From XML to XML: the why and how of making the biodiversity literature accessible to researchers

Alistair Willis[1], David King[1], David Morse[1], Anton Dil[1], Chris Lyal[2], Dave Roberts[3]

[1] Department of Computing, The Open University, Milton Keynes, UK
[2] Department of Entomology, The Natural History Museum, London, UK
[3] Department of Zoology, The Natural History Museum, London, UK
A.G.Willis@open.ac.uk, D.J.King@open.ac.uk, D.R.Morse@open.ac.uk, A.Dil@open.ac.uk,
C.lyal@nhm.ac.uk, dmr@nomencurator.org

Abstract

We present the ABLE document collection, which consists of a set of annotated volumes of the Bulletin of the British Museum (Natural History). These were developed during our ongoing work on automating the markup of scanned copies of the biodiversity literature. Such automation is required if historic literature is to be used to inform contemporary issues in biodiversity research. We consider an enhanced TEI XML markup language, which is used as an intermediate stage in translating from the initial XML obtained from Optical Character Recognition to taXMLit, the target annotation schema. The intermediate representation allows additional information from external sources such as a taxonomic thesaurus to be incorporated before the final translation into taXMLit. We give an overview of the project workflow in automating the markup process, and consider what extensions to existing markup schema will be required to best support working taxonomists. Finally, we discuss some of the particular issues which were encountered in converting between different XML formats.

1. Introduction

Biological taxonomy is the discipline that manages the names of 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). Publication through peer-reviewed journals is a relatively recent phenomenon, with scientific observations appearing in a variety of publications (e.g. learned societies such as the Proceedings of the Royal Society, institutional annual reports and encyclopaedias) until the 1930s. However, unlike most other sciences, taxonomic research and usage require access to the full range and history of publications on the subject. 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) and is seen as a major impediment to implementing the Convention on Biological Diversity (SCBD, 2008).

In this paper, we discuss the ABLE (Automatic Biodiversity Literature Enhancement) project, a collaboration between Natural Language Processing researchers and the Natural History Museum, London, which aims to improve access to collections of scanned documents from the taxonomic literature. We are providing mechanisms to automatically annotate documents from existing large scale scanning projects, such as the Biodiversity Heritage Library (BHL)[1]. The scale of BHL, which scans pages at the rate of 600,000 a month (Freeland, 2008), demonstrates the need for automatic mark-up. The current rate of scanning makes it impractical to process the output manually. For example, two biologists took nearly a year to annotate 2,500 pages even when using a tool to assist their work (Sautter et al., 2009).

The ultimate goal of the project is to support the automatic mark-up of scanned documents in taXMLit, an XML schema specialised for the biodiversity informatics community, and make the resulting document collection publicly available. By marking up information such as taxonomic names and bibliographic citations in the documents, the collection should also be of value to the Information Extraction and Information Retrieval communities. The major design decision has been to implement an interim conversion from DjVu XML to TEI XML[2] rather than attempt the production of taXMLit files in one step.

2. Background

The historical biodiversity literature can inform management practices in modern concerns, especially biodiversity loss, land-use patterns, sustainability and climate change. However, in order for this information to be usable, it is necessary to be able to access the documents electronically, in particular for searching. This requires that the collections be digitised, for which industrial-scale scanning projects are essential (Curry and Connor, 2007). 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.

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). However, 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 recog-

[1] www.biodiversitylibrary.org

[2] www.tei-c.org

nition accuracy is consequently worse. Errors introduced by the OCR process can give potential variations in recognised taxonomic names. For example, erroneous recognition of '*o*' in place of '*c*' might propose the taxon *Pioa*, not a known name, rather than *Pica* (European magpie). External data sources such as the 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. BHL have observed that 35% of taxon names in scanned documents contain an error and 50% of those errors are in one or two characters[3].

