SEVENTH FRAMEWORK PROGRAMME

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)**

D3.3.2. Report on progress during the coordination process of partners and non consortium partners

Deliverable Status:	**Draft**
File Name:	**pro-iBiosphere_WP3_MfN_D3.3.2_VFF30042014.pdf**
Due Date:	**April 2014 (M20)**
Submission Date:	**30 April 2013 (M20)**
Dissemination Level:	**Public**
Authors:	**MfN (Daniel Mietchen)**

Consisting of:

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

REVISION CONTROL

Version	Author	Date	Status
0.1	Daniel Mietchen (MfN)	11.02.2014	Initial draft
0.5	Daniel Mietchen (MfN)	21.04.2014	Draft
0.6	Sabrina Eckert (FUB-BGBM)	22.04.2014	Draft
0.7	Soraya Sierra (Naturalis), Scott Edmunds (BGI), David King (Open University)	23.04.2014	Draft
0.8	Gregor Hagedorn (MfN), Anton Güntsch, Andreas Müller (both FUB-BGBM), Quentin Groom (NBM)	24.04.2014	Draft
	David Patterson (Plazi)	26.04.2014	Comments
0.9	Lyubomir Penev (Pensoft), Chris Maloney (A-Tek/ NCBI), Terry Catapano (Plazi)	26.04.2014	Comments, additions
1.0	Daniel Mietchen (MfN) Gregor Hagedorn (MfN)	30.04.2014	Final

Table of Contents

Executive summary

A pro-iBiosphere workshop on markup of biodiversity literature was held in February 2014 (MS12) to discuss and coordinate with partners and non-consortium partners the implications of the report D3.3.1 on "Semantic integration of the biodiversity literature" (submitted in December 2013 to the EC). The aim of the workshop was to build consensus on the recommendations and on the interoperability of markup approaches used for legacy and recent biodiversity literature across organismic domains. The present deliverable summarizes the outcomes of the workshop and builds on D3.3.1 to review use cases for markup of the biodiversity literature; to consider alternative approaches to semantic enrichment and to generate an initial work plan and roadmap for the semantic integration of biodiversity literature. The workshop also inspired several of the projects tackled at the Biodiversity Data Enrichment Hackathon co-organized by pro-iBiosphere and Naturalis in March.

The main recommendations are that the biodiversity informatics community should concentrate on:

- semantic enrichment of selected revisionary works, capturing the most recent knowledge efficiently, rather than the taxonomic literature as a whole;
- differentiating between the different use cases for semantic markup and annotations and choosing appropriate resource investments;
- investing into the re-invention of publishing workflows, such that they automatically produce semantically integrated publication forms;
- working towards an integrative and semantically enabled Open Knowledge System that integrates well with past knowledge, while allowing for new agile and collaborative approaches to publishing and curating, possibly building on the Wikidata model.

Introduction

Long before the concept of "markup" and digital text management was developed, we structured texts in a variety of ways (including elements now considered markup, like italics, paragraphs, chapters, or footnotes; as well as creating sections to documents, or distinguished taxonomic treatments). Similarly, we referenced and linked objects or texts in more or less formal ways (e.g. scientific taxon names or citations of other written works). Markup is a further formalization of these practices. The term refers to computer-readable structuring of texts and cross-referencing between texts or between texts and objects (e.g. images and other media, or geographic points associated with occurrence records). Beyond that, markup can be used at levels of granularity and with semantic precision that is impossible in traditional printed texts.

This document explores the use cases of, and a vision for, semantic annotation and integration of biological publications. Semantic annotation is the practice of associating statements as to the meaning of a piece of text or other information in the context of a document. Semantic integration is the process of combining information from various sources on a given topic, so as to help organize, discover and retrieve information related to that topic. The foundation for scientific knowledge management is the sharing of knowledge produced by scientific investigations through the scientific literature - that is journal articles, monographs, conference proceedings - and information systems (such as software, databases, and knowledge bases).

Among the long-term goals of knowledge management are to accelerate the work, to increase the amount of content, to monitor and enhance the quality of content, and that machines assist these processes significantly, to the point that some kinds of information about the topic at hand can be inferred automatically on the basis of known relationships to other pieces of knowledge.

Semantic enrichment through markup is one way to approach automation, and several approaches to marking up the biodiversity literature have been discussed in report D3.3.1 (Semantic integration of the biodiversity literature)[1]: (1) fully automated natural language processing, (2) basic markup complemented by automated processing and specialist correction, (3) social crowdsourcing models, and (4) pre-publication markup.

Given that the vast majority of the biodiversity literature has so far not been marked up in any of these ways, questions arise as to

- whether and how to close this gap,
- for what purposes that would make a difference,
- what the demand is, and
- how it can be met.

[1] http://wiki.pro-ibiosphere.eu/w/media/2/2a/Pro-iBiosphere_WP3_MfN_D3.3.1_VFF20122013.pdf

To explore these questions further, a workshop[2] on markup of biodiversity literature (MS12) was organized at the Museum für Naturkunde (MfN) in Berlin on February 10-11, 2014, in conjunction with a pro-iBiosphere workshop on alternative business models (MS23) and the fifth management meeting of the pro-iBiosphere task leaders.

