A biodiversity terminological inventory

1. BASIC INFORMATION

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),

The inventory is organised as a set of "source terms", which are both scientific and common names of species (about 288,000 names). Each source term is accompanied a set of semantically-related terms. These related terms have been automatically extracted using text mining methods from the English pages of the Biodiversity Heritage Library (BHL)[1], which are about 26 million pages, dated back from the sixteenth century.

Each related term can be:

- **synonyms of the source term, i.e., scientific names or common names denote the same taxon.** For example, "*Iridoprocne bicolor"* and "*Tachycineta bicolor"* are identified as terms related to "tree swallow" since they are scientific names of "tree swallow".
- **species that share habitat, taxonomic class, or geographic location with the source species name**. For example, since "Arctic fox" and "Polar bear" share the same geographic locations of Alaska, Greenland and Spitzbergen, we consider them as semantically-related species names.

The term inventory is helpful in automatic query expansion. Specifically, the inventory can be useful in terms of increasing the number of relevant documents retrieved by a search engine. In this way, the inventory facilitates the suggestion of terms that are semantically related to the species name supplied as a query, which are in turn appended to the query. For instance, when a user searches for relevant documents of "*Aquila chrysaetos*", the search engine can return X hits. By integrated with the term inventory, the advanced search engine can suggest "golden eagle" as a related name of "*Aquila chrysaetos*". Therefore, when a user selected "golden eagle", the search engine will automatically perform query expansion and additionally return Y relevant documents to "golden eagle". Thus, the user will receive X+Y relevant documents in total just by one click on the suggested term.

Another potential application of the inventory could be taxonomy curation, in which existing taxonomies are populated in a semi-automatic manner, i.e., with a user-in-the-loop manually validating automatically suggested names.

1.2 Representation of the lexicon (flat files, database, markup)

The terminological inventory is publicly available in the JSON format--a lightweight data-interchange format.

[1] www.biodiversitylibrary.org

1.3 Character encoding

The characters have been encoded in UTF8.

2. ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Sophia Ananiadou
Address: Manchester Institute of Biotechnology,131 Princess Street, Manchester M1 7DN, UK
Affiliation: National Centre for Text Mining, School of Computer Science, University of Manchester
Position: Professor
Telephone: +44 161 306 3092
Fax: +44 161 306 5201
e-mail: Sophia.Ananiadou@manchester.ac.uk

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource is available as an archive from the META-SHARE platform.

2.3 Copyright statement and information on IPR

3. TECHNICAL INFORMATION

3.1 Directories and files

The inventory is presented in one JSON file named bhl_full_inventory.json.

3.2 Data structure of an entry

The inventory is organised as a set of source terms. Each source term corresponds to a species name. The list of used source terms was obtained by collecting all available species names from the Catalogue of Life[2], Encyclopedia of Life[3], and the Global Biodiversity Information Facility[4]. The complete archives were then further processed to determine the various textual contexts in which each of the source terms can appear; the 20 terms which appear in the most similar contexts to the source term were extracted as related terms and accompany each source term. The source term is provided with additional information including URI, UUID and LSID, generated by the

[2] www.catalogueoflife.org/
[3] www.eol.org
[4] www.gbif.org

Global Name Index (GNI)[5] so that further applications can easily link the term to other resources. The frequency of each source term in BHL is also included in each entry. For each related term, in addition to the same information as the source term, the similarity score of the source and the related terms is given. The structure of each entry is detailed in Table 1.

Table 1. The structure of each entry in the inventory

Name	Type	Value	Description
sourceTerm	Pair	string	The source species name
GN_URI	Pair	string	A URI of the source species name assgined by GNI
GN_uuid_hex	Pair	string	A UUID in hexa value of the source species assigned by GNI
GN_lsid	Pair	string	A LSID of the source species name assgined by GNI
frequency	Pair	number	The frequency of the source term in the Biodiversity Library History
relatedTerms	Array	object	An array of 20 topmost semantically related terms of the source species
targetTerm	Pair	string	One related term of the source term
score	Pair	number	The similarity score of the source and the related species.

3.3 Lexicon size (nmb. of lexical items, KB occupied on disk)

The terminological inventory contains about source terms. The approximate size of the file is about 1.4GB without compression.

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

The language of the biodiversity terminological inventory is English.

4. 2 Entry Type

In the final inventory, each source term is accompanied by the following information:

- The URI, UUID and LSID assigned by GNI
- The frequency in BHL
- The set of the 20 terms that are considered to be most closely related to the source term, according to contextual similarity.

[5] http://gni.globalnames.org/

- For each of the 20 most related terms, URI, UUID and LSID are provided. The term is also accompanied by a numerical score that represents the degree of similarity between the contexts of the related term and the contexts of the source term. This score is called the cosine similarity.

An example of one entry in the inventory is presented in Table 2.

