Project Acronym: **pro-iBiosphere**

Project Full Title: **Coordination & policy development in preparation for a European Open Biodiversity Knowledge Management System, addressing Acquisition, Curation, Synthesis, Interoperability & Dissemination**

Grant Agreement: **312848**

Project Duration: **24 months (Sep. 2012 - Aug. 2014)**

D.4.2 Strategy for improvement & interoperability of the XML schemas

Deliverable Status: **Final**
File Name: **pro-iBiosphere_WP4_Plazi_D4.2_VFF31_01_2014.PDF**
Due Date: **December 2013 (M16)**
Submission Date: **January 2014 (M17)**
Dissemination Level: **Public**
Authors: **Plazi, Pensoft, BGBM, NBGB, RBGK, Naturalis**

Consisting of:

FUB-BGBM	Freie Universität Berlin, Germany
MfN	Museum für Naturkunde, Germany
Naturalis	Stichting Nederlands Centrum voor Biodiversiteit Naturalis, Netherlands
BGM	Botanic Garden, Meise, Belgium
Pensoft	Pensoft Publishers Ltd, Bulgaria
Plazi	Plazi, Switzerland
RBGK	Royal Botanic Gardens Kew, United Kingdom
Sigma	Sigma Orionis, France

REVISION CONTROL

Version	Author	Date	Status
01	Donat Agosti	15.10.2013	First draft structure
02	David Patterson	23.12.2013	Read through and revise
03	Lyubomir Penev	21-31.12.2013	Re-structuring the draft, additions, edits
04	József Geml	02.01.2014	Content on fungi pilot
05	Jordan Biserkov, Teodor Georgiev	04.01.2014	Description of DwC-A
06	Donat Agosti	07.01.2014	Content
07	Terry Catapano	07.01.2014	Content
08	Peter Hovenkamp	07.01.2014	Review
09	Donat Agosti	10.01.2014	Review of draft
10	Peter Hovenkamp	12.01.2013	Review of draft
11	Soraya Sierra, Peter Hovenkamp	13.01.2014	Review of draft
12	Thomas Hamann, Soraya Sierra	14.01.2014	Review of draft
13	Lyubomir Penev, David Patterson, Thomas Hamann, Soraya Sierra	16.01.2014	Editing draft, resolving comments
14	Donat Agosti, Robert Morris, Terry Catapano	16-19.01.2014	Editing and resolving comments
15	Andreas Müller, Soraya Sierra, two anonymous reviewers	24.01.2014	Review of draft
16	Donat Agosti, David Patterson, Lyubo Penev	25-26.01.2014	Resolving comments
17	P. Hovenkamp, Soraya Sierra, Donat Agosti	27-31.01.2014	Review and editing of final draft
FF	Soraya Sierra	31.01.2014	Final Draft converted to Portable Document Format (PDF)

Authors:

Donat Agosti (Plazi)
Christine Barker (RBGK)
Terry Catapano (Plazi)
Hong Cui (University of Arizona)
Sabrina Eckert (BGBM)
Jordan Bisrekov (Pensoft)
Teodor Georgiev (Pensoft)
Quentin Groom (BGM)
Anton Güntsch (BGBM)
Gregor Hagedorn (MfN, Plazi)
Thomas Hamann (Naturalis)
Peter Hovenkamp (Naturalis)
Patricia Kelbert (BGBM)
Paul Kirk (RBGK)
Don Kirkup (RBGK)
Jeremy Miller (Naturalis)
Robert Morris (Plazi)
Anton Müller (BGBM)
Sylvia Mota de Oliveira (Naturalis)
Lyubomir Penev (Pensoft)
Soraya Sierra (Naturalis)
Tony Walduck (RBGK)

Contributions from:

Paul Kirk (RBGK)
Christine Barker (RBGK)
Vincent Roberts (Centraalbureau voor Schimmelcultures)
Rich Pyle (Bishop Museum)
Daniel Mietchen (Museum für Naturkunde, Berlin)

Mark-up Pilot Participants:

Donat Agosti (Plazi)
Hong Cui (University of Arizona)
Teodor Georgiev (Pensoft)
Pavel Stoev (Pensoft)
Lyubomir Penev (Pensoft)
Jordan Biserkov (Pensoft)
Quentin Groom (NBGB)
Thomas Hamann (Naturalis)
Peter Hovenkamp (Naturalis)
Don Kirkup (RBGK)
Jeremy Miller (Naturalis)
Soraya Sierra (Naturalis)
Sylvia Mota de Oliveira (Naturalis)
Tony Walduck (RBGK)

Contents

Executive summary

The pro-iBiosphere Task 4.2 “Promote and monitor the development and adoption of common mark-up standards and interoperability between schemas” has used pilot studies to analyse the currently applied XML schemas for mark-up of taxonomic information. Four interoperability pilots were conducted, drawing on a broad spectrum of organisms (i.e. animals, higher plants, fungi and bryophytes).

In the course of these pilots, we have made the following new additions and improvements to the e-infrastructure that is available for use within Europe:

1. adopted the TaxPub Journal Article Tag Suite (JATS) to serve as a standard system for the semantic mark-up of publications in the area of biological systematics;
2. developed an Extensible Stylesheet Language Transformations (XSLT) mechanism to inter-convert content from TaxPub and TaxonX;
3. developed a common XML query model to automatically register names in nomenclatural registries such as the International Plant Name Index (IPNI) for plants and ZooBank for animals;
4. adapted the DarwinCore-Archive to transfer data on the citations of materials and taxonomic treatments (Plazi, Pensoft);
5. developed a proof-of-concept workflow to link publishers (in this case Pensoft), repositories (Plazi), and dedicated taxonomic platforms (CDM);
6. explored and established feasibility of the use of Resource Description Framework (RDF) for modelling and transfer of information in taxonomic treatments.

Recommendations

The results show that mark-up requires substantial effort and human resources. In the case of many types of legacy literature, these are the most important limiting factors in achieving a sufficient degree of interoperability. We conclude that a strategy to improve machine-based dialogue of semantic enhanced biodiversity information relating to taxonomic treatments should contain at least the following elements:

1. adoption of TaxPub JATS for mark-up of new taxonomic literature;
2. adoption of TaxonX for mark-up of legacy literature, and investment in the further development that will overcome current limitations;
3. development of instruments for lossless conversion among environments such as TaxPub and TaxonX that will ensure integration of data from historical and recent literature;
4. development of industrial strength digitisation workflows that will format the corpus of legacy literature into useful machine-readable content;
5. endorsement of Darwin Core Archive as an interchange format for occurrence data and taxonomic treatments;

6. utilisation of RDF as a format for distribution of data extracted from the content of marked-up treatments;
7. substantial investment in mark-up tools with improved user interfaces and efficiency so as to achieve high accuracy and reduced labour costs.

This work complements the outcomes reported in the "*Report on ongoing biodiversity related projects, current e-infrastructures and standards* (Agosti et al., 2013)[1].

[1] http://tinyurl.com/nhxsnju

Introduction

The Open Biodiversity Knowledge Management System (OBKMS) is a vision for the future management of biodiversity data. It relies on access to machine-readable data to discover, extract, mine and reuse content so as to create new insights. The corpus of scientific literature is the largest source of biodiversity data. Publications codify and create units of biodiversity knowledge in descriptions, definitions, and discussions of taxonomic concepts. A taxonomic concept is the meaning of a name as established by a treatment in a scientific publication. This process is consistent across all life (plants, animals, fungi, protists, prokaryotes, and viruses) and applies to historical accounts as well as contemporary ones. The treatments may vary in detail and structure, but will include a minimum set of elements that is specified by the taxonomically relevant nomenclatural codes (e.g. McNeill et al. 2012; ICZN 2012). The components of a treatment that accompany a name will tend to include sections on traits, distribution, ecology, observations, images, and reference to the literature and specimens. Because of the consistency of treatments, the structure can be formally represented for machine use using mark-up and semantic technologies.

Open Biodiversity Knowledge Management requires new mechanisms to make content ready for machine use and to ensure interoperability (see pro-iBiosphere report D2.1.1 *Report on ongoing biodiversity related projects, current e-infrastructures and standards*[2]). This implies the atomisation of the content into sufficiently granular semantic elements (statements, facts, observations, values of variables), which need to be described using widely accepted metadata and ontologies. While progress is being made, the goal is not yet fully achieved because different systems are being used by different users.

One of the steps in mobilising content such as treatments, is the mark-up of structural and semantic items within the treatment. This allows these elements to be identified in ways that makes them machine accessible. Mark-up can be conducted by hand, semi-automated tools, or scripts developed to identify the elements in the source document. Mark-up is often conducted in compliance with particular data schemas. The goal is to unify the output from the different mark-up tools; improve the merger of data from legacy and recent publications; promote data harvesting; and ensure preservation of treatments for all taxa. Here we review the Extended Mark-up Language (XML) schemas for mark-up of taxonomic information, the mark-up processes, and address strategies that will streamline data exchange. We rely on the use of the pro-iBiosphere pilots to evaluate schemas and processes. In particular, we address the integration of names registration with the publication process. This work complements deliverable D2.1.1 *Report on ongoing biodiversity related projects, current e-infrastructures and* standards[2] and reports that deal with general workflows (D 3.3.1[3]) and costs involved in mark-up (D 6.1.2, due in May 2014).

Tools, formats, and protocols already exist for the exchange of formalised elements such as names, bibliographic records, and references to materials. They have been developed by members of the pro-iBiosphere team, and are deployed and reviewed in this project. Examples are the Pensoft / ZooBank / IPNI XML query model for names and the Pensoft / Plazi GBIF (Global Biodiversity

[2] http://tinyurl.com/nhxsnju

[3] http://wiki.pro-ibiosphere.eu/wiki/D3.3.1_Semantic_integration_of_the_biodiversity_literature

Information Facility) /EOL (Encyclopedia of Life) DarwinCore-Archive for materials citations described below. Within this project, we have extended the available toolkit with new tools to deal with citations of treatments. The process is far from complete, and tools to work with more complex and/or less rigidly formalised elements, such as character traits, need further development.

A current standard for machine-readable, structured, semantically enhanced and linked taxonomic article is TaxPub developed as an extension of the widely used Journal Article Tag Suite (JATS). It is deployed in a series of Pensoft journals (Penev *et al*. 2012). TaxPub articles demonstrate the vision of dissemination well beyond the traditional channels of paper-based publications. TaxPub-based journals are integrated into the world's largest Biomedical archive, PubMed Central, guaranteeing connection of content well beyond the immediate target domain. The approach enables the content to be acquired by aggregators like GBIF (observation records), EOL, Plazi, or CDM (Common Data Model). Content can be extracted for import into databases or converted into other literature mark-up schemas.

As the potential of semantically enhanced publications is now becoming clear, it is now beginning to be adopted by taxonomic journals, but most information is still available in hard copy only or in unstructured Portable Document Format (PDF). This presents the challenge of how to capture the content in this huge corpus of literature encompassing hundreds of millions of pages in ways that will make it accessible to an OBKMS. The criteria for schemas to convert this corpus into machine-readable formats include: full compatibility with TaxPub and compatibility with existing vocabularies such as Taxon Concept Schema and Darwin Core. Important considerations for digitising and marking up this corpus include: costs of conversion (with an emphasis on human resources); how well the data and mark-up meet the needs of end-users; and the nature of the associated toolkits and support. Costs and fitness for use will promote or deter the changes to working practices that must accompany OBKMS (Thessen & Patterson 2011).

