Literature-driven Curation for Taxonomic Name Databases

Hui Yang, Alistair Willis, David R. Morse, Anne de Roeck
Department of Computing and Communications
{Hui.Yang, Alistair.Willis, David.Morse,
Anne.deRoeck}@open.ac.uk

Abstract

Digitized biodiversity literature provides a wealth of content for using biodiversity knowledge by machines. However, identifying taxonomic names and the associated semantic metadata is a difficult and labour intensive process. We present a system to support human assisted creation of semantic metadata. Information extraction techniques automatically identify taxonomic names from scanned documents. They are then presented to users for manual correction or verification. The tools that support the curation process include taxonomic name identification and mapping, and community-driven taxonomic name verification. Our research shows the potential for these information extraction techniques to support research and curation in disciplines dependent upon scanned documents.

1 Introduction

Our understanding of the natural world is rapidly increasing. At the same time, issues in biodiversity are shown to be relevant to many important policy areas, such as climate change, food security and habitat management. Biological taxonomy is a discipline that underlies all of these areas; understanding species, their behaviours and how they interact is of critical importance in being able to manage commercially important land and environment use (SCBD, 2008).

A major difficulty facing the curation of comprehensive taxonomic databases is incorporating the knowledge that is currently contained only in the printed literature, which spans well over one hundred million pages. Much of the literature, especially old taxonomic monographs that are both rare documents and extremely valuable for taxonomic research, are almost entirely in paper-print form and are not directly accessible electronically. Recent large-scale digitization projects like the Biodiversity Heritage Library [1] (BHL) have worked to digitize the (out of copyright) biodiversity literature held in natural history museums and other libraries' collections. However, due to the lack of semantic metadata, the tasks of finding, extracting, and managing the knowledge contained in these volumes is still a primarily manual process and remains extremely difficult and labour-intensive. The difficulty in accessing the existing taxonomic literature is a severe impediment to research and delivery of the subject's benefits (Godfray, 2002). Semantic tagging of organism mentions in biodiversity literature has recently been regarded as a pivotal step to facilitate taxonomy-aware text mining applications, including species-specific document retrieval (Sarka, 2007), linking biodiversity databases (White, 2007), and semantic enrichment of biodiversity articles (Penev et al, 2010).

Semantic web is a potential solution to the problems of data fragmentation and knowledge management if the appropriate metadata can be created (Page, 2006). However, manually creating this metadata is an enormous and unrealistic task. The verification process of checking the validity of a taxonomic name is a specialist task requiring expert skills.

In this paper, we present a semi-automated system that aims to develop a literature-driven curation process among practicing taxonomists, by providing tools to help taxonomists identify and validate appropriate taxonomic names from the scanned historical literature. Potential taxonomic names are automatically extracted from scanned biodiversity documents with their associated contextual information. These are presented for validation to taxonomic curators via an online web service. The manually verified or corrected names can then be indexed, and the semantic data stored, using the Darwin Core biodiversity data standard.

2 System Framework

Figure 1 shows the process for obtaining metadata and curating taxonomic names. Publishers who specialise in biological taxonomy often add appropriate metadata (Penev et al. 2010), but for scanned literature, this is generally not available.

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

Proceedings of the Joint Workshop on NLP&LOD and SWAIE, pages 25–32,
Hissar, Bulgaria, 12 September 2013.

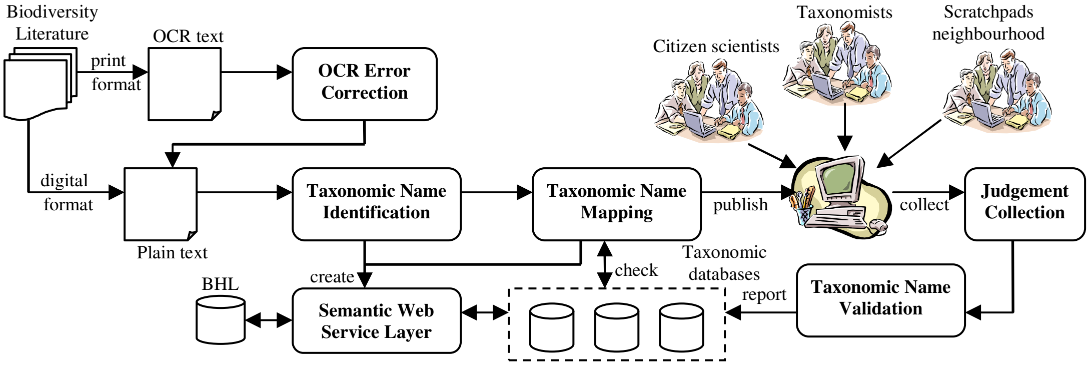


Figure 1. System framework of literature-driven curation for taxonomic databases

The image files from scanned literature are processed through the ABBY FineReader or PrimeReader Optical Character Recognition (OCR) software to generate a plain text file.

Next, we identify those tokens in the plain text that may be taxonomic names (possibly containing errors through imperfect OCR, or other transcription errors). We use an information extraction tool for Named Entity Recognition (NER) based on Conditional Random Fields (CRFs) (Lafferty et al., 2001). The detected names are then mapped onto unique identifiers across a range of taxonomic databases such as uBio Name Bank[2], Encyclopaedia of Life (EoL)[3], and Catalogue of Life (CoL)[4].

Taxonomic names that cannot be found in online databases will be validated manually. Potential unknown taxonomic names are presented for validation or correction to the research community via the Scratchpads social network[5] (e.g., professional taxonomists, experienced citizen scientists and other biodiversity specialists) in a community-driven verification process. The newly verified taxonomic name, along with additional metadata recording the user who verified the name, its context and bibliographic details is published as a semantic web service layer (currently a Scratchpads portal).