The workshop brought together a variety of producers and users of markup. These included the pro-iBiosphere partners Naturalis, Plazi, Pensoft, BGBM, RBGK and BGM as well as

- the Biodiversity Heritage Library (BHL),
- PubMed Central (PMC),
- University of Glasgow, and
- independent consultants.

A total of 21 participants attended the workshop. For the complete list, see the workshop page on the pro-iBiosphere wiki.

Central to the discussions during the workshop was the identification of use cases for the markup of biodiversity literature. The use cases related to:

- Data integration across multiple sources
- Repurposing and redistribution
- Metadata extraction and processing
- Facilitating data mining
- Disambiguation and machine reasoning

In the following, we will briefly discuss each of these use cases and illustrate them with examples from the biodiversity literature that are drawn from presentations given during the workshop. Presentations are available from the workshop's page on the pro-iBiosphere wiki.

Use cases for markup

Data integration across multiple sources

The largest platform for biodiversity literature is the Biodiversity Heritage Library[3] and its global network of partners (cf. D3.3.1). They compile content from a multitude of sources. Initially, markup was not a priority for BHL, and even achieving a machine readable identification of articles within journals is an ongoing effort. However, taxon names within the text are automatically identified, and tools developed by uBio[4] and Global Names[5] call upon the names found to create taxonomic indexes. Known problems are spelling errors, inconsistent spellings of technical names and terms, and Optical Character Recognition (OCR) errors.

[2] http://wiki.pro-ibiosphere.eu/wiki/MS12_-_Workshop_on_mark-up_of_biodiversity_literature
[3] http://www.biodiversitylibrary.org/
[4] http://www.ubio.org/
[5] http://www.globalnames.org/

BHL's legacy and specialist documents are a particular challenge for OCR engines, because the original print quality is often poor, and because the engines use dictionaries that don't include technical vocabulary and abbreviations.

In recent months, the focus of BHL activities has moved closer to markup, with gaming,[6] social media,[7] hackathons[8] and code sprints[9] all being explored, for example improving text accuracy of OCR output, crowdsourcing metadata, and building semantic search tools. Some of these topics were tackled during the Biodiversity Data Enrichment Hackathon[10] in March 2014 at Naturalis, e.g. a tool chain has been produced that would allow BHL users to correct OCR output while reading BHL materials. [11]

A second approach to integrate multiple data sources is to use a standard interchange format, into which the files from various sources can be converted for aggregation. An example of this is the Journal Article Tag Suite (JATS)[12] that is used to aggregate articles in PubMed Central (PMC), the largest full-text repository of biomedical literature.

As mentioned in D3.3.1, a taxonomy-specific extension to JATS has been developed (TaxPub)[13], which allows, for instance, to mark up taxon names. TaxPub was designed for prospective markup of contemporary manuscripts, and while an integration of journal articles from BHL into JATS is feasible in principle, a TaxPub version suitable for legacy content - e.g. based on the Archiving tag set of JATS - does not exist yet. Content from BHL books could in principle be integrated using the Book Interchange Tag Suite (BITS)[14], a JATS extension for books.

On that basis, it would be technically possible to integrate a suitably licensed subset of BHL into PMC. However, no steps have been taken in this direction, nor are any anticipated, since the two platforms overlap only partially in terms of their missions, their user communities and their technical infrastructure.

Another use case for aggregated data from a diverse set of sources is the World Flora Online initiative[15]. It is the first of 16 targets of the UN Global Strategy for Plant Conservation[16] 2011-2020. World Flora Online requires input from a diverse set of sources. Mobilizing literature data from existing floras and bringing them into a format that is consumable will play an important role. This includes all data

[6] http://biodivlib.wikispaces.com/Purposeful+Gaming
[7] http://blog.biodiversitylibrary.org/2014/03/first-meeting-of-mining-biodiversity.html
[8] https://www.idigbio.org/content/citscribe-hackathon
[9] http://blog.eol.org/post/75726397084/eol-bhl-research-sprint-at-nescent-update-1
[10] http://wiki.pro-ibiosphere.eu/wiki/Data_enrichment_hackathon,_March_17-21_2014
[11] D. P. Shorthouse, R. Page, K. Richards, M. Tähtinen (2014). *Hacking OCR for pro-iBiosphere*. http://www.pro-ibiosphere.eu/news/4643_hacking%20ocr%20for%20pro-ibiosphere/
[12] http://jats.nlm.nih.gov/
[13] Penev L, Catapano T, Agosti D, et al. Implementation of TaxPub, an NLM DTD extension for domain-specific markup in taxonomy, from the experience of a biodiversity publisher. 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/.
[14] http://jats.nlm.nih.gov/extensions/bits/
[15] http://www.missouribotanicalgarden.org/plant-science/plant-science/world-flora-online.aspx
[16] http://www.cbd.int/gspc/

typically found in floras, whether relating to taxonomy, identification ecological or traits data (inter alia). In contrast to previous use cases, this requires high degrees of granularity. The approach followed by using the FlorML[17] schema might be exemplary here. Examples for repositories using the FlorML approach are the EDIT Platform instances for Flora Malesiana[18] and Flore du Gabon[19].

