



Open Research Online

The Open University's repository of research publications and other research outputs

One document, many users: what happens when you re-purpose a document?

Conference or Workshop Item

How to cite:

King, David; Morse, David and Lyal, Chris (2013). One document, many users: what happens when you re-purpose a document? In: BioCuration 2013, 07-10 Apr 2013, Churchill College, Cambridge, UK.

For guidance on citations see [FAQs](#).

© 2013 ViBRANT

Version: Version of Record

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data [policy](#) on reuse of materials please consult the policies page.

oro.open.ac.uk

One document, many users

1 Biologia Centrali-Americana

To assess the challenge around issues such as climate change and invasive species requires a baseline of historic data. We are fortunate in biodiversity that such data does exist in a rich body of literature. One such source of historic data is the Biologia Centrali-Americana (BCA), which documents the plant and animal life in Central America one hundred years ago, and which can be compared to contemporary species distributions. This valuable resource has recently been re-keyed and manually marked up by the INOTAXA project (<http://www.inotaxa.org/>) and is now being curated before wider release.

4 The taxonomist's view

```
<div type="taxon synonymy">
<p element id="BCA-aves-v3p1-2240">
<hi rend="genus">
<hi rend="italic">Vi reol ani us</hi>
</hi>
<hi rend="species">
<hi rend="italic">mel itophrys</hi>
</hi>,
<bi bl rend="primary">
<author>DuBus</author>,
<title>Esq. Orn. </title>
```

The taxonomist needs to know the provenance of the taxon. Hence the mark-up is more than just the taxon name. In this example the taxon name is linked to the original describer of the taxon.

The taxonomist is also interested in typographical cues, such as the use of italic text.

Several information extraction tasks must be linked to provide a complete record.

7 Not marked up

A restricted range of entities is marked up. For example, Mexico and Guatemala are not recorded as countries.

8 What do computer scientists want?

Computer scientists prefer stand-off annotation so as to preserve the original text intact. This approach makes reuse of the text easier too.

The focus is on extracting chosen data. Frequently though this involves treating the text as independent tokens. Typographical cues are not considered, and collocation of terms is a specific task.

Note, computational linguists have a different view.

2 The re-purposing

Text mining has had some success in the recent, born-digital, bio-medical literature. Applying these approaches to the historic biodiversity literature is still in its infancy. One barrier is the lack of suitable corpora against which to develop and then test automated solutions. The ViBRANT project (<http://vbrant.eu/>) seeks to re-purpose the large volume of re-keyed data produced by INOTAXA to support the development of text mining solutions. However, this apparently straightforward task has thrown up many issues because biodiversity and computer scientists have different requirements of the mark up.

3 Additional challenges

This poster does not consider other challenges such as:

- Rekeyed data omits running headers, in the example below the re-keyed text omits VIREOLANIUS. 209
- OCR induced errors, in the example below the running header is identified as 'VIEEOLANIUS. 209', when it should read 'VIREOLANIUS. 209'; and the next line, which should read 'VIREOLANIUS.' is identified as 'VIKEOLANIUS.' Hence, we have two different incorrect recognitions.

5 Use or lose?

The genus name *Laniarius* is not marked up in the taxonomist's XML because it compares an African species to the Central American species being described. This work is concerned with documenting Central American species only.

For text mining purposes all taxonomic names are useful as training and testing data.

6 The computer scientist's view

T25 genus 1647 1658 Vi reol ani us
T26 speci fi cepi thet 1659 1670 mel itophrys

The computer scientist is concerned with one text mining operation at a time. The taxon name is not associated with an author, for example. These represent different name extraction challenges.

Contemporary computer science tools do not capture textual cues such as italics. Hence, potential semantic enhancements are more difficult to apply.

10 Conclusion

Taxonomists and Computer Scientists have different approaches to mark-up, one using in-line the other stand-off.

It is possible to develop translation tools to re-format in-line XML to stand-off mark-up and *vice versa*. A simple mapping of elements across the mark up schemes is not sufficient.

Additional stages are needed to re-purpose the mark-up to meet the different requirements of the two groups as to which elements are marked up and how they relate to other elements.

See King *et al*, in preparation.

VIREOLANIUS. 209

VIREOLANIUS.

Vireolanus, Bonaparte, Consp. Av. i. p. 330 (1850) (ex Du Bus); Baird, Rev. Am. B. i. p. 350.

This genus, with the next, form a distinct section of the *Vireolanidae*, by reason of their stout beaks and their more robust build. They approach the Shrikes (*Laniidae*); 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 *Malconotus*, calling it *M. leucotis*, and in the same genus he placed several species now considered to belong to *Laniarius*.

From *Cyclorhis Vireolanus* 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 *Vireolanus* 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.

Vireolanus 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.

a. *Subtus albus torque pectorali castaneo.*

1. **Vireolanus melitophrys.**

Vireolanus melitophrys, DuBus, Esq. Orn. p. 213; Baird, Rev. Am. B. i. p. 350; Consp. Av. i. p. 330; *Sci. & Salt. Hist.* 1860, p. 31; *Esq. Orn.* p. 13, t. 7; Baird, Rev. Am. B. i. p. 306; *Sunichrast, Mem. Bost. Soc. N. H. i.* p. 548.

Lanius chrysophrys, Licht. Mus. Berol. (fide Bonaparte).

Supra olivaceus, capite summo et cervice postica plumbeis, superciliis late flavis, stris per oculos a rictu ad nuquam extensa, altera angusta rictali nigris; subtus albus, pectore late castaneo, hypochondriis eodem colore lavatis; iride (ave viva) viridescens albo, pedibus carnis. Long. tota 6.0, alis 3.0, cauda 2.6, rostri a rictu 0.4, tarsi 0.65. (Descr. maris ex Volcan de Fuego, Guatemala. Mus. nostr.)

♀ capite summo ochraceo tincto, stris capitis lateribus fusco-nigris, subtus pectore castaneo diluore, abdomine toto ochraceo lavato a mare differt. (Descr. feminae ex Volcan de Fuego, Guatemala. Mus. nostr.)

Hab. Mexico ^{1 2 10}, Orizaba (*Botteri*²), Jalapa (*de Oca*⁴), Capulalpam (*Boucard*⁵), temperate region of Vera Cruz (*Sunichrast*⁶), valley of Mexico (*le Strange*); GUATEMALA⁶, Volcan de Fuego (*O. S. & F. D. G.*^{7 8}).

This species, which has no near ally in the genus *Vireolanus*, is restricted in its range to the highlands of Mexico and Guatemala. In the former country it has been BIOL. CENTR. AMER., Aves, Vol. I., March 1883. 27



ViBRANT is funded by the European Union 7th Framework Programme within the Research Infrastructures group. Contract no. RI-261532. Period, Dec. 2010 to Nov. 2013. Coordinator: Dr Vince Smith. E-mail: enquiries@vbrant.eu

David King¹, David Morse¹, Chris Lyal²

¹ Department of Computing, The Open University, UK.

² The Natural History Museum, London, UK.

{David.King, David.Morse} @open.ac.uk

c.lyal@nhm.ac.uk