3 Taxonomic Name Recognition

Automatic identification of taxonomic names from biodiversity text has attracted increasing research interest over the past few years, but is difficult because of the problems of erroneous transcription and synonymy. There may be orthographic and other term variation in names assigned to the same species (Remsen, 2011). For example, *Actinobacillus actionomy*, *Actinobacillus actionomyce*, and *Actinobacillus actionomycetam* could all be variants of the same name. In addition, scanned documents can cause many OCR errors due to outdated fonts, complex terms, and aspects such as blemishes and stains on the scanned pages. Wei et al (2010) have observed that 35% of taxonomic names in scanned documents contain an error, and this creates difficulties for term recognition (Willis et al. 2009). For example, erroneous OCR might propose '*o*' in place of '*c*' for the taxon *Pioa*, not a known name, rather than *Pica* (European magpie).

Approaches to taxonomic name recognition (TNR) span a broad range from traditional dictionary lookup (Gerner et al., 2010; Koning et al., 2005; Leary et al., 2007) combined with linguistic rule-based (Sautter et al., 2006) to pure machine learning (Akella et al., 2012).

In our system (Figure 1), the first stage is identifying potential taxonomic names. We used a supervised learning algorithm implemented by the CRF++ Package[6]. Compared to other machine learning algorithms, CRFs are good at sequence segmentation labeling tasks such as Named Entity Recognition, which have been shown to be effective for biological entity identification in the biomedical literature (Yang et al. 2008). They can be easily adapted to similar tasks like Taxonomic Name Recognition (TNR).

3.1 Dataset Preparation and Annotation

To assess the performance of the CRFs on taxonomic texts, we generated training and test sets from scanned volumes between 1879 and 1911 from the Biodiversity Heritage Library (BHL).

Annotations were carried out using the BRAT Rapid Annotation Tool (BRAT)[7] (Stenetorp et

[2] http://www.ubio.org/

[3] http://eol.org/

[4] http://www.catalogueoflife.org/

[5] http://scratchpads.eu

[6] http://crfpp.googlecode.com/svn/trunk/doc/index.html

[7] http://brat.nlplab.org/

al., 2012) to mark up taxonomic elements in biodiversity literature. All mentions of taxonomic names in the text were manually tagged and linked to identifiers in external taxonomic databases (i.e. uBio Name Bank, Catalogue of Life, and Encyclopaedia of Life) where possible. Annotated mentions were also assigned to several categories that indicate specific linguistic or semantic features (e.g. taxonomic rank, genus abbreviation or omission) for evaluation analysis. The manually annotated dataset consists of:

(a) **Training data.** We selected three BHL volumes of different animal groups: *Coleoptera* (Beetles)[8], *Aves* (Birds)[9], and *Pisces* (Fish)[10] as the training data to build a CRF-based taxon recogniser. The volume text used for the annotation is *clear text*, i.e. text from which OCR errors are removed, which was obtained from the INOTAXA Project[11]. Table 1 reports the statistical annotation information about these three volumes.

	#Pages	**#Taxonomic Names**
Coleoptera	324	7,264
Aves	553	8,354
Pisces	234	4,915

Table 1. The statistics on the training data

(b) **Test data.** The dataset used for the evaluation of the taxon recogniser is another BHL volume about *Coleoptera* (Beetles)[12]. Taxonomic names are annotated in two datasets of different quality text, one is *clear text* (high-quality text) and the other is the original *OCR text* with scanning errors (poor-quality text). The reason for building this comparative corpus is to estimate the impact of OCR errors on taxon name recognition. The statistics about this corpus are given in Table 2. More taxonomic names are found in the OCR text than the clear text because the OCR text includes page headings that may contain the scientific name of an organism.

	Clear Text	**OCR Text**
#Pages	373	373
#Taxonomic Names	5,198	5,414
#Taxonomic Names (with OCR error)	--	2,335 (43.1%)

Table 2. The statistics on the test data

3.2 Taxonomic Name Identification

To train the CRF-based recogniser, we used a variety of linguistic and semantic features to characterise the semantics of taxonomic names. The features used for taxon recognition were grouped into the following five categories:

- **Word-token Feature.** This type of feature includes word lemma, Part-of-Speech (POS) tag, and chunk tag of the word, which are obtained from the Genia Tagger[13].
- **Context Features.** The features for the lemma and POS tag of the three neighbouring words before and after the current word token are also considered.
- **Orthographic Features.** Taxonomic names tend to be case sensitive, e.g. *Agelaus phaenicio*. Moreover, much taxonomic literature employs abbreviations as standard like *A. phaenicio*. Some special tokens, e.g. Greek symbols (*α*, *β*, *γ*) and Roman numbers (*I.*, *II.*, *iv.*) also frequently occur in the text.
- **Morphologic Features.** Some taxonomic names contain typographic ligatures, e.g., *æ* (*ae*), *œ* (*oe*), *Æ* (*AE*). We observed that some mentions contain the same suffix strings such as *-us*, *-um*, *-eus*.
- **Domain-specific Features.** Taxonomic rank markers and their abbreviations, e.g., *species*, *genus*, *sp.*, *subg.*, *fam.*, etc., frequently occur in the text preceding taxonomic names. This is a binary property. *Y* if the word is a rank marker or *O* otherwise.

The training data file for the CRFs consists of a set of word token instances, each of which contains a feature vector that is made up of five groups of features described above together with an entity class label – BIO tags.

