Name-based Approach to Build a Hub for Biodiversity LOD

Yoshitaka Minami[a], Hideaki Takeda[a*], Fumihiro Kato[a], Ikki Ohmukai[a], Noriko Arai[a], Utsugi Jinbo[b], Shoko Kawamoto[c], Satoshi Kobayashi[d], and Motomi Ito[e]
[a] *National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan*
[b]*National Museum of Nature and Science, 7-20, Ueno-Koen, Taito-ku, Tokyo, Japan*
[c]*Database Center for Life Science, Research Organization of Information and Systems, 2-11-16, Yayoi, Bunkyo-ku, Tokyo, Japan*
[d]*Transdisciplinary Research Integration Center, National Institute of Polar Research, 10-3, Midori-cho, Tachikawa-shi, Tokyo, Japan*
[e]*Department of General Systems Studies, Graduate School of Arts and Sciences, The University of Tokyo, 3-8-1, Komaba, Meguro-ku, Tokyo, Japan*

Abstract. Because of a huge variety of biological studies focused on different targets, i.e., from molecules to ecosystem, data produced and used in each field is also managed independently so that it is difficult to know the relationship among them. We aim to build a data hub with LOD to connect data in different biological fields to enhance search and use of data across the fields. We build a prototype data hub on taxonomic information on species, which is a key to retrieve data and link to databases in different fields. The core of this hub is the dataset for species and taxa. We adopted the database called "Building Dictionary for Life Science (BDLS)" that contains relationship between scientific names and common Japanese names. Based on this dataset, we integrate various datasets such as domain-specific taxonomies and specimen databases.

Keywords: Linked Open Data (LOD), biodiversity, taxonomy, data integration

1. Introduction

Biodiversity [1] becomes a big scientific and social problem according to global awareness to the environmental problems. Biodiversity is related to many research fields, in particular biology, but biology itself consists of many research fields focused on various targets, from molecules to ecosystem. Thus, there are many biological disciplines, from molecular and cell biology, to ecology, evolution and taxonomy. Data collected from biological studies are highly diverse in contents and formats. Each research field can yield and use data for own field, but such data often lacks information on relationships to one another. As collaborative research projects across different fields are developing, demands for a data coordination system is increasing.

When focusing on data, diversity can be categorized in the following three ways. Firstly, there is diversity in subjects of biological researches. There are different fields depending on hierarchical level of focus ranging from molecular biology to ecology, and analysis using multi-scale data is often required. There are large databases for molecular data (e.g. DDBJ, NCBI) and for specimen and observation data (GBIF), but their relationships are rather weak. On the other hand, some specialists of specific groups of organisms build their own specific databases. Such databases contain valuable data but are often independent from each other.

Secondly, representation in local languages is needed for wide range of people, for example, governmental people working on biological resource management or biodiversity conservation in individual countries.

* Corresponding author. E-mail: takeda@nii.ac.jp.

Thirdly, there is diversity by people. In addition to researchers and governmental people, general people are also looking for biological data for their activities. Recently, Citizen Science programs, namely, researches and studies in collaboration with general people, are emerging and biology is in its forefront.

We aim to build a data hub to connect data collected in various biological fields to enhance researchers to search and use data across fields. Therefore, the information infrastructure for biodiversity should be required to treat heterogeneous, multi-scale and multilingual data in scattered databases and to provide them for various people. We aim to build a data hub to absorb the above diversity.

2. Related work

In biodiversity informatics [2], ensuring interoperability of the various databases specialized in individual purpose, is one of the most important issues [3]. Several researchers and groups have started to research about Linked Data in biodiversity information but they have not reached standard or consensus fully [4]. Peterson et al. [5] emphasized that the integration of scientific names using linked data approach has a big potential and enables to create rich services that biologist can benefit. Darwin Core [6] is a well-known standard for metadata for biodiversity information, but Linked Data for Darwin Core is under discussion. TaxonConcept[1] provides ontology and data for species but data is limited to the specific geographical area.

