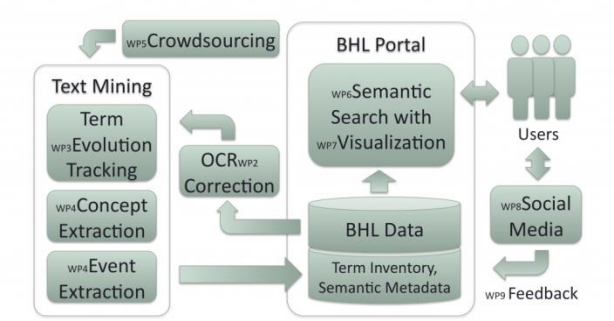
Mining Biodiversity

Enriching Biodiversity Heritage with Text Mining and Social Media

Abstract: The *Mining Biodiversity* project aims to transform the Biodiversity Heritage Library into a next-generation social digital library resource to facilitate the study and discussion (via social media integration) of legacy science documents on biodiversity by a worldwide community and to raise awareness of the changes in biodiversity over time in the general public. The project will integrate novel text mining methods, visualization, crowdsourcing and social media into the BHL. The resulting digital resource will provide fully interlinked and indexed access to the full content of BHL library documents, via semantically enhanced and interactive browsing and searching capabilities, allowing users to locate precisely the information of interest to them in an easy and efficient manner.

The nine Working Packages that compose this project can be grouped in the following primary tasks:



1) **Automatic error correction** in extracted text from legacy biodiversity literature via optical character recognition (OCR).
2) **Corpus annotation** by crowdsourcing a gold standard for training and evaluating automatic annotation tools in legacy texts with semantic metadata (i.e., terminology, entities and significant events).
3) **Adaptation of text mining technologies** to extract metadata automatically and to track terminology evolution over time.
4) **Facilitate BHL's Semantic Search** allowing users to explore search results according to multiple information facets.
5) **Use interactive visualization techniques** to help users to make sense of search results through the integration of next generation browsing capabilities, assisted by a semantic similarity network of important terms and entities.
6) **Design of a social media layer**, serving as an environment for diverse users to interact and collaborate on science, public education, awareness and outreach.

Timeline:

Jan 2014 - December 2015

Deliverables:

- Requirements report and design mock-ups of the enhanced BHL Portal (July 2014)
- Tool for OCR correction based on algorithms like Google n-grams and fast string matching (October 2014)
- A compilation of automatically extracted normalized terms (August 2014) and term inventory with time boundary information (November 2014).
- Named entity recognisers, annotations for named entities. (Feb.2015) and event extractors, annotations for events (July 2015).
- Annotation guidelines, annotation tool configuration (Sep. 2014) and an annotated corpus (Dec. 2014).
- BHL database enriched with semantic metadata (May 2015), a method for retrieving results based on the semantic similarity between the query and BHL documents (Jul. 2015), and BHL's search interface extended by facets. (Set. 2015) and visualization modules (Dec. 2015)
- Report on the results of the survey of social media sites currently used by biodiversity scholars (Dec. 2014), BHL fully integrated with social media sites (May 2015) and BHL enhanced with a collaborative curation environment (Dec. 2015)

Funder and Partners

Funder: Institute of Museum and Library Services ($174,724 for US partners)

Partners: Center for Biodiversity Informatics, Missouri Botanical Garden, (US); National Centre for Text Mining, University of Manchester, (UK); Big Data Analytics Institute and Social Media Lab, Dalhousie University, (CAN). Also participating: Smithsonian Institutions and Encyclopedia of Life.

Public Webpage:

http://miningbiodiversity.org/