


<b>Cover Sheet for Proposals</b> <i>(All sections must be completed)</i>			
<b>Name of Initiative:</b>		Enhancing Digital Resources	
<b>Name of Lead Institution:</b>		The Open University	
<b>Name of Proposed Project:</b>		Automatic Biodiversity Literature Enhancement	
<b>Name(s) of Project Partner(s):</b>		The Natural History Museum, London	
<b>Full Contact Details for Primary Contact:</b>			
<b>Name:</b> Dr David Morse			
<b>Position:</b> Senior Lecturer			
<b>Email:</b> d.r.morse@open.ac.uk			
<b>Address:</b> Department Computing Department, The Open University, Walton Hall, Milton Keynes, MK7 6AA			
<b>Tel:</b> 01908 858463			
<b>Fax:</b> 01908 652335			
<b>Length of Project:</b>		12 months	
<b>Project Start Date:</b>		1 <sup>st</sup> October 2008	
		<b>Project End Date:</b>	
		30 <sup>th</sup> September 2009	
<b>Total Funding Requested from JISC:</b>			
<b>Total Institutional Contributions:</b>			
<b>Total Funding Broken Down over Financial Years (April-March):</b>			
<b>April 08 – March 09</b>		<b>April 09 – March 10</b>	
<b>Outline Project Description:</b>			
<p>We will extend and establish the generality of the mark-up and meta data extraction from scanned literature developed by Lu et al (2008), targeting the biodiversity domain. Meta-data will focus on proper nouns (taxon, people and place names) and dates: we will enhance the searchability of those terms using associative techniques from Natural Language Processing (NLP) combined with likely Optical Character Recognition (OCR) errors, for example by allowing the recovery of Pioa against a search for Pica, provided the context of Pioa is a bird, ideally a magpie. The project will work with approximately 10 volumes that will be scanned (approx. 3000 pages) which will be rendered into several alternative XML structures. The project deliverables will be made available on the project website and through the Biodiversity Heritage Library (BHL) as exemplar data sets which will, hopefully, stimulate further research into automatic extraction of meaning from scanned literature. If fully successful the software developed here will be applied to the BHL library of over 6 million pages. BHL scanners produce a structural XML output and a small part of the project will look at the feasibility of developing software to create compatible files starting from plain image scans.</p>			
<b>I have looked at the example FOI form at Appendix A and included an FOI form in the attached bid (Indicate in relevant Box)</b>		<b>YES</b>	
<b>I have read the Circular and associated Terms and Conditions of Grant at Appendix B (Indicate in relevant Box)</b>		<b>YES</b>	

## C. Quality of Proposal and Robustness of Workplan

### Project overview

1. This 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. This project aims 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. We plan to demonstrate the generality and to extend the procedures developed by Lu et al (2008), applying them to the Biodiversity Heritage Library ([BHL](#)) scans from the Natural History Museum in London.
2. The primary goal 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. Thus the project is compatible with the programme scope (b), *Enhancement of existing collections*. The project will also generate approximately 10 volumes of scanned documents, which will be made freely available to the research community.