Table 2. An example of one entry in the inventory


```
[{"sourceTerm":"Phidippus auctus",
"GN_URI":"http://gni.globalnames.org/name_strings/4190896.xml",
"GN_uuid_hex":"46d3d5e9-7a11-5863-b2ff-28c93996f357",
"GN_lsid":"urn:lsid:globalnames.org:index:46d3d5e9-7a11-5863-b2ff-28c93996f357",
"frequency":5,
"relatedTerms":[
        {"targetTerm":"Toxeuma fuscicorne",
        "GN_URI":"http://gni.globalnames.org/name_strings/9095648.xml",
        "GN_uuid_hex":"02a1d2ee-2c20-5346-aec5-1722457f37c6",
        "GN_lsid":"urn:lsid:globalnames.org:index:02a1d2ee-2c20-5346-aec5-1722457f37c6",
        "score":0.7135},
        ...
        ]
} ]
```


4.3 Coverage of the lexicon

The inventory was created automatically by applying GloVe (Pennington et al., 2014)--a distributional semantic model, to a set of about 26 million English pages of BHL. 288,562 species names that appear in BHL documents with a frequency of at least five were selected to be included in the inventory.

4.5 Intended application of the lexicon

The term inventory can be useful in terms of increasing the number of relevant documents retrieved by a search engine. In this way, the term inventory facilitates the suggestion of terms that are semantically related to the species name supplied as a query, which are in turn appended to the query. As opposed to other automatic query expansion methods (Carpineto et al., 2012), the inclusion of additional terms to expand a query depends upon the user's preferences, making a search interface's behaviour similar to that of a recommender system (Bobadilla et al., 2013). Furthermore, with the integrated term inventory, the interface can show users not only documents matching the original query but also those that mention semantically relevant species which have, for example, shared characteristics such as habitat or taxonomic classification. For instance, as a user searches for documents pertaining to "lion", it might be useful for him or her to find

documents mentioning not only “lion” but also *Panthera leo*---the scientific name of “lion”---and “jaguar”---a big cat similarly belonging to the *Panthera* genus. In contrast, without the term inventory, the search results will include only documents that match exactly the same species, e.g., “lion” and *Panthera leo*.

