## Mining Biodiversity

### Jisc Final Report

<table>
<thead>
<tr>
<th><strong>Project Information</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Project Identifier</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Project Title</strong></td>
<td>Mining Biodiversity</td>
</tr>
<tr>
<td><strong>Project Hashtag</strong></td>
<td>#mibio_did</td>
</tr>
<tr>
<td><strong>Start Date</strong></td>
<td>7th April 2014</td>
</tr>
<tr>
<td><strong>End Date</strong></td>
<td>30th April 2016</td>
</tr>
<tr>
<td><strong>Lead Institution</strong></td>
<td>National Centre for Text Mining, (NaCTeM) University of Manchester</td>
</tr>
<tr>
<td><strong>Project Director</strong></td>
<td>Sophia Ananiadou</td>
</tr>
<tr>
<td><strong>Project Manager</strong></td>
<td>Sophia Ananiadou</td>
</tr>
<tr>
<td><strong>Contact email</strong></td>
<td><a href="mailto:sophia.ananiadou@manchester.ac.uk">sophia.ananiadou@manchester.ac.uk</a></td>
</tr>
<tr>
<td><strong>Partner Institutions</strong></td>
<td>Missouri Botanical Garden (US), Dalhousie University (Canada), Ryerson University (Canada)</td>
</tr>
<tr>
<td><strong>Project Web URL</strong></td>
<td><a href="http://miningbiodiversity.org">http://miningbiodiversity.org</a></td>
</tr>
<tr>
<td><strong>Programme Name</strong></td>
<td>Digging into Data Phase 3</td>
</tr>
<tr>
<td><strong>Programme Manager</strong></td>
<td>Christopher Brown</td>
</tr>
</tbody>
</table>

### Document Information

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Date</strong></td>
<td>24/09/2016</td>
</tr>
<tr>
<td><strong>Filename</strong></td>
<td>JISC_NaCTeM_MiningBiodiversity</td>
</tr>
<tr>
<td><strong>Access</strong></td>
<td>This report is for general dissemination</td>
</tr>
</tbody>
</table>

### Document History

<table>
<thead>
<tr>
<th><strong>Version</strong></th>
<th><strong>Date</strong></th>
<th><strong>Comments</strong></th>
</tr>
</thead>
</table>

---

Document title: Jisc Final Report Template
Last updated: Feb 2011 - v11.0
Table of Contents

MINING BIODIVERSITY .................................................................................................................................................. 1
1 ACKNOWLEDGEMENTS .............................................................................................................................................. 4
2 PROJECT SUMMARY .................................................................................................................................................. 4
3 MAIN BODY OF REPORT .......................................................................................................................................... 4

3.1 PROJECT OUTPUTS AND OUTCOMES .................................................................................................................. 4
3.2 HOW DID YOU GO ABOUT ACHIEVING YOUR OUTPUTS / OUTCOMES? ......................................................... 5
3.3 WHAT DID YOU LEARN? ......................................................................................................................................... 13
3.4 IMMEDIATE IMPACT .............................................................................................................................................. 14
3.5 FUTURE IMPACT .................................................................................................................................................... 15

4 CONCLUSIONS ......................................................................................................................................................... 15
5 RECOMMENDATIONS .............................................................................................................................................. 16
6 IMPLICATIONS FOR THE FUTURE .......................................................................................................................... 17
7 REFERENCES ............................................................................................................................................................... 17
8 APPENDICES .............................................................................................................................................................. 18

APPENDIX A - REPORT: SURVEY ON ENHANCEMENTS TO BHL SEARCH FUNCTIONALITIES ... 18
1 Acknowledgements

Behind the Mining Biodiversity project is a consortium consisting of the following partners: (1) the National Centre for Text Mining (NaCTeM) of the University of Manchester, UK; (2) the Missouri Botanical Garden in the US, a member of the Biodiversity Heritage Library; (3) the Big Data Analytics Institute of Dalhousie University, Canada; and (4) the Social Media Lab of Ryerson University, Canada. The partners gratefully acknowledge the funding provided by Jisc under the Digging into Data 3 (DiD3) Program.

2 Project Summary

The Mining Biodiversity project is an international collaboration between the National Centre for Text Mining (UK), Missouri Botanical Garden (US), Dalhousie University’s Big Data Analytics Institute (Canada) and Ryerson University’s Social Media Lab (Canada). Its overarching goal is to transform the Biodiversity Heritage Library (BHL) [1], a digital library of over 40 million pages of taxonomic literature, into a next-generation social digital resource to facilitate the collaborative study and discussion of legacy biodiversity documents by a worldwide community. In this project, methods for text mining, visualisation and social media analysis were developed to effectively serve BHL users with semantically enriched content. The resulting digital resource provides access to the full content of BHL library documents via semantically enhanced, interactive browsing and searching capabilities, allowing users to more efficiently locate information of interest to them.

3 Main Body of Report

3.1 Project Outputs and Outcomes

<table>
<thead>
<tr>
<th>Output / Outcome Type</th>
<th>Brief Description and URLs (where applicable)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Report</td>
<td>User requirements analysis</td>
</tr>
<tr>
<td>Knowledge built</td>
<td>Terminological inventory (<a href="https://nactem.ac.uk/bhl_inventory/">https://nactem.ac.uk/bhl_inventory/</a>)</td>
</tr>
<tr>
<td>Knowledge built</td>
<td>Corpus of annotated documents</td>
</tr>
<tr>
<td>Software</td>
<td>Tool for correcting errors generated by optical character recognition (OCR)</td>
</tr>
<tr>
<td>Software</td>
<td>Suite of text mining-based curation tools</td>
</tr>
<tr>
<td>Software</td>
<td>Suite of text mining-based semantic metadata generation tools</td>
</tr>
<tr>
<td>Software</td>
<td>Social media widgets</td>
</tr>
</tbody>
</table>
3.2 How did you go about achieving your outputs / outcomes?