	Precision	Recall	F-measure
Clear text	0.9285	0.8642	0.8952
OCR text	0.4450	0.3716	0.4050

Table 3. The overall performance of taxonomic name identification on a comparative dataset

Performance Evaluation. The trained CRFs were evaluated on the test corpus. We compared the results of the clear text with those of the OCR text in order to test the OCR-error toleration capability of the trained CRFs. As shown in Table 3, the trained CRFs performs well and achieves an F-measure as high as 0.8952 on the clear text

[8] http://www.biodiversitylibrary.org/ia/mobotbca_12_01_01

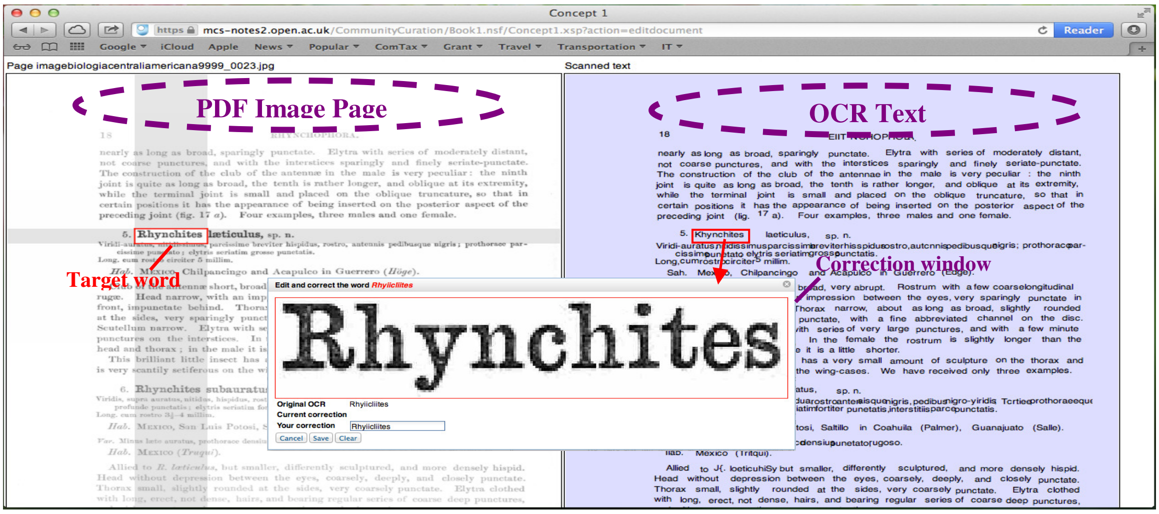
[9] http://www.biodiversitylibrary.org/ia/mobotbca_03_01_00

[10] http://www.biodiversitylibrary.org/ia/mobotbca_05_00_00

[11] http://www.inotaxa.org/jsp/index.jsp

[12] http://www.biodiversitylibrary.org/ia/mobotbca_12_04_03

[13] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/



Page imagebiologiacentraliamericana9999_0023.jpg

18 RHYNCHOPHORA.

nearly as long as broad, sparingly punctate. Elytra with series of moderately distant, not coarse punctures, and with the interstices sparingly and finely seriate-punctate. The construction of the club of the antennæ in the male is very peculiar: the ninth joint is quite as long as broad, the tenth is rather longer, and oblique at its extremity, while the terminal joint is small and placed on the oblique truncature, so that in certain positions it has the appearance of being inserted on the posterior aspect of the preceding joint (fig. 17 *a*). Four examples, three males and one female.

5. Rhynchites lætículus, sp. n.

Scanned text

18 EIIT NCHOPHOLA.

nearly as long as broad, sparingly punctate. Elytra with series of moderately distant, not coarse punctures, and with the interstices sparingly and finely seriate-punctate. The construction of the club of the antennae in the male is very peculiar : the ninth joint is quite as long as broad, the tenth is rather longer, and oblique at its extremity, while the terminal joint is small and placed on the oblique truncature, so that in certain positions it has the appearance of being inserted on the posterior aspect of the preceding joint (lig. 17 a). Four examples, three males and one female.

5. Khynchites laeticulus, sp. n.

Edit and correct the word *Rhyiiclites*

Original OCR Rhyiiclites

Current correction

Your correction Rhyiiclites

Cancel Save Clear

Figure 2. A web service for OCR error correction

(good-quality text). On the "dirty" OCR text the performance is worse and the F-measure drops to 0.405. This shows that OCR errors are a potential threat that greatly affects the effectiveness of taxonomic name identification. Therefore an OCR error correction tool is necessary for searching and processing the OCR-scanned text.

OCR Error Correction. To reduce the impact of OCR errors on the identification of taxonomic names, we developed a mechanism for error checking and correction. Figure 2 shows a screen shot of the web service[14] used to highlight a potential taxon to a user. The left-hand panel is the image of the original page, and the right-hand panel is the corresponding text, which is extracted from the DjVu XML file created by the OCR software. When a word is selected using the navigation content in the right-hand panel, a small error-correction window pops up, and the user is allowed to make possible modifications, based on the enlarged image of the target word appearing in the pop-up window.

3.3 Taxonomic Name Mapping

Taxonomic name mapping or normalization is to map the detected mentions in the text into standardised taxonomic identifiers (Gerner et al. (2010). It aims to generate correct lists of unique identifiers (typically from referent taxonomic databases) for each taxonomic name. There are two potential factors that affect mapping accuracy. First, taxonomic names are not completely stable, and may change due to taxonomic revision. There may be multiple names (synonyms) for the same organism, and the same name may refer to different taxa (homonyms). Moreover, there is lexical and terminological variation among taxonomic names. Second, currently there is not a complete taxonomic database that covers all the organisms in the world so multiple taxonomic databases are needed to complement each other.

