Mining Georeferences from Biodiversity Literature: A Pilot Study

By: Gretchen Stahlman
PhD Candidate
School of Information
University of Arizona
gstahlman@email.arizona.edu

In collaboration with:
Martin R. Kalfatovic, BHL Program Director and Associate Director of Digital Programming & Initiatives
Carolyn A. Sheffield, BHL Program Manager
Smithsonian Biodiversity Heritage Library
https://www.biodiversitylibrary.org/

In fulfillment of:

LEADS-4-NDP Fellowship
http://cci.drexel.edu/mrc/research/leads/

September 2018

Table of Contents

1 Introduction

This report describes the outcomes of a short pilot study that was undertaken by a doctoral student through the LEADS-4-NDP fellowship program, between May and August 2018. LEADS-4-NDP, hosted by Drexel University and funded by the Institute of Museum and Library Services, aims to educate future Library & Information Science (LIS) faculty in current data science issues and techniques. For the 2018 program, ten Fellows were placed with partner institutions to assist with projects involving a data science learning opportunity. The pilot project presented here has focused on exploring methods for automating the identification and disambiguation of geographic names within BHL collections (~55 million scanned pages) and, where possible, translating toponyms to polygons or point locations for visual browsing. By the conclusion of the project, a survey of the literature and related projects was conducted, and a variety of possible techniques were explored, ranging from simple to complex strategies and heuristics for identifying and linking georeferences. A test corpus of 50 documents was evaluated.

1.1 Summary of Recommendations

Among possible solutions, this report chiefly recommends gathering stakeholders and researchers together through a workshop or other targeted forum to identify community needs, priorities, and possible methods for accomplishing the objectives of this project. Scientific workshops are a common practice for gathering feedback from experts to inform directions for future research and development, and methodologies exist for accurately capturing ideas and project management strategies through workshops. For example, the Delphi Method (Keeney et al., 2001; Heidorn, et al., 2018) has been adapted to solicit consensus from experts through an iterative process of breakout groups and whole-group discussions, with initial questions for deliberation posed by workshop organizers, and subsequent questions identified through the outcomes of each breakout session to arrive to a final consensus. Considering the extreme diversity of BHL texts (in text quality, language, topic area, geographic representation, and historical context), a scientific workshop could prioritize practical and potentially fruitful areas of focus for initial efforts with available methods, while brainstorming feasibility and possible workflows for conducting information extraction activities with more complicated sectors of the corpus.

The report further outlines a possible workflow for a multi-phased, collaborative project to thoroughly and accurately extract and disambiguate place names in association with scientific names, considering the diversity of the BHL collections, the varying quality of Optical Character Recognition (OCR) in the scanned documents, and the historical contexts of these texts. The outcomes of this exploratory study indicate that machine learning could be used to create more nuanced entity recognition in BHL texts than is currently possible with generic Named Entity Recognition (NER) classifiers. Additionally, place name tags could be verified by human curators and/or citizen scientists through components of existing systems and gazetteers and linked to

parsed taxonomic names as if "collecting virtual specimens" (D. Schigel, C. Sheffield, personal communication, August 16, 2018). Finally, human-verified georeferences in the texts could augment BHL metadata, allow users to browse the collections spatially, and provide new occurrence data to the Global Biodiversity Information Facility (GBIF). The conclusion of the pilot study is that this project is both feasible and scientifically useful, and – as other communities are converging on similar objectives – the time is ripe for development and collaboration in this area in support of open science and public engagement.

1.2 Summary of Approaches and Deliverables

Following several initial meetings between the mentors and intern in May 2018, tasks and goals for the project were agreed upon and tracked in a Google spreadsheet. A communication plan was established, and regular weekly meetings took place, typically on Thursdays. Additional meetings were arranged with relevant stakeholders for information gathering and coordination. All BHL meetings were held virtually. The project was intended to be an opportunity for hands-on development of technical skills geared towards augmenting LIS education, and so the outcomes naturally shifted throughout the internship period as new knowledge was gained and relevant information discovered.

Approaches for tackling the problem are described briefly below:

- *Literature review, exploration of methods.* A thorough survey of the literature was initially conducted, and review of the literature and other relevant projects and methods was ongoing throughout the project. See section 15 for some key sources.
- *Interviews with stakeholders and researchers.* A series of introductory meetings was initiated by BHL with stakeholders and researchers internal and external to Smithsonian. Other formal and informal meetings were held with University of Arizona experts in Natural Language Processing and biodiversity informatics, as well as with researchers at other institutions working on similar projects. A summary of key meetings is provided in section 14.
- *Machine learning and crowdsourcing identified as promising methods.* The BHL collections represent a large full-text dataset. Current computing technology could be used to extract information from the collection on a large scale. Human effort for information extraction and quality control could be crowdsourced by leveraging existing volunteer and citizen science communities. See section 4 for more information.
- *Annotated test corpus for assessment and training.* For evaluation purposes, a test corpus of 50 documents was selected based on the following criteria: publication in the last 200 years (for better OCR), English language, and diversity of topics. A subset of this test corpus was annotated using brat annotation software. See sections 3 and 12 for more information.
- *Used StanfordNER classifier and Mordecai Geoparser to extract and visualize places names as a demonstration.* Named Entity Recognition (NER) is a natural language processing task, enabled by a variety of available software tools and statistical methods. For this project, two promising resources – Stanford's NER classifiers and Mordecai

Geoparser - were used for demonstration of a possible workflow pipeline leading to visualization of the place names in a BHL document. See section 4 for more information.

By August 2018, feasible **deliverables** of the pilot study coalesced to include:

- *Assessment and recommendations.* This report documents the LEADS project activities and final recommendations based on the approaches outlined above.
- *NER and visualization prototype.* See section 8 for an example of an NER and visualization pipeline output. Corresponding code is in Github.
- *Annotated test corpus.* The first 10 pages of some documents in the test corpus were annotated using brat annotation software (approximately 70 pages total). The following types of terms are annotated: toponyms, taxonomic names, type specimens, and relationships between toponyms and taxonomic names, and between type specimens and taxonomic names. Annotated data can be found in Github. Possible methods for evaluation of OCR quality are included. See the appendices at the end of this report for more information.
- *Annotation environment.* A Virtual Machine has been configured in CyVerse with Python 2.7 and the brat annotation software, along with the test corpus for verification by other annotators. See section 12 for details on accessing and using the VM.
- *Connections made for possible collaboration and citizen science.* The intern has reached out to the Natural Language Processing research group at the University of Arizona School of Information, to the Global Biodiversity Information Facility (GBIF), to Plazi, and to Zooniverse. These connections, among others, could be leveraged to collaborate on a full project. See section 14 for more information.

1.3 Summary of Learning Outcomes

Current approaches in data science emphasize an intersection between domain knowledge, computer science and statistics. LIS researchers, educators and practitioners must unify these sectors to identify, negotiate and provide access to relevant information. As a PhD student intern without prior deep knowledge of biodiversity, only superficial knowledge of computer science, and passing knowledge of statistical methods and computer programming (primarily applied to social science), the project was both a challenge and opportunity for rapid development of data science skills and new insight into domain-specific data practices. In the case of this particular internship assignment, the LEADS pairing of project with fellow successfully illuminated the crux of the data science movement towards domain-agnostic practices, inter- and trans-disciplinary research, and reproducible software. For the continued relevance of LIS as a field that transcends domains amidst rapidly-developing technology, documentation and standardization of platforms are imperative to the utility and longevity of data science applications across domains, as well as new education paradigms as exemplified by the LEADS program.

2 Background

As access to curated historical data is increasingly possible through advances in technology and community-driven initiatives, enabling time domain research is a focus for development of resources across disciplines, including species distribution over time in biodiversity research. Scientific literature is potentially a rich source of such information, where extracting data from the text of historical publications can produce insight and augment newer data. For biodiversity and ecological research in particular, Bowker (2000) illustrates the importance of data diversity across disciplines for cataloging and studying life itself across vast time periods and geographies. Similarly, Thompson, et al. (2001) emphasize that studying complex biological and ecological systems requires techniques for mining historical data to address cross-cutting research questions. A survey on the needs and practices of biodiversity scientists conducted in 2010-11 supports these recommendations (Davis, et al., 2014), including the following findings relevant to the current study on georeferencing historical biodiversity texts:

> Respondents were also asked what biodiversity information they would like to have, but were unable to find. For general biodiversity information, the more common responses mirror what is most often referenced in the literature. For example, needs include information concerning specific taxa, species, or geographic locations, occurrence and distribution information, and increased access to historic literature, specimen databases, and museum collections. **Seventy-one nearly identical responses were provided for the raw data respondents would like to have but have been unable to find: 48% concern species or taxon level occurrence, distribution and/or abundance data, 13% concern geographic raw data such as geo-referenced collection data**, and 13% concern aquatic related raw data. (p. 694)

Georeferenced databases are particularly useful for studying the distribution of biodiversity over time. The Global Biodiversity Information Facility (GBIF, 2018) supports a growing international database of curated species occurrence data, providing standards and tools for sharing and citation of georeferenced information. As an Associate Participant in GBIF, the Biodiversity Heritage Library (BHL) is a central resource for the research community and embodies a shared mission to make biodiversity data openly accessible and discoverable. BHL also partners with the Encyclopedia of Life (EOL), a rich resource curating descriptive information on all known life on Earth. Both EOL and BHL cross-link taxonomic references across their platforms.

BHL collections span 500 years and are contributed by institutions worldwide, consisting of millions of scanned pages with corresponding OCR text files. Currently, BHL offers full text searching, as well as scientific name recognition using the Global Names Architecture tools and services to display scientific names identified on each page. In order to further support discovery of content related to species occurrence and distribution data within the BHL corpus—specifically as relates to the geographic context of species observations or collection events—various annotation, text mining, machine learning, and visualization tools could be

used with the derivative files (e.g., plain text) available for each page in BHL to identify and disambiguate place names on BHL pages on a large scale, and to further link places with scientific names as potential new occurrence data as an additional service to the global biodiversity research community.

