Minutes from the workshop on the implementation of the TaxPub XML schema

Paris, 19.01.2011

Milestone 6.12: Workshop on mark up tagging tools and implementations at TDWG

Leading partner: Pensoft

Related to:

M7.11 - Review of options for interactive mark up tools within the Scratchpad infrastructure

M7.10 - Agreement of standard format for community contributed bibliographies in conjunction with WP4

Partipants: Vince Smith, David Roberts, Sandra Knapp, Simon Ryrcroft, Hartwig Thomas, Terry Catapano, Donat Agosti, Vishwas Chavan, David Remsen, Teodor Georgiev, David King, David Morse, Lyubomir Penev

Compiled by:

Donat Agosti, Lyubomir Penev

January 2011

AGENDA

TaxPub Meeting Paris, 19th of January, 9.00-18.00

TaxPub – a working XML schema for prospective publishing and its compliance to other standards/schemas with regard to VIBRANT's tasks and objectives

Location: "Tour Jussieu" (the room is located in the big tower of the University on the map at http://www.upmc.fr/fr/universite/campus_et_sites/a_paris_et_en_idf/jussieu.html).

Room 1006, 10th Floor (booked from 8.30 – 18.30)

Goal: Outline the current status of development and use of TaxPub and its role in VIBRANT; outline relationship of TaxPab with other standards/tools for taxonomy mark up and publishing.

Expected participants (or representatives from the respective workpackages): Vince Smith, David Roberts, Sandra Knapp, Simon Ryrcroft, Ben Scott, David Morse, Dauvit King, Christos Arvanitidis, Terry Catapano, Donat Agosti, Guido Sautter, Vishwas Chavan, David Remsen, Teodor Georgiev, Lyubomir Penev

Organisation of the meeting: each of the topics below will be given 10 to 20 minutes for a introductory presentation (maximum 5-10 slides , in very few cases up to 15 slides!) and 30-60 minutes for discussion.

9.00-9.15 Outline of the workshops tasks: TaxPub as a cross-point between text and data (Lyubomir Penev)

9.15-10.00 Introduction to TaxPub: current status and development; TaxPub as extension of the NLM DTD; TaxPub and TaxonX - Terry Catapano (20 min)

10.00-10.30 One year experience of testing and using TaxPub – Teodor Georgiev (10 min); identifying potential future problems of TaxPub and authoring tools

10.30-11.00 Coffee break

11.005-11.30 TaxPub in Scratchpads: - Simon Ryrcroft, 10 min

11.30-12.30 DarwinCore and TaxPub; Introduction to DwC - Dave Remsen, 20 min; TaxPub compliance to DwC - Terry Catapano, 10 min.

12.30-14.00 Lunch

14.05-15.00 TaxPub and the GBIF-Pensoft Data Publication project - Vishwas Chavan, 10 min; Teodor Georgiev, 10 min

15.00-16.00 TaxPub (prospective publishing) and TaxonX (legacy literature) mark up in practice (Guido Sautter, 10 min; David King, 10 min)

16.00-16.30 Coffee Break

16.30- 17.00 TaxPub and ecology publications: identify terms of terms to be identified (open discussion)

17.00-18.00 The near future: conclusions and planning for further development and release of next TaxPub version

MINUTES:

Following are points that have been raised and decisions (bold). The agenda is attached and the ppt will be available through WP6 of ViBRANT.

Topics

Shall materials citation be marked-up? Cost is an issue; Would be advantageous if at least the type material is marked-up to a stage that includes all the elements needed to create an available name sensu the Codes? Mark-up to elements within the materials citation should be the standard for those publications being produced straight from databases (Johnson's pubs).

Peer review process: register users so that new names do not get public too early. In the proposed open peer review system, there is a fear that names could be taken and published before the author has his work published. For that reason, some sort of registration process for the peer review might be helpful.

Botany vs zoological code vs other Codes?! (how to make this happen?!)

At what point do you want break off parsing names (600+ potato names as an example); shall we forget this and look into the future. Just use names as verbatim.

Latin as language is unique to the Botanical literature, but needs to be inserted into TaxPub: **create the possibility to insert an attribute that allows to indicate the language used.**

Looking into the future of TaxPub

LSID for treatments? If treatments are treated as the core element in publishing taxonomy, then it ought to have a LSID/DOI attached to it. Where is this going to be issued? In the publication, or when it gets into a repository like Plazi? How will they be resolved and what is being served?

Abbreviations: There should be a normalized term or URI, and an expansion of the term, such as for AMNH (American Museum of Natural History). Is there such a server available?

For instamce **TaxTreat**, or **TaxTrtm**

The issue of **taxonomic names** needs some consideration.

Why is author differently tagged in nomenclature and nomenclature citation? Consistency should be maintained. Use sensu for the citation authority?