To resolve the problem of orthographic and term variations between taxonomic names, we exploited a generic and effective cascaded matching method that consists of two stages:

- **Stage I - Exact Matching:** string matching between original identified mentions and database entries. If a name mention is a known synonym in the curated list of a taxonomic database, the unique identifier of a taxon entry associated with the synonym will be assigned to the mention. It is possible that the mention might be mapped to synonyms of different organisms. In these ambiguous cases, additional information such as the surrounding context of the mention and the attributes of its neighboring mentions are needed to help determine the selection of the most appropriate organism.
- **Stage II - Rule-based approximate matching:** First, a set of transformation rules that capture morphologic features of name variations are generated to produce more potential extended mentions. Second, for each unmatched mention filtered at the first stage, the possible extended candidate names created by the transformation rules (described later) are sent to the taxonomic databases again to find the possible matched synonyms of a known taxon entry.

[14] http://mcs-notes2.open.ac.uk/CommunityCuration/Book1.nsf/Concept1.xsp

Construction of transformation rules. According to the observations on our manually-annotated taxonomic dataset, we roughly group name variations into the four categories below:

(a) **Ligature replacement:** typographic ligatures (e.g., æ, œ, Æ, Œ, etc.) that appear in taxonomic names of the old literature are generally replaced with the corresponding two or more consecutive letters. For example, *Agelæus phœniceus* -> *Agelaeus phoeniceus*, *Dendrœca* -> *Dendroica*

(b) **Latin declension:** the scientific name of an organism is always written in either Latin or Greek. A Latin noun can be described in different declension instances (e.g., First-declension, Second-declension) by changing its suffix substring. For instance, *puellae*, *puellarum*, *puellis*, *puellas*, can be normalized as the same root word *Puella*.

(c) **Parenthesized trinomial names:** some taxonomic names consist of three parts. These are usually represented as a species names with a subgenus name contained within parentheses, e.g., *Corvus (Pica) beecheii*, *Tanagra (Aglaia) diaconus*. However, the parenthesized subgenus name is not used very much, and some taxonomic databases do not contain the information at this low rank level. Therefore, the subgenus name can be ignored when mapping, e.g., *Corvus (Pica) beecheii* -> *Corvus beecheii*

(d) **Taxon variety names:** taxon variety names are another special case, which can appear in various name forms like *Peucæa æstivalis arizonæ*; *Peucæa æstivalis var. arizonæ*; *Peucæa æstivalis, var arizonæ*; *Peucæa æstivalis, β. Arizonæ*; even *Peucæa arizonæ* due to taxonomic inflation in which known subspecies are raised to species as a result in a change in species concept (Isaac et al. 2004).

A set of linguistic rules are expressed as regular expressions to record the syntactic and semantic clues found in the name variations discussed above. These rules are used to transform the original mentions to possible extended candidates for string matching in taxonomic databases.

External taxonomic databases. To link more identified mentions to existing external taxonomic databases, we chose three widely-used large-scale taxonomic databases: uBio Name Bank, Encyclopaedia of Life (EoL), and Catalogue of Life (CoL), which separately curate 4.8 million, 1.3 million, and 1.6 million taxonomic names respectively. In each database, each species has exactly one entry with a unique identifier, a name classified as *scientific name* (i.e. the "correct" canonical name), as well as other possible variants (e.g., synonyms, common misspellings, or retired names if the organism has been reclassified). Moreover, these three databases provide relevant web services to users for online search of taxonomic names. For each candidate name, we send a name query to different databases, and automatically extract the relevant unique identifier from the returned result.

Mapping Results. We collected a total of 8,687 distinct candidate names from four annotated BHL volumes and mapped them to the chosen taxonomic databases. Table 4 shows the statistical information of name matches in the individual databases. It is interesting to note that the matched names in EoL usually can be found in uBio, whereas CoL can find some names that do not appear in either uBio or EoL. Nearly a half of the names (4, 273 names) could not be found in any of the taxonomic databases. This suggests that machine-learning based TNR can find quite a lot of new names that a simple dictionary approach cannot identify. Moreover, biodiversity literature is a potentially useful resource to enrich the existing taxonomic databases.

	uBio	EoL	CoL
Mapped Names (total names: 8,687)	3,565 (41.1%)	2,893 (33.3%)	3,354 (38.6%)

