Use Cases of Existing Standards of XML Mark up Tagging and Semantic Enhancements Collected and Reviewed

Milestone 6.10 report
February 2011
Leading partner: Pensoft)

Related to:

M7.11 - Review of options for interactive mark up tools within the Scratchpad infrastructure – due May 2011

M7.10 - Agreement of standard format for community contributed bibliographies in conjunction with WP4 – due May 2011

M7.12 - A suite of test cases that will be used to test the de-duplication software - due May 2011

Compiled by:

Lyubomir Penev, Terry Catapano, Christopher Lyal, John Morse, Guido Sautter, Donat Agosti

Table of Contents

Introduction

XML mark up in taxonomy has a relatively short history. First attempts to digitize legacy taxonomic literature date back some 15 years ago using the TEI- Lite (http://www.tei-c.org/Guidelines/Customization/) schema. TEI-Lite appeared as very generic to properly model the complexity of taxonomic texts. While the broader TEI tag set could certainly be customized for retrospective conversion of legacy taxonomic literature, TEI-Lite per se is not a great fit. It is a specific customization of TEI itself and is highly generic. It has no taxonomy specific tags, and if it were to be extended, it would not be proper to call it "TEI-Lite". A version created as part of the INOTAXA project with some taxonomy tags has not been used outside that project.

A completely new world of data mining and processing of taxonomic texts through semantic XML mark up has been recently advanced by the efforts of a group of enthusiasts around Plazi (http://www.plazi.org, see also http://en.wikipedia.org/wiki/Plazi and Agosti and Egloff 2009) and INOTAXA (www.inotaxa.org). Plazi articulated some truly innovative concepts and tools, such as an electronic form of the "taxon treatment" concept (Sautter et al. 2007, Agosti et al. 2007), TaxonX and TaxPub XML schemas for either marking up legacy literature (http://www.taxonx.org, http://sourceforge.net/projects/taxonx) or to serve prospective publishing (http://sourceforge.net/projects/taxpub), respectively. A special software tool, GoldenGATE, was also developed by Plazi (together with IPD Böhm at the Karlsruhe Institute of Technology, Germany) to facilitate the process of marking up of published taxonomic works (http://plazi.org/?q=GoldenGATE).

Major efforts in this direction were also invested by Anna Weitzman (Smithsonian) and Christoper Lyal (NHM London) to create and launch the taXMLit schema within the Literature Working Group of TDWG (http://wiki.tdwg.org/Literature) and to elaborate the schema as a future TDWG standard (see also (http://www.sil.si.edu/digitalcollections/bca/documentation/taxmlitv1-3intro.pdf). TaXMLit prepared the ground for the INOTAXA project tested and implemented through an extensive mark up of Biologia Centrali-Americana (BCA) volumes and other more recent papers, currently providing access to more than 800 taxon treatments of insects and fish (Weitzman & Lyal, 2006; Lyal & Weitzman, 2008).

In an earlier paper, Sautter et al. (2007) compared seven different schemas with regard to their attractiveness for mark up of taxonomic publications: ABCD, SDD/UBIF, TaxonX, taXMLit, LinneanCore, DarwinCore and NCD (Natural Collection Description). The authors concluded that only four of them – ABCD, TaxonX , taXMLit and SDD/UBIF are appropriate for mark up of taxonomic documents; the first three of them have been evaluated as more "document-centric" oriented and the last one as clearly "data-centric". Later, TaxonX and taXMLit have been analysed comparatively in order to maximize interoperability and mapping between them (e.g., Catapano & Weitzman, presentation at TDWG 2007: http://wiki.tdwg.org/twiki/pub/Literature/WebHome/Catapano_Weitzman_Markup_Final.pdf and Weitzman, http://wiki.tdwg.org/twiki/pub/Literature/WebHome/comparisonTaxonXtaXMLit22Oct07.pdf)

The present report aims at summarizing the experience and use cases accumulated during the four years following the aforementioned analysis. The report focuses on XML schemas that have been used for mark up of taxonomic publications, that is on **schemata dealing with taxonomic literature**, being it either legacy or prospectively published. While we analyse here the three schemas most widely used for mark up of taxonomic literature, namely TaxonX, taXMLit and TaxPub, their relation to the data-centric schemas, such as SDD and DarwinCore[1], should certainly be explored as being of primary interest to ViBRANT's and Scratchpads' data management practices.

The present report is NOT a technical evaluation of the appropriateness of the different schemata. It aims at tracing forth the process of increasing interoperability between the three schemas for the goals of the ViBRANT project and beyond.

The user needs within ViBRANT for mark up will be different depending on the direction of flow for data: from within the scratchpad out (WP6) and from outside the scratchpad in (WP7). For the latter we would expect a need for greater atomization than the former, since data/information within legacy literature needs to be made available to the appropriate places within the scratchpads in a manner that enables it to be integrated with other content (i.e. seamless import). Export, however, needs only the content to be atomized to the extent the publisher requires to construct a manuscript.

The concept of „taxon treatment"

The concept of taxon treatment is fundamental for understanding the logic and goals of the taxonomy mark up process. The term has been used historically a lot by taxonomists, but its application as a concept in the mark up process of taxonomic literature has been exploited by Plazi to „atomize" taxonomic publications and explore how much of the text tagging can be done by machine either before or after publication. Following taxonomic paper publishing traditions, an initial description of a „electronic taxon treatment" can include a formal description of a taxon including sections on nomenclature, morphological characteristics, behaviour, ecology, distribution, and specimens examined (Sautter et al. 2007, Catapano 1010, Penev et al. 2010).

According to Norman Johnson's (in litt.) definition adopted by Catapano (2010), taxon treatment is a „publication or (more frequently) section of a publication documenting the features or distribution of a related group of organisms (called a "taxon", plural "taxa") in ways adhering to highly formalized conventions." Some of these are over a century old and are maintained by scientific commissions accepted by the profession. Two of the most significant are the international standard for naming animals, the International Code for Zoological Nomenclature (ICZN), and the corresponding code for plants, the International Code for Botanical Nomenclature (ICBN).

[1] DarwinCore, as well as MODS, is embedded in TaxonX as a namespace for taxonomy-specific details

The features and structure of treatments have varied across time as well as across and within publications. Despite the variation, however, a few key features are commonly found. First, and most important, is a section usually displayed as a heading presenting information related to the naming of the described species or other taxon of higher rank in one or another standard hierarchy. This "nomenclature" section contains at minimum the name of the taxon. Often following the name is an indication of whether the taxon is new to science or there has been a change in its taxonomic status, the name or names of the persons responsible for the naming, and the year of publication. Following the Heading is often one or more citations of earlier treatments, if the taxon is not new, including the original treatment citation and subsequent citations, often giving different taxonomic concepts, and including synonyms – other taxa that have been united with the taxon given in the heading. Other information, such as standard identifiers and references to physical specimens, may also be found.

A number of other sections may follow the nomenclature section. For example, an important section, frequently titled "Materials Examined" lists the specimens or other materials (e.g.., DNA sequences) used as the basis of the treatment. This section often includes the circumstances of collection and/or deposition at a museum or other institution. Historically, these details allowed scientists to visit the holding institution (or seek a loan) for further scientific investigation of the very material that was described by the treatment. Currently, several models describing specimen data tend to follow the Darwin Core standard (http://rs.tdwg.org/dwc/index.htm). Also common is a "Description" section providing information—often in highly structured language, and sometimes in tabular form—on the distinctive features of the collected organisms, with an aim toward specifying a characterization of the entire taxonomic class such material represents.

Similar to a Description section is the "Diagnosis", which contains descriptions of only those features "that distinguish that species from others, in the same way that the disease identification you receive when you visit the doctor is called the diagnosis because the doctor has distinguished your illness from all other possibilities based on the basis of your symptoms and tests." (Winston 1999). Other treatment sections may include an "Etymology" section explaining the origins of the taxon name, sections summarizing the spatial and temporal distribution of the taxon, an "Ecology" section discussing behaviour and relationships to habitat and discussions of range or distribution, relationships to other taxa discussions of other life stages and a variety of other topics. In some cases a paragraph may include different topics. For infraspecific or higher level taxa (such as genera and families) a "Key" presenting a set of instructions for distinguishing lower level taxa from one another is also very common.

The launch of the electronic taxon treatment concept[2] played a key role in the development of taxonomic tagging methodology. Moreover, it is expected that its influence will increase in the near future. Thus, we consider it necessary to describe the concept here in more detail.

[2] There can be two approaches here: Top down: Identify the boundaries of taxon treatments (either automatically or manually) then mark up the sections and other information within taxon treatments. Bottom up: Identify

From the text-processing perspective, a taxon treatment is any "block of text" containing several paragraphs with information on a given taxon, that can be delimited from other taxon treatments within the same document by specifying the treatment's start and end tags. From the viewpoint of the publishing tradition in systematics, the treatment is a block of information on a given taxon that may include some elements of the following (Penev et al. 2010):

1. New taxon description or re-description of a known taxon
2. Change of a nomenclatorial status of a taxon (a nomenclatural act)
3. Summary of some or all previous knowledge on a taxon from literature sources, usually structured in logical pieces, e.g., nomenclature, morphological description, distribution, ecology, biology
4. Summary of some or all previous knowledge plus newly published data on the same taxon, e.g., localities, ecological/biological observations
5. Summary of newly published data on an already known taxon
6. Summary of treatments of subordinated taxa, for instance a revision or catalog of a genus listing treatments of ALL or SOME of its species is a treatment of that genus
7. Listing of subordinated taxa, e.g., a checklist of a family from a region forms a treatment of that family.

Taxon treatments usually have the form of published conventional texts that could be enhanced by a wide array of tags and external links. More importantly, taxon treatments may be archived, searched, harvested, or linked as separate pieces of information directly related to their respective taxa.

A publication may consist of one or many treatments of different taxa of different taxonomic ranks. One taxon may have more than one treatment within a publication, although the tradition of systematics publishing usually assumes one "core" treatment per taxon within a document.

