PRO-iBIOSPHERE
WWW.PRO-iBIOSPHERE.EU

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

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)**

D2.1.1 Report on ongoing biodiversity related projects, current e-infrastructures and standards

Deliverable Status:	**Final**
File Name:	**pro-iBiosphere_WP2_PLAZI_D2.1.1_VFF_30062013.pdf**
Due Date:	**June 2013 (M10)**
Submission Date:	**June 2013 (M10)**
Dissemination Level:	**Public**
Task Leader:	**Donat Agosti (Plazi)**
Authors:	**D. Agosti, L. Penev, T. Catapano, S. Eckert, T. Georgiev, Q. Groom, A. Güntsch, G. Hagedorn, P. Hovenkamp, D. Kirkup, E. Kralt, D. Mietchen, J. Miller, S. Sierra**

European Commission

Copyright
© Copyright 2012-2014 The pro-iBiosphere Consortium

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 rights reserved - OpenAccess reproduction are permitted under Creative Commons (CC-BY) 3.0".

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)	10 April 2013	Start of drafting
2.0	Lyubomir Penev (Pensoft)	25 June 2013	First draft completed
3.0	Soraya Sierra, Peter Hovenkamp, Jeremy Miller (Naturalis), Quentin Groom (NBGB), Daniel Mietchen (Pensoft), Donat Agosti, Terry Catapano (Plazi), Alan Paton, Don Kirkup (RBGK) Anton Güntsch, Sabrina Eckert (BGBM), Gregor Hagedorn (Mfn)	27 June 2013	Corrections and additions
4.0	Lyubomir Penev (Pensoft)	28 June 2013	Final draft assembled
5.0	Teodor Georgiev (Pensoft)	28 June 2013	Deliverable formatted
6.0	Eva Kralt, Soraya Sierra (Naturalis)	29 June 2013	Final corrections and additions; Deliverable format revised
7.0	Soraya Sierra (Naturalis)	30 June 2013	Deliverable submitted

Executive Summary

This report should be regarded as a baseline coordination platform for general high-level policy and strategy coordination. As a prerequisite, the report has been collaboratively prepared, with participation of all partners. The present activities, strategies, gaps, goals, use cases, interests and visions as well as cooperation and interrelations of the various European and international partners interested in participating in an Open Biodiversity Knowledge Management System (i.e., a *taxon treatment-like knowledge management system*) have been documented and updated. The report focuses on the (i) existing digital infrastructures; (ii) past publications and curation systems, including regional or global monographs; and (iii) data elements constituting taxonomic treatments, such as specimen data, images, sequences, taxon treatments, taxon names and their concepts, morphological characters, ecological and biological traits. Of special importance are further potential routes for cooperation between European and non-European biodiversity projects and platforms. An Advisory Board (AB) was created with the purpose of elaborating further recommendations for the improvement of data integration and interoperability; at present, the AB consists of four members.

Most important gaps, challenges and recommendations are summarised in a white paper (section 9 of the present report). The white paper includes the following topics:

- Transition to full and gold open access publishing models for biodiversity information;
- Policies towards open data publishing, sharing and re-use;
- Increased cooperation between EU and non-EU bioinformatics initiatives;
- Support digitisation, markup, data mining and re-publishing of legacy literature in advanced open access, semantically enhanced mode;
- Maximum interoperability and integration between traditionally produced printed works (so called "legacy literature") as in Biotas (e.g., Floras, Faunas, Mycotas, Protozoas) and newly produced data;
- Principle of machine readability: pressure to adopt advanced open access publishing models that allow automated data harvesting and re-use;
- Wide adoption of universal persistent and resolvable identifiers for biodiversity informatics elements;
- Further development and implementation for standards of data and associated metadata and data management infrastructures;
- Linked Data and associated tools and infrastructure (e.g., ontologies, vocabularies, registries) to become a predominant model of data management in the biodiversity domain;
- Promoting the use of existing directories for registration of tools and services derived from EU-funded projects.