Table 4. Name mapping in taxonomic databases

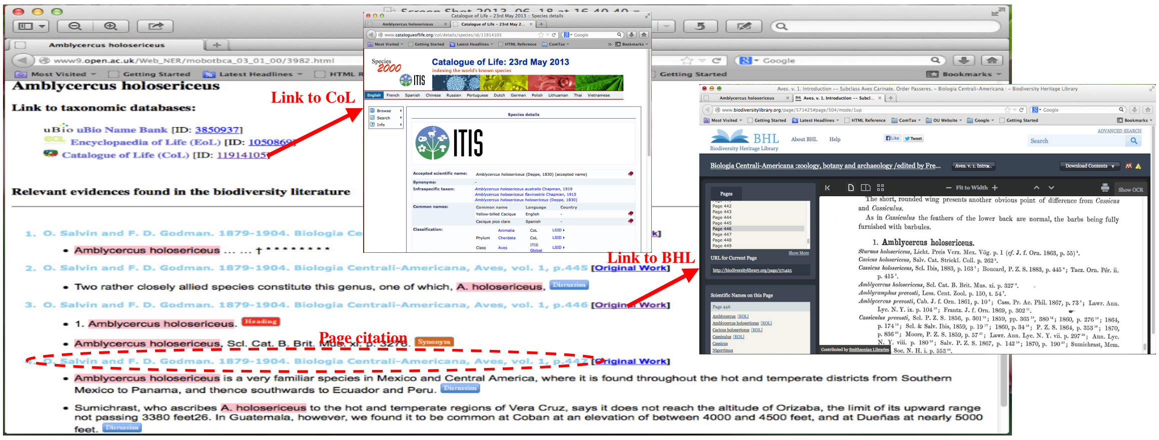
4 Community Metadata Collection

Biodiversity communities have come to the consensus that converting unstructured biodiversity literature into semantically-enabled, machine-readable structured data is essential to use the currently highly fragmented data sources. The main semantic metadata system is the Darwin Core biodiversity data standard[15], maintained by the Biodiversity Information Standards group (TDWG)[16], and based on Dublin Core. The main objects in Darwin Core represent an organism's scientific name, information pertaining to its classification, and the geographical and geological contexts of the organism.

For this research, key information can be stored in the **dwc:Taxon** class, which has terms defined for the taxonomic name itself (**dwc:scientificName**), as well as a unique identifier, the Life Sciences ID (LSID) to locate the

[15] http://rs.tdwg.org/dwc/index.htm

[16] http://www.tdwg.org/



Amblycercus holosericeus

Link to taxonomic databases:

uBio Name Bank [ID: 3850937]
Encyclopaedia of Life (EoL) [ID: 1050869]
Catalogue of Life (CoL) [ID: 11914105]

Relevant evidences found in the biodiversity literature

1. O. Salvin and F. D. Godman. 1879-1904. Biologia Ce
- Amblycercus holosericeus … … † * * * * * * * * *

2. O. Salvin and F. D. Godman. 1879-1904. Biologia Centrali-Americana, Aves, vol. 1, p.445 [Original Work]
- Two rather closely allied species constitute this genus, one of which, A. holosericeus, Discussion

3. O. Salvin and F. D. Godman. 1879-1904. Biologia Centrali-Americana, Aves, vol. 1, p.446 [Original Work]
- 1. Amblycercus holosericeus. Heading
- Amblycercus holosericeus, Scl. Cat. B. Brit. Mus. xi. p. 3276. Synonym

4. O. Salvin and F. D. Godman. 1879-1904. Biologia Centrali-Americana, Aves, vol. 1, p.447 [Original Work]
- Amblycercus holosericeus is a very familiar species in Mexico and Central America, where it is found throughout the hot and temperate districts from Southern Mexico to Panama, and thence southwards to Ecuador and Peru. Discussion
- Sumichrast, who ascribes A. holosericeus to the hot and temperate regions of Vera Cruz, says it does not reach the altitude of Orizaba, the limit of its upward range not passing 3380 feet26. In Guatemala, however, we found it to be common at Coban at an elevation of between 4000 and 4500 feet, and at Dueñas at nearly 5000 feet. Discussion

Catalogue of Life: 23rd May 2013
indexing the world's known species

Species details

Accepted scientific name:	*Amblycercus holosericeus* (Deppe, 1830) (accepted name)		
Synonyms:	-		
Infraspecific taxon:	*Amblycercus holosericeus australis* Chapman, 1919 *Amblycercus holosericeus flavirostris* Chapman, 1915 *Amblycercus holosericeus holosericeus* (Deppe, 1830)		
Common names:	Common name	Language	Country
	Yellow-billed Cacique	English	-
	Cacique pico claro	Spanish	-
Classification:	Animalia	CoL	LSID
Phylum	Chordata	CoL	LSID
Class	Aves	ITIS Global	LSID

Biologia Centrali-Americana :zoology, botany and archaeology /edited by Fre...

URL for Current Page
http://biodiversitylibrary.org/page/571425

Scientific Names on this Page
Page 446
Amblycercus [EOL]
Amblycercus holosericeus [EOL]
Cacicus holosericeus [EOL]
Cassiculus [EOL]
Cassicus
Nigerrimus

The short, rounded wing presents another obvious point of difference from *Cassicus* and *Cassiculus*.

As in *Cassiculus* the feathers of the lower back are normal, the barbs being fully furnished with barbules.

1. Amblycercus holosericeus.

Sturnus holosericeus, Licht. Preis Verz. Mex. Vög. p. 1 (cf. J. f. Orn. 1863, p. 55).
Cacicus holosericeus, Salv. Cat. Strickl. Coll. p. 262.
Cassicus holosericeus, Scl. Ibis, 1883, p. 163; Boucard, P. Z. S. 1883, p. 445; Tacz. Orn. Pér. ii. p. 415.
Amblycercus holosericeus, Scl. Cat. B. Brit. Mus. xi. p. 327.
Amblyramphus prevosti, Less. Cent. Zool. p. 150, t. 54.
Amblycercus prevosti, Cab. J. f. Orn. 1861, p. 10; Cass. Pr. Ac. Phil. 1867, p. 73; Lawr. Ann. Lyc. N. Y. ix. p. 104; Frantz. J. f. Orn. 1869, p. 302.
Cassiculus prevosti, Scl. P. Z. S. 1856, p. 301; 1859, pp. 365, 380; 1860, p. 276; 1864, p. 174; Scl. & Salv. Ibis, 1859, p. 19; 1860, p. 34; P. Z. S. 1864, p. 353; 1870, p. 836; Moore, P. Z. S. 1859, p. 57; Lawr. Ann. Lyc. N. Y. viii. p. 180; Salv. P. Z. S. 1867, p. 142; 1870, p. 190; Sumichrast, Mem. Soc. N. H. i. p. 553.