Likewise, the TaxonX[20] schema has been used by Plazi for over seven years to mark up legacy taxonomic literature with a focus on identification and enhancement of taxonomic treatments and their components. Plazi has marked up in TaxonX over 33,000 treatments from more than 2,500 publications.

Another format used to integrate data from multiple sources, that is gaining popularity, is the Darwin Core Archive[21] star schema. Currently, it has been used to deliver taxon treatment and occurrence data to the Global Biodiversity Information Facility (GBIF)[22] and the Encyclopedia of Life (EOL)[23] from both newly published articles in the Biodiversity Data Journal[24] and legacy literature, marked up and stored at Plazi[25].

The EU funded project OpenUp[26] has released about 1.6 Mio digitized specimen media data to GBIF and Europeana. Identifying and harvesting media data from multiple literature sources could additionally contribute to this result and make biodiversity media data available and retrievable through Europeana, which has a much broader scope than scholarly literature.

Other approaches that integrate data from multiple sources not only offer enhanced access to information from a large number of sources but try to aggregate and merge data in ways that allow them to be used for completely new purposes. These are reported in detail in the following section.

Repurposing and redistribution

The simplest uses of repurposing involve the separation of content from presentation information, to address the needs of human readers on different devices or the need to display diverse content in aggregating portals like EoL or PubMed Central (PMC). In PMC, all articles are available in two HTML layouts that are specific to PMC and different from the layouts in use at the various publishers that deliver content to PMC. This is made possible by the markup of document sections like abstract, methods, results, discussion and supplementary materials, as well as by the markup of equations, tables, figures and other elements within these sections.

As another example, the commercial aggregator PubChase mirrors openly licensed materials from PMC

[17] Hamann, T. D., Müller, A., Roos, M. C., Sosef, M., & Smets, E. (2014). Detailed mark-up of semi-monographic legacy taxonomic works using FlorML. Taxon 63(2):377-393. doi: 10.12705/632.11
[18] http://dev.e-taxonomy.eu/dataportal/flora-malesiana/
[19] http://dev.e-taxonomy.eu/dataportal/flore-gabon/
[20] http://www.taxonx.org/
[21] http://code.google.com/p/gbif-ecat/wiki/DwCArchive
[22] http://www.gbif.org/
[23] http://eol.org/
[24] http://biodiversitydatajournal.com
[25] http://plazi.org/
[26] http://open-up.eu/

on its site, and renders the content in yet another display format, using the open-source tool Lens[27], which is developed and maintained by the publisher eLife[28].

Pensoft journals like ZooKeys[29] and PhytoKeys[30], have marked up taxon names, references and geolocations for several years now using TaxPub. The online versions of such articles offer additional functionalities (e.g. links to cited sources, occurrence maps and taxon profiles) when hovering over such items. The journal Standards in Genomics[31] has also marked up species names for microbial genome papers, albeit using NamesforLife[32], which is patented.

A major advantage of markup is that significant components of documents can be addressed and extracted at sub-article level and distributed or recombined in new ways to enhance discoverability. For instance, Pensoft journals have established automated workflows to:

- post taxon treatments to the Plazi repository and to Species ID[33];
- post images to Species ID and EoL;
- post the full text of articles to PubMed Central, BHL and CLOCKSS[34];
- register taxonomic acts with ZooBank[35], the International Plant Names Index (IPNI)[36] and Index Fungorum[37];
- post occurrence records to GBIF.

Another example is the pro-iBiosphere Pilot 3, "Interoperability model between Plazi and the EDIT Platform for Cybertaxonomy based on transformations between XML-repositories and CDM-stores"[38]. The system establishes a transformation pipeline between marked-up literature from Plazi (an XML-based repository) and the CDM-stores of the EDIT Platform for Cybertaxonomy (a highly granular object-oriented data store). With this, the aggregated data become both searchable via the Portal system of the EDIT platform and accessible through its highly capable web-services layer. Also, the data may be reorganized, improved and enriched and therefore used for future revisionary work.

The web-service layer of platforms like the EDIT Platform allows to use data in completely new contexts, such as workflows developed in BioVel[39] (an EU funded project). These workflows are used, e.g., for ecological niche modelling, or for modelling changes over time by comparing old specimen data with new one. As old data are usually underrepresented in such workflows, opening up literature data is urgently needed for certain tasks. Also the EU funded project EU-BON[40] has a task (3.4) for creating a

[27] https://github.com/elifesciences/lens
[28] http://elifesciences.org/
[29] http://www.pensoft.net/journals/zookeys/
[30] http://www.pensoft.net/journals/phytokeys
[31] http://standardsingenomics.org
[32] https://services.namesforlife.com/
[33] http://species-id.net/wiki/
[34] http://www.clockss.org/
[35] http://zoobank.org/
[36] http://www.ipni.org/
[37] http://www.indexfungorum.org/
[38] See also D4.1 Report on strategies for improved cooperation and interoperability between infrastructures, http://wiki.pro-ibiosphere.eu/w/media/1/13/Pro-iBiosphere_WP4_FUB-BGBM_VFF_20122013.pdf
[39] https://www.biovel.eu/
[40] http://www.eubon.eu/

workflow to mark up literature to provide input into the modeling activities for observation and traits data which may use the pipeline described above and intends to use the data for completely new purposes.