Despite the basic structural similarities among treatments, their authors and those who will reuse data have different needs that affect the structure and granularity of the content. Reuse of the data may thus result in a need for more granularity or structure and so may drive the evolution of mark-up tools towards adaptability to meet the needs of a broad community of users. For instance, GoldenGATE is a semi-automatic open source tool developed by members of the pro-iBiosphere group to accelerate the semantic mark-up of scientific literature. Participants of the pilot studies were trained in its use to provide feedback for further development and to give users the flexibility to customise the process to meet their own needs.

What is 'interoperability'?

Baumann (2011) identifies three types of data exchange. "Negotiated Interchange" involves human communication and intervention to establish needs and to adjust data formats or systems to achieve compatibility. "Blind Interchange" requires human effort to find and acquire content, with the responsibility for ensuring that the data are fit for purpose being shouldered by the user of the data alone. 'Interoperation', finally, is distinguished as the approach that requires little human intervention, if any. Rather, the recipient accesses data directly from a source, and the data are

automatically used by a system's processes. Bauman concludes, "...an interoperable text is one that does not require any direct human intervention in order to prepare it to be used by a computer process other than the one(s) for which it was created."

We conducted pilot studies to review and analyse XML schemas and striving for as much interoperability as possible. The schemas included TaxPub, TaxonX and ABCD (Access to Biological Collection Data) (Catapano 2010, Penev *et al*. 2011). Different approaches were used in the pilots to explore different aspects of data exchange. The Extensible Stylesheet Language (XSLT) was employed to convert TaxPub documents to GoldenGATE's internal format in one pilot; DarwinCore-Archive was used as data transfer format for treatment-based data in another. Further information about the schemas is provided in the section on 'Methods and Approach'.

Data exchange among XML schemas

We addressed three technologies for different aspects of data exchange: Extensible Stylesheet Language (XSLT) for conversion among schemas, Darwin Core Archive as format to transfer data across systems, and we started discussions about how the Resource Description Framework (RDF) could be employed for data representation utilised by multiple systems.

The first is automatic transformation using the Extensible Stylesheet Language (XSLT). XSLT was used to convert TaxPub documents produced by Pensoft into the GoldenGATE internal XML format, from which it was transformed as TaxonX for download from Plazi's treatment repository. An XSLT transformation essentially takes data from locations in the source document and places it in the proper and valid location in the target document. So, for instance, the name of a taxon in a TaxPub element "<taxon-name>" is converted to the "<taxonomicName>" element in the GoldenGATE (GG) XML format and then to the "<name>" element in TaxonX. To the extent that two or more schemas have similar structures and conceptual understandings of the objects being modelled, transformation of instance documents from one schema to another proceeds fairly easily with little loss of information.

Secondly, the DarwinCore-Archive (DwC-A) is a platform-independent standard developed by GBIF to provide a machine-readable format for exchange of biodiversity data in Darwin Core standard[4] [5]. It is extensible, self-documenting and easy to produce. It is a 'star schema' with, at its core, a central data file, for example a taxon checklist or a set of occurrence records, to which any number of extensions files can be linked to the core file through links to identifiers for records in that file (Fig. 1). The format is popular and is endorsed as the interchange format for Scratchpads with other platforms such as EOL, EDIT CDM and others (Baker et al. 2014). DwC-A was tested as a conversion format from text-based checklists into machine-readable formats (Remsen et al. 2012). The human effort required makes it unsuitable as a routine method. Its star schema is not able to satisfy the richer relational model required for the EOL, the Global Genome Biodiversity Network (GGBN) and

[4] http://rs.tdwg.org/dwc/terms/index.htm
[5] http://code.google.com/p/gbif-ecat/wiki/DwCArchive

others. Work has started on generalising the DarwinCore-Archive to a “Biodiversity Data Archive” (Wieczorek et al. in press).

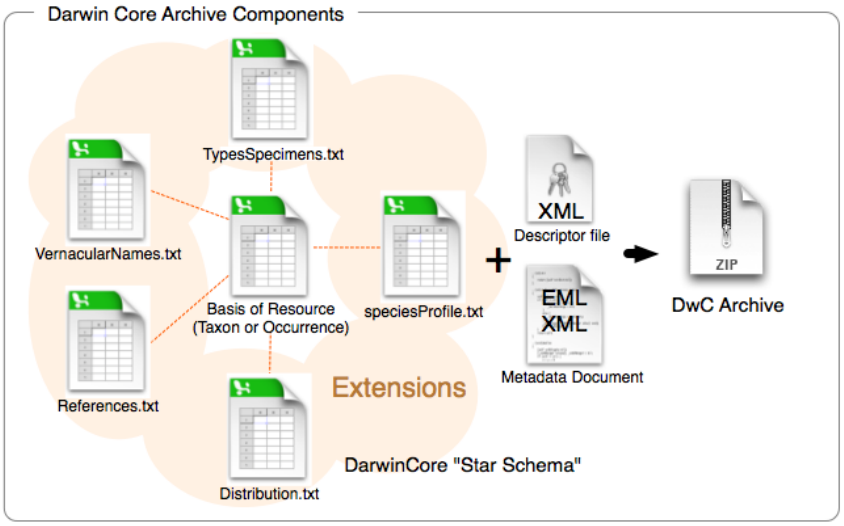

Figure 1. The Darwin Core Star schema showing the relations of different data files to the core file[6].

The pro-iBiosphere team explored the potential of DwC-A to deliver machine-readable files with information from treatments. Plazi adopted a ‘Taxon treatments’ core to which other files are linked via *taxonID* with a value of ‘treatment ID + ".taxon"’. This allows **one** treatment to have **many** occurrences, media items, references, vernaculars, distributions or descriptions.

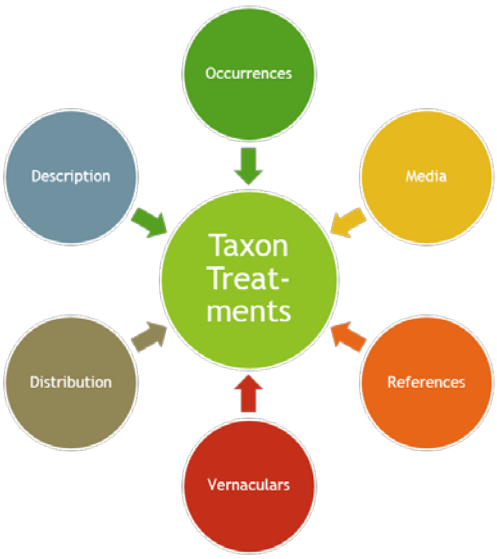

Figure 2. DarwinCore-Archive schema used by Plazi.

[6] http://code.google.com/p/gbif-ecat/wiki/DwCArchive

Pensoft has explored a similar DwC-A - based approach to structuring data in a way that best satisfies the needs of data transfer to the EOL (Fig. 3). This approach allows the consumer, such as EOL, to show only the references relevant to an image, or to a particular section of the treatment.

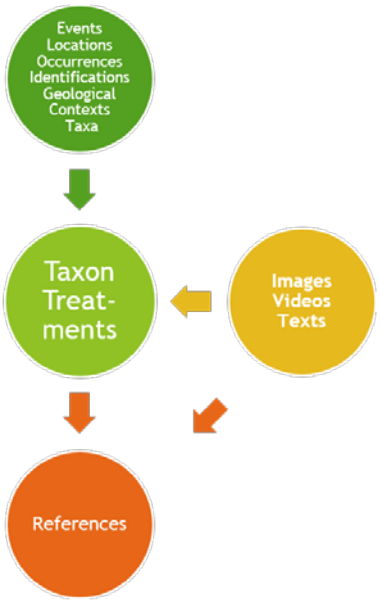

Figure 3. DarwinCore-Archive schema used by Pensoft.

The third mechanism we considered as a possible means of improving access to information from different schemas is RDF[7]. Users are faced with a greater workload if they have to access structured data in different schemas. It is simpler if all schemas export content in RDF as a common data structure, matched to a common ontology (Fig. 4).

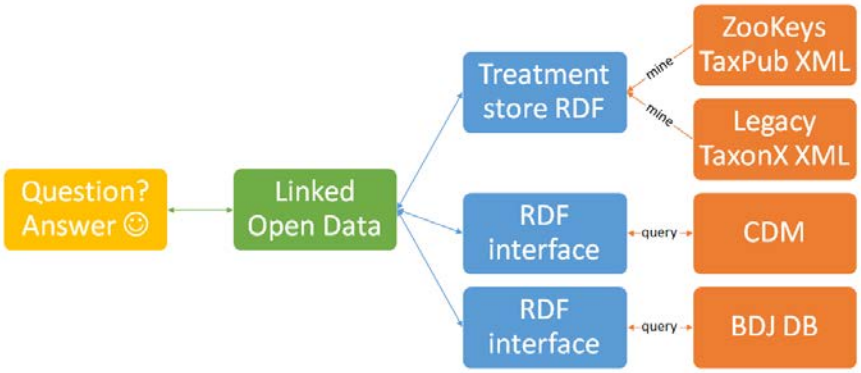

Figure 4. Improving interoperability among schemas by using RDF as a common access mechanism.

The RDF information is available as triples that make up 'Linked Open Data' and can be queried by SPARQL (SPARQL Protocol and RDF Query Language) services. Abbreviations: RDF - Resource

[7] http://www.w3.org/RDF/

Description Framework; CDM - Common Data Model; BDJ DB - Biodiversity Data Journal database shown here as an example of a full XML-based publishing workflow. “ZooKeys” is an example of a journal that publishes marked-up text allowing extraction of treatments. “Legacy” means the source of historical literature from where treatments can be extracted through mark-up techniques and tools.

The advantage of using RDF is that it removes variation in the syntax of data in records and documents as an obstacle to interchange, by representing the facts being asserted by the data sources and not their structure or expression. To the extent that the concepts being used by different sources can be resolved to a common definition and understanding of the relationships to other concepts, those concepts can be formally expressed in an ontology. So, for example, TaxPub may encode a reference to a taxonomic name with a <taxon-name> element, while TaxonX uses a <name> element. Both can be related to a term in an ontology, for example Darwin Core’s “scientificName” so that the assertion that a treatment is about a certain taxonomic concept with the identifier http://example.com/treatment/xyz can be expressed by the RDF triple:

http://example.com/treatment/xyz dwc:scientificName “Aus Bus”

This assertion can then be used by any RDF-aware application and thus conversion between elements mapped to the same concept can be facilitated. Discussions were initiated at the project meetings and workshops about the extent to which data from treatments could be represented in RDF. Plazi has begun development of a small ontology. Modelling treatments and existing ontologies, such as DarwinCore and DublinCore, could be used to construct triples representing taxonomic information in treatments. One advantage of RDF is that, with provision of suitable ontologies such as the Treatment Ontology under development, various mature machine reasoners can, without further programming, provide sophisticated semantics extracted from data even though it is not explicit. For example, if a biological taxonomy expressed in RDF as an ontology were presented along with a query like “Find all treatments of genera in the plant family Rubiaceae”, the correct answers would be returned even if the underlying data are incomplete as to which species are in that family. Once represented in RDF, triples can be merged into graphs which can then be used by different systems capable of ingesting and processing RDF data. RDF presents a flexible and standardised means for representing assertions. As with any data standard, successful application of RDF will depend upon the fitness of data to applications, and applications will depend upon development of the questions and analyses within the context of a Knowledge Management System. Despite its potential, RDF was not tested in the pilots as its implementation requires significant resources that had not been included in the original budget.