3. The basic policies for integration

In order to fulfill the requirements mentioned in Section 1, we set up two basic policies. The first one is that we focus on taxonomic information on species since they are common and mandatory fields for most biological information. It provides very basic information on each species including classification, and scientific and general (English and Japanese) names. It also provides links to entries on other database such as NCBI (National Center for Biotechnology Information), EOL (Encyclopedia of Life), and DBpedia based on scientific name.

The second is that we treat names as the first class entities. A taxon can be represented as a set of names that are linked to each other. It is a different approach in other studies like TaxonConcept and Darwin Core [6] where taxa are treated as the first class entities. There are two main benefits for our approach. The first is that it is easy to build and maintain the database since identification of the authorized names can be postponed when building the database. The second is that linking to other databases is relatively easy including non research-based databases like those maintained by citizen since variation of names is naturally included. There are also the drawbacks. It is not easy to provide authorized names since it needs some processing on the network. The other is that homonymies, i.e., some names are used for different taxa.

4. The core dataset: BDLS (Building Dictionary for Life Science)

We selected the dataset called BDLS (Building Dictionary for Life Science) for a core dataset for taxa. BDLS[2] is an integrated dictionary for biology which is built from nearly 100 sources which include various illustrated books and specimen dataset in museums, the latter of that are provided by the Science Museum Net[3] in Japan.

There are two main parts, i.e., one is the dictionary for taxa (mainly species) and the other is the dictionary for terminology. We used the former mainly.

It contains scientific names, Japanese common names, and common names in other languages (mainly English) for taxa and their relations (relation between scientific names and common names). Every relation is annotated with provenance, i.e., the source like a name of book or database.

It contains 55,759 scientific names and 57,929 Japanese common names. Among them, there are 55,245 relations between scientific and Japanese common names. It is probably the largest dictionary for Japanese common names for taxa. The analysis

[1] http://www.taxonconcept.org

[2] It is developed by Database Center for Life Science (DBCLS), Research Organization of Information and Systems (ROIS), Japan and available from http://lifesciencedb.jp/bdls/

[3] The network is maintained by the National Museum of Nature and Science, Tokyo. http://science-net.kahaku.go.jp/

about correspondence with other databases like Species2000 and NCBI is available[4].

5. Other databases

There exist various datasets even just concerning species information. Among them we selected mainly three datasets to examine feasibility of integration as a necessary condition to publish and to contain species data. We used the following two databases to test integration.

1. A domain-specific dataset for taxon names: the Current Checklist of Japanese Butterflies

We selected the Current Checklist of Japanese Butterflies [7] as a domain-specific dataset. It is a checklist (a list of species names) created and authorized by the butterfly taxonomists. It covers all butterfly species which number is 327 found in Japan, and describes each species by scientific and Japanese general names and higher taxa.

2. A domain-specific dataset for specimens: Bryophytes Specimen Collection

We selected the Bryophytes Specimen Collection that National Institute of Polar Research developed and maintained as another domain-specific dataset. This data has 56,590 specimen data.

6. The data model for species information

These dataset that we have selected have common information about scientific name and taxon. Then, we made a data model shown in Fig.1 in order to link to entries among the databases through the common information and to be used from outside conveniently. This data model was expressed in Named Graph for the data sources, i.e., the sub datasets in BDLS, Butterflies and Bryophytes.

One of the big issues in species information is treatment on various names. Each species has its scientific name but can be represented differently. One case is caused by different citation forms (e.g., use of abbreviation for genus, omission of authors). The other case is derived from multiple names for one species. The valid species name and the combination of genus and species might be changed as taxonomic studies proceed. Furthermore a species may have multiple general names in local languages. It is a delicate problem in taxonomy to choose a unique valid name for each species and it is beyond our scope[5]. Rather we represent each name as a node and associate nodes by relationship such as synonyms.