Taxon profiles generated "on the fly" or extracted through web "scrapers" have several features of treatments (e.g., EOL, NCBI, Wikipedia, or ispecies.org taxon profiles). To be called treatments, however, they have to be published in a static and citable form. It seems necessary to distinguish these two types of taxon profiles (published in accordance with scholarly publishing practices and dynamic, often generated "on the fly" through web-based compilations), although the border between them may sometimes seem vague. The essential feature of a treatment is that it encompasses information published in accordance with both present-day publishing standards and the requirements of nomenclatural codes.

What is not a taxon treatment?

1. A citation of a taxon name within a text, although such a citation usually holds information linked to the particular taxon. For instance, listing of a species within a

---

information within a document then identify the smallest logical unit that contains all the information that should be in a taxon treatment.

"plain" checklist cannot be a treatment of that species; a sentence within a text paragraph stating that "taxon X is parasitic on taxon Y" is neither a treatment of taxon X nor of taxon Y

2. A key[3][4], because in some cases keys are constructed for related taxa that do not form a taxon (they may form a "species-group" or "taxa-group", but this is not a taxon unless a name is given to that group). Identification keys, even they are exhaustive for a named taxon, are usually tagged separately from taxon treatments.
3. A single picture or group of pictures of a taxon[5]
4. A single map or group of maps of a taxon
5. Gene sequence(s) of a taxon
6. SDD (Structured Descriptive Data) (or any) matrices, or raw data, or databases. Treatments can be relatively easily generated from databases, however, information on a taxon becomes a treatment when (a) it is published, and (b) corresponds to the aforementioned description of taxon treatment.

The TaxonX schema and the TaxPub DTD largely follow the above restrictions which arise from a community of practice rooted in paper publishing. taXMLit is less restrictive, although its recognition of treatments encompasses the more limited formats discussed above. However, more open-ended concepts of what makes treatment have proven necessary in marking up complete papers, where authors have been found to publish nomenclatural and taxonomic changes in a very wide variety of ways. In the electronic era, broader notions of a treatment can easily be added to the electronic forms by simple extension of the schema or DTD, in ways that do not make useless publications with the narrower form.

Why are taxonomic treatments important? What role do they play in various disciplines? Taxonomic treatments are important because they allow "atomising" taxonomic texts, that is they permit labelling and delimiting a dedicated piece of information (e.g., a block of text) describing a a taxon within a document from other similar pieces of information, describing other taxa. Taxonomic treatments allow a rapid transition from conventional, article-level publishing in the biodiversity science, to treatment-level (or content- or data-level) taxonomic publishing. XML encoded taxonomic treatments facilitate future use, re-use and collation (harvesting and indexing, mashups, linkouts) of data, because computers can recognise data elements within treatments and relate such data to taxon names. Treatments, especially those of new or re-described taxa summarize and comment all the knowledge of a particular taxon. Thus, at the time of publishing, treatments for the basis from which some relevant information could be found by external links, and which can be used as crystalline core for the expansion, reuse, rewriting of the information on this taxon.

---

[3] It is still to decide if keys should be tagged separately and when they need to be tagged within taxon treatments. In some cases keys are part of a taxon treatment but in some cases they are a stand-alone publication encompassing several taxa which could be marked as separate treatments

[4] In some cases with legacy literature taxon treatments occur within keys – taxa may be described as new (or synonymised) as part of a key lug.

[5] In some early literature the only record is an illustration which, together with the name in the figure legend, comprises the original description of the species.

Taxonomic treatments are important because they allow mobilization, retrieval and re-use of taxonomic data published not only in the present day, but in most cases also in historical taxonomic literature. Recent and historical treatments can be interlinked through taxon names, given the proviso that name strings can change between treatments of the same taxon in different sources; a solution to the latter problem will be found through Global Unique Identifiers (GUIDs) for each taxon (applicable through Life Science Identifiers, or LSIDs) expected to be realized through the Global Names Architecture framework (www.gna.org).

Finally, treatments are important because in a straightforward way they relate information on organisms to the oldest and most widely used identifiers in the history of biology – the taxonomic names of organisms. Through names, and especially through the recently developed global index of taxon names (Global Names Architecture, or GNA, Global Names Index, or GNI, Global Names Usage Bank, or GNUB, see http://www.globalnames.org and http://www.gbif.org) treatments may be linked to any other information in any other branch of science that uses taxonomic names.

To facilitate "atomizing" of taxonomic texts into retrievable and machine-readable forms, we need a computer language and sets of rules and protocols in taxonomic publishing, such as XML (see below for more details). TaxonX and TaXMLit are markup XML schemas developed to encode historical, or legacy, taxonomic literature. They are therefore robust enough to retrieve a great variety of styles used in such literature. TaxPub was developed as an extension of the general Document Type Definitions (DTD) format of the National Library of Medicine of the US (NLM, http://dtd.nlm.nih.gov) to facilitate markup of prospective taxonomic publishing. Whilst TaxonX has been developed to model treatments, taXMLit and Taxpub model the entire publication. TaXMLit as a selfstanding schema and TaxPub is built upon and an existing widely used DTD of NLM. TaxonX and TaxPub build on the concept of using as much of existing elements, whilst taXMLit is self-contained. The intention is that all the three schemas/DTD can be mapped to TDWG vocabularies, at least regarding to domain specific content.

TAXONX

Sources: http://sourceforge.net/projects/taxonx/; http://www.taxonx.org/schema/v1/taxonx1.xsd; www.plazi.org, Sautter et al. 2007

Description

Taxonx (http://taxonx.org/schema/v1/taxonx1.xsd) is a XML schema for encoding taxonomic literature in order to:

- Create open, stable, persistent, full text digital surrogates of taxonomic treatments
- Identify taxonomic treatments and their major structural components to enable networked reference and citation

- Identify lower level textual data such scientific names, localities, morphological characters, and bibliographic citations to facilitate their extraction by, and integration with external applications and resources
- Study and describe the structure of systematics publications by creating few typical corpora of literature, such as entire journal (e.g., AMNH Novitates, Zootaxa), across taxa (e.g., all ant systematics papers post 1995), or faunistic (e.g. all ant systematics paper covering Madagascar ranging from 1758 to 2011)

TaxonX is a lightweight (with only 30+ elements) and flexible (because very little of the elements have to be used to create a valid document) schema which should be quickly learned and may be applied to the wide variety of formatting present in legacy document as well as for new publications. It relies on (see use of MODS for file-level bibliographical metadata), use of external schemata. It has loose content requirements allows for instances to be encoded over time and at many levels of granularity, while maintaining validity through iterations. Additionally, TaxonX contains mechanisms for semantic normalization of the data contained in treatments. The schema can be readily converted to or expressed as an extension of the NLM/NCBI Journal Archiving DTD.

Design and Development

Development of TaxonX, an XML-Schema for markup of treatments had begun at AMNH and continued through the duration of a subsequent NSF/DFG grant (see below). As the project was concluding, participants established Plazi Verein, a Switzerland-based independent not-for-profit organization aiming to help remove technological, social, and legal barriers to the creation of and access to taxonomic literature. Among its many activities Plazi maintains the TaxonX schema and a repository of XML-encoded publications, develops the semi-automatic markup tool, GoldenGate (Sautter et al., 2007), and strenuously advocates for open access to scientific literature (Agosti and Eglof, 2009). As part of these efforts, Plazi has encoded approximately 600 publications containing roughly 13000 treatments using the TaxonX schema and currently harvests all the treatments published in ZooKeys since issue 50 (Sautter et al., 2009; http://plazi.org (accessed Feb 28, 2011).

TaxonX provides for the encoding of taxonomic treatments, with elements for the major structural components of treatments (e.g., Nomenclature, Materials Examined, Description, etc...) and phrase-level features of interest in taxonomy (e.g., scientific names, locality names, characters, etc...) as well as mechanisms for linking to external resources and the semantic normalization of terms mentioned in the source document. The TaxonX instances contain a moderate degree of markup. Bibliographic metadata for the source documents are provided in each instance. Other sections of treatments are identified and named when they occur, but are not always present due to the wide variability of the structure of the source documents. All scientific names are marked and associated with an LSID, but other features may not always be identified. Materials Examined can be broken down to individual materials citations which in turn can be normalized and linked to external resources, such as a type specimen, through LSIDs or other html links.

Implementations

Use Case 1: The GOLDEN GATE (GG) software tool. GG development was lead by Guido Sautter (Sautter et al. 2007) to serve mark up of legacy literature. GG is in fact an integrated set of tools and modules under a single umbrella application and UI that allow highly automated large scale output. The use cases listed below have been performed with the use of GG. In 2010 GG launched a web interface to allow social networking elements in the mark up process.

Use Case 2: Ants of Madagascar. In 2006-2008, all available literature on the ants of Madagascar has been OCR-ed, marked up to a treatment level and stored on Plazi's treatment repository; the collections covered ca. 4,000 taxon treatments, 119 publications of a total of about 2500 pages of legacy publications with. The project formed the basis for the subsequent development of Plazi's mark up projects (see below).

Use Case 3: The Zootaxa-TaxonX-ZooBank Project. In 2007, GBIF approved a Seed Money Award project entitled "Extracting Nomenclatural Data, Species Descriptions and Collecting Events from Legacy Publications: The Zootaxa-TaxonX-ZooBank Project" (GBIF Tracking Number 2007-94). The latter provided, as one of its outputs, biodiversity literature encoded as TaxonX XML documents through the use of the GoldenGate markup tool. The Species Profile Model (SPM) extension provides XSLT scripts for the creation of SPM documents expressed in RDF from the TaxonX documents. The RDF files are mounted on the Plazi web server for harvesting by SPM aware agents, particularly those to be created by the EOL project. SPM data are served through TAPIR.