Methods and Approach

Pilots

The pilots in the project were designed and selected to address two major issues: the feasibility of allowing end-users mark-up legacy taxonomic literature, and the interoperability between different mark-up schemas.

The subjects for the mark-up pilots were selected from a broad taxonomic spectrum (see pro-iBiosphere Wiki[8] and Appendix) and used both legacy sources of various ages and prospective literature. The pilots were designed with the purposes of: (i) assessing the available tools and costs of markiup and extracting semantically ready content from source documents; and (ii) establishing the extent of taxon-related problems in the mark-up process - are there (for example) issues that are uniquely associated with sources that relate to plants versus sources related to animals.

To discuss Interoperability pilot 4 (see below), Plazi attended an initial three day technical workshop in Tucson, on December 5-7, 2012.

Table 1. Summary of mark-up pilots. The rows distinguish the target taxa, and the columns indicate the number of journals in which the treatments occurred, the number of treatments, and whether the treatments were prepared with (prospective) or without (legacy) semantic enhancement.

Taxon	Number of journa	Number of treatmen	Type of treatmen
Ants (animals)	40	486	Legacy
Chenopodium (plants)	15	174	Legacy
Spiders (animals)	59	334	Legacy
Fungi (Mycota)	5	5	Legacy
Nephrolepis (plants: fern	1	35	Legacy
Bryophyta (plants: mosse	2	25	Legacy
Eupolybothrus (animals)	152	154	Legacy / prospect
Mistletoes (plants)	4	35	Legacy
TOTAL	**274**	**1050**	

[8] http://wiki.pro-ibiosphere.eu/wiki/Pilots

Training

To introduce two open source tools (GoldenGATE[9] and CharaParser) developed to convert unstructured taxonomic documents into semantically enhanced documents two trainings were organised by pro-iBiosphere. The first one in January 2013 in Leiden [10], and second one during the pro-iBiosphere meeting held in February 2013.

For the former, a pro-iBiosphere version has been developed, which includes a module allowing the direct import of PDFs, thus circumventing a separate Optical Character Recognition (OCR) pre-processing step, with a customised manual [11]. Online support was provided, and progress was reviewed at the pro-iBiosphere meeting in Berlin, October 8, 2013.

The participants of the pilots were all involved in marking up their own documents, all of which finally ended up on the Plazi server.

XML schemas tested

Three XML schemas - TaxonX, TaxPub and ABCD - were evaluated to assess how well XML-encoded data were exchanged to serve the needs of both the sender and recipient. The performance of schemas has previously been compared (Penev et al. 2011, Sautter et al., 2007a). The present report assesses the exchange of taxonomic names, taxonomic treatments, localities, and bibliographic references; areas previously identified as important (D2.1.1 *Report on ongoing biodiversity related projects, current e-infrastructures and standards*[12]). TaxonX[13] and TaxPub[14] were developed and are maintained by Plazi as open source projects for legacy and contemporary publications respectively. Below, we describe the schemas developed by Plazi and tested for interoperability issues within the pro-iBiosphere consortium; ABCD has been developed by Taxonomic Databases Working Group (TDWG) [15]. Other schemas are available, such as the FlorML schema (Hamann et al., in press), TCS XML, and schemas in use at RBGK. Comparison of these formats will be conducted in the next months by the FUB-BGBM and will be documented in the pro-iBiosphere report D3.3.2 (due in April 2014).

TaxonX was developed for data extraction and not to model an entire publication. It was designed for use with taxonomic literature. It is lightweight with about 30 elements appropriate to treatments (e.g. Nomenclature, Materials examined, Description, etc.) and phrase-level features of interest in taxonomy (e.g. scientific names, locality names, characters, etc.). It deals with bibliographic metadata, links to external resources, and semantic normalisation of terms in the source document. The section "Materials examined" can be broken down to individual material citations, normalised and linked to external resources such as a type specimen through LSIDs (Life Science Identifiers). In

[9] http://plazi.org/?q=GoldenGATE
[10] http://wiki.pro-ibiosphere.eu/wiki/Pilots#Golden_Gate_mark-up_training_.28Task_3.29
[11] http://biowikifarm.net/goldenGATE/Main_Page
[12] http://tinyurl.com/nhxsnju
[13] http://www.taxonx.org
[14] https://github.com/tcatapano/TaxPub/releases/tag/v0.5-beta
[15] http://www.tdwg.org/standards/115/

many cases, it relies on use of external schemas such as the Metadata Object Description Schema (MODS) for modelling file level bibliographical metadata and Darwin Core for observation data. TaxonX is flexible, is quickly learned and may be applied to the wide variety of formatting present in legacy documents as well as new publications. TaxonX schema is used by the Plazi Treatment Repository (PTR) of XML-encoded publications and by the semi-automated mark-up tool GoldenGATE (Sautter *et al*. 2007b).

TaxPub was designed for use with semantically enhanced publishing. It is a domain-specific extension of the widely used JATS (Journal Archiving Tag Suite of the US National Library of Medicine), and adapted to biodiversity literature. As it is based on JATS, TaxPub-based articles can be readily imported into PubMed Central (Catapano 2010, Penev et al. 2012). TaxPub was designed primarily for publications containing taxonomic treatments. Treatments are formal descriptions of taxa, and usually include sections on nomenclature, morphological characteristics, behaviour, ecology, distribution, and specimens examined. TaxPub models these features, providing a container element with a required highly structured nomenclature element that contains the essential data about the named species. Other optional elements model the other included sub-sections, such as scientific names, citations of specimens and other materials, and descriptions of organisms. TaxPub relies on the elements in the Journal Publishing Tag Set[16] for all other features. In particular, a <named-content> element that can be applied to a wide range of phrase-level data which may be of interest in taxonomy (e.g. locality information such as latitude, longitude, elevation, etc.). Since July 2009, TaxPub has been implemented in the routine publishing practice of Pensoft. It delivers: (1) semantically enhanced, domain-specific XML versions of articles for archiving in PubMedCentral (PMC); (2) visualisation of taxon treatments on PMC; and (3) export of taxon treatments to various aggregators, such as EOL, PTR, and the Wiki Species-ID.net (Penev et al. 2010, 2012).

Mark-up strategies

Alternative mark-up strategies are either manual or script-based such as FlorML[17] developed by Hamann et al. (in press) or others developed at Kew (Kirkup et al. 2005). Script-based mark-up is semi-automatic, often works best with large taxonomic works such as Floras and Faunas which follow a common though complex structure. It typically requires training. The script-based method will be tested in the next months, and its potential for Floras and Faunas will be compared to other techniques within the upcoming phase of the pro-iBiosphere project. Approaches that are more manual, tend to be more flexible, but are more time-consuming. Further options are macros in word processors such as MS-WORD (Kirkup et al. 2005). Given that there are strengths in both approaches, a combined approach is often best, for instance by including the regular expression based algorithms developed for FlorML into the function rich environment of GoldenGATE. As most treatments can be divided into sections, mark-up can be done incrementally. This allows delivery of treatments fit for a specified use. Progressive mark-up can be done by downloading documents in GoldenGATE from the Plazi SRS server, and re-uploading them when the desired level of mark-up is complete. A progressive approach allows for additions to be made after the initial mark-up, which

[16] http://jats.nlm.nih.gov/publishing/
[17] https://github.com/thoha/FlorML

might include links to external resources, such as type material (see ant pilot); and for more targeted use of other tools like CharaParser (see Interoperability pilot 4) to extract traits.

Criteria to assess pilots and tasks

Taxonomic scope. We evaluated the suitability of the mark-up tools for ants, spiders, centipedes, fungi, flowering plants, ferns, and mosses. These allowed evaluation across a spectrum of styles and in the context of different codes of nomenclature.

Heterogeneity within sources. The sources included different languages (including Latin) and scripts. The sources included a wide array of styles and level of detail, from those such as contemporary botanical documents which are highly compliant with emerging conventions to others with considerable variation in their approach to descriptions. The scope included variation in referencing materials, specimens, and geoLocation. Older literature opened up issues with inaccurate OCR. The approach had to contend with elements which varied significantly in complexity, the simplest being names, and becoming increasingly complex with literature references, treatments, the way materials were cited and the description and handling of traits (characters).

Granularity: Users were given open opportunity to comment on the schemas in use (see presentations of the authors of the pilots, Fall 2013 pro-iBiosphere meeting Berlin[18], and Appendix on Pilot Studies). In the Requirements Survey19, users were asked to assess whether the cost in human resources of marking up content with greater detail (finer granularity) is justified by the enhanced fitness for purpose compared to less costly mark-up that delivers less detail. This information is summarised in the report D6.1.1 Measuring and Constraining Costs[20], and will be a vital component in the second T6.1 report (D6.1.2.).

Conclusion

We addressed methods and strategies to tackle the interoperability of XML schemas used for biodiversity data. Our insights are based on the results of four interoperability pilots, drawing on content from eight mark-up pilots. We established the feasibility of mark-up using different tools, and of data exchange linking publishers (Pensoft), repositories (Plazi), and dedicated taxonomic platforms (CDM).

In respect of compatibility of XML schemas, there is as yet no universal system (Penev et al. 2011, Sautter et al. 2007). Users will need to assess each schema for coverage and detail in conjunction with the mark-up systems which they need to apply.

TaxPub, GoldenGate XML, and TaxonX have a high degree of similarity, so the development of XLST transformations was relatively straightforward and the interchange of Pensoft's data in TaxPub to Plazi's GoldenGate XML and subsequently TaxonX was minimally lossy.

[18] http://tinyurl.com/q35nz4p
[19] http://tinyurl.com/pn3gfz8
[20] http://tinyurl.com/o57sne7

Completeness (rather than granularity) is a major challenge encountered during the pilots. Most target schemas have mechanisms to allow low level atomisation, however SQL based database approaches such as the CDM may encounter problems with only partly marked-up data, while text-based approaches may have less problems here.

When documents with a coarse mark-up are sent to another system requiring a stricter and more granular schema, a decision must be made to either reject the documents, or remediate them so that the data can be used. The behaviours and design of systems, and the particular uses of schemas here are more significant to achieving lossless and automatic interchange than are the schemas themselves. As a strategy to overcome this impediment, it may be argued that all parties in a community of interest should adopt strict and finely granular schemas and encode their documents to a high degree of granularity to ensure that all applications in the community will be served. However, this approach can be very expensive, as it requires all parties to mark-up to a standard which may be beyond their needs. Effectively it is a transfer of burden and resource expenditure from the party with the most requirements to those with less. A more effective approach is one which does not place a heavy reliance on a schema requirement for interchange, but rather recognises the inevitability of discrepant needs and accordingly considers the design of applications, and the protocols of exchanges as key components of interchange. A system might be, for example, designed to tolerate data which is "lower quality" (from its perspective) and be able to "fail gracefully", using what it can from sub-optimal incoming data, or allowing routines to upgrade it.

