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] The Open University, Walton Hall,
Milton Keynes, MK7 6AA, UK.

[2] Department of Entomology and [3] Department of Zoology,
The Natural History Museum,
London, SW7 5BD, UK

Corresponding author: a.g.willis@open.ac.uk

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 [1]. 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. The older literature, dating from 15th century, can inform management practices in modern concerns, especially biodiversity loss, land-use patterns, sustainability and climate change. However, for the information to be useful to taxonomists, it needs to be available in searchable, electronic formats. The difficulty of accessing taxonomic information is a severe impediment to research and delivery of the subject's benefits [2] and is a major impediment to implementing the Convention on Biological Diversity [3].

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) [4]. The scale of BHL, which scans pages at the rate of 600,000 a month [5], 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 [6].

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 taxon name and citation 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 [7] rather than attempt the production of taXMLit files in one step.

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 [8] and Plazi [9]). In practice, there are four key entity types required by biodiversity researchers:

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

Some of this data can be derived from the physical layout of the documents. For example, articles can be identified from page breaks and numbers. Similarly, 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).

Traditionally, 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.

The ABLE project follows INOTAXA in using taXMLit, which is an active schema in that it enables researchers to add annotation on:

- links to collections and use of Globally Unique Identifiers (GUIDs),
- 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.

TaXMLit has been successfully used in the INOTAXA project [10], but because of the complexities of taXMLit the INOTAXA project used TEI XML mark up 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 though 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.

ABLE Project Workflow

The ABLE project workflow is shown in Figure 1.

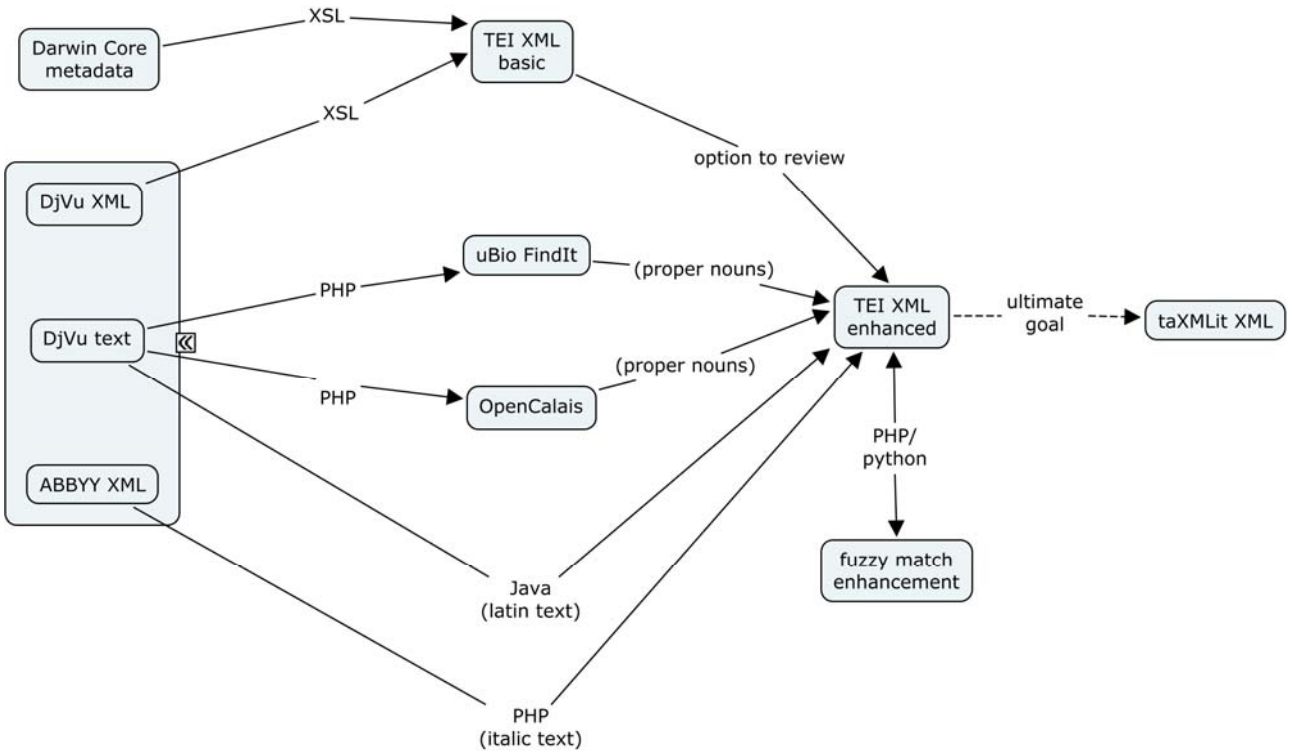

Figure 1. ABLE project enhanced mark-up workflow

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 co-ordinates for each word's bounding box. This format has previously been used by Lu *et al* [11] 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 metadatata 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 [12] identifies genus names and OpenCalais [13] 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.

