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)**

D3.2.2 Report on the state and quality of biosystematics documents and survey reports

Deliverable Status:	**Final**
File Name:	**pro-iBiosphere_WP3_Plazi_VFFa_31082013.pdf**
Due Date:	**31 August 2013 (M12)**
Submission Date:	**31 August 2013 (M12)**
Dissemination Level:	**Public**
Task Leader:	**Donat Agosti (Plazi)**
Authors:	**D. Agosti, T. Catapano, J. Cora, A. Güntsch, Q. Groom, G. Hagedorn, D. Kirkup, J. Macklin, D. Mietchen, J. Miller, R. Morris, A. Patton, L. Penev, D. Patterson, S. Sierra**

Copyright

© Copyright 2012-2014, the pro-iBiosphere Consortium. Distributed under the terms of the Creative Commons Attribution 3.0 License.

Consisting of:

Naturalis	Naturalis Biodiversity Center	Netherlands
NBGB	Nationale Plantentuin van België	Belgium
FUB-BGBM	Freie Universität Berlin	Germany
Pensoft	Pensoft Publishers Ltd	Bulgaria
Sigma	Sigma Orionis	France
RBGK	The Royal Botanic Gardens Kew	United Kingdom
Plazi	Plazi	Switzerland
Museum für Naturkunde Berlin	Museum für Naturkunde Berlin	Germany

Disclaimer

All intellectual property rights are owned by the pro-iBiosphere consortium members and are protected by the applicable laws. Except where otherwise specified, all document contents are: "© pro-iBiosphere project".

All pro-iBiosphere consortium members have agreed to full publication of this document. The commercial use of any information contained in this document may require a license from the owner of that information.

All pro-iBiosphere consortium members are also committed to publish accurate and up-to-date information and take the greatest care to do so. However, the pro-iBiosphere consortium members cannot accept liability for any inaccuracies or omissions nor do they accept liability for any direct, indirect, special, consequential or other losses or damages of any kind arising out of the use of this information.

Revision Control

Version	Author	Date	Status
1.0	Donat Agosti (Plazi)	01 June 2013	First Draft
2.0	Terry Catapano (Plazi)	14 August 2013	Draft
3.0	Terry Catapano, Donat Agosti (Plazi)	20 August 2013	Draft
4.0	Donat Agosti (Plazi), Daniel Mietchen (Mfn)	21 August 2013	Draft
5.0	Terry Catapano, Donat Agosti (Plazi)	22 August 2013	Draft
6.0	Donat Agosti (Plazi), Anton Güntsch (FUB-BGBM), Eva Kralt (Naturalis)	23 August 2013	Draft
7.0	Soraya Sierra (Naturalis)	24 August 2013	Draft
8.0	Jeremy Miller (Naturalis), Lyubomir Penev (Pensoft), Donat Agosti (Plazi), David Patterson (Maple Ferryman and Plazi)	25 August 2013	Draft
9.0	Soraya Sierra (Naturalis), Quentin Groom (NBGB)	26 August 2013	Draft
10.0	Eva Kralt (Naturalis)	28 August 2013	Draft formatted
11.0	David Patterson (Maple Ferryman and Plazi), James Macklin (AAFC)	26 August 2013	Draft
12.0	Donat Agosti (Plazi)	28 August 2013	Draft
13.0	Eva Kralt (Naturalis)	29 August 2013	Draft incorporated and formatted
14.0	David Patterson (Maple Ferryman and Plazi), Donat Agosti (Plazi)	30 August 2013	Final Draft reviewed
FF			Final Draft converted to Portable Document Format (PDF)

Table of Contents

Executive summary

The present document is a deliverable of the pro-iBiosphere project, funded by the European Commission's Directorate-General Information Society and Media (DG INFSO), under its 7th EU Framework Program for Research and Technological Development (FP7).

Biosystematics has a two hundred and fifty year old tradition of documenting the world's living species and higher taxa in highly standardized taxonomic treatments. The convention for taxonomic treatments consists of a scientific Latin name for each taxon, a list of citations to previous references to the described taxon (including any synonyms), a list of exemplar specimens, illustrations, a diagnosis, a summary of its distribution and behaviour and ecology, and other relevant information. These treatments have been published in articles and monographs to create a corpus of biosystematic literature of tens of millions of pages. The target audience for this literature has been the human reader. We can now extend this model with metadata and attached digital objects, with the potential to transform the biodiversity literature into a gateway to the content of collections of specimens, sounds, images, descriptions, interactive maps of occurrences, and DNA information. Here we address the challenge of how to enhance the publication process to make these rich data accessible, computable and re-usable.