Use Case 4: SPM (Species Profile Model) export from Plazi to EOL. In collaboration with Encyclopedia of Life (EOL) and GBIF, Plazi has now developed a web service providing treatments in Species Profile Model (SPM) format allowing EOL and other interested parties to consume and automatically publish such content. Plazi received a small grant from EOL (managed by GBIF) to implement a service based on the Species Profile Model for the provision of taxonomic descriptions to EOL to complement a previous GBIF Seed Money Award to Zootaxa and Plazi that mobilised species occurrence records for the GBIF network from such source data. The data for the project were taxonomic publications related to Ants. The original publications had been scanned, with the text captured via OCR, and encoded by Plazi using GoldenGate (http://plazi.org/?q=GoldenGATE) and the TaxonX XML schema (http://TaxonX.org/schema/v1/TaxonX1.xsd). An XSLT conversion to SPM RDF/XML was developed and deployed as a web service using the eXist XML database (www.exist-db.org) so that SPM files generated dynamically from the TaxonX files can be retrieved via an HTTP GET request. A documented API is provided for the service, which allows the client applications latitude on tailoring the service. Sufficient documentation is provided so that clients can use the service for altogether different and unique processing of the underlying XML document. At the date of finishing this report (in February 2011), 5892) treatments have been made accessible

on EOL, including fish, ant and platygasteroid wasps. By the end of February 2011, 12,842 treatments from 605 publications have been available, and the numbers will increase steadily.

Problems Encountered and Lessons Learned

Based on accumulated experience, the following success factors of TaxonX can be encountered:

- TaxonX is a lightweight and flexible schema which should be quickly learned and may be applied to the wide variety of formatting present in legacy documents
- Relies on use of external schemata (see use of MODS for file-level bibliographical metadata).
- Loose content requirements allows for instances to be encoded over time and at many levels of granularity, while maintaining validity through iterations.
- Contains mechanisms for semantic normalization of the data contained in treatments. See TaxonX's use of Darwin Core (soon perhaps LinneanCore, SDD, etc…) to normalize phrase level data, and xid elements for inclusion of LSID's, ITIS, HNS, or other external identifiers.

There are some hurdles for adoption of TaxonX, such as:

- The heterogeneity and structural looseness of the data contained in legacy taxonomic treatments nevertheless defies encoding and semantic normalization by even a lightweight and flexible schema.
- The flexibility of the schema may present difficulties both in authoring and in profiling the encoded data for use by external applications as well as in converting into other schemas/DTD.
- Dependence on external schemata requires vigilance and active maintenance of the schema; may complicate validation of instances over long-term; namespace wrangling makes authoring difficult
- Mark-up, even in a light way, needs some domain specific expertise, specific quality controls to assure that the elements are properly idenified and therefore costs time.

Potential users of TaxonX could be:

- Biodiversity Heritage Library would become tremendously more useful, if at least treatment boundaries, nomenclatural elements and respective names are marked up and *linked up* to the respective scan on BHL.
- Ultimately, one could envision this to be an intermediary step to extract and store the treatments in more powerful structures, such as databases. All the treatments are primarily linked to genetic, distributional or nomenclatorial and other data via the taxonomic name of which to which the treatment refers. At antbase/HNS this link is in a simple form already implemented by a link from each citation to the respective pdf copy of the referring page.

- Future agregators of treatments might be institutions like Zoobank, but essentially dedicated databases allowing specific applications, like ispecies, to collect the treatments and use them for specific purposes.
- All aggregators that will benefit from improved search, information retrieval and information extraction.

TAXPUB

Sources: http://sourceforge.net/projects/taxpub/; Catapano 2010; Penev et al. 2010a

Description

TaxPub is an extension of the NLM/NCBI Journal Publishing DTD (Version 3.0) for the encoding of the literature of biological taxonomy. A key feature of this literature is the taxonomic description: publications or sections of publications that name and describe species and other taxonomic information. Given that it is estimated that the majority of all species have yet to be described, and that some 15-20,000 new species are described each year, and that markup might be applied prior to publication at less expense than applying markup to existing publications, TaxPub aims at providing a tagset for the encoding of new taxonomic literature. TaxPub extends the Publishing ("Blue") DTD parsimoniously. A few phrase-level elements are available at the relevant places throughout the entire DTD. Most of the extension, however, occurs in a single section-level element <tp:taxon-treatment>. The development of the extension proceeded smoothly, but several challenges have been encountered: lack of consensus on the components of taxonomic descriptions; relationship and alignment of TaxPub to other related schemas in the field; decisions on creating new elements or using existing NLM DTD elements and how to document and validate the usages; resistance to DTD as the XML schema language; and the efficiency of creating a superset extension rather than utilizing other simpler profiling mechanisms.

The development of TaxPub is an outgrowth of an earlier effort to digitize the taxonomic literature of ants for purposes of developing data mining techniques for the extraction of species data from taxonomic literature with TaxonX (see above) as one of its main results. The work was originally performed as part of a joint U.S. National Sciences Foundation and Deutsches Forschungsgemeinschaft (German Research Foundation) grant awarded to the American Museum of Natural History (AMNH) and the University of Magdeburg (later to Karlsruher Institut für Technologie/Karlsruhe Institute of Technology).

Given the complexity and difficulty of digitizing existing taxonomic literature, and that it covers a minority of all species, greater benefit at less cost might be found in the encoding of new, born digital, taxonomic literature. Increasingly, treatments are derived from data maintained in databases, whether for names, specimens, or bibliographic references. This information could be exported into XML directly, saving an enormous amount of time and ensuring accuracy. The idea to generate publishable natural language treatments from databases arose in the early 1970's and was unambiguously in place by 1980 (Dallwitz, 1980). The rise of XML has provided more tools to produce and exploit structured treatments, but often these tools are used backwards with time wasted by experts providing markup to published literature. Indeed, in the

case of recently published literature information originating in parsed form in citation managers and databases becomes converted by an author to unstructured text for publication, only to be parsed out once again during the markup process.

Consensus on an XML schema often fosters development of tools, services, and applications utilizing suitably encoded data. TaxPub is an attempt to catalyze this process in the hope that the community will be intrigued, and find it useful enough to adopt and sustain.

Design and Development

In the second half of 2008, with the assistance of Jeff Beck, Laura Kelly and Scott Federhen of NCBI, Plazi lead by Terry Catapano developed the first draft of the extension now called TaxPub. Since then development has been assisted by Donat Agosti, an ant systematist, President of Plazi and research scientist at the American Museum of Natural History, and by Robert Morris, Emeritus Professor of Computer Science at University of Massachusetts at Boston and an Information Technology Associate of the Harvard University Herbaria. The project is hosted on SourceForge (http://sourceforge.net/projects/taxpub/) with the first release in December, 2008.

The first version release of TaxPub is scheduled for March 2011. A call for comments have been sent out in December 2010 soliciting feedback and requests for new features. Subsequent releases will be backwards compatible until the next version release.

Rather than adapting TaxonX for publishing applications it was more efficient to extend the NLM/NCBI DTD. The Journal Publishing DTD already included elements for document features, so it was necessary only to add elements and attributes relevant to taxonomic descriptions. TaxPub extends the Publishing ("Blue") DTD parsimoniously.

To better distinguish TaxPub elements from those of the base DTD, elements from the extension have been put into their own namespace, with element names starting with the prefix "tp:". A few phrase-level elements are made available at relevant places throughout the DTD. There are elements for scientific names, <tp:taxon-name>, citations of specimens and other materials, <tp:material-citation>, and descriptions of organisms' physical characteristics, <tp:descriptive-statement>.

The <tp:taxon-name> and <tp:descriptive-statement> elements have simple content models, each allowing any number of optional "part" elements allowing for tagging of the element's components. Required "-part-type" attributes[6] provides further semantics. Because the field of biodiversity has many published vocabularies, URIs are available for many concepts and entities

[6] The fact that the type attibute of element XYZ-part is named „XYZ-part-type" and not simply „type" is pretty much of a nuissance from a processing type of view, namely so because the same attribute (from a semantic point of view) is named differently depending on the name of the element it belongs to, forcing applications to construct the attribute name from the element name, at no gains in expressiveness at all.

of interest. The addition of “-type-uri” attributes to all TaxPub elements with “-type” attributes is under consideration so that, if available, semantics may be provided through use of a URI as a value instead of, or in addition to, a string value.

Of course an additional attribute is not strictly necessary[7] as users may already use URIs in the existing “-type” attributes. We encourage that usage.

Additionally, as in many TaxPub elements, the <object-id> element from the base DTD is available, again with the intention of allowing semantic enhancement through linkage to standard identifiers. <tp:taxon-name> also has additional special attributes: <tp:taxon-name> with “auth-code” to report the nomenclatural code to which the tagged name is conformant; “rank” to explicitly indicate the taxonomic rank (e.g.., genus, species, etc...) of the named taxon; and a “reg” attribute (shared by <tp:taxon-name-part>) to contain a regularized form of an element's contents.

The other element available throughout the DTD, <tp:material-citation>, has a richer content model. Like bibliographic citations, specimen citations can be complex, with many pieces of information. To accommodate granular encoding, <tp:material-citation> allows #PCDATA, the Publishing DTD elements <named-content>, <xref>, and <object-id>, and TaxPub elements <tp:taxon-name>, <tp:material-location> for information on the institution currently housing the referenced material, and <tp:collecting-event> for information on where, when, and by whom the specimen was found. The <tp:collection-event> element has a number of sub-elements: <named-content>, <object-id>, as well as <date>, and extension elements <tp:taxon-name> and <tp:collecting-location>. <tp:collection-location> itself permits zero or more <object-id> and an optional <comment> element, and zero or more <tp:location> element which has a “location-type” attribute to specify whether tagged location is a country, city, province, etc...

Most of the extension occurs in a single section-level element <tp:taxon-treatment>, available in the body of an NLM document. The <tp:taxon-treatment> element contains elements for metadata about the treatment itself, <treatment-meta>, and its component sub-sections: a required <tp:nomenclature> section and zero or more <tp:treatment-sec> elements. Originally, two other named treatment sections were included in the extension, <tp:description> and <tp:materials-examined>, but as their content models did not differ from that of <treatment-sec>, they were removed. A “treatment-sec-type” attribute is available to provide specific semantics for <treatment-sec>, but aside from the inclusion of the other TaxPub elements available throughout the DTD, the content model of treatment-sec is essentially the same as a generic section.

The only required element in the TaxPub extension is <tp:nomenclature>. Its content model is more complicated than other extension elements because it must model and conform to the very formal structure required by the aforementioned nomenclatural codes. <tp:nomenclature> must contain a <tp:taxon-name>, which includes the name of the organism being described by the treatment. Indication that a taxon is a new species or genus is handled

[7] It is still a question why have two attributes at all?

by a <tp:taxon-status> element. A <tp:taxon-authority> element may be used for a “brief bibliographic reference to the original publication of the [taxon] name” (Winston 1999) required by nomenclatural codes and typically in the form of an author’s last name followed by the year of publication. For more granular markup, <tp:taxon-authority-part> elements with “tp:taxon-authority-part-type” attributes are available.

