

Project Acronym: ABLE  
 Version: 1.0  
 Contact: d.r.morse@open.ac.uk  
 Date: 4<sup>th</sup> November 2008



## Project Document Cover Sheet

Project Information			
<b>Project Acronym</b>	ABLE		
<b>Project Title</b>	Automatic Biodiversity Literature Enhancement		
<b>Start Date</b>	1 <sup>st</sup> October 2008	<b>End Date</b>	30 <sup>th</sup> September 2009
<b>Lead Institution</b>	The Open University		
<b>Project Director</b>	David Morse		
<b>Project Manager &amp; contact details</b>			
<b>Partner Institutions</b>	The Natural History Museum, London		
<b>Project Web URL</b>			
<b>Programme Name (and number)</b>	Circular 09/08: JISC Digitisation Programme - Enriching Digital Resources		
<b>Programme Manager</b>	Ben Showers		

Document Name			
<b>Document Title</b>	Project Plan		
<b>Reporting Period</b>			
<b>Author(s) &amp; project role</b>	ABLE Project Team		
<b>Date</b>	December 2008	<b>Filename</b>	ABLE Project Plan v1.0.pdf
<b>URL</b>			
<b>Access</b>	<input checked="" type="checkbox"/> Project and JISC internal		<input type="checkbox"/> General dissemination

Document History		
Version	Date	Comments
1.0	05/12/08	



## JISC Project Plan

### Overview of Project

#### 1. Background

The science of natural history began in the Renaissance and from it the various modern life-science disciplines have developed. Publications from the 15th century onwards provide a wealth of information, rich in observation, as natural science moved from descriptive to the hypothesis-driven science that dominates today's publication landscape. The older literature can inform management practices in modern concerns, especially biodiversity loss, land-use patterns, sustainability and climate change.

Biological taxonomy is the discipline that manages the names for living and fossil organisms, defining the relationships within and between them. It therefore provides the central infrastructure for information management in the biological sciences (Knapp et al, 2004). However, unlike most other sciences, taxonomic research and usage require access to the full range and history of publications on the subject. Publication through peer-reviewed journals is a relatively recent phenomenon. Until the 1930s, scientific observations appeared in a wide variety of publications, including learned Societies [e.g. Proceedings of the Royal Society], Institutional annual reports [e.g. *Abhandlungen der Akademie der Wissenschaften der DDR Berlin*] and encyclopaedias [e.g. Bronn's *Thier-riechs*]. Many of these publications are only held in a few libraries and are difficult to access.

The difficulty of accessing taxonomic information is a severe impediment to research and delivery of the subject's benefits (Godfray, 2002). It has also been seen as a major impediment to implementing the Convention on Biological Diversity (SCBD, 2008). Taxonomic names change over time (Roberts, 1996) and while this is both inevitable and desirable as knowledge advances, it makes information management more challenging. For example, the taxonomic hierarchies used by Catalogue of Life<sup>1</sup> and the NCBI<sup>2</sup> are different, so the collective groups that might be used in a search comprise different actual organisms.

To 'liberate' the information and data contained in the literature of the last 500 or so years, it is necessary to be able to search the documents electronically. This requires that the collections be digitised (Curry & Conner, 2007, Lyal & Weitzman, 2008), for which industrial-scale scanning projects are essential. However, current OCR (Optical Character Recognition) technology is not perfect. Errors are introduced at the scanning stage so that key words may be unrecognised by standard search techniques. To maintain, or better, increase the rate of scanning it is not practical to engage in manual validation and error checking of documents. Therefore a mechanism to reduce the impact of OCR errors and to flag such errors for human correction is necessary.

The Biodiversity Heritage Library<sup>3</sup> (BHL) is pursuing a programme to improve accessibility by digitising such works. The industrial scale of the project means that scanning takes place by volume rather than by article, so in BHL, the original scanned material must be identified by its volume without being able to identify individual articles within that volume. Although scientific tradition uses the article as the basic unit of reference, BHL cannot currently deliver that level of resolution. Lu et al (2008) have recently made substantial headway using rule-based pattern matching to recognise and analyse

---

<sup>1</sup> <http://www.catalogueoflife.org>

<sup>2</sup> <http://www.ncbi.nlm.nih.gov/Taxonomy>

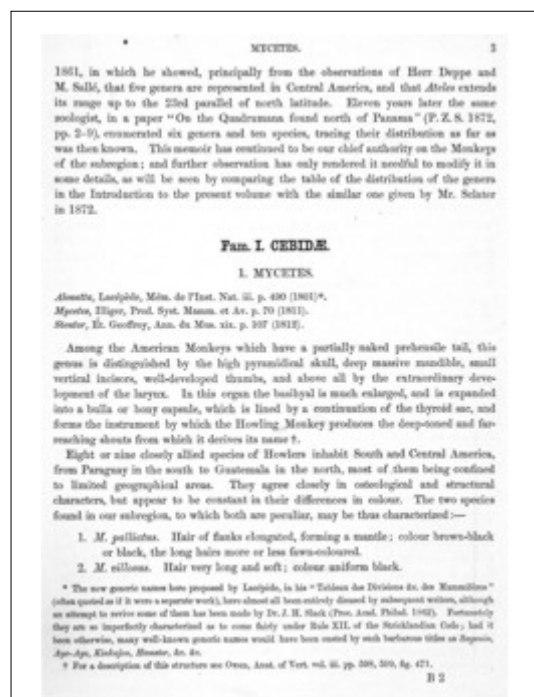
<sup>3</sup> <http://www.biodiversitylibrary.org>

volume- and issue-title pages and a machine-learning approach to detect article title blocks and thus to generate article metadata.

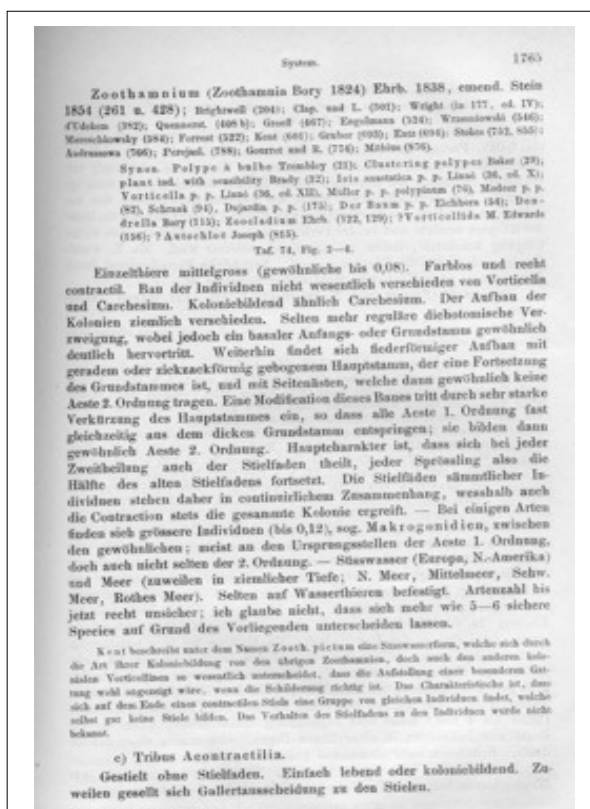
Typographical layout is an integral part of the information structure (Bringhurst, 2005), but often obeys conventions that have developed within a particular field of study (Hollingsworth et al., 2005). This structural information is independent of the language in which the text is written, so someone familiar with the principles of layout within the field of study can readily identify the section of a work that needs to be translated (Figs 1 & 2).