Related efforts have been undertaken to extract robust biodiversity data with geographic information from the literature. Plazi's GoldenGATE project has developed software for annotating and extracting treatment data from text and PDF files, and retrieving associated metadata (Faulwetter, et al., 2016). The GEOLocate project supports biodiversity research by translating descriptive text references of place names into geographic coordinates (see Appendix A for a demonstration of the workflow). The BioNames project links publication references and taxonomic data across a diversity of resources and platforms, including BHL (Page, 2011; 2013). Nguyen et al. (2017) present a sophisticated method of mining BHL documents for biodiversity terms using context and similarity measures to extract information and update taxonomic databases with previously unknown terms. Nguyen's research group – the National Centre for Text Mining (NaCTeM) at University of Manchester - is also currently working on the "COPIOUS" text mining project to efficiently extract occurrence data related to the Philippines from BHL literature (Gabud, et al., 2017; Batista-Navarro, et al., 2017).

Automatically detecting, disambiguating and visualizing place name references in large text datasets is a computing challenge across disciplines. However, this task also presents a shared opportunity for cutting edge research and development of scalable technology and techniques. For example, using machine learning Weissenbacher, et al. (2015) created an automated method for retrieving geospatial metadata from full-text phylogeology literature, successfully implementing toponym detection and disambiguation using dictionary-based and rule-based heuristics, and referring to the GeoNames database of locations for toponym resolution. In contrast to modern scientific literature with generally predictable structural characteristics and topic areas across publications, heterogeneous collections of historical texts are particularly challenging and unsuited for generic named entity recognition algorithms and modern gazetteers alone. To perform toponym resolution on a corpus of Civil War memos, DeLozier, et al. (2016) tackled this issue by creating the GeoAnnotate tool for collaborative annotating of locations and coordinates, with likely place names pre-annotated using Stanford's NER package to minimize the workload of annotators. As an example of a solution to the problem of discrepancies between historical and contemporary maps, Cura, et al. (2017) extracted information from historical maps of Paris and matched them to current addresses in modern Paris. For biodiversity, Cardoso et al. (2014) used several existing online geographic databases to create a new open gazetteer with superior recall for adding coordinates to species occurrence records. These projects highlight applicable methods – and difficulties – associated with attempting to accurately and automatically extract place names from the BHL corpus, which spans not only centuries, but the entire globe.

3 Opportunities and Challenges

Natural language processing (NLP) and machine learning represent a potential method of recognizing geographic references and occurrence relationships in the text of BHL collections. Recent advances in deep learning and computational linguistics make it possible to implement complex algorithms to extract contextualized information from an entire corpus of literature with reasonable confidence. In the case of BHL's scanned documents, OCR text files could be used for machine learning, following some preliminary assessment and processing of the text files, along with annotation of a subset of the corpus for training and testing. Through the course of this exploratory project, some **NLP challenges particular to the BHL corpus** have been identified and are listed below:

1) **OCR quality**.

OCR is a challenge for NLP tasks across research areas (Alex & Burns, 2014; Klein, et al., 2014; Kumar, 2016), including for scanned biodiversity texts – which often include handwritten field notes - (Heidorn & Zhang, 2013; Paul, et al., 2013; Paul & Heidorn, 2013). Particularly for older documents, the text will likely require additional human effort to clean errors and/or transcribe the documents prior to processing.

2) **Diversity of the corpus**.

The diversity of a corpus can affect the performance, and thus the suitability, of machine learning for extracting accurate information (Shmanina, et al., 2013; Sokolova & Bobicev, 2018). BHL collects a tremendous diversity of texts, ranging from handwritten documents to structured scientific articles in a variety of languages. Named Entity Recognition efforts could be clustered according to similarity of document topics, types and languages, while leveraging community input to prioritize areas with higher likelihood of resulting in useful occurrence data.

3) **Historical context**.

Aside from linguistic variations in historical documents, toponym resolution is particularly challenging for older texts (DeLozier, et al., 2016). Cura, et al. (2017) refer to historical maps of the city of Paris in order to translate place name references to current addresses – a tactic that would not be practical considering the international scope of BHL documents. Furthermore, political boundaries and geographic features (such as rivers and coastlines) also change over time. This creates difficulty in extracting accurate data from the texts and requires contextual insight and customized heuristics, especially considering that the BHL corpus spans five centuries.

4) **Ambiguity of geographic references**.

Some BHL documents contain very structured information – such as lists of species and treatments with predictable geographic references in close proximity to species and specimens within the text. However, even within these structured areas, references to places can be ambiguous. Examples of such ambiguity from the test corpus evaluated here include:

- Ranges, such as "Puno: Agapata to Chatapata";
- Negation, such as "South American genre not represented in United States";
- 3-dimensional space, such as "elevation: 3,200 meters".

Considering the challenges for automation outlined above, some level of human effort in the project explored by this pilot study is unavoidable. Any larger-scale effort must utilize human effort for software and heuristic development, information extraction and curation efficiently. **Crowdsourcing is a potential solution** to aspects of resolving the associated linguistic and contextual challenges. Citizen science research and practice have shown that while machines are successful at certain tasks, humans possess superior decision-making skills (Sorensen, et al., 2016; Xue, 2014). These intuitive abilities can be leveraged by mobilizing online communities to perform classification tasks, where a large number of individual efforts are assessed algorithmically to produce accurate results when aggregated (Lintott, 2008; Mugar, et al., 2015). In the case of this project, citizen scientists could feasibly perform quality control on automatically-generated placename and species tags to extract usable data from BHL documents (see section 5 for possible paths forward).

4 Summary of Tools and Methods

A first step towards the envisioned BHL geoparsing project is to assess existing efforts and software. Numerous tools and methods were explored for this pilot study (see appendices for an annotated literature review, summary of meetings and contacts, code and visualization description, and other products). Key potentially-useful resources that were distilled from this broad exploration are listed and briefly described below:

4.1 Toponym Recognition and Disambiguation

4.1.1 Named Entity Recognition Tools

StanfordNER. The Stanford NER[1] (Finkel, et al., 2005) software package comes with several trained classifiers for extracting named entities (including places) from text. The software is free to download and use, but requires web hosting on a local machine or server. The NLTK toolkit for Python provides a programmatic interface to StanfordNER[2].

Edinburgh Geoparser. As described by Grover, et al. (2010), the Edinburgh Geoparser NER model is based on heuristics to disambiguate placenames based on context, font and lexical rules. A corresponding visual interface – the Edinburgh Geo-annotator – was also developed to assist with annotation and human correction of extracted place references (Alex, et al., 2014). The Edinburgh Geoparser is available for download[3], but requires installation of additional

[1] https://nlp.stanford.edu/software/CRF-NER.shtml
[2] http://www.nltk.org/api/nltk.tag.html#module-nltk.tag.stanford
[3] https://www.ltg.ed.ac.uk/software/geoparser/

software dependencies and is not compatible with all versions of common operating systems. The software can be wrapped in a Docker container for portability[4].

GeoDeepDive. The GeoDeepDive[5] project links a substantial digital library of full text publications with High Throughput Computing for large-scale automated text and data mining projects. Text in GeoDeepDive is pre-processed for NLP tasks using Stanford CoreNLP, and all documents are OCR'd with Tesseract for quality control and uniformity. Due to access agreements with publishers, results of GeoDeepDive queries only contain relevant snippets of text surrounding terms of interest. Applications can be developed and deployed using an existing template and in coordination with the GeoDeepDive team. Named entities – geonames and scientific names - could be quickly extracted from a large corpus (including all BHL texts if contributed to the corpus) using a dictionary-based method[6] along with proximity parameters for linking entity types (see section 9 for an example using the GDD API). Detailed notes from the recent GeoDeepDive Workshop in Madison, WI (August 20-21, 2018) can be found here: https://docs.google.com/document/d/1DBg1q0XfVA5_umf2wZDhTg5za_SE8r6pM3r5cvn8Xvs/edit?usp=sharing

Mordecai. Development of the Mordecai[7] Python library for geoparsing was funded by the Defense Advanced Research Projects Agency (DARPA), the US Army Research Office, and the National Science Foundation (NSF), Mordecai uses spaCy NER to extract place names. Mordecai requires a copy of the GeoNames gazetteer running as a local web service and indexed with Elasticsearch (which also requires installation of Docker). Mordecai was developed with neural networks and trained using Prodigy. Mordecai output produces detailed geographic information, along with a country-level confidence score (parameters can be adjusted to only return results with a higher confidence score). A batch_geoparse option was recently developed, though speed may be an issue for a very large corpus: https://github.com/openeventdata/mordecai/issues/55. See section 8 for an example of placename extraction using Mordecai.

4.1.2 Machine Learning Tools

Stanford CoreNLP. This widely-used suite of NLP tools[8] facilitates text processing and analysis tasks, as well as training new NER models based on annotated corpora.

spaCy. This is a powerful Python library[9] for Natural Language Processing, particularly suited for deep learning projects and neural network models.

---

[4] https://github.com/metazool/edinburgh-geoparser-docker
[5] https://geodeepdive.org/
[6] https://github.com/UW-Deepdive-Infrastructure/dictionary_example
[7] https://github.com/openeventdata/mordecai
[8] https://stanfordnlp.github.io/CoreNLP/index.html
[9] https://spacy.io/

Prodigy. This software package[10] provides an annotation tool that works throughout the annotation process to dynamically create machine learning models while training data. Prodigy is not free, and licenses are available at a range of prices and number of users.

TolstoyAI. Several conversations were held with TolstoyAI[11] throughout the period of this study. The startup company provides user-friendly text classification applications and has achieved better-than-human accuracy on certain tasks. Using annotations of 178 pages of Joseph Grinnell's field notes, the company achieved ~80% accuracy on entity extraction. TolstoyAI indicated that they would like to assist museums with digitizing collections, hoping to be included in future grant proposals. It was discussed that TolstoyAI could test their models on the annotations produced by this LEADS-BHL project. It is a strong recommendation of this report that any useful annotations produced by the LEADS fellowship be made available for other researchers to work with as well.

4.1.3 Human Effort

brat. brat[12] was selected as the annotation tool for this project for several reasons. First, it is used for several collaborative NLP annotation projects by University of Arizona School of Information researchers, as well as by other related projects (such as the Arizona State university phylogeology project discussed in the Background section above – see Weissenbacher, et al., 2015). Second, the brat annotations are stored in a standoff format, preserving the original data. Third, it provides a simple visual interface for highlighting text to be annotated, and drawing relationships between annotations. Finally, the structured format of the annotations can be easily used for computing tasks. brat runs as a web service, so must be hosted locally or on a server. See section 8 for a visualization example of placename extraction using brat annotations.

