semantic framework for the integration of trait data, TraitBank simply links data records to relevant ontologies and controlled vocabularies. These links improve the discoverability and queriability of the data and provide interoperability with other semantic resources, but more principled inference is left to end users. This lightweight semantic approach allows for the efficient management of a large and diverse data store and ensures scalability as the system grows.

TraitBank is designed for use by a wide audience including biodiversity researchers, information and data scientists, but also teachers, students, and the public. It provides both human and machine accessible query interfaces, and trait data are displayed on EOL taxon pages making them readily accessible to the EOL user base of about 6 million unique users per year[13].

2.1. Data model

To represent trait data, TraitBank uses and extends TDWG Darwin Core [42] (Figure 1), the most widely used standard for exchange of biodiversity data. Darwin Core Archives are already the preferred method for sharing media, references, and taxonomic data with EOL. Other prominent initiatives like GBIF, OBIS, and the Atlas of Living Australia (ALA)[14] support Darwin Core, and it has gained wide acceptance in the natural history collection and citizen science communities. Adoption of this standard by an increasing number of projects will enable data providers to efficiently share their resources with multiple biodiversity information systems [2].

[3] http://www.diversitas-international.org/
[4] http://www.earthobservations.org/geobon.shtml
[5] http://eol.org
[6] http://www.gbif.org/
[7] http://globalbioticinteractions.org/
[8] http://www.iobis.org
[9] http://paleobiodb.org/
[10] http://esapubs.org/archive/default.htm
[11] http://www.pangaea.de/
[12] http://creativecommons.org/licenses/by
[13] Data from 1 October 2013 to 30 September 2014
[14] http://www.ala.org.au

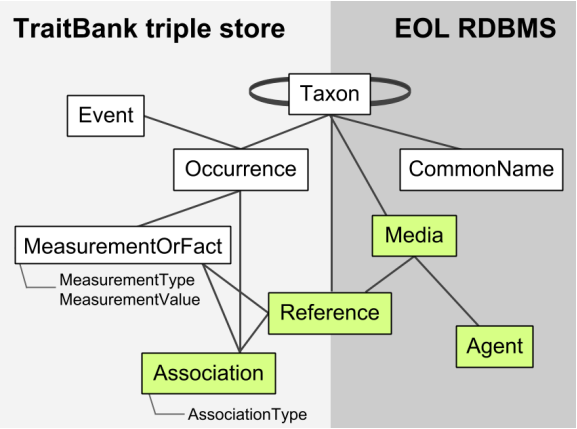

Figure 1. Data model and architecture for TraitBank/EOL. Elements are from Darwin Core except for the following extensions developed by EOL: Media (with Audubon Core)[15], References (with BIBO)[16], Associations (under development), and Agents[17]. Only the most important properties are indicated. TraitBank elements may hold only pointers to elements managed in the EOL relational database management system (RDBMS), like taxon names and references.

Each TraitBank record is associated with an `Occurrence`, which links to the taxon identifier. The `Occurrence` may also include the context in which the trait was recorded (e.g. geospatial information, dates, sex, life stage, individual count). `LifeStage` and `Sex` values are standardized whenever possible through links to terms from the Phenotypic Quality Ontology (PATO) [25] or the Uber Anatomy Ontology (UBERON) [26].

The Darwin Core `MeasurementOrFact` extension holds information about the trait measured and other metadata. `measurementType` describes the trait that was measured using a Uniform Resource Identifier, URI, from a domain ontology (e.g., Plant Trait Ontology [20] or Vertebrate Trait Ontology [28]). `measurementValue` holds either a number or a categorical value represented by a URI from an ontology, if possible (e.g., PATO or Environments Ontology (ENVO) [7]). Associated measurement metadata may include `measurementUnit` (mapped to the Units of Measurement Ontology, UO [18]), `measurementAccuracy` and `measurementMethod` (not yet standardized) , and `StatisticalMethod`, (e.g. mean or maximum), mostly mapped to the Semanticscience Integrated Ontology (SIO)[19]. In addition to these frequently documented parameters, custom fields can be created to accommodate any metadata extracted from the source.

Interactions among species (e.g., predator-prey relationships) are handled using a new `Associations` Darwin Core extension which is still under development. This extension references two records in the `Occurrence` extension, with `AssociationType` indicating the type of relationship (e.g., X *feeds on* Y, A *parasitizes* B).

As with other content on EOL, provenance of TraitBank data is handled using rich attribution metadata via fields from Dublin Core[20] (e.g., `bibliographicCitation`, `contributor`, `source`) and Darwin Core (e.g., `measurementDeterminedBy`, `identifiedBy`, `recordedBy`), with structured references supported using an EOL extension based on the Bibliographic Ontology (BIBO)[21].

2.2. Taxonomic semantics

Taxonomic names reconciliation is at the heart of any effort to integrate biodiversity information [32]. Since there is no comprehensive consensus classification for organism names, EOL maps each data record to names in multiple taxonomic hierarchies from several scientific providers. Synonyms, misspellings, ranks, and parent taxa are taken into account during the reconciliation process. Rather than attempt to fully capture these complex interactions semantically [14], TraitBank reflects data structures already developed to represent the multiple classifications managed in the EOL relational database [31].

Scientific names in TraitBank are designated with the Darwin Core property `scientificName` and are typically associated with Taxon URIs that have the rdf:type of `Taxon`. These in turn are associated to an EOL taxon page URL (e.g., http://eol.org/pages/328615) using the `taxonConceptID` predicate. These Taxon URIs associate a data point with a particular page and describe the parent/child relationships between the taxa. The parent/child relationships use the `parentNameUsageID` predicate.