Contributed by Smithsonian Libraries

Figure 3. A sample web page to show how extracted semantic features link to the BHL and external taxonomic databases

taxon across remote databases (**dwc:taxonID**) and various terms giving taxonomic information and provenance. For example, the metadata identifying the LSID and publication data for the species *Anthus correndera* might be represented using the standard Darwin Core terms:

```
<dwc:Taxon>
   <dwc:taxonID>urn:lsid:catalogueoflife.org:
     taxon:f000e838-29c1-102b-9a4a-
      00304854f820:col20120721</dwc:taxonID>
   <dwc:scientificName>Anthus
                  correndera</dwc:scientificName>
   <dwc:class>Aves</dwc:class>
   <dwc:genus>Anthus</dwc:genus>
   <dwc:specificEpithet>correndera
                              </dwc:specificEpithet>
   <dwc:namePublishedIn> London Med.
         Repos., 15: 308.</dwc:namePublishedIn>
</dwc:Taxon>
```

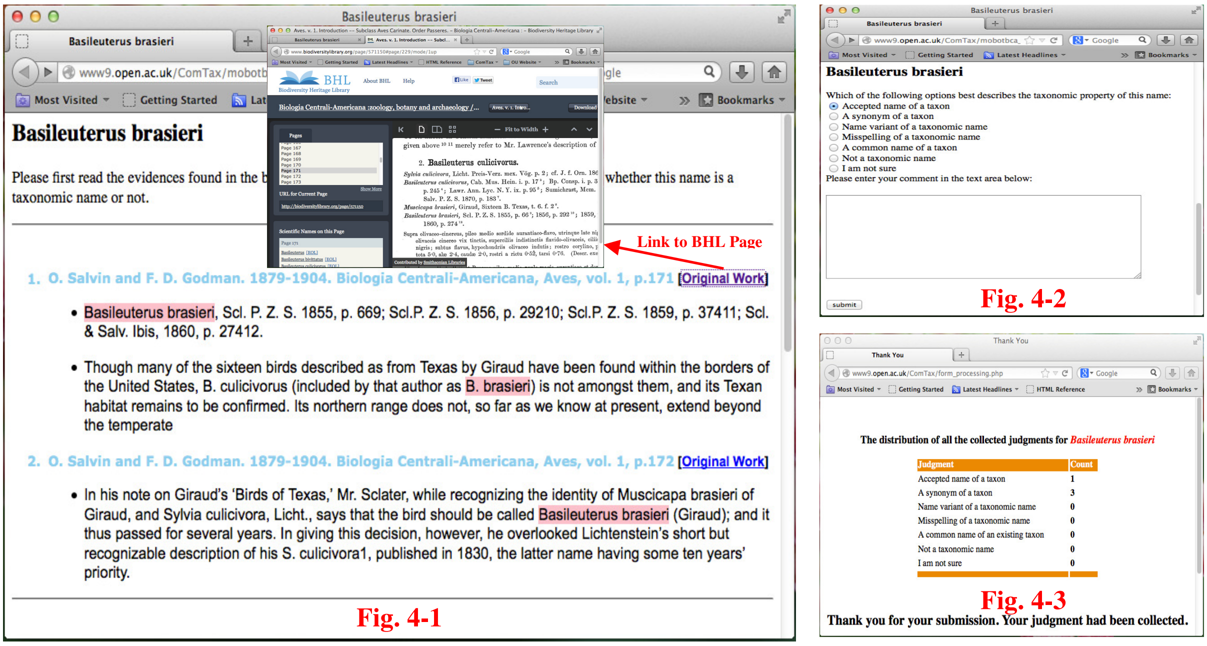
The basic Darwin Core terms can be extended to represent the information obtained via the original document and the curation tools. Labels should be used to represent information about the verified form of the name and the identity of the verifier, with the verifier's Scratchpad login name being the obvious choice. Adding the name of the verified form and the verifier with appropriate labels to the metadata (**dwc:nameVerifiedBy** and **dwc:dateVerified**) would then give:

```
<dwc:Taxon>
   <dwc:taxonID>urn:lsid:catalogueoflife.org:
     taxon:f000e838-29c1-102b-9a4a-
      00304854f820:col20120721</dwc:taxonID>
   <dwc:scientificName>Anthus
                  correndera</dwc:scientificName>
   <dwc:nameVerifiedBy>Scratchpad user:
            Michael Smith</dwc:nameVerifiedBy>
   <dwc:dateVerified>2013-06-15
                              </dwc:dateVerified>
   <dwc:genus>Anthus</dwc:genus>
   <dwc:specificEpithet>correndera
                              </dwc:specificEpithet>
   <dwc:namePublishedIn>London Med.
         Repos., 15: 308.</dwc:namePublishedIn>
</dwc:Taxon>
```

Further contextual information can be stored, providing species description information including morphological features, biogeographic distribution, and ecology. Figure 3 shows a web page[17] corresponding to a species in which the contexts surrounding the occurrence of the target mention are extracted from the text. Each piece of evidence is given a bibliographic citation that is linked to the respective copy of the referring page (here, the BHL). Unique database identifiers and hyperlinks to external taxonomic databases are provided on the web page if possible. Connections to external databases increase the understanding and analysis of the behaviour of the target species. These bibliographic linkages allow the system to identify and track back the raw data across the range of remote databases.