In this section, we outline details of the work undertaken by firstly presenting our project’s aims and objectives, followed by an overview of the roles of the different project partners. Lastly, we present details of our consortium’s activities according to our nine work packages.

**Aims and objectives**

We at the National Centre for Text Mining (NaCTeM) of the University of Manchester has partnered with three other institutions, namely the Missouri Botanical Garden (US), Dalhousie University’s Big Data Analytics Institute (Canada), and Ryerson University’s Social Media Lab (Canada). With their expertise, we accomplished the objectives that we set out in the beginning of the project as follows.

1. **Improving BHL content and enriching it with semantic metadata.** BHL contains digital legacy documents both in their original form (e.g., as scanned images) and in plain text, as produced by optical character recognition (OCR) tools. Error correction was applied on the OCR-generated texts to improve their quality. Text mining (TM) techniques were then be used to automatically annotate these documents with semantic information. The original BHL system contains only annotations for taxonomic names, aside from the standard bibliographic metadata. One of our aims was to automatically enrich the documents in the library with further semantic information, e.g., identified mentions of geographic locations, anatomical entities, morphological characteristics, and the associations between them.

2. **Construction of a biodiversity term inventory.** A collection of biodiversity-relevant terms built based on the BHL documents which were processed by our text processing tools. This was motivated by our desire to enrich the few existing controlled vocabularies on biodiversity, e.g., Catalogue of Life [2], by linking any given term with other semantically related terms, for the purpose of enabling the BHL search system with automatic query expansion.
3. **Development of a BHL semantic search system.** A search engine over BHL documents that leverages text mining results was developed. Originally, the BHL search system supports only basic search functionalities, i.e., keyword-based searches, searching in bibliographic fields, and browsing. The incorporation of TM-generated semantic metadata and biodiversity term inventory, however, enabled our proposed search engine to return results based on semantic similarity. To facilitate efficient navigation of search results, the search system was also enhanced with capabilities for visualisation.

4. **Provision of a social media environment for BHL users.** Existing social media sites were integrated with the BHL to foster the collaborative discussion and curation of biodiversity digital artifacts. Integrating the BHL with social media sites will potentially allow the BHL to reach an even wider audience of scholars and at the same time, facilitated the enrichment of the library with community-curated digital media, e.g., biodiversity-relevant images, video clips.

**Roles of project partners**
NaCTeM was responsible for all tasks relating to the automated semantic processing of BHL textual content. This includes the automatic annotation of BHL documents with semantic metadata, construction of the biodiversity term inventory, and development of the BHL search engine. Being a leading research centre for text mining, NaCTeM is equipped with state-of-the-art biomedical text mining methods which were adapted for the domain of biodiversity.

The Missouri Botanical Garden is a member of the BHL consortium of libraries. They guided the development of the BHL into a social digital library by providing use case scenarios and specifications. They provided user feedback throughout the duration of the project, especially towards the final phase in which close engagement with users was necessary to carry out qualitative evaluation of our results.

Dalhousie University's Big Data Analytics Institute lent their expertise on processing volumes of data specifically for the tasks of OCR error correction (building upon their previous work on Google n-grams) and interactive visualisation. The Social Media Lab of Ryerson University, on the other hand, developed the social media environment for BHL. Firstly, they used their expertise on analysis of social networks to identify existing social media sites which serve as discussion for a for biodiversity scholars. The identified sites, e.g., Twitter, Facebook, were then integrated into the BHL through application programming interfaces (APIs).

**Work packages**

**Work package 1: Requirements analysis**
In order to identify the improvements that will best enhance the search experience of BHL users, we designed several mock-ups (i.e., wireframes) showing proposed functionalities that could potentially enable BHL users to more efficiently discover and share information. To gather user preferences and suggestions, the mock-ups
were published as an online survey\(^1\) in which various types of users participated, e.g., taxonomists, ecologists, biologists, informatics researchers and librarians. A total of 31 respondents contributed their feedback and suggestions. We summarised results of the survey in a report that we provide as Appendix A of this document.

Twelve of the survey respondents became members of a focus group for further guiding the development of the proposed functionalities. Based on the focus group discussions, we finally identified the functionalities and types of semantic metadata which are of highest priority. These results informed the annotation schema that we devised as part of Work package 5 (Corpus annotation).

**Work package 2: OCR error correction**

Like in most historical archives, textual content in BHL suffers from errors brought about by the application of optical character recognition (OCR) tools on scanned images of pages. To rectify those errors, our partners from the Institute for Big Data Analytics in Dalhousie University developed statistical methods for detecting and subsequently correcting errors from OCR. The tool is underpinned by an unsupervised statistical method that makes use of frequencies in the Google Books Ngram Corpus to guide the selection of the most likely correction based on lexical and contextual similarities. The OCR error detection and correction tool has been packaged as an interactive web application.

**Work package 3: Term harvesting and tracking**

A list of terms pertinent to the biodiversity domain was compiled, containing names of various species, e.g., mammals, birds and flowering plants. We made use of methods based on distributional semantic models (DSMs) in order to keep track of expressions that are semantically related to each of these terms. By using DSMs, which take into account the similarity of expressions in terms of context, we are ensuring that as many semantic variants are being captured, apart from just lexical ones. We investigated various count-based and prediction-based DSMs, and conducted a comparative evaluation of the following models: count-based [3], continuous bag-of-words (CBOW) [4], global vectors (GloVe) [5], and their basic additive model (BAM) variants [6], i.e., count-based BAM, CBOW-BAM, GloVe-BAM. This allowed us to identify prediction-based methods, i.e., CBOW and GloVe, as best performing in majority of the cases. We thus applied the CBOW and GloVe models to automatically retrieve semantically related terms for a total of more than 288,000 mammal, bird and flowering plant names in our term inventory. The resulting publicly available resource ([http://nactem.ac.uk/bhl_inventory](http://nactem.ac.uk/bhl_inventory)) supports BHL users in finding more documents of interest by means of query expansion, i.e., the automatic suggestion of keywords related to an initial search term, as described below in Work package 7.