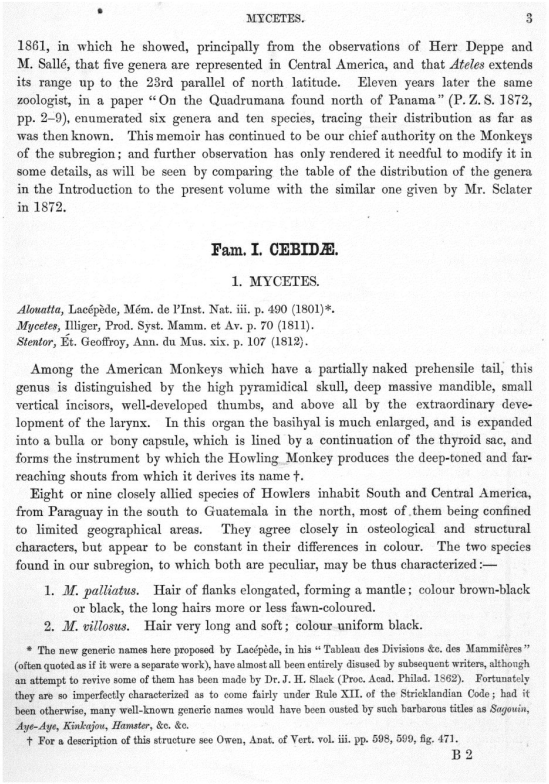
In addition, the biodiversity literature makes extensive use of layout as an integral part of its information structure (Bringhurst, 2005), but often obeys conventions that have developed within a particular field of study (Hollingsworth et al., 2005). This structural information is independent of the 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 (Figure 1). Automatic techniques that are being used increasingly to manage the huge volume and variation of terminology across scientific literature (for example GoldenGATE[4] for taxonomic Named Entity Recognition) have not generally focussed on the initial stage of obtaining the documents through OCR and the subsequent possibility of incorrectly scanned terminology.

The INOTAXA (Weitzman and Lyal, 2006) project and others have found that OCR from scanned page images recovers certain typographical features, such as paragraphs and headings, but cannot reliably determine other features, especially the indent position and the distinction between normal, bold and italic text (Bapst and Ingold, 1998). It is vital to be able to distinguish all of these features in the taxonomic literature since taxonomic hierarchies are often typeset using indent position to indicate nested groupings and taxonomic names are typically typeset in an italic font. Indeed, when attempting to mark up volumes of the *Biologia Centrali-Americana*, the process of reliably interpreting the OCR text was found to be intractable and the cheaper option was to have the content re-keyed.

The aim of the ABLE project is to support the translation of scanned documents from the XML obtained from OCR systems, to the target markup of taXMLit, automating the steps as far as possible. Scans obtained from the BHL contain markup both in DjVu and from Abbyy FineReader. DjVu produces a light XML output which contains only word and paragraph coordinates. Abbyy produces a more detailed markup, containing typographical information (such as italicisation) and a confidence measure of the proposed characters and words. Lu et al. (2008) have recently made substantial headway using rule-based pattern matching to recognise and analyse volume- and issue-title pages, although typographical cues such as paragraphs or columns are generally not always a sufficiently accurate discriminator for identifying all scientifically important terms (Caracciolo and de Rijke, 2006).

[3]Chris Freeland, personal communication

[4]idaho.ipd.uni-karlsruhe.de/GoldenGATE



MYCETES. 3

1861, in which he showed, principally from the observations of Herr Deppe and M. Sallé, that five genera are represented in Central America, and that *Ateles* extends its range up to the 23rd parallel of north latitude. Eleven years later the same zoologist, in a paper "On the Quadrumana found north of Panama" (P. Z. S. 1872, pp. 2–9), enumerated six genera and ten species, tracing their distribution as far as was then known. This memoir has continued to be our chief authority on the Monkeys of the subregion; and further observation has only rendered it needful to modify it in some details, as will be seen by comparing the table of the distribution of the genera in the Introduction to the present volume with the similar one given by Mr. Sclater in 1872.

Fam. I. CEBIDÆ.

1. MYCETES.

Alouatta, Lacépède, Mém. de l'Inst. Nat. iii. p. 490 (1801)*.
Mycetes, Illiger, Prod. Syst. Mamm. et Av. p. 70 (1811).
Stentor, Ét. Geoffroy, Ann. du Mus. xix. p. 107 (1812).

Among the American Monkeys which have a partially naked prehensile tail, this genus is distinguished by the high pyramidical skull, deep massive mandible, small vertical incisors, well-developed thumbs, and above all by the extraordinary development of the larynx. In this organ the basihyal is much enlarged, and is expanded into a bulla or bony capsule, which is lined by a continuation of the thyroid sac, and forms the instrument by which the Howling Monkey produces the deep-toned and far-reaching shouts from which it derives its name †.