The codes address other complexities of citations (e.g., multiple authors, a species being moved to a different genus since the original publication, etc...), but the current <tp:taxon-authority> model ought to be sufficient. Following the citation of taxon authorship will frequently be a series of citations “of all the names that have been used in published references to [the described] taxon” (Winston, 1999). TaxPub provides a <tp:nomenclature-citation-list> element to group <tp:nomenclature-citation> elements for these citations. The citations may consist of several parts. First is a reference to a name, consisting of a required <tp:taxon-name>, followed by zero or more <tp:taxon-author> elements. Next is a bibliographic reference to the publication in which the taxon was named, for which <mixed-citation> (for an inline citation) or <xref> (for links to an entry in a reference list>) may be used. A reference to specimens may be present for which <tp:material-citation> is available. Other information may be included in an optional <comment> element. As it models perhaps the most complex, least standardized component of taxonomic descriptions, <tp:nomenclature-citation> will no doubt be subject to further review and criticism, and will likely be revised frequently until a stable element definition is achieved. One goal however is, that schematron allows to assess whether the criteria are fulfilled to create an available name or nomenclatorial acts sensu the Codes.

Implementations

In 2009 initial tests using TaxPub were performed. Norman Johnson, of Ohio State University (OSU) and a Plazi member, produced treatments from a database tracking morphological features of wasp species described as part of the NSF-sponsored Planetary Biodiversity Inventories program. The resulting TaxPub encoded treatments contain nomenclature sections, a description section containing standardized descriptions of morphological features, a listing of specimens used as the basis of the treatment including the locations of collection and of deposition, and a link to a map showing the distribution of the specimens. Significantly, the marked up text was generated by software directly from the database. The OSU implementation realized one of the primary objectives of TaxPub: database-driven publication of species descriptions in order to enable less loss, more rapid publication of data rich descriptions.

Soon after the initial release of TaxPub, Plazi was joined by Pensoft, the publisher of the online open access taxonomy journal ZooKeys, in a collaboration to integrate TaxPub into its publication workflow. The approach differed from OSU’s in applying markup to submitted manuscripts. Pensoft faced a set of challenges similar to those for retrospective conversion. Among them was the identification and encoding of treatments, scientific names, and bibliographic references. Developing their own software tool (Pensoft Mark up Tool, or PMT – see Penev et al. 2010), in 2010 ZooKeys began to publish TaxPub versions of their articles. Although lacking a very fine level of markup granularity (for example <material-citation> is not used), the ZooKeys articles accomplish many of the goals of the TaxPub extension. Treatments

are identified, and thus are directly and easily machine addressable, as are treatment sub-sections. All scientific names and name parts are tagged with <tp:taxon-name> elements. <tp:nomenclature-citation> elements include <tp:taxon-name> and link to full bibliographic entries, themselves marked up with <mixed-citation>. Significantly, because TaxPub motivated and enabled its use of the NLM DTD, ZooKeys articles will be archived in PubMed Central.

The ZooKeys working examples (Stoev et al. 2010, Blagoderov et al. 2010b, Brake and von Tschirnhaus 2010, Taekul et al. 2010) are entirely based on revision #123 available from the SVN trunk of TaxPub (http://sourceforge.net/projects/taxpub). In fact, the present exemplar papers are the first published TaxPub articles in biodiversity science, intended to demonstrate the advantages of the XML-based markup and editorial workflow in the way biodiversity information is being published and disseminated.

Use Case 1: TaxPub is used to mark up taxonomic papers to publications process through the Pensoft Mark Up Tool (PMT) created by Pensoft to embed semantic tagging technologies within the editorial process. As a result, through PMT and INDesign, 3 electronic versions of a paper are generated and routinely published: (1) PDF identical to the printed version; (2) HTML to provide links to external resources and semantic enhancements to published texts for interactive reading; (3) XML version compatible to PubMedCentral archiving NLM DTD TaxPub extension), thus providing a machine-readable copy to facilitate future data mining.

Currently TaxPub is used routinely in the editorial process of five journals published by Pensoft:

- ZooKeys – www.pensoft.net/journals/zookeys
- PhytoKeys – www.pensoft.net/journals/phytokeys
- International Journal for Hymenoptera Research – www.pensoft.net/journals/jhr
- International Journal of Myriapodology – www.pensoft.net/journals/ijm
- Comparative Cytogenetics – www.pensoft.net/journals/compcytogen

Use Case 2: Export of new taxa to EOL. All new species description in Pensoft journals are exported to EOL through a specialized tool that addresses the requirements of the recipient on the day of publication; the file contains bibliographic metadata, taxonomic classification, species description and links to the species images. The export XML file is being harvested and uploaded on EOL on a daily basis.

Use Case 3: Export of all taxon treatments to Plazi. All taxon treatments identified within the XML file of a published paper are harvested by Plazi and uploaded on their servers. Thereafter, treatments are available for use by various organisations and individuals, e.g. EOL.

Use Case 4: Archiving in PubMedCentral. ZooKeys was accepted for indexing and archiving in PubMedCentral in August 2010. From that time, TaxPub XML output of issues 50-52 of ZooKeys passed 4 rounds of testing at NLM. All suggestions have been implemented in the XML export and when needed, corrections were also reflected in changes in TaxPub.

Use Case 5: Use of TaxPub XML files to create a semantically enhanced HTML version of the publication. Several phrase-level or tp:taxon-treatment element and its sub-elements have been implemented in creation of semantic enhancements to taxonomic texts. The process has

been described and exemplified in issue 50 of ZooKeys (Penev et al. 2010a,b) and from that point turned into a routine practice for five of Pensoft's journals.

Use Case 6: In the same issue 50 (Blagoderov et al. 2010a, Penev et al. 2010b) ZooKeys piloted acceptance of manuscripts in XML format generated from two independent sources, Scratchpads (sample papers: Blagoderov et al. 2010b, Brake and Tschirnhaus 2010) and SysLab tool of the Hymenoptera Online database (Taekul et al. 2010). The process should be turn into a routine practice, based again on TaxPub, during the ViBRANT project.

Problems Encountered and Lessons Learned

While taxonomic treatments do appear to follow conventional patterns, there is in fact no consensus on what the structural components of a treatment are, nor even what they are named. This is a problem even within the domain of zoology (the primary focus of TaxPub to date), but more so if one seeks consensus that simultaneously encompasses the domains of zoology, botany, and bacteriology. More discussion needs to take place beyond the limited circle of TaxPub if there is any hope that the extension will be useful for any purposes beyond Plazi's own.

A number of XML schemas (e.g., for names, specimens, descriptive data, and phylogenetics) are in use or under development in biological taxonomy and related fields. There is significant interest in integrating or harmonizing TaxPub with these related schemas. Care must be taken in the design of TaxPub in aligning with external schemas without compromising its integrity or complicating maintenance. One such schema is Darwin Core (http://rs.tdwg.org/dwc/index.htm), a representation-free controlled vocabulary with implementations in XML Schema and in RDF used for the exchange of specimen data. One option for alignment with Darwin Core was simply to incorporate its elements in TaxPub at the appropriate points in the DTD as, for instance, MathML elements are included. Ultimately this approach was rejected due to the maintenance burdens of synchronizing TaxPub with Darwin Core, a schema not under Plazi's control. Also, the likelihood that valid instances of TaxPub would become invalid if changes to Darwin Core were to be incorporated would complicate maintenance of applications developed against TaxPub. The approach decided upon was to not include Darwin Core at all, but to eventually document the use of URIs of Darwin Core terms as values for type attributes of relevant TaxPub and Publishing DTD elements, e.g. <named-content> and <tp:location>. Further testing and implementation of TaxPub will reveal whether this approach is effective or expressive enough.

The choice of whether to create a new element or to expect use of existing Publishing DTD elements occurred often in the design of the extension. In fact, most TaxPub elements (<tp:treatment>, <tp:nomenclature> and <tp:nomenclature-citation> excepted) closely resemble some Publishing DTD element that could be used instead. In modelling keys to taxa, for example, it was decided to rely on generic Publishing DTD elements, particularly those for tables, rather than create new special purpose elements. When use of existing elements is preferred, however, it becomes necessary to somehow express intentions regarding usage of the Publishing DTD. In implementing TaxPub for ZooKeys, Pensoft, for example, found nothing regarding keys but were eventually provided instructions and examples on tagging of keys using

generic table elements by the TaxPub editors. Extension of the DTD thus requires more than simply editing DTD and entity files. Those planning on creating an extension should plan on producing at minimum written documentation on not just the extension, but on use of the base DTD as well. Going beyond this, it is advantageous to express the intended or endorsed usages in a Schematron schema, if only to provide an example for implementers to do so themselves for their applications. Defining such profiling mechanisms again places burdens on those developing and maintaining an extension.

The issues of both documentation and alignment with other schemas were complicated by working with DTD as the XML Schema language. The lack of robust namespace support in DTD removed the option of importing external schemas into TaxPub. This would make synchronization less onerous, for example, were it decided to include Darwin Core elements in TaxPub. It also would enable the inclusion of XML data in TaxPub instances themselves rather than on linking to them as external documents. It may be argued that there is little practical difference between inclusion of XML data and linking to that data in external documents, but it is nevertheless a limitation to which extension editors must adjust. And it must be explained to users who either have XML data they wish to include in articles, or consumers who would prefer access such data directly.

Though the Publishing DTD itself has excellent documentation, the means to include notes on usage (through use of comments) in the DTD is limited and ad hoc. Far more useful would be the built-in mechanisms for inline documentation as are available in W3C XML Schema and RelaxNG. Especially helpful is the possibility of XML encoded documentation in schema annotations. This would allow for far richer, more readily processed and easily maintained documentation than is currently the case. Given the importance of documentation, it is frustrating and burdensome to have to work around the limitations of annotation in DTD.