**Work package 4: Entity and relation extraction**

In order to facilitate the generation of semantic metadata over the textual collections in BHL, a suite of tools for automatically extracting biodiversity-relevant named entities (and relationships between them) were developed using the Argo text mining

---

\(^1\) [http://miningbiodiversity.polldaddy.com/s/enhancements-to-bhl-search-functionalities](http://miningbiodiversity.polldaddy.com/s/enhancements-to-bhl-search-functionalities)
workbench (http://argo.nactem.ac.uk) [7]. These tools were implemented as automatic processing workflows, comprised of natural language processing (NLP) components for sentence segmentation (LingPipe [8]), tokenisation, part-of-speech (POS) tagging, lemmatisation (GENIA Tagger [9]), deep syntactic parsing (Enju [10]), gazetteer matching, and for machine learning-based sequence labelling based upon the conditional random fields (CRF) [11] algorithm (NERsuite [12]). As a result, we can now automatically recognise expressions pertaining to taxonomic names, locations, habitats, persons, anatomical entities, qualities and time.

Work package 5: Corpus annotation
To support the training and development of our automated tools for semantic metadata extraction, e.g., the named entity recognisers mentioned above, we have constructed a corpus of BHL documents consisting of pages enriched with semantic annotations. As a purely manual annotation process is tedious and time-consuming, we developed a strategy that lends semi-automatic annotation support to our human annotators. Specifically, we took an active learning-like approach that iteratively: (1) leverages machine learning-based tools to automatically generate annotations in order to lessen the effort required from human annotators, (2) takes the corrections done by the human annotations on the pre-supplied annotations as feedback, and (3) retrains the tools, thus incrementally improving their performance as well as the quality of their automatically generated annotations. In this manner, two sets of annotations over 155 BHL pages were produced. These were then compared against each other by a third human annotator, a biodiversity specialist, who acted as an adjudicator deciding on what was incorporated into the final set of gold standard annotations.

Work package 6: Building of semantic search engine
Employing Apache Lucene Solr (https://lucene.apache.org/solr/), a well-known search engine, all of the documents in BHL were indexed with two types of information: (1) bibliographic metadata, e.g., title, author names, publication year; and (2) semantic metadata corresponding to any of the species names in our term inventory (described in Work package 3), that are mentioned in the documents. The result is a search engine over millions of pages of optical character-recognised text.

Work package 7: Integration with visualisation modules
To facilitate the visual exploration of BHL documents, a user-interactive web-based interface, depicted in Figure 1, was developed with the semantic search engine developed in Work package 6 as its backend. The resulting search system (http://nactem.ac.uk/BHLQueryExpansion/) supports automatic query expansion, allowing the user to discover documents pertaining to concepts that are semantically related to his or her search query. The user could, for example, specify the common name of a species (e.g., lion) and the search system will suggest other names with which the original query could be expanded (e.g., the scientific name Panthera leo), in order to broaden the set of relevant documents. Apart from strict synonyms, names of semantically related species, i.e., those that belong to the same taxonomic family (e.g., jaguar, cougar), could also be suggested by the search system based on the underlying term inventory. The user, however, has full control over what gets included in the search results. That is, the result set contains only documents matching the user’s original query, and is expanded only if the user explicitly selects
any of the suggested names. The system’s suggestions are presented in a visual manner: a thumbnail image from the Encyclopedia of Life [13] is shown for every suggested name, as well as indicator bars signifying the name’s frequency within the BHL corpus and its relatedness to the original query. If a user chooses to expand the search results by choosing any of the suggested names, again there are visual widgets (e.g., a context viewer with colour-coded items) to help him/her distinguish the documents retrieved by the original query from those that were matched as part of the expansion. A detailed user guide/manual is available for download from the web interface.

**Work package 8: Social media integration**

To better support BHL users in sharing information or findings of interest with the rest of the community, and to foster interactions and collaborations between them, our partners from the Social Media Lab of Ryerson University have developed and/or exploited several social media tools. The AddThis widget (http://www.addthis.com/) has been incorporated into the BHL web pages, allowing users to instantly publish URLs of interest in social media sites such as Twitter, Facebook, GooglePlus and Pinterest. In this way, users are able to bookmark and share any of the BHL pages or images on any social media platform, in a centralised manner. In order to determine how we can best support collaborations between BHL users, however, we needed to fully understand the profile of BHL users, e.g., which biodiversity subdomains or types of content are already getting the most attention or interest, or which ones could benefit from more social media exposure. To this end, the ready-made Netlytic application (https://netlytic.org/) was used to analyse the BHL Twitter network. Additionally, Altmetric tools (http://www.altmetric.com/) were exploited in order to track which of the BHL documents are being mentioned on the web. Most importantly, an application called MyTweeps was developed, facilitating the analysis of Tweets from one’s followers (called a community). Configuring this application for BHL’s Twitter account (http://mytweeps.com/community/BioDivLibrary) allowed us to gain insight on what interests BHL’s followers.
Figure 1. Visual search interface incorporating suggested semantically related terms for query expansion. A: Initial and expanded query. B: Search result list and context viewer. The context viewer on the left-hand side shows a zoomed out view of the retrieved list. Documents retrieved according to the expanded query are shown with a light blue background. C: Thumbnails with suggested terms for query expansion. Apart from a relevant image, each thumbnail depicts the suggested term’s frequency within BHL documents, its relatedness to the query term, and the provenance of the suggestion, i.e., our term inventory or other external resources, e.g., the Catalogue of Life [14], Encyclopedia of Life [13], and the Global Biodiversity Information Facility [15].

Work package 9: Evaluation
We carried out both quantitative and qualitative evaluation. For the former, the performance of each version of the OCR error detection and correction tools and term extractors has been measured using standard metrics of precision, recall and F-score, shown in Table 1. Similarly, the performance of our text mining tools was systematically evaluated. Table 2 presents the performance of our named entity recognisers on each entity type, in terms of precision, recall and F-score. Meanwhile, Table 3 provides details on our distributional semantic models’ performance on the task of automatically extracting semantic variants of species names, in terms of mean average precision (MAP).