We do not specify a comprehensive semantic schema, but target the process that will build a system that can scale to all challenges, can evolve to increased sophistication, and can call upon and link to existing and emerging external semantic data management systems. This approach can evolve to ensure that content users get the information they need.

We address the need to convert legacy literature into semantically enhanced documents or database records. By relying on pre-existing vocabularies, we will avoid duplication of effort. Existing schemas such as TaxonX schema provide a starting point for semantic enhancement. The use of the TaxPub extension of the Journal Article Tag Suite (JATS) Document Type Definition (DTD) will guarantee integration of the corpus of future and enhanced publications. The discipline will need new enhancement tools that enable more fine grained structuring of taxonomic descriptions and conversion among schemas. Finally, criteria should be established for the next generation Biosystematics Literature to facilitate machine readability, text and data mining, integration into emerging Semantic Web environments and genomics, and for the infrastructure to manage this information.

We make the following recommendations:

1. All biosystematic (= taxonomic) literature needs to be openly accessible to the maximum extent possible. At least publicly funded institutions should refrain from claiming intellectual property rights for biosystematic information and in respect of material which is protected by copyright or database rights, they should be commit it to the public domain by publishing it under a CC0 or similar license.

2. Biosystematics documents should be encoded in an open, platform-independent XML or an equivalent language.

3. The semantic elements of XML encoded documents should be cross-mapped to corresponding terms and concepts in external vocabularies.

4. Markup conventions should complement existing standards. The following elements should be marked up to the finest degree of granularity possible:

    - Scientific taxon names
    - Author names
    - Georeferenced observations
    - Type and voucher materials
    - Bibliographic references
    - Species traits
    - Treatments
    - Visual and audio material
    - Identification keys
    - DNA references

5. Markup conventions should complement existing standards.

6. Markup should be as explicit as possible and in open documentation to improve access to legacy literature and ease of their future use.

7. Nomenclatural acts and synonymies should be semantically enhanced to improve usability.

8. Semantic enhancement should allow progressive markup as an iterative process.

9. Funding agencies should support the development of tools for markup of biosystematic documents, especially of names, materials cited, bibliographic references, traits and treatments.

10. The community must develop and maintain registries of sources and repositories for semantically enhanced biosystematic publications, treatments and data to ensure visibility of and open persistent access to this corpus of material.

11. Stable globally unique identifiers should be used for semantic elements.

12. Reference databases must be developed, be easily accessed, and must be maintained.

13. iBiosphere should minimally export metadata relating to biodiversity data objected to the Linked Open Data Cloud.

Introduction

Overview

Documentation provides the core record and the primary means of communication of science. Unlike other sciences, biosystematic (i.e., taxonomic) publications include quasi-legal documents that must comply with the Codes of Nomenclature. They remain relevant for an indefinite period. This corpus of publication starts in 1754 and 1758 for plants and animals respectively.

Printed publication is increasingly being replaced by digital online publication. The Portable Document Format (PDF) is the dominant form of digital publication. PDF, however, though electronically transmittable and coarsely searchable, essentially preserves the conventions and limitations of static, printed documents, and are suited primarily to human readers. While this has brought progress in the dissemination of information, it falls well short of taking advantage of the full potential of highly linked digital publication.

The Internet and digital publications provide a framework in which we can call on machines to perform many of the tasks humans do. In a digital networked environment, the previously painstaking process of compiling lists of species of a given region or parasitizing a particular host, can be replaced with automatic extraction of information and enhanced with graphs of relationships among species or to reveal digital objects that are associated with a particular taxon (Parr et al., 2012). The data and associated metadata will be stored in dedicated databases, linked and referenced to the original publications, and be available for new uses beyond those originally intended by the authors (Bourne, 2005; Shotton, 2009).

There are already technologies to describe the content of biosystematics documents in a formal way. Chief among them is the Extensible Markup Language (XML) which uses schemas or Document Type Definitions (DTD) to specify the “grammars” within types of documents. Machine-readable controlled vocabularies or ontologies that define the terms and concepts and allow access to and cross-linking of XML encoded publications. Yet, at this time, less than 10% of the descriptions of new taxa are published in this advanced way. Significant time and effort is required to make enhanced publication widespread.