In mathematics, the markup of formulas in MathML[41] has provided a basis for MathWebSearch,[42] an open-source search engine that finds related mathematical formulas, based on an algorithm used in automatic theorem proving, which is integrated into the search facilities in use at the zbMATH platform.[43] In principle, similar search engines could be built on the basis of other markup vocabularies, and in the pro-iBiosphere context, traits (as covered by Pilot 4) would be particularly interesting.

The Optical Society of America have recently digitized all legacy issues of their journals and marked them up using JATS.[44] Once this was accomplished, they could offer enhancements to the HTML versions of their articles (similar to what was discussed in the layout section above) as well as a database of all their images[45] as an additional product for their subscribers. Since all their equations are marked up in MathML as well, they are now considering to offer MathWebSearch, too.

Another approach to rearranging content is to mine content repositories to extract - and possibly re-aggregate - specific kinds of content. This is exemplified by the Open Access Media Importer,[46] an automated tool that crawls openly licensed articles in PubMed Central on an ongoing basis for audio and video materials and uploads these to Wikimedia Commons, from where they can be embedded in Wikipedia articles.

While the tool has successfully transferred over 16 000 files from hundreds of journals so far, it has uncovered multiple markup-related issues in the process: ambiguities in the NISO JATS standard, or in the PMC Tagging Guidelines, as well as inconsistencies and multiple classes of errors in the JATS XML content that publishers deliver to PMC.[47]

Finally, markup efforts in one field can benefit similar efforts in other fields. For instance, JATS was originally developed for biomedical research articles but has now been adopted by physics and geoscience journals as well.

Metadata extraction and processing

High-quality bibliographic metadata are essential for the scientific knowledge infrastructure to function and to enable linking through citations.

[41] http://www.w3.org/Math/
[42] http://trac.mathweb.org/MWS
[43] http://zbmath.org/formulae/
[44] Dineen MS, Gross M, Ashlem D, et al. NLM conversion to build "atomic" physics content in an agile fashion. 2013. In: Journal Article Tag Suite Conference (JATS-Con) Proceedings 2013 [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2013. Available from: http://www.ncbi.nlm.nih.gov/books/NBK159728/
[45] http://imagebank.osa.org/
[46] https://commons.wikimedia.org/wiki/User:Open_Access_Media_Importer_Bot
[47] Mietchen D, Maloney C, Moskopp ND. Inconsistent XML as a Barrier to Reuse of Open Access Content. In: Journal Article Tag Suite Conference (JATS-Con) Proceedings 2013 [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2013. Available from: http://www.ncbi.nlm.nih.gov/books/NBK159964/

Structured metadata traditionally exist separately from the main body of text. However, some document metadata (e.g. authors, addresses, title, journal, abstract) may be marked up within the document itself and thus available for metadata harvesting. Some journals, e.g. Earth science journals published by Copernicus, have been providing some basic JATS-flavored markup of article metadata - including abstracts and references - for about a decade, even though they have only recently started to work on providing full-text XML (also using JATS).

The transition from print to digital is affecting metadata handling in multiple ways. For instance, article pagination - a hallmark of print-based metadata - is irrelevant in born-digital, reformattable documents, which allow optimal viewing experience on a large set of diverse devices. Furthermore, citations to formal publications are increasingly complemented by citations to other research artifacts (some of which may be versioned), for which citation practices have only begun to be standardized (e.g. for data[48] and code[49]). Similarly, attribution is complicated by the increasingly collaborative nature of research in many fields (including biodiversity-related ones), to the point that taxonomies for author contributions have started to emerge.[50]

Depending on markup granularity and uniformity, metadata can be used to generate dashboards, e.g. of recent publications on a given taxon or by researchers from a given institution or acknowledging support from a specific funding source. With suitable markup, research fronts can be studied as publication networks based on co-words, on co-citations, on bibliographic coupling, or combinations of these techniques (Van den Besselaar and Heimeriks 2006)[51], without the need for additional text mining. Unique identifiers for concepts like taxa, institutions, authors and funding sources will be key enablers of such applications. Prototype visualizations based on marked-up treatments from Plazi have been produced during the Biodiversity Data Enrichment Hackathon in March 2014 at Naturalis.[52]

A closely related topic is the linking between publications by means of citations and references. References may include a globally unique identifier, or solely rely on quality metadata from which the referenced object can be inferred. Where persistent identifiers are missing (which may be the case even for articles published after DOIs had become the de-facto standard for that purpose) the semantic linking is hindered by the plethora of citation styles which are in use across the scholarly literature. These have many idiosyncrasies (and thus potential sources of ambiguities and errors), for example, inconsistent abbreviations of author and journal names, date formats, punctuation, and text styling (bolding and italics).

Facilitating data mining

Since markup can be produced at multiple levels of granularity, existing coarse-grained markup can be

[48] http://www.force11.org/datacitation
[49] http://thenextweb.com/dd/2014/03/17/mozilla-science-lab-github-figshare-team-fix-citation-code-academia/
[50] http://dx.doi.org/10.1038/508312a
[51] P. van den Besselaar, G. Heimeriks, 'Mapping research topics using word-reference co-occurrences: a method and an exploratory case study', Scientometrics 68 (2006) pp 377-393. doi: 10.1007/s11192-006-0118-9
[52] D. King, J. Miller, G. Sautter, S. Pereira (2014). *Data visualisation task for pro-iBiosphere.* http://www.pro-ibiosphere.eu/news/0_4_2014#4651

used to guide - and typically speed up - more fine-grained markup procedures. For instance, once the beginning and end of a taxon treatment are marked up, the markup process for traits can operate within those boundaries (see also the pro-iBiosphere pilot 4[53] on generating identification keys by re-using morphological characters from published species descriptions), as could a data mining tool for traits.