Table 1. Performance of OCR error detection and correction tools

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error detection</td>
<td>86.53</td>
<td>95.74</td>
<td>90.90</td>
</tr>
<tr>
<td>Error correction</td>
<td>59.61</td>
<td>65.95</td>
<td>62.62</td>
</tr>
</tbody>
</table>
### Table 2. Performance of named entity recognisers on each entity type

<table>
<thead>
<tr>
<th>Entity Type</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxon</td>
<td>81.09</td>
<td>46.64</td>
<td>59.22</td>
</tr>
<tr>
<td>Location</td>
<td>83.11</td>
<td>70.00</td>
<td>75.99</td>
</tr>
<tr>
<td>Habitat</td>
<td>76.00</td>
<td>46.34</td>
<td>57.58</td>
</tr>
<tr>
<td>Person</td>
<td>50.00</td>
<td>5.00</td>
<td>9.09</td>
</tr>
<tr>
<td>Anatomical part</td>
<td>79.61</td>
<td>48.47</td>
<td>60.25</td>
</tr>
<tr>
<td>Temporal expression</td>
<td>100.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Overall</td>
<td>81.11</td>
<td>52.93</td>
<td>64.06</td>
</tr>
</tbody>
</table>

### Table 3. The mean average precision (MAP) obtained by each distributional semantic model (DSM) on bird, mammal and plant names. Bold numbers denote the highest scores obtained per general DSM type in each gold standard used (BirdLife=BirdLife International [16], CoL=Catalogue of Life [14], ITIS=Integrated Taxonomic Information System [17], PLANTS=Plants Database [18]).

<table>
<thead>
<tr>
<th>Approach</th>
<th>Model</th>
<th>Birds</th>
<th>Mammals</th>
<th>Plants</th>
<th>CoL</th>
<th>ITIS</th>
<th>CoL</th>
<th>PLAN TS</th>
<th>CoL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count-based</td>
<td>Count-based</td>
<td>44.28</td>
<td>49.86</td>
<td>39.04</td>
<td>39.27</td>
<td>38.79</td>
<td>36.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Count-based BAM</td>
<td>34.15</td>
<td>36.90</td>
<td>30.54</td>
<td>30.25</td>
<td>44.36</td>
<td>45.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction-based</td>
<td>CBOW</td>
<td>69.56</td>
<td>73.34</td>
<td>52.17</td>
<td>53.12</td>
<td>62.14</td>
<td>51.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CBOW-BAM</td>
<td>13.28</td>
<td>16.86</td>
<td>19.46</td>
<td>19.08</td>
<td>49.03</td>
<td>51.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GloVe</td>
<td>72.79</td>
<td>77.11</td>
<td>70.76</td>
<td>73.30</td>
<td>28.30</td>
<td>33.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GloVe-BAM</td>
<td>22.86</td>
<td>24.06</td>
<td>27.44</td>
<td>27.82</td>
<td>46.71</td>
<td>46.65</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Qualitative evaluation of our tools was carried out by conducting a survey\(^2\) amongst 20 domain experts in which we collected user feedback on the effectiveness and usability of the web-based search interface developed as part of Work package 7. We requested the experts to explore the search system using six different scientific or vernacular species names as queries. Whilst three of the queries were pre-supplied by us, for the other three, the users were asked to enter their own preferred species names belonging to any of the following three taxonomic classes: birds, mammals and flowering plants. Scientific names had to be provided according to the binomial nomenclature, consisting only of the genus and specific epithet, e.g., *Lepus timidus* and *Spizella passerina*. In exploring the results of each query, we encouraged them to select a few of the terms automatically suggested by the interface, in order for them to have an appreciation of the effects of the query expansion feature. We designed a questionnaire sheet containing eleven

\(^2\) [http://goo.gl/forms/3mO5fWd7Y4](http://goo.gl/forms/3mO5fWd7Y4)
questions, eight of which were compulsory; the other three were optional. The eight compulsory questions consist of two subsets. The first four questions were formulated to evaluate whether the expansion of queries in a semi-automatic manner is useful or not. In particular, one of the questions aims to estimate the helpfulness of suggested species names that are not necessarily semantic variants but are semantically related, i.e., through shared habitat, geographic location or taxonomic class as in the case of “jaguar” and “lion”. The second subset of four questions sought to obtain feedback on the visual functionalities of the interface. In each question, users were asked to assess a specific functionality by rating its usefulness/helpfulness from 1 (not useful/helpful at all) to 5 (very useful/helpful). Out of the 20 invited experts, 13 submitted complete responses. Their ratings for the eight compulsory questions are depicted in Figure 2. We can see that the median ratings for the first four questions are consistently high, i.e., from 4 to 5, which can be generally thought of as an indication that our users find the term inventory-based query expansion functionality to be useful. Based on the responses to Question 1, most users agreed that the inventory’s suggested terms are useful. In terms of broadening the scope of search results (covered by Questions 2 and 4), responses ranged from 3 to 5, indicating that users are generally satisfied with this system feature. For Question 3, which assessed whether automatically suggested semantically related terms are useful although not all of them are exact synonyms, all of our users responded positively (with answers ranging from 3 to 5), except for one who gave a rating of 2. These responses reveal that most users share the opinion that it is helpful to be presented with suggestions of species names that are not necessarily synonyms but may be related to the original query in other ways, e.g., in terms of shared habitat, taxonomic class, or geographic location.

![Figure 2. Ratings for the visual search interface given by thirteen users. The red line indicates the median rating for each question.](image-url)
3.3 What did you learn?
The lessons that we learnt from the Mining Biodiversity project are summarised by the four points outlined below.