[15] http://eol.org/schema/media_extension.xml
[16] http://eol.org/schema/reference_extension.xml
[17] http://eol.org/schema/agent_extension.xml
[18] http://code.google.com/p/unit-ontology/
[19] http://semanticscience.org
[20] http://dublincore.org/
[21] http://bibliontology.com/

3. Implementation

To ensure that TraitBank would meet the needs of the scientific community and to build a stakeholder base ready to use it, EOL convened workshops and an advisory panel early in the development process. Scientists who attended workshops sponsored by EOL's Biodiversity Synthesis Center at the Field Museum over a period of four years provided high-level community requirements. A workshop in Washington, DC in September 2012 brought together more than twenty experts from biology and computer science, including semantics, to focus on the questions that could be addressed with a comprehensive, integrated trait repository and associated software and infrastructure requirements. Teleconferences with an 11-person panel of scientists and technologists drawn from the above workshops informed iterative design and development. Following the first production release of TraitBank in January 2014, futher refinements to the technology were implmeneted on an as-needed basis, and the focus of the development team shifted to increasing the amount of content aggregated into TraitBank.

The initial data sets targeted for ingestion into TraitBank were chosen to quickly achieve broad taxonomic coverage for a number of commonly studied ecological and life history traits. In addition to iconic data sets like Pantheria [21], IUCN Redlist[22], and the Global Wood Density Database [9], we looked especially for trait data that would be useful for marine biodiversity science, a focus of one of TraitBank's sponsors. The TraitBank corpus has since grown to include more than 11 million data records sourced from over 50 data sets. They represent more than 300 attributes for over 1.7 million taxa (Table 1). Moving ahead our strategy for data acquisition is guided by the needs of our audiences, sponsors, and partners.

Table 1. TraitBank contents as of 27 January 2015 as retrieved from http://eol.org/statistics. Trait types include both MeasurementTypes and AssociationTypes. An overview of TraitBank data sets is available at http://eol.org/collections/97700

Data sets	52
Trait types	331
Individual data records	11,063,667
Taxa with at least one data record	1,730,789
Total triples	218,893,457

[22] http://www.iucnredlist.org/

3.1. Data import

Most TraitBank data are imported from other databases via PHP connectors or uploaded directly via Darwin Core Archive files[23]. A custom spreadsheet template is also available to support conversion of tabular data to a Darwin Core Archive[24].

If a data set introduces new concepts (attributes, values or metadata) to TraitBank, the new terms and their defintions must be added to the TraitBank URI registry before the data can be harvested [10]. Each attribute is mapped to broad subject categories (Distribution, Physical Description, Ecology, Life History and Behavior, Evolution and Systematics, Physiology and Cell Biology, Molecular Biology and Genetics, Conservation, Relevance to Humans and Ecosystems, Notes, Names and Taxonomy, Database and Repository Coverage), and basic semantic relationships are entered into the system (see below). Attributes are also ranked based on their putative audience appeal, so that attributes of greater interest to EOL audiences can be displayed more prominently in the EOL interface (see below).

3.2. Semantic annotation

[23] http:// eol.org/info/structured_data_archives
[24] http://eol.org/info/cp_spreadsheet

Table 2. Some of the frequently referenced ontologies in TraitBank

Subject Areas	Ontology	Example terms
Statistics	Semanticscience Integrated Ontology (SIO) [12]	mean, minimal value, standard deviation
Units of measure	Units of Measurement Ontology (UO) [15]	meter, years, degree Celsius
Habitat information	Environments Ontology (EnvO) [7]	wetland, desert, snow field
Attributes of organisms	Phenotype Quality Ontology (PATO) [25]	aerobic, conical, evergreen
Plant attributes	Plant Trait Ontology (TO) [20]	flower color, life cycle habit, salt tolerance
Animal attributes	Vertebrate Trait Ontology (VT) [28]	body mass, total life span, onset of fertility
Animal natural history	Animal Natural History and Life History Ontology (ETHAN) [30]	nocturnal, oviparous, scavenger

If a provider supplies semantic annotations with their data, these mappings are preserved in TraitBank. However, only three TraitBank data partners, Environments-EOL [27], Global Biotic Interactions [35], and PolyTraits [13] fall into this category. Most of the resources we aggregate are not "born semantic," i.e., the data come to us with labels, some metadata, and sometimes an associated article explaining the rationale and methods of the study. In these cases, EOL staff analyze the meaning of each attribute and select formally-defined semantic terms to represent them. Terms from ontologies under active development by engaged communities are preferred. These include Open Biological and Biomedical Ontologies (OBO) Foundry ontologies such as Molecular Function (GO)[25], Plant Ontology (PO)[26], Phenotypic Quality (PATO)[27] and Chemical Entities of Biological Interest (CHEBI)[28], as well as OBO Foundry candidate ontologies such as Environment Ontology (ENVO) [29], Plant Trait Ontology (TO) [30], Uber Anatomy Ontology (UBERON)[31] and Ontology of Biological Attributes (OBA)[32] (Table 2).

Not all concepts encountered in TraitBank data sets can be matched to terms in current ontologies or controlled vocabularies. Especially in the life history and ecology domains ontology coverage is still sparse. EOL staff therefore regularly propose new terms for adoption into ontologies like PATO and CHEBI, and we are involved in efforts to extend the Relations Ontology (RO) [33], Population and Community Ontology (PCO) [34], and Biological Collections Ontology (BCO)[35] to improve coverage of the different dimensions of biotic interactions.

Many traits are highly complex and require referencing of more than one class, potentially from multiple ontologies. Some new terms are therefore created through Term Genie[36], a post-composition tool that formally constructs composite attributes by combining classes from PATO and GO or UBERON. For example, secondary xylem volumetric density[37] (i.e., wood density) and cell shape[38] are attributes from TraitBank data sets that have been added to OBA.

The goal is for new TraitBank terms to become part of the most relevant ontologies so that they can be managed by domain experts and readily discovered by users and semantic web developers. Since adding new terms to ontologies can often take a considerable amount of time, EOL creates provisional URIs while term requests are under review.