Table of Contents

1 – Introduction

Biodiversity data and information constitutes an important source of knowledge for many disciplines - for example all life sciences, including key areas for humans' well-being, such as medicine, conservation and agriculture - and generally, its understanding contributes to human welfare and prosperity.

Biodiversity, the living component of our biosphere, is complex and can be described and classified in many different ways. A widely accepted view proposed by the Convention on Biological Diversity (CBD, 1992) partitions biodiversity into three domains: diversity of genes, species and ecosystems. This report follows this view and - while well aware of the intricate nexus of the three levels - focuses at the level of species diversity. The core of the biodiversity knowledge at this level is composed of information on species and specimens.

A wealth of such knowledge has been implicitly stored traditionally in the form of hundreds of millions of preserved specimens in Natural History collections, very recently in huge numbers of DNA-based records in genomic databases and in observations by conservation projects and citizen scientists stored as images and other recordings. Explicitly, this knowledge is available in several hundred millions of pages of taxonomic literature, notably Biotas (e.g., Floras, Faunas, Mycotas, Protozoas). These have been brought together over hundreds of years by European natural history institutions, herbaria, botanic gardens, through a continuous scientific synthesis and publishing process, and increasingly also through ongoing biodiversity monitoring programs.

Taxonomic literature provides the authoritative synthesis of the information that is available on a particular group of organisms (= taxon) in a particular region, compiled and evaluated by specialist taxonomists. It thus provides access to the highest quality data available. This includes links - originally implicit in paper-based publications, now increasingly explicit in electronic documents - to external resources upon which these analyses are based.

However, external and internal factors call for an adaptation of the production workflows and accessibility of these data, information and knowledge.

At present, the traditional practice of print-based biodiversity literature is opening up to accept e-publications as one of the (quasi-legal) formats to describe new taxa, that is to make new scientific names available for taxa discovered during the scientific process. With that, the name and all content linked to this new taxon are published. The current trend to request open access of research results offers a perspective that this information will not only be open for free use, but also linked and collated to other data. These links, either to other taxa or to different information sources on the same taxon, create the "big data" pool that can be used to attain a qualitatively higher level of knowledge.

This report is building on (i) two meetings convened within the pro-iBiosphere project, i.e., workshops on "Legacy Literature - Semantic markup generation, data quality and user-participation infrastructure" (held on the 13th of February 2013, in Leiden, the Netherlands) and "Coordination and routes for cooperation" (held on the 23rd of May 2013, in Berlin, Germany); (ii) the analysis of a questionnaire sent to the participants of the workshops (data available here); (iii) literature review of the articles “An appraisal of megascience platforms for biodiversity information” (Triebel et al., 2012); (iv) review of existing analyses on the biodiversity landscape, i.e., the report “Data & Modelling Tool Structures - Status Report on Infrastructures for Biodiversity Research” (LifeWatch, 2008); (v) the review paper of Berendson et al. (2011) Biodiversity information platforms: From standards to interoperability; (vi) the special volume of Smith and Penev (2011) e-Infrastructures for data publishing in biodiversity science; (vii) the Global Biodiversity Information Outlook (in prep., outcome from the Global Biodiversity Informatics Congress, Copenhagen 2012); (viii) the pro-iBiosphere’s List of other biodiversity projects and initiatives (2013); and (ix) the many years of experience of pro-iBiosphere partners in the field of biodiversity informatics.

2 – The concept of 'taxon treatment'

Perhaps the most significant component of taxonomic literature since the time of Linnaeus is the 'taxon treatment', i.e., information about a single taxon, typically headed by the taxon name and including morphological, distributional, taxonomic and other information. According to a definition by Norman Johnson (pers. commun.), adopted by Catapano (2010) and Penev et al. (2010), a taxon treatment is a "publication or (more frequently) section of a publication documenting the features or distribution of a related group of organisms (called a "taxon", plural "taxa") in ways adhering to highly formalized conventions". Some of these conventions (those pertaining to a subset of the treatment dealing with nomenclature) are maintained by scientific commissions accepted by the taxonomic profession, including the *International Code for Zoological Nomenclature* (ICZN) for animals, and the *International Code of Nomenclature for algae, fungi, and plants (ICNafp)*.