Eight or nine closely allied species of Howlers inhabit South and Central America, from Paraguay in the south to Guatemala in the north, most of them being confined to limited geographical areas. They agree closely in osteological and structural characters, but appear to be constant in their differences in colour. The two species found in our subregion, to which both are peculiar, may be thus characterized:—

1. *M. palliatus*. Hair of flanks elongated, forming a mantle; colour brown-black or black, the long hairs more or less fawn-coloured.
2. *M. villosus*. Hair very long and soft; colour uniform black.

* The new generic names here proposed by Lacépède, in his "Tableau des Divisions &c. des Mammifères" (often quoted as if it were a separate work), have almost all been entirely disused by subsequent writers, although an attempt to revive some of them has been made by Dr. J. H. Slack (Proc. Acad. Philad. 1862). Fortunately they are so imperfectly characterized as to come fairly under Rule XII. of the Stricklandian Code; had it been otherwise, many well-known generic names would have been ousted by such barbarous titles as *Sagouin*, *Aye-Aye*, *Kinkajou*, *Hamster*, &c. &c.

† For a description of this structure see Owen, Anat. of Vert. vol. iii. pp. 598, 599, fig. 471.

B 2

Figure 1: A sample page from the *Biologia Centrali America*. 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.

We are building on previous work to allow markup of entities of interest to taxonomists (such as species names) by incorporating additional layout markup through extensions to existing XML schema such as DjVu XML and SciXML (Lewin, 2007). The ultimate target is full mark-up in the taXMLit schema[5].

3. Biodiversity Literature Mark-up

The key aim of marking up the biodiversity literature is to facilitate information retrieval and information extraction (see for example INOTAXA[6] and Plazi[7]). XML schemas used for biodiversity markup have generally focussed on particular elements for specific applications. For example, Linnean Core marks up taxonomic names and concepts, while SDD (Structure of Descriptive Data) focuses on particular subsets of taxonomic information. In practice, there are four key entity types required by biodiversity researchers:

[5]wiki.tdwg.org/twiki/pub/Literature/WebHome/taXMLit_v3-017a-15Jan08.xsd

[6]www.inotaxa.org

[7]www.plazi.org

- taxonomic names,
- author names (as authority for nomenclature),
- geographical locations, and
- dates

As noted in section 2., some of this data can be derived from the physical layout of the documents. For example, indentation is used to display taxonomic hierarchies as indented lists, and italicisation used to identify taxa (rendering species names such as *Escherichia coli*, or *E. coli* in the abbreviated form).
The ABLE project follows INOTAXA in using taXMLit. This is an active XML schema, whose ongoing development is to enable researchers to annotate new and legacy literature with information on:

- links to collections and use of Globally Unique Identifiers (GUIDs) to allow links between literature content, collections and collection data,
- changes at specimen level (e.g. uploading of images, comments on data) to be linked dynamically to treatment, and
- use of GUIDs to allow curators to be updated on status of specimens.

Because taXMLit is designed to be extensible and interoperable with other standard formats, it will facilitate links with additional resources such as images of specimens, and databases of names. It has been successfully used in the INOTAXA project, but because of the complexities of taXMLit the INOTAXA project has used TEI XML markup as an interim stage in the conversion process, which is an approach we have also adopted.
The TEI XML schema provides a basic set of tags focused on document structure. TEI XML is free format, although a hierarchy can be implemented within the document content by the use of `<div>` tags to identify different sections of the document. TEI XML includes features we have found to be beneficial while developing our approach to enhanced mark-up including:

- an `<expan>` tag to record the expansion of an abbreviation entered by the encoder. So, *A. viridens* can become `<expan>Attelabus</expan> viridens`,
- numerous date and time formats,
- bibliographic citations,
- semantic enhancement through the `@type` attribute,
- simple support for images and diagrams, including the ability to embed digitized versions of a graphic,
- cross-references as used extensively in taXMLit.