The goal of this report is to outline a strategy on how to enable and accelerate the semantic enhancement of documents.

Requirements of the users and creators of biodiversity information: syntax and semantics

The requirements of creators and users of biosystematics publications are very different and wide-ranging (Kirkup, 2013). With the increased access made possible by digital publication, the audience for Biosystematics Literature will expand beyond taxonomy into dependent fields such as ecology, environmental science and medicine.

There are many kinds of biosystematics publications that serve different purposes (Winston, 1999; Miller et. al., 2012). Monographs present thorough comprehensive treatments of all taxa in a taxonomic area. Flora and Fauna and similar works target the diversity in a geographic context, often synthesizing more than taxonomic information and including indigenous names, usage, conservation status, and so on (Marhold & Stuessy, 2013). A large corpus of articles include descriptions that comply with the nomenclatural codes of an estimated 17,000 new taxa each year. The slow process of conventional description is increasingly being complemented with discovery based on DNA analyses (Parr et al., 2012; Ratnasingham & Hebert, 2013). Creators are also users of biosystematic data, and all will benefit from greater efficiency and quality by having access to a seamless body of digital knowledge of biodiversity.

Markup

In order to promote the seamless body of information, we must define the useful elements of data clearly so that they are readily accessible for machine consumption. This can be accomplished with XML markup. With XML, tags are used to mark and categorize information of interest in a document. For example:

<taxon-name>Trichoteleia irwini</taxon-name> is named after Dr. Michael Irwin for his excellent field work...

The XML tag “taxon-name” indicates that the segment of text enclosed by <taxon-name> and </taxon-name> is the name of a taxon. Markup that uses agreed terms makes it easy for algorithms to extract information and to reference the data in publications. XML markup accomplishes two things. First, it provides a syntax for indicating the boundaries of text segments and so defines structure within a document. Second, it asserts provides semantics through the names of the semantic elements, asserting that the delineated text is an instance of or is about some concept; in this case, a taxonomic name. By defining classes and concepts, XML adds semantic meaning to the document.

XML syntax is able to provide more structure and semantics to encoded documents. For example:

<taxon-name>
<taxon-name-part taxon-name-part-type="genus">Trichoteleia</tp:taxon-name-part> <taxon-name-part taxon-name-part-type="species">hemlyae</tp:taxon-name-part> <object-id>urn:lsid:zoobank.org:act:A8E68218-7683-45D8-A71B-4E0662F53BC9</object-id> <object-id>urn:lsid.biosci.ohio-state.edu:osuc_concepts:241293</object-id>
</tp:taxon-name>

XML elements may be nested within other elements. Here, the taxon-name element is declared as consisting of two taxon-name-parts (genus and species) and two object-id elements identify remote digital objects. As illustrated here, XML elements may also have "attributes" which qualify the nature of the element (e.g., the taxon-name-part elements are of type "genus" and "species"). Through mechanisms such as nesting and attributes, a marked up document can achieve a high degree of syntactic granularity and semantic specificity.

Types of documents may be formally defined using one of several Schema Languages (e.g., DTD, RelaxNG, W3C XML Schema) which provide a mechanism to define the names of the semantic elements and their relationships. Such schemas can range from simple to complex, from small to large, for general purposes or for particular domains (including taxonomy, Penev et al. 2012). One of the most important schemas for Scientific, Technical, and Medical (STM) publications is the JATS (http://jats.nlm.nih.gov/about.html) formerly developed and maintained by the US National Library of Medicine (NLM) and the National Center for Biotechnology Information (NCBI). It is now an US National Institute for Standards Organization (NISO) standard, maintained by an independent team of experts. JATS is widely used by STM publishers, and is the internal format used by the NLM's PubMed Central repository of journal articles. For the biodiversity domain, JATS offers a solid foundation against which markup for Biosystematics Literature may be defined. It was with this rationale that Plazi developed the TaxPub extension of the JATS DTD (Catapano, 2010). TaxPub inherits all of the "generic" article elements from JATS, but adds a few elements and structures to model components of interest in biosystematics publications (e.g., treatments, taxonomic names, materials-examined, etc.). By extending JATS, rather than creating a new schema from the ground up, we saved an enormous amount of time, allowed users to benefit from the infrastructure, documentation, and tools existing for the widely adopted JATS, and facilitated sharing of TaxPub encoded articles with those already familiar with JATS. Indeed, by adopting the TaxPub extension of JATS, Pensoft was able to have its journals included in PubMedCentral, the first biosystematics publisher to have done so (Penev et. al., 2012).