Formica rufa Linné vs *Formica rufa* Linné. Linné, 1758:354, or *Solanum nigriceps* (Linné) Knapp. Linné after rufa is the taxon authority, and Linné 1758:354 a citation of the work where the name has been

authored. In the case of botany, this is a little bit more complex since the authors of subsequent emendations are all listed, not known in zoology.

Taxon authority outside the name: Joined to the name in the same nomenclature element

Name authority needs an element to a citation? Within the tag should be an expansion out to the object.

Botany: Author name is part of the name. Xus Agosti vs Xus Knapp: there is an issue of homonymy.

Use or create <Tp:nomenclature-citation>

List of citations, that contain a name and a citation; Easier in Botany because it is there more formalized. Always the paper is cited.

What about all the **elements that refer to a specific relation** of the citation to the nomenclature species (nov.comb; stat.rev., etc)?

Make nomenclatorial element so it includes all the elements required to create an available name sensu the Codes.

Naming convention of sections of treatments

<treatment-sec> if possible, all the sections of treatment should be covered with the treatmens-sec attribute; if possible, a list of value should be created

NB! **Vocabulary of treatment sections guide; make sure that the vocabulary is linked to both, TaxPub and TaxonX, and well described and documented**

Could we use the SPM vocabularu here?

Keys

Keys use JATS elements, and they can be formatted using NLM elements (essentially to format tables)

Why not use SDD? No way known to handle SDD within TaxPub. Focus on TaxPub is to represent the key; if interest in data then make reference to an external SDD file.

Descriptive statement: not defined yet an element <character>, <state>

<Materials-citation>

Use named-content to define elements with in materials-citation.

Materials citation: Can Scratchpad deliver fully tagged materials citation? In Scratchpad the elements are in DWC and thus ought be possible the export

Release of Taxpub

March, comments and published version not later than April/Mai 2011

Get somebody to maintain taxpub

Create an advisory board

TaxonX - TaxPub relation

TaxonX for legacy publications; loose schema

Taxpub for prospective publications; very rigid; not ment to be applied retrospectively

Teodor: mark-up applications

Constraints: Requirments from NLM

Data paper

How to describe a taxonomic range

Literature references:

How to assure that the literature references are commensurable with NLM

Why data paper?

Data paper not authored through GBIF? How to integrate a data paper that has not been originated in GBIF, into the GBIF records later?

Stabilize file name, so you can find in LOKSS

How to use NLM TaxPub:

Use existing NLM DTD and see what you can do, and if possible without any extensions. You can formalize using a subset and or define certain attributes of elements. You can use a schematron to validate it.

What is a minimal threshold to qualify as a paper?

Could one visualize its content? Provide an assessment tool for data in the data paper.

Input could be a DWC-archive

Use a controlled vocabulary for the data paper; use NLM DTD (not creating an extension); use a schematron to control the document for validity; Or you can use a subset of NLM DTD. Do a profile of NLM TaxPub for data papers WITHOUT the need to establish a new element.

An application of NLM for data papers; Do a profile of TaxPub for data paper (Terry)

Scratchpads

Creating a description

Have an element in TaxPub that shows the version, especially the final published

Where is TaxPub created? You should have the option to export in various formats, also to create a DWC-archive

Import of specimen data in DwC into taxpub.

DWC-Archive (Remsen)

Sandy: use a Solanum checklist of Paraguai in coordination with DWC-Archive to be published in Phytokeys.

Remsen: shows what else could be done with a DWC-Archive.

How much does DWC-archive match up with TaxPub? Occurrence data; taxon names are modeled in both the same way? Embed DWC in taxpub: possible

DWC-Archive versioning: is happened

Terry is looking into how to assure that the name elements in TaxPub conform with DwC-Archive. An issue is the explicitation of elements that are implicit in the text. Explicit made data is not displayed, but needed for detached information. This information is stored in the little DWC-metadata file.

Linkout to a DWC-archive from TaxPub, similar to SDD.

How should authors publish their data? EXCEL, that then it will be converted to DWC-Archive. How to minimally impose us on this person who gives us the data. BUT: we should avoid to publish in supplementary file, because it has a big likelyhood to become detached.

What is happening to supplementary data at Pensoft? It goes on Pensoft server, gbif and PubMed Central?

Run an example file to create a DWC-Archive from a TaxPub paper.

Incentivation of additional data. Is creating a paper the right way to go? Seriousness comes in with publishing.

To create a DWC-Archive you need to have introductory remarks included.

Morse: Mark up

What is the effort to get something out of the mark-up? BCI used as gold standard; TaXMLit conversion, then run 3,000+pages of Bulletin Brit Mus (Nat.Hist).

The goal is to improve OCR using AI.

Compare costs of what each element costs running different methods, and or even re-OCR to get a better text.

JATS Journal Archiving Tag Suite (formerly NLM publishing and archiving DTD)