The metadata can potentially encode many semantic aspects of the data. Identified taxonomic names and hyperlinks to repositories will improve species-specific document retrieval. Encoding different names for organisms will improve synonym detection so reconciliation techniques are needed to connect multiple names. Also, linkages to the unique identifiers of organisms facilitate the reconciliation process. Future work will consider citation information, which improves the traceability of naming authorities.

[17] http://www9.open.ac.uk/Web_NER/mobotbca_03_01_00/3982.html


Basileuterus brasieri
Basileuterus brasieri
www9.open.ac.uk/ComTax/mobotb
Most Visited
Getting Started
Bookmarks
Aves. v. 1. Introduction -- Subclass Aves Carinate. Order Passeres. – Biologia Centrali-Americana : – Biodiversity Heritage Library
www.biodiversitylibrary.org/page/571150#page/229/mode/1up
BHL
Biodiversity Heritage Library
About BHL
Help
Search
Biologia Centrali-Americana :zoology, botany and archaeology /...
Aves. v. 1. Intro...
Download
Fit to Width
Pages
Page 167
Page 168
Page 169
Page 170
Page 171
Page 172
Page 173
Show More
URL for Current Page
http://biodiversitylibrary.org/page/571150
Scientific Names on this Page
Page 171
Basileuterus [EOL]
Basileuterus bivittatus [EOL]
Basileuterus culicivorus [EOL]
Contributed by Smithsonian Libraries
Link to BHL Page

given above 10 11 merely refer to Mr. Lawrence's description of

2. Basileuterus culicivorus.

Sylvia culicivora, Licht. Preis-Verz. mex. Vög. p. 2; cf. J. f. Orn. 186
Basileuterus culicivorus, Cab. Mus. Hein. i. p. 17 3; Bp. Consp. i. p. 3
p. 245 4; Lawr. Ann. Lyc. N. Y. ix. p. 95 5; Sumichrast, Mem.
Salv. P. Z. S. 1870, p. 183 7.
Muscicapa brasieri, Giraud, Sixteen B. Texas, t. 6. f. 2 8.
Basileuterus brasieri, Scl. P. Z. S. 1855, p. 66 9; 1856, p. 292 10; 1859,
1860, p. 274 12.

Supra olivaceo-cinereus, pileo medio sordide aurantiaco-flavo, utrinque late ni
olivaceis cinereo vix tinctis, superciliis indistinctis flavido-olivaceis, cilii
nigris; subtus flavus, hypochondriis olivaceo indutis; rostro corylino, p
tota 5·0, alæ 2·4, caudæ 2·0, rostri a rictu 0·52, tarsi 0·76. (Descr. exe

Basileuterus brasieri

Please first read the evidences found in the b[...] whether this name is a taxonomic name or not.

1. O. Salvin and F. D. Godman. 1879-1904. Biologia Centrali-Americana, Aves, vol. 1, p.171 [Original Work]
   - Basileuterus brasieri, Scl. P. Z. S. 1855, p. 669; Scl.P. Z. S. 1856, p. 29210; Scl.P. Z. S. 1859, p. 37411; Scl. & Salv. Ibis, 1860, p. 27412.
   - Though many of the sixteen birds described as from Texas by Giraud have been found within the borders of the United States, B. culicivorus (included by that author as B. brasieri) is not amongst them, and its Texan habitat remains to be confirmed. Its northern range does not, so far as we know at present, extend beyond the temperate
2. O. Salvin and F. D. Godman. 1879-1904. Biologia Centrali-Americana, Aves, vol. 1, p.172 [Original Work]
   - In his note on Giraud's 'Birds of Texas,' Mr. Sclater, while recognizing the identity of Muscicapa brasieri of Giraud, and Sylvia culicivora, Licht., says that the bird should be called Basileuterus brasieri (Giraud); and it thus passed for several years. In giving this decision, however, he overlooked Lichtenstein's short but recognizable description of his S. culicivora1, published in 1830, the latter name having some ten years' priority.

Fig. 4-1

Basileuterus brasieri

Which of the following options best describes the taxonomic property of this name:
● Accepted name of a taxon
○ A synonym of a taxon
○ Name variant of a taxonomic name
○ Misspelling of a taxonomic name
○ A common name of a taxon
○ Not a taxonomic name
○ I am not sure
Please enter your comment in the text area below:

submit

Fig. 4-2

The distribution of all the collected judgments for *Basileuterus brasieri*

Judgment	Count
Accepted name of a taxon	1
A synonym of a taxon	3
Name variant of a taxonomic name	0
Misspelling of a taxonomic name	0
A common name of an existing taxon	0
Not a taxonomic name	0
I am not sure	0

Fig. 4-3

Thank you for your submission. Your judgment had been collected.

Figure 4. A sample web page for taxonomic name verification (4-1: Extracted context evidence from the text; 4-2: A multi-option form for judgment collection; 4-3: The distribution of human judgments)

4.1 Taxonomic Name Validation

Taxonomic name validation task is to present unknown names for human validation. Validating taxonomic names is a specialist process, requiring extensive human involvement and expertise. Non-professional taxonomists and citizen scientists are an essential part of this effort. We aim to demonstrate how small, lightweight plug-ins integrated to existing web-based collaboration tools can facilitate the semantic annotation of open biodiversity resources via crowdsourcing techniques.

Scratchpads (Smith et al. 2009) are a content management system that is optimised for handling biological taxonomy data. Scratchpads are widely used amongst professional and amateur taxonomists, and so are a useful portal for validation.