JATS is not the only general purpose schema available to serve as a basis biosystematics markup. The Text Encoding Initiative (TEI) (http://www.tei-c.org/index.xml) provides a rich, extensible set of tagsets for a variety of document types. The ABLE project (http://able.open.ac.uk/corpus) extended the TEI by including taxonomy-specific elements from the earlier taXMLit schema (Weitzman & Lyall: http://www.sil.si.edu/digitalcollections/bca/documentation/taXMLitv1-3Intro.pdf) used for the markup of publications from the Bulletin of the British Museum (Natural History) (Willis et al, 2010).

Another option that is not widely employed would be to extend XHTML, the XML version of the Hypertext Markup Language, or HTML5 by using elements in these languages to mark up basic document features and to add semantics through the use of "microformats" (http://www.w3.org/html/wg/drafts/microdata/master/) in the attributes of these elements. For example:

> <p>Trichoteleia irwini is named after Dr. Michael Irwin for his excellent field work...<p>

In this case the span element <span... has been given semantic value by including the URI (Uniform Resource Identifier) for the DarwinCore (dwc - see below) term "scientific name". While microformats offer a powerful mechanism, they lack some important features of XML. For example, it is difficult to define rules for structure (such as, where elements may appear, in what order, how many times, etc...). The semantics are very weak for special purposes as XHTML and HTML5 and very general and rely entirely on external vocabularies for specific semantics. This approach might prove to be useful in converting richer XML encoded documents for publishing on the web.[1]

Vocabularies and ontologies

While XML is strong on syntax, it is weak on semantic features. Beyond defining element names in a schema, there is little information on the relationships among terms and concept. The formalization of semantic relationships can be achieved through machine-readable controlled vocabularies and ontologies. They are created in languages such as the Simple Knowledge Organization System (SKOS) (http://www.w3.org/2004/02/skos/) for vocabularies, and the Resource Description Framework schema (RDF) (http://www.w3.org/TR/1998/WD-rdf-schema/) and Web Ontology Language (OWL) (http://www.w3.org/TR/owl2-overview) for ontologies. Vocabularies define, organize, and relate terms in much the same way as a thesauri do, and, crucially, they identify terms by assigning URIs to them. Likewise, ontologies identify "concepts" or "things" and relate them as either subclasses or properties of one another. Identified with URIs and with definitions and relations formally expressed, external resources may refer to the concepts or terms by using their URIs. Indeed, terms and concepts and resources may all reference each other and combine to make meaningful assertions in a vast web of relations.

Just as XML schemas have proliferated over the past decade, so have ontologies and vocabularies. One of the most important in the area of biosystematics is DarwinCore (http://rs.tdwg.org/dwc/). DarwinCore "includes a glossary of terms (in other contexts these might be called properties, elements, fields, columns, attributes, or concepts) intended to facilitate the sharing of information about biological diversity by providing reference definitions, examples, and commentaries." The primary focus of DarwinCore is on taxa and their occurrences. It provides a solid terminology to provide clear semantics for biosystematics. Many other ontologies and vocabularies extend and refine what is available in DarwinCore.

[1] For a radical proposal and compelling discussion on the use of microformats more generally, see Piez, Wendell. "Abstract generic microformats for coverage, comprehensiveness, and adaptability." Presented at Balisage: The Markup Conference 2011, Montréal, Canada, August 2 - 5, 2011. In Proceedings of Balisage: The Markup Conference 2011. Balisage Series on Markup Technologies, vol. 7 (2011). doi:10.4242/BalisageVol7.Piez01.

The XML schemas, DarwinCore, and other ontologies and vocabularies provide the means by which we can create structured, granular, and conceptually explicit machine-readable biosystematics documents.

What data elements and what granularity?