Recogito. The Recogito[13] project is an initiative of Pelagios Commons[14], which is a community focused on supporting linked open data, particularly linking historical information to geographic places. The Mellon-funded Recogito tool is open-source and provides a seamless environment for annotating, visualizing and exporting placenames, spatial information and entity relationships. Using four existing gazetteers (Pleiades Gazetteer of the Ancient World, Digital Atlas of the Roman Empire, a subset of GeoNames, and China Historical GIS)[15], Recogito pre-annotates placenames, making it easy for a user to verify or correct the placename annotations with a map interface. Data can be downloaded in a variety of formats, including JSON for easy incorporation into existing databases and interface with other applications. Recogito is working with the gazetteer community to add new coverage, and unique installations of this open

[10] https://prodi.gy/features/named-entity-recognition
[11] http://tolstoy.ai/
[12] http://brat.nlplab.org/
[13] https://recogito.pelagios.org/
[14] http://commons.pelagios.org/
[15] https://recogito.pelagios.org/help/faq

source platform can import customized gazetteers for more nuanced placename extraction pertaining to specific geographic areas and periods of time.[16]

GoldenGATE. As a steward of - and advocate for - open biodiversity data, Plazi[17] maintains the GoldenGATE[18] XML markup tool for document annotation, particularly to extract taxonomic treatments from text. While the GoldenGATE-Imagine version of the software permits editing of PDF documents, the process is currently prone to a variety of errors. A 2016 workshop found that workshop that GoldenGATE-Imagine could be developed specifically for occurrence extraction, representing potential for new contributions of taxonomic names not currently recognized by GBIF. However, the process is currently tedious and prone to errors, though tools exist that could assist in mobilizing data more easily. See Faulwetter, et al. (2016) for more information.

GeoAnnotate. Similar to the Edinburgh Geo-annotator mentioned above, GeoAnnotate[19] was developed as an annotation interface for the toponym resolution project of DeLozier, et al. (2016) using a corpus of Civil War documents. The interface is open source and can be installed locally with a Parse Server backend.

Zooniverse. As discussed above, communities of citizen scientists can be leveraged to perform computing tasks on large datasets. With more than 1.5 million users, Zooniverse[20] currently hosts 90 citizen science classification projects, including several focused on extracting information from text. Standard tools are available for straightforward classification projects, while specialized software can be developed according to the needs of a project.

4.1.4 Mapping and Visualization

Python libraries. A number of geocoding libraries are written for Python, including Geocoder[21] and GeoPy[22]. These tools facilitate communication with popular web services such as GeoNames[23] and Google[24] to obtain the coordinates of named places and recognizable addresses for mapping and other data manipulation.

GEOLocate. Developed for georeferencing natural history collections data, GEOLocate parses text references into point coordinates and polygons. An API exists for batch processing of text. Output is delivered in JSON format. When submitting a locality description, GEOLocate requires

[16] https://github.com/pelagios/recogito2
[17] http://plazi.org/
[18] http://plazi.org/resources/treatmentbank/goldengate-editor/
[19] https://github.com/utcompling/GeoAnnotate
[20] https://www.zooniverse.org/
[21] https://geocoder.readthedocs.io/
[22] https://geopy.readthedocs.io/en/latest/
[23] http://www.geonames.org/
[24] https://developers.google.com/maps/documentation/geocoding/start

country and state for US locations. A sample R script for batch processing using the web services with a CSV file returns two output files – one with all returned locations, and the other with only the first returned location: http://www.geo-locate.org/files/glc.r.txt. A collaborative web client exists for community efforts.

Python - Leaflet + Folium. Folium[25] enables visualization of geographic data with Leaflet (through a leaflet.js library). An example of visualization using folium and Leaflet[26] mapping software in a Jupyter Notebook is described in section 8.

5 Possible Paths Forward

The insight, software, and techniques identified above are all resources that could be applied to components of an effort to extract geographic information and biodiversity data from BHL collections. A full project should be designed according to the needs of the community, and considering the institutional capabilities of BHL and its partners and stakeholders. As mentioned in section 1.1 above, initiatives with potential scientific impact and that require investment and buy-in from stakeholders can be informed by targeted opportunities for dialogue and by soliciting expert feedback (see Keeney et al., 2001 and Heidorn, et al., 2018 for examples). Automating the extraction of information from BHL collections is complicated by the structural, linguistic, and contextual diversity of the corpus. Particularly for NLP and machine learning methods that depend on training a computer to recognize features of the documents, certain areas of the corpus may be identified as more promising sources of new data. Compiling an initial set of use cases could help maximize the use of available resources, target selection of content to focus on for real world research questions, and build a group of stakeholders that might be willing or even eager to participate in testing or otherwise contribute time, effort, or expertise. Collaborators could be identified to assist with the effort, and with developing and submitting funding proposals if necessary. Such a workshop or targeted meeting could bring stakeholders together to obtain community feedback and outline priorities and methods. Depending on the scale of the effort, possible workflows (and budgets) range from simple to complex, and two potential workflows are outlined below.

5.1 (Relatively) Simple/Inexpensive Workflow:

Rod Page – author of iPhylo[27] and creator of BioNames – has been thinking about strategies for this project, as well. Page recently created an online tool that searches GBIF data to georeference strings of text: https://lyrical-money.glitch.me/

[25] https://python-visualization.github.io/folium/docs-v0.6.0/
[26] https://leafletjs.com/
[27] http://iphylo.blogspot.com/

This concept is described in the application's ReadMe file: https://glitch.com/edit/#!/lyrical-money?path=README.md:1:0

For BHL, Page recommends: "Use [succinct] data structures http://stevehanov.ca/blog/index.php?id=120 and list of countries and admin level 1 localities from geonames to find locations, grab surrounding text, georeference, store centroid and polygon in database, index using Elastic search" (Page, 2018).

This strategy could perhaps be facilitated using the GeoDeepDive infrastructure (see section 4.1.1 above). Since GeoDeepDive can match text to dictionary terms efficiently, extensive dictionaries of place names and taxon names could be applied to the BHL corpus. Assuming that taxon names which are very likely to be associated with placenames appear within a concise number of characters from each other (which is not universally the case in the BHL corpus as encountered during the annotation process for this exploratory project), GeoDeepDive could return the related snippets of text corresponding to each document in BibJSON format for further evaluation and processing. As part of GeoDeepDive, the BHL corpus could then be available for further discovery through text mining by other researchers.

5.2 (Relatively) Complex Workflow:

A more complicated (but perhaps more thorough) workflow would be to begin with an annotation project, leading to machine learning to extract likely place names and taxonomic references, and incorporating a citizen science project to verify and correct tagged entities, which can then be used as data and to augment BHL metadata for spatial browsing. Pustejovsky & Stubbs (2012) outline the steps towards NLP annotation for machine learning (discussed further in section 11). The process of annotating a corpus and training a model is iterative, but typically involves an initial evaluation, followed by annotation by multiple annotators towards high agreement, and reevaluation throughout as the model is being created and tested. When entities can be extracted with reasonable accuracy, the project could pass to citizen science for further processing as an opportunity for public engagement.

Conversations with Zooniverse during the course of this pilot study indicate that it would be practical to create a citizen science project to crowdsource the extraction of data and metadata from the BHL texts. The project could involve a workflow in which text would be pre-annotated with taxonomic and geo-names, and citizen scientists would verify/correct this information and indicate relationships between places and species. After "collecting a specimen", a button could be clicked to visualize the occurrence on a map in relation to other occurrences in GBIF[28]. Working with the PDF files could be complicated, particularly for the older texts, so this issue

[28] This workflow is somewhat analogous to the Astronomy Rewind project in Zooniverse, which is extracting astronomical images and associated coordinates from pages of old journal articles, and allowing users to then visualize the extracted images on the sky using the WorldWide Telescope: https://www.zooniverse.org/projects/zooniverse/astronomy-rewind

would need to be considered carefully, and perhaps could involve the expertise and/or development of Plazi/GoldenGATE and Pelagios.

In considering a possible budget for the more complex workflow tentatively envisioned here, the following categories should be included for a multi-year effort:

- Project manager (NLP expert, perhaps affiliated with a university research group, 2+ months/year to supervise the project)
- Trained annotators (perhaps university students)
- Software/web developer(s)
- Librarian (perhaps partial FTE)
- Infrastructure needs

Assuming that a 10-page BHL document takes one hour or more to annotate in brat (where GeoNames references must be looked up and added manually for disambiguation), a minimum of 3 annotator hours would be needed to complete a single document, at approximately $12/hour. At this time, it is uncertain how many documents would be necessary for the proposed project, particularly considering the diversity of the BHL corpus. Tools such as Recogito, the Edinburgh Geo-annotator, and GeoAnnotate could speed up the annotation process. A subscription to Prodigy could also be cost-effective (group licenses cost $490/seat in packages of 5 seats)

Zooniverse developer costs depend on the project's customization needs, but resources should be available for a BHL developer or data scientist to prepare and upload the data (either manually or creating an automated pipeline into Zooniverse via the Panoptes API[29]). Resources must also be allocated to work with the results from Zooniverse and aggregate them back into BHL and GBIF systems. Furthermore, a Zooniverse project requires the engagement of a science team to promote the project and interact with citizen scientists regularly on the dynamic Zooniverse message boards. Finally, Zooniverse recommends dedicating time for researchers to study the volunteer experience itself and the online citizen science community the project creates - for example, whether certain features motivate more classifications, and what the learning outcomes are for participants.

6 Conclusion and Summary of Recommendations

The project explored by this pilot study could result in potentially useful biodiversity data if implemented fully. Furthermore, efforts towards a full project could have implications for additional discoveries – text mining in GeoDeepDive to answer research questions devised by other researchers or disciplinary communities, for example, and/or broad advances in the NLP field by developing new training data and models for historical documents, which could also

[29] https://github.com/zooniverse/panoptes-cli

have cross-cutting impact. In trying and testing a wide variety of software platforms, this project has also encountered with the ephemeral nature of technology. Developments occur rapidly and are as quickly replaced by new developments, where documentation and software dependencies sometimes fall by the wayside in the process (an advantage of Docker[30] and other newer initiatives like Code Ocean[31] and the Open Science Framework[32]). In designing a socio-technical project such as this, it is necessary to make a decision whether to push forward and participate in the evolution, or wait for a more ideal time and better technology.