TraitBank terms, their definitions, and URIs are listed in the TraitBank Data Glossary[39] which is populated automatically from the TraitBank URI registry. The entry for each attribute features a quick link to a data search for all relevant TraitBank

[25] http://geneontology.org/
[26] http://www.plantontology.org/
[27] http://wiki.obofoundry.org/wiki/index.php/PATO:Main_Page
[28] http://www.ebi.ac.uk/chebi/
[29] http://environmentontology.org/
[30] http://archive.gramene.org/plant_ontology/
[31] http://uberon.github.io/
[32] http://wiki.geneontology.org/index.php/Extensions/x-attribute
[33] https://code.google.com/p/obo-relations/
[34] https://code.google.com/p/popcomm-ontology/
[35] https://github.com/tucotuco/bco
[36] http://www.berkeleybop.org/software/termgenie
[37] http://purl.obolibrary.org/obo/OBA_1000040
[38] http://purl.obolibrary.org/obo/OBA_0000052
[39] http://eol.org/data_glossary

records (see below). The URIs of EOL provisional terms resolve to relevant entries in this Data Glossary. As domain ontologies increase their coverage, fewer terms and definitions will have to be maintained in the TraitBank Data Glossary.

For terms imported from ontologies and controlled vocabularies, the Data Glossary entry can serve as a backup when the original resource is moved or temporarily unavailable. If the definition of a term changes in the source ontology, the Data Glossary entry also serves as a record of the definition implied in the TraitBank annotation. Links to individual glossary entries can be generated based on URIs (e.g. the OBA URI for cell shape is http://purl.obolibrary.org/obo/OBA_0000052, but the definition of this term can also be accessed in the EOL Data Glossary via this URL: http://eol.org/data_glossary#http___purl_obolibrary_org_obo_OBA_0000052).

3.3. Reasoning

Because of the complexity of semantic reasoning and the challenges of reasoning across highly heterogeneous or web-scale data sets [33, 41] the availability of semantic reasoning capabilities was limited in the first release of TraitBank, with the goal to add additional reasoning later as the system matures and as demand requires. However, conversion relationships of units (e.g., from *g* to *kg*), logarithmic transformations, and some equivalent and inverse relationships (e.g., *preysUpon* and *hasPredator*) are already implemented. Eventually, reasoning can be expanded to infer values based on phylogeny, or to leverage semantic similarity for searches. As the corpus of data in TraitBank grows the value of this work will increase, and it is therefore a priority for the next phase of development.

3.4. Data quality

The quality of the data represented in TraitBank is highly variable. Early in the planning process, we made the decision to not only aggregate tightly curated data but to also recruit data in need of review (e.g., data from citizen science and text mining projects) and data of questionable provenance (e.g., summary statistics without original sources). Such provisional data can make important contributions to the biodiversity knowledge base in cases where no data from scientific studies are available, where such data cannot be shared and reused freely, or where the expert curated data are of limited scope. Feedback from stakeholders has since confirmed that, at least for some applications, provisional data are better than no data at all.

Data quality concerns may also extend to the accuracy of the semantic annotations in TraitBank. Most of these links are created by trained biologists, but not necessarily by domain experts. Also, when data sources provide only vague descriptions of attributes, values, and metadata there will be some conjecture involved in the selection of the appropriate semantic context.

Finally, taxonomic name reconciliation relies on algorithms that may yield suboptimal results if there are unresolved homonyms, unrecognized synonym relationships, contradictory taxonomic data from

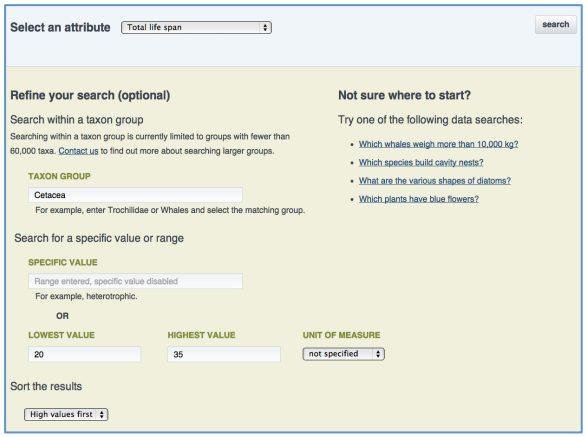

Figure 2. EOL data search interface for TraitBank, accessible at http://eol.org/data_search

different providers or undocumented lexical variants of taxon or author names. As a result data records may sometimes not be associated with the most appropriate EOL taxon page.

TraitBank users in need of high quality data are advised to thoroughly check data sources, semantic annotations, and taxon mappings before employing the data in scientific analyses. The metadata needed to perform these assessments are provided alongside TraitBank records in all data delivery interfaces (see below).

3.5. Data search, download, and API

TraitBank data can be queried and downloaded through the EOL data search interface[40] which is

[40] http://eol.org/data_search

accessible through numerous links on EOL web pages. A JSON-LD service is provided for machine access to the data, and relevant records are displayed on taxon pages throughout the EOL web site.

The EOL data search (Figure 2) supports queries based on individual attributes. A generic search returns all TraitBank records for a given attribute like *tail length* or *plant growth habit*. Searches can be refined by specifying a value or range of values, and they can be restricted to a particular taxonomic group. Filtering by group currently relies on parent/child relationships in the National Center for Biotechnology Information (NCBI)[41] and Catalogue of Life [37] classifications, so only records for taxa that are featured in one or both of these hierarchies are returned for taxon-restriced queries.