4. ABLE Project Workflow

The ABLE project workflow is shown in Figure 2. Documents in the BHL have mark-up in at least two forms, both of which are obtained from the ABBYY OCR package. The two formats are DjVu XML and an associated (much larger) native XML format. The DjVu XML format contains the full document text, the logical structure of the text such as page information, and the coordinates for each word's bounding box. This format has previously been used by Lu et al. (2008) to identify article boundaries within scanned volumes and so generate metadata about the volume's content.
BHL sources are shown on the left. Each scanned document is accompanied by metadata files which contain information such as the journal title and volume number. This information is encoded in a Dublin Core XML metadata file. We use XSL to extract this data from the metadata file and insert it into the `<teiHeader>` metadata elements of our TEI format output file. The source document's DjVu XML file is transformed using XSL to produce the `<text>` elements of a TEI file. This produces a valid, well-formed TEI file of the scanned original, available for manual review if desired, or in our project, for automated semantic enhancement. This file is then passed to two web services which identify proper names in the text: uBio[8] identifies genus names and OpenCalais[9] identifies country names. These services return XML files containing the identified names. We process the returned files to add the identified names to the basic TEI XML file.
uBio provides a name thesaurus to assist information retrieval over taxons. Passing the TEI file to the uBio Taxonomic Name Server allows recognition of taxa within uBio and subsequent markup into the enhanced TEI. Where a taxon is recognised in the body of the text, its presence is captured with the attribute `Explicit`:

```
<txm:TaxonHeadingParagraph
                Explicit="true">
    Trichodectes canis
</txm:TaxonHeadingParagraph>
```

with the attribute taking a value of `false` to represent additional taxonomic information obtained from uBio but not physically present in the document:

```
<txm:TaxonName Explicit="false">
    Trichodectes canis
</txm:TaxonName>
<txm:GenusName Explicit="false">
    Trichodectes
</txm:GenusName>
<txm:SpeciesEpithet Explicit="false">
    canis
</txm:SpeciesEpithet>
```

The associated GUID information is represented as an element tagged as `GUID`:

[8] www.ubio.org
[9] www.opencalais.com

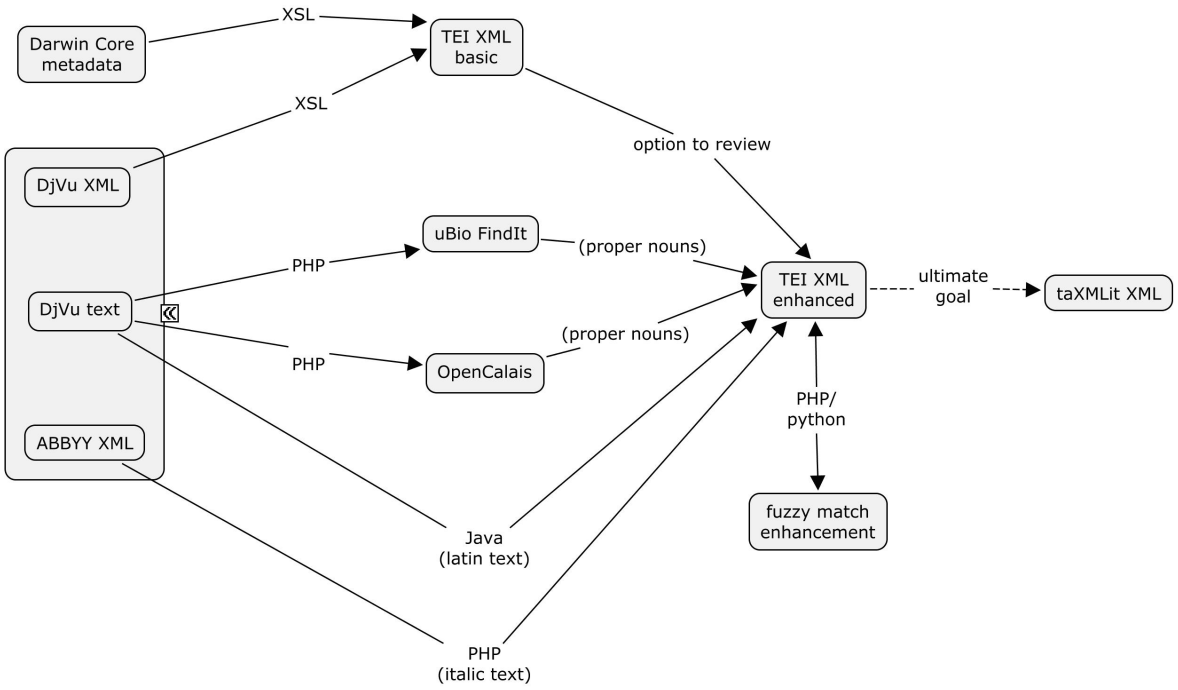


Figure 2: ABLE project enhanced mark-up workflow

```
<txm:GUID Source="uBio"
          Kind="namebankID"
          Explicit="false">
    3509767
</txm:GUID>
```