Despite the conventions of form developed over two centuries, the documents which describe biodiversity information are extremely heterogeneous. This heterogeneity makes discovery of content more challenging. It can impede awareness of, and access to, the content. The diversity has many causes, from the diversity of goals of the scientists, varying editorial policies, and changes in the publication processes over time. To overcome the problems of access to content, there is pressure to achieve seamless integration of data by adopting a single XML schema for all Biosystematics Literature. While we acknowledge the appeal of this approach, we are not convinced that it will best serve the means and interests of all parties involved, nor that it can be effectively imposed on legacy literature. Rather than exclusively focusing effort on standardizing schemas, we should also allow the "market" decide which schema or schemas will emerge as de facto standards. There is a great advantage to newcomers to select from one of the existing schemas. JATS/TaxPub is arguably the most robust in terms of current implementation and maintenance infrastructure, but the TEI-based taXMLit can certainly serve many users purposes well. Others may need to develop their own schema to meet special needs. Whatever the names chosen for the elements, inconsistencies and incompatibilities will emerge but they can be fixed over time. Indeed, it will be those schemas that adapt quickly and intelligently to user demands that will emerge as the leaders.

At this time, it would be useful to understand what data are needed by the user community, how best they should be represented for current and future applications; or to contend with challenges such as the increasing separation of traditional taxonomy and DNA-based discovery of taxa (Parr et al., 2012). While no document will contain all data, each source will require some data. Our aim is an evolving mechanism that flexibly enhances any combination of data in ways that make the data available to serve any current or future purpose.

Experience suggests that there is a core set of widely useful data elements for Biosystematics Literature (see e.g. Winston, 1999 or Marhold & Stuessy, 2013). These are:

- Scientific Names
- (Geocoded) observation data
- Type (including voucher) materials
- Bibliographic references
- Traits
- Treatments (i.e., descriptions of a particular taxon)
- Images and audio

- Keys
- DNA sequences

In any biosystematic publication, as many of these elements as are available should be marked up. With basic markup in place, it will be possible to add additional layers of more granular markup for specific purposes later on. The degree of granularity of markup will often be determined by the resources, expertise, and needs of the producer. Some of the data elements will lend themselves to fine grained markup. For example, GPS technology is so common that observational data may be presented with a high degree of granularity. At the same time, most of the legacy data is not geocoded but it can be asserted and is thus valuable nevertheless. Likewise, bibliographic references, often coming out of reference manager databases, can be parsed easily into at least title, author, and year. The boundaries of treatments, and the types of sections they contain are easily delineated. Markup can be added retrospectively. Further, tools exist and are being developed to help automate the process of adding granular markup to names -such as the Global Names Recognition and Discovery tools (http://gnrd.globalnames.org/). Others can identify and markup observation data, traits, or bibliographic references. Abbreviations and other sorts of textual compression, such as journal titles in bibliographic references or location names, often defy automated tools, and we would be well served if such practices ceased. The development of automated feature-extraction tools should be a priority as the savings in labour costs are enormous, especially in relation to legacy data.

In addition, any data set on which the publication depends should be included as supplementary material or linked to from the publication.

Further, as controlled vocabularies and ontologies emerge, it will help if the schema or the tagged elements point to published identifiers for the concepts or terms. For example, each tagged scientific name should include or be linked to a persistent identifier for a corresponding record in a name registry (e.g., ZooBank, IPNI, etc.) (Patterson et al., 2010). If the names do not have identifiers in the registry, the future system should include a device for the streamlined registration of new names. Similarly, each bibliographic reference should include a DOI or other identifier for the cited work. For those works which do not have standard identifiers, the biosystematics community should take responsibility for establishing the identifiers. Various approaches can be followed: (i) make arrangements with a DOI issuer (e.g., Cross-Ref) to assign DOIs; (ii) leverage the investment in the Biodiversity Heritage Library and use their identifiers; (iii) use a stable “http URI” (Hyam et al., 2012; http://wiki.pro-ibiosphere.eu/wiki/Best_practices_for_stable_URIs); and (iv) publishers assign identifiers to treatments and keys with an associated effort (as is being pursued by Plazi) to assign them retrospectively.

In conclusion, there is no standard profile for biosystematic documents. Their content should be annotated, directly referenceable, and made suitable for analysis, retrieval, and processing, especially by machines. We should link to existing and new data sources or to supplementary material using persistent dereferenceable URIs that will open up a network of data for subsequent re-use.

Why do we need the pro-iBiosphere vision?