Search results can be explored in the EOL interface, or they can be downloaded as a CSV (comma-separated values) file. The CSV format is easily parsed and can be imported into common spreadsheet applications for manual or semi-automatic processing. The downloaded file features comprehensive information about each data record. It includes the unique EOL identifier for the associated taxon along with its scientific name and a common name if available. Each data row specifies the attribute label (e.g., *egg size* or *leaf shape*), the value (e.g., 38.5 or *acicular*), and units (e.g., *mg* or *km*) when appropriate. Most unit types are automatically normalized into comparable values. However, the raw value and units are also provided. In addition to attribute and value labels, all relevant URIs are provided. The metadata include the data provenance and context information such as *life stage* or *geographical location*.

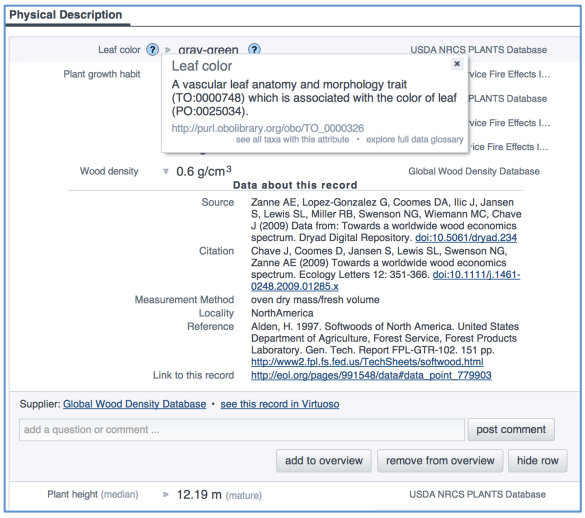

Fig. 3. Part of a data tab of an EOL taxon page. Wood density is expanded to show rich metadata. Users can select info buttons (?) to access definitions of terms, URIs, and links to the glossary and data search interface.

To support data-driven web-applications, a JSON-LD application programming interface (API)[42], is available [24]. Based on EOL page identifiers (which are accessible through the EOL Search API[43]) this service returns all TraitBank records for a given taxon; e.g., a url of the form http://eol.org/api/traits/328067 will return all data for the kinkajou, *Potus flavus*, which has EOL page id #328067.

3.6. TraitBank data on EOL taxon pages

TraitBank data are also displayed prominently on EOL taxon pages where they enrich the experience of millions of visitors each year. On many pages, these data fill important gaps by providing information that is not yet available in narrative form. Ubiquitous links to term definitions and data searches also encourage users to explore biodiversity data and give students and teachers easy access to sample data sets for instruction and projects

The Overview tab, which is the information center of each EOL taxon page, features a sample of relevant data records. By default, these records are selected automatically based on global, dynamic attribute rankings. The principal criterion for these rankings is the relative level of interest expected in a general audience. For example, attributes like *flower color* or *habitat* are presumed to be of greater interest than things like *outer ear length* or *germinative response to heat stimuli*.

A comprehensive presentation of TraitBank data is provided in the Data tab of EOL taxon pages. The default view of this tab shows a simple list of attribute labels, values, and data providers, ordered by subject (Distribution, Physical Description, Ecology, etc.). A dynamic user interface (Figure 3) gives access to the metadata for each record as well as URIs and definitions for attributes and categorical data values. Access to curation and commenting tools (see below), the data glossary, and data search interface are also provided.

Most TraitBank data are at the level of species or subspecies. For select physical, ecological, and life history attributes, the EOL Data tabs for higher taxa (genera, families, etc.) also feature summaries of the data represented among the taxonomic children of the group. Maximum and minimum values are displayed along with record and taxa counts and a quick link to a data search that yields relevant records.

[42] http://eol.org/traitbank#reuse
[43] http://eol.org/api/docs/search

3.7. Data curation

Any registered EOL member can review TraitBank content and report problems by adding comments to individual data records. EOL Curators – individuals with validated professional credentials - have the power to remove incorrect or suspect TraitBank records from public view. Flagged records remain visible to other curators and can be restored if flagged in error. Currently, TraitBank data providers do not receive notifications of comments and curator actions, but this feature will soon be available on an opt-in basis. This will allow data providers to benefit from the quality control activities of the EOL community.

EOL curators also participate in the selection of data for the Overview tabs of individual taxon pages. This activity is particularly important to ensure that the most interesting and informative records are highlighted for taxa of interest to a wide audience.

3.8. Architecture and technology

TraitBank is built on the RDF triple store integrated into the open source edition of the OpenLink Virtuoso Universal Server [44]. This datastore is accessed by EOL's application servers and backend data harvesting engine [31]. Virtuoso was selected over other candidate technologies such as neo4j[45] because using an RDF triple store made it easier to import and blend standard URI-based ontologies, URIs provided by content partners, and when necessary newly minted EOL URIs. The SPARQL[46] query language works well to efficiently query complex chains of relationships including recursive queries needed for traversing taxonomic hierarchies.

All code is available under an MIT open source license and is published to the EOL project on GitHub[47].

4. Evaluation and Conclusions

The amount of available biodiversity information has transcended our ability to process and analyze it. TraitBank addresses this impediment with an efficient, pragmatic approach to trait data integration that bridges taxon-specific and technology-specific systems. By organizing distributed knowledge from diverse sources into a lightweight, scalable framework, we facilitate its retrieval and reuse for a variety of applications, ranging from large-scale synthetic analyses of biodiversity to linked data products like the Knowledge Graph[48] and hands-on data science in the classroom.

4.1. Feedback from stakeholders

TraitBank was released in January 2014 after private (September 2013) and public (October 2013) beta test releases, with each test followed by a survey. Informal demonstrations to communities at several conferences have also been used to gather feedback. Some of the most valuable insights about the needs of TraitBank users were gained during the EOL-NESCent-BHL research sprint [29]. This event, scheduled only a week after TraitBank's public launch, brought together a diverse group of biologists and informaticians to tackle large-scale questions ecological and evolutionary questions with the aid of resources provided by EOL and the Biodiversity Heritage Library (BHL)[49]. During the four-day meeting, members of the TraitBank team had the opportunity to interact with users while they explored the TraitBank corpus and used it to assemble their own data sets.