By representing the source of the information in the GUID tag, it is possible to identify alternative sources of taxon which do not appear in uBio (which will be the case for the majority of taxa). We have proposed elsewhere (Willis et al., 2009) that multiple OCR scans of a document can be used to identify possible taxa, as different OCR systems tend to interpret unknown characters differently. For example, the taxon *RHYNCHOPHOBA* was interpreted as *KHYNCHOPHOBA* and *BHYNCHOPHOKA* by ABBYY and PDF maker respectively. Similarly, italicisation information and recognition of latinate suffices (*ae*, *us*, *ii* etc) can be used to provide an indication of further potential taxa. The enhanced TEI (and so subsequent taXMLit) allows the source of such proposals to be represented, with a confidence measure to reflect the degree to which subsequent analyses might rely on that information. TaXMLit incorporates a global `SpecialistReview` attribute. This attribute can be used to highlight uncertain identifications, and comments can be added explaining the need for review.
For Named Entity Recognition beyond taxonomic names, we have attempted to exploit existing name recognition services. We have centred these attempts on the free web service OpenCalais, a set of tools developed and provided by Thomson Reuters to create semantic metadata for content. Although the service is primarily intended to enable semantic enhancement of general internet text, such as blogs, rather than scientific works, it has proved valuable in recognising entities such as Countries and other geographical information. OpenCalais has addressed many obvious difficulties, such as correctly recognising countries that no longer exist (eg. the D.D.R.), which are common when dealing with historic literature.
DjVu XML does not contain any information about the typography of the source document's text. However, we can extract this from the native ABBYY XML files. This is a character level XML schema that not only identifies the characteristics of each character but records the confidence that the identification is correct. This results in large, and to some extent unwieldy, XML files. Thus, in our work we have encountered plain text files of 1.1Mb that have ABBYY XML files in excess of 240Mb associated with them. We are particularly interested in highlighting italic text, because of the convention that taxon names are italicised. Thus, by analysing the ABBYY XML we can produce a list of candidate names for comparison against those returned by the uBio service.

4.1. taXMLit as target

The preferred goal of our project would be to mark up the source texts in an universally agreed standard. No such agreed standard yet exists within the biodiversity community, although this issue is being addressed by a working group in the Biodiversity Information Standards organisation[10]. This lack of a standard is one motivation for our decision to mark up with enhanced TEI XML, following the precedent of the INOTAXA project.
A result of this is the overloading of the TEI `hi` tag. The tag was originally intended for typographic highlights, such as italic text. In biodiversity documents marked up in TEI, this results in the italicisation of the taxon *Scutocyamus parvus* (for example) marked up as:

[10]http://www.tdwg.org

```
<tei:hi rend="italic">
    Scutocyamus parvus
</tei:hi>
```

where the `tei` namespace is used to distinguish from the taXMLit tags `txm:TaxonName`, `txm:GenusName`, `txm:SpeciesEpithet` and other taxonomic data obtained from uBio. As the taXMLit tags describe only the taxonomic information, projects such as INOTAXA have included further important semantic information in the TEI `hi` tag. For example, country names, which are not generally distinguished typographically are represented within the enhanced TEI as:

```
<tei:hi rend="Country">
    Bulgaria
</tei:hi>
```

with the purported rendering of `Country` used to represent the semantic information that *Bulgaria* is a country, and the namespace `tei` still used to distinguish the tag from a taXMLit tag.
Future work should address this issue further, following final decisions by the Standards organisation.

5. Issues in Automation

Automating the process highlighted several issues in combining and converting between the different XML formats. The information obtained from the automation is added to our TEI files, but not as TEI. As previously noted, the ultimate goal is for full markup in taXMLit. The taXMLit schema is a very detailed, highly atomised schema. So for example, all taxon names are broken down to the appropriate rank such as genus name or species epithet. Sufficient information is obtained from the source text and uBio to achieve this with the taxon names. Therefore, the taxon name information is encoded with the taXMLit `<TaxonHeading>` element to the TEI file through use of XML namespaces.

Namespaces Tags are drawn from several XML schemas within one file, so the usual practice of placing all tags within the default namespace will not work. We have adopted `tei` for TEI tags, and `txm` for taXMLit tags, as shown in Figure 3.

Matters become more complicated when language tags are applied because TEI makes use of the XML tag set for the language attribute (Figure 4).