Considering these circumstances and the background research presented here, for reasons outlined in detail above, **the overall recommendation of this report is to begin with a workshop or meeting to obtain community feedback and stakeholder engagement, and to brainstorm with a group of experts the methods, workflows, funding streams and other support, and targeted areas of the collection towards which to focus initial and longer-term information extract efforts.**

Other **recommendations summarized from the above discussion** are:

- **Use machine learning and/or NLP techniques** to create customized NER classifiers or dictionaries for searching and tagging clusters of similar BHL documents;
- **Strongly consider citizen science** as an option for engaging the public with "virtual specimen" collection through crowdsourced verification and correction of tagged entities and relationships;
- **Work with stakeholders** who are already involved in aspects of the linked data challenge - such as GBIF, Plazi, and Pelagios – to develop resources necessary to accomplish the project; and
- **Make annotations public** as open data to support the NLP community.

Products accompanying this report are detailed below in the Appendices.

[30] https://www.docker.com/
[31] https://codeocean.com/
[32] https://osf.io/

7 References

Alex, B., & Burns, J. (2014, May). Estimating and rating the quality of optically character recognised text. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage* (pp. 97-102). ACM.

Batista-Navarro R, Nguyen N, Soto A, Ulate W, Ananiadou S (2017) Argo as a platform for integrating distinct biodiversity analytics tools into workflows for building graph databases. Proceedings of TDWG 1: e20067. https://doi.org/10.3897/tdwgproceedings.1.20067

Bowker, G. C. (2000). Biodiversity Datadiversity. *Social Studies of Science*, *30*, 643–683. doi:10.1177/030631200030005001

Cardoso, S. D., Serique, K. J., Amanqui, F. K., Dos Santos, J. C., & Moreira, D. A. (2014, June). A gazetteer for biodiversity data as a linked open data solution. In *WETICE Conference (WETICE), 2014 IEEE 23rd International* (pp. 435-440). IEEE.

Cura, R., Dumenieu, B., Perret, J., & Gribaudi, M. (2017). Historical collaborative geocoding. *arXiv preprint arXiv:1703.07138*. https://arxiv.org/pdf/1703.07138.pdf

Davis, M. L. S., Tenopir, C., Allard, S., & Frame, M. T. (2014). Facilitating Access to Biodiversity Information: A Survey of Users' Needs and Practices. *Environmental management*, *53*(3), 690-701.

DeLozier, G., Wing, B., Baldridge, J., & Nesbit, S. (2016). Creating a novel geolocation corpus from historical texts. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)* (pp. 188-198).

Faulwetter, S., Pafilis, E., Fanini, L., Bailly, N., Agosti, D., Arvanitidis, C., ... & Georgiev, T. (2016). EMODnet Workshop on mechanisms and guidelines to mobilise historical data into biogeographic databases. *Research Ideas and Outcomes*, *2*, e9774.

Finkel, J.R., Grenager, T., and Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005),* pp. 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf

Finlayson, M. A., & Erjavec, T. (2017). Overview of Annotation Creation: Processes and Tools. In *Handbook of Linguistic Annotation* (pp. 167-191). Springer, Dordrecht.

Freeland, C., Kalfatovic, M., Paige, J., & Crozier, M. (2008). Geocoding LCSH in the Biodiversity Heritage Library. *The Code4Lib Journal*, (2), 11.

Gabud R, Yap S, Batista-Navarro R, Ananiadou S (2017) Developing a knowledge base on the habitats and reproductive conditions of Dipterocarps through information extraction. Proceedings of TDWG 1: e20066. https://doi.org/10.3897/tdwgproceedings.1.20066

GBIF: The Global Biodiversity Information Facility (2018) *What is GBIF?*. Available from https://www.gbif.org/what-is-gbif [1 September 2018].

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., & Ball, J. (2010). Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *368*(1925), 3875-3889.

Heidorn, P. B., Stahlman, G. R., & Steffen, J. (2018). Astrolabe: Curating, Linking, and Computing Astronomy's Dark Data. *The Astrophysical Journal Supplement Series*, *236*(1), 3

Heidorn, P. B., & Zhang, Q. (2013). Label annotation through biodiversity enhanced learning. *iConference 2013 Proceedings* (pp. 882-884). doi:10.9776/13450

Keeney, S., Hasson, F., & McKenna, H. P. (2001). A critical review of the Delphi technique as a research methodology for nursing. *International Journal of Nursing Studies*, *38*(2), 195–200. doi:10.1016/S0020-7489(00)00044-4.

Klein, E., Alex, B., Grover, C., Tobin, R., Coates, C., Clifford, J., ... & Fieldhouse, I. (2014). Digging Into Data White Paper: Trading Consequences.

Kumar, A. (2016). A survey on various OCR errors. *International Journal of Computer Applications*, *143*(4), 8-10.

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., ... & Murray, P. (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, *389*(3), 1179-1189.

Mugar, G., Østerlund, C., Jackson, C. B., & Crowston, K. (2015, June). Being present in online communities: learning in citizen science. In *Proceedings of the 7th International Conference on Communities and Technologies* (pp. 129-138). ACM.

Nguyen, N. T., Soto, A. J., Kontonatsios, G., Batista-Navarro, R., & Ananiadou, S. (2017). Constructing a biodiversity terminological inventory. *PloS one*, *12*(4), e0175277.

Page, R. D. (2013). BioNames: linking taxonomy, texts, and trees. *PeerJ*, *1*, e190. Describes development of the BioNames project, which links publication references (full text where possible) and taxonomic data across a diversity of resources and platforms.

Page, R. D. (2011). Extracting scientific articles from a large digital archive: BioStor and the Biodiversity Heritage Library. *BMC bioinformatics*, *12*(1), 187.

Page RDM (2008) Biodiversity informatics: the challenge of linking data and the role of shared identifiers. *Brief Bioinform 9*: 345–354).

Page, R.D.M. (2018). *GBIF georeferencing*. Retrieved from https://glitch.com/edit/#!/lyrical-money?path=README.md:1:0

Paul, D., & Heidorn, P. B. (2013). Augmenting optical character recognition (OCR) for improved digitization: Strategies to access scientific data in natural history collections. *iConference 2013 Proceedings*, Fort Worth, Texas (pp. 514-518). doi:10.9776/13266 http://hdl.handle.net/2142/39427

Paul, D., Heidorn, P. B., Best, J., Gilbert, E., Neill, A., & Ulate, W. (2013). Help iDigBio reveal hidden data: iDigBio Augmenting OCR working group needs you -Part II. *iConference 2013 Proceedings* (pp. 1066-1068). doi:10.9776/13517. http://hdl.handle.net/2142/42515.

Pustejovsky, J., & Stubbs, A. (2012). *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.".

Shmanina, T., Zukerman, I., Yepes, A. J., Cavedon, L., & Verspoor, K. (2013). Impact of corpus diversity and complexity on ner performance. In *Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013)* (pp. 91-95).

Sokolova, M., & Bobicev, V. (2018). Corpus Statistics in Text Classification of Online Data. *arXiv preprint arXiv:1803.06390*.

Sørensen, J. J. W., Pedersen, M. K., Munch, M., Haikka, P., Jensen, J. H., Planke, T., ... & Sherson, J. F. (2016). Exploring the quantum speed limit with computer games. *Nature*, *532*(7598), 210.

Thompson, J.N., Reichman, O.J., Morin, P.J., Polis, G.A., Power, M.E., Sterner, C.A. … Strauss, S.Y. (2001). Frontiers of ecology. *Bioscience, 51* (1), 15-24. http://bio.research.ucsc.edu/people/thompson/PublPDFs/095Frontiers.pdf

Weissenbacher, D., Tahsin, T., Beard, R., Figaro, M., Rivera, R., Scotch, M., and Gonzalez, G. (2015). Knowledge-driven geospatial location resolution for phylogeographic models of virus migration, *Bioinformatics*, Volume 31, Issue 12, 15 June 2015, Pages i348–i356, https://doi.org/10.1093/bioinformatics/btv259

Xue, K. (2014). Popular Science. *Harvard Magazine, 116*(3), 54-59.

8 Appendix A: Toponym Extraction and Visualization

As an example of toponym extraction and visualization, a Jupyter Notebook was created to display placenames extracted from the same BHL document[33] with Leaflet using three methods: human annotations with brat (requiring manual look-up of coordinates in GeoNames), human annotations in Recogito (with pre-annotated placenames and visual GeoNames interface for correction of locations), and automatic extraction with Mordecai. Python was used to interface with Mordecai and to automatically extract coordinates in a .csv file. Figures 1-3 show images of the three maps. Code for this demonstration can be found through the project's GitHub: https://github.com/LEADS-BHL

Figure 1: Place names extracted with human annotations in Recogito

[33] The BHL document used for this example is:
https://www.biodiversitylibrary.org/item/19764#page/7/mode/1up