Language used in the biodiversity domain poses challenges to natural language processing. Whilst attempting to develop methods for extracting relationships between biodiversity-relevant concepts, we realised that biodiversity literature is quite different from other types of scientific literature, e.g., biomedical articles. Whereas scientific findings in the latter are mostly expressed in the form of full sentences, sentence fragments or phrases are very prevalent in biodiversity literature. Many documents, for example, would mention the species being described only at the beginning, and then use sentence fragments to describe its physical characteristics or behaviour, purposely without reiterating the subject, i.e., the name of the species. Upon processing a set of 155 BHL pages with the Enju syntactic parser [10], more than 76% of the sentences were detected as sentence fragments. We note, however, that this issue is not only due to incomplete sentences; errors brought about by OCR also account for this issue.

Figure 3. Top-20 accuracy of the various distributional semantic models (DSMs) on bird, mammal and flowering plant names using the Catalogue of Life as gold standard. On bird names (a), the continuous bag-of-words (CBOW) model obtained optimum performance, closely followed by the global vectors (GloVe) model. In contrast, GloVe was top performing on mammal names (b). However, on flowering plant names, GloVe obtained very poor performance, with CBOW demonstrating the strongest performance.

Different taxonomic groups should be treated differently. Results of our evaluation of our distributional semantic models imply that each taxonomic group, e.g., birds, mammals, flowering plants, should be investigated independently of the others. Figures 3a-3c depict the top-20 accuracy of the various models we explored, respectively, on bird, mammal and plant names from the same resource (i.e., the Catalogue of Life). In the case of bird names, the word2vec continuous bag-of-words (CBOW) model was top performing, with the global vectors (GloVe) model close behind. With mammal names, this ranking is slightly reversed, with GloVe obtaining superior performance over CBOW. Surprisingly, however, GloVe obtained lowest performance on plant names, compared to all of the other models considered. The
above findings imply that one cannot blindly apply just any distributional semantic model (DSM) on a particular species or taxonomic category. Instead, a thorough and comparative investigation is required in order to determine the best performing one for every category.

**Employ visualisation tools to give users control over the search system.** During the user requirements survey and focussed group discussions that we carried out at the beginning of the project, our respondents were a bit skeptical about how related concepts will be presented by a search system to a user. They were mostly concerned about introducing noise in the search results, i.e., including results that might be irrelevant to the original query, and thus they originally preferred a conventional faceted search interface that does not automatically incorporate any text mining-derived results. However, after careful studies and a series of discussions, we were able to choose a suitable means for presenting related concepts, i.e., by taking inspiration from recommendation systems (Figure 1C). That is, the incorporation of text mining-derived results is done semi-automatically, based upon the user’s preference to include or exclude the system’s suggested terms. In this way, users are informed and guided by text mining but still have full control over what becomes included in the search results. The survey that was conducted to evaluate the final version of the interface demonstrated our users’ acceptance of our recommendation system-inspired search interface.

### 3.4 Immediate Impact

The Mining Biodiversity project officially kick-started NaCTeM’s involvement in text mining in the domain of biodiversity. Previously, NaCTeM has focussed on biomedical text mining and supporting health applications. This project, however, provided us with the opportunity to undertake research in a new challenging area, i.e., biodiversity, and to develop research work that supports semantic searching over a well-known digital library, the Biodiversity Heritage Library. Owing to this project, we have gained exposure and at the same time created impact on the biodiversity informatics and digital libraries communities, as evidenced by our participation and presence in two annual meetings of the Taxonomic Database Working Group (TDWG 2014 and 2015) as well as the 20th International Conference on Theory and Practice of Digital Libraries (TPDL 2016).

We believe that this project has benefitted the wider community in that it served as a vehicle for us to build text mining capacity within the digital libraries community. The tools which were developed as part of this project have been made available in the web-based text mining workbench Argo ([http://argo.nactem.ac.uk/](http://argo.nactem.ac.uk/)) which allows even non-text mining experts to build their own text mining solutions by designing workflows using a graphical user interface. We submitted a proposal for a text mining tutorial at TPDL 2016, which eventually was one of the conference’s three accepted tutorials ([http://www.tpdl2016.org/tutorials](http://www.tpdl2016.org/tutorials)). It served as a knowledge transfer activity in which we aimed to enable digital librarians in the use of Argo for enriching their respective textual collections with semantic metadata. Indeed, the tutorial was well received by the audience. We received not only positive feedback but also interest from institutes, i.e., the German National Library of Science and Technology (TIB)
and the Information Library Complex of Peter the Great St. Petersburg Polytechnic University, in utilizing our text mining tools in Argo for the purposes of enriching and extending their libraries. Furthermore, we have been invited by one of the TPDL pioneers to organize a full workshop (on the same topic of text mining to support biodiversity informatics and digital libraries) in TPDL 2017 in Greece.

We believe that this project has instilled a deeper appreciation for text mining in taxonomists, biodiversity informaticians and digital librarians. Although the text mining paper that we have submitted to PLoS One is still under review as of this writing, preliminary comments indicate that owing to this project, taxonomists have started to appreciate the advantages of going beyond keyword- and string-matching when indexing biodiversity literature. Our comprehensive review of related work showed that all efforts for indexing documents with species names have thus far focussed on the use of string similarity measures and capturing lexical variants and strict synonyms. Meanwhile, our methods capture semantic variants in addition to strict synonyms. Despite our methods’ deviation from conventional string-based methods and that they return more than strict synonyms, feedback from our reviewers indicate their acceptance of our semantics-based approach and their recognition of its usefulness in the context of literature search. One of the reviewers, for example, has even suggested the eventual integration of our methods into the well-known Global Names Index [19], a resource for indexing taxonomic knowledge.

3.5 Future Impact

Apart from the impact that we have already created, described above, we foresee that the work we carried out as part of the Mining Biodiversity project will have a strong impact on the creators and developers of other systems for searching biodiversity knowledge. Our work so far has enhanced searching over the Biodiversity Heritage Library. However, other biodiversity knowledge repositories, e.g., Pensoft [20] and GBIF [15], still rely on keyword-based search. We hope that with the publication of our text mining papers and resources, we can enable such stakeholders to develop advanced search systems. We will be able to track this impact in two ways: (1) by monitoring the number of accesses (through the publishers’ websites) and citations (through Google Scholar) for our journal and conference papers; and (2) by monitoring how many times our resources have been downloaded from the NaCTeM website. We will be responsive to any enquiries (e.g., by email or social media) pertaining to the integration of our tools and resources in other applications.