Identified names The output from the online services needs further processing before being applied to the TEI file to remove false identifications such as the one shown in Figure 5.

Choice of matching elements Where there is a choice, representing this can be straightforward, as the two examples in Figure 6 show.

Matters are more complicated when semantic enhancements are applied to the basic TEI file. To make future conversion to taXMLit easier, mark up uses the appropriate taXMLit tags, and namespaces permit the use of the different XML schemas within one file (Figure 7).

```
<tei:p>Table 1 Genera of Triozidae
with type-species, numbers of
species, distribution and host plant
data.  Numbers of <tei:lb/>species
recorded in parenthesis under one
zoogeographical region also occur in
another region.  For the purposes of
this<tei:lb/>table species previously
included under the generic names
Megatrioza, Heterotrioza and Smirnovla
are here included<tei:lb/>under
Trioza.  (Heterotrioza Dobreanu
&amp;amp; Manolache, 1962:  258;
type-species Trioza obliqua Thomson.
Megatrioza<tei:lb/>Crawford, 1915:
264; type-species M. armata Crawford.
Smirnovia Klimaszewski, 1968:  13;
type-species Trioza<tei:lb/>femoralis
Foerster.)</tei:p>
```

Figure 8: Successful markup of correct data in TEI

6. Delivered Documents

The project has produced an eleven volume document corpus containing both structural and content mark-up. The use of TEI as our XML schema means that the individual files are relatively small and hence are amenable to further enhancement through open source XML authoring tools such as XML Copy Editor. The collection consists of 11 volumes of the Bulletin of the British Museum (Natural History), which are marked up in TEI-Lite with the taxonomic enhancements previously described. The volumes that were marked up are:

Bulletin of the British Museum (Natural History)

Entomology series

Volume	49	50	51	52	53
Pages	436	384	422	400	328

Zoology series

Volume	27	28	35	36	44	50
Pages	430	524	408	384	428	360

The collection consists of a total of 4,504 marked up pages. The electronic versions of the marked up volumes are available from the ABLE project website `http://able.open.ac.uk/`.

7. Conclusions

We have made considerable progress towards the fully automated markup of biodiversity documents. The creation of TEI XML files from documents held by BHL is part of an established workflow within the project and produces output such as that shown in Figure 8.
We are now enhancing the basic TEI XML file, which provides document-centric information, with data-centric information through semantic mark-up using taXMLit. However, this is particularly problematic for legacy literature, as scans of originals (which are possibly several hundred years old) is significantly more error-prone than for born-digital

```
<tei:TEI xmlns:tei="http://www.tei-c.org/ns/1.0"
         xmlns:txm="http://taxonomic-trial/namespace">
```

Figure 3: Namespace Definitions

```
<tei:foreign xml:lang="la">quispiam</tei:foreign>
```

Figure 4: Latin language identified by a TEI tag with an XML attribute

```
<tei:div>
<tei:p>&gt; n i</tei:p>
<tei:p>Bulletin of the</tei:p>
<tei:p>British Museum (Natural</tei:p>
<tei:p>BRITISH Mi;<tei:lb/>(NATURAL
HISTORY!<tei:p>
<tei:p>29JUN1984</tei:p>
<tei:p>Afro tropical jumping plant
lice<tei:lb/>of the family
Triozidae<tei:lb>(Homoptera:
Pslloidea)</tei:p>
<tei:p>David Hollis</tei:p>
<tei:p>Entomology series<tei:lb/>Vol 49
No 1</tei:p>
<tei:p>28 June 1984</tei:p>
<tei:pb/>
</tei:div>
```

Figure 9: Successful mark-up of erroneous data in TEI

documents. We are addressing the task of working with documents which may contain many OCR errors. However, as this refinement is an ongoing process, it is important that current markup allows uncertainty to be represented; Figure 9 demonstrates how the automated routines mark up erroneously read documents in TEI format.

Documents generated by this project are being made publicly available via the Scratchpad biodiversity network[11]. We hope that by providing an initial collection of marked up documents, and associated means for automatic document annotation, future scanned documents can be made better available for search across multiple digital libraries.

8. Acknowledgements

The authors would like to thank JISC, the UK's Joint Information Systems Committee, who funded this work under the JISC Digitisation Programme – Enriching Digital Resources.

9. References

F. Bapst and R. Ingold. 1998. Using typography in document image analysis. In *Electronic Publishing, Artistic Imaging, and Digital Typography*, Berlin/Heidelberg. Springer.