As we move forward, the value of biosystematics documents does not simply reside in the data that are included, but in how well prepared they are open access, how linkable they are, and whether their content is available for re-use, especially by machines (Harditsy & Roberts, 2013). This applies to both legacy publications and to prospective publications. Only if we make this transition, can we ensure that the wealth of existing knowledge is available to enrich the new approaches (barcoding and genomics) that rely on the molecular discovery of biodiversity.

No catalogue of the named taxa of the world has emerged after 250 years of scientific naming of species, about 20 years of the World Wide Web, and several years of compiling the Biodiversity Heritage Library. This vividly illustrates that traditional publishing conventions have left us with a major problem of extracting information, in this case - scientific names. Another on-going issue relates to data quality. Global Biodiversity Information Facility (GBIF) invests heavily in cleaning their 400 million occurrence records (Boakes et al., 2010). The path forward again lies in formalizing the semantics and granularity of observation data as part of the publication process.

Our vision targets both the process of enhancing documents AND the need to develop and maintain repositories for semantically enhanced and linked taxonomic publications or treatments in order to assure access to this new, unique corpus of literature (see recommendation nr. 9). An appropriate infrastructure must be available to capture, manage and integrate data and information from biosystematic sources. Without such an infrastructure, and its use by the community (Thessen and Patterson, 2011), users will not gain the full potential benefit from markup.

At present, only fragmentary elements of such an infrastructure exists. The vision of the pro-iBiosphere project is to prepare the ground for an integrating global system for the intelligent management of biodiversity knowledge (i-Biosphere). Such system needs to:

- Offer a robust service-oriented architecture for distributed taxon-level information
- Include a central registry of sources and services, with documentation, so that they can be discovered
- Provide open and free access to all names and taxonomic information from a single source to all persons who need biodiversity data, without legal barriers, copyright and database protection rights, nor requiring the consent of other individuals or institutions (WP2)
- Facilitate the re-use of biodiversity data and information
- Be interoperable with closely related initiatives
- Be fully aware of user requirements so that it serves the community as a whole (WP2)
- Have a solid long term sustainability plan to maintain the infrastructure and the services (WP6)

The system will need to be implemented in the near future. It will require additional funding as a second i-Biosphere project.

Recommendations

1.

Case. Biosystematics documents, especially taxonomic documents, are very rich in content. They are highly structured, and are the foundation of our knowledge of biodiversity. This corpus includes tens of millions of pages in the legacy literature; includes descriptions of approximately 17,000 new taxa being added annually to the estimated 2.3 million described living and extinct species, and complemented with many improved descriptions in floras, faunas, mycotas, and other monographs. This information is essential to progress in the biological sciences and especially to taxonomy, phylogeny, and ecology.

Recommendation: All biosystematic (= taxonomic) literature needs to be openly accessible to the maximum extent possible. Publicly funded institutions should refrain from claiming intellectual property rights for biosystematic information and in respect of material which is protected by copyright or database rights, they should be committed to the public domain by publishing them under a CC0 or similar license.

2.

Case. Biosystematics literature, both legacy and prospective, is rich in data. The data should be structured and made available for linking, analysis, retrieval, and re-use and not locked and unstructured inside binary formats such as PDF.

Recommendation: Biosystematics documents should be encoded in an open, platform-independent XML or adequate language.

3.

Case. Biosystematics literature is rich in data. If the documents are semantically enhanced, data can more easily be mined, extracted and reused, especially if the semantic elements are linked to widely accepted external vocabularies.

Recommendation: The semantic elements of XML encoded documents should be cross-mapped to corresponding terms and concepts in external vocabularies.

4.

Case. Biosystematics publications, though exhibiting a high degree of formal heterogeneity, share many core data elements of interest not only to taxonomists but to other domains such as ecology

Recommendation:. Markup conventions should complement existing standards. The following elements should be marked up to the finest degree of granularity possible:

- Scientific taxon names
- Author names
- Georeferenced observations
- Type and voucher materials
- Bibliographic references

- Species traits
- Treatments of taxa
- Visual and audio material
- Identification keys
- DNA references

5.

Case. Some elements of the biosystematic literature are very specific to taxonomy (such as taxonomic treatments and nomenclature), but others are shared with disciplines like ecology, physiology. We can use existing vocabularies in overlapping areas to integrate information across domains.

Recommendation: Markup conventions should be developed in coordination with and to complement existing standards. Existing vocabularies should be used as widely as possible and new elements introduced only for content that is specific to biosystematics.