In the biodiversity informatics community we have encountered other resistance to the use of DTD as the schema language for TaxPub on several other grounds. There is general unfamiliarity with DTD and perception that it is “old fashioned” and complicated (e.g., “do I change *models.ent or *modules.ent?”). While not technical hindrances, nevertheless such perceptions do impede acceptance of TaxPub by its target community. As a result there has been a need to educate on XML Schema languages, something not envisioned as a task at the outset.

Finally, at a more fundamental level the question of whether to superset the Publishing DTD is a major consideration. TaxPub, as an extension, provides semantics beyond what is available in the base DTD through creating newly named elements—thus lending itself to domain-specific application. However, TaxPub instances may not be easily processed by applications already familiar with the Publishing DTD. TaxPub does not add many new elements with content models that could not be modelled using ordinary Publishing DTD elements. So why create a superset at all? Much could be accomplished through other methods of profiling. Most important for profiling is written documentation, detailing usage—already a necessary task when creating an extension. A controlled vocabulary for “-type” attribute of <named-content>, <sec> and similar generic elements (perhaps published as a machine readable form such as SKOS) can effectively provide semantics for document features not already addressed in the Publishing DTD. Rules on usage and checking of controlled values of type attributes can be

expressed in a Schematron schema and provide validation. While its customizability is a well developed feature of the Publishing DTD, it may not ultimately be the most effective or efficient approach for TaxPub.

TAXMLIT

Sources: Weitzman & Lyal, http://www.sil.si.edu/digitalcollections/bca/documentation/taXMLitv1-3Intro.pdf; Sautter et al. (2007)

Description

The **taXMLit** schema is designed to accommodate taxonomic literature but was developed particularly in the context of Zoological and Botanical taxonomic literature and should also accept fungal and paleontological publications, but this has yet to be tested. The schema does not take into account the kinds of data needed for viral or bacterial publications. It covers all of the components of taxonomic publications and the taxon treatments contained within them other than the actual characters, which are dealt with by other projects such as SDD. This explanation is written in conjunction with version 1.3 of taXMLit, which should be consulted for additional information.

The schema is extremely atomised, permitting both recovery of publication components (e.g. taxon treatments, bibliographic entries, discussion paragraphs) and also of data within those components, such as specimen data and biological associations. The full taXMLit contains data elements extracted from the text that permit detailed data query, browse and download; a version that does not include these and is more document-centric has also been developed ('taXMLite').

Implementation of the schema in an appropriate system ('INOTAXA' has been designed for this purpose) will allow the text of marked up taxonomic texts to be fully searchable. Users may chose to extract data (e.g. taxonomic names, specimen data, citations, biological association data, persons' names) for use in analysis or other function, or access taxon treatments, keys, images or other content components as reference resources. In conjunction with the appropriate system the schema would also facilitate static links from the text to other data sources (e.g. specimen databases on the web, ZooBank). The use of the schema for multiple taxonomic works will allow multiple sources to be searched or browsed simultaneously, and permit links between different works that include the same taxa or their synonyms. Moreover, this opens the way for virtual compilations of taxon treatments to be made up by the user, comprising components of more than one original work, e.g. checklists, faunas, and floras. These functions will require that the schema should, in the appropriate parts, be using the same or similar elements to schemas used by other relevant partners, and be mappable to them.

Design and Development

taXMLit and the interface designed for querying content, INOTAXA, grew out of a Mellon-funded meeting in 2001 at which a number of major museums and herbaria determined to demonstrate the potential of combining information, literature and research data held within their collections. It was funded by the Atherton Seidell Fund in 2002. As a testbed, not only because of its intrinsic interest but also because of the wide coverage of animals and plants and the variety of editorial styles applied, the project used the 57 Zoological volumes of the *Biologia Centrali-Americana*, although a number of other texts have also been marked up and are accessible through the INOTAXA.org pilot (currently accessible are a two papers from Zootaxa, the more recent being Pyle *et al* (2008) on *Chromis*, a paper that has been used by a number of initiatives to enable comparison.

The emphasis during the development of taXMLit has been meeting user needs, and the prioritisation was (i) the schema to accommodate the needs, (ii) the implementation (INOTAXA) to demonstrate and test delivery against the needs and (iii) only following that the development of mark-up techniques. The project is currently refining INOTAXA following feedback and review, and moving into the third phase of mark-up.

Mark-up has to accommodate the source, which for most legacy literature is not born digital. Tests against BHL OCR content show a success rate against scientific names of only 14-35%. The ABEL project 9Morse et al, 2009) has examined this problem. To date the mark-up to taXMLit has been undertaken in two stages: a preliminary mark-up to a document-centric format (generally TEI-Lite with a systematic 'flavour' created by Weitzman & Lyal) and a subsequent parsing to taXMLit. The latter stage has been done by hand for a number of publications, but automated through use of a specially-written script for a volume of the BCA.

Within taXMLit paragraphs of text in the original work are captured as a whole to facilitate the order of the text components as subsequently reconstructed. In some cases reconstruction will use a different order than the original (e.g. keys can be spread throughout the text with lugs separated by treatments or other complex elements, but need to be reconstructed without these interruptions). Individual paragraphs may then be parsed into more or less detailed elements. Each paragraph (i.e. any text component terminated by the stroke of an 'Enter' key) element is given an ElementID, which run sequentially through the text. In addition to facilitating text reconstruction it also allows the use of an IDREF attribute (a cross-reference within the mark-up) elsewhere in the marked-up text, for example with dates, key lugs and images, to make the referred information available where required.

The root element of the schema is the TaxonomicPublication. This can be any kind of publication, including multivolume works, articles in journals, and books. It has an attribute of 'TaxonomicPublicationID'. In order to accommodate publications that appeared in multiple volumes or fascicles there is a child element 'IndividualPublication'. PublicationFrontMatter and PublicationBackMatter are separated from the PublicationTaxonomicMatter, which contains the bulk of the taxonomic components. There is, of course, overlap between front and back matter, given that some publication components (e.g. glossary, acknowledgements) may occur either before the taxonomic content of a work or after it. Consequently, some elements appear (optionally) in both PublicationFrontMatter and PublicationBackMatter. The PublicationTaxonomicMatter is unbounded (may repeat as a group), since within some

publications (such as the BCA) there are several components each with a separate author, title and even introductory matter ('PublicationTaxonomicSubhead'), which do not neatly lie within separate fascicles or volume parts. Using the method adopted allows us to take account of this. Each publication also has an element 'NamesCitedList', created in mark-up, to provide identification for all names in the individual work and facilitate cross-reference. The full set of elements is large, designed to accommodate the detail required for search, browse and download of identified components. While taXMLit uses elements that cover the same concepts as those used in other schemas (e.g. ABCD, designed for specimen data), the individual elements are not exactly the same. This is because the data as presented in the literature is often different in format to that recovered from specimen labels, and may not be as easily interpreted. However, the XML is designed to permit easy mapping to ABCD and DwC. In many elements the content is atomised; taxonomic names, for example, are fully atomised and rank to each component assigned.

Throughout much of taxonomic literature abbreviations are used (e.g. for genus names) or descriptors are omitted (e.g. for hierarchical rank levels above genus, or for components of label data if these are repeated). While this information is simple to derive for a human reader it is less accessible to machine treatment or amenable to database storage. For this reason many elements in the schema have the attribute 'Explicit', to denote whether the information included is explicitly stated or implicit and derived either by programming code or by a human in the final verification of the mark-up. In all cases, these elements may be derived from the text content itself and are needed to build a complete database of the text that can be searched and made interoperable.

The implicit information referred to in the previous paragraph is intended to accommodate only that which is unequivocal (e.g. the abbreviation "A." in front of a species name where the only genus name in context is "A-us" is marked as that name). However, TaXMLit is not designed to include wider interpretations of that text drawing on information and knowledge from outside the text itself. For example, although locality names may be outdated (e.g. 'Burma' instead of modern-day 'Myanmar'), or subsequent information indicates that the cited publication date is incorrect, such information is not accommodated in the current schema. To accommodate this and similar information an 'interpretation layer' (TIL) will be created in a subsequent phase of the project. The TIL will function as both a simple layer around the publication that will allow authors to contribute information such as that above, and as a complex proxy layer between the digitized publication and other digitized information sources, including a variety of kinds of authority files. This facility will enable data to be entered independently after the digitization, in the same manner that any piece of text or data requires and receives interpretation in the normal course of use. The TIL will also allow interpretations to be attributed and dated, and allow for multiple interpretations of the same information by the same or different people or publications. Another function of the TIL will be in facilitating linkages between different taxonomic treatments, and between the treatments and other data sources.

Some of the original formatting is retained (e.g. underlining, italics, bold etc) although font and line indentation, for example, are not. Page numbers are retained.

An interface, INOTAXA, has been developed to enable the queries and browse functions prioritised in the user-needs survey. To minimimise search speed over multiple large XML documents the marked-up text is stored in relational database form. To date upload to the database has been via individual scripts; now the project is moving beyond development stage the database has been simplified and an upload tool is being built.

Implementations

Use Case 1: Mark-up of 'old' taxonomic literature. Literature used for this is primarily the *Biologia Centrali-Americana* (BCA) (http://www.sil.si.edu/digitalcollections/bca) but also include parts of Linnaeus 1758 *Systema Naturae* and Linnaeus 1752 *Species Plantarum*. The intent was to discover the flexibility required for older literature written before more modern standardisation and the advent of the nomenclatural codes.

Use case 2: Mark-up of recent taxonomic zoological and botanical taxonomic literature, including different formats. Texts span 1992-2008, include 'standard' taxonomic papers, part of a synonymic catalogue and a book chapter; taxonomic papers are from the Coleopterist's Bulletin, Mosquito Systematics, Proceedings of the Biological Society of Washington, Systematic Botany, Transactions of the American Entomological Society and Zootaxa. The intent was to determine variation in modern literature and how to deal with multiple publications including treatments of the same taxa under the same and different names and in different taxonomic positions.