OCR can have high accuracy when applied to born-digital text, i.e. modern literature, where the target image has been computer-generated, as demonstrated by the PaperBrowser project (Karamanis et al., 2008), which supports curation of the FlyBase genomic database. PaperBrowser has demonstrated the value of representing layout information in a suitable markup language (SciXML). Such layout is normally self-consistent, but varies between publications.

OCR performs markedly less well on scanned pages, especially of older publications. These have old typefaces and, to the modern eye, odd layout conventions (Lu et al., 2008) so recognition accuracy is consequently worse. Errors introduced by the OCR process give potential variations in recognised



**Fig.1** A sample page from the Biologia Centrali America (Alson *et al.* 1879). This layout includes a page heading (centred capitals) on the same level as the page number; a continuation of body text from the previous page; two centred headings, one in bold and the other in capitals; a set of synonyms (not indented); body text (first line indented); two identification key questions (to differentiate species), strongly indented with outdented first lines; and two footnotes in smaller font.



**Fig 2.** A sample page from Bütschli's Protozoa (Bütschli 1887-89). Note that this has been scanned on a standard flat-bed scanner (darkening background towards the spine, on the left) and has not been de-skewed.

Project Acronym: ABLE  
Version: 1.0  
Contact: David Morse  
Date: November 2008

taxonomic names. For example, erroneous recognition of 'o' in place of a 'c' might propose the taxon *Pioa*, not a known name, rather than *Pica* (European magpie). External data sources, e.g. Catalogue of Life and NameBank associate known latinised names with common names and synonyms, but these are under active development and are incomplete, and so cannot form the only basis for term recognition. In addition, mistaking an 'o' for an 'a' can change the genus *Homa* (a hemipteran insect) into *Homo* (mankind), so that non-appearance in an existing database cannot be used to identify errors. BHL observe 35% of taxon names in scanned documents contain an error and 50% of those errors are in one or two characters (C. Freeland, pers. comm.).

Terminological variation has also been recognised as a significant problem for the management of terms in biomedical curation (Nenadić et al., 2004) where orthographic and other linguistic variations can make automated recognition of similar terms difficult (e.g. for searching document collections). The genus name *Pieris* is a valid name for both a plant (*Ericaceae*) and a butterfly (including the cabbage white), so a single name can represent two quite separate concepts. Abbreviation within text is also common, so we would seek to associate *E. coli*, for instance, with *Escherichia coli*, if it is a bacterium, or *Entamoeba coli*, if it is a protozoan. A further aim of the project, therefore, is to develop and implement a fuzzy matching system that will allow effective searching of the document collection in the face of such terminological variation based on association of terms within the document and an external reference of equivalence and membership. For instance, the noun 'magpie' is known to be a bird that carries the latinised name *Pica* which is a plausible error from the recovered term 'Pioa'. A match of *Pioa* to *Pica* is fuzzy because there is no direct relationship between *Pioa* and *Pica*. The matching will be based on ontologies to be built during the project.

## References

- Alston, E. R., Sclater, P. L., Keulemans, J. G., Smit, J., Wolf, J., Godman, F. D. C. & Salvin, O. (1879). *Biologia Centrali-Americana : Mammalia*. London.
- Bringhurst, R. (2005) *The Elements of Typographic Style*. 3rd Edition. Hartley and Marks Publishers
- Bütschli, O. (1887-89). Protozoa. Abt. III. Infusoria und System der Radiolaria. In *Klassen und Ordnung des Thiersreichs*, pp. 1098-2035. Edited by H. G. Bronn. Leipzig.
- Curry, G. B. & Connor, R. J. (2007). Automated extraction of biodiversity data from taxonomic descriptions. *Systematics Association Special Volume Series 73*, 63-81.
- Godfray, H. C. J. (2002). Challenges for taxonomy. *Nature*, Lond 417, 17-19.
- Hollingsworth, B., Lewin, I. & Tidhar, D. (2005). Retrieving Hierarchical Text Structure from Typeset Scientific Articles – a Prerequisite for E-Science Text Mining. In *Proceedings of the 4th UK e-Science All Hands Meeting*, pp. 267-273. Nottingham, UK.
- Karamanis, N., Seal, R., Lewin, I., McQuilton, P., Vlachos, A. & Gasperin, C., Drysdale R. & Briscoe, E. (2008) *Natural Language Processing in aid of FlyBase Curation*. *BMC Bioinformatics* 9.
- Knapp, S., Lamas, G., Lughadha, E. N. & Novarino, G. (2004). Stability or stasis in the names of organisms: the evolving codes of nomenclature. *Phil. Trans. Roy. Soc. Series B*: 359, 611-622.
- Lu, X., Kahle, B., Wang, J. & Giles, L. (2008). A metadata generation system for scanned scientific volumes. In *Proceedings of the 8th ACM/IEEE joint conference on Digital libraries*, pp. 167-176.
- Lyal, C.H.C. & Weitzman, L., 2008. Releasing the content of taxonomic papers: solutions to access and data mining.
- Nenadić, G., Ananiadou S. & McNaught, J. (2004). Enhancing automatic term recognition through recognition of variation. *Proc. 20th International Conference on Computational Linguistics*.
- SCBD (2008). *Guide to the Global Taxonomy Initiative*. *CBD Technical Series*, 30, pp viii + 195

## 2. Aims and Objectives

The project seeks to enhance access to a large body of scanned literature in the biodiversity domain by developing fuzzy matching of search terms, so that searching the literature is robust to errors introduced by OCR and other sources. Biological knowledge, especially taxonomic knowledge, is often presented in a stylised form, generally using typographical clues to its meaning. Therefore we aim to use typographical information and other contextual clues to identify and tag document content by type. This combination of Natural Language Processing (NLP) with typographical information extraction should be applicable in other fields that historically use structured data.