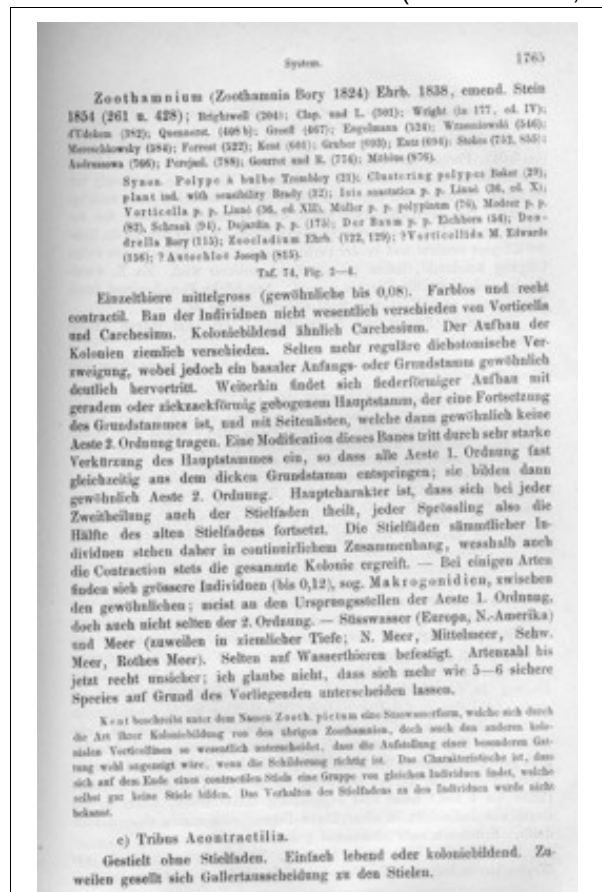
### Background

3. 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.
4. 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.
5. 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](#) and the [NCBI](#) are different, so the collective groups that might be used in a search comprise different actual organisms.
6. 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.
7. The 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 volume- and issue-title pages and a machine-learning approach to detect article title blocks and thus to generate article metadata.

8. 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).
9. 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.
10. 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 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.).
11. Terminological variation has also been recognised as a significant problem for the management of terms in biomedical curation (Nenadić et al.,



**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.

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 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.

## Project plan

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

12. 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.
13. 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](#). OCR of image segments will be undertaken with the open source package [Tesseract](#). Tesseract does not capture layout structure.
14. 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.
15. 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.
16. The hierarchical structure of scientific publication (Hollingsworth et al., 2005) makes identification of narrative blocks fairly straightforward<sup>1</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](#) which can be resolved using [ViTAL](#).

---

<sup>1</sup> A volume analysed by Lu et al. is held at: <http://tinyurl.com/6dtcth> & <http://tinyurl.com/65ecbq>

17. 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](#) 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](#) 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.

### **Step 2: Fuzzy matching**

18. 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.
19. 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, 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.
20. 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.
21. 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

22. The final end-user interface will be hosted on the project web server and, if possible, will also be delivered through the [BHL website](#), with the intention of incorporating it as part of their search algorithm and reflected in through-linking of bibliographic citations.
23. A description of the processes and any code written 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](#) (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.

### Timetable

24. The major project tasks are shown in the Gantt chart below. Document mark-up will continue as a background task from March to July in order to expand the document corpus.

Task	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Project startup (website & scan volume selection)												
Document mark-up of narrative blocks (Step 1)												
Fuzzy matching software development (Step 2)												
Search interface software development (Step 3)												
Consolidation of document corpus and ontologies												
Project (interface) evaluation												
Final report to JISC												

### Deliverables

25. Documents scanned as part of the project (approximately 10 volumes, 3000 pages). These will be selected from volumes held at the NHM (August 2009).
26. 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 (July 2009).
27. Ontologies of terms for which associations have been discovered (July 2009).
28. Web-based user interface providing search based on fuzzy matching. (July 2009)
29. 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. (Started October 2008 continuously updated until project end: September 2009).
30. The final project report to JISC. (September 2009)

### Performance metrics

31. 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.
32. The project will aim to improve the recovery of taxon names from 65% (the current rate) to better than 85%.

## Project management arrangements

33. The project will make extensive use of electronic means of communication to support day-to-collaboration and management of the project. However, we will capitalise upon the relatively close proximity of the two institutions to make quarterly visits by staff based at the OU to the Natural History Museum for project management and progress review meetings. Further visits will be made to facilitate collaboration with Museum taxonomists (one of the user communities – see below) and to use the Museum’s facilities, including the library and BHL scanner. Given the condition of some of the source texts, none of these will be removed from the Museum, hence visits may be necessary to check and/or re-scan some documents. We will apply project management arrangements that are standard to both the OU and NHM, commensurate with the scale of the project.

## Risk assessment

34. The major project-specific risks have been identified and are tabulated below. Of these risks, the first is the most serious, since significant delay in recruitment will impede progress and make it harder to produce project deliverables in a timely fashion.

Risk	Probability (1 = low; 5 = high)	Severity (1 = low; 5 = high)	Score ( P x S)	Action
The project will be unable to recruit a research associate at the appropriate grade or for the required period	2	4	8	Searching for a suitable candidate will begin immediately in anticipation of a positive decision by JISC. We will be able to draw upon applicants from computing and biological science as long as they demonstrate aptitude or experience in the other discipline.
The feasibility and practicality of the approach has been demonstrated (Lu et al, 2008) but to achieve that level of success 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.
The project will take longer than the 12 months planned	1	2	2	The size of the document collection can be reduced if progress is slow and project milestones are at risk.
Identifying document structure is more challenging than anticipated (Step 1)	1	3	3	The PIs are in contact with Chris Freeland (BHL) and Xiaonan Lu (Penn. State) who are all 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.