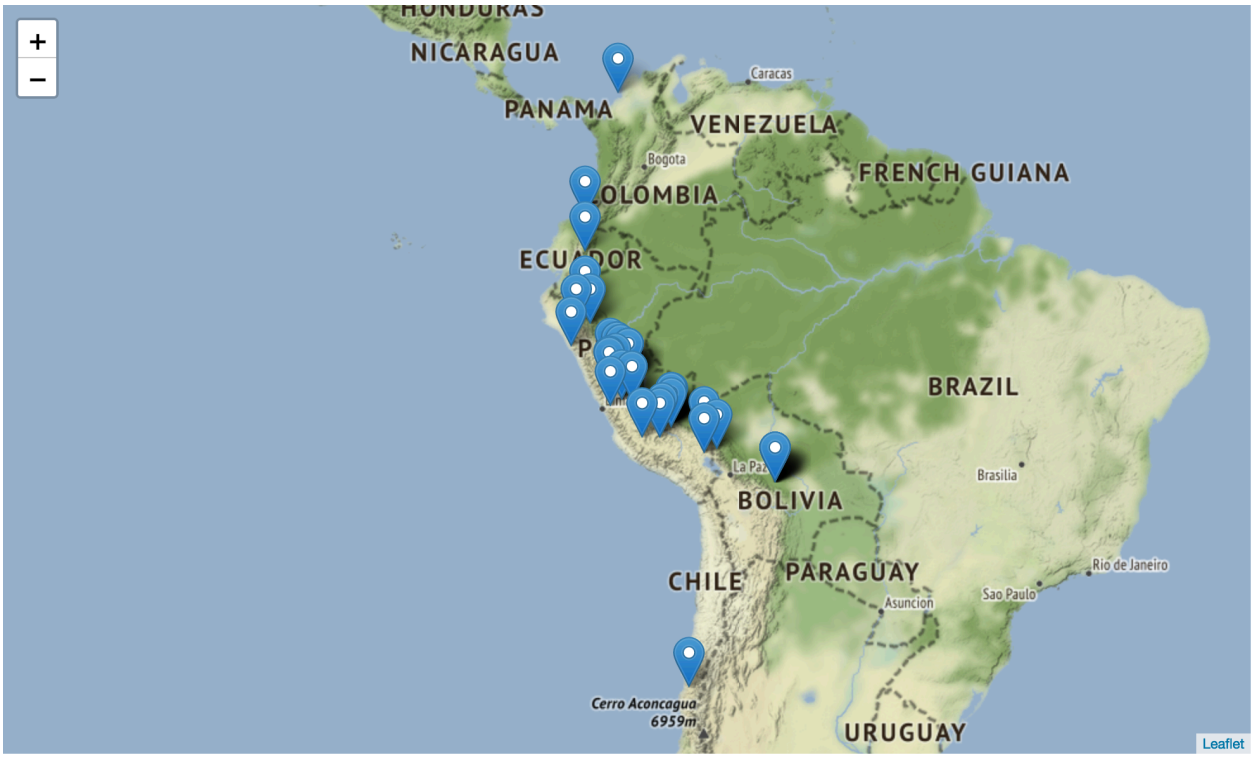

Figure 2: Place names extracted with human annotations in brat

Figure 3: Place names extracted automatically with Mordecai

Figure 4: Screenshot of Recogito map view

9 Appendix B: GeoDeepDive Example

A quick example was created using the GeoDeepDive API to search for “NMNH” and “USNM” references in the entire GDD corpus. The standard GDD processing with Stanford CoreNLP includes named entities, so the detected locations were extracted (with a small 100-article limit), along with coordinates, in a .csv file listing latitude, longitude and publication name. Code for this demonstration can be found in Jupyter Notebook through the project’s GitHub: https://github.com/LEADS-BHL

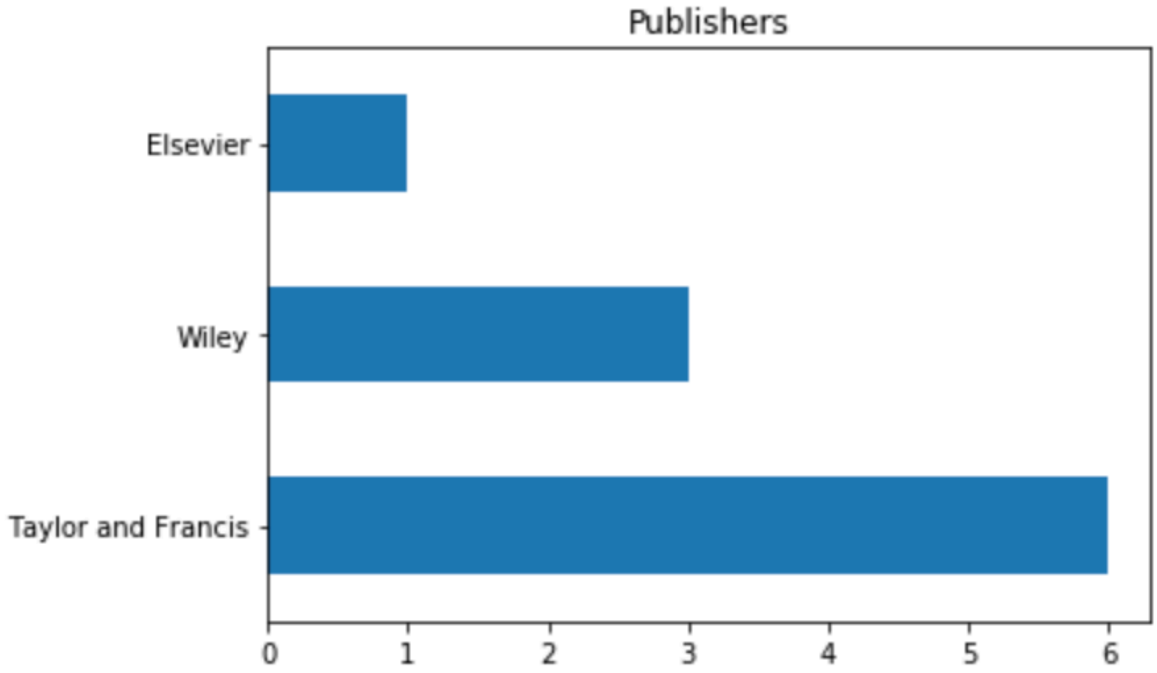

Figure 4: Top publishers of articles mentioning NMNH and USNM

10 Appendix C: OCR Quality Evaluation Software

Given the varying quality of Optical Character Recognition in BHL documents, a method of automatically assessing OCR quality is desirable, to assess which documents require additional human effort prior to machine processing. Email communication with Plazi on this topic resulted in the following possible strategy (summarized from a July 16 email exchange):

A dictionary can be created using the corpus itself. Word tokens appearing 5 or more times can be assumed to be correct. A self learning dictionary can be extended with the known-correct tokens. Work tokens appearing twice are likely correct and can be added to the dictionary. Users can manually verify word tokens appearing only once. The OCR text can be compared to this dictionary to identify tokens that are not words and to assess the quality of the text.

For demonstration purposes, an outside software package was located in GitHub (and forked to the page for this project[34] at https://github.com/LEADS-BHL). The software compares tokens to dictionary matches and was tested with a selection of the test corpus. A generic dictionary was used to test, but a custom dictionary as described above could provide greater accuracy. As an example, the output for one document looks like this:

> Evaluating /Users/gretchenstahlman/Documents/Gretchen's Files/Geoparsing/BHL Geoparsing - Test Corpus/Full text/Expanding Access/americanflowerga1852buis_djvu.txt
> Attempting token matching
> Generating Tokens, breaking words at hyphens
>
> 2751 items discarded
> 1115 hyphenated line fragments fused
> 115326 tokens processed
> **11822 total capitalized tokens**
> **9743 total capitalized dictionary matches**
> 0 total capitalized valid substitutions
> **103504 total lower-case tokens**
> **101229 total lower-case dictionary matches**
> 10 total lower-case valid substitutions

[34] https://github.com/LEADS-BHL/ocreval

11 Appendix D: Annotation Guidelines and Test Corpus

The test corpus used for this project (and corresponding annotations) can be found in this project's GitHub: https://github.com/LEADS-BHL.

For annotating in brat, the annotation guidelines of the Arizona State University phylogeography project were adopted (see http://diego.asu.edu/). All geographic annotations included a separate note with GeoNames ID and coordinates where possible. To summarize:

- Named geographic locations should be tagged.
- Abbreviations of states and countries should be tagged.
- Indirect mentions of places should not be tagged.
- Where there is ambiguity about whether the text refers to for example - a capital of semi-political entity, the less specific location should be chosen.
- Phrases referring to multiple locations should not be annotated.
- Official names should be annotated entirely.
- Labels after an entity name such as "County" or "Province" should not be tagged.
- Descriptive words such as "northern" or "southern" should not be tagged.
- Latitude and longitude references in the text should be tagged.
- Street names should be tagged, but numbers and postal codes should be not.
- Terms that cannot be specifically reference as named locations in GeoNames (such as "in the mountains") should not be tagged as toponyms.
- Adjectives such as "American" should not be tagged.
- Names of organizations and locations that appear within organization names should not be tagged.

To this we add:

- General terms that give orienting information about a place should be tagged as "Other". UA iSchool postoc Egoitz Laparra is working on a project to disambiguate general references in text such as "grassland regions of the prairie provinces"; "in the interior of British Columbia"; and "on the Alberta side of West Reflex Lake".
- Taxonomic names should be labeled.
- Type specimens should be labeled.
- Relations should be made between type specimens and taxonomic names.
- Relations should be made between taxonomic names and locations.
- Because brat often encounters errors with annotations spanning multiple lines, it is best to annotate only up until a line break, and include a note with the full reference.
- Misspellings should include a note "misspelling".
- There is no mechanism for copying annotations to other instances of the same reference. It is a good idea to keep a notepad with information about common place names.

- Colloquial/common names should not be annotated; only scientific names.

A summary of annotation best practices is adapted here from Finlayson & Erjavec (2017) and overall derived from Pustejovsky & Stubbs (2012):

The motivating question for this project is: Can NLP and machine learning be used to facilitate automatic extraction and parsing of references to geographic places in the text of BHL publications?

The current effort presented in this report represents a "version 1" corpus; in other words, a first pass at annotation scheme specification and annotation guidelines. Eventually, a gold standard corpus should be created, with inter-annotator agreements (IAA calculations), tools for text evaluation and correction, and a processing pipeline.

Moving forward, the "MATTER" Framework should be followed:

M: Model
Preliminary evaluation to develop a model, characterize issues, provide recommendations.

A: Annotate
Update model, normalize artifacts and remove typos, write guidelines for annotation, define skills, etcetera.

T: Train&
T: Test
In this case, prepare data for others to use.

E: Evaluation
Get a sense of data quality and scope.

R: Revise
Repeat as necessary.