R. Bringhurst. 2005. *The Elements of Typographic Style*. Hartley and Marks, 3 edition.

C. Caracciolo and M. de Rijke, 2006. *Generating and Retrieving Text Segments for Focused Access to Scientific Documents*. Number 3936 in Lecture Notes in Computer Science. Springer-Verlag.

G. B. Curry and R. J. Connor. 2007. Automated extraction of biodiversity data from taxonomic descriptions. *Systematics Association Special Volume 73*, pages 63–81.

Chris Freeland. 2008. Evaluation of taxonomic name finding and next steps in BHL developments. Taxonomic Database Working Group.

H. Charles J. Godfray. 2002. Challenges for taxonomy. *Nature*, 417, May.

B. Hollingsworth, I. Lewin, and D. Tidhar. 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*, pages 267–273, Nottingham, UK.

N. Karamanis, R. Seal, I. Lewin, P. McQuilton, A. Vlachos, C. Gasperin, Drysdale R., and E. Briscoe. 2008. Natural language processing in aid of flybase curation. *BMC Bioinformatics*, 9.

Sandra Knapp, Gerardo Lamas, Eimear Nic Lughadha, and Gianfranco Novarino. 2004. Stability or stasis in the names of organisms: the evolving codes of nomenclature. *Philosophical Transactions of the Royal Society of London, Series B*, 359(1444), April.

I. Lewin. 2007. Using hand-crafted rules and machine learning to infer SciXML document structure. In *Proceedings of the 6th UK e-science All Hands Meeting*.

Xiaonan Lu, Brewster Kahle, James Z. Wang, and C. Lee Giles. 2008. A metadata generation system for scanned scientific volumes. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 167–176.

Secretariat of the Convention on Biological Diversity (SCBD). 2008. Guide to the global taxonomy initiative. Technical Report 30, CBD.

Guido Sautter, Klemens Böhm, Donat Agosti, and Christiana Klingenberg. 2009. Creating digital resources from legacy documents: An experience report from the biosystematics domain. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications*, Heraklion, Crete.

Anna L. Weitzman and Christopher H. C. Lyal. 2006. INOTAXA - INtegrated Open TAXonomic Access and the "Biologia Centrali-Americana". In *Proceedings of the contributed papers sessions, Biomedical and Life Sciences Division, SLA*. DBIO.

Alistair Willis, David Morse, Anton Dil, David King,

[11] http://scratchpads.eu

```
<entity>
       <nameString>The major</nameString>
       <parsedName canonical="The major">
              <component type="name" rank="genus">The</component>
              <component type="name" rank="species">major</component>
       </parsedName>
</entity>
```

Figure 5: False taxon name identification returned from uBio

Dublin Core	**TEI**

`<dc:title>`
`Bulletin of the British Museum`
`(Natural History).`
`</dc:title>`
	`<tei:titleStmt>`
	`<tei:title>`
	`Bulletin of the British Museum`
	`(Natural History).`
	`</tei:title>`
	`</tei:titleStmt>`

`<dc:publisher>`
`LONDON : BM(NH)`
`</dc:publisher>`
	`<tei:publicationStmt>`
	`<tei:publisher>`
	`LONDON : BM(NH)`
	`</tei:publisher>`
	`</tei:publicationStmt>`

Figure 6: Examples of matching Dublin Core to Text Encoding Initiative tags

Dave Roberts, and Chris Lyal. 2009. Improving search in scanned documents: Looking for OCR mismatches. In *Workshop on Advanced Technologies for Digital Libraries*, Trento, Italy.

FindIt	TXM
<entity> <nameString> Simophion calvus </nameString> </entity>	<txm:TaxonHeading> <txm:TaxonHeadingParagraph Explicit="true"> Simophion calvus </txm:TaxonHeadingParagraph> </txm:TaxonHeading>
<entity> <parsedName canonical="Simphion calvus"> <component type="name" rank="genus"> Simphion </component> </entity>	<txm:TaxonHeading> <txm:TaxonHeadingName> <txm:AlternateUsedInWork Source="current context"> <txm:GenusName Explicit="false"> Simphion </txm:GenusName> </txm:AlternateUsedInWork> </txm:TaxonHeadingName> </txm:TaxonHeading>

Figure 7: Examples of matching FindIT to taXMLit tags