Based on user feedback and observations of user behavior, new features are added to TraitBank (e.g., JSON-LD access on a taxon by taxon basis) and the data search and download functions have been revised. In addition, new data sets were imported to TraitBank in response to specific user requests.

Several improvements suggested by users are still in the planning stages. These include support for more complex data queries, with multiple facets across traits, metadata, values, and taxa,, improved presentation of results including visualizations, an R-interface for access to TraitBank data, and better performance of searches filtered by taxonomic group. Also, TraitBank's geographic keyword vocabulary is not yet standardized. Most locations are currently stored as text strings, preventing reasoning on geographic distribution data. These records need to be mapped to gazetters like GAZ[50], Geonames[51] and MarineRegions.org.

[44] http://virtuoso.openlinksw.com/
[45] http://www.neo4j.org/
[46] http://www.w3.org/TR/rdf-sparql-query/
[47] http://github.com/eol
[48] http://www.google.com/insidesearch/features/search/knowledge.html
[49] http://biodiversitylibrary.org
[50] http://bioportal.bioontology.org/ontologies/GAZ
[51] http://www.geonames.org/

4.2. Implications for interoperability

TraitBank fosters semantic interoperability both within and across domains by using URIs from ontologies that are also used in other systems. As the use of semantic technologies is already prevalent in genomics, morphology, ecology, and developmental biology communities, it makes sense to link newly exposed and annotated biodiversity trait information to these efforts. On the other hand, where existing ontologies do not yet capture knowledge adequately (e.g., missing terms, missing relations, missing definitions, complex taxonomic and nomenclatural semantics), our approach still allows progress in knowledge management and sharing in the most practical sense, even if not all elements of the system are interoperable.

Recent efforts to automate the description and measurement of organisms [3, 6, 22] accelerate the pace of data generation. While semantic annotation and open access publishing are likely to become an integral part of modern scientific workflows, standardization across data sets and domains remains in its infancy [11]. We expect that the semantic annotation of TraitBank resources will long remain a work in progress. The rapid growth and diversification of the corpus of data frequently requires the exploration of new subject areas. Even the annotation of existing data sets is often an iterative process as best practices develop in response to evolving needs for integration, new ontology resources, and feedback from domain and knowledge representation experts.

4.3. Impact on semantic community, data providers and research community

TraitBank is a starting point for the untangling of the vast riches compiled through centuries of biodiversity exploration. It will take time for it to mature into a comprehensive, consistent knowledge management platform that can supply highly curated, analysis-ready data products. Based on our experience so far, domain ontologies will have to become much more detailed if they are to be applied to the backlog of biodiversity data. Achieving the desired level of complexity without sacrificing interoperability will be an ongoing challenge. Because of its broad scope, TraitBank is in an ideal position to provide the stewards of many relevant domain ontologies with use cases that can help to optimize the development of their resources. We also anticipate that the prominent use of semantics in TraitBank will result in increased usage of ontologies in research applications.

TraitBank complements taxon or subject-specific trait databases by filling gaps (both in taxonomic and attribute space), by recruiting new types of data (e.g., from text-mining, citizen-science, and specimen data digitization efforts) and by integrating knowledge across the tree of life and multiple scientific domains. To promote progress in the aggregation of comprehensive data sets of particular interest to scientists and the public, EOL has funded projects like GloBi (Global Biotic Interactions) [35] and Environments-EOL [27]. For these communities and other ongoing projects like PolyTraits and OBIS, TraitBank provides a live platform for distribution and re-use that exposes their data to broader audiences and promotes significant community curation. For legacy data providers, such as the authors of literature-derived data sets, TraitBank improves discoverability of data that otherwise would not be exposed to the Linked Open Data (LOD) community [5]. Once provisioned to TraitBank, data can be discovered and re-used for a wide range of use cases, from simple fact-finding to "big data" modeling studies. Through its association with the Encyclopedia of Life web site, TraitBank also brings awareness of data science and interoperability efforts to novel audiences. Some of these new data users may themselves become data providers, e.g., through participation in citizen science[52] or transcription crowdsourcing projects[53].

With TraitBank only a year old, it is somewhat premature to assess its impact on scientific research. The TraitBank data search interface has so far been accessed over 5,000 times, and more than 1,500 data packages have been downloaded. Also, papers citing TraitBank as a data source are starting to appear in the literature (e.g., [1, 4, 8, 27, 35, 36, 43]). Future development efforts will focus on improving TraitBank's utility for reasearch by improving the search interface, exposing the data in more advanced machine-readable formats, employing standardized data quality descriptors, replacing provisional EOL terms with community-managed terms, and exploring the best use of reasoning within the EOL-TraitBank framework.

[52] http://inaturalist.org
[53] http://www.notesfromnature.org/

5. Acknowledgements

Support for TraitBank was provided by the Alfred P. Sloan Foundation, the Smithsonian Institution, the Marine Biological Laboratory, and the John D. and Catherine T. MacArthur Foundation. The production hardware infrastructure for the EOL website was supported by the Harvard Faculty of Arts and Sciences (FAS) Sciences Division Research Computing Group and the Smithsonian Institution. The TraitBank development team wishes to specifically thank Dr. Jesse Ausubel for his support and for his commitment to the entire Encyclopedia of Life initiative.