Therefore, the primary goal of the project is structural recognition, disambiguation and mark-up, from which metadata (taxon names, people's names, locations and dates) will be extracted to build indices and ontologies from the rapidly growing digital content of the BHL.

The specific objectives that the project intends to achieve include:

1. The project will take sample pages from the 10 scanned volumes to develop methods. Performance will be measured by the number of pages marked-up to the two defined levels of structure and content. The target is to process at least 2 full volumes (~600 pages) during the project.
2. The project will aim to improve the recovery of taxon names from 65% (the current rate) to better than 85%.

### 3. Overall Approach

#### Step 1: Scanning, document structure & mark-up

To make a large volume of scanned literature accessible, processing to extract index terms must be automatic, and although the BHL text is processed by Optical Character Recognition (OCR), this is an automatic, rather than a corrected, treatment. For the purposes of this project, contextual similarity will be estimated from the typographic features of the term, and surrounding linguistic cues.

The detection of text blocks on a page is normally achieved by pre-processing in the OCR package, for instance the detection of left and right margins and columns, so we expect that these image features can be quickly determined (Lebourgeois & Emptoz, 1999). We doubt that there is significance in inter-word spacing where text is justified, but in cases where it is not, such as the synonymy block in Fig. 1 – that is a significant feature, indicating that the synonymy text is not body-text. Image analysis will be undertaken using the open source application NIH Image<sup>4</sup>. OCR of image segments will be undertaken with the open source package Tesseract<sup>5</sup>. Tesseract does not capture layout structure.

In our experience OCR from scanned pages recovers certain typographical features, such as paragraphs and headings, but it does not reliably determine other features, especially indent position and the distinction between normal, bold and italic text (Bapst & Ingold, 1998). The very best modern OCR systems available, such as JSTOR, are more accurate than the desktop versions but such software is expensive and even the JSTOR system does not accurately capture typographical elements. The INOTAXA project found that scanned images of the *Biologia Centrali-Americana* to be intractable and the cheaper option was to have the content re-keyed (C. Lyal: pers. comm.). BHL scanning uses Abbyy FineReader and produces a light XML output (no styles, only words and paragraphs co-ordinates). Different disciplines tend to develop a preferred layout style (Hollingsworth et al., 2005) and the first research goal will be to identify narrative blocks, use pattern matching, machine learning and NLP techniques to identify putative functions for these blocks (e.g. title, authors, citations, heading level n, etc.; Lu et al, 2008) and add this structural mark-up to the XML file. In effect this process is conceptually equivalent to reverse-engineering a functional DTD.

Figure 2 demonstrates taxonomic information that can be obtained from the typographical structure of a document. The taxon heading (*Zoothamnium* ...) is presented in a typographical structure very similar to the body text, except that it includes a list in smaller font. The synonymy statement is also in list-form but further-indented with an aligned first line. The single centred line below the synonymy statement is a direction to the illustrations which, typically for publications of this age, are gathered into a set of plates rather than presented near the referencing text. The single paragraph of body text is followed by a comment, logically equivalent to a footnote, with the same typography as the body text except in a smaller font. This comment is at the end of the section and is followed by a heading and finally more body text.

---

<sup>4</sup> <http://rsbweb.nih.gov/nih-image>

<sup>5</sup> <http://code.google.com/p/tesseract-ocr>

The hierarchical structure of scientific publication (Hollingsworth et al., 2005) makes identification of narrative blocks fairly straightforward<sup>6</sup> (Lu et al, 2008), from which we wish to extract index information. Further disambiguation can take place by expanding abbreviated terms, for taxon names (see above), but also for standard author names defined in the botanical literature (Greuter et al, 2000). Furthermore, in-text citations can be linked to the citation listed in the bibliography or, better, to a digital version of the text, if available. This will involve parsing the citation into components (various routines exist to do this), building an OpenURL<sup>7</sup> which can be resolved using ViTAL.

In the current project this information will be included in the XML mark-up that the user interface can handle as required. We will seek to understand and mark-up individual narrative blocks such as citation objects. Other content will be marked up as and when the element can provide a contextual meaning. This will be a progressive process and will involve multiple parsing using the GoldenGATE<sup>8</sup> tool. Such techniques are being increasingly used to manage the huge volume and variation of terminology across scientific literature (Cohen and Hersh, 2005), in particular for the (difficult) task of Named Entity Recognition. Availability of the abstract collection Medline<sup>9</sup> has meant that research has generally focussed on the identification of biomedical terminology (typically gene and protein names) within plain text records; the preliminary stage of obtaining the documents through OCR and the subsequent possibility of incorrectly scanned terminology has received relatively little attention. We expect to modify an existing XML schema to accommodate the additional information described above, but we will provide translation services into at least DjVu XML, SciXML (Lewin, 2007) and NLM DTD (used by BHL). Ultimately we will work towards full mark-up in the taXMLit schema<sup>10</sup>.

## Step 2: Fuzzy matching

The fuzzy matching algorithm, based on existing work in the field of biomedical terminology, will be a two-stage mechanism (Tsuruoka and Tsujii, 2004), in which an initial match is made using the concept of edit distance (two similar terms are candidates for being variants of the same term if few edits are required to transform one to the other, for example replacing no more than two characters, or removing a hyphen). The match is then refined by considering the neighbouring terms of the various candidates. The refinement stage can be carried out with different degrees of linguistic analysis, in particular by looking at the distributional similarity of the term (Weeds et al., 2007), where a high distributional similarity means that both words are surrounded by other similar terms. For example, consider the earlier example in which the taxon *Pica* has been incorrectly interpreted as *Pioa*. If the surrounding terms have contextual link with birds (or *Aves*, *Passeriformes*, *Corvidae*) or magpies, then the name is likely to be *Pica* (European magpie) and the term can be sensibly returned against a search for *Pica*. Similarly, the context should allow a distinction to be drawn between *Pieris* as used for a plant or for a butterfly. In this last case there is no error in the OCR or the original typography but a single name representing quite separate concepts. Again, the context of the name usage should be able to resolve these instances. Weeds et al. (2007) discuss possible distributional similarity measures that could form the basis for the current project. While both authors consider deep grammatical analyses as well as shallow measures, grammatical analysis is computationally expensive, and so in the first instance this project would use only a measure of co-occurrence of neighbouring terms to estimate term similarity.