Similarly, if citations to materials in collections are marked up in the literature, a project on digitizing collections can build on that to identify cases where specimen labels from a particular collection have already been transcribed (as is common in taxonomic revisions), which may in turn help with the creation of occurrence maps or investigations into the seasonality of the organisms covered by the collection.[54]

Disambiguation and machine reasoning

Search engines should lead their users to the information they are looking for, if it exists at all. In the context of biodiversity research - for which names of taxa, geographic entities, authors, institutions, genes, characters, traits, and features are significant objects of interest - the original data may be enhanced by identifying and normalizing these names through computer-assisted techniques such as named entity recognition, associating the recognized entities with terms in controlled vocabularies and ontologies. The resulting *semantic* markup can then facilitate retrieval of semantically related concepts despite variation in form or presentation.

While some level of noise in the search results may be tolerable for human users, it often breaks automated tools. Here, semantic markup can help to disambiguate taxonomic from non-taxonomic uses of ambiguous strings like *Victoria*, which may refer to people, place names or taxa, amongst other things.

To distinguish between taxonomic uses of the valid term for the plant genus
Victoria Lindl., 1837 (Magnoliophyta: Nymphaeaceae)
and the homonymic valid term for the animal genus
Victoria Warren, 1897 (Lepidoptera: Geometridae),
a fine-grained level of markup and semantic annotation is necessary, using ontologies[55] specific to biological taxonomy or subsets thereof.

During the Biodiversity Data Enrichment Hackathon, a use case was presented that called for the creation of an RDF knowledge base of plant phenotypes by extracting trait data from digitised floras. The resulting ontology, the Flora Phenotype Ontology (FLOPO)[56], consists of more than 25,000 classes describing plant traits and phenotypes, and every class in FLOPO has at least one taxon annotation in

[53] http://wiki.pro-ibiosphere.eu/wiki/Pilot_4
[54] Dikow, T., Meier, R., Vaidya, G. G., & Londt, J. G. H. (2009). Biodiversity research based on taxonomic revisions—A tale of unrealized opportunities. *Diptera Diversity: Status, Challenges and Tools. Leiden: Brill Academic Publishers*, 323-345. doi: 10.1163/ej.9789004148970.I-459.55

[55] Schulz, S.; Stenzhorn, H.; Boeker, M. (2008). "The ontology of biological taxa". *Bioinformatics* **24** (13): i313. doi:10.1093/bioinformatics/btn158.
[56] https://bioportal.bioontology.org/ontologies/FLOPO

one of the processed floras. These floras have since been annotated with terms obtained from the Environment Ontology (ENVO)[57], providing a proof of concept of an ontology-mediated knowledge base that can be brought to bear on a variety of research questions on phenotype/environment interactions; functional diversity; and community ecology. Work is now under way to standardize the extracted taxon names in the floras using IPNI identifiers and represent them together with their environment and FLOPO annotations as Linked Data in an RDF store.

Besides FLOPO and ENVO, a number of other ontologies relevant to taxonomy and systematics are available through the BioPortal[58] operated by National Center for Biomedical Ontology (NCBO)[59], such as the Biological Collections Ontology (BCO)[60] and the Taxonomic Rank Vocabulary (TAXRANK)[61]. These are complemented by ontologies hosted elsewhere, e.g. the Ontology for the Museum Domain (MAO)[62] and the Ontology of the Birds of the World (AVIO)[63], both available through the Finnish Ontology Library Service (ONKI)[64].

Most existing ontologies have a very clearly defined scope, within which they have been crowdsourced in some way, though the respective crowds tended to be tiny subsets of the corresponding communities of practice. They also often have some mechanism in place for updates, but these tend to be slow. The situation is similar with annotation schemes like the Open Annotation Data Model[65].

In this regard, the Wikidata[66] project - already briefly introduced in D3.3.1 and more recently reviewed[67] - is of particular interest. It is a collaboratively editable knowledge base intended to eventually cover the breadth of topics available across all Wikimedia[68] projects (Wikipedia[69] and its sister projects like Wikisource[70], Wikiversity[71] or Wiktionary[72]) in all their languages, and it can be edited by anyone at any time. It is built on top of MediaWiki (the wiki engine that runs all Wikimedia projects) by way of Wikibase, a MediaWiki extension for handling structured data.