On interoperability

Through Interoperability pilots 1, 2 and 3, Pensoft and Plazi demonstrated that interoperable exchange of content from taxon treatments extracted from both legacy and prospectively published literature is achievable. The approach was effective for sources relating to three organismic domains (fungi, plants and animals).

Through the pilot on "Common query/response model for automated registration of higher plants (IPNI), fungi (Index Fungorum, MycoBank) and animals (ZooBank)" Pensoft successfully established a common XML query model for automated registration of nomenclatural acts between publishers and the electronic registers for higher plants (IPNI), fungi (Index Fungorum, MycoBank) and animals (ZooBank). This is available for use by other publishers.

We also successfully transferred content between Plazi and the EDIT Platform for Cybertaxonomy based on transformations between XML-repositories and CDM-stores, thereby preparing for services that give access to Plazi data through an EDIT platform portal.

Finally, we successfully applied the language processing tool CharaParser to extract character-related information from treatments.

While feasibility of the pathways and data exchange has been established, all processes are far from meeting Baumann's definition of interoperability: "...an interoperable text is one that does not require any direct human intervention in order to prepare it to be used by a computer process other

than the one(s) for which it was created". That is, the mark-up used in our pilots fell into Baumann's categories of Negotiated and Blind Interchange.

Our primary conclusion is that improvement of schemas is not the most pressing to improve interoperability; rather our focus should be on the applications, improving them to reduce the demand on expert human resources to achieve successful mark-up. Until that investment is made, 'interoperability' as defined above is not feasible.

Impediments to full automation occur at all levels. They include diversity in source documents because of OCR problems; varying styles of different authors; changes in conventions across time; different conventions; languages or vocabularies for different taxa. Older literature often falls short in providing the data required by current biodiversity applications, as illustrated by the *Chenopodium* pilot. This pilot targeted biogeographic information, but the older literature was imprecise in respect of dates and locations. Consequently, considerable expert effort was needed to provide useful, but still often imprecise, records. Correct page numbering is a particular problem, but given that it used to be available from commercial products (e.g. ABBYY Fine Reader OCR package v8.0), dialogue with them may make this re-available. Use of insufficiently experienced users to carry out mark-up to save costs generates mistakes that have to be corrected later.

Schemas provide a syntax and some semantics in the form of labels for data items. Compatibility among schemas in terms of scope, variables, and units contributes to but does not guarantee "interoperability". Interoperability is much more significantly influenced by the need for human intervention during mark-up. To a certain degree, we can design schemas with interoperability in mind, for example, we can narrow the field of negotiation through improved semantic labels and syntactic structures to represent data in a particular domain. The community will need to improve the definitions of schemas, and build validation tools that will provide feedback to both sources and recipients. However, a shift from blind and negotiated interchange to full interoperability will require a costly investment in applications. Script-based applications for automated mark-up are optimised for application to bodies of literature with a more or less consistent structure. They lack the flexibility to apply to a diversity of sources.

It may be attractive to combine all available tools as a suite of modules, with a shell that first asks users to establish if the source is suited to one of the script-based approaches, and if not, then to employ the more general GoldenGATE. There may be opportunities to develop products in collaboration with commercial organisations, such as developers of OCR software.

We draw attention to the importance of applications because from this project it appears that the cost, in terms of human resources, to prepare sources for use in an OBKMS is much greater than the cost involved in negotiating between different schemas. In the end, these developments may prove to be too costly, and that blind and negotiated procedures may be most appropriate. Progressive mark-up may be an effective strategy to minimising the mark-up costs. Users mark-up minimal elements plus any additional elements that they need for their own purpose. Later users may add to and extend the scope of the mark-up. Each community will need to define what the required threshold of mark-up should be.

While we emphasise the post-publication procedures, we also see a role for changes at the level of authoring, taxonomic data environments and publishing. Catalogues like the World Spider Catalog[21] can be exploited as systems for organising treatments, the fundamental building blocks of taxonomic knowledge. They can evolve to play a greater role in cybertaxonomy, and can do so by collaborating with the literature mark-up initiatives. Given that PDF based documents also present problems, every effort should be made to avoid continuation of current publishing practices, and to promote the publication of semantically enhanced document based on TaxPub JATS. Publications can be further enhanced by the addition of identifier-to-identifier links to other data environments, such as the nomenclatural names registries (see Appendix: Interoperability pilot 2).

Lessons learned from the training

It is important to look into means to collaborate with professional vendors taking over all steps in the conversion process that can be faster processed by non-domain specialists. For the mark-up that clearly requires domain expertise, a GoldenGATE version or equivalent needs to be developed that is as user-friendly as possible, and can be customised for the tasks needed (e.g. mark-up of location, names, traits). The decoupling of OCR from the mark-up was also an important element in the Flora Malesiana project coordinated by Naturalis. The delimitation of tasks between a commercial vendor and domain specialists might furthermore be shifted towards more domain-specific mark-up if a vendor could be trained to recognise some basic domain-specific characteristics like treatment boundaries, taxonomic names or block of references. A strategy to mark-up entire Floras or journals rather than a corpus of articles from various journals covering one specific taxon as has been done in the pilots, might be advantageous (see D3.1.1 [22]).

Achievements

This work has led to the following additions to and recommendations for the future OBKMS.

We have established an Extensible Stylesheet Language Transformations (XSLT) mechanism to: (i) exchange content between TaxPub and TaxonX. We will need to build a library of conversion tools, ideally converting to a single high granularity common format. (ii) inter-convert content from TaxPub and TaxonX. We will need to build a library of interconversion tools, ideally converting to a single high granularity common format.

Pensoft partners have created a common XML query model that fits into the publication workflow, and will automatically register names in nomenclatural registries such as IPNI for plants and ZooBank for animals. The system allows for dialogue with the registries, such that identifiers for nomenclatural acts are automatically included within publications.

We have adapted the DarwinCore-Archive star schema so that it can be used to exchange data on citations of materials and taxonomic treatments. This investment in a GBIF-endorsed environment is

[21] http://research.amnh.org/iz/spiders/catalog/
[22] http://wiki.pro-ibiosphere.eu/wiki/File:Pro-iBiosphere_WP3_MfN_D3.3.1_VFF20122013.pdf

now available for widespread community use. Similarly, we have envisioned application of RDF for modelling and transfer of information in taxonomic treatments. We track and participate in the TDWG RDF Task Group.[23]

Recommendations

We make the following recommendations that will improve machine-based dialogue of semantic enhanced biodiversity information relating to taxonomic treatments:

1. adopt TaxPub JATS for mark-up of new taxonomic literature;
2. adopt TaxonX for mark-up of legacy literature, and invest in the further development that will overcome current limitations;
3. develop instruments for lossless conversion among environments such as TaxPub and TaxonX that will ensure integration of data from historical and recent literature;
4. develop industrial strength digitisation workflows that will format the corpus of legacy literature into useful machine-readable content;
5. endorse and extend DarwinCore-Archive as an interchange format for occurrence data and taxonomic treatments;
6. utilise RDF as a format for distribution of data extracted from the content of marked-up treatments;
7. invest in mark-up tools with improved user interfaces and efficiency so as to achieve high accuracy and reduced labour costs.

[23] https://code.google.com/p/tdwg-rdf/

Appendix: Pilot studies

Reports of the mark-up pilots

Eight mark-up pilot studies were conducted to explore the contribution of taxonomy to the information flow “Publication -> treatment repository -> data mining tool -> new information”. These pilots are summarised in Table 1 and in Y1 resulted in 1 published publication. Additional publications are being prepared.

Formicidae, Hymenoptera (Animals, Ants)

Team: Plazi; State University of Arizona, California Academy of Sciences, Ohio State University.

Goal: Demonstrate a workflow that extracts data including morphological traits from printed record using GoldenGATE and Charaparser, transfer the data to a dedicated database (Hymenoptera Name Server, HNS), edit and add new data including new characters, observation records, digital imagery, edit the marked-up cited publications and upload them to Plazi to provide access to treatments and export the result as TaxPub JATS file for publication in ZooKeys.

Results: All treatments cited in the source publication (Fisher & Smith, 2008) were extracted, mark-up of traits in progress due to heterogeneous morphological descriptions (see Interoperability pilot 3 below). The entire workflow is in place and a publication is planned that includes character data, all names linked to ZooBank and HNS, all treatment citations linked to the cited treatments, and all type material linked through stable HTTP-URIs (see D4.1) to the respective specimens imaged at the California Academy of Sciences. The transfer of extracted character data from CharaParser to the Hymenoptera Online is planned in the second year.

Comments: The very heterogeneous publication structures of how descriptions are provided, from discrete blocks to a mixture of description proper and discussion, to a comparative discussion of characters, make an automated mark-up process very difficult. Different languages complicate matters even more. The process lends itself for work in a particular journal or biota for which the process can be adjusted.

Chenopodium (Plants)

Team: Botanic Garden, Meise; Botanischer Garten und Botanisches Museum, Berlin

Goal: Legacy botanical literature is a large repository of phytogeographic information. Using *Chenopodium vulvaria* (Amaranthaceae) as a model, this pilot project aimed to evaluate the extraction, availability and quality of this information. *C. vulvaria* is a widespread, European and Middle-eastern species. Its unique foul smell makes it unmistakable and it has historically been used as a medicine, though it is more commonly known as a weed of towns and cities. There is a large body of literature on this species covering hundreds of years and in many languages. At the Botanic Garden, Meise, a variety of texts were marked-up using the GoldenGATE editor. These texts spanned

167 years of publishing and were in English, Danish, French, Portuguese, Italian and Spanish. All elements of these treatments were marked-up, but particular attention was paid to the geographic elements. Finished treatments were uploaded to the Plazi repository. At the Botanischer Garten und Botanisches Museum, Berlin, treatments were extracted from the Plazi repository and imported into a CDM datastore using the TaxonX schema. This datastore can be view on the Chenopodium Portal[24].

Results: Heterogeneity in structure and language of texts presented difficulties in accommodating the information within the stricter, modern interpretations required for the TaxonX and the CDM. Mark-up was primarily manual, and semi-automated steps (such as identification of Latin names) needs to be checked and often corrected.

Comments: Legacy literature, especially the older literature, is a valuable source for biogeographic records. Nevertheless, the data are imperfect as they often lack exact dates or location. Adding this information adds an additional significant overhead to mark-up, and we advocate that users assign priority to those works that should be marked-up and georeferenced. Users should also be aware that this information may be present in collection databases such as BRAHMS (a flexible database management system for botanical researchers and herbaria), and may wish to evaluate the cost-effectiveness of discovering and mobilising such information.

Araneae (Animals: Spiders)

Team: Naturalis, Plazi, Naturhistorisches Museum (Bern), University of Bern

Goal: This pilot explored how to visualise the data contained in a taxonomic treatment, and to interlink data elements (text description, literature references, materials citations, and figures) through the framework of the World Spider Catalog[25]. Taxonomic literature is full of references to literature, specimens, figures, tables, etc. The approach adopted here was to convert references into links and thereby improve access from the catalog to treatments, page images, specimen data, and so on to generate a unique information-rich profile for every treatment.