Use case 3: Human search and browse of content. This is the INOTAXA interface, built in several phases with testing of each phase by primarily taxonomist users who were new to the system. The prototype includes three publications (a BCA volume on Coleoptera: Curculionidae, a Zootaxa paper on Curculionidae also included in the BCA volume, and Pyle et al 2008 on *Chromis* fish, together including more than 800 taxon treatments). Two additional sources of information were added: the digitised contents of Vaurie & Selander, 1971 (georeferenced localities for specimens in the BCA) and a list of person names in all possible formats (e.g. Smith, Smith, J., J. Smith, J Smith etc) – this allows synonymy of different name strings representing the same individual without editing / changing the original text. The interface permits:

a) a simple search on any term, delivering results under the categories 'All', 'Treatments', 'Keys' and 'Other' (any mention of the term in Front or Back Matter or image captions).

b) a simple search of images, using terms in the captions.

c) a Boolean search on 57 indexed fields, permitting complex questions to be posed to the content. Results are delivered under the headings All', 'Treatments', 'Keys', 'Other' and 'Specimens' (downloadable specimen data generated from the text, where meeting the criteria of the query).

d) a taxonomic browse through all taxa within in all publications included in the INOTAXA content. Results are presented under the headings Classifications, Treatments, Keys, Images, Specimen records, Geography, People, Publications and Other.

e) a geographic browse through all places mentioned in all publications within the INOTAXA content. Results are presented under the headings Keys, Images, Specimen records, Gazetteer, People, Publications Accepted Names, All Names and Other.

f) a Person browse through all people mentioned in all publications within the INOTAXA content (as taxon authors, treatment authors, collectors or editors). Results are presented under the headings Person detail, Treatments, Keys, Images of, Images by, Specimens collected, Specimens Determined, Collecting sites, Publications, Treatments, Publications and Other.

Use case 4: Provision of content to Encyclopedia of Life. Currently marked-up text is delivered to EOL and mapped to their schema for display on their pages.

Problems Encountered and Lessons Learned

1. Interoperability. In order to maintain the potential to deliver data in a format that could be used by other applications we could either have incorporated elements of existing schemas or built our own in a manner that allowed easy mapping to others. We chose the latter route to enable us to provide simple versioning of taXMLit and not have to deal with independent and unplanned changes by embedded schemas.
2. GUIDs. Initially GUIDs were not explicitly included; as biodiversity informatics has moved towards implementation a placeholder for GUIDs has been included in many elements.
3. Accommodating multiple formats of legacy literature. Although taxonomic literature is reputedly standard in content experience with many different papers and books has demonstrated the extreme variability of formatting applied, even within single papers. Most complex elements of taXMLit are optional and available in many different places within the schema to accommodate the observed variation.
4. Implicit content. Much content is implicit in nature (see discussion above). Care must be taken in recognising such content but it is necessary to do so to facilitate search and browse functionality in the interface. Such implicit content must be indicated as such. Because spelling errors and other infelicities may have nomenclatoral significance, and because correction relies on individual expertise these are not changed in the current implementation. Such change or annotation must be explicitly authored, and the ability to do this wil be introduced in a later implementation.
5. Mark-up. Semi-automated mark-up has been achieved using a purpose-written script, incorporating rules developed to accommodate the structure of the publication. Even with this there are places where specialist knowledge is required. To facilitate this a SpecialistReview attribute has been introduced throughout. Means of providing a rule library to assist mark-up are being considered.

6. Recovery of original formatting. Only some original format is retained, where this aids in understanding (e.g. italicisation). INOTAXA delivers content in a standardised format to aid comprehension, but allows (subject to copyright) access to the original text.
7. Hierarchies. Each work has an independent hierarchy, which is displayed in INOTAXA. Where a work is produced in multiple fascicles it is assumed unless stated otherwise that the hierarchy does not change.

Criteria for evaluation, comparison and cross-points between the taxonX, TaxPub and taXMLit

As mentioned earlier, the main aim of the current report is to identify the most practical ways to increase the compatibility and accessibility between the schemas. Compatibility is understood here as *ability of the schemas to identify, mark up and export a specific number of elements, most commonly used in both legacy and prospective taxonomic literature and needed for further data mining and reuse by various end users*. An important criterion of compatibility is that schemas can be mapped to a shared (TDWG) vocabulary, and thus would allow conversion from one to another schema. In other words, we should identify presence/lack of features within schemas that will allow generating a unified output from different taxonomic sources (and marked up with different schemas) in a form of:

1. Overall structure of the document: ability of schemas to capture it
2. Bibliographic metadata: relevance to the current widely adopted standards, such as NLM, BibTex, MARC, MODS and others
3. Taxon names: granularity level of mark up; identifying ranks and authorities
4. Nomenclatural acts: presence of controlled vocabularies and normalized (standardized) tags for the different nomenclatoral acts
5. Taxon treatments – unifying the definition of taxon treatment; ways of delimitation within texts
6. Internal structure of taxon treatments – level of mark up granularity
7. Nomenclature section of treatments – names, authorities, synonyms
8. Localities: compliance to DarwinCore; formats for use of geographical coordinates
9. Reference lists, in-text citations and links between both
10. How Pertinent identifiers (UUID, GUID, LSIDs, DOI etc.) are used to identify in an unified way the different elements of taxonomic papers – taxon names, publications, treatments, datasets, keys, phylogenetic trees etc.

Why is this needed? The main reason is of course the necessity to unify the different elements of taxonomic information coming from different sources so that they could be made interoperable, extracted, indexed, collated, used and reused by one and the same user. The format of the final output of the schemas – in a form of XSLT stylesheets for instance – will be determined by the expectations and needs of the end users as shown on Fig. 1.

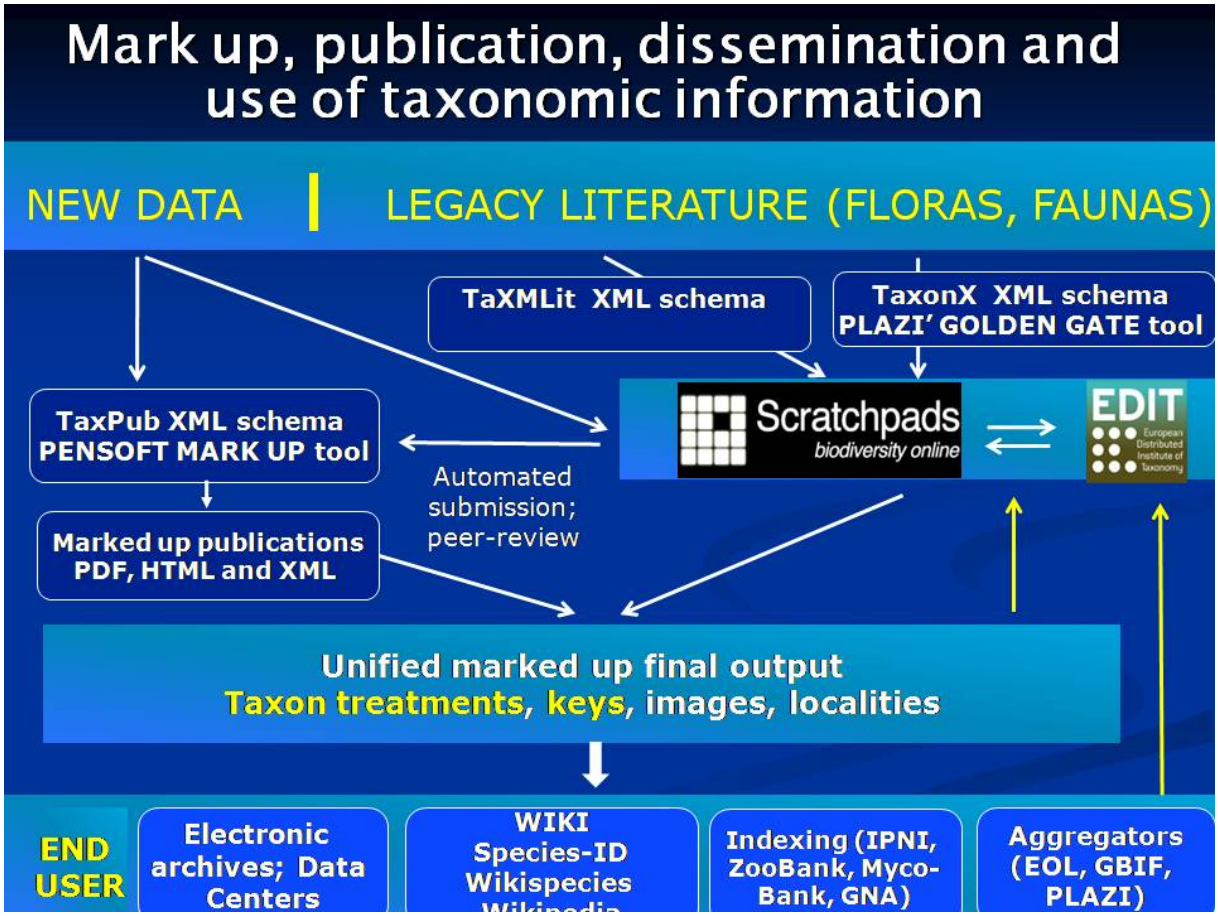

Fig. 1. Flowchart of mark up, publication, dissemination and use of taxonomic information

As for taxonX and TaxPub, these schemas are largely interoperable due to the fact that both have been developed by the same author and contributors and because both schemas have been used together in some of the cases mentioned above. The challenge will be to ensure an output compatibility between taXMLit with the rest, particularly with TaxonX. A comparison between the two schemas has already taken place (Carapano and Weitzman, 2007: http://wiki.tdwg.org/twiki/pub/Literature/WebHome/Catapano_Weitzman_Markup_Final.pdf and Weitzman 2007: http://wiki.tdwg.org/twiki/pub/Literature/WebHome/comparisonTaxonXtaXMLit22Oct07.pdf (Fig, 2.)

