Document Mark-Up for Different Users and Purposes

David King, and David R. Morse

Department of Computing and Communications,
The Open University,
Milton Keynes,
MK7 6AA, UK
{david.king, david.morse}@open.ac.uk

Abstract. Semantic enhancement of texts aids their use by researchers. However, mark-up of large bodies of text is slow and requires precious expert resources. The task could be automated if there were marked-up texts to train and test mark-up tools. This paper looks at the re-purposing of texts originally marked-up to support taxonomists to provide computer scientists with training and test data for their mark-up tools. The re-purposing highlighted some key differences in the requirements of taxonomists and computer scientists and their approaches to mark-up.

Keywords: mark-up, XML annotation, stand-off annotation, biodiversity

1 Introduction

To assess global challenges surrounding issues such as climate change and invasive species requires a baseline of historical data. One source of historical data is the Biologia Centrali-Americana (BCA). The BCA was privately issued in installments between 1879 and 1915 by F. Ducane Godman and Osbert Salvin of The Natural History Museum, London. As described in its prospectus "The work consists of 63 volumes containing 1,677 plates (of which more than 900 are coloured) depicting 18,587 subjects. The total number of species described is 50,263 of which 19,263 are described for the first time." This record of Central America's plants and animals can usefully be compared to contemporary species distributions. The BCA is available in scanned form from the Biodiversity Heritage Library [1, 2]. It has recently been re-keyed and manually marked-up by the INOTAXA [3, 4] project to help taxonomists search the contents of its 63 volumes. Curation of the marked-up volumes is continuing pending their public availability.

The manual annotation of large-scale works like the BCA is time consuming and demands expert review to curate the results. The task could benefit from automation, but attempts to automate the process face the problem of not having suitable corpora against which to develop and test the required text-mining tools.

One project, ViBRANT [5], seeks to use INOTAXA's re-keyed data to produce a corpus to support the development of text-mining tools for biodiversity documents. However, the apparently straightforward task of re-purposing INOTAXA's mark-up

has highlighted several issues because of the different audience requirement of the mark-up.

In this short paper we will describe the different needs of scientists in biodiversity and computing, how this affects the mark-up made to the documents, and how this in turn affects the re-working of annotations to meet the differing requirements.

2 Taxonomists' Requirements

XML is intended to bring structure to unstructured text and can be applied to scientific biodiversity documents [6, 7, 8]. As the prevailing mark-up technology, it was adopted by taxonomists, often in collaboration with colleagues from their supporting library services, to result in three leading XML schemas today [9]. All are applied directly to the source text so that the XML mark-up is inline with the original text.

TaxonX [10] is a lightweight mark-up focused on taxon treatments (description of species). It was created by an interdisciplinary group as part of Plazi [11] with the goal of modeling taxon treatments to provide a basis for data mining and extraction.

taXMLit [12] is a detailed mark-up focused on data curation, extraction and analysis. This schema was developed from TEI [13] as part of the INOTAXA project with the ambitious goal of covering all document and data content types. Hence, it offers very flexible possibilities for data mining though tagging a wide range of components within taxonomic papers.

TaxPub [14] is an extension of the National Library of Medicine DTD focused on layout and taxon names [15]. The schema was developed by Plazi in collaboration with U.S. National Center for Biotechnology Information [16]. Whereas TaxonX and taXMLit are mark-up XML schemas developed primarily to encode historical taxonomic literature, TaxPub aims to facilitate mark-up of new, born digital taxonomic publications as part of the publication process [17].

Each schema has its own strengths and weaknesses arising from the priorities of the taxonomists who developed them. TaxonX primarily models treatments, which are key data for taxonomists, but only records other data at a generic level. In contrast, the extensive tag sets of taXMLit and TaxPub permit detailed mark-up of all content elements. In practical terms, TaxonX requires the user to investigate documents at a treatment level, whereas the other two schemas enable other forms of enquiry to be accomplished as easily, such as searching by habitat. However, this flexibility is at the cost of complexity in mark-up and time required to produce it.

Achieving the full potential of XML marked-up documents requires supporting queries tailored to the schema's specific elements. These can be incorporated into a portal for ease of human use, as well as being built into web services. For TaxonX the portal is Plazi and for taXMLit the portal is INOTAXA [3]. TaxPub is not used this way, but as an enhanced archive format. TaxonX publications can be archived in PubMed Central [18] for subsequent retrieval.