There are four main categories of interest to modern research which are significant for contextual analysis: the scientific name (taxon), geographical location and personal names (e.g. authors, collectors or expedition members) and observation date. The first three categories are outside standard language, in that they are unlikely to be found in dictionaries available to OCR software, so are the most likely areas in which OCR errors will occur (Tong & Evans, 1996). The routines in GoldenGATE to identify potential personal and place names will be used, along with additional clues,

---

<sup>6</sup> A volume analysed by Lu et al. is held at: <http://www.archive.org/details/annalsmagazineof15lond>

<sup>7</sup> <http://en.wikipedia.org/wiki/OpenURL>

<sup>8</sup> <http://idaho.ipd.uni-karlsruhe.de/GoldenGATE/>

<sup>9</sup> <http://www.nlm.nih.gov>

<sup>10</sup> <http://research.amnh.org/informatics/taxlit/schemas/taXMLit-v1-3.xsd>

Project Acronym: ABLE  
Version: 1.0  
Contact: David Morse  
Date: November 2008

such as that personal names are often associated with an in-text citation, and taxon names are generally italicised. As given strings could match against more than one potential meaning, the local context is used to determine which concept is added to the XML mark-up. As strings potentially contain OCR errors, like the Pioa example given above, it would be imprudent to try to guess the correct form in all cases. It is better to return potential matches against a user query, so Pioa should be returned against a search for *Pica*, but it is also a plausible match for *Rea*, also a passerine bird but not a magpie.

In addition to the disambiguation discussed above, the linkage information should enable association tables to be built so that a search for 'magpies' also recovers *Pica pica*, for instance. Further external data sources, particularly Catalogue of Life, NameBank and Global Names Architecture (GNA) will be used to associate latinised names with both common names and synonyms.

A further step between typographical recognition and linguistic analysis is to identify the passage in which a term appears, because certain terms are restricted or to particular types of narrative block; typographical cues such as paragraphs or columns are generally not a sufficiently accurate discriminator (Caracciolo and de Rijke, 2006). The (very efficient) TextTiling algorithm (Hearst, 1997) can be used to provide a decomposition of a document into its argumentation components rather than its physical components. The argumentation passages identified by TextTiling have been shown to be more appropriate for such linguistic analyses than the typographical structural information.

### Step 3: User interface

The final end-user interface will be hosted on the project web server and, if possible, will also be delivered through the BHL website<sup>11</sup>, with the intention of incorporating it as part of their search algorithm and reflected in through-linking of bibliographic citations.

### References

- Bapst, F. & Ingold, R. (1998). Using Typography in Document Image Analysis. In Electronic Publishing, Artistic Imaging, and Digital Typography, pp. 240. Berlin / Heidelberg: Springer.
- Caracciolo, C. & de Rijke, M. (2006) Generating and Retrieving Text Segments for Focused Access to Scientific Documents. Lecture Notes in Computer Science 3936, Springer-Verlag.
- Cohen, A. M. and Hersh, W. R. (2005) A survey of current work in biomedical text mining. Briefings in Bioinformatics 6(1):57-71.
- Greuter, W., McNeill, J., Barrie, F. R. & other authors (2000). International Code of Botanical Nomenclature ( Saint Louis Code) adopted by the 16th International Botanical Congress St. Louis, July 1999. In Regnum Vegetabile, 138, pp. XVIII, 474 p. Königstein: Koeltz Scientific Books.
- Hearst, M. A. (1997) TextTiling: segmenting text into multi-paragraph subtopic passages. Computational Linguistics 23:1.
- Hollingsworth, B., Lewin, I. & Tidhar, D. (2005). Retrieving Hierarchical Text Structure from Typeset Scientific Articles – a Prerequisite for E-Science Text Mining. In Proceedings of the 4th UK e-Science All Hands Meeting, pp. 267-273. Nottingham, UK.
- Lebourgeois, F. & Emptoz, H. (1999). Document Analysis in Gray Level and Typography Extraction Using Character Pattern Redundancies. In Fifth International Conference on Document Analysis and Recognition (ICDAR'99), pp. 177.
- Lewin, I. (2007). Using hand-crafted rules and machine learning to infer SciXML document structure. Proceedings of the 6<sup>th</sup> UK e-science All Hands Meeting.
- Lu, X., Kahle, B., Wang, J. & Giles, L. (2008). A metadata generation system for scanned scientific volumes. In Proceedings of the 8th ACM/IEEE joint conference on Digital libraries, pp. 167-176.
- Tsuruoka, Yoshimasa and Jun'ichi Tsujii. Improving the performance of dictionary-based approaches in protein name recognition. Journal of Biomedical Informatics 37 (2004).
- Tong, X. & Evans, D. A. (1996). A Statistical Approach to Automatic OCR Error Correction In Context. In Proceedings of the Fourth Workshop on Very Large Corpora, 88-100. Copenhagen, Denmark.
- Weeds, J., Dowdall, J., Schneider, G., Keller, W. & Weir, D. (2007) Using Distributional Similarity to Organise BioMedical Terminology. In F. Ibekwe-SanJuan, A. Condamines and M. T. Cabre Castellvi (eds.) Application-Driven Terminology Engineering.

---

<sup>11</sup> <http://www.biodiversitylibrary.org>

## 4. Project Outputs

The project deliverables will include:

1. Documents scanned as part of the project (approximately 10 volumes, 3000 pages). These will be selected from volumes held at the NHM.
2. A document corpus (containing both structural and content mark-up in XML) such that scientific names, geographical location, personal names and dates will be tagged. Placeholders will be left for alternatives to be inserted automatically by fuzzy matching or through the user interface. It is hoped that this document corpus will stimulate research and development of tools and technologies for information extraction from the biodiversity literature, much as the GENIA corpus (Kim et al. 2003) has in stimulating interest in information extraction from the biomedical literature.
3. Ontologies of terms for which associations have been discovered.
4. Web-based user interface providing search based on fuzzy matching.
5. A public web site which promotes the project and holds all public outputs from the project, including: the documents and indexes identified above; software developed on the project (search interfaces to the document collections, and software to perform fuzzy matching); interim and final reports; presentations etc. This website will be hosted and maintained by the Open University both during and after the project.
6. Interim and final project reports as specified by JISC.

The knowledge, and experience gained during the project will also be disseminated via conference and journal papers.

## References

Kim, J.D., Ohya, T, Tateisi, Y & Tsujii, J. (2003), GENIA corpus-a Semantically Annotated Corpus for Bio-textmining, *Bioinformatics*, 19, Suppl. 1.

## 5. Project Outcomes

There is particular urgency in the fields of climate change and biodiversity loss, where biodiversity literature can provide base-line occurrence data and reveal historical patterns of change that can inform current management practices. The work pioneered by Lu et al (2008) needs to be extended to make searching the scanned literature more straightforward for the non-specialist, both within the HE sector and in the broader scientific community.