6.

Case. Taxonomic publications are based on observations of material from various sources, often collected over centuries and housed in the natural history museum collections. Some of the records are only digitally discoverable through the explicit listings in publications.

Recommendation: Markup should be as explicit as possible and in open documentation to improve access to cited legacy literature and observation records and the ease of their future use.

7.

Case. Taxonomic publications are quasi-legal publications in the sense that the Codes of Nomenclature define conditions under which a name is published, how other nomenclatorial acts should be carried out, and in what form and what elements have to be presented. This highly structured data can easily semantically marked up, allowing machines to not only harvest the new nomenclatoral act but also to decide whether an act is valid in a nomenclatoral sense.

Recommendation: 7. Nomenclatoral acts and synonymies should be semantically enhanced to improve usability.

8.

Case. Semantic enhancement of publications imposes a cost on authors who will favor markup that addresses the needs of their own research questions. Later users might use parts of publications for different purposes. We cannot foresee all future usage, and so must approach markup with flexibility of future use in mind.

Recommendation: Semantic enhancement should allow progressive markup as an iterative process.

9.

Case. Semantic enhancement in the form of fine grained markup and linkage to external resources may be assisted by automatic means. We need more tools to make automatic markup possible, and add to them tools those that export content from source databases and deliver them to the publication process.
Recommendation: Funding agencies should support the development of tools for markup and exchange of biosystematic documents.

10.

Case. Taxonomic treatments are at the core of Biosystematics Literature. Each name is originally linked to a treatment that includes the diagnosis of the taxon and can encompass the entire knowledge about it. Treatments appear in publications that range from a treatment of one species to an extensive revision. A semantic environment will require repositories for treatments, similar to Genbank for DNA sequences or the Biodiversity Heritage Library for legacy publications. An example is the Plazi (http://plazi.org) repository used in this project's pilot studies in Workpackage 4 to demonstrate, among others, data transfer between Plazi and the EDIT platform CDM (FUB-BGBM) and its reuse.
Recommendation: The community must develop and maintain registries of sources and repositories for semantically enhanced biosystematic publications, treatments and data to ensure visibility of and open persistent access to this corpus of material.

11.

Case. Taxonomic publications are rich in data. To make fullest use, the data has to semantically enhanced and linked to the original source. The semantics used must be linked to domain-specific vocabularies or ontologies. The data (such as bibliographic references, authors, observations, images, cited treatments, and names) have to be linked to their sources. Appropriate identifiers should be used that are stable and widely used in the management of digital information.
Recommendation: Use stable globally unique identifiers for semantic elements.

12.

Case. Taxonomic publications are rich in references and acronyms that are well established in the biodiversity domain. These include bibliographic references, standard variables or indices used for descriptions and measurement, acronyms of collections, author names and increasingly domain-specific ontologies of, for example, descriptive data. External database that define and resolve these elements can be linked to from within publications. The use of shared global databases will reduce duplication and more quickly lead to a seamless knowledge management system.
Recommendation: Reference databases must be developed, be easily accessed, and must be maintained.

13.

Case. The Linked Open Data Cloud is an increasingly powerful environment through which data can be discovered and interlinked. Certain classes of biological data are well represented in this environment and are very actively used.

Biodiversity data are not well represented. Greater benefits will come from enhanced publications if the marked up content is accessible from the LODC.

Recommendation: iBiosphere should minimally export metadata relating to biodiversity data objects to the Linked Open Data Cloud

Acknowledgements

The pro-iBiosphere consortium acknowledges the participants in the workshops that were held in Leiden (February 2013) and Berlin (May 2013), all respondents to the Questionnaires, and the reviewers of this report.

References

Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Chang-qing, D., Clark, N.E., O'Connor, K. and Mace, G.M. (2010). Distorted views of biodiversity: spatial and temporal bias in species occurrence data. PLoS Biology 8(6): e1000385. http://www.plosbiology.org/article/info%3Adoi%2F10.1371%2Fjournal.pbio.1000385

Bourne, P. (2005). Will a Biological Database Be Different from a Biological Journal? PLoS Comput Biol 1(3): e34. doi:10.1371/journal.pcbi.0010034. http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.0010034

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

Hardisty, A. & Roberts, D. (2013). A decadal view of biodiversity informatics: challenges and priorities. BMC Ecology 13:16. http://www.biomedcentral.com/1472-6785/13/16

Hyam, T., Drinkwater, R.E. and Harris, D.J. (2012). Stable citations for herbarkum specimens on the Internet: an illustration from a taxonomic revision of Duoscia (Malvaceae). Phytotaxa 73: 17-30. http://www.mapress.com/phytotaxa/content/2012/f/pt00073p030.pdf

Kirkup, D.. (2013). D2.2 Report on user feedback, 7th Framework Programme, pro-iBiosphere project (Coordination & policy development in preparation for a European Open Biodiversity Knowledge Management System, addressing Acquisition, Curation, Synthesis, Interoperability & Dissemination).