Results: We designed a number of graphs and other tools for visualising materials citations. These include charts to show the number of specimens referenced broken down by gender or institutional collection, by year or time of year, maps for georeferenced records, and country list. A growing number of taxonomic papers have been marked-up with sufficient detail to be able to generate these visualisations. This remains a work in progress and is a priority for 2014. The new database version of the World Spider Catalog will share their identifiers with the Plazi "treatment bank", making it possible to link treatments directly to their citation in the catalog.

Comments: Taxonomic catalogues like the World Spider Catalog can provide the nucleus and framework for mobilising taxonomic knowledge. Taxonomic catalogues will need to extend their activities to if they are to contribute to taxonomic literature mark-up initiatives.

[24] http://dev.e-taxonomy.eu/dataportal/chenopodiumPilot/node/1
[25] http://research.amnh.org/iz/spiders/catalog/

Basidiomycota (Fungi)

Team: Naturalis, The University of Arizona

Goal: The goal was to provide tools that will make comparisons of morphological and ecological data more efficient. We built a workflow that began with scanning published literature, submitting it to character recognition, and marking up to the granularity of morphological characters. In addition, species descriptions already in digital format were added to test character-parsing tools to establish the feasibility of using this approach to build searchable databases of morphological characters.

Results: Descriptions of 15 species have been digitised using GoldenGATE and 33 additional species description from already digital formats were used for database construction by CharaParser.

Comments: The process of scanning, OCR, and mark-up is very time-consuming because of the amount of materials, because of the need to correct errors introduced during the OCR step, and to to produce comparable mark-up schemes among differently structured morphological descriptions. CharaParser is a useful tool but differences in vocabulary used in different publications or for different taxa influence the results. To address this, we created text files with hierarchical structure of the morphological traits as a vocabulary for use by the software.

Nephrolepis (Ferns)

Team: Naturalis

Goal: mark-up one relatively recent article with a very extensive list of scientific names including hundreds of synonyms and combinations, and containing treatments with a wide variety of detail (to accommodate treatments at levels ranging from species to cultivar), to assess the general usability of the tool used for the mark-up and delivery of data to Plazi and from there to CDM and EOL.

Results: Document has been marked-up and uploaded to Plazi and CDM. Mark-up refinements are more or less continuously being made.

Comments: mark-up was a time-consuming process requiring a degree of attentiveness that is difficult to maintain for the time required for many of the tasks. This emphasises the need for further investment in the user interface and usability of GoldenGATE. Further changes that offer users more flexibility in the revision or refinement of existing mark-up would be welcomed.

Additionally, this pilot also demonstrated that much of the time is spent on solving problems due to a lack of guidance for the required mark-up structure, e.g. with regard to nesting of tags or treatments. Attention should be given to stricter specifications of these in the mark-up schema's.

Bryophyta (Mosses)

Team: Naturalis

Goal: Test the integration of taxonomic data from hard copy relating to the Flora of the Guianas and the Flora of Suriname and the contribution this can make to transforming these Floras into dynamic online environments. The three countries of the Guianas - Guyana, Suriname and French Guiana - have more than 80% of their political territories covered by ca. 50 million ha of pristine Amazon forest. It is estimated that there are 15,000 to 18,000 plant species. The Flora of the Guianas Programme was created in the 1980s as a follow-up of Flora of Suriname, to generate and publish accurate taxonomic data of plant species occurring in the Guianas, by collecting, identifying, cataloguing and describing plant specimens. After 30 years of work, the published fascicles of the Flora of the Guianas cover around 25% of the species occurring in the region. To what extent can we increase the rate of species described by recovering information from taxonomic treatments published in Flora of Suriname, in theses, and old publications? Most of this legacy literature is out of print and no digitised text is available. Through a process of scanning, OCR and mark-up of these texts, all basic elements of the taxonomic treatments – species descriptions, nomenclature, collections studied, etc. – can be merged within the new treatments. We tested the feasibility of data integration between the two Floras, by using the treatments of the genus *Campylopus* (Dicranaceae, Bryophyta), published as hard copy, to create a single complete online treatment. We (1) compared the adequacy of two mark-up tools; (2) estimated time required for training and executing each mark-up procedure; (3) estimated time and costs of extending the approach to the complete content of the Flora of the Guianas.

Results: A total of 25 species treatments from two book series have been marked-up using the program GoldenGATE and were uploaded to Plazi depository. The complete procedure took about three work days by an inexperienced user, plus one workday of training time. The material could not be imported into the Platform for Cybertaxonomy (CDM platform) in the current status, because the quality of the mark-up was not adequate. After improvement, another import trial will be carried out (and more results will be added to this section). Currently, the treatments are also being marked-up using custom Perl scripts, which were written initially for Flora Malesiana.

Comments: Digitisation and mark-up procedures are time-consuming; medium-level technical support is required. The mark-up using both tools – GoldenGATE and Perl scripts – has a steep learning curve and becomes cost-effective only if a large body is being marked-up. GoldenGATE seems more effective because it has greater flexibility to manage taxonomic treatments that arise from different sources with different components and formatting, but greater flexibility with this tool would be welcomed. Perl scripts require more learning time and are designed to work with particular formats. This approach works best with consistent text, such as taxonomic treatments from a single Flora or from different Floras with similar formatting. Tasks that need continued effort are marking up the history of names, and linking treatment and bibliographic citations and to digital

treatments and papers (respectively). We envisage problems in importing nomenclatural information into CDM where conflicts with data from newer nomenclatural acts might occur.

Eupolybothrus (Animals, Centipedes)

Team: Pensoft

Goal: To digitise, extract and mark-up all original descriptions and important subsequent treatments of species and subspecies belonging to the centipede genus *Eupolybothrus* (Chilopoda: Lithobiidae) from legacy and prospective literature. There are about 200 major treatments of *Eupolybothrus* related to 104 taxon names, published in 50 different articles between 1847 and 2011. The centipede genus *Eupolybothrus* comprises 25 valid species and 15 doubtful (sub-)species assigned to 7 subgenera ranging from mostly southern Europe, North Africa, the Near and Middle East, including the largest Mediterranean islands of Corsica, Sardinia, Sicily, Crete and Cyprus. The marked-up taxon treatments will be uploaded and made public at Plazi and CDM platforms. A publication in the Biodiversity Data Journal will demonstrate for the first time how a species checklist can be enhanced by linking to the respective digitised species treatments in Plazi. A new species described in the paper will present how a new treatment will be integrated with those from the legacy literature in a single treatment repository (Plazi) and also combined in a contemporary publication.

Results: Within the pilot, a total of 140 taxon treatments have been marked-up by Pensoft, of them 100 uploaded to the Plazi platform. The description of a new species of *Eupolybothrus* - *E. cavernicolus* - was published in BDJ, with rich data accompanying the morphological description (Stoev et al. 2013). This publication is only a test for the pilot and provides links to original descriptions at Plazi of 8 *Eupolybothrus* species. One more publication called "Cybertaxonomic checklist of *Eupolybothrus* will be submitted in April 2014 (see Milestone MS25[26]).

Comments: The study achieved its goals, but revealed that the value of Golden Gate could be enhanced, especially in respect of the user interface. The introduction of 'hot keys', lists of general tags used in the subsequent XML processing that can be saved and not lost after the program is closed, and better recognition of manually entered known tags, such as 'quantity' and 'locality'.

Loranthaceae (Mistletoes, Plants)

Team: RBG Kew.

Goal: To demonstrate the reuse of data extracted from hard copy publications (Floras) by marking up scanned texts. Using a variety of approaches (word processing software (MS-Word), Visual Basic and GoldenGATE), the following examples of reuse will be provided: Presentation of species information via web portals (EDIT CDM,Kew efloras database); Extraction of selected nomenclatural, morphological, geographical information for incorporation into the TRY Plant Traits database; Re-

[26] http://wiki.pro-ibiosphere.eu/wiki/MS25

formatting and addition of new information and media to provide an interactive identification tool; Combine overlapping treatments and to facilitate gap filling for new Flora accounts; Provide access to treatments by uploading them to Plazi. Investigate ways that links between associated organisms might work (hosts, parasites, pollinators, dispersers).

Results: 34 species treatments and 1 genus were marked-up for the mistletoe genus *Helixanthera* (Loranthaceae) from Africa and Asia. The accounts were taken from four different publications (with some taxonomic overlap between): Polhill and Weins “The Mistletoes of Africa” (12 species using “Kew method”, custom XML format), Flora of Tropical East Africa (5 species using “Kew method”, custom XML format), Flora of China (7 species using GoldenGATE, in TaxonX format), Flora Malesiana (11 species from Naturalis Thomas Harmann, custom XML format). All of the mark-up methods used were able to produce a basic level of atomisation equivalent to what might be described as the treatment level (i.e. nomenclature, literature citation, description, geography etc.). Further structuring of the descriptive (morphological) information in the texts was achieved using the Kew method. The resulting schemas for each publication are slightly different; currently only the Flora of China document which is in TaxonX format is suitable for upload to Plazi or to import into the CDM. For the other samples, a small amount of work is required (renaming of elements to the ones GoldenGATE normally uses, to facilitate transformation into TaxonX atomisation of the taxon names extension of the TaxonX generation XSLT to output the "morph" and "char" extensions). The next phase of this pilot will concentrate on atomisation of the descriptive information.

Comments: Our experience with the automated functions of the GoldenGATE software e.g. for marking the bibliography and the production of standard TaxonX output was positive. Tasks which in GoldenGATE require a large amount of user intervention, such as marking the treatment boundaries, we found much easier to do using the methods which utilise the document formatting (e.g. using MS-Word macros). For large documents in particular, a combination of pre-processing to mark-up those elements based on the format, then to import into GoldenGATE in order to run its automated processes and output standardisation, might be more scalable. Extensions to the current GG scripts which could be used to automatically mark-up things like colours, morphology, habitats, (or any custom category if provided with a dictionary of terms) would be really useful.

Reports of the interoperability pilots

Interoperability Pilot 1: Interoperability model between taxon treatments from both legacy and prospective literature from three organismic domains (fungi, plants and animals)

Goal: The goal of this pilot is to demonstrate the interoperability of management of treatments originating from legacy and prospective publications. It calls on materials from mark-up pilots dealing with animals, plants, and fungi.
Interoperability, in this context, is seen as the successful integration of the parts of treatments (such as bibliographic references, names, citations of materials citation) from both legacy and prospective (semantically enhanced) sources, their AUTOMATIC upload to the Plazi repository with the capacity to data in response to queries, whether or not the data are from legacy or prospective sources. capacity to prove responses to query queried, providing requested data irrespective of its origin.

Approach: Plazi, Pensoft and other partners discussed the strategy of representing the data in RDF. Adoption of this approach (see Fig. 4) makes exchange and processing easier, and opens up the content to a large number of tools and subsidiary standards that have been built and adopted by the Semantic Web community over a period of the last dozen years. Plazi is developing an OWL ontology that will make the content usable by behind-the-scene machine reasoners.

Results: We were able to successfully implement cross treatment queries that included legacy and prospective content by region, year, author, taxonomy, etc. (Fig. 5-7). This established the possibility of integration of data originating from different XML schemas and across systems. An exemplar query for the country returns immediately 209 treatments from 31,688 treatments currently residing in the Plazi repository. Compilations of lists like these previously required years of staff effort and did not provide link to the content of the cited treatments. The use of DWC-A presents the opportunity to exchange data among a greater number of partners, such as GBIF.