BHL currently scans material in units of a volume without being able to identify individual articles within a volume. Scientific tradition uses the article as the basic unit of reference and, at present, BHL is not able to deliver that level of resolution, creating a barrier to access that this project will lower.

This research, combining NLP with typographical information extraction could be applicable in other fields that historically use structured data and can be expected to reveal other avenues of subsequent development. We will work with the JISC Digitisation Programme to seek other bodies of digitised literature to which the techniques implemented in this project can be applied.

Any tools and protocols developed by the project will be made available on the project web site for broad engagement by the scientific community. The GENIA corpus (Kim et al., 2003) has been instrumental in founding the field of Biomedical NLP, and it is hoped that the document collection created by the project will stimulate similar interest in information extraction from the biodiversity literature.

## References

Kim, J.D., Ohya, T, Tateisi, Y & Tsujii, J. (2003), GENIA corpus-a Semantically Annotated Corpus for Bio-textmining, *Bioinformatics*, 19, Suppl. 1.

Lu, X., Kahle, B., Wang, J. & Giles, L. (2008). A metadata generation system for scanned scientific volumes. In *Proceedings of the 8th ACM/IEEE joint conference on Digital libraries*, pp. 167-176.

## 6. Stakeholder Analysis

Stakeholder	Interest / stake	Importance
Research scientists and educators active in biodiversity-related disciplines who need to access the legacy taxonomic literature.	The difficulty in accessing taxonomic information is a barrier to research and delivery of the subject's benefits (the so-called 'taxonomic impediment') and it has also been seen as a major impediment to implementing the Convention on Biological Diversity. This project aims to facilitate access to the old taxonomic literature.	High
The JISC Digitisation Programme	As the agency funding the project, JISC have great interest in seeing that the project runs smoothly, delivers on time and within budget, that it has an impact within the Digitisation Programme, and on the wider community that JISC serves.	High
Research scientists and educators in the Natural Language Processing (NLP) community, particularly those working on information extraction.	The project will seek to scale up recent results from the information extraction research community to production use. The project outputs (see Section 5 above) will stimulate further research in information extraction from the taxonomic literature by providing a high quality document corpus that can be used as an exemplar data set.	Medium

## 7. Risk Analysis

Risk	Probability (1-5)	Severity (1-5)	Score (P x S)	Action to Prevent/Manage Risk
<b>Staffing</b>				
The project will be unable to recruit a research associate at the appropriate grade or for the required period	3	5	15	We will draw upon applicants from both computing and biological science as long as they demonstrate aptitude or experience in the other discipline.  If the delay is prolonged we will identify staff from the existing RA pool of the OU who could fill the gap until the OU's RA is appointed.
<b>Organisational</b>				
The project intends to work with natural scientists at the NHM and other institutions to identify their needs and priorities for the scanned volumes that the project will work with.	3	1	3	There may be limited opportunity to influence the choice of volumes that are scanned. Choice of volumes appears to be based on pragmatic as well as scientific grounds (the physical shape, condition and rarity of the volume being three such factors), but see External Suppliers below.

<b>Technical</b>				
The feasibility and practicality of the approach has been demonstrated (Lu et al, 2008) but to achieve the level of success desired will require human intervention	2	2	4	The software products will be designed to minimise and simplify intervention, so that even if fully autonomous operation is not achieved, the number of pages a person can process per hour will be enhanced.
Identifying document structure is more challenging than anticipated	1	3	3	The PIs are in contact with Chris Freeland (BHL) and Xiaonan Lu (Penn. State) who are generously supportive. We will develop this relationship as the project proceeds, reducing the risk of running into unanticipated problems.
Developing the fuzzy matching algorithm is more challenging than anticipated (Step 2)	2	2	4	We will make a form-filling interface available for interactive searching and annotation of the document collection so users can contribute to development of the ontologies.
<b>External suppliers</b>				
Appropriate document scans will not be available.	1	5	5	Documents scanned at the Natural History Museum are hosted by the Internet Archive and backed up at the Museum. To date, 22,700 volumes have been scanned so there are many to choose from.
<b>Legal</b>				
None identified.				

## 8. Standards

<b>Name of standard or specification</b>	<b>Version</b>	<b>Notes</b>
The principal standards for XML-based mark-up of taxonomic literature are taXMLit and TaxonX (further information on the standards and a comparison of the two are available at the links below)		We intend to work with the taXMLit standard rather than TaxonX because the former has been designed to support the mark-up of entire works. We will work towards full mark-up in the taXMLit schema, tracking developments in the schema as and when they happen. In order to accommodate the additional information that we intend to extract, we may need to propose modifications to the taXMLit schema ourselves.
Scans available from the BHL are in a variety of formats including JPEG-2000 and DjVu XML, with bibliographic details in XML MARC format.		We will work with these standards since our source material is represented in these formats.
Programming standards Java PHP		We will be developing software, websites and web services which will be implemented using a variety of means. We will adopt standard software development conventions and open standards wherever possible.
<b>TaxonX and taXMLit</b>		

## 9. Technical Development

We will work with natural scientists at the Natural History Museum and other institutions to identify their needs and priorities for the project. The intention is, therefore, to develop the prototypes by involving the domain experts in their design, and allow users to test and evaluate every iteration of the prototypes. Our approach to software development involves demonstrations and a release-early and release-often development framework (characteristic of eXtreme Programming). We will use this model to collect feedback on the functionality and usability of the prototypes we will be developing. However, the approach will remain flexible enough to accommodate changes in development as required by user requirements or changes in project deadlines, etc.

A description of the processes and any software developed will be published in the usual way but also mounted, together with the meta-data gathered through this project, on a community-based web site developed on the Scratchpad model<sup>12</sup> (again hosted on the project server). This is intended to serve as exemplar data for further research in the field. In addition the website will allow the user community to build and record exceptions and typographical rules for particular publication runs.

Note that the prototypes developed will be built to open standards to ensure generalisability and sustainability.

## 10. Intellectual Property Rights

BHL specialises in scanning literature published before 1923, which is therefore out of copyright. Such literature is also often hard to access and BHL supports a mechanism for users to nominate particular volumes for priority scanning. The tools will be available to initiatives like JSTOR to enhance accessibility to more recent, copyright material.

All outputs of the project will be placed into the public domain. The source documents (scanned images and text obtained by OCR) are licensed under a Creative Commons (Attribution-Noncommercial 2.5) licence which means that they are royalty free. Authors' rights for publications originating from the project will be retained.

## *Project Resources*

### 11. Project Partners

The project partners are The Open University and The Natural History Museum.

The Open University will be responsible for software development (particularly Steps 2 and 3 above).  
Main contact: Dr David Morse

The Natural History Museum will be the data provider (via the Biodiversity Heritage Library); assist with mark-up development and application (Step 1 above); represent one of the major user communities and provide taxonomic expertise.  
Main contact: Dr Dave Roberts

We intend to sign a Memorandum of Understanding in December 2008.