4 Conclusions

The Jisc-funded Mining Biodiversity project has provided us with an opportunity to undertake research in the development of text mining methods to facilitate semantic searching over the Biodiversity Heritage Library (BHL) textual collections. Firstly, with our partners from Dalhousie University, we were able to develop tools for optical character recognition (OCR) error detection and correction. Based on their output, we were able to develop a gold standard corpus containing documents with manually annotated named entities (e.g., taxa, geographic locations, persons, habitats, anatomical parts, qualities, temporal expressions). This resource allowed us to develop and evaluate an active learning-like framework for semi-automatically
curating biodiversity literature, which we have made available in our web-based text mining workbench Argo. In parallel, we developed distributional semantic models (DSMs) for automatically constructing a biodiversity terminological inventory. The semantic metadata generated by processing documents with these tools was used to index BHL documents. To enable users to search and explore the knowledge contained in these documents in an interactive manner, a visual interface was developed. Domain experts who evaluated the interface found the text mining-based functionalities for recommending semantically related concepts useful and beneficial. In parallel, our partners from Ryerson University have developed social media widgets that facilitate the discussion of BHL items by members of the community in an interactive and collaborative manner.

Outcomes of our work have been disseminated to the wider community by means of user surveys, publication of research articles, presentations in conferences, organisation of a text mining tutorial in a leading digital libraries conference. In this way, we maximise the impact of our work on the wider community. Specifically, we have put forward a semantics-driven approach based on text mining, for enriching digital libraries with semantic metadata, which in turn, facilitates more meaningful searching through textual collections. The resources that we have made publicly available, e.g., the biodiversity terminological inventory, text mining workflows, will allow members of the taxonomic/biodiversity informatics and digital libraries communities to generate and incorporate semantic metadata in their own literature collections.

This successful collaboration with our partners in the US and Canada would not have been possible without funding support from Jisc under the Digging into Data 3 Program. It opened opportunities for our research group to undertake interdisciplinary research in a problem area where text mining and natural language processing are as yet underexplored. It also allowed us to carry out dissemination activities, e.g., paper publications, conference presentations and tutorials, in order to maximise our project’s impact.

5 Recommendations

With the funding support from Jisc, we have successfully achieved our aims for this project. Nevertheless, we recommend that further and follow-up work be carried out in this area, especially as some problems are yet unsolved. The correction of OCR-generated errors in legacy biodiversity literature, for example, can still be improved. As discussed above, one of the main points we realised as part of this project is the greater difficulty that comes with dealing with biodiversity literature, compared to textual content from other domains such as biomedicine. Literature describing taxonomic findings, for instance, mostly consists of sentence fragments, leading to many syntactic parsing failures. Further research can thus be carried out in order to better address the extraction of relationships between different types of biodiversity-relevant concepts.

We have demonstrated that our text mining-based approaches can facilitate the semantic enrichment of the Biodiversity Heritage Library (BHL). We recommend that other digital libraries, even those in other domains (e.g., Europeana [21]) also
explore the use of text mining to enrich their textual collections with semantic annotations.

It is, however, understandable for some digital libraries, especially the already well established ones, to have reservations about taking a text mining-based framework to enrich and annotate their collections. With the availability of financial support for digital humanities research from funding bodies such as Jisc, however, more studies that will investigate the feasibility and effectiveness of such approaches within the context of individual digital libraries can be undertaken, and could lead to a step-change in how digital libraries are developed, updated and explored. We thus highly recommend the continuation of programs similar to Digging into Data, with a view of supporting such interdisciplinary research into digital humanities. A unique aspect of Digging into Data is its support for research involving partners from various countries, e.g., United Kingdom, US, Canada and the Netherlands, thus fostering the development of new collaborations. To further support the creation of new linkages especially at the project proposal stage (for future rounds of Digging into Data), we recommend that Jisc encourages participants to make use of tools such as Piirus (https://www.piirus.ac.uk/) to help them forge new research partnerships.

6 Implications for the future
As with all of the projects undertaken by NaCTeM, all of the tools and resources developed as part of the Mining Biodiversity project will be continuously made available to the community, for as long as the Centre is running. Being a self-sustainable Centre, NaCTeM owns and maintains its own servers which are hosting all of the tools, services and resources relevant to this project. Furthermore, resources such as corpora and the biodiversity term inventory will be made available via the META-SHARE repository (http://www.meta-net.eu/meta-share). NaCTeM is committed to making all of those available to the community even after the completion of Mining Biodiversity.

For any queries, potential linkages and collaborations, all members of the Mining Biodiversity research team can be reached through the details provided on the project’s NaCTeM page (http://nactem.ac.uk/DID-MIBIO) which will be continuously accessible from NaCTeM’s main website (https://nactem.ac.uk/). An in-house software engineering team is always ready to take queries and provide support to users encountering any issues with accessibility of our tools and resources.

7 References
8 Appendices

Appendix A - Report: Survey on Enhancements to BHL Search Functionalities
Report: Survey on Enhancements to BHL Search Functionalities

The survey was run from the 15th of August to the 14th of September. We received a total of 31 responses, 23 of which are complete, i.e., the respondent answered all of the 25 questions. In this report, we summarise the results from the survey and come up with some propositions for the enhanced BHL search engine which will be developed.

Respondent Profile

Area of expertise (24 responses)
The most common expertise amongst our respondents is Taxonomy (16), followed by Ecology (7) and Biodiversity Informatics (6).

Preferred search engines (31 responses)
Majority (23) of our respondents rely on other search systems aside from BHL. These include Google Scholar (20), JSTOR (18), Web of Science (15), bioOne (12) and PubMed (11).