Our curation web service is a Scratchpads plug-in. Text for validation is selected via a simple recommender system[18] (Figure 4). Users are presented with one or more potential taxonomic names found by the CRFs, as text "snippets" containing the proposed name with the surrounding context of the original scan (Figure 4-1). To collect specialists' judgments, a multiple-option form (Figure 4-2) is used to request a judgment of whether the text snippet represents a potentially new taxon, a synonym or a name variant of an existing organism.

The validation information is collected in a back-end MySQL database in a metadata format that contains the curator's name, verification time stamp, the target name, the associated publication, along with appropriate page citations and associated URI page linkages to make the support evidence traceable. By ensuring that this data is available to the community via the semantic web service layer, the judgment is exposed to the community for further validation or modification (distribution illustrated in Figure 4-3).

Our aim in the medium term is to link the validation task to search results within the Scratchpad portal[19]. This will allow us to investigate whether the output of document searching can be used as a reward for carrying out the validation exercise, and so whether the task can be presented in a (relatively) unobtrusive manner to users.

5 Conclusions and Future Work

Increasing numbers of older documents are scanned and made available online, through digital heritage projects like BHL. It will become more important to annotate those documents with semantic data in order to curate and manage the information contained in the documents.

We have described how information extraction techniques can be used as part of a curation system to improve the mechanisms for collecting

[18] http://www9.open.ac.uk/ComTax/mobotbca_03_01_00/4136.html

[19] http://taxoncuration.myspecies.info/node/77

this metadata. Although we have focused on identifying taxonomic names, the same techniques could be used to recognise any data of interest, such as geographical data in historic land documents, or proper names in census data. The critical part of the system, of course, is to be able to find suitable user groups to provide the appropriate semantic markup, as the data can rapidly become very large.

The semantic web can provide a portal to this data, if the metadata can be reliably collected. We believe that IE-supported curation techniques can be used to bring this collection about.

Future work includes: (1) The datasets were annotated by one computer scientist. It would be interesting to compare the annotated data with the verification results from biodiversity experts. (2) We need more annotated OCR text for the development of an automated OCR-error correction tool and a TNR tool built for OCR text. (3) Our project is in its early stages and requires more time for the collection of validation judgments; to conduct the evaluation of the validation tool and to analyse the validation results.

Acknowledgments

The work was funded by JISC through the Digital Infrastructure programme, Managing Research Data call. We thank Mr Jef Lay for developing the OCR error correction tool shown in Figure 2.

References

L. M. Akella, C. N. Norton, and H. Miller. 2012. NetiNeti: discovery of scientific names from text using machine learning methods. *BMC Bioinformatics,* 13 (211):1471-2105.

M. Gerner, G. Nenadic, and C. M. Bergman. 2010. LINNAEUS: A Species Name Identification System for Biomedical Literature. *BMC Bioinformatics* 11:85.

H. C. Godfray. 2002. Challenges for taxonomy. *Nature,* 417 (6884):17-19.

N. J. Isaac, J. Mallet, and G. M. Mace. 2004. Taxonomic inflation: its influence on macroecology and conservation. *Trends in Ecology and Evolution.* 19: 464–469.

D. Koning, N. Sarkar, and T. Moritz. 2005. Taxongrab: extracting taxonomic names from text. *Biodiversity Informatics,* 2:79–82.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning (ICML-2001)*, pages 282-289.

P. R. Leary, D. P. Remsen, C. N. Norton, D. J. Patterson, and I. N. Sarkar. 2007. uBioRSS: tracking taxonomic literature using RSS. *Bioinformatics,* 23(11):1434–1436.

R. D. M. Page. 2006. Taxonomic Names, Metadata, and The Semantic Web. *Biodiversity Informatics*, 3, pp. 1-15.

L. Penev, D. Agosti, T. Georgiev, et al. 2010. Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. *ZooKeys,* 50:1–16.

D. Remsen. 2011. Biodiversity Informatics - GBIFs role in linking information through scientific names. In *The Symposium of Anchoring Biodiversity Information: From Sherborne to the 21st Century and Beyond.*

I. N. Sarkar. 2007. Biodiversity informatics: organizing and linking information across the spectrum of life. *Briefings in Bioinformatics,* 8(5):347-357.

G. Sautter, K. Böhm, and D. Agosti. 2006. A combining approach to find all taxon names (FAT) in legacy biosisystematics literature. *Biodiversity informatics,* 3:41-53.

Secretariat of the Convention on Biological Diversity (SCBD). 2008. *Guide to the global taxonomy initiative*. CBD, Technical Report.

V. Smith, S. Rycroft, K. Harman, B. Scott, and D. Roberts. 2009. Scratchpads: A Data-Publishing Framework to Build, Share and Manage Information on the Diversity of Life. *BMC Bioinformatics* 10(14):S6.

P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, pages 102-107.

Q. Wei, P. B. Heidorn, and C. Freeland. 2010. Name Matters: Taxonomic Name Recognition (TNR) in Biodiversity Heritage Library (BHL). In *Proceedins of* 2010 *iConference,* pages 284-288.

R. J. White. 2007. Linking Biodiversity Databases. *Systematics Association Special Volume,* 73:111–128.

A. Willis, D. Morse, A. Dil, D. King, D. Roberts, and C. Lyal. 2009. Improving search in scanned documents: Looking for OCR mismatches. In *Proceedings of the workshop on Advanced Technologies for Digital Libraries*, 2009.

H. Yang, G. Nenadic, and J. Keane. 2008. Identification of Transcription Factor Contexts in Literature Using Machine Learning Approaches. *BMC Bioinformatics* 9:S11.