Marhold, K. & Stuessy, T. (eds.) in collaboration with Agababian M., Agosti D., Alford M.H., Crespo A., Crisci J.V., Dorr L.J., Ferencová Z., Frodin D., Geltman D.V., Kilian N., Linder H.P., Lohmann L.G., Oberprieler C., Penev L., Smith G.F., Thomas W., Tulig M., Turland N. and Zhang X-C. (2013). The Future of Botanical Monography: Report from an international workshop, 12–16 March 2012, Smolenice, Slovak Republic. Taxon 62: 4–20. http://www.ingentaconnect.com/content/iapt/tax/2013/00000062/00000001/art00003

Miller, J., Dikow, T., Agosti, D., Sautter, G., Catapano, T., Penev, L., Zhang, Z-Q., Pentcheff, D., Pyle, R., Blum, S., Parr, C., Freeland, C., Garnett, T., Ford, L., Muller, B., Smith, L., Strader, G., Georgiev, T. and Benichou, L. (2012). From taxonomic literature to cybertaxonomic content. BMC Biology 10:87, doi:10.1186/1741-7007-10-87. http://www.biomedcentral.com/1741-7007/10/87

Parr, C.S., Guralnick, R., Cellinese, N. and Page, R.D.M. (2012). Evolutionary informatics: unifying knowledge about the diversity of life. TREE 27(2): 94-103 doi:10.1016/j.tree.2011.11.001. http://download.cell.com/trends/ecology-evolution/pdf/PIIS0169534711003247.pdf?intermediate=true

Patterson, D..J, Cooper, J., Kirk, P.M., Pyle, R.L. and Remsen, D.P. (2010). Names are key to the big new biology. TREE 25(12): 686-691. doi:10.1016/j.tree.2010.09.004. http://www.cell.com/trends/ecology-evolution//retrieve/pii/S0169534710002181?_returnURL=http://linkinghub.elsevier.com/retrieve/pii/S0169534710002181?showall=true

Penev, L., Catapano, T., Agosti, D., Georgiev, T., Sautter, G. and Stoev, P. (2012). Implementation of TaxPub, an NLM DTD extension for domain-specific markup 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/

Piez, W. (2011). Abstract generic microformats for coverage, comprehensiveness, and adaptability. In Proceedings of Balisage: The Markup Conference 2011. Balisage Series on Markup Technologies, vol. 7 (2011). doi:10.4242/BalisageVol7.Piez01.http://www.balisage.net/Proceedings/vol7/html/Piez01/BalisageVol7-Piez01.html

Ratnasingham, S., Hebert, P.D.N.(2013). A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. PLoS ONE 8(7): e66213. doi:10.1371/journal.pone.0066213. http://www.plosone.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0066213&representation=PDF

Shotton, D. (2009). Semantic Publishing: the coming revolution in scientific journal publishing. Learned Publishing 22 (2): 85–94; doi:10.1087/2009202. http://www.ingentaconnect.com/content/alpsp/lp/2009/00000022/00000002/art00002

Thessen, A. E. and Patterson, D. J. (2011). Data issues in the life sciences. ZooKeys 150: 15–51. doi: 10.3897/zookeys.150.1766. http://www.pensoft.net/journals/zookeys/article/1766/abstract/data-issues-in-the-life-sciences

Willis, A., King, D., Morse, D., Dil, A., Lyal, C. and Roberts, D. (2010). From XML to XML: the why and how of making the biodiversity literature accessible to researchers, LREC 2010 (7th International Conference on Language Resources and Evaluation), Mediterranean Conference Centre, Valletta, Malta, 17–23 May 2010.

Winston, J.E. (1999). Describing species: Practical taxonomic procedure for biologists. Columbia University Press, New York.