## Intellectual Property Rights

35. 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.
36. 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.

## D. Impact

37. 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.
38. 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.
39. 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.
40. The project web site will be maintained for a period of five years beyond the end of the project. Any tools and protocols developed by the project will be made available on the 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 this project will stimulate similar interest in information extraction from the biodiversity literature.

## E. Partnership and dissemination

41. The major stakeholders in this project are (i) JISC and the JISC Digitisation Programme; (ii) scientists active in biodiversity-related disciplines, and (iii) the Natural Language Processing (NLP) community, particularly those working on information extraction. The project team will work with JISC to engage with, capitalise upon and contribute to the experience and expertise of the JISC Digitisation Programme. We will work with natural scientists at the NHM and other institutions to identify their needs and priorities for the project. Through demonstrations and a release early and often development model (characteristic of eXtreme Programming) we will collect feedback on the functionality and usability of the prototypes we will be developing. Finally, we will scale up recent results from the information extraction research community to production use and stimulate further research in this area through the project outputs (see Section D above).
42. Our project dissemination activities include setting up, maintaining and actively promoting a project website which will be used as a vehicle for publicising the project and as a repository for the project deliverables identified in Section C above. We will promote the project through our individual networks of contacts, through attendance at scientific meetings where we will demonstrate the project deliverables and through full participation in JISC events. Key scientific meetings include e-biosphere 09, a conference on biodiversity informatics hosted by the Natural History Museum, the annual TDWG (Taxonomic Databases Working Group) conference and an expanding series of workshops in Biomedical Natural Language Processing that are loosely grouped under the collective BioNLP.

## F. Budget

43. Please see Appendix A for details of the budget.
44. The Institutional Contribution includes the directly allocated costs (OU & NHM) and some indirect costs. Other contributions include: academic and research staff time as advisors to and users of the project's outputs (OU & NHM); use of the scanner and library (NHM); technical support (OU & NHM) and recruitment costs. The benefits to the partner institutions include: gains in their reputations and publicity through successful completion of the project (OU & NHM). An outcome of the project will be improved access to the digitised literature which will provide direct benefit to the NHM. The OU will benefit from improved research and postgraduate teaching (primarily research dissertation) opportunities, through exposure to the problem domain and the project outputs (see Section D above).

## G. Previous Experience of the Project Team

45. Dr Chris Lyal leads the beetle diversity and evolution group in the Department of Entomology at the Natural History Museum, and is also the UK Focal Point for the CBD Global Taxonomy Initiative. He is a Co-PI on the *Electronic Biologia Centrali-Americana* and INOTAXA projects, and until recently was a member of the GBIF Science Committee and chair of its Electronic Catalogue of Names of Known

Organisms (ECAT) subcommittee. He is also an active taxonomist, focussing on the hyperdiverse beetle group Curculionoidea (weevils). Recent publications include:

Lyal, C.H.C. & Weitzman, L., 2008. Releasing the content of taxonomic papers: solutions to access and data mining.

Smith, R.D., Aradottir, G.I., Taylor, A. & Lyal, C., 2008, *Invasive species management – what taxonomic support is needed?* Global Invasive Species Programme, Nairobi, Kenya. 52pp.

Lyal, C.H.C., Kirk, P., Smith, D & Smith, R. (2008) The value of taxonomy to biodiversity and agriculture. *Biodiversity*, 9, 8-13.

Weitzman, A.L. & Lyal, C.H.C., 2006, INOTAXA — INTeGrated Open TAXonomic Access and the “Biologia Centrali-Americana”. *Proceedings Of The Contributed Papers Sessions Biomedical And Life Sciences Division, SLA*. 8pp.

Lyal, C.H.C., Douglas, D. & Hine, S.J., 2006, Sclerolepidia: an under-utilized character system in Curculionoidea. *Systematics and Biodiversity*, 4 (2), 203-241.