Frequency of use (24 responses)
Whilst majority (13) of the respondents indicated that they use BHL occasionally, i.e., when the need to look up anything biodiversity-specific arises, seven said they are heavy users and three said they use it only moderately, i.e., two-three times a week.

Language of content of interest (24 responses)
Most of our respondents are interested in content in English (21). However, a significant number is also interested in content in French (13), German (11), Latin (8) and Spanish (8).

Satisfaction with BHL’s current search functionalities (27 responses)
The current version of BHL comes with both basic and advanced search functionalities. In the former, a user inputs his keyword(s) in one field (Figure 1) and BHL searches for those keywords in any of the currently searchable bibliographic metadata, e.g., title, author name, year, subject. With the latter, a user can carry out finer-grained search by supplying his keywords under specific fields. For instance, a user can choose which of the searchable items (i.e., books/journals, articles/chapters, scientific names) and fields (i.e., title, author name, year) his keywords should be matched against (Figure 2).

Figure 1. BHL’s basic search functionality
Majority of the respondents (13) use basic search. When asked whether they are happy with the results they are getting using the current search features (basic or advanced), many of them (13) are non-committal, i.e., they are not especially happy but at the same time not complaining about them. It is worth noting, however, that there are more respondents who explicitly said they are satisfied (8) with the current features than those who said they are not (only 1).
Feedback on proposed features
To elicit feedback from our respondents, we came up with a few mock-ups, i.e., wireframes, to show our proposed functionalities and interfaces. In all of the mock-ups presented below, a common feature is the time slider at the top, that the user can adjust according to his time period of interest.

Searching by facets (26 responses)
The first enhancement which was presented by the survey to the user is the use of facets for narrowing down search results (Figure 3). We are proposing a number of pre-defined top-level facets, including Taxonomy, Geography, Habitat, Traits and Publication. However, each of these facets, presented in a tree-like structure, is expandable --- the lower-level facets can be dynamically generated, depending on the clusters that will be automatically induced over the matching documents. The corresponding tick boxes can then be used to indicate the specific dimensions of interest.

Figure 3. A mock-up of the proposed feature for searching by facets, shown on the left-hand side.
When asked whether such faceted search could be useful to them, 17 respondents (out of 26) said they definitely are, and that they are relevant to the kind of work they are doing. Six said they can see how they can be useful but not for what they are currently working on.

Out of the 17 respondents who answered the question of which other facets they would like to see, six suggested Life History and three would like to see Ecosystems. There also seems to be a strong inclination towards Nomenclatural Act.

**Searching by automatically suggested questions (25 responses)**

The survey also presented an alternative way for users to narrow down search results, i.e., through the automatic generation of relevant questions, as shown in Figure 4.

![Figure 4. Automatically generated questions (on the right-hand side) to help the user find other relevant information.](image)

Ten out of the 25 respondents said that they can see how this feature can be useful, but not for their current purposes. Nine users said it will be definitely useful to them.
Searching by subject-verb-object queries (25 responses)

We also proposed a way for users to search for a concept and associated specific relations, i.e., by specifying subject-verb-object queries (Figure 5).

Figure 5. Searching by subject-verb-object. In this example, the user is interested in discovering whatever species (object) *Leptinotarsa* (subject) has been known to feed on (verb), and the corresponding supporting literature. Using the interactive graph on the right-hand side, a user can narrow down the results, e.g., to only documents about *Leptinotarsa* feeding on *eggplants*.

As with the feature for searching using suggested questions, there is no strong sentiment on whether this functionality is something that is definitely useful. Whilst 12 respondents said this is very relevant to their work, 11 indicated that they can see how it can be useful but not to them currently.

As for the graph-based visualisation of results (shown on the right-hand side of Figure 5), 12 out of 18 respondents said that they can see how this feature can be useful but not for their current purposes.
Searching for direct associations (25 responses)
In the survey, we also offered the option of finding concepts directly associated with a query concept (Figure 6).

Figure 6. Searching for directly associated concepts. The results are categorised according to type (i.e., Taxa, Geographical location, Habitat and Trait). In this example, a user is interested in concepts associated with *Eltanin*.

Similar to the last two proposed features, there is no strong indication of whether this feature is definitely wanted by our respondents. Twelve out of 25 respondents said this is something very relevant to what they are doing, whilst ten said it might be useful but not for their current purposes.

When asked what other concept types they would like to see aside from Taxa, Geographical location, Habitat and Trait, the respondents showed a strong preference for Taxon authority and Collector (as is the case with additional facets of interest in the first suggested feature).
Searching for indirect associations (23 responses)

Going beyond searching for direct associations, finding concepts indirectly associated with a query is another feature that we have proposed. With this feature, the user specifies the target concept type of interest, i.e., Taxa, Geographical location, Habitat, Trait, and the pivot concept type (Figure 7).

Figure 7. Searching for concepts indirectly associated with the query. In this example, the user wants to find species which could be related to *tarsier* in terms of their occurrences in the same geographical locations.

Out of the 23 respondents to our question of whether searching by indirect associations could be useful, 11 indicated that they can see its benefits but not to what they are currently doing, whilst six signified it will be definitely useful.

Improved page view (23 responses)

In terms of displaying results, we have also proposed to enhance BHL’s page view. Currently, each document page in BHL has an accompanying list of taxon names which are referred to
within the content. However, the exact occurrence of the name in the text is not shown. We propose to improve this in two ways (Figure 8). Firstly, we shall add more concept types as annotations, e.g., Geographic locations, Person names. Secondly, the actual locations of these mentions will be highlighted in the text.

Results from the survey indicate that majority (16) found this feature definitely useful, while five said it seems useful but not for their current purposes. For the question of what other concept types they would like to see highlighted on BHL pages, we received 15 responses, 12 of which indicated Taxon authority and nine specified Collector.

Figure 8. The proposed improved page view, with more semantic annotations, and with their actual locations highlighted within the text.