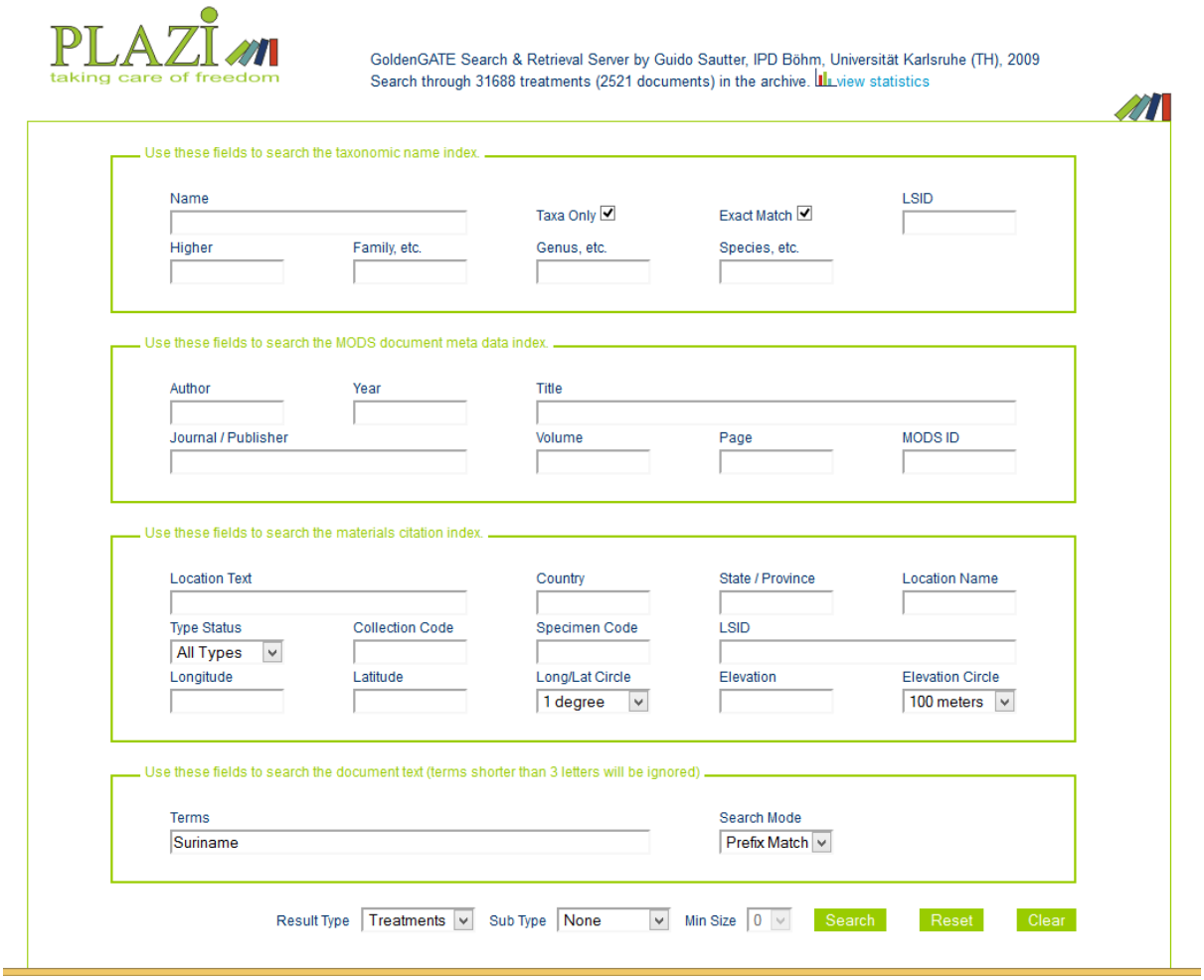

Figure 5. Plazi search interface. Search example for Terms =”Suriname”

Figure 6. Plazi Search return for Terms="Suriname". 209 treatments are returned including plants (e.g. *Solanum*), animals (*Anochetus*), mosses (*Compylopus*) and ferns (*Nephrolepis*) used in the pilots. All the colored terms are hyperlinked and all the treatments can be viewed (Fig. 7)

Figure 7. Treatment for *Nephrolepis rivularis* linked from Fig. 6.

Interoperability Pilot 2: Common query/response model for automated registration of higher plants (IPNI), fungi (Index Fungorum, MycoBank) and animals (ZooBank)

Goal: To achieve an automated workflow that enriches the publication workflow so that new nomenclatural acts in publications can be transmitted to nomenclatural registries; with sufficient dialog for the publication process to acquire identifiers for inclusion in the paper.

Introduction: Registration is expected to emerge as the default mechanism by which nomenclatural practices exploit the digital media to register authors, works, and nomenclatural acts (Knapp et al. 2011). Different ways for registration have been reviewed by Pyle and Michel (2008) and Morris et al. (2011). Implementation of automated registration of nomenclatural acts that comply with the International Code of Zoological Nomenclature (ICZN 2012) and the International Code for Nomenclature of Algae, Fungi and Plants (McNeill et al. 2012) has been pursued in the last few years (Penev et al., in press).

Registration is extremely time-consuming process if done "by hand", especially if required as part of "turbo-taxonomic" publication which targets accelerated descriptions that combine molecular data, morphological descriptions, and digital imaging (Butcher et al. 2012, Riedel et al. 2013). The numbers of new taxa described in such papers may be high; for example Butcher et al. (Butcher et al. 2012) described 178 new species of parasitic wasps and Riedel et al. (Riedel et al. 2013) described 101 new species of *Trigonopterus* weevils. The record is held by Marsh et al. (2013) who described 277 new braconid wasps from Costa Rica. This paper is remarkable also because it became the first "turbo-taxonomic" paper where all 277 new species were registered in ZooBank automatically in just a few seconds, saving a great deal of time to the authors, publisher and the registry. If papers such as this have to use traditional means to register new species, then it will slow down the speed of discovery of our biodiversity and the nomenclatural work may simply not be completed.

Automated registration accelerates the taxonomic workflow, reduces opportunities for errors, clarifies the distinction between dates of acceptance and publication of a manuscript, efficient and accurate validation of final published data and metadata by automated communication from the publisher to the registry on the day of publication.

The registration workflow: The registration of nomenclatural acts and quality control should be a responsibility of publishers and registry curators and, to a lesser extent, of authors. Registration of a nomenclatural act could be initiated by an author, but a publisher-initiated model avoids registry curators finding, vetting data for compliance with the appropriate code, and entering it. "Journal-centric" registration has already been implemented in the Pensoft's journals ZooKeys[27] and PhytoKeys[28]. The model presented below can be adapted for author initiation, though we envisage that there would be a greater curatorial overhead and a greater likelihood of errors being created.

[27] http://www.pensoft.net/journals/zookeys
[28] http://www.pensoft.net/journals/phytokeys

In the "journal-centric" model, the registration of taxonomic and nomenclatural acts involves two main classes of actors: (1) publishers, and (2) registry curators. The publisher initiates the registration process with the following workflow to achieve interoperability of multiple environments (see also Fig. 8):

Step 1. XML message sent from the publisher to the registry on acceptance of the manuscript containing the type of act, taxon names, and preliminary bibliographic metadata; the registry will store the data *but not make these publicly available before the final publication date*;
Step 2a. Response by the registry of an XML report containing the unique identifier(s) of the act(s) and/or any relevant error messages;
Step 2b. Error correction and de-duplication performed manually: human intervention, at registry or publisher side (or at both);
Step 3. Inclusion of registry supplied identifiers in the published treatments;
Step 4. Making the information in the registry publicly accessible upon publication, providing a link from the registry record to the article.

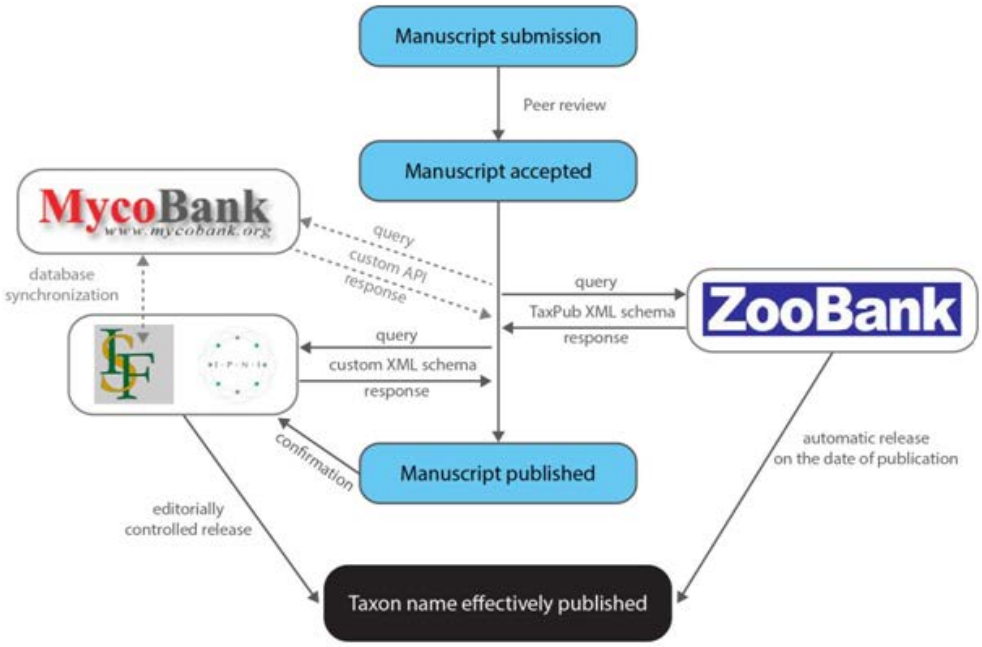

Figure 8. Automated registration process and validation of finally published data and metadata between publisher and registry. Abbreviations: IPNI - International Plant Name Index, IF - Index Fungorum.

Results:

Automated registration with IPNI. The pre-publication registration of new plant taxa and nomenclatural acts in IPNI and inclusion of the IPNI identifiers in the protologues was first tried with the journal PhytoKeys since publication of its first issue in 2010 (Penev et al. 2010). With the pro-iBiosphere-project the workflow has been refined to include an automated registration module. The

pilot project uses a custom XML format with a proof-of-concept use for the description of the new genus *Lettowia* description and a new combination *Lettowia nyassae* (Oliv.) H. Rob., comb. nov. in the paper of Robinson and Skvarla (Robinson and Skvarla 2013). The emphasis of the pilot was to understand the workflow. As this is scaled up to production use with a broader range of partners, IPNI will move to use the Taxon Concept Schema standard to encode the data exchanged. This will enable broader adoption. The XML query is submitted to IPNI's Application Programming Interface (API) through a POST request and receives responses with automatically inserted IPNI registration identifiers.

Automated registration with Index Fungorum. The registration workflow of Index Fungorum (IF) will adopt that of IPNI as the IF system moves to Royal Botanical Gardens Kew.