6. References

[1] Angeli, N. F., Otegui, J., Wood, M., & Gomez-Ruiz, E. P. (2014). A process to support species conservation planning and climate change readiness in protected areas. *PeerJ PrePrints* 2:e492v2 http://dx.doi.org/10.7287/peerj.preprints.492v2

[2] Baker, E., Rycroft, S., & Smith, V. S. (2014). Linking multiple biodiversity informatics platforms with Darwin Core Archives. *Biodiversity Data Journal* 2: e1039 http://dx.doi.org/10.3897/BDJ.2.e1039

[3] Balhoff, J. P., Dahdul, W. M., Kothari, C. R., Lapp, H., Lundberg, J. G., Mabee, P., Midford, P. E., Westerfield, M., & Vision, T. J. (2010). Phenex: Ontological Annotation of Phenotypic Diversity. *PLoS ONE*, *5*(5),10. http://dx.doi.org/10.1371/journal.pone.0010500

[4] Barnagaud, J.-Y., Papaïx, J., Gimenez, O., Svenning, J.-C. (2014), Dynamic spatial interactions between the native invader Brown-headed Cowbird and its hosts. *Diversity and Distributions*. http://dx.doi.org/10.1111/ddi.12275

[5] Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, *5*(3),1–22. http://dx.doi.org/10.4018/jswis.2009081901

[6] Burleigh, J. G., Alphonse, K., Alverson, A. J., Bik, H. M., Blank, C., Cirranello, A. L., Cui, H., Daly, M., Dietterich, T. G., Gasparich, G., Irvine, J., Julius, M., Kaufman, S., Law, E., Liu, J., Moore, L., O'Leary, M. A., Passarotti, M., Ranade, S., Simmons, N. B., Stevenson, D. W., Thacker, R. W., C Theriot, E., Todorovic, S., Velazco, P. M., Walls, R. L., Wolfe, J. M., & Yu, M. (2013). Next-generation phenomics for the Tree of Life. *PLoS Currents*. http://dx.doi.org/10.1371/currents.tol.085c713acafc8711b2ff7010a4b03733

[7] Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., Lewis, S. E., & the ENVO Consortium. (2013). The environment ontology: contextualising biological and biomedical entities. *Journal of Biomedical Semantics* 4,43. http://dx.doi.org/10.1186/2041-1480-4-43

[8] Caldwell, I. R. & Hart, E. M. (2014). Using Encyclopedia of Life's TraitBank to identify plant traits associated with vulnerability. *PeerJ PrePrints* 2:e491v1 http://dx.doi.org/10.7287/peerj.preprints.491v1

[9] Chave, J., Coomes, D., Jansen, S., Lewis, S. L., Swenson, N. G., & Zanne, A. E. (2009). Towards a worldwide wood economics spectrum. *Ecology Letters* 12,351–366. http://dx.doi.org/10.1111/j.1461-0248.2009.01285.x

[10] Courtot, M., Gibson, F., Lister, A. L., Malone, J., Schober, D., Brinkman, R. R., & Ruttenberg, A. (2011). MIREOT: The minimum information to reference an external ontology term. *Applied Ontology*, *6*,23–33. http://dx.doi.org/10.1038/npre.2009.3576.1

[11] Deans, A. R., Lewis, S. E., Huala, E., Anzaldo, S. S., Ashburner, M., Balhoff, J. P., Blackburn, D. C., Blake, J. A., Burleigh, J. G., Chanet, B., Cooper, L. D., Courtot, M., Csösz, S., Cui, H., Dahdul, W., Das, S., Dececchi, T. A., Dettai, A., Diogo, R., Druzinsky, R. E., Dumontier, M., Franz, N. M., Friedrich, F., Gkoutos, G. V., Haendel, M., Harmon, L. J., Hayamizu, T. F., He, Y., Hines, H. M., Ibrahim, N., Jackson, L. M., Jaiswal, P., James-Zorn, C., Köhler, S., Le-

cointre, G., Lapp, H., Lawrence, C. J., Le Novère, N., Lundberg, J. G., Macklin, J., Mast, A. R., Midford, P. E., Mikó, I., Mungall, C. J., Oellrich, A., Osumi-Sutherland, D., Parkinson, H., Ramírez, M. J., Richter, S., Robinson, P. N., Ruttenberg, A., Schulz, K. S., Segerdell, E., Seltmann, K. C., Sharkey, M. J., Smith, A. D., Smith, B., Specht, C. D., Squires, R. B., Thacker, R. W., Thessen, A., Fernandez-Triana, J., Vihinen, M., Vize, P. D., Vogt, L., Wall, C. E., Walls, R. L., Westerfeld, M., Wharton, R. A., Wirkner, C. S., Woolley, J. B., Yoder, M. J., Zorn, A. M., & Mabee, P. (2015). Finding our way through phenotypes. *PLOS Biology* 3(1), e1002033. http://dx.doi.org/10.1371/journal.pbio.1002033

[12] Dumontier, M., Baker, C. J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N. R., Duck, G., Furlong, L. I., Keath, N., Klassen, D., McCusker, J. P., Queralt-Rosinach, N., Samwald, M., Villanueva-Rosales, N., Wilkinson, M. D., & Hoehndorf, R. (2014). The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *Journal of Biomedical Semantics* 5(1),14. http://dx.doi.org/10.1186/2041-1480-5-14

[13] Faulwetter, S., Markantonatou, V., Pavloudi, C., Papageorgiou, N., Keklikoglou, K., Chatzinikolaou, E., Pafilis, E., Chatzigeorgiou, G., Vasileiadou, K., Dailianis, T., Fanini, L., Koulouri, P., & Arvanitidis, C. (2014). Polytraits: A database on biological traits of marine polychaetes. *Biodiversity Data Journal*, *2*, e1024. http://dx.doi.org/0.3897/BDJ.2.e1024

[14] Franz, N. M., & Thau, D. (2010). Biological taxonomy and ontology development: Scope and limitations. *Biodiversity Informatics*, *7*, 45–66.