The portals are also necessary for general work with the marked-up documents, because the portals can remove the inline mark-up that otherwise makes the text difficult for humans to read.

The subtly different purpose can make it difficult to convert marked-up documents across these schemas [19]. For example, taXMLit provides for divides location into three levels (locality, country and continent) whereas TaxonX and TaxPub have only 'location' as one entity to cover all levels. Hence, it is possible to convert from taXMLit to the others automatically, but it may not be possible to do the reverse. However, all three XML schemas permit the addition of data that is not in the source document. In the location example, it is unlikely that the source text explicitly mentions all three tiers of location, but this enhancement can be provided in the XML mark-up. The choice of how to enhance a source document is one of the key differences between taxonomists' view of the text and computer scientists'.

3 Computer Scientists' Requirements

Computer scientists prefer to preserve the original text intact. This allows further analysis on the text without the complications of having to allow for changes caused by the presence of inline mark-up. This approach makes reuse of the text easier too. It also permits the application of several layers of annotation covering different purposes to the text.

To meet these needs computer scientists prefer to use stand-off mark-up, in which the mark-up is held in a separate file to the source text. This does raise document management issues, such as version control across files that are avoided if both text and mark-up are in the one document. Arguably, data scientists should be able to handle such issues though.

At one time much work in this domain used XML-based stand-off annotation, following the ISO Linguistic Annotation Framework [20]. Of late however, there is a move towards a lighter weight form of annotation, exemplified in the biodiversity domain by the brat stand-off format [21] and accompanying mark-up tool [22].

Concerns such as multiple views of the document, are generally of little concern to the taxonomic community because they are focused on one use of the document, even if they do have different working practices to achieve that one use. In contrast, the authors, who are data scientists, have been looking to apply other forms of analysis to the text to determine if additional cues for accurate information extraction are available. As the original text is unaltered, it is relatively easy to apply a second layer of analysis over the existing taxonomic mark-up and search for significant overlapping patterns. This multiple application of different annotations would be far more difficult if working with inline XML.

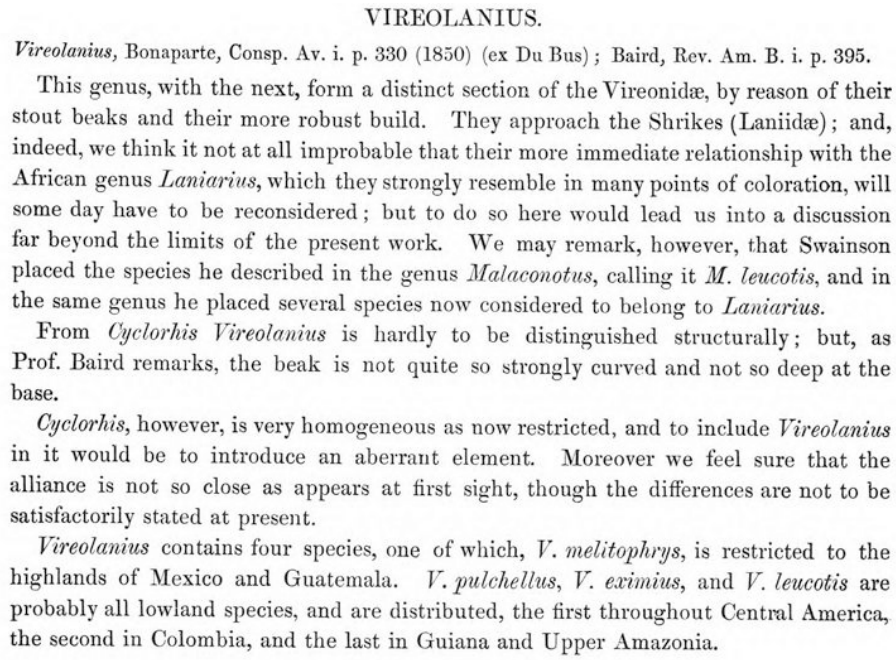
4 Working Differences in Practice – Some Examples

Figure 1 shows part of a page from the BCA's first volume about birds. It is a conventional discussion piece on a species.