46. Dr David Morse is a Senior Lecturer in the Computing Department at the Open University and a Scientific Research Associate in the Zoology Department at The Natural History Museum, London. Prior to joining the OU, he was a lecturer in the Computing Laboratory at the University of Kent, where he was co-investigator on the very successful JISC Technology Applications Programme funded project [Mobile Computing in a Fieldwork Environment](#). Thus the project team has prior experience of managing JISC projects. David Morse has collaborated with Dave Roberts for several years on the Nomenclator project (see <http://www.nomenclator.org/>) so this project proposal builds upon an already established collaboration. As the Primary Contact for the project, David will manage the project and provide the main channel of communication with the JISC community.

Ytow, N., Morse, D.R. & Roberts, D.McL. (2001). Nomenclator: a nomenclatural history model to handle multiple taxonomic views. *Biological Journal of the Linnean Society*, 73(1): 81-98.

Pascoe, J., Ryan, N.S. & Morse, D.R. (2000). Using While Moving: HCI Issues in Fieldwork Environments. *ACM Transactions on Computer Human Interaction*, 7(3): 417-437.

47. Dr Dave Roberts is head of the Protista & Mathematics Division in the Department of Zoology at the Natural History Museum, London. He leads a workpackage, 'Unifying revisionary taxonomy on the Web', in the EU-funded project EDIT (<http://www.e-taxonomy.eu>) and currently hosts in Incoming International Fellow under the Marie Curie scheme. He currently has a user-friendly guide to identify ciliates in activated sludge in beta-testing (<http://ciliateguide.myspecies.info>).

Yi, Z., Chen, Z., Warren, A., Roberts, D. M. & other authors (2008). Molecular phylogeny of Pseudokeronopsis (Protozoa, Ciliophora, Urostylida), with reconsideration of three closely related species at inter- and intra-specific levels inferred from the small subunit ribosomal RNA gene and the ITS1-5.8S-ITS2 region sequences. *Journal of Zoology* 275, 268–275.

Roberts, D. & Chavan, V. (2008). Standard identifier could mobilize data and free time. *Nature, Lond* 453, 449-450.

Mayo, S. J., Allkin, R., Baker, W. & other authors (2008). Alpha e-taxonomy: responses from systematics community to the biodiversity crisis. *Kew Bulletin* 63, 1–16.

Tillier, S. & Roberts, D. (2006). Taxonomy on the fly in a European web project. *Nature*, 440, 24.

48. Dr Alistair Willis has been a lecturer in the Open University's Computing Department since 2003 and is a member of the Natural Language Processing theme. His background is in the representation of ambiguity, and determining correct meaning. He was a member of the Semantic Mining Network of Excellence to facilitate cross-fertilisation between scientific disciplines in medical informatics.

Willis, A., Chantree, F.J. & De Roeck, A.N. (In Press) Automatic Identification of Nocuous Ambiguity. *Research on Language and Computation*. Springer.

Willis, A. (2007) NP Coordination in Underspecified Scope Representations. *Proceedings of IWCS-7*.

Chantree, F.J., Willis, A.G., Kilgarriff, A. & De Roeck, A. (2006). Detecting dangerous coordination ambiguities using word distribution. In *Recent Advances in Natural Language Processing IV*, John Benjamins.

49. A Research Associate who will be appointed to the project, based at the OU. The Research Associate will contribute to most aspects of the work. This post will be advertised.

## 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.
- Bapst, F. & Ingold, R. (1998). Using Typography in Document Image Analysis. In *Electronic Publishing, Artistic Imaging, and Digital Typography*, pp. 240. Berlin / Heidelberg: Springer.
- 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.
- 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.
- 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.
- 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.
- 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.
- Kim, J.D., Ohya, T, Tateisi, Y & Tsujii, J. (2003), GENIA corpus-a Semantically Annotated Corpus for Bio-textmining, *Bioinformatics*, 19, Suppl. 1.
- 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.
- Nenadić, G., Ananiadou S. & McNaught, J. (2004). Enhancing automatic term recognition through recognition of variation. *Proc. 20th International Conference on Computational Linguistics*.
- Roberts, D. M. (1996). Explaining taxonomy to kids. In *Society for General Microbiology Quarterly*.
- SCBD (2008). Guide to the Global Taxonomy Initiative. *CBD Technical Series*, 30, pp viii + 195
- 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*.