[15] Gkoutos, G. V, Schofield, P. N., & Hoehndorf, R. (2012). The Units Ontology: a tool for integrating units of measurement in science. *Database : The Journal of Biological Databases and Curation*, *2012*, bas033. http://dx.doi.org/10.1093/database/bas033

[16] Guisan, A. (2014). Biodiversity: Predictive traits to the rescue. *Nature Climate Change*, *4*(3), 175–176. http://dx.doi.org/10.1038/nclimate2157

[17] Harfoot, M., & Roberts, D. (2014). Taxonomy: Call for ecosystem modelling data. *Nature*, *505*(7482),160. http://dx.doi.org/10.1038/505160a

[18] Harmon, L. J., Baumes, J., Hughes, C., Soberon, J., Specht, C. D., Turner, W., Lisle, C., & Thacker, R. W. 2013. Arbor: Comparative Analysis Workflows for the Tree of Life. *PLOS Currents*. http://dx.doi.org/10.1371/currents.tol.099161de5eabdee073fd3d21a44518dc.

[19] Heidorn, P. (2008). Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends* 57(2) Fall 2008. Institutional Repositories: Institutional Repositories: Current State and Future. Edited by Sarah Sheeves and Melissa Cragin. http://hdl.handle.net/2142/9127

[20] Jaiswal, P., Ware, D., Ni, J., Chang, K., Zhao, W., Schmidt, S., Pan, X., Clark, K., Teytelman, L., Cartinhour, S., Stein, L., & McCouch, S. (2002). Gramene: development and integration of trait and gene ontologies for rice. *Comparative and Functional Genomics*, *3*(2), 132–136. http://dx.doi.org/10.1002/cfg.156

[21] Jones, K. E., Bielby, J., Cardillo, M., Fritz, S. A., O'Dell, J., Orme, C. D. L., Safi, K., Sechrest, W., Boakes, E. H., Carbone, C., Connolly, C., Cutts, M. J., Foster, J. K., Grenyer, R., Habib, M., Plaster, C. A., Price, S. A., Rigby, E. A., Rist, J., Teacher, A., Bininda-Emonds, O. R. P., Gittleman, J. L., Mace, G. M., & Purvis, A. 2009. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* 90:2648.

[22] Kao, R. H., Gibson, C. M., Gallery, R. E., Meier, C. L., Barnett, D. T., Docherty, K. M., Blevins, K. K., Travers, P. D., Azuaje, E., Springer, Y. P., Thibault, K. M., McKenzie, V. J., Keller, M., Alves, L. F., Hinckley, E.-L. S.,

Parnell, J., & Schimel, D. (2012). NEON terrestrial field observations: designing continental-scale, standardized sampling. *Ecosphere*, *3*(12),1–17. http://dx.doi.org/10.1890/ES12-00196.1

[23] Kattge, J., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Bönisch, G., Garnier, E., Westoby, M., Reich, P. B., Wright, I. J., Cornelissen, J. H. C., Violle, C., Harrison, S. P., Van Bodegom, P. M., Reichstein, M., Enquist, B. J., Soudzilovskaia, N. A., Ackerly, D. D., Anand, M., Atkin, O., Bahn, M., Baker, T. R., Baldocchi, D., Bekker, R., Blanco, C. C., Blonder, B., Bond, W. J., Bradstock, R., Bunker, D. E., Casanoves, F., Cavender-Bares, J., Chambers, J. Q., Chapin Iii, F. S., Chave, J., Coomes, D., Cornwell, W. K., Craine, J. M., Dobrin, B. H., Duarte, L., Durka, W., Elser, J., Esser, G., Estiarte, M., Fagan, W. F., Fang, J., Fernández-Méndez, F., Fidelis, A., Finegan, B., Flores, O., Ford, H., Frank, D., Freschet, G. T., Fyllas, N. M., Gallagher, R. V., Green, W. A., Gutierrez, A. G., Hickler, T., Higgins, S. I., Hodgson, J. G., Jalili, A., Jansen, S., Joly, C. A., Kerkhoff, A. J., Kirkup, D., Kitajima, K., Kleyer, M., Klotz, S., Knops, J. M. H., Kramer, K., Kühn, I., Kurokawa, H., Laughlin, D., Lee, T. D., Leishman, M., Lens, F., Lenz, T., Lewis, S. L., Lloyd, J., Llusià, J., Louault, F., Ma, S., Mahecha, M. D., Manning, P., Massad, T., Medlyn, B. E., Messier, J., Moles, A. T., Müller, S. C., Nadrowski, K., Naeem, S., Niinemets, Ü., Nöllert, S., Nüske, A., Ogaya, R., Oleksyn, J., Onipchenko, V. G., Onoda, Y., Ordoñez, J., Overbeck, G., Ozinga, W. A., Patiño, S., Paula, S., Pausas, J. G., Peñuelas, J., Phillips, O. L., Pillar, V., Poorter, H., Poorter, L., Poschlod, P., Prinzing, A., Proulx, R., Rammig, A., Reinsch, S., Reu, B., Sack, L., Salgado-Negret, B., Sardans, J., Shiodera, S., Shipley, B., Siefert, A., Sosinski, E., Soussana, J.-F., Swaine, E., Swenson, N., Thompson, K., Thornton, P., Waldram, M., Weiher, E., White, M., White, S., Wright, S. J., Yguel, B., Zaehle, S., Zanne, A. E., & Wirth, C. (2011). TRY - a global database of plant traits. *Global Change Biology*, *17*(9), 2905–2935. http://dx.doi.org/10.1111/j.1365-2486.2011.02451.x

[24] Lanthaler, M., & Gütl, C. (2012). On using JSON-LD to create evolvable RESTful services. In *Proceedings of the 3rd International Workshop on RESTful Design WSREST 2012 at WWW2012* (pp. 25–32). ACM Press. http://dx.doi.org/10.1145/2307819.2307827