Designed to move data curation from multiple language versions of Wikipedia and its sister projects into one central and multilingual environment, Wikidata covers a very broad range of topics that includes all formally described taxa. Each concept (or item) receives a permanent identifier: for some of the *Victoria* examples mentioned above, these would be Q9439 for the British queen who reigned

[57] https://bioportal.bioontology.org/ontologies/ENVO
[58] https://bioportal.bioontology.org
[59] https://www.bioontology.org/
[60] https://bioportal.bioontology.org/ontologies/BCO
[61] https://bioportal.bioontology.org/ontologies/TAXRANK
[62] http://onki.fi/en/browser/overview/mao
[63] http://onki.fi/en/browser/overview/avio
[64] http://onki.fi/en/
[65] http://www.openannotation.org/spec/core/
[66] https://www.wikidata.org/
[67] Denny Vrandečić, Markus Krötzsch. Wikidata: A Free Collaborative Knowledge Base. In Communications of the ACM (to appear). ACM 2014. http://korrekt.org/page/Wikidata
[68] http://www.wikimedia.org/
[69] http://www.wikipedia.org/
[70] https://wikisource.org/
[71] http://www.wikiversity.org/
[72] http://www.wiktionary.org/

throughout most of the 19th century, Q36687 for the Australian state, Q312194 for the waterlily, and Q7926540 for the moth.

Each concept can be annotated with statements about specific properties[73], the range of which is continuously expanding as Wikidata is being developed and populated, and each statement can have multiple sources (e.g. due to multiple formal taxonomic revisions). Wikidata activities related to biological taxonomy are being coordinated by the Taxonomy task force[74], which anyone can join. A taxon that is well annotated with currently available properties is Cactaceae (Q14560).

In the long run, it will become possible to add statements that go beyond bibliographic and treatment-style morphological and phylogenetic information for a given taxon. For instance, information could be included as to whether individuals of a species have successfully passed a mirror self-recognition test[75], or doubts about the methodology underlying such results[76].

It is possible to map existing ontologies onto Wikidata. For instance, Property:P1060[77] for the transmission process of a pathogen has been modeled after the pathogen transmission ontology[78] from Open Biomedical Ontologies[79].

Since Wikidata uses a Creative Commons CC0 waiver to maximize potential use and reuse, only ontologies that are in the Public Domain or available under CC0 can be integrated this way. Many existing ontologies, however, are available under restrictive terms, e.g. the Plant Ontology[80] under a Creative Commons Attribution-NoDerivs license (CC BY-ND), which prohibits derivative works.

Provisioning the biodiversity literature with permanent identifiers to collaboratively annotatable and versioned items (either directly from Wikidata or from a Wikibase instance dedicated to biodiversity information) would allow for machine reasoning of much richer flavours than markup that points at static items.

Outlook

The range of use cases illustrated by the examples above triggers the question of prioritization: how should markup efforts be balanced between the desire to cover the vast quantities of legacy literature and the desire to quickly gain a comprehensive knowledge base for the most recent state of knowledge?

---

[73] https://www.wikidata.org/wiki/Wikidata:Properties
[74] https://www.wikidata.org/wiki/Wikidata:Taxonomy_task_force
[75] Prior, H.; Schwarz, A.; Güntürkün, O. (2008). "Mirror-Induced Behavior in the Magpie (*Pica pica*): Evidence of Self-Recognition". *PLoS Biology* **6** (8): e202. doi:10.1371/journal.pbio.0060202.
[76] Soler, M.; Pérez-Contreras, T. S.; Peralta-Sánchez, J. M. (2014). "Mirror-Mark Tests Performed on Jackdaws Reveal Potential Methodological Problems in the Use of Stickers in Avian Mark-Test Studies". *PLoS ONE* **9** (1): e86193. doi:10.1371/journal.pone.0086193.
[77] https://www.wikidata.org/wiki/Property:P1060
[78] http://www.berkeleybop.org/ontologies/trans.owl
[79] https://en.wikipedia.org/wiki/Open_Biomedical_Ontologies
[80] http://www.plantontology.org/

We conclude that pre-publication markup of "in press" materials within current publishing workflows should have the highest priority. However, this is not sufficient to timely achieve the breadth of a knowledge base as required for the future of biodiversity research.

Here, large-scale revisionary taxonomic works like Floras, Faunas and Mycotas - which are one of the focus topics of the pro-iBiosphere project - can play a crucial role, for two main reasons:

- They are essentially reviews of their subject (albeit they frequently add original material), comprehensively representing contemporary expert knowledge on the topic (with some delays incurred through the publication process).
- Their sequence over time provides a version history of scholarly knowledge about their respective taxa. If these works were fully marked up from their most recent (or the next upcoming) edition onwards, this semantic enrichment and its integration into a global semantic knowledge system would serve as a useful set of pointers to past literature on the topic. This would reduce the need to mark up those legacy parts of the literature, and facilitate their markup should it become necessary.

Once revisionary works are marked up at a reasonable level of detail, markup efforts of the biodiversity literature at large can be focused on taxa or geographic regions that are not or not sufficiently covered yet, perhaps in combination.

What a 'reasonable level of detail' is, will certainly depend on usage (and vice versa). Taxon names, bibliographic references and geolocations would seem like a basic set that is useful for many purposes, perhaps along with identifiers for collections, specimens, funding sources and so forth, as discussed above.