12 Appendix E: Annotation Environment

For collaborative annotation, a brat annotation environment (with the test corpus included as pre-loaded data) was configured using a virtual machine within CyVerse cyberinfrastructure. To access the virtual machine, it is necessary to create an account and obtain access to the "Atmosphere" component:

https://user.cyverse.org/
https://atmo.cyverse.org/application

For detailed instructions, CyVerse tutorials are located here:

https://wiki.cyverse.org/wiki/dashboard.action

Access to the "image" of this preconfigured system is currently restricted to LEADS-BHL project members.

Once an account is created and access is granted, search for and start the image "Geoparsing Annotation" using the default configuration.

Once launched, select "Open Web Desktop" to access a visual desktop interface to the VM in your browser.

Open a terminal window, and run the following commands using "sudo":

```
cd /var/www/html/brat-v1.3_Crunchy_Frog/
python standalone.py
```

Copy the address returned by the final command ("serving at […]")
Open an internet browser and enter the address to access the locally-hosted brat interface.

Login: editor
Password: annotate

After annotating, to share annotated data with collaborators, enter the following command into a terminal window:

$ cyverse_backup.sh

Respond to the following questions according to these examples:

Do you want to backup or restore local instance/volume data?[backup/restore]: backup

Specify local file/directory's full path on the instance [eg: /home/gstahlman,

/home/gstahlman/file.txt]: /var/www/html/brat-v1.3_Crunchy_Frog/data/

Specify file/directory path on iPlant datastore [eg: /iplant/home/gstahlman, /iplant/home/gstahlman/file.txt][Default is /iplant/home/gstahlman/atmosphere] : /iplant/home/gstahlman/atmosphere

Enter CyVerse password, hit enter

The file/folder already exists on the datastore. Do you want to overwrite?[yes/no]: yes

Enter your current iRODS password:

Navigate to Data Store, share with collaborators

Email gstahlman@email.arizona.edu with questions.

13 Appendix F: Other Potentially Useful Software

Other outside software that could be applicable has been forked into this project's GitHub: https://github.com/LEADS-BHL

14 Appendix G: List of Meetings

Smithsonian

- Luis Villanueva, Smithsonian Informatics Program Officer
- Rosa Lin and Arjun Kumar, Tolstoy AI
- Nhung Nguyen, National Centre for Text Mining
- Effie Kapsalis (Chief of Content & Communication Strategy), Adam Soroka (Senior Architect, Research Computing), and Beth Stern (Enterprise Architect), Smithsonian Office of the CIO
- Dmitry Schigel, Global Biodiversity Information Facility (GBIF)
- Rebecca Dikow (Research Data Scientist), Smithsonian Office of the CIO
- Bianca Crowley and Grace Constantino, BHL Secretariat
- BHL Tech Team
- David Thau, Google Earth Engine (email discussion only so far)
- Donat Agosti, President, Plazi (email discussion only so far)
- Adam Mansur (IT Specialist/Data Manager) Smithsonian Department of Mineral Sciences (email introduction, scheduling meeting)
- Carolyn Sheffield and Martin Kalfatovic (weekly mentor check-ins and other virtual meetings since May 14, 2018)

U of A

- Bryan Heidorn, School of Information Director
- Steven Bethard, School of Information (NLP)
    - Vikas Yadav (PhD Student)
    - Egoitz Laparra (Postdoctoral Researcher)

Other

- Rosa Lin and team (Tolstoy AI)
- Daniel Wieferich and Brandon Serna (USGS)

At GeoDeepDive

- Brandon Serna, USGS
- Jo Walsh, British Geological Survey
- Shanan Peters, GeoDeepDive PI

15 Appendix H: Annotated Literature Review

Methods/Techniques

Cardoso, S. D., Serique, K. J., Amanqui, F. K., Dos Santos, J. C., & Moreira, D. A. (2014, June). A gazetteer for biodiversity data as a linked open data solution. In *WETICE Conference (WETICE), 2014 IEEE 23rd International* (pp. 435-440). IEEE.

Describes Brazilian researchers' effort to create an open gazetteer with superior recall for adding coordinates to species occurrence records. Gazetteer could not be found online, but perhaps the researcher group could be contacted directly.

Cura, R., Dumenieu, B., Perret, J., & Gribaudi, M. (2017). Historical collaborative geocoding. *arXiv preprint arXiv:1703.07138*. https://arxiv.org/pdf/1703.07138.pdf

Extracts information from historical maps of Paris and matches to current addresses.

Davy Weissenbacher, Tasnia Tahsin, Rachel Beard, Mari Figaro, Robert Rivera, Matthew Scotch, Graciela Gonzalez; Knowledge-driven geospatial location resolution for phylogeographic models of virus migration, *Bioinformatics*, Volume 31, Issue 12, 15 June 2015, Pages i348–i356, https://doi.org/10.1093/bioinformatics/btv259

Phylogeology tracks spread of viruses with genetic sequence data related to time of sample collection and location of host. Geolocation can be used to recreate virus migration path. Developed automated method to retrieve geospatial metadata from full-text articles. Toponym resolution in two steps: detection (finding place names) and disambiguation (assigning coordinates). Dictionary-based and rule-based heuristics used. GeoNames dictionary detects mentions of places. Physical properties (population and distance heuristics) and context (lexical content) of references inform disambiguation. Gold standard corpus for toponym resolution on phylogeography texts was created. Lack of common frameworks and few standard copora hinder progress on toponym resolution. Also, gazetteers do not have standard practices for assigning coordinates to names of places. Also, variation in corpora makes it difficult to evaluate heuristics. Prior efforts include Tamames and de Lorenzo (2010), searching for locations of bacterial habitats. Developed corpus of 50 full text and 200 abstracts, performed toponym detection with GeoNames and Google Maps with capitalized words in noun phrases. Then trained a classifier to distinguish environmental references/context. Then disambiguated by searching for locations in lexical contexts and then querying GeoNames. Created an index of GeoNames for efficiency (with tokens and clusters, noisy entries blacklisted, filtered out acronyms and names). Preprocessed documents using ANNIE sentence splitter (gate.ac.uk) and Genia for POS tagging and chunking (n-actem.ac.uk/tsujii/GENIA/tagger/). XML tag substituted for toponym in document and all possible coordinates are inserted for further disambiguation. Geospatial metatdata in GenBank improves toponym resolution (metadata heuristic/rules). 60 articles annotated from sample of 598, associated with 5730 GenBank records. Output of toponym detector formatted to be compatible with Brat annotator (brat.nlplab.org). Manually

corrected and disambiguated toponyms, removed toponyms not related to phylogeography. If toponym not found in GeoNames, used Google Maps and Wikipedia. Annotation guidelines developed prior to beginning. Gold standard has 379 toponyms for 1881 occurrences. Prioritized recall over precision. Analyzed errors of toponym detector by hand using random selection. Leads to problems with toponym resolver. Similarly evaluated performance of disambiguator.

DeLozier, G., Wing, B., Baldridge, J., & Nesbit, S. (2016). Creating a novel geolocation corpus from historical texts. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)* (pp. 188-198).

Good example of toponym resolution with historic documents. Developed GeoAnnotate tool for annotation. Document geolocation and toponym resolution are separate but related tasks – difference is size of span of text. Toponym resolution is for potentially ambiguous toponyms while document geolocation is for larger spans of text. Toponym resolution suffers from lack of training material. Domain adaptation is sometimes used, Wikipedia for example. OCR was hand-corrected for data preparation. Preprocessing stitched up page breaks and removed headers and footers (program was written to do majority of this, but also required hand effort). GUI tool created so annotators could indicate spans of document and document locations on a map. Sequence model automatically split text into documents, which were trained on the manually annotated documents. GUI tool: points and polygons can be added by drawing on a map or entering coordinates. Annotators encouraged to make use of a single point vs a polygon if possible (allowed for more annotations). Place names must be relevant to subject. Annotators were encouraged to look up toponym coordinates, for example in Wikipedia. Geographic theme of documents was difficult to determine due to many place name mentions. System was designed for a single point; centroids taken to handle multiple point documents. References to places often ad-hoc or in reference to previously mentioned locations. For toponym annotation, a subset of volumes annotated with document geolocations were selected. Stanford's NER was run, producing a pre-annotated set, and annotators corrected and added. 1,644 annotations across 100 page subsections of 15 volumes. All originally done by a single annotator, then team of three annotators corrected problems. Annotators given guidelines, asked to quickly scan for place names and include point, multi-point, polygon, and multi-polygon geometries where appropriate. Excluded metonymic and demonymic names from annotation. Only selected toponym encompassing named entity (ie exclude organization mentioning place in name). Toponym hierarchies marked as separate entities (ex Richmond, VA, CSA). Challenges included: conjunctive toponyms; possessive toponyms; difficult toponyms; rivers and physical features (mark only points on rivers most relevant or shown on Wikipedia); geographically vague regions (attempted to draw reference given context); referring expressions (tried to reference whole expression, but often marked embedded city point); embedded named entities ("Cavalry of Virginia" vs "Cavalry from Virginia", VA only annotated in the latter). Performance evaluated for document geolocation and toponym resolution using different statistical methods and comparing (SPIDER, TopoCluster, TopoClusterGaz, etc).

Gazetteer dependent systems were less successful making predictions. This corpus presents innovative challenge and is the first published document geolocation corpus for historical text. Also shows joint reference marked with toponym and docgeo annotations.

Faulwetter, S., Pafilis, E., Fanini, L., Bailly, N., Agosti, D., Arvanitidis, C., ... & Georgiev, T. (2016). EMODnet Workshop on mechanisms and guidelines to mobilise historical data into biogeographic databases. *Research Ideas and Outcomes*, *2*, e9774.

Report of workshop exploring utility of GoldenGATE-Imagine for annotating and extracting treatment data from PDF texts. The system permits markup of taxonomic mentions and locations, then retrieving Darwin Core Archives for exporting treatments and publishing of metadata. Overall goal is to enable tracking of species occurrence over time. Obstacles include errors in the text leading to extensive human effort, inconsistent distribution of information in the publications, and unstructured and mixed information, OCR quality and typography, complexity of natural language references in text, among others. Location information is also a challenge: "Thus, location information cannot always be checked against commonly used gazetteers, and to resolve these names and georeferenced them, additional, tedious and time consuming research into other literature is often needed. In other cases, coastlines or river courses may have changed over the decades or centuries, and given coordinates of a marine expedition may now fall on land and not in the sea" (p. 9). In evaluating semi-automated annotation, the workshop concludes that GoldenGATE-Imagine could be developed specifically for occurrence extraction, representing potential for new contributions of taxonomic names not currently recognized by GBIF. However, the process is currently tedious and prone to errors, though tools exist that could assist in mobilizing data more easily.