A basic model is as follows. We provide a class for taxon name and classes for scientific name and common name as its subclass. All nodes on names are instances of these classes. A node can have a "hasTaxonRank" property of which value is an instance of Class "TaxonRank", i.e., either kingdom, phylum, class, order, family, subfamily, tribe, subtribe, genus, subgenus or species. Another type is a node for specimen which is an instance of Class "Specimen" providing specific properties on specimen. All triples on instances are associated to data source URIs in Named Graph.

The benefit of name-based approach is rapid integration of data from different data sources. Its drawback is complexness of representation since a single specimen is represented as a network but it can be compensated by inference in RDF.

Representation of species name common to three datasets is defined as follows. First, nodes of the ScientificName type and CommonName type are generated for species name and for common name respectively. Next, the hasCommonName property links a node of the ScientificName to a node of the CommonName, and the hasScientificName property vice versa. A node representing taxon is linked to a node of the TaxonRank type by the hasTaxonRank property. And, other items are literal.

Nodes representing specimen in Bryophytes dataset describe ID, collected date, collector, latitude, longitude, floral region, floral subsection, locality, sporophyte, altitude, determiner, and herbarium housed[6]. Nodes representing specimen are defined as follows. First, the node is generated assigning a URI to each specimen because there can multiple specimens for one species. Next, a specimen node is linked to the node of ScientificName by the species property. A specimen node has a link by the

[4] http://lifesciencedb.jp/bdls/8abf67fb72b66cc153ca7d50679c

[5] The exception in our dataset is Butterflies dataset where a list of valid scientific and common names is authorized by taxonomic experts.

[6] The list of herbarium index is available from Index Herbariorum: (http://sciweb.nybg.org/science2/IndexHerbariorum.asp)

crm:has_current_location property to a node representing a facility to represent herbarium housed relationship. And, other Items are literal.
Source information in BDLS dataset is represented by name in Named Graph. RDF triples representing data in a data source has a name which has a link by the dcterms:source property to the node representing the data source. Then it has properties such as rdfs:label and dcterms:publisher.

7. The Results

In accordance with the data model, we generated LOD from the selected data. As a result, the number of taxon name is 443,248, scientific name of species is 226,141, common name of species is 219,865, hasScientificName property node is 87,160 and hasCommonName property node is 84,610. The numbers of names become roughly four times larger than those in BDLS due to introduction of other databases. But the number of the relation between scientific and common names does not increase so much. It indicates that many of the relations between them are left to be added.

Our approach is successful in integration basically. But it causes some problems when the dataset is used. For example, we implemented a simple taxon search interface by name. We can show the results by matching names but we can just a set of taxon names but not show representative names. One of the potential problems of our approach is homonymy, i.e., two taxa may share a name. Though the naming rule of scientific name is not essentially permitted a name sharing two or more taxa, there are some exceptional cases e.g., an animal and a plant can share one genus name. We checked how it is in the real dataset by using the NCBI taxonomy. We found 1797 homonymical names. By using this data, we can distinguish taxon names properly.

8. Publishing data as Linked Data

We created the model for species information and translated the data into Linked Data according to the model.

The whole data is available from SPARQL Endpoint[7] and a simple search interface[8]. It also includes the out-going links to LOD such as DBpedia (en) and to non-LOD data sites such as NCBI taxonomy and Encyclopedia of Life. But since licensing for some of the original datasets is not clear as open, the whole dataset itself has not been registered to the Data Hub yet. We are currently working towards open license for them.

BDLS dataset is clearly open with CC-BY-SA license. It is registered in the Data Hub[9] and available as the dumped data and the SPARQL Endpoint[10]. It is alone valuable as the database for species with scientific names and Japanese common names, which can work a hub of biodiversity information by interlinking various datasets by not only scientists but ordinary people.

9. Conclusion

We described the concept of the data bub for species based on names and the prototype system with translating the datasets into Linked Data. Name-based approach is well suited to Linked Data since different names for species which may appear in different datasets can be linked to each other. We publish the integrated dataset as a prototype and also the core dataset as LOD to prompt integration with more datasets.