Automated registration with MycoBank. The services: 1 'SearchMycoBankWithFilters', 2 'InsertUserProfile', 3 'UpdateUserProfile', 4 'InsertMycobankRecord', and 5 'UpdateMycobankRecord' of the MycoBank API are called on for implementation. Using combinations of (1, 2, 3) and (1, 4, 5) one can implement the Upsert (Update if exists, Insert otherwise) semantics required for the Common query/response registration model. As there are multiple registries for fungi (MycoBank, IndexFungorum, Fungal Names), another approach would be to perform the registration with only one of them and rely on their synchronisation mechanisms currently being built to propagate the information to the other databases.

Automated registration with ZooBank. ZooKeys was the first journal that implemented a mandatory registration of new taxon names in zoology from publication of its first issue in 2008 (Penev et al. 2008). The automated registration with ZooBank (Polaszek et al. 2005) is based on a slightly different approach than that with IPNI and uses the TaxPub XML schema (Catapano, 2010) as a basic standard. Upon acceptance, the XML version of the manuscript is uploaded to the ZooBank server through the ZooBank's interface. Then a software tool at ZooBank harvests the TaxPub XML and registers the title, authors and new taxon names. The tool also checks if some or all authors have been previously registered and inserts their current (or newly registered) ZooBank UUIDs. The ZooBank database offers editorial intervention if there are ambiguities s to the identity of the author. The revised TaxPub XML is sent back to the publisher with inserted UUIDs for the article, authors and new names. If the manuscript XML has been changed after registration, it can be uploaded again and the new data will replace the previous ones. At the day of publication, the names and the bibliographic metadata are made publicly available in ZooBank.

Supplementary file 1[29]. XML query sent from Pensoft to IPNI on the day of acceptance of the manuscript for publication [exemplified with the paper of Robinson and Skvarla (2013)].

Supplementary file 2[30]*.* XML response of IPNI to the query in Appendix 1. The response is sent back to Pensoft and contains the registration numbers of the new genus name and the new combination [exemplified with the paper of Robinson and Skvarla (2013)].

[29] http://wiki.pro-ibiosphere.eu/w/media/7/71/IPNI_query.xml
[30] http://wiki.pro-ibiosphere.eu/w/media/d/de/IPNI_response.xml

Supplementary file 3[31]. TaxPub XML of a ready-to-publish manuscript submitted from Pensoft to ZooBank [exemplified with the paper of Morffe and Rodríguez (2013)].

Supplementary file 4[32]. TaxPub XML returned from ZooBank to Pensoft containing UUIDs of the article, authors and new taxon names [exemplified with the paper of Morffe and Rodríguez (2013)].

Comments: The automated registration of content as part of a publication workflow has been implemented, and can be adopted by other publishers. The registration workflow established here is free for use by any publisher who would like to implement it. To ensure broader adoption of the registration model, the data exchanged through the workflow should be encoded in a standard. At this time, we recommend that *zoology* journals adopt the TaxPub XML schema (Catapano 2010; open source available at: TaxPub[33]) which encodes publications as required by the zoological code. For plants and fungi, the registration workflow will implement the XML Taxonomic Concept Schema (TCS)[34] which encodes names. The supplementary files 1-4 show some data encoded for the pilot project using a custom XML format – whilst this shows the kind of data that will be exchanged, it should not be used as a template – the TCS and TaxPub standards should be used as reference. Once the editorial workflow is defined, and structured data can be produced according to these standards, journal editors should contact registries for access to their APIs.

Interoperability Pilot 3: Interoperability model between Plazi and the EDIT Platform for Cybertaxonomy based on transformations between XML-repositories and CDM-stores

This has been reported independently as report D4.1 (Strategies for improved cooperation and interoperability between infrastructures)[35]. The executive summary of D4.1 is as follows.

The pro-iBiosphere project recognises that there are many challenges to effective interoperability among biodiversity-related infrastructures. In order to avoid duplication of effort with other ongoing initiatives (e.g. TDWG Biodiversity Information standards and Lifewatch), pro-iBiosphere has decided to focus on two main topics of primary importance, which have the potential to significantly improve interoperability between e-infrastructures, in line with the recommendations of the "Bouchout Declaration" drafted in pro-iBiosphere D2.1.2, "Towards a draft strategy for increased cooperation'. These topics are:

- stable identifiers for biodiversity collection objects;
- a global registry for biodiversity-related services.

[31] http://wiki.pro-ibiosphere.eu/w/media/0/00/ZooBank_query.xml
[32] http://wiki.pro-ibiosphere.eu/w/media/4/40/ZooBank_response.xml
[33] https://github.com/tcatapano/TaxPub/releases/tag/v0.5-beta
[34] http://www.tdwg.org/standards/117/
[35] http://wiki.pro-ibiosphere.eu/w/media/1/13/Pro-iBiosphere_WP4_FUB-BGBM_VFF_20122013.pdf

Stable identifiers: In cooperation with the Information Science and Technology Committee of the Consortium of European Taxonomic Facilities (CETAF-ISTC), a Linked (open) Data compliant system for using HTTP-URIs as stable identifiers for collection objects and their associated metadata has been defined and documented by pro-iBiosphere: see Best Practice Guidelines[36].

The project piloted a working identifier system that we expected to be adopted by the wider community. At present, the following institutions are committed to implementing the system:

1. Botanischer Garten and Botanisches Museum Berlin-Dahlem, Germany;
2. Harvard University Herbaria;
3. Harvard Museum of Comparative Zoology, US;
4. Muséum national d'histoire Naturelle, Paris, France;
5. Museum fur Naturkunde, Berlin, Germany;
6. National Botanic Garden Belgium, Belgium;
7. Naturalis Biodiversity Center, the Netherlands;
8. Royal Botanic Gardens Kew, UK;
9. Royal Botanic Gardens Edinburgh, UK.

Biodiversity Service Registry: The evolving biodiversity informatics infrastructures being assembled around the world are un-coordinated and web services are often poorly documented. A global registry of biodiversity-related services will help potential users discover and understand service functions, interfaces, and behaviour; and will foster the growth of service-based applications and workflows. The 7th Framework project BioVel has implemented the Biodiversity Catalogue to meet this need. pro-iBiosphere has made recommendations for enhancements that will secure the registry of biodiversity-related web services at the heart of the future OBKMS.

We explored interoperability between biodiversity-related e-infrastructures with the development of a pilot workflow between the document-centric Plazi system and the CDM-based EDIT Platform for Cybertaxonomy. This exercise has demonstrated the feasibility of the concept and has improved both technical and "human" interoperability. The pilot elaborated an evolving workflow through cooperation between taxonomists, biologists, bioinformaticians and computer scientists. This workflow will be used not only for mark-up of legacy data from historical literature, but more important, to database the marked-up text so that legacy data joins other data within a common data model platform, to be used for the generation and mobilisation of new knowledge.

Interoperability Pilot 4: Revision of a tool (CharaParser) that generates identification keys by reusing morphological characters from published species descriptions

[36] http://wiki.pro-ibiosphere.eu/wiki/Best_practices_for_stable_URIs

Goal: Upgrade CharaParser to extract morphological descriptive data, locality and bibliographic citations as a starting point to automatically generate identification keys.

Software Description: CharaParser is a semantic annotation software specifically designed to parse semi-structured, morphological descriptions of organisms written in English. It aims to turn text descriptions to computable trait data such as taxon-character matrices, The data can then be used to facilitate taxon concept analysis or build phylogenies. The software is developed by Hong Cui's research team at University of Arizona.

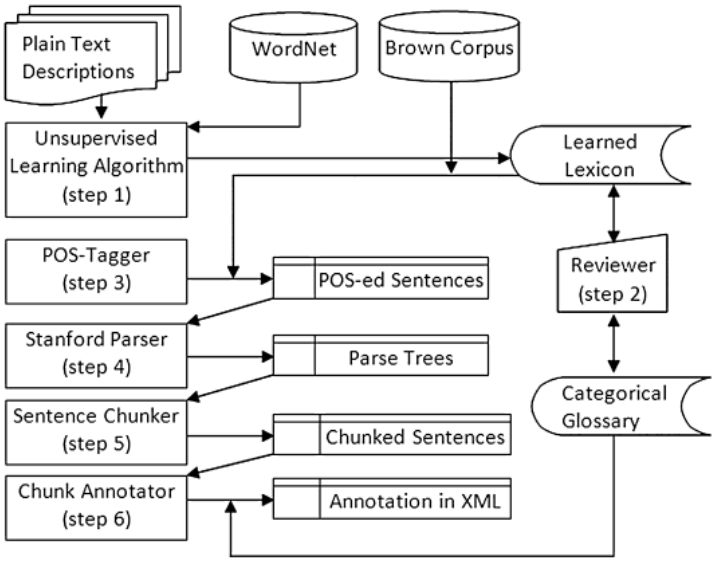

Figure 9. CharaParser Architecture

CharaParser has a six-step annotation process (Fig. 9). Starting with plain-text descriptions, Step 1 uses an unsupervised learning algorithm (Cui, Boufford & Selden, 2010) to create (a) a set of domain terms that are structure names (e.g. "stems") and have POS (Part of Speech) tags of NNS (singular noun), NNP (plural noun), or NN (noun); and (b) a set of domain terms that are character states (i.e. character values), for example, "ovate" is a state/value for character "shape", and terms like "ovate" will have a POS tag of JJ (adjective). These two sets of domain terms form a domain lexicon containing domain terms and their POS tags. The domain lexicon (i.e. outputs a and b) are then reviewed, corrected, and enriched by a user (Step 2). Using the domain lexicon, CharaParser tags a sentence with POS tags before it is passed to Stanford Parser (Step 3). The domain lexicon can also be added to an optional categorical glossary to expand it, which in turn can be used by CharaParser to improve parsing. Because the unsupervised algorithm learns only domain nouns and adjectives, common English nouns, adjectives, and all other word groups are left to Stanford Parser to assign POS tags. After sentences are POS-tagged, CharaParser feeds them to the Stanford Parser and obtains parse trees from it. Each parse tree corresponds to a sentence (Step 4). Next, CharaParser chunks parsed sentences to separate semantic chunks of various types: structure-related, character-related, relation-related, and syntactic-based (Step 5). Chunked sentences are read from left to right and each chunk is subsequently processed according to its type to produce the final annotation

results (Step 6). The character-related chunks are associated with appropriate structure-related chunks and final XML output is generated according to the schema published[37].

The input CharaParser requires are XML files with taxon names and morphological description paragraphs marked-up, such as TaxonX files output by GoldenGATE. The XML output of CharaParser can be converted to other standard arrangements, for example, taxon-character matrices. A prototype software tool, called Matrix Generation, takes CharaParser's XML output and outputs taxon-character matrices (Fig. 7).