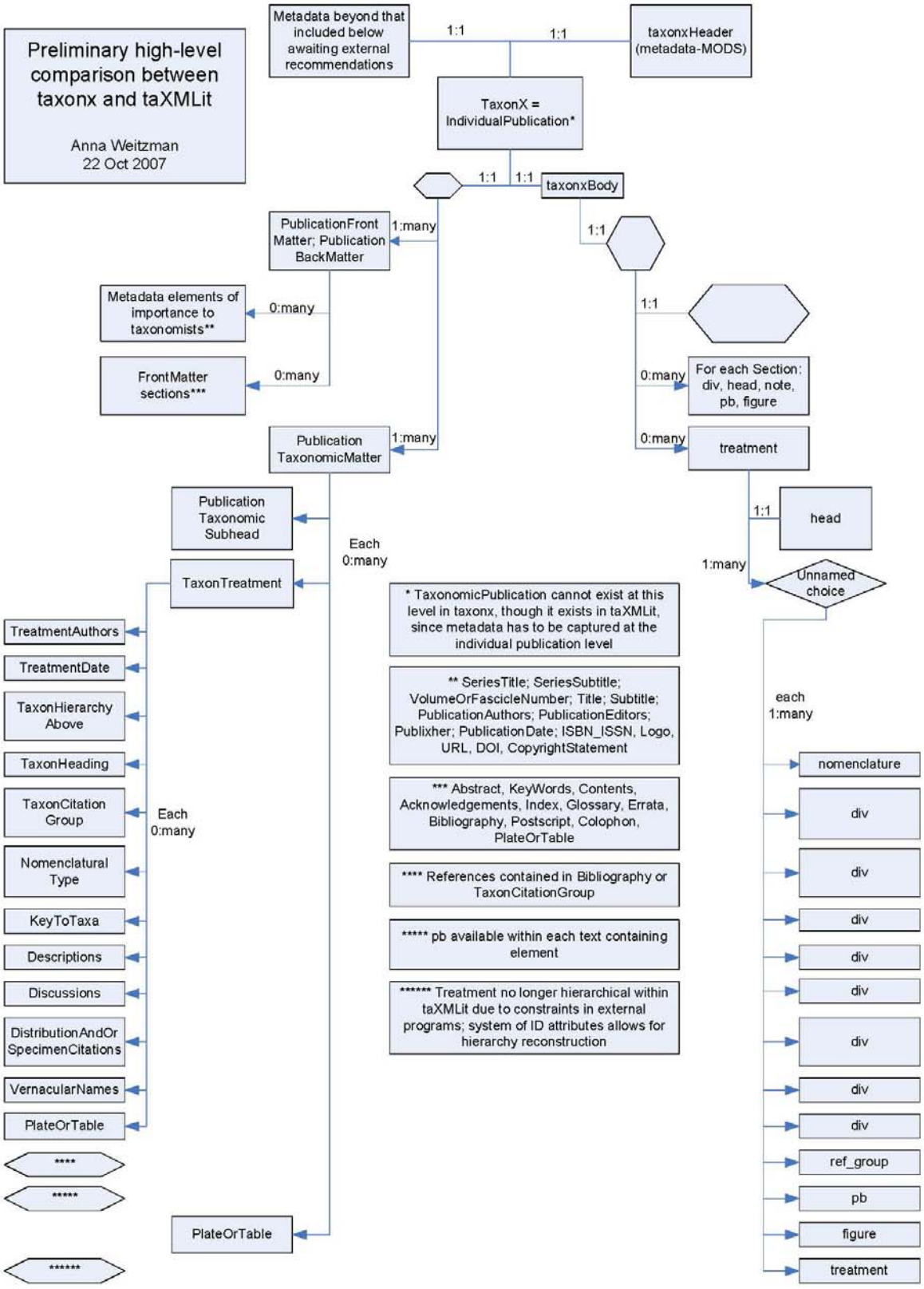

Fig. 2 . Comparison between taXMLit and taxonX (after Carapano and Weitzman, 2007)

Further, there should be one more set of criteria for comparison and evaluation between the schemas designed for retrospective mark up. We should bear in mind now the many complexities and challenges faced in the process of **retrospective markup**, many of which are not technical at all. For any of these projects the following will have to be considered and clarified:

1) What are the tolerances for text accuracy?
2) What are the editorial policies for, among others:
    a) corrections/retention of typos and other errors in the text
    b) interpretation of unclear text
    c) choice of "copy-text", i.e., the exemplar from which the digitized version of the text will be made. It is *highly* unlikely that every copy of any edition of a work will have exactly the same text
3) What are the policies and practices for normalization and other annotation, such as:
    a) expansion of abbreviations
    b) normalization of taxon names, personal names, corporate names, etc...
    c) modernization of archaic or changed place names (e.g., Rhodesia/Zimbabwe)
    d) annotation and other editorialization, as for example, correction of incorrect taxon names, assignment of coordinates to location names, etc...
4) What are the textual objects of interest which will be encoded (i.e., do not aim to tag everything). What is in scope, and what is not? What has the highest priority?
5) What are the purposes of the markup? Just one cannot "tag everything" no single encoding of a text is going to be equally suitable for any purpose. Three main categories can be seen as:
    a) rendition/representation of the text in HTML, PDF, ePub, or other formats
    b) archiving of the text for long term preservation
    c) analysis, data mining, and other processing
6) What are the policies and practices for the handing of non-textual features such as illustrations, inserted plates, fold-out maps, etc...?
    a) how should multicolumn text be handled?
    b) what are the policies and practices regarding overlapping hierarchies in the text?

We consider such a comparative analysis very important, not to say crucial, for the success of VIBRANT and beyond.

Semantic tagging and semantic enhancements to taxonomic papers

Semantic tagging is generally considered to be a method of assigning markers, or tags, to text strings to identify their meaning so that the string and its meaning can be made discoverable and readable not only by humans but also by computers. Further, relations between objects could be modelled as formalized expressions of relationship (triples). There are several computer languages developed to provide text tagging, the most popular of them being the eXtensible Markup Language (XML) (see next section). Special machine-readable XML documents called "XML schemas" constrain the valid use of each tag, and so provide the background for semantic tagging. For example, in basic XML one can tag the name *Drosophila melanogaster* with the tag TaxonName. Provided users' tools take care to uniformly use this for an actual taxon name, there will be no semantic discord among or within documents about what is a taxon name, and software tools can easily be built to exploit these implicit community agreements about meaning. Special languages, namely XML-Schema and the XML Document Type Definition (DTD) can express syntactic restrictions on documents that enforce some context on the use of community-designed controlled vocabulary. When documents comply with these restrictions, it is then possible to write and support software to perform meaningful searches within or across documents, to transform documents from one form to another (e.g. from XML to PDF or HTML), or to facilitate a standardised way for archiving and computer retrieval of the whole document.

At the forefront of informatics research, visions of a fully Semantic Web are advancing (http://en.wikipedia.org/wiki/SemanticWeb) but these seem to remain over the horizon for robust scientific publishing. It is beyond the scope of the present paper to cover in fine detail the vast and extremely dynamic area of semantic tagging, even in the sense we use it. We illustrate how tagging works in taxonomic publications with the following simple example (Fig. 1). Thanks to tagging, computers can recognise portions delimited between the start and end tags to have a certain meaning, thus they can retrieve tagged texts, extract information from them, direct elements to databases and so on.

Semantic tagging is often related to semantic enhancements providing a good basis for the latter. The terms, however, are not identical. Semantic enhancement to scientific texts can be determined as "anything that enhances the meaning of a published journal article, facilitates its automated discovery, enables its linking to semantically related articles, provides access to data within the article in actionable form, or facilitates integration of data between articles" (Shotton et al. 2009).

In the current mature XML technologies, semantic enhancements are typically used for a better visualization and utilization of published text through various hyperlinks, either within the text or to external resources, while tagging is mostly used to transform a text into a computer-readable form. In case of treatment mark-up, a search for taxa living in a particular area would not list any pages containing the particular taxon and locality name (e.g. what Google does) but would provide a list of taxa or treatments for the particular area. Tagged text could be presented in a simple, "non-enhanced" form, and vice versa, semantically enhanced papers need not necessarily be based on XML-tagged text. Important new and rapidly developing areas of semantic enhancements include the so-called "mashup" and "linkout" technologies created to utilize data from different online resources (e.g., mapping geographical localities of a taxon

harvested from different articles, datasets and websites. Linkout software tools locate strings or identifiers within certain Web resources (e.g., through a taxon name or its persistent identifier), receive back the information (often in XML or JavaScript Object Notation [JSON] formats) and represent a summary of that information on a resulting webpage. Harvesting web resources with the help of so-called "scraper" or "harvester" software can be made dynamically, that is in real time (mostly through APIs, Application Programming Interfaces, when these are available on the source website) or by search/provide functions.

Use cases of semantic enhancements

There are not that many cases of usage of semantic enhancements to taxonomic literature, and even fewer when we look for application of them as routine practice in the publishing process. The topic is relatively new and is mostly discussed at the level of research or proposals for uses in the "articles of the future". Among the few papers that treat semantic enhancements to biodiversity papers we should mention Shotton (2009), Shotton et al. (2009), and Page (2010).

The topic has been reviewed in a special issue of ZooKeys (Penev at al. 2010) and illustrated through several sample papers in the same issue 50 of ZooKeys. Issue 50 established the following methods of semantic enhancements to biodiversity papers and few more have been added afterwards:

1. For the first time, a newly published taxonomic revision can be searched and retrieved for taxon treatments.
2. Treatments , taxon names and citations can be identified through the papers in different color so that to easily identify them during the reading process
3. Georeferenced localities can be mapped on Google Maps for separate treatments, or collated for groups of treatments (e.g, for all species in a genus treated in the papers)
4. Occurrence data can be published as supplementary KML file and visualized on Google Earth
5. Citations in the text are cross-linked with the reference lists; each citation can be visualized as full text reference by pointing out the cursor on it;
6. Figure citations are cross-linked with the figures themselves; each figure can be visualized just by pointing the cursor on its citation
7. Each taxon name published in the paper, independently of its rank, is linked to its dynamic online profile (Pensoft Taxon Profile, PTP, www.ptp.pensoft.eu), created on the fly. PTP links the taxon name to a number of searched biodiversity resources, such as GBIF, EOL, NCBI, BHL, IPNI, Index Fungorum, ZooBank, Tropicos, PLANTS database, Morphbank, Wikipedia, Wikispecies, Yahoo images etc.
8. Each new taxon name is registered with IPNI (plants) or ZooBank (animals) and the respective LSIDs are listed in the published paper.