In many contexts, especially General Ecosystem Models, such a basic level of markup may not be sufficient, and more detail is necessary, including morphological, genetic, behavioral, ontogenetic and ecological information, as well as experimental evidence data. This will allow comprehensive modeling and machine-assisted investigations into relationships, e.g. between angiosperm leaf venation patterns, growing season temperature, and atmospheric CO_2 concentration.[81]

Fine-grained markup is resource-intensive to produce. Regardless of the form of presentation, techniques, and environments of semantically enriched data, the great challenge - and what should be the focus of expense and effort - is in developing the tools, approaches, mechanisms, workflows and environments to perform the semantic enhancement efficiently, easily, accurately, and extensibly at the scale required.

Natural Language Processing techniques, approaches such as Named Entity Recognition, text and data mining, and tools such as Google Refine[82] have been employed for years in other disciplines scientifically as well as commercially. What remains is the adaptation of these tools and techniques to the needs of biodiversity. The goal is to deploy them in - ideally customizable - workflows where the required input from domain experts to produce high quality data is minimized.

[81] Blonder, B.; Enquist, B. J. (2014). "Inferring climate from angiosperm leaf venation networks". *New Phytologist*: in press. doi:10.1111/nph.12780.
[82] http://code.google.com/p/google-refine/

The pilots within the pro-iBiosphere project go in this direction, but given that the project is a coordination action, they could only be explored, not implemented or deployed at the scale required for a future Open Biodiversity Knowledge Management System.

In this regard, it is noteworthy that a crowd of non-experts can under some conditions be trained to perform expert tasks, as demonstrated by numerous citizen science initiatives that range from the exploration of protein structures[83] to the genre annotation of oil paintings[84] or - bringing artwork, ontologies and crowdsourcing together - the identification of bird species depicted in paintings[85].

Markup within documents may for many use cases be only a limited solution, preparing the path towards novel knowledge publishing systems that integrate publishing of free-form articles, of data providing the necessary evidence, and fine-grained provision and curation of semantic statements.

The Wikidata model, wherein any given biological entity (e.g. ecosystem, taxon, gene) can be annotated with essentially unlimited level of detail, accommodates conflicting and contradictory scientific conclusions and supports workflows towards consensus as well as options for permanent dissent (due to different fitness for a different use). It may be a viable alternative to prospective markup and could serve as a component in workflows to enrich legacy literature.

A major aspect in favour of the Wikidata model are the timing, scope and location of corrections or updates. Major revisionary works often take years to compile. With the dominant publishing mode being still paper-centric, it often takes decades for updates or corrections to be published. Where updates are available, earlier versions have no way to point to such modifications. In a Wikidata model model, such modifications could be provided at any time, right in the location where users would be looking for them, and in case a specific version of the information in a Wikidata model is needed (a rough equivalent of references like "*sensu* Miller et al. 2013"), each version of the entry has a permanent identifier that can be used for that purpose.

Given that major revisionary works have so far been carried out almost exclusively by experts on the respective taxa (and often *the* experts), a system that allows for others - let alone anyone, as in Wikidata - to participate might be prone to inconsistencies in terms of quality. Further investigations are necessary to explore how to navigate these issues and how to use available versioning or qualifier options to manage this. A Wikidata model could potentially revolutionize the way taxonomic revisions are being performed and communicated.

[83] Cooper, S.; Khatib, F.; Treuille, A.; Barbero, J.; Lee, J.; Beenen, M.; Leaver-Fay, A.; Baker, D.; Popović, Z.; Players, F. (2010). "Predicting protein structures with a multiplayer online game". *Nature* **466** (7307): 756–760. doi:10.1038/nature09304.
[84] M. C. Traub, J. van Ossenbruggen, J. He, L. Hardman (2014). Measuring the Effectiveness of Gamesourcing Expert Oil Painting Annotations. Lecture Notes in Computer Science Volume 8416:112-123. doi: 10.1007/978-3-319-06028-6_10.
[85] http://sealincmedia.wordpress.com/2014/03/31/linking-birds/

Roadmap and first elements of a work plan

In recent years, a number of perspectives on the near future of biodiversity informatics have been put forward, including the Global Biodiversity Informatics Outlook (GBIO)[86], the "Decadal view of biodiversity informatics"[87] white paper, and the observation[88] that the differences between the biodiversity literature and biological databases are steadily diminishing.

Before that background, the focus has to be on the ultimate goal of a new, capable, efficient knowledge management system for biodiversity information, one that transparently leads from original data to reviewed conclusions, provides full version history as well as attribution, and is semantically interlinked between different knowledge domains.

In order to achieve this, constraints on the openness of biodiversity data need to be tackled. Three paths have to be pursued in an increasingly coordinated fashion:

1. Publishing new knowledge in a semantically aware system. This will be a collaborative and distributed system, and its implementation and maintenance a permanent activity requiring stable funding (by participating institutions with a long-term knowledge preservation commitment and ability like museums, external funding bodies, or the private sector).
2. Cost-efficient, largely automated basic enrichment of large - and especially revisionary - parts of the legacy literature with the primary goal of increased discoverability of taxon names, person names, geolocations, or citations. This will be an ongoing activity requiring funding until the need for enrichment of legacy publications is saturated.
3. Fine-grained markup of selected literature with the goal of seeding the semantic knowledge management system. These activities will have to transition from the currently very large-scale but rare publications of Faunas and Floras to a more agile system.