[25] Mabee, P. M., Ashburner, M., Cronk, Q., Gkoutos, G. V, Haendel, M., Segerdell, E., Mungall, C., & Westerfield, M. (2007). Phenotype ontologies: the bridge between genomics and evolution. *Trends in Ecology & Evolution*, *22*(7), 345–50. http://dx.doi.org/10.1016/j.tree.2007.03.013

[26] Mungall, C. J., Torniai, C., Gkoutos, G. V, Lewis, S. E., & Haendel, M. A. (2012). Uberon, an integrative multi-species anatomy ontology. *Genome Biology* 13,R5. http://dx.doi.org/10.1186/gb-2012-13-1-r5

[27] Pafilis, E., Frankild, S. P., Schnetzer, J., Fanini, L., Faulwetter, S., Pavloudi, C., Vasileiadou, A., Leary, P., Hammock, J., Schulz, K., Parr, S., Arvanitidis, C., and Jensen, L. J. (in press). ENVIRONMENTS and EOL: identification of Environment Ontology terms in text and the annotation of the Encyclopedia of Life. *Bioinformatics*.

[28] Park, C. A., Bello, S. M., Smith, C. L., Hu, Z.-L., Munzenmaier, D. H., Nigam, R., Smith, J. R., Shimoyama, M., Eppig, J. T. & Reecy, J. M. (2013). The Vertebrate Trait Ontology: a controlled vocabulary for the annotation of trait data across species. *Journal of Biomedical Semantics*, *4*(1),13. http://dx.doi.org/10.1186/2041-1480-4-13

[29] Parr, C. S. and McClain, C. R. (2014) EOL-BHL-NESCent Research Sprint Report. *PeerJ PrePrints* 2:e503v1 http://dx.doi.org/10.7287/peerj.preprints.503v1

[30] Parr, C., Sachs, J., Parafiynyk, A., Wang, T., Espinosa, R., & Finin, T. (2006). *ETHAN: the Evolutionary Trees and Natural History Ontology*. University of Maryland, Baltimore County.

[31] Parr, C. S., Wilson, N., Leary, P., Schulz, K. S., Lans, K., Walley, L., Hammock, J. A., Goddard, A., Rice J., Studer, M., Holmes, J. T. G., and Corrigan Jr., R. J. 2014. The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth. *Biodiversity Data Journal* 2: e1079, http://dx.doi.org/10.3897/BDJ.2.e1079.

[32] Patterson, D. J., Faulwetter, S. and Shipunov, A. (2008) Principles for a names-based cyberinfrastructure to serve all of biology. *Zootaxa* 1950,153–163.

[33] Payne, P. R. O. (2012). Chapter 1: Biomedical knowledge integration. *PLoS Computational Biology*, *8*(12), e1002826. http://dx.doi.org/10.1371/journal.pcbi.1002826

[34] Pereira, H. M., Ferrier, S., Walters, M., Geller, G. N., Jongman, R. H. G., Scholes, R. J., Bruford, M. W., Brummitt, N., Butchart, S. H. M., Cardoso, A. C., Coops, N. C., Dulloo, E., Faith, D. P., Freyhof, J., Gregory, R. D., Heip, C., Höft, R., Hurtt, G., Jetz, W., Karp, D. S., McGeoch, M. A., Obura, D., Onoda, Y., Pettorelli, N., Reyers, B., Sayre, R., Scharlemann, J. P. W., Stuart, S. N., Turak, E., Walpole, M., & Wegmann, M. (2013). Essential Biodiversity Variables. *Science* , *339* (6117), 277–278. http://dx.doi.org/10.1126/science.1229931

[35] Poelen, J. H., Simons, J. D. and Mungall, C. J. (2014). Global Biotic Interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecological Informatics*, 24,148–159 http://dx.doi.org/10.1016/j.ecoinf.2014.08.005

[36] Quintero E, Thessen AE, Arias-Caballero P, Ayala-Orozco B. (2014) A statistical assessment of population trends for data deficient Mexican amphibians. *PrePrints* 2:e703 http://dx.doi.org/10.7717/peerj.703

[37] Roskov Y., Kunze T., Orrell T., Abucay L., Paglinawan L., Culham A., Bailly N., Kirk P., Bourgoin T., Baillargeon G., Decock W., De Wever A., Didžiulis V., eds. (2013). Species 2000 & ITIS Catalogue of Life, *2013 Annual Checklist*. Digital resource at www.catalogueoflife.org/annual-checklist/13. Species 2000: Naturalis, Leiden, the Netherlands.

[38] Rotman, D., Procita, K., Hansen, D., Parr, C. S., & Preece, J. (2012). Supporting Content Curation Communities : The Case of the Encyclopedia of Life. *Journal of the American Society for Information Science and Technology*, 63(6), 1–29. http://dx.doi.org/10.1002/asi.22633

[39] Thessen, A. E., & Parr, C. S. (2014). Knowledge Extraction and Semantic Annotation of Text from the Encyclopedia of Life. *PLoS ONE*, 9(3), e89550. http://dx.doi.org/10.1371/journal.pone.0089550

[40] Vision, T. (2010). The Dryad Digital Repository: Published evolutionary data as a part of the greater data ecosystem. *Nature Precedings*, (713),1–1. http://dx.doi.org/10.1038/npre.2010.4595.1

[41] Urbani, J. (2013). Three Laws Learned from Web-scale Reasoning. In *2013 AAAI Fall Symposium Series*.

[42] Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertsib, T., & Vieglais, D. (2012). Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE*, 7(1), e29715. http://dx.doi.org/10.1371/journal.pone.0029715

[43] Wright, C., & Seltmann, K. (2014). Usage patterns of blue flower color representation by Encyclopedia of Life content providers. *Biodiversity Data Journal* 2: e1143. http://dx.doi.org/10.3897/BDJ.2.e1143