10/06/14

A re-publication of Flora Malesiana in semantically enriched open access edition

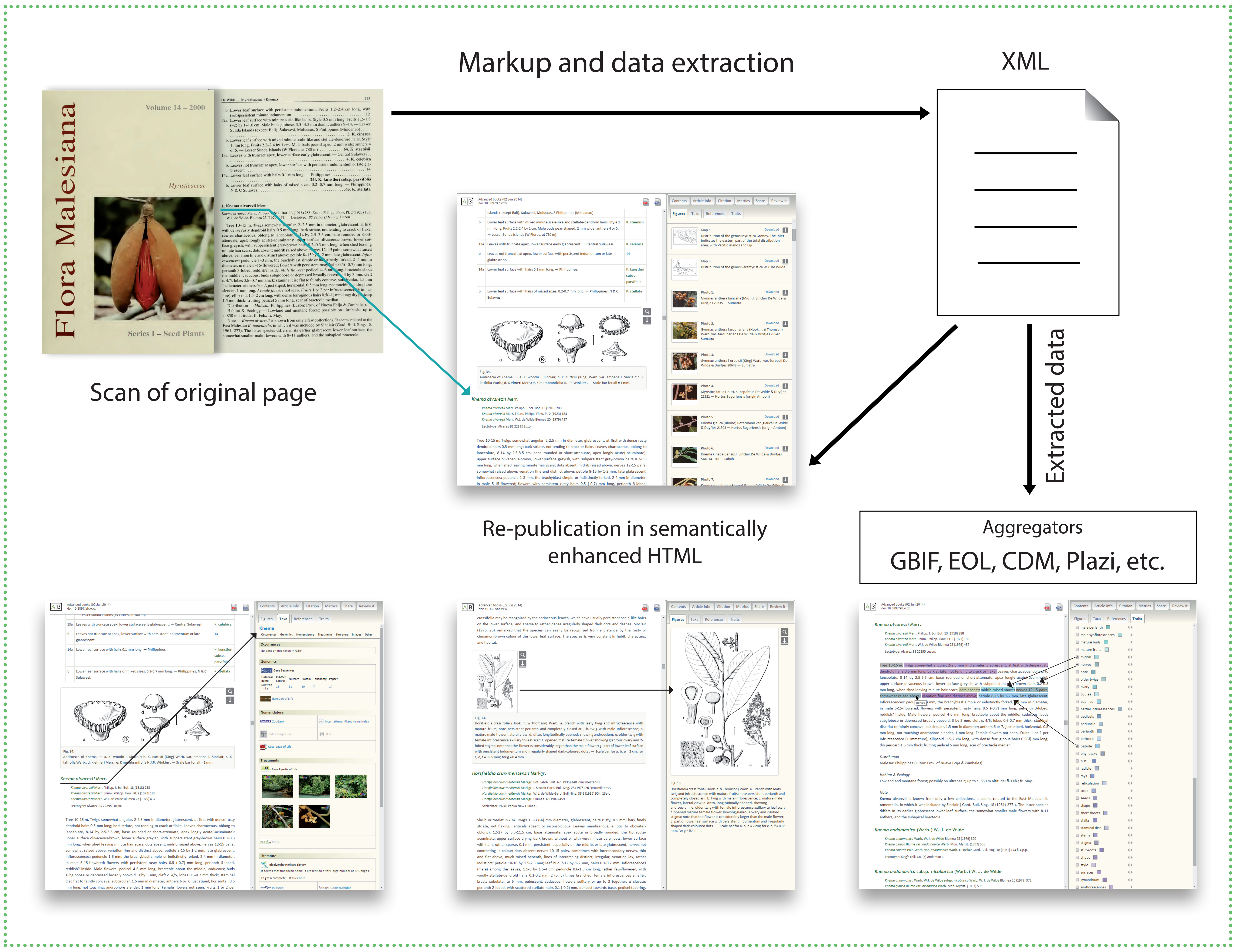
Lyubomir Penev[1,2], Teodor Georgiev[2], Jordan Biserkov[2], Thomas Hamann[3], Peter Schalk[3], Andreas Müller[4], Anton Güntsch[4], Terry Catapano[5], Donat Agosti[5]

1 Institute for Biodiversity and Ecosystem Research, Bulgarian Academy of Sciences, Sofia, Bulgaria, **2** Pensoft Publishers, Sofia, Bulgaria, **3** Netherlands Centre for Biodiversity Naturalis, Leiden, The Netherlands, **4** Freie Universität Berlin – Botanischer Garten und Botanisches Museum, Berlin-Dahlem, Germany, **5** Plazi, Zinggstrasse 16, Bern, Switzerland

Contact: Lyubomir Penev, info@pensoft.net

Background

The bulk of the taxonomic information is closed in paper-based legacy literature, especially in fundamental regional treatises such as Flora, Fauna and Mycota series. The current pilot demonstrates a workflow that enhances the marked up content of Flora Malesiana and re-publishes it into an open access, semantically enriched HTML edition available on the newly launched, Advanced Books publishing platform (http://advancedbooks.org) (Fig. 1).