The focus here has to be on efficient operation for a purpose rather than just-in-case operations. Capturing the newest taxon treatments in semantically enabled formats is more important than capturing the complete history of taxonomic publishing in the same way. How the different pilots are contributing to this workplan is discussed in the Appendix.

[86] http://www.biodiversityinformatics.org/

[87] Hardisty, A.; Roberts, D.; Biodiversity Informatics Community; Addink, W.; Aelterman, B.; Agosti, D.; Amaral-Zettler, L.; Ariño, A. H.; Arvanitidis, C.; Backeljau, T.; Bailly, N.; Belbin, L.; Berendsohn; Bertrand, N.; Caithness, N.; Campbell, D.; Cochrane, G.; Conruyt, N.; Culham, A.; Damgaard, C.; Davies, N.; Fady, B.; Faulwetter; Feest, A.; Field, D.; Garnier, E.; Geser, G.; Gilbert, J. (2013). "A decadal view of biodiversity informatics: Challenges and priorities". BMC Ecology 13: 16. doi: 10.1186/1472-6785-13-16.

[88] Bourne, P. (2005). "Will a Biological Database Be Different from a Biological Journal?". PLoS Computational Biology 1 (3): e34. doi: 10.1371/journal.pcbi.0010034.

Appendix A - Contributions from other pro-iBiosphere deliverables and pilots

For most of the components of an open biodiversity knowledge management system, some basic functionality exists, but not at or near production level. Some of the gaps have been investigated in the pro-iBiosphere project - especially in the pilots - and will briefly be outlined below.

The pro-iBiosphere deliverable D6.1.2, "Report on costs" (due in May 2014) further investigates the cost issues around markup activities. It will include information on the core products and services identified during the business model workshop in February 2014, as well as their associated costs and economic viability.

In Pilot 1[89], interoperability between TaxPub and TaxonX has been demonstrated. TaxonX is the basis of the successful markup efforts driven by Plazi as well as others using the GoldenGATE toolkit, while TaxPub integrates taxonomy-specific terms into the JATS standard in use by many publishers as well as PubMed Central. This pilot contributes to the roadmap by showing that - despite different workflows for prospective and legacy markup - the results can be integrated in a consistent fashion, paving the way for an integrative Open Knowledge System.

In Pilot 2[90], a common query and response model has been developed for automated registration of higher plants (International Plant Names Index, IPNI), fungi (Index Fungorum, MycoBank) and animals (ZooBank), which contributes to the roadmap through pipelining registration and linking registry identifiers of newly published names or nomenclatural acts to their respective treatments, which are published in marked-up TaxPub format. Such an approach will save future costs and effort associated with post-publication markup and is the only way to put newly published information straight into a structured form that can be exported and aggregated automatically and re-used for different purposes.

As detailed in the "Repurposing and redistribution" section, Pilot 3[91] (Interoperability model between Plazi and the EDIT Platform for Cybertaxonomy based on transformations between XML-repositories and CDM-stores) contributes to this work plan by allowing fine-grained markup to feed the knowledge management system by loss-free transformation of markup data into the EDIT Platform. The latter allows further revisioning of the data, including intensive quality checking and data enrichment. Therefore, it also results in the publishing of new knowledge. The EDIT Platform is expected to either become part of the distributed knowledge management system or to allow synchronization with it.

In Pilot 4[92], the focus is on testing the level of semantic markup and annotation that will be required for re-using the information, e.g. for the purpose of identifying unknown organisms. This is highly relevant not only for identification. The granularity of knowledge necessary for identification is very similar to the level required for ecosystem modeling.

[89] http://wiki.pro-ibiosphere.eu/wiki/Pilot_1
[90] http://wiki.pro-ibiosphere.eu/wiki/Pilot_2
[91] http://wiki.pro-ibiosphere.eu/wiki/Pilot_3
[92] http://wiki.pro-ibiosphere.eu/wiki/Pilot_4

The pilot combines methods of advanced automatic markup with tests how to convert this to commonly used structured biodiversity knowledge systems and tools (e.g. xper[93], EDIT Platform/CDM[94]) using SDD [95] as an exchange format. It is a highly demanding test for the creation of a future knowledge system that opens the information on properties, traits, characters, etc. of organism for modeling and machine reasoning.

Pilot 9 - Spiders[96] (*From Catalog to Digital Fauna*). The goal of this pilot is to prepare XML treatments from spider taxonomic literature for integration with the new World Spider Catalog (WSC). The WSC references treatments, figures, and literature for all of spider taxonomy. This pilot is advancing the roadmap by providing fine-grained taxonomic treatments on Plazi. Database identifiers have been shared between WSC and Plazi, which will allow the catalog to link directly to the treatment on Plazi. In parallel, methods of visualization for primary specimen data are being developed. These data summaries facilitate the comprehension, comparison, and aggregation of primary specimen data.

---

[93] http://infosyslab.fr/lis/?q=en/resources/software/xper3
[94] http://dev.e-taxonomy.eu/trac/wiki/CommonDataModel
[95] Hagedorn, G.; Thiele, K.; Morris, R. & Heidorn, P. B. 2006. The Structured Descriptive Data (SDD) w3c-xml-schema, version 1.1. http://rs.tdwg.org/UBIF/2006/rddl.html
[96] http://wiki.pro-ibiosphere.eu/wiki/Pilots#Spiders