9. A tool was developed that exports taxon treatments to the wiki platform www.species-id.net; the link to the respective wikipage is situated under the taxon name in the nomenclature section of the treatment
10. Occurrence (and other) datasets are published either as supplementary data files or through the IPT (Integrated Publishing Toolkit) of GBIF

Summary and conclusions

This report should serve as a basis for choosing the right strategy in implementation of different XML schemas for mark up of taxonomic texts within the ViBRANT project and beyond. The three reviewed schemas – taxonX, TaxPub and taXMLit – cover the main tasks of taxonomy mark up rather well. TaxonX is a lightweight, object-centred schema focusing on taxonomic treatments extracted from legacy literature; taXMLit is a document-centred, very detailed schema covering mark up of legacy literature; TaxPub is an extension to the NLM journal publishing DTD and has been created to support prospective publishing in taxonomy.

All three schemas have advantages and shortcomings outlined in the text; they also have passed the stages of creation, testing and implementing through a number of use cases.

The report does not recommend choosing one or some of the schemas. Rather, it proposes several cross-points that can be used to match common elements present in both legacy and present-day taxonomic literature. A common output, when needed, from documents marked up in the different schemas could be achieved through XSLIT conversions. Most important common elements in differently tagged text are: taxonomic names, taxon treatments, nomenclatural acts, literature references, as well as the overall structure of the published document and its bibliographic metadata.

The report also outlines several questions to be answered when evaluating a certain schema to be used for the different goals of ViBRANT, Scratchpads and beyond.

References

ABCD - Access to Biological Collection Data - a joint CODATA and TDWG initiative. (http://www.bgbm.org/TDWG/CODATA/)

Agosti D, Egloff W (2009) Taxonomic information exchange and copyright: the Plazi approach. BMC Research Notes 2: 53. doi: 10.1186/1756-0500-2-53

Agosti D, Klingenberg C, Sautter G, Johnson N, Stephenson C, Catapano T (2007) Why not let the computer save you time by reading the taxonomic papers for you? Biológico, São Paulo 69 (suplemento 2): 545-548.

Blagoderov V, Brake I, Georgiev T, Penev L, Roberts D, Rycroft S, Scott B, Agosti D, Catapano T, Smith VS (2010a) Streamlining taxonomic publication: a working example with Scratchpads and ZooKeys. ZooKeys 50: 17–28. doi: 10.3897/zookeys.50.539

Blagoderov V, Hippa H, Nel A (2010b) *Parisognoriste*, a new genus of Lygistorrhinidae (Diptera, Sciaroidea) from the Oise amber with redescription of Palaeognoriste Meunier. ZooKeys 50: 79–90. doi: 10.3897/zookeys.50.506

Brake I, von Tschirnhaus M (2010) Stomosis arachnophila sp.. n., a new kleptoparasitic species of freeloader flies (Diptera, Milichiidae). ZooKeys 50: 91–96. doi: 10.3897/zookeys.50.505

Chavan VS, Ingwersen P (2009) Towards a data publishing framework for primary biodiversity data: challenges and potentials for the biodiversity informatics community. BMC Bioinformatics 2009, 10 (Suppl 14): S2. doi: 10.1186/1471-2105-10-S14-S2

Costello MJ (2009) Motivating online publication of data. BioScience 59: 418-427. doi: 10.1525/bio.2009.59.5.9.

Dallwitz MJ A (1980) general system for coding taxonomic descriptions. Taxon 291-43 Also available at http://delta-intkey.com.

Erwin TL, Johnson PJ (2000) Naming species, a new paradigm for crisis management in taxonomy: rapid journal validation of scientific names enhanced with more complete description on the Internet. The Coleopterists Bulletin 54 (3):269-278.

Fisher BL, Smith MA (2008) A Revision of Malagasy Species of *Anochetus* Mayr and *Odontomachus* Latreille (Hymenoptera: Formicidae). PLoS ONE 3(5): e1787. doi: 10.1371/journal.pone.0001787

Johnson NF, Masner L, Musetti L, van Noort S, Rajmohana K, Darling DC, Guidotti A, Polaszek A (2008) Revision of world species of the genus *Heptascelio* Kieffer (Hymenoptera: Platygastroidea, Platygastridae). Zootaxa 1776:1-51.

Lyal, C.H.C. & Weitzman, L., 2008. Releasing the content of taxonomic papers: solutions to access and data mining. Proceedings of the BNCOD Workshop "Biodiversity Informatics: challenges in modelling and managing biodiversity knowledge" http://biodiversity.cs.cf.ac.uk/bncod/LyalAndWeitzman.pdf

Mengual X, Ghorpadé K (2010) The flower fly genus *Eosphaerophoria* Frey (Diptera, Syrphidae). ZooKeys 33: 39–80. doi: 10.3897/zookeys.33.298

Miller JA, Griswold CE, Yin CM (2009) The symphytognathoid spiders of the Gaoligongshan, Yunnan, China (Araneae, Araneoidea): Systematics and diversity of micro-orbweavers. ZooKeys 11: 9-195. doi: 10.3897/zookeys.11.160

Morse, D., Dil, A., King, D., Willis, A., Roberts, D. & Lyal, C., 2009, *Improving search in scanned documents: Looking for OCR mismatches*. Pp 58-61 in Bernardi, R., Chambers, S. & Gottfried, B (eds), Proceedings of the workshop on advanced technologies for digital libraries (AT4DL 2009) http://purl.org/bzup/publications/9788860460301

Page RDM (2006) Taxonomic names, metadata, and the Semantic Web. Biodiversity Informatics 3:1-15.

Page RDM (2010) Enhanced display of scientific articles using extended metadata. Web Semantics: Science Service Agents World Wide Web. doi: 10.1016/j.websem.2010.03.004

Penev L, Erwin T, Miller J, Chavan V, Moritz T, Griswold C (2009a) Publication and dissemination of datasets in taxonomy: ZooKeys working example. ZooKeys 11: 1-8. doi: 10.3897/zookeys.11.210

Penev L, Sharkey M, Erwin T, van Noort S, Buffington M, Seltmann K, Johnson N, Taylor M, Thompson FC, Dallwitz MJ (2009b) Data publication and dissemination of interactive keys under the open access model: ZooKeys working example. ZooKeys 21: 1–17. doi: 10.3897/zookeys.21.274

Polaszek A et al. A universal register for animal names 2005. Nature 437477. DOI: 10.1038/437477a.

Pyle RL, Earle JL, Greene BD (2008) Five new species of the damselfish genus *Chromis* (Perciformes: Labroidei: Pomacentridae) from deep coral reefs in the tropical western Pacific. Zootaxa 1671:3-31.

Sautter G, Böhm K, Agosti D (2007) A Quantitative Comparison of XML Schemas for Taxonomic Publications. Biodiversity Informatics 4: 1–13. https://journals.ku.edu/index.php/jbi/article/view/36

Sautter G Böhm K Padberg F Tichy W Empirical Evaluation of Semi-Automated XML Annotation of Text Documents with the GoldenGATE Editor Proceedings of European Conference on Research and Advances in Digital Libraries 2007. Budapest, Hungary .

Sautter G Böhm K Agosti D Klingenberg C Creating digital resources from legacy documents: An experience report from the biosystematics domain Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications 2009. Heraklion, Crete .

Sharkey MJ, Yu DS, van Noort S, Seltmann K, Penev L (2009) Revision of the Oriental genera of Agathidinae (Hymenoptera: Braconidae) with an emphasis on Thailand including interactive keys to genera published in three different formats. ZooKeys 21: 19–54. doi: 10.3897/zookeys.21.271

Shotton D (2009) Semantic Publishing: the coming revolution in scientific journal publishing. Learned Publishing 22(2): 85–94. doi: 10.1087/2009202

Shotton D, Portwin K, Klyne G, Miles A (2009) Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article. PLoS Comput Biol 5(4): e1000361. doi: 10.1371/journal.pcbi.1000361

Smith V (2009) Data publication: towards a database of everything. BMC research Notes 2: 113. doi: 10.1186/1756-0500-2-113

Stoev P, Akkari N, Zapparoli M, Porco D, Enghoff H, Edgecombe GD, Georgiev T, Penev L (2010) The centipede genus *Eupolybothrus* Verhoeff, 1907 (Chilopoda: Lithobiomorpha: Lithobiidae) in North Africa, a cybertaxonomic revision, with a key to all species in the genus and the first use of DNA barcoding for the group. ZooKeys 50: 29–77. doi: 10.3897/zookeys.50.504

Taekul C, Johnson NF, Masner L, Polaszek A, Rajmohana K (2010) World species of the genus Platyscelio Kieffer (Hymenoptera, Platygastridae). ZooKeys 50: 97–126. doi: 10.3897/zookeys.50.485

Talamas EJ, Johnson NF, van Noort S, Masner L, Polaszek A (2009) Revision of world species of the genus *Oreiscelio* Kieff er (Hymenoptera, Platygastroidea, Platygastridae). ZooKeys 6: 1-68. doi: 10.3897/zookeys.6.67

TDWG (2007 onwards) TDWG: standards. Biodiversity Information Standards. http://www.tdwg.org/standards/ [accessed 31.VIII.2009].

Weitzman AL, Lyal CHC. An XML schema for taxonomic literature – taXMLit - http://www.sil.si.edu/digitalcollections/bca/documentation/taXMLitv1-3Intro.pdf

Weitzman AL & Lyal CHC, 2006, INOTAXA — INtegrated Open TAXonomic Access and the "*Biologia Centrali-Americana*". *Proceedings Of The Contributed Papers Sessions Biomedical And Life Sciences Division, SLA*. 8pp. http://units.sla.org/division/dbio/Baltimore/index.html

Weitzman AL, Lyal CHC. An XML schema for taxonomic literature – taXMLit - http://www.sil.si.edu/digitalcollections/bca/documentation/taXMLitv1-3Intro.pdf

Willis A, King D, Morse D, Dil A, Lyal C, Roberts D, From XML to XML: The why and how of making the biodiversity literature accessible to researchers Language Resources and Evaluation Conference (LREC) 19-21May2010. Malta.

Winston J. (1999) Describing Species. New York: Columbia University Press.