Taxonomists need to know the provenance of the species being discussed. Hence the mark-up includes more than just the taxon name in the text. Typically it will contain additional information such as the name of the authority (the person who first

identified the species). An example of this form of enhanced mark-up, using a simplified version of taXMLit, is shown in figure 2. [Note the overloaded use of TEI's rend attribute which includes font rendering and taxonomic rank information.] In this example, the species Vireolanius melitophrys was first described by Du Bus, and that is recorded in the mark up of the taxon name. The mark-up is embedded in the text.

Computer scientists are interested in taxonomic names for information extraction. The originating authority is of no concern. Figure 3 shows the brat stand-off annotation format. This format gives the location of each species name in the document's page, expressed in terms of a character offset from the beginning of the page.



VIREOLANIUS.

Vireolanius, Bonaparte, Consp. Av. i. p. 330 (1850) (ex Du Bus); Baird, Rev. Am. B. i. p. 395.

This genus, with the next, form a distinct section of the Vireonidæ, by reason of their stout beaks and their more robust build. They approach the Shrikes (Laniidæ); and, indeed, we think it not at all improbable that their more immediate relationship with the African genus *Laniarius*, which they strongly resemble in many points of coloration, will some day have to be reconsidered; but to do so here would lead us into a discussion far beyond the limits of the present work. We may remark, however, that Swainson placed the species he described in the genus *Malaconotus*, calling it *M. leucotis*, and in the same genus he placed several species now considered to belong to *Laniarius*.

From *Cyclorhis Vireolanius* is hardly to be distinguished structurally; but, as Prof. Baird remarks, the beak is not quite so strongly curved and not so deep at the base.

Cyclorhis, however, is very homogeneous as now restricted, and to include *Vireolanius* in it would be to introduce an aberrant element. Moreover we feel sure that the alliance is not so close as appears at first sight, though the differences are not to be satisfactorily stated at present.

Vireolanius contains four species, one of which, *V. melitophrys*, is restricted to the highlands of Mexico and Guatemala. *V. pulchellus*, *V. eximius*, and *V. leucotis* are probably all lowland species, and are distributed, the first throughout Central America, the second in Colombia, and the last in Guiana and Upper Amazonia.

Fig. 1. Part of a page scan from the BCA.

```
<div type="taxon synonymy">
 <p elementid="BCA-aves-v3p1-2240">
  <hi rend="genus">
   <hi rend="italic">Vireolanius</hi>
  </hi>
  <hi rend="species">
   <hi rend="italic">melitophrys</hi>
  </hi>,
  <bibl rend="primary">
  <author>Du Bus</author>,
  <title>Esq. Orn.</title>
```

Fig. 2. A taxonomist's view of the taxon name

```
T25 genus 1647 1658 Vireolanius
T26 specificepithet 1659 1670 melitophrys
```

Fig. 3. A Computer Scientist's view of the taxon name

Hence, when re-purposing the INOTAXA marked-up documents to provide gold standard data for training and testing text-mining tools, some marked-up information is lost. This is also important when attempting to provide meaningful text-mined texts for taxonomists to use, if possible the text-mining tool needs to collocate the authority name in the text to add it to the taxon name mark-up.

A second discrepancy is apparent on this sample page. The genus name *Laniarus* is not marked-up in the INOTAXA supplied XML because it is an African species to which the Central American bird, that is the object of the discussion, is being compared. This work is concerned with documenting Central American residents only; hence, the African bird is not marked-up. In contrast, to train and test a text-mining tool that can accurately identify taxonomic names, all such names must be marked up; the geographical location of the species is irrelevant for this task. Therefore, the INOTAXA supplied data could not be automatically converted to a text-mining training set in stand-off format, but had to be manually curated too, looking for omissions such as this.

5 Conclusion

The two groups of scientists have different purposes for the mark-up. Taxonomists see mark-up as a means to exploit the documents' contents. Computer scientists see mark-up as part of a process to explore the documents. For taxonomists ease of document management outweighs concerns about future reuse, the opposite is true for computer scientists. Hence, the different preferences for inline and stand-off mark-up.

