PROCEEDINGS OF TDWG

Biodiversity Information Science and Standards 1: e17354
doi: 10.3897/tdwgproceedings.1.17354

OPEN ACCESS

Conference Abstract

Crowdsourcing Data Enhancements to Improve Named Entity Recognition in the Biodiversity Heritage Library

Katie Mika [‡]

[‡] Museum of Comparative Zoology, Harvard University, Cambridge, MA, United States of America

Corresponding author: Katie Mika (kmika@fas.harvard.edu)

Received: 20 Jul 2017 | Published: 25 Jul 2017

Citation: Mika K (2017) Crowdsourcing Data Enhancements to Improve Named Entity Recognition in the Biodiversity Heritage Library. Proceedings of TDWG 1: e17354. https://doi.org/10.3897/tdwgproceedings.1.17354

Abstract

The Biodiversity Heritage Library's holdings include dozens of manuscript collections that are largely hidden due to minimal descriptive metadata and the absence of machine readable facsimiles. Transcription projects for collections are time consuming, intellectually intensive, and expensive for an organization to facilitate. Crowdsourcing has been identified as a sustainable model for generating transcriptions for large collections and institutions with diverse holdings and may improve data collection from a diverse range of users to enhance descriptive metadata. Manuscript transcriptions also must fit into the larger BHL objective of producing and making available large scale datasets for users and researchers to study and manipulate. While generating only full-text transcriptions will improve discoverability, items remain isolated from other literature and collections. The GNA machine learning and named entity recognition algorithms allow BHL to extract, index, and attach page records to scientific names in order to create additional access points to biodiversity literature. These tools depend on imperfect optical character recognition (OCR), which contain spelling and layout errors and outdated naming conventions and can be added to BHL's larger crowdsourcing platform to solicit corrections. Transcriptions, similarly, introduce antiquated taxonomic data and common names and must be optimized for use with GNA's name finding tools. Improving named entity recognition by correcting OCR output and enhancing transcribed manuscript items will allow BHL to better connect

content across collections and ultimately provide a broader and more complete picture of biodiversity.

Keywords

Crowdsourcing, Named Entity Recognition, Machine Learning, Optical Character Recognition, Manuscript Transcription

Presenting author

Katie Mika

Grant title

Foundations to Actions: Extending Innovations in Digital Libraries in Partnership with NDSR Learners