---

<sup>12</sup> <http://scratchpads.eu>

Project Acronym: ABLE  
Version: 1.0  
Contact: David Morse  
Date: November 2008

## 12. Project Management

Given the scale of the project we have not formalised reporting lines although David Morse (OU) and Dave Roberts (NHM) are the primary contacts on the ABLE project for their respective institutions. We have had and will continue to hold regular, informal project meetings at which progress will be reviewed, problems will be discussed and future actions will be identified. Between these meetings, the project continues to make extensive use of electronic means of communication (chiefly electronic mail) to support day-to-collaboration and management of the project. We have not formed local management committees but the project has full support of our respective Heads of Departments.

### Team members at The Open University

Anton Dil (a.dil@open.ac.uk) brings image analysis and text processing expertise to the project. He will be investigating ways in which the accuracy of OCR can be improved by pre-processing page images; the possibility of fine-tuning the OCR process itself, and exploring ways in which taxonomic names can be identified in free-text.

David Morse (d.r.morse@open.ac.uk) will manage the project and act as the liaison with and reporting line to JISC. He will be managing the project part-time along with other duties on this and other projects.

Alistair Willis (a.g.willis@open.ac.uk) has research interests in Natural Language Processing. He will advise on the development of fuzzy matching algorithms.

A Research Associate will be appointed to work on the project. She/he will be responsible for carrying out many of the tasks identified in the work packages.

### Team members at The Natural History Museum

Dave Roberts (workpackage6@googlemail.com) is the project lead at the Natural History Museum. He leads a workpackage, '*Unifying revisionary taxonomy on the Web*', in the EU-funded project EDIT (<http://www.e-taxonomy.eu>) which gives him the contacts and user-base within which the project can carry out usability testing of its prototypes, developed under the release-early and release-often framework identified above.

Chris Lyal (c.lyal@nhm.ac.uk) is one of the architects of the TaXMLit mark-up schema and has considerable experience of marking-up the taxonomic literature. He will advise on and help the team carry out mark-up of the volumes identified in Step 1.

Bernard Scaife (b.scaife@nhm.ac.uk) manages the Biodiversity Heritage Library's scanning operation that is based at the Natural History Museum. He will liaise with the Biodiversity Heritage Library as and when necessary, advise on scanning issues, and the choice of volumes to mark up.

## 13. Programme Support

The project has not identified any skills gaps and hence training needs yet. However, the project has a limited lifespan and strictly limited funds for training. Therefore, if training needs are identified that cannot be met in-house then we will need to receive support from the Programme in respect of training and workshops. Areas that may become an issue are project dissemination, for example.

## 14. Budget

See Appendix A.

## *Detailed Project Planning*

## 15. Workpackages

See Appendix B.

Page 12 of 22  
Document title: JISC Project Plan  
Last updated: April 2007

## 16. Evaluation Plan

The project has identified two key metrics by which the success of the project can be measured. These are the number of pages marked up, and the extent to which the recovery of taxonomic names has been improved from the current level of 65%.

Timing	Factor to Evaluate	Questions to Address	Method(s)	Measure of Success
July 2009	The number of pages marked-up to the two defined levels of (1) structure and (2) content.	How many pages have been marked up and to what level?	Count of the number of pages marked up. Assessment by the project team and others of the quality and completeness of the mark-up applied.	The target is to process at least 2 full volumes (~600 pages) during the project. (Although note that the Biodiversity Heritage Library are now estimating the average volume length to be 416 pages.)
September 2009	The recovery of taxonomic names	Has the recovery of taxonomic names been improved from its current level of 65%?	Standard metrics used in information retrieval research such as precision and recall.	The amount to which the recovery of taxonomic names has been improved from its current level of 65% to, ideally, better than 85%.
Ongoing during dissemination activities	Overall quality of the outputs being presented as part of the dissemination activities	How does the work stand up to critical appraisal by informal and formal, external peer review?	Informal and formal external (to the project) peer review	External critique of the work is minor with constructive, positive feedback provided.

## 17. Quality Plan

Output	Documents scanned during the imaging process				
Timing	Quality criteria	QA method(s)	Evidence of compliance	Quality responsibilities	Quality tools (if applicable)
March 2008	Standards applied by the Biodiversity Heritage Library	Visual inspection during and after the scanning process.	Documents made publically available through Biodiversity Heritage Library	Biodiversity Heritage Library / Internet Archive	Not known.

<b>Output</b>	Corpus of marked up documents				
<b>Timing</b>	<b>Quality criteria</b>	<b>QA method(s)</b>	<b>Evidence of compliance</b>	<b>Quality responsibilities</b>	<b>Quality tools (if applicable)</b>
August 2008	Primarily, acceptability of the corpus to peers and fitness for purpose.	Quality assurance of corpora is an active research area - we know of no widely accepted quality standards, criteria or methods	Positive feedback from the community on quality and usefulness	Project Team	N/A

<b>Output</b>	Ontologies of term associations				
<b>Timing</b>	<b>Quality criteria</b>	<b>QA method(s)</b>	<b>Evidence of compliance</b>	<b>Quality responsibilities</b>	<b>Quality tools (if applicable)</b>
August 2008	Fitness for purpose.	Formal techniques for evaluation of quality or completeness of ontologies are an under-researched area.	Positive feedback from the community on quality and usefulness	Project Team	N/A

<b>Output</b>	Web-based interface providing search based on fuzzy matching				
<b>Timing</b>	<b>Quality criteria</b>	<b>QA method(s)</b>	<b>Evidence of compliance</b>	<b>Quality responsibilities</b>	<b>Quality tools (if applicable)</b>
Ongoing	Usability; improvement in results returned over default search (See Section 16)	Usability evaluation and evaluation metrics given in Section 16.	Satisfactory results of usability evaluation and test results obtained.	Project Team	Metrics given in Section 16

<b>Output</b>	Project documentation				
<b>Timing</b>	<b>Quality criteria</b>	<b>QA method(s)</b>	<b>Evidence of compliance</b>	<b>Quality responsibilities</b>	<b>Quality tools (if applicable)</b>
September 2008	Fitness for purpose, quality, completeness.	Review by JISC Programme Managers	Sign-off by JISC Programme Managers	Project manager	JISC Project Management Guidelines