Leetaru, K. H. (2012). Fulltext geocoding versus spatial metadata for large text archives: Towards a geographically enriched Wikipedia. *D-lib Magazine*, *18*(9), 5. http://www.dlib.org/dlib/september12/leetaru/09leetaru.html

Presents workflow of full text geocoding (also known as geoparsing or GIR/Geographic Information Retrieval) and overview of GNS and GNIS gazetteers, and a case study of Wikipedia. Not practical to use human editors to assign geographic tags. Available commercial tools include MetaCarta and Yahoo Placemaker, but systems are black boxes. First stage is to parse and identify words and phrases that may be georeferences. More sophisticated is POS tagging. Algorithm eliminates all information that can't be a location (common words compared to gazetteers). Country-specific gazetteers provide high resolution coverage. Common gazetteers are Alexandria, Getty, GeoNames, GNS and GSGS GNIS. Final stage is disambiguation and confirmation of candidates from earlier stages. Linguistic context is important. ML categorizes text as containing location references (large computational requirements and per feature training). Centroid ranking is a method. Problematic if locations in a country share the same name. Also differences in spelling. Matches with three or fewer characters are usually false positives and should be on black list. Human entered geotags in Wikipedia found to be biased, and more even coverage was found with automated methods.

Xu, B., & Fan, G. (2015, September). Multimodal topic modeling based geo-annotation for social event detection in large photo collections. In *Image Processing (ICIP), 2015 IEEE International Conference on* (pp. 3319-3323). IEEE.
https://ieeexplore.ieee.org/document/7351418/

Estimating geographic coordinates in digital media where GPS information is not available – social event detection with multimodal clustering algorithm. Measures semantic similarity of images with temporal, spatial, visual and textual information. Algorithm propagates geographic information from geo-tagged photos to geo-missing photos.

Mindat.org
https://www.mindat.org/

Highlights a shared problem across disciplines. This database serves the mineral science community, similar to GBIF. A recent talk at the GeoDeepDive workshop describes georeferencing of mineral observations in text. Notes from the talk are copied below:

David Von Bargen (MinDat): Rock & Mineral Entity Recognition

Extract mineral and rock names from text - mindat.org
Wiki-based program with database in the back end
Eclectic user base - professionals and amateurs

Four databases: mineral, locality (with 300k locations), mineral/locality database (1.2 million entries), photographic database (900k photos)

Adding rocks and minerals to locality page
"Locality Editor"

Mineral names database contains 51k entries (minerals, rocks, varietal, synonyms, also have weird spellings, foreign languages, etc)
Pre-processing
Then stemming and tokenization
Most minerals are singular, rocks more likely to be plural - have to take care of that
Hyphens - lots of problems
End of line (need to eliminate), separates minerals in list, compound words, part of name
Initial processing
Where array element contains hyphen, remove and create two or more words; keep original to check against
Plurals stemmed
New array with unique array elements

Problems to solve:
Match mineral and rock lists to locations
Eliminate minerals used in methods
Minerals mentioned in references
Minerals not occurring at location
Minerals that are in a group or structural related
Native elements vs chemical
Advantage: lots of articles that are structured alike

90% of database is literature constructed - goes back to reference in the literature

People add site identification

Can see how good data is, has information about how they determined it

Human constructed resources to automated workflows

Software Tools/Resources

Geograpy2 package: Extract places names and add context. Uses NLTK, jellyfish, pycountry, GeoLite2, ISO3166ErrorDictionary
https://github.com/Corollarium/geograpy2/blob/master/README.md

Creating and Visualizing Neural Network in R
https://www.analyticsvidhya.com/blog/2017/09/creating-visualizing-neural-network-in-r/

Neural networks are adaptive and are characterized by interconnected information processing units. Information is processed through a first layer, then passed to a hidden layer, then to a final layer that produces output. Connections are weighted. Example uses dataset describing properties of cereals. Dependent variable is Rating. 60% training set is used to find relationship between dependent and independent variables and test set assesses model performance. Neural network model predicts Rating.

Deep Learning with R
https://www.r-bloggers.com/deep-learning-with-r/

Introduces Keras interface for R and shows how to perform image classification.

Read and write ESRI Shapefiles with R
https://www.nceas.ucsb.edu/scicomp/usecases/ReadWriteESRIShapeFiles

R and GIS – working with shapefiles
https://www.r-bloggers.com/r-and-gis-working-with-shapefiles/

Reading shape files with maptools package, and also rgeos and sp packages. Shape files contain polygons. Can plot, transform shape files. Can check if a point falls inside a polygon.

Simplifying polygon shapefiles in R
https://www.r-bloggers.com/simplifying-polygon-shapefiles-in-r/

Some shape files more detailed than necessary, so need to simplify map.

Naïve Bayes Classifier: theory and R example
https://rpubs.com/riazakhan94/naive_bayes_classifier_e1071

Simple classifier, popular for text classification. Student data as example. Making a naïve bayes classifier to predict association. 70/30 partition for training and validation. Contingency table to check accuracy.

SemEval-2019
http://alt.qcri.org/semeval2019/index.php?id=tasks

International Workshop on Semantic Evaluation. Has Task 12: Toponym Resolution in Scientific Papers.

Babelfy
http://babelfy.org/

Graph-based approach to semantic interpretation. Entity linking and word sense disambiguation. Has API to programmatically disambiguate text. Associates words with concept or named entity.

TopoCluster
https://github.com/grantdelozier/TopoCluster

Python application. Geostatistical based approach to describing lexical variation and perform toponym resolution.

GeoSequence
https://github.com/grantdelozier/GeoSequence

Sequence modeling approach to toponym resolution. Creates variable to point to path containing observation counts.

TextGWR
https://github.com/grantdelozier/TextGWR

Python application. Text geolocation/georeferencing using geographically weighted regression. Includes utilities for geostatistics on geolocated texts.

Argo: A Web-based Text Mining Workbench
http://argo.nactem.ac.uk/

Can be used for workflow development for text analysis. Recommended by Nhung Nguyen.

National Center for Text Mining
http://www.nactem.ac.uk/software.php

Research group of Sophia and Nhung. Many possible tools and other resources here.

Edinburgh Geoparser
https://www.ltg.ed.ac.uk/software/

Package for entity recognition in British Geological Survey texts. Should be effective for international georeferences. Available on Github, but needs to be Dockerized. Requires installation of software dependencies. A recent talk at the GeoDeepDive workshop describes the project. Notes from the talk are copied below:

Joanna Walsh (British Geological Survey): Geosemantics at BGS

Issues with placename extraction from geological texts

Experimented with open source tools and techniques
BGS used to produce Geological Memoirs, long narrative reports
Few hundred documents, had 60 processed - scanned, OCRd, manually corrected, marked up as XHTML, intensive manual labor by retired librarian
Threw collection into Stanford CoreNLP
    Picking out named entities, trying to classify
Custom model for CoreNLP done a few years ago in earlier annotation project (2500 sentences produced good trained model)
Tried training Spacy with same data, results more mixed

Extracted term is linked to a linked data URL and kept in JSON document store
Have 150,000 extracted entities, many references to place
Many pitfalls, want to know how others approached same problems

Used mostly open source toolkit - Edinburgh Geoparser, extracts information and looks up place names, how they cluster together, helps to pick out right one

Problem out of the box: overlap between rock formations and place names - Geoparser would pick these out

Needed to develop a model that had the ability to recognize technical terms and locations to avoid false positives

Rather than customize Edinburgh Geoparser, instead just took out second stage - Georesolution stage
Took Stanford output, fed through Georesolver
Used Nominatim, OpenStreetMap project
Set up in-house version of Nominatim, got peculiar results
    Only had UK coverage, but found erroneous mismatches
Decided not to use all OpenStreetMap data - got false positives, went back to official UK mapping agency, with some enhancements internal to BGS
Took only top 100 street names from OpenStreetMap and added to the data

Visualized results - will be publically available in the next week or two
Shortcomings show that documents need more pre-processing
Shows clustering of references found in document, with overlap with places - more spread with false positives
Shortcomings of gazetteer coverage
False positives can be eliminated quickly - poorly prioritized places, names in references section (eliminate references) - need to minimize labor of librarians

Looking at correlations between named entities
Looking at distances inside paragraph
But often feel - is it stating the obvious?
Has been a fascinating innovation project - hope to learn more, collaborate with GeoDeepDive, to prove value of putting more work into the effort

Next Steps? Don't know where to take it
Do more with Linked Data - semantic search with knowledge graph
Richer and more subtle entity correlations
    Try training a model to classify positive and negative correlations between named entities (looking at using Spacy toolkit)
Explore real value of work for/with scientists
    Are we naive idealists?
    Need to collaborate, share training data

Work was conceived and put in motion by Rachel Heaven (reh@bgs)

Questions:

Named entity for stratigraphic terms - extension of StanfordCore?? How could others build this
    It is a cinch with sentences classified, add-on model to StanfordNER - 20 lines of code training algorithms
    Output, dropped custom model into CoreNLP server, started adding customizations

Regional ontologies in GDD not crossmapped,
Ike Nikisi-Orji - PhD student working on this

GeoDeepDive
Mansur, A. (2018). *Tracking current and historical use of the NMNH collections.* Draft 2018-05-31.
https://geodeepdive.org/
https://github.com/adamancer/speciminer

GeoDeepDive allows searching large corpus by writing custom scripts. Author developed speciminer tool. Retrieves and parses text to associate with scientific department within NMNH, then matches to catalog number and then record with defined criteria: 1. Match keywords in snipped; 2. Match keywords in document title; 3. Match topic of paper; 4. Match topic of journal. ~61,000 distinct specimen citations found, subset confidently matched to NMNH. Issues include non-standard formulations of catalog numbers, typos and OCR errors, assigning topics, missing records, incomplete corpus (bias towards recent publications, unequal coverage of disciplines). Speciminer works reasonably well. Strong matches should be considered for integration into NMNH collections database. Documentation of provenance of citation record will be important.