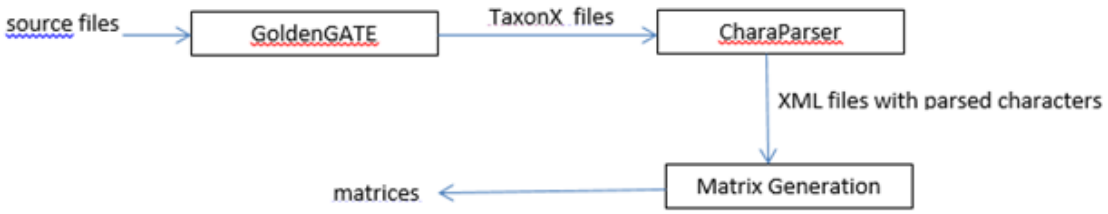

Figure 10. Pipeline used in the pilot studies

Comments:

Ant descriptions. Cui received seven publications marked-up in TaxonX format on ants. There were many OCR errors and inaccurate mark-ups. The OCR errors originated from the conversion from source PDF files to TaxonX files. The inaccurate mark-ups include non-morphological descriptions marked-up as morphological descriptions. Cui used CharaParser to process all input files and generated matrices. The results showed, not surprisingly, that non-description sentences were not parsed correctly, while description sentences were. These results are pending review by ant experts to further review.

Fungi. Cui received one file from a mycologist and used CharaParser to process the file. The biologist reviewed the results and provided feedback. Based on the feedback, CharaParser has been enhanced so the output meets the requirements of the mycologist. The feedback helped to improve CharaParser in general. This pilot was used to start the integration of the ontological component into CharaParser. More Fungi descriptions are required to further test the robustness of the system.

Conclusions. CharaParser uses unsupervised learning to adapt Stanford Parser to parse semi-structured descriptive sentences. The user can interact with CharaParser and review the terms learned and by adding the terms CharaParser failed to learn. As a consequence, the users of CharaParser enrich domain-specific glossaries and that could be used to create domain ontologies. The time and effort users spend on parsing descriptions can be seen as investment of long term value and of benefit to others. The approach shows considerable promise, but further development is needed:

[37] http://biosemantics.googlecode.com/svn/trunk/characterStatements/CharaParserSchema.xsd

1. We need better systems to generate clean input for CharaParser.
2. Integration of ontologies into CharaParser parsing process (ontologies enhance information, so that, e.g. the use of the term ‘surface to refer to “Cap of mushroom”and “stem of mushroom”can be distinguished and made semantically more meaningful.)
3. Introduction of quality control at each step to ensure GoldenGATE, CharaParser, and Matrix Generation tools produce what is required, and for corrections to be made. A tool that helps the user to review CharaParser’s results is now being developed.
4. We need to train the users on using these tools. We strive to make the tools as user-friendly as possible, but they are highly specialised tools that requires training to use them effectively.

References

Agosti D et al. (2013) D2.1.1 Report on ongoing biodiversity related projects, current e-infrastructures and standards. http://www.pro-ibiosphere.eu/getatt.php?filename=pro-iBiosphere_WP2_PLAZI_D2.1.1_VFF_30062013_4303.PDF

Baker E, Rycroft S, Smith V (2014) Linking multiple biodiversity informatics platforms with Darwin Core-Archives. Biodiversity Data Journal 2: e1039. doi: 10.3897/BDJ.2.e1039.

Bauman S (2011) Interchange vs. Interoperability. In: *Proceedings of Balisage:* The Mark-up Conference Montréal, Canada, August 2 - 5, 2011. Balisage Series on Mark-up Technologies, vol. 7 doi: 10.4242/BalisageVol7.Bauman01.

Butcher B, Smith M, Sharkey M, Quicke D (2012) A turbo-taxonomic study of *Thai Aleiodes* (*Aleiodes*) and *Aleiodes* (*Arcaleiodes*) (Hymenoptera: Braconidae: Rogadinae) based largely on COI barcoded specimens, with rapid descriptions of 179 new species. Zootaxa 3457: 1-232.

Catapano T (2010) TaxPub: An extension of the NLM/NCBI Journal Publishing DTD for taxonomic descriptions. Proceedings of the Journal Article Tag Suite Conference 2010. http://www.ncbi.nlm.nih.gov/books/NBK47081/#ref2

Cui H, Boufford DE, Selden P (2010) Semantic annotation of biosystematics literature without training examples. Journal of the American Society for Information Science and Technology 61: 522–542. doi: 10.1002/asi.21246.

Hamann TD, Müller A, Roos MC, Sosef M, Smets E (in press) Detailed mark-up of semi-monographic legacy taxonomic works using FlorML. Taxon.

International Commission of Zoological Nomenclature (ICZN) (2012) Amendment of Articles 8, 9, 10, 21 and 78 of the *International Code of Zoological Nomenclature* to expand and refine methods of publication. ZooKeys 219: 1-10. doi: 10.3897/zookeys.219.3944

Kirkup D, Malcolm P, Christian G, Paton A (2005) Towards a digital African Flora. Taxon 54: 457-466.

Knapp S, McNeill J, Turland N (2011) Changes to publication requirements made at the XVIII International Botanical Congress in Melbourne - what does e-publication mean for you? PhytoKeys 6: 5-11. DOI: 10.3897/phytokeys.6.1960

Marsh P, Wild A, Whitfield J (2013) The Doryctinae (Braconidae) of Costa Rica: genera and species of the tribe Heterospilini. ZooKeys 347: 1-474. DOI: 10.3897/zookeys.347.6002

McNeill J, Barrie FR, Buck WR et al., eds. (2012) International Code of Nomenclature for algae, fungi, and plants (Melbourne Code), Adopted by the Eighteenth International Botanical Congress Melbourne, Australia, July 2011 (electronic edition). International Association for Plant Taxonomy.

Miller JS, Funk VA, Wagner WL, Barrie F, Hoch PC, Herendeen P (2011) Outcomes of the 2011 Botanical Nomenclature Section at the XVIII International Botanical Congress. PhytoKeys 5: 1-3. doi: 10.3897/phytokeys.5.1850

Morffe J, Rodríguez N (2013) *Batwanema* gen. n. and *Chokwenema* gen. n. (Oxyurida, Hystrignathidae), new nematode genera as parasites of Passalidae (Coleoptera) from the Democratic Republic of Congo. ZooKeys 361: 1-13. doi: 10.3897/zookeys.361.6351

Morris PJ, Macklin JA, Croft J, Nicolson N, Whitbread G (2011) Fungal nomenclature 4. Letter of concern regarding Props.(117119) to amend the ICBN to require pre-publication deposit of nomenclatural information. Mycotaxon, 116(1), 513-517. doi: 10.5248/116.513

Patterson DJ, Cooper J, Kirk PM, Pyle RL, Remsen DP (2010). Names are key to the big new biology. Trends in Ecology & Evolution, 25(12): 686-691. doi: 10.1016/j.tree.2010.09.004

Penev L, Erwin T, Thompson FC, Sues H-D, Engel MS, Agosti D, Pyle R, Ivie M, Assmann T, Henry T, Miller J, Ananjeva NB, Casale A, Lourenzo W, Golovatch S, Fagerholm H-P, Taiti S, Alonso-Zarazaga M (2008) ZooKeys, unlocking Earth's incredible biodiversity and building a sustainable bridge into the public domain: From "print-based" to "web-based" taxonomy, systematics, and natural history. ZooKeys Editorial Opening Paper. ZooKeys 1: 1-7. doi: 10.3897/zookeys.1.11

Penev L, Kress WJ, Knapp S, Li DZ, Renner S (2010) Fast, linked, and open – the future of taxonomic publishing for plants: launching the journal PhytoKeys. PhytoKeys 1: 1-14. doi: 10.3897/phytokeys.1.642

Penev L, Lyal C, Weitzman A, Morse D, King D, Sautter G, Georgiev T, Morris R, Catapano T, Agosti D (2011) XML schemas and mark-up practices of taxonomic literature. ZooKeys 150: 89-116. doi: 10.3897/zookeys.150.2213

Penev L, Catapano T, Agosti D, Sautter G, Stoev P (2012) Implementation of TaxPub, an NLM DTD extension for domain-specific mark-up in taxonomy, from the experience of a biodiversity publisher. In: Journal Article Tag Suite Conference (JATS-Con) Proceedings 2012 [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2012. Available from: http://www.ncbi.nlm.nih.gov/books/NBK100351/

Penev L, Paton A, Nicolson N, Kirk P, Pyle R, Whitton R, Georgiev T, Barker C, Hopkins C, Robert V, Biserkov J, Stoev P (in press) A common registration–to–publication automated pipeline for nomenclatural acts for higher plants (International Plant Names Index, IPNI), fungi (Index Fungorum, MycoBank) and animals (ZooBank). ZooKeys.

Polaszek A, Alonso-Zarazaga M, Bouchet P, Brothers DJ, Evenhuis N, Krell F-T, Lyal CHC, Minelli A, Pyle RL, Robinson NJ, Thompson FC, van Tol J (2005) ZooBank: the open-access register for zoological taxonomy: Technical Discussion Paper. Bulletin of Zoological Nomenclature 62(4): 210-220.

Pyle RL, Michel E (2008) ZooBank: Developing a nomenclatural tool for unifying 250 years of biological information. Pp. 39-50. In: Minelli, A., Bonato, L. & Fusco, G. (eds.) Updating the Linnaean Heritage: Names as Tools for Thinking about Animals and Plants. Zootaxa 1950: 1-163.

Riedel A, Sagata K, Surbakti S, Tänzler R, Balke M (2013) One hundred and one new species of *Trigonopterus* weevils from New Guinea. ZooKeys 280: 1-150. doi: 10.3897/zookeys.280.3906

Remsen D,Knapp S, Georgiev T, Stoev P, Penev L (2012) From text to structured data: Converting a word-processed floristic checklist into DarwinCore-Archive format. PhytoKeys 9: 1. DOI: 10.3897/phytokeys.9.2770

Robinson H, Skvarla J (2013) *Lettowia*, a new genus of Vernonieae from East Africa (Asteraceae). PhytoKeys 25: 47-53. doi: 10.3897/phytokeys.25.5556

Sautter, G. Böhm, K. and Agosti, D. (2007a) A quantitative comparison of xml schemas for taxonomic publications. Biodiversity Informatics, 4: 1-13

Sautter G, Agosti D, Böhm K (2007b) Semi-Automated XML mark-up of Biosystematics Legacy Literature with the GoldenGATE Editor. Proceedings of PSB 2007, Wailea, HI, USA, 2007. http://psb.stanford.edu/psb-online/proceedings/psb07/sautter.PDF

Stoev P, Komerički A, Akkari N, Liu S, Zhou X, Weigand A, Hostens J, Hunter C, Edmunds S, Porco D, Zapparoli M, Georgiev T, Mietchen D, Roberts D, Faulwetter S, Smith V, Penev L (2013) *Eupolybothrus cavernicolus* Komerički & Stoev sp. n. (Chilopoda: Lithobiomorpha: Lithobiidae): the first eukaryotic species description combining transcriptomic, DNA barcoding and micro-CT imaging data. – Biodiversity Data Journal, 1: e1013. doi: 10.3897/BDJ.1.e1013

Thessen, AE, and Patterson DJ 2011. Data issues in the life sciences. ZooKeys 150: 15–51. doi: 10.3897/zookeys.150.1766

Wieczorek J, Bánki O, Blum S, Deck J, Döring M, Dröge G, Endresen D, Goldstein P, Leary P, Krishtalka L, Otuama É, Robbins R, Robertson T, Yilmaz P (in press) Meeting Report: GBIF hackathon-workshop on Darwin Core and sample data (22-24 May 2013) http://www.gbif.org/resources/2245