## 18. Dissemination Plan

Timing	Dissemination Activity	Audience	Purpose	Key Message
October – December 2008	Create project website	All interested parties – taxonomists at project start but increasingly the Natural Language Processing community	Raise awareness of the project; maintain interest in the project, and host resources and outputs from the project.	Project news and information about the project.
November 2008	Article describing the project for the EDIT Newsletter	Taxonomists	Raise awareness of the project	What the project aims to achieve.
July – September 2009	Present paper at a Biomedical Natural Language Processing (BioNLP) workshop – to be identified	Computer Scientists and BioNLP researchers	Raise awareness of the project and disseminate interim project results.	Describes what the project has achieved and discovered to date.
August – October 2009	Announce availability of document corpus on relevant mailing lists such as the BioNLP mailing list.	Computer Scientists and BioNLP researchers	Encourage experimentation with and development of the document corpus.	That there is a new resource that researchers in NLP can explore.
August – October 2009	Announce availability of “fuzzy matching” web service on relevant mailing lists	Researchers interested in biodiversity issues	Raise awareness of project outputs and elicit feedback from early users of the service	Describes project achievements and solicit volunteers for informal trials of the service.
October 2009	Paper / Poster presented at the annual TDWG (Taxonomic Databases Working Group) conference.	The biodiversity informatics research community.	Disseminate project outputs.	Describes what the project has achieved and discovered to date.

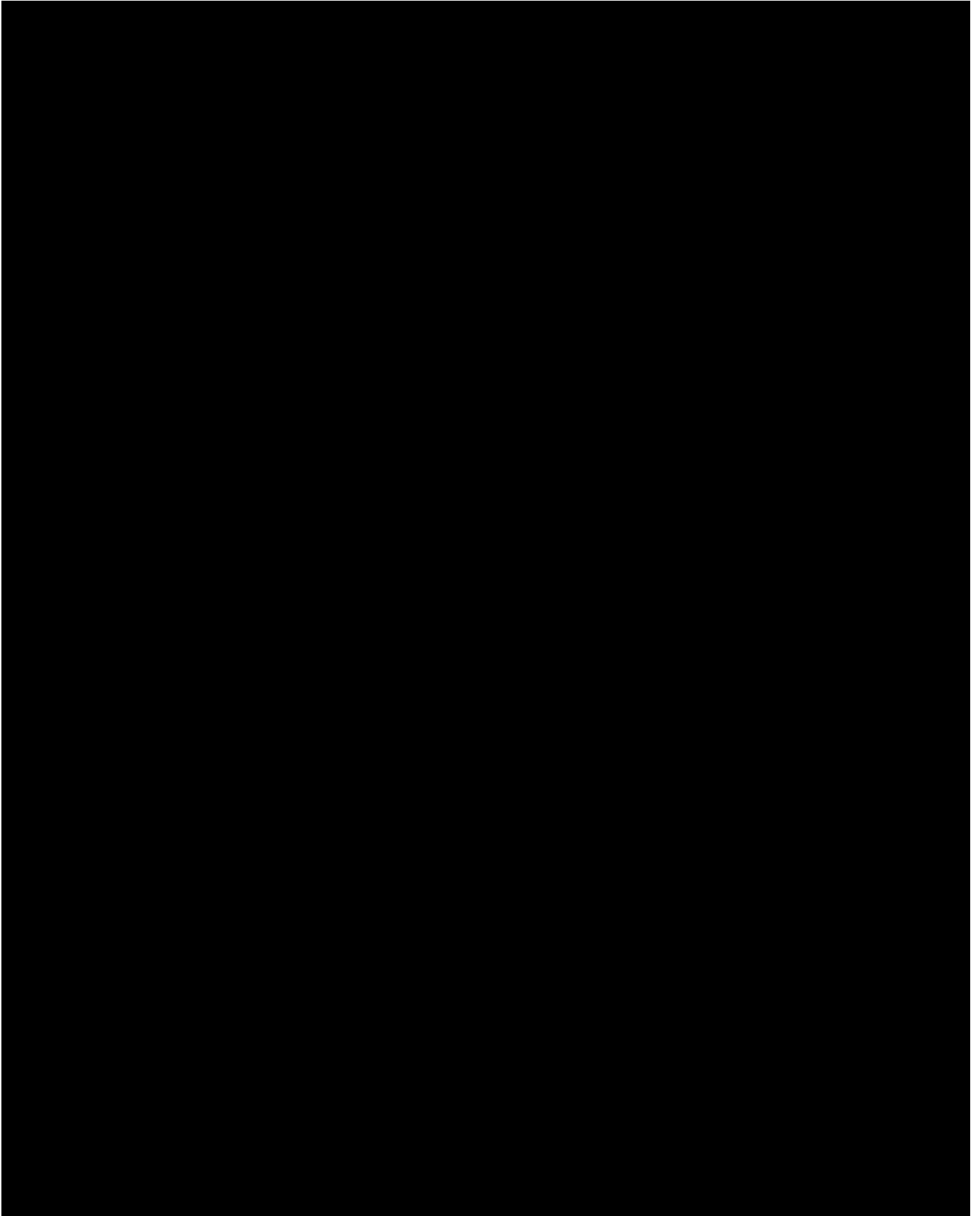
## 19. Exit and Sustainability Plans

Project Outputs	Action for Take-up & Embedding	Action for Exit
Documents scanned as part of the project	None – these will become part of the much larger collection of documents hosted by the Biodiversity Heritage Library. Careful choice of documents (linked to the marked-up versions below) will ensure that they are used by the scientific community.	None.
The corpus of marked up documents (containing both structural and content mark-up in XML).	This document corpus will stimulate research and development of tools and technologies for information extraction from the biodiversity literature. Key actions are to host the corpus on a publically accessible web-site and publicise it through conference /workshop presentations	See Dissemination Plan above.

	and relevant mailing lists.	
Ontologies of terms for which associations have been discovered.	Ensure that it is well documented and that facilities for maintenance, update and extension exist. Probably through forms on the web-based interface to the fuzzy matching search below.	Make available and document on project website.
Web-based user interface providing search based on fuzzy matching.	Ensure that good help facilities are available to support users. See also Dissemination Plan above.	Needs linking to or hosting on a portal at the Natural History Museum for maximum visibility and effectiveness.
Project web site.	Implicitly part of the Dissemination Plan in that it will be the first port of call for many respondents to project publicity.	Ensure that the project web site is maintained and hosted for the foreseeable future; linked to or is subsumed into projects in related areas.
Project documentation	Experience reports on procedures, technical developments, software products, corpora and ontologies should be documented and made available on the project web site.	Ensure that the latest versions of all project documentation are available on the project web site for reference by those engaged in similar projects.