Markup and data extraction

XML

Scan of original page

Extracted data

Re-publication in semantically enhanced HTML

Aggregators

GBIF, EOL, CDM, Plazi, etc.

Fig. 1. Re-published edition of volume 14 of Flora Malesiana on advancedbooks.org

The workflow

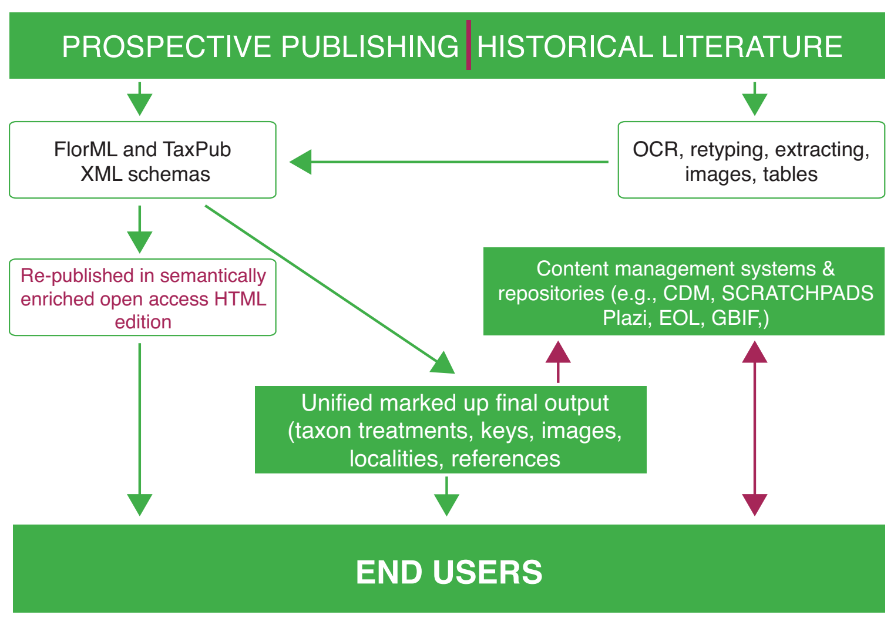
Step 1. Conversion of the printed volumes into digital text format, through scanning and OCR (Naturalis).

Step 2. Markup of generic document features and domain-specific information following the FlorML (Naturalis); export of extracted data into EDIT CDM, (BGBM).

Step 3. Conversion of the FlorML XML files into TaxPub-based XML (Plazi, Pensoft); export treatments to the Plazi Treatment Repository.

Step 4. Markup, convert and publish the XML into semantically enriched open access HTML edition (Pensoft).

Step 5. Browse, search, export and re-use of the atomized content (taxon treatments, images, morphological characters, etc.).



PROSPECTIVE PUBLISHING | HISTORICAL LITERATURE

FlorML and TaxPub XML schemas

OCR, retyping, extracting, images, tables

Re-published in semantically enriched open access HTML edition

Content management systems & repositories (e.g., CDM, SCRATCHPADS Plazi, EOL, GBIF,)

Unified marked up final output (taxon treatments, keys, images, localities, references

END USERS

Fig. 2. Multiplying the impact of the markup effort: (1) content digitized, data extracted and collated with other data; (2) content linked to external sources and re-published in semantically enriched open access; (3) Re-use and re-cycle of biodiversity data from both legacy and recently published literature

Key outputs

The present pilot demonstrates how scientifically important historical monographs, enriched with additional information from up-to-date external sources related to taxon names, species treatments, morphological characters, etc., become freely usable for anyone at any place in the world, in addition to other benefits of the digitization and markup effort such as data extraction and collation, distribution and re-use of atomized content, and archiving of different data elements in relevant repositories (Fig. 2).

PENSOFT. Naturalis Biodiversity Center BGBM PLAZI

PRO-iBIOSPHERE
WWW.PRO-iBIOSPHERE.EU

facebook.com/proibiosphere twitter.com/proibiosphere plus.google.com/108695805977454304422 linkedin.com/groups/PRO-iBiosphere-4682845