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. We attempt to identify previously unseen taxa (which do not appear in uBio) by using italicisation and recognition of latinate suffices (eg. *ae, us, ii* etc). TaXMLit allows a confidence measure to be associated with such proposed taxa, which also reflects the possibility of misreadings by OCR error.

Issues in Automation

We can give some examples of the markup that deal with issues in the automation.

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 2.

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

Figure 2. Namespace definitions

Matters become more complicated when language tags are applied because TEI makes use of the XML tag set for the language attribute (Figure 3).

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

Figure 3. Latin language identified by a TEI tag with an XML attribute

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 4.

```
<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 4. False taxon name identification returned from uBio

Choice of matching elements – Where there is a choice, representing this can be straightforward, as the two examples in Figure 5 show.

DC	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 5: Examples of matching Dublin Core to Text Encoding Initiative tags

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.

FindIT (no namespace)	TXM
<entity> <nameString> Simophion calvus </nameString> </entity>	<txm:TaxonHeading> <txm:TaxonHeadingParagraph Explicit="true"> Simophion calvus </txm:TaxonHeadingParagraph> </txm:TaxonHeading>
<entity> <parsedName canonical="Simophion calvus"> <component type="name" rank="genus"> Simophion </component> </entity>	<txm:TaxonHeading> <txm:TaxonHeadingName> <txm:AlternateUsedInWork Source="current context"> <txm:GenusName Explicit="false"> Simophion </txm:GenusName> </txm:AlternateUsedInWork> </txm:TaxonHeadingName> </txm:TaxonHeading>

Figure 6: Examples of matching FindIT to taXMLit tags

The FindIT namestring represents the taxon name present in the source file and this is relatively straightforward to mark up. However, the genus name is more complicated because we need to record the appropriate context in taXMLit form too, because GenusName as a tag can appear in various places in a taXMLit file. Hence the concern raised by Sautter *et al* [14] about the deep element hierarchy in taXMLit arising from the degree of atomisation that the schema permits. This hierarchy makes searching and manipulating taXMLit encoded files relatively difficult, but equally because the mark up is so detailed it permits searches and other processing not possible in other, less detailed, schemas.

Conclusion

The ABLE project has made considerable progress towards the fully automated mark up 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 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 7. An example of successful mark up of correct data in TEI

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 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 8 demonstrates how the automated routines mark up erroneously read documents in TEI format.

```
<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: Psylloidea)</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 8. An example of successful mark-up of erroneous data in TEI

Documents generated by this project are being made publicly available via the Scratchpad [15] biodiversity network. 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.

References

[1] Knapp, S., Lamas, G., Lughadha, E.N., Novarino, G.: Stability or stasis in the names of organisms: the evolving codes of nomenclature. Philosophical Transactions of the Royal Society. Series B, 359. (2004)

[2] Godfray, H.C.J.: Challenges for taxonomy. Nature, 417. (2002)

[3] Secretariat of the Convention on Biological Diversity (SCBD): Guide to the Global Taxonomy Initiative. CBD Technical Series, 30 (2008).

[4] Biodiversity Heritage Library, http://www.biodiversitylibrary.org

[5] Freeland C.: An evaluation of taxonomic name finding and next steps in BHL developments. *Taxonomic Database Working Group*, Fremantle, Australia (2008).

[6] Sautter G, Böhm K, Agosti D, Klingenberg C.: Creating digital resources from legacy documents: An experience report from the biosystematics domain. In 6th European Semantic Web Conference, LNCS Springer-Verlag, Berlin. (2009)

[7] http://www.tei-c.org/

[8] http://www.inotaxa.org

[9] http://plazi.org

[10] Weitzman, A. L. and Lyal, C. H. C.:INOTAXA—INtegrated Open TAXonomic Access and the "Biologia Centrali-Americana" (2006)

[11] Lu, X., Kahle, B., Wang, J. Z. and Giles, C. L.: 'A metadata generation system for scanned scientific volumes', In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, ACM New York. (2008)

[12] http://www.ubio.org

[13] http://www.opencalais.com/

[14] Sautter, G., Böhm, K. and Agosti, D.: A Quantitative Comparison of XML Schemas for Taxonomic Publications, *Biodiversity Informatics*, 4, (2007)

[15] http://scratchpads.eu