<b>Project Outputs</b>	<b>Why Sustainable</b>	<b>Scenarios for Taking Forward</b>	<b>Issues to Address</b>
Documents scanned as part of the project	These are useful in and of themselves. Taxonomists and others interested in biodiversity have ongoing need for access to the old taxonomic literature.	The Biodiversity Heritage Library (BHL) is committed to scanning and making available as much of the old taxonomic literature as possible. Any documents scanned specifically for the project will be made available through the BHL.	None.
The corpus of marked up documents (containing both structural and content mark-up in XML).	There is considerable interest amongst biodiversity informatics researchers in mining the taxonomic literature.	The GENIA corpus has demonstrated that the existence of an annotated corpus is instrumental in advancing Natural Language Processing and Information Extraction research in particular application areas.	Ensuring that the corpus is marked up to a sufficiently high standard and that it is publicised adequately to the relevant communities.
Web-based user interface providing search based on fuzzy matching.	The interface will significantly improve the performance of searches for taxonomic names over simple string matching.	It is anticipated that this will operate alongside existing searches of the Biodiversity Heritage Library, probably across a certain subsets of the BHL document collection.	Where the web site should be hosted to maximise performance while reducing network traffic and impact on local infrastructure.
GENIA Corpus project home page: <a href="http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/">http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/</a>			

Project Acronym: ABLE  
Version: 1.0  
Contact: David Morse  
Date: November 2008



## Appendix B. Workpackages

### JISC WORK PACKAGE

WORKPACKAGES	Month	1	2	3	4	5	6	7	8	9	10	11	12
		Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
1: Project set-up													
2: Document mark-up of narrative blocks (Step 1)													
3: Fuzzy matching software development (Step 2)													
4: Search interface software development (Step 3)													
5: Project dissemination													
6: Project evaluation													
7: Formal reporting to JISC													

Project start date: [October 2008](#)

Project completion date: [September 2009](#)

Duration: [12 months](#)

Project Acronym: ABLE  
 Version: 1.0  
 Contact: David Morse  
 Date: November 2008

Workpackage and activity	Earliest start date	Latest completion date	Outputs (clearly indicate deliverables & reports in bold)	Milestone	Responsibility
--------------------------	---------------------	------------------------	--	-----------	----------------

<b>YEAR 1</b>					
<b>WORKPACKAGE 1:</b>					
<b><i>Objective: Project set-up</i></b>					
1. Establish framework of meetings and working relationships between project partners	Oct 2008	Nov 2008			NHM + OU
2. Recruit research associate to work on the project	Oct 2008	Dec 2008			NHM + OU
<b>WORKPACKAGE 2:</b>					
<b><i>Objective: Document mark-up of narrative blocks</i></b>					
3. Select documents to mark-up from volumes scanned by BHL but hosted at NHM	Oct 2008	Dec 2008	List of volumes and / or articles within selected volumes that will be marked up		NHM
4. Software tool and schema selection to support document mark-up	Nov 2008	Jan 2009			NHM + OU
5. Automated and manual mark-up of documents to identify structural components	Nov 2008	March 2009			NHM + OU
6. Automated and manual mark-up of documents to identify content	Jan 2009	August 2009	<b>Scanned document corpus (Project Output 1)</b> <b>Marked-up document corpus (Project Output 2)</b>	Sept 2009	NHM + OU

Project Acronym: ABLE  
 Version: 1.0  
 Contact: David Morse  
 Date: November 2008

Workpackage and activity	Earliest start date	Latest completion date	Outputs (clearly indicate deliverables & reports in bold)	Milestone	Responsibility
<b>WORKPACKAGE 3:</b>					
<b><u>Objective:</u> Fuzzy matching software development</b>					
7. Implement software to compute edit distances between terms	March 2009	May 2009			NHM + OU
8. Develop prototype software to refine match based on considering neighbouring terms	April 2009	July 2009			NHM + OU
9. Compile linkage information and build ontologies of terms for associations discovered	April 2009	July 2009	<b>Ontologies of term associations (Project Output 3)</b>		NHM + OU
<b>WORKPACKAGE 4:</b>					
<b><u>Objective:</u> Search interface software development</b>					
10. Implement prototype search software interface	May 2009	June 2009			NHM + OU
11. Integrate fuzzy matching software into search software interface	July 2009	August 2009	<b>A web-based user interface providing search based on fuzzy matching (Project Output 4)</b>		OU

Project Acronym: ABLE  
Version: 1.0  
Contact: David Morse  
Date: November 2008

Workpackage and activity	Earliest start date	Latest completion date	Outputs (clearly indicate deliverables & reports in bold)	Milestone	Responsibility
--------------------------	---------------------	------------------------	--	-----------	----------------

<b>WORKPACKAGE 5:</b>					
<u>Objective:</u> Project dissemination					
12. Create project website	Oct 2008	Dec 2008	<b>Project website (Project Output 5)</b>	Dec 2008	OU
13. Article describing the project for the EDIT Newsletter	Nov 2008	Nov 2008	Article in the EDIT Newsletter		NHM
14. Prepare a paper for Natural Language Processing workshop / conference	March 2009	Sept 2009	A paper, proposal for a workshop or poster will be submitted to a conference		NHM + OU
15. Prepare a paper for workshop / conference in taxonomy / biodiversity (e.g. TDWG 2009)	March 2009	Sept 2009	A paper, proposal for a workshop or poster will be submitted to a conference		NHM + OU
<b>WORKPACKAGE 6:</b>					
<u>Objective:</u> Project evaluation					
16. Assess evaluation criterion: The number of pages marked-up to the two defined levels of (1) structure and (2) content.	July 2009	August 2009			NHM + OU
17. Assess evaluation criterion: The recovery of taxonomic names	Sept 2009	Sept 2009			NHM + OU
18. Internal review of project progress towards achieving project goals and deliverables (primarily through quarterly meetings and review of reports prior to submission to JISC)	Dec 2008	Sept 2009			NHM + OU

Project Acronym: ABLE  
 Version: 1.0  
 Contact: David Morse  
 Date: November 2008

Workpackage and activity	Earliest start date	Latest completion date	Outputs (clearly indicate deliverables & reports in bold)	Milestone	Responsibility
<b>WORKPACKAGE 7:</b>					
<u>Objective:</u> Formal reporting to JISC					
19. Submit project plan & website description etc.	Oct 2008	Nov 2008	<b>Project plan, website description</b> (Project Output 6)	Nov 2008	NHM + OU
20. Submit interim project report	March 2009	April 2009	<b>Interim project report</b> (Project Output 6)	April 2009	NHM + OU
21. Prepare draft final project report	Sept 2009	Oct 2009	<b>Draft final report</b>	Oct 2009	NHM + OU
22. Submit final project report	Sept 2009	Oct 2009	<b>Final report</b> (Project Output 6)	Oct 2009	NHM + OU