GeoDeepDive Workshop

Workshop was held in Madison, WI on August 20-21, 2018. Detailed notes from workshop talks are available here (including several discussions of automatically extracting place references from text and linking to other types of data): https://docs.google.com/document/d/1DBg1q0XfVA5_umf2wZDhTg5za_SE8r6pM3r5cvn8Xvs/edit?usp=sharing

PROTAX

https://www.helsinki.fi/sites/default/files/atoms/files/protaxusermanual.pdf

Bayesian process for classification. Not applicable.

Leaflet

https://leafletjs.com/

https://github.com/muziejus/leaflet-quickstart/blob/master/README.md

JavaScript library for interactive map tools

Metadata

DCMI Box Encoding Scheme: specification of the spatial limits of a place, and methods for encoding this in a text string

http://www.dublincore.org/documents/dcmi-box/

Encoding scheme for identifying region of space using geographic limits and representing as a value string.

Gazetteers/Ontologies

GeoNames

http://www.geonames.org/

Database of geospatial locations with API

socrata.com dataset - Country List ISO 3166 Codes Latitude Longitude

https://opendata.socrata.com/dataset/Country-List-ISO-3166-Codes-Latitude-Longitude/mnkm-8ram

Global Names Recognition and Discovery
http://gnrd.globalnames.org/api

Finds scientific names. Has API.

GeoLocate
http://www.geo-locate.org/developers/default.html

Tool for georeferencing collections.

Google Geocoding API
https://developers.google.com/maps/documentation/geocoding/start

Converts addresses into geographic coordinates. Also reverse geocoding – converts coordinates into human readable addresses for mapping.

Annotation

Pustejovsky, J., & Stubbs, A. (2012). *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.".

Comprehensive guide to annotation. Includes MATTER cycle for project design and implementation. Discusses standards and models.

Open Annotation Data Model
http://www.openannotation.org/spec/core/

Interoperable framework for creating associations

brat standoff format
http://brat.nlplab.org/standoff.html

Annotations stored separate from document text (for each document a corresponding annotation file). Brat can perform sentence segmentation whether or not there are newlines. General structure - each line contains one annotation with an ID, ex. entity, event trigger, event, relation. Text-bound annotations – same structure, continuous and discontinuous sentence structure supported. Also includes attribute and modification annotations, normalization annotations, and note annotations.

Annotation Tools

Dataturks
https://dataturks.com/

All data uploaded must be open.

Pelagios Recogito
https://recogito.pelagios.org/

Semantic annotation - draws connections between people, places, things, events. Can keep data in private workspace. Linked data format. Available gazetteers are: Pleiades

Gazetteer of the Ancient World, Digital Atlas of the Roman Empire, a subset of GeoNames, China Historical GIS. Integrates Stanford CoreNLP for NER. Should cite if used in publications. Open source and can be downloaded from GitHub: https://github.com/pelagios/recogito2. Pelagios how-to guide: https://github.com/pelagios/pelagios-cookbook/wiki

GeoAnnotate
https://github.com/utcompling/GeoAnnotate

Javascript application to collect toponym and document level annotations. Need to set up local web server and install Parse backend with MongoDB.

brat rapid annotation tool
http://brat.nlplab.org/

Online environment for text annotation

GoldenGate
http://plazi.org/resources/treatmentbank/goldengate-editor/

Imagine markup editor for PDFs. Can propose to call GeoLocate and return record with pointer to place name, species and spatial information.

Corpora

ASU Diego Lab
http://diego.asu.edu/index.php?downloads=yes

Toponym Resolution Resources (includes annotated corpus) and other projects. Toponym Resources download includes Annotation Guidelines document and 60 annotated PubMed articles.

War of the Rebellion
https://github.com/utcompling/WarOfTheRebellion

Annotated corpus of Civil War archives. Consists of toponym corpus (JSON test-train, and simplified XML to be fed into Toponym Resolution systems) and document-geolocation corpus (JSON files). TR systems of Topocluster and Fielspring required simplification of geojson geometry to a centroid.

BioNLP-Corpora
http://bionlp-corpora.sourceforge.net/

Includes Annotation Schema document.

Biodiversity Literature and Data

Freeland, C., Kalfatovic, M., Paige, J., & Crozier, M. (2008). Geocoding LCSH in the Biodiversity Heritage Library. *The Code4Lib Journal*, (2), 11.

Library of Congress subject headings are a source of geographic information about BHL documents. To explore new services for BHL users, subject headings are used to geocode documents with the subfield $z (geographic subdivision) and visualized with Google Maps API to request coordinates in KML format. Latitude and longitude are stored in the BHL database. An issue with this workflow is that LCSH values are only as complete as the BHL partners' cataloging practices. Also, point locations are returned for polygons as centroids, which can create confusion with placemarks for city and state. User studies are needed to explore future services for spatially browsing collections.

Guala, G. F. (2016). The importance of species name synonyms in literature searches. *PloS one*, *11*(9), e0162648.

Describes ITIS web service for including synonyms in literature searches.

Nguyen, N. T., Soto, A. J., Kontonatsios, G., Batista-Navarro, R., & Ananiadou, S. (2017). Constructing a biodiversity terminological inventory. *PloS one*, *12*(4), e0175277. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0175277

Presents text mining methods for biodiversity term inventory. Applied to all of BHL. Can identify semantically similar names that are not in existing taxonomies; can be used to update taxonomies. Also improving search. And developed visual interface.
Catalog of Life contains more than 1.6 million species names. GBIF has data related to occurrences of species. GNA promotes interoperability between taxonomies. BHL holds a vast collection, needs automated methods for processing textual information. Many tools have been developed for facilitating text based analysis of taxonomic names. Distributional Semantic Models (DSMs) computed relatedness based on models different from knowledge based approaches, encoding lexical context as a vector and calculating similarity between terms. Prediction based DSMs are used for the first time to compile large scale terminological resources automatically. English-language data obtained with BHL API. 20 biologists assessed usefulness of the interface, leading to query enhancements.

Page, R. D. (2013). BioNames: linking taxonomy, texts, and trees. *PeerJ*, *1*, e190.

Describes development of the BioNames project, which links publication references (full text where possible) and taxonomic data across a diversity of resources and platforms.

Thessen, A. E., & Parr, C. S. (2014). Knowledge extraction and semantic annotation of text from the encyclopedia of life. *PloS one*, *9*(3), e89550.

Biodiversity research is challenged by data heterogeneity, distributed nature of data, and the need to consider the entire body of knowledge (Page RDM (2008) Biodiversity informatics: the challenge of linking data and the role of shared identifiers. *Brief Bioinform 9*: 345–354). Paper describes annotation workflow for identifying taxonomic references and using algorithms to extract information into machine readable formats. JSON text was parsed and cleaned with Beautiful Soup python library, then passed to GlobalNames. Human annotators verified automatic extraction.

Wei, Q., Freeland, C., & Heidorn, P. B. (2008). Taxonomic Name Recognition in Biodiversity Heritage Library.
https://www.researchgate.net/publication/41492665_Name_Matters_Taxonomic_Name_Recognition_TNR_in_Biodiversity_Heritage_Library_BHL
iConference paper. Evaluates NER algorithms.

Citizen Science

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., ... & Murray, P. (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, *389*(3), 1179-1189.
Description of original citizen science project that inspired creation of Zooniverse online citizen science community and platform for project development.

Mugar, G., Østerlund, C., Jackson, C. B., & Crowston, K. (2015, June). Being present in online communities: learning in citizen science. In *Proceedings of the 7th International Conference on Communities and Technologies* (pp. 129-138). ACM.
Main question: how do online community members learn to become valuable contributors to the community? Examines online citizen science projects as collaboration between laypeople and researchers. Participants may gather primary data, such as classifications in Zooniverse. Participants are deliberately not allowed to see others' work. Limited number of experts are on-hand to teach and provide guidance. Provides information that is useful for understanding online citizen science dynamics, and also for development of crowdsourcing infrastructures like Zooniverse.

Sørensen, J. J. W., Pedersen, M. K., Munch, M., Haikka, P., Jensen, J. H., Planke, T., ... & Sherson, J. F. (2016). Exploring the quantum speed limit with computer games. *Nature*, *532*(7598), 210.
When computers can beat us at many tasks, there is a battle between man and machine. Machine consciousness has not been achieved, but computers can make choices without human input. Human skills are superior in some areas. Research group at Aarhus University identifies our strength as: "our skill in approaching problems heuristically and solving them intuitively". Even supercomputers struggle with quantum physics problems. The CODER project uses gamification to solve large scale problems. People playing games is considered computing power. Players are more efficient than machines at solving these problems towards understanding quantum physics challenges

Xue, K. (2014). Popular Science. *Harvard Magazine*, *116*(3), 54-59.
Certain scientific problems require computational resources – the structure of proteins, for example. But challenging with infinite structural possibilities. The Rosetta@home project outsourced prediction to distribution of home computers. People started to get involved and want to participate more. The result was "protein folding" activities for ordinary citizens. "Citizen Science" enables ordinary people to contribute to research. Beyond donating idle time on computers, citizen scientists can participate more deeply. It is not especially difficult to explain basic scientific questions to people who want to

help. Galaxy Zoo was founded in 2007. The project was overwhelmed with initial traffic due to interest. It used humans to do what computers could not – to rapidly classify galaxies with a large number of images captured by telescopes. Humans are as good as computer programs for certain tasks, and humans can train computers if a lot of people look at data. Many scientific disciplines have more data than they can handle. Combinations of human and computer analyses can scale to process big data; this type of research is called “crowdsourcing” - small contributions from a large number of users. This still requires training citizen scientists to perform computational tasks, and there are concerns about trustworthiness/accuracy of results. Citizen scientists often find ways to make projects their own. Ordinary people provide fresh eyes on data, which can be a learning opportunity for scientists as well. The question is: Is citizen science for the citizens or the scientists? “Most researchers involved with citizen science believe this vision is one worth seeking, whatever the way forward may be. ‘To what degree does citizen science bring the lay community closer to the interface of science and society?’ asks Eva Guinan. ’In a world where so many people say and feel that they are being left behind by science and technology, does citizen science help? Or does it act like just another online game?’”