BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e35013
doi: 10.3897/biss.3.35013

OPEN ACCESS

Conference Abstract

Text-mining BHL: towards new interfaces to the biodiversity literature

Roderic Page ‡

‡ University of Glasgow, Glasgow, United Kingdom

Corresponding author: Roderic Page (roderic.page@glasgow.ac.uk)

Received: 30 Mar 2019 | Published: 13 Jun 2019

Citation: Page R (2019) Text-mining BHL: towards new interfaces to the biodiversity literature. Biodiversity Information Science and Standards 3: e35013. https://doi.org/10.3897/biss.3.35013

Abstract

The taxonomic literature is one of the largest resources of information on biodiversity, both current and in the past. Unlike many scientific disciplines this literature remains perpetually relevant as successive taxonomic work builds upon those earlier foundations. Projects such as the Biodiversity Heritage Library (BHL) have greatly increased access to that literature, as have numerous independent digitisation efforts by museums, herbaria, and publishers. But the focus of this access has been human readers, with limited use of text mining tools, mostly focussed on extracting taxonomic names. This talk explores other kinds of data that can be extracted from text on BHL and elsewhere, focusing on taxonomic names, geographic localities and specimen codes in the context of the BioStor project (https://biostor.org, Page 2011).

The problem of finding taxonomic names in text has been well studied (e.g., Akella et al. 2012), and new BHL content is continuously indexed by names. Despite this, there is only weak linkage between taxonomic name databases and BHL. Even projects that create these links (e.g., BioNames, Page 2013) do not enable links in the reverse direction. In other words, a BHL reader is unaware whether the appearance of a name on a page is the first publication of that name, nor are they told of the fate of a name in subsequent research. The absence of these links reduces the value of BHL to working taxonomists.

In addition to taxonomic names, a typical taxonomic paper often contains specimen codes. Extracting these from text and linking them to digital representations, such as occurrence

records in GBIF, opens up the possibility to provide detailed provenance for occurrence data, as well as citation-based metrics for the utility of natural history collections.

Taxonomic papers are also often rich in geographic information. A simple method for extracting locality information from text is to search for latitude and longitude coordinates, and BioStor currently does this. To date some 83,000 individual point localities have been extracted (Fig. 1). These are used to provide a simple geographic search interface in BioStor, and are also harvested by JournalMap (Karl et al. 2013). But these localities are not linked to the original location in the source text, nor are they linked to any associated specimens, so they cannot be interpreted as occurrences that could be harvested by GBIF. If the goal is to contribute to GBIF then we need tools that can parse locality information and link that to associated specimens.

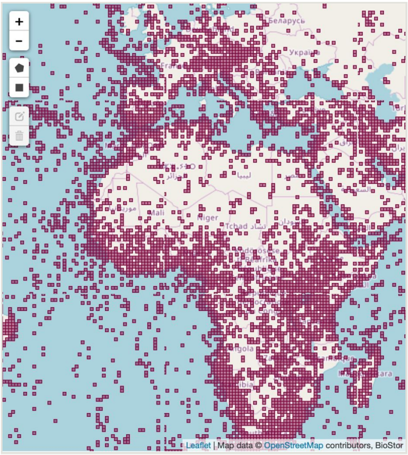

Figure 1.

Point localities extracted from articles in BioStor.

A general framework for handling data on taxonomic names, specimens, and geographic localities in text is to treat them as annotations (Batista-Navarro et al. 2017). By modelling annotations using the Web Annotation Data Model (https://www.w3.org/TR/annotation-model/) we can incorporate these annotations into biodiversity knowledge graphs (Page 2016). We can also combine these annotations with new standards for describing digitised content, such as the International Image Interoperability Framework (IIIF, https://iiif.io). The implications of this approach for developing new interfaces to the biodiversity literature will be discussed.

Keywords

text mining, BHL, BioStor, taxonomic names, specimens, geocoding

Presenting author

Roderic Page

References

- Akella LM, Norton C, Miller H (2012) NetiNeti: discovery of scientific names from text using machine learning methods. BMC Bioinformatics 13 (1): 211-211. https://doi.org/10.1186/1471-2105-13-211
- Batista-Navarro R, Zerva C, Nguyen NH, Ananiadou S (2017) A Text Mining-Based Framework for Constructing an RDF-Compliant Biodiversity Knowledge Repository. Information Management and Big Data. [ISBN 978-3-319-55209-5]. https://doi.org/10.1007/978-3-319-55209-5_3
- Karl J, Herrick J, Unnasch R, Gillan J, Ellis E, Lutters W, Martin L (2013) Discovering Ecologically Relevant Knowledge from Published Studies through Geosemantic Searching. BioScience 63 (8): 674-682. https://doi.org/10.1525/bio.2013.63.8.10
- Page RD (2011) Extracting scientific articles from a large digital archive: BioStor and the Biodiversity Heritage Library. BMC Bioinformatics 12 (1): 187-187. https://doi.org/10.1186/1471-2105-12-187
- Page RDM (2016) Towards a biodiversity knowledge graph. Research Ideas and Outcomes 2: e8767-e8767. https://doi.org/10.3897/rio.2.e8767
- Page RM (2013) BioNames: linking taxonomy, texts, and trees. PeerJ 1: e190-e190. https://doi.org/10.7717/peerj.190