References

[1] Bisby, F.A.: The quiet revolution: biodiversity informatics and the Internet, Science, Vol. 289, No. 5488, pp. 2309-2312, (2000) Edwards, J.L., Meredith A.L., and Nielsen E.S.: Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop. Science, Vol. 289, No. 5488, pp. 2312-2314, (2000)

[2] Jenkins M (2003) Prospects in biodiversity. Science 302: 1175-1177.

[3] Global Biodiversity Information Facility (2011) Recommendations for the Use of Knowledge Organisation Systems by GBIF. accessible online at http://links.gbif.org/gbif_kos_whitepaper_v1.pdf

[4] Patterson DJ, Cooper J, Kirk PM, Pyle RL and Remsen DP (2010) Names are key to the big new

[7] http://lod.ac/species/sparql

[8] Access from http://lod.ac/ by selecting SPECIES as dataset.

[9] http://datahub.io/dataset/lodac-bdls

[10] http://lod.ac/wiki/LODAC_BDLS

biology. Trends in Ecology and Evolution 25: 686-691. doi:10.1016/j.tree.2010.09.004

[5] Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, et al. (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7(1): e29715. doi:10.1371/journal.pone.0029715

[6] Inomata T, Uémura Y, Yago M, Jinbo U and Ueda K (2010) The Current Checklist of Japanese Butterflies. http://binran.lepimages.jp/

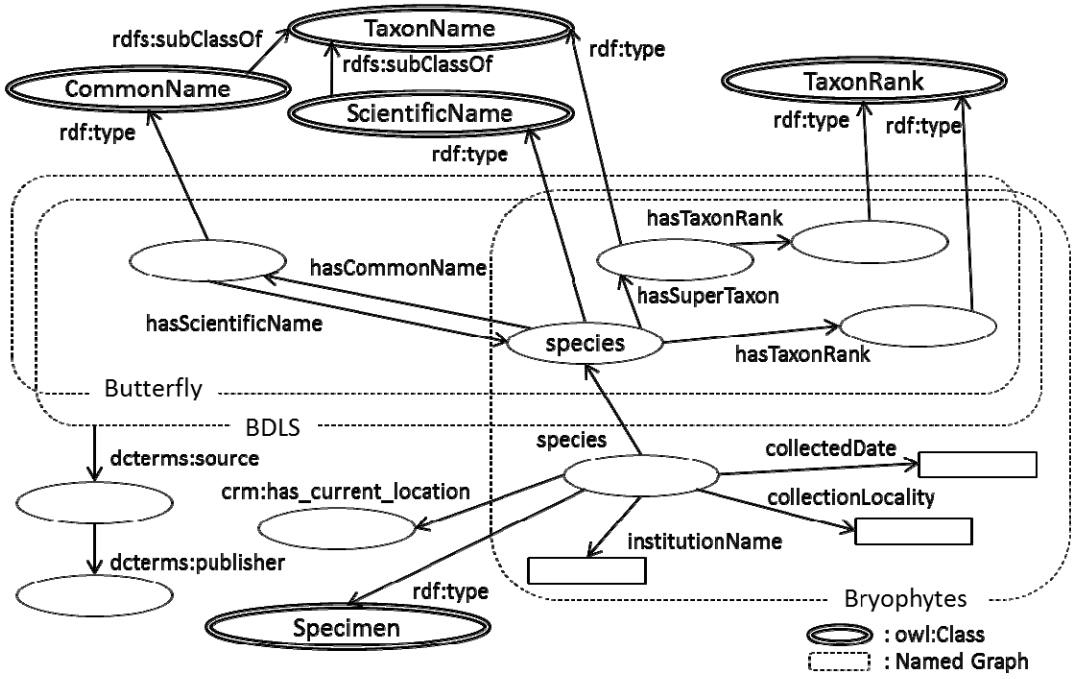

Fig. 1. Data model