Ranking of proposed search functionalities (23 respondents)
In the survey, we asked our respondents to rank the four search features presented to them according to their preference. The calculated weighted rank (which took into account all 23
responses to this question) indicates that the feature for searching by facets is most preferred. Searching by subject-verb-object and searching for direct associations are tied as the next preferred feature, whilst searching for indirect associations is the least preferred.

We do realise that a probable reason why the last three search features that we proposed were not very convincing (apart, of course, from the possibility that some users simply do not need them at all) is our presentation of only one visualisation option for each search capability. This might have given our respondents the impression that the search feature under consideration is tightly coupled with the suggested visualisation, e.g., that with searching by subject-verb-object queries, a user will automatically get results in a graph (when perhaps he prefers another visual representation). In reality, the visualisation tools can be treated separately from the search capability. Whilst the latter deals with how a user will formulate his queries and the types of information he can expect, the former is a tool for analysing and exploring the actual results. To follow-up on this issue, we will revisit these three search features (for which we received inconclusive results) during focus group discussions, but this time with more options for visualisation.

Also during the focus group meetings, we will be open for new suggestions trying to learn more about what type of advanced queries or analytical needs would be important for users, so that new interfaces can be developed to address these needs.

Use cases
We list down below some of the use cases that we have gathered from comments provided by our survey respondents.

1. Finding the original description (for taxonomic research).
2. Finding host plants, for example (for ecological research).
3. Finding illustrations and plates.
4. Finding taxon name usage instances (taxonomic treatment, nomenclatural act).
5. Capturing spelling variants.
6. Viewing versions of the OCR-generated text.
7. Exposing the semantic metadata as a SPARQL endpoint.
8. Being able to access the search functionalities through APIs.
9. For the graph-based visualisation, to make the red leaves “explorable”.
10. In terms of searching for indirect associations, viewing only the most relevant results in case too many results come up.
11. Allowing users to highlight keywords they have entered.
12. Allowing users to annotate concepts if incorrectly recognised or missed.

Additionally, we shortlisted some competency questions that we found from the iBiosphere wiki (http://wiki.pro-ibiosphere.eu/wiki/Competency_Questions_for_RDF_Treatments) which illustrate several use cases which we imagine some BHL users might also be interested in:
1. Show me the treatments that describe the first publication of the name "Aus bus".
2. Which parasites of Oak trees in Central America are known?
3. Extract morphological features from a description to filter treatments based on morphological criteria (identification scenario).
4. Find treatments from a given taxonomic group that contain images (or video, or sound...).
5. What ant is eating seeds?
6. Are red taxa living above 1000 meter elevation?
7. What are the taxa living in a particular location?
8. What specimens have lat long coordinates in UTM projection?
9. Which other taxon concepts does the organism interacts with?
10. I want to have all the taxa which have been observed between 1990 and 2000 that live in Europe and for which geographic coordinates or collection codes are given.

Recommendations

Considering these results, we have come up with the following recommendations:

1. To implement the following features for which majority of our respondents have a strong preference:
   a. Faceted search (Figure 3).
   b. Improved page view (Figure 8).
2. To include additional facets such as Nomenclatural act and Authority.
3. To consult with our focus group regarding the gathered use cases, i.e., which of them should be eventually supported by the enhanced BHL system.
4. Taking into consideration outcomes from (3), to also seek our focus group’s advice on whether we should pursue the implementation of the three other features: (1) searching by automatically-generated questions, (2) searching for direct and (3) indirect associations. In the survey, each of these features was presented together with only one option for visualising search results, but for the focus group discussions, we will be suggesting more visualisation widgets. We will also emphasise that the search feature and visualisation are not tightly coupled.

Vocabularies for incorporation

The proposed search functionalities will be driven by semantic metadata which will be generated over the entire BHL collection by means of text mining methods. These will be largely based on certain vocabularies. We have listed below some of those that are currently being considered.

Taxon names and authorities/Nomenclatural acts
EOL vocabularies, e.g., Species 2000 & ITIS Catalogue of Life, NCBI Taxonomy, IUCN Red List

Geographic locations
Gazetteer Ontology, GeoNames, Marine Regions
Morphology
Uber Anatomy Ontology, Plant Structure Ontology, Phenotype and Trait Ontology, Flora Phenotype Ontology, Ontology of Biological Attributes

Habitats/Ecological interactions
Environment Ontology, Relations Ontology

Alternative visualizations

Subject-verb-object
Sunburst:
http://demeter.research.cs.dal.ca/~soto/bhl/Sequence%20sunburst-SVO/

After hovering on “feeds on” and “apples”: 
Sankey diagram
http://demeter.research.cs.dal.ca/~soto/bhl/Sankey_SVO/

Searching by subject-verb-object queries
Radial collapsible tree
http://demeter.research.cs.dal.ca/~soto/bhl/RadialReingoldTilfordTreeSVO/

Searching by subject-verb-object queries
Direct and indirect associations

http://demeter.research.cs.dal.ca/~soto/bhl/Treemaps/

I am looking for:
- [ ] Taxa
- [x] Geographical locations
- [x] Traits
- [x] Habitats

Matching documents

- Indirectly associated with [Enter]
- Via [Enter]

Reported between (via Taxa: T. geversianus)

1842 - 1900

Western Seab.  Kerguelen  Scotia Seab  Ross Seab

Draike Passage

Subantarctic

Other

Antarctic

Sub Antarctic

McMurdo Sound

South Georgia Islands

Filter
Interactive Bar charts
http://demeter.research.cs.dal.ca/~soto/bhl/Barchart-Associations/
Ask a question
Code by Jason Davies
http://www.jasondavies.com/wordtree/?source=e5e8d11a4d57bf64ccbe03aa489b8955&phrase-line=1&prefix=what&reverse=0
Opisthroproctus soleatus

characteristics of

What are the

What other species share the same

Which species/taxa are related to

In which geographical locations can I find

What other species are co-located with

In which environment does

What other species are in the same habitat as

Map-based visualisation (for geographical locations)
http://demeter.research.cs.dal.ca/~soto/bhl/maps/