The same text, and even apparently the same type of entities within a text, can be interpreted differently for there can be subdivisions that are applicable to only one discipline. Taxonomists further complicate the issue by including data that is not present in the source text in their mark-up. Highlighting again the fundamental difference that taxonomists want the mark-up to support their work exploiting the documents, indeed going beyond the documents, whereas computer scientists are content to study the documents as artifacts in their own right. These differences in requirements open up interesting problems when converting from one mark-up regime to the other, as elements need to be discarded or added appropriately; a challenge to inform our continuing research within ViBRANT.

Acknowledgements. The ViBRANT project for funding this work. ViBRANT is funded by the European Union 7th Framework Programme within the Research Infrastructures group. The INOTAXA project for generously making their materials available for this work.

6 References

1. Biodiversity Heritage Library, http://www.biodiversitylibrary.org/
2. BHL Book of the Week: Biologia Centrali-Americana, http://blog.biodiversitylibrary.org/2012/09/biologia-centrali-americana-hispanic.html
3. INOTAXA, INtegrated Open TAXonomic Access, http://www.inotaxa.org/
4. Weitzman A.L., and Lyal C.H.C.: INOTAXA — INtegrated Open TAXonomic Access and the "*Biologia Centrali-Americana*". In: Proceedings Of The Contributed Papers Sessions Biomedical And Life Sciences Division, SLA. 8pp. http://units.sla.org/division/dbio/Baltimore/index.html (2006)
5. ViBRANT, Virtual Biodiversity Research and Access Network for Taxonomy, http://vbrant.eu/
6. Murray-Rust, P., and Rzepa, H.S.: Scientific publications in XML - towards a global knowledge base. Data Science 1, 84–98 (2002)
7. Cui, H.: Approaches to Semantic Mark-up for Natural Heritage Literature. In: Proceedings of the iConference 2008. http://ischools.org/conference08/pc/PA5-2_iconf08.doc (2008)
8. Parr, C.S., and Lyal, C.H.C.: Use cases for online taxonomic literature from taxonomists, conservationists, and others. In: Proceedings of TDWG Annual Conference http://www.tdwg.org/proceedings/article/view/269. (2007)
9. Penev, L., Lyal, C.H.C., Weitzman, A., Morse, D., King, D., Sautter, G., Georgiev, T., Morris, R.A., Catapano, T., and Agosti, D.: XML schemas and mark-up practices of taxonomic literature. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150, 89–116 (2011)
10. TaxonX, http://www.taxonx.org/
11. PLAZI, http://www.plazi.org/
12. Weitzman A.L., and Lyal C.H.C.: An XML schema for taxonomic literature – taXMLit - http://www.sil.si.edu/digitalcollections/bca/documentation/taXMLitv1-3Intro.pdf (2004)
13. TEI, Text Encoding Initiative, http://www.tei-c.org/index.xml
14. TaxPub, http://sourceforge.net/projects/
15. Catapano, T.: TaxPub: An extension of the NLM/NCBI Journal Publishing DTD for taxonomic descriptions. Proceedings of the Journal Article Tag Suite Conference 2010. http://www.ncbi.nlm.nih.gov/books/NBK47081/#ref2 (2010)
16. US National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/
17. Penev, L., Agosti, D., Georgiev, T., Catapano, T., Miller, J., Blagoderov, V., Roberts, D., Smith, V., Brake, I., Ryrcroft, S., Scott, B., Johnson, N., Morris, R., Sautter, G., Chavan, V., Robertson, T., Remsen, D., Stoev, P., Parr, C., Knapp, S., Kress, W., Thompson, C., and Erwin, T.: Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. ZooKeys 50, 1–16 doi: 10.3897/zookeys.50.538 (2010)
18. PubMedCentral, http://www.ncbi.nlm.nih.gov/pmc/
19. Willis, A., King, D., Morse, D., Dil, A., Lyal, C., and Roberts, D.: From XML to XML: The Why and How of Making the Biodiversity Literature Accessible to Researchers. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Valletta, Malta: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/787_Paper.pdf (2010)
20. Ide, N., and Romary, L.: International standard for a linguistic annotation framework." Journal of Natural Language Engineering 10(3‑4), 211–225. (2004)
21. brat standoff format, http://brat.nlplab.org/standoff.html
22. brat rapid annotation tool, http://brat.nlplab.org/