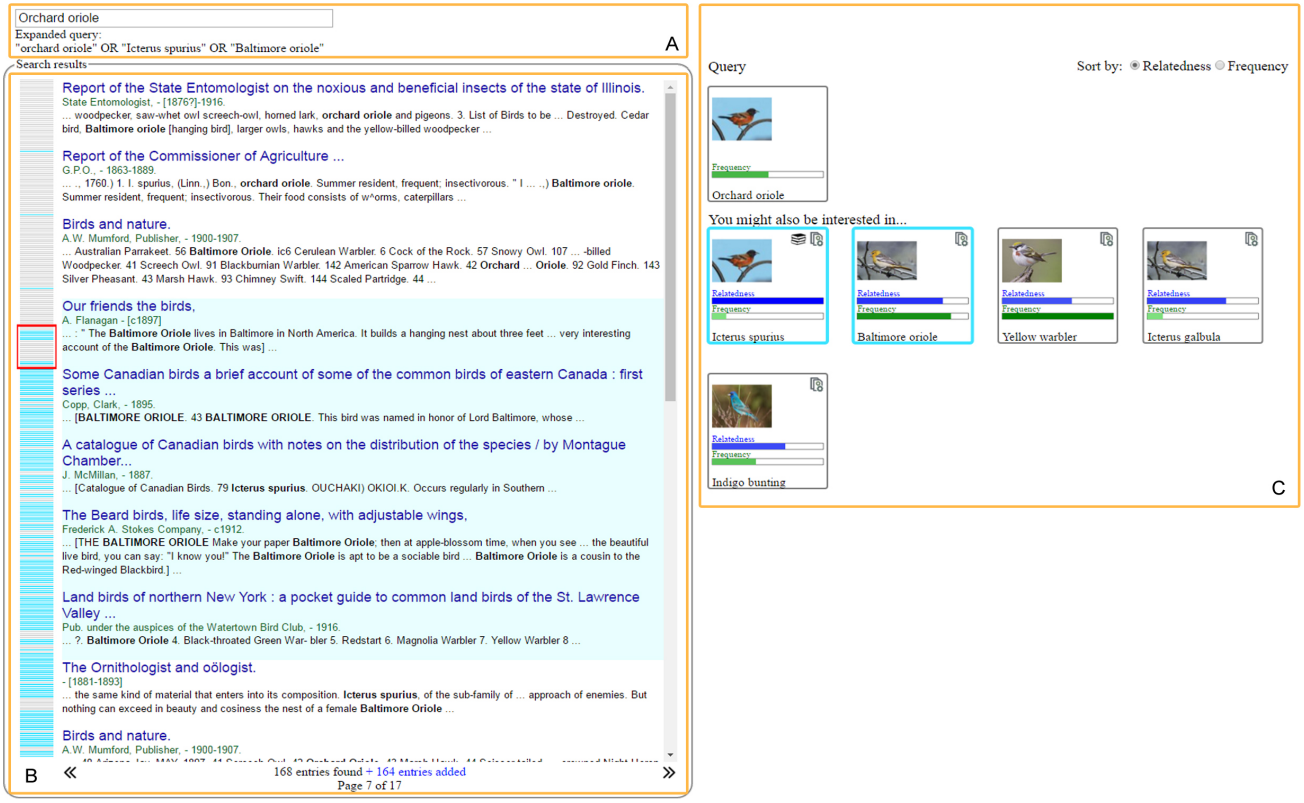

Figure 1. **Visual search interface incorporating suggested semantically related names for query expansion**.} **A-** Initial and expanded query. **B-** Search result list and context viewer. The context viewer on the left-hand side shows a zoomed out view of the retrieved list. Documents retrieved according to the expanded query are shown with a light blue background. **C**- Thumbnails with suggested names for query expansion. Apart from a relevant image, each thumbnail depicts the suggested term’s frequency within BHL documents, its relatedness to the query term, and the provenance of the suggestion, i.e., our term inventory or other external resources such as CoL, EoL and GBIF.

The web-based interface[6] is shown in Figure 1, together with a description of its main components. On the left-hand side is a context viewer (inspired by the overview+detail approach (Cockburn et al., 2009) that provides an overview of the results, indicating the portion of the ranked list that the user is currently looking at. Unlike regular search interfaces, terms semantically related to the query species name are also presented on the right-hand side of the screen as thumbnails. We aimed to present the suggested terms in a simple yet informative manner. To this end, each thumbnail is displayed with two parallel indicator bars, one representing the term’s frequency in the BHL corpus and the other its relatedness to the query term. This eliminates the need to present their exact values to the user, facilitating a more intuitive visual comparison of these measures. Within the thumbnails, there are also small icons that

[6] http://nactem.ac.uk/BHLQueryExpansion/

indicate whether the name exists as a semantically related term in our term inventory, in an external resource (e.g., CoL, EoL, and GBIF) or in both. For example, in Figure 1, there are five species names returned by our interface (under the heading ``You might also be interested in…") for the query *Orchard oriole*, the first of which was suggested by both our term inventory and external resources, and the remaining four recommended by just our term inventory (as indicated by the icons at the upper right-hand corner of each thumbnail). Hovering over a thumbnail reveals detailed information about the suggested term. Images of the species retrieved from EoL (via their web services[7]) are shown for reference purposes.

4.6 Reliability (automatically/manually constructed)

The inventory is constructed using fully automatic methods. In order to evaluate the methods, we compared the inventory with four taxonomies, i.e., the Catalogue of Life, the BirdLife Taxonomic Checklist, the Interagency Taxonomic Information System (ITIS) and the PLANTS Database, for the compilation of reference standard data sets containing semantic variants. In this work, for every preferred name *n*---the binomial name that is most commonly used to denote a certain species *s*---we define *semantic variants* as the set of names containing: (1) scientific names synonymous to *n*, and (2) vernacular names that denote *s*. In BirdLife, for instance, the scientific name *Actitis macularius* has two strict-sense synonyms, i.e., *Actitis macularia* and *Tringa macularia*, and one vernacular name, i.e., "spotted sandpiper"'. *Actitis macularius* is the preferred name, and *Actitis macularia*, *Tringa macularia* and "spotted sandpiper"' are its corresponding semantic variants.

We then selected preferred names that appear in both the BHL corpus and the reference standards. We retained only those that have at least one synonym or vernacular name according to the reference standards, and which appear in the BHL corpus frequently enough, i.e., at least 50 times. We finally randomly selected 500 preferred names for each of the bird, mammal and plant categories. The highest mean average precisions of 20 topmost candidates for each category were 77.11%, 73.30%, and 62.14%, respectively.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Pennington J, Socher R, Manning CD. **GloVe: Global Vectors for Word Representation**. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014); 2014. p. 1532--1543.

Carpineto C, Romano G. **A survey of automatic query expansion in information retrieval**. *ACM Computing Surveys (CSUR)*. 2012;44(1):1.

Bobadilla J, Ortega F, Hernando A, Gutiérrez A. **Recommender systems survey**. *Knowledge-Based Systems*. 2013;46:109--132.

[7] http://eol.org/api

Cockburn A, Karlson A, Bederson BB. **A review of overview+ detail, zooming, and focus+ context interfaces**. *ACM Computing Surveys (CSUR)*. 2009;41(1):2.

The inventory is introduced in the following article:

Nguyen, N. T. H., Soto, A. J., Kontonatsios, G., Batista-Navarro, R. and Ananiadou, S.. (In Press). **Constructing a Biodiversity Terminological Inventory**. *PLOS ONE*. DOI: 10.1371/journal.pone.0175277