Taxonomic treatments are important because they permit labeling and delimiting a dedicated piece of information describing a taxon within a document from other similar pieces of information, describing other taxa. The retrieval of this content type has been identified as valuable to users of marked-up text through formal and informal assessment and the importance of enabling the user to retrieve a digitised taxon treatment as a core element has been recognised by most projects employing XML for taxonomic publications (e.g., Weitzman and Lyal 2004; Kirkup et al. 2005; Agosti et al. 2007; Sautter et al. 2007; Parr and Lyal 2007; Lyal and Weitzman 2008; Catapano 2010, Penev et al. 2010, 2011). Subsequent usage of the marked-up paper, for example dissemination of content to various aggregators, can in some cases be performed at the level of treatments. In addition, marked-up text or data can be retrieved by machine from either within or outside treatments. Inevitably the concept of the taxon treatment is incorporated in most, if not all, schemas developed for taxonomic literature, both in the markup process and to inform user queries.

Determining the boundaries of taxon treatments in the markup process can be problematic and may require manual intervention. Curry and Connor (2008) described the automatic identification and tagging of elements that typically occur within treatments, using stylistic rules to parse the text; they seem to have identified treatment boundaries *a priori*. More extensive algorithms also based on publication-specific stylistic rules (but not requiring *a priori* identification of treatment boundaries) were employed in a trial markup of a large single volume of the *Biologia Centrali-Americana* into taXMLit (Weitzman and Lyal 2006; Lyal and Weitzman 2008). The Plazi project atomises the publication into taxon treatments and seeks to maximise the number and consistency of tags by machine, either before or after publication (Agosti et al. 2007; Catapano 2010; Penev et al. 2010a). The concept of taxon treatments from the viewpoint of their markup in taxonomic literature has been described by Catapano (2010) and Penev et al. (2010a). Therefore, we shall only briefly summarise the main features of treatments.

There is considerable structural diversity in taxon treatments across taxonomic literature, the main sources of variation being historical differences in the approach to treatments between different groups of taxonomists and across time, and different editorial and publishers' formats. Nevertheless, it is possible to identify a few key features commonly found in treatments, such as the "Nomenclature" section, containing names and synonyms; "Material examined", containing data on the studied specimens; "Type designation", for new or revised taxa; "Morphological description"; "Etymology", on the origin of the newly proposed names; "Differential diagnosis", separating the taxon from similar taxa; as well as data on biology, ecology, or conservation status, etc.

Penev et al. (2010a) listed the following cases in which a logically delimited block of text within a taxonomic paper can be regarded as a taxon treatment:

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

At the same time, the following cases do not usually constitute a treatment:

1. A citation of a taxon name within a text, although such a citation usually holds information linked to the particular taxon. For instance, listing a species within a "plain" checklist cannot usually be a treatment of that species (in early literature under the ICZN such an instance must be considered a treatment in certain circumstances); a sentence within a text paragraph stating that "taxon X is parasitic on taxon Y" is neither a treatment of taxon X nor of taxon Y.
2. An identification key, because in some cases keys are constructed for related taxa that do not form a taxon (they may form a "species-group" or "taxon-group", but this is not a taxon unless a name is given to that group). Identification keys, even if they are exhaustive for a named taxon, are usually tagged separately from taxon treatments. However, some keys include all of the information within a publication about a given taxon, and the practice may be to consider them treatments. In some cases, keys contain taxon treatments, including those of new taxa, or synonymies. How keys are tagged is probably an editorial matter.
3. A single picture or group of pictures of a taxon. In some early publications, however, a taxon is based exclusively on an image and its caption, a source which is available under the relevant nomenclatural Code, and therefore the picture and caption have to be regarded as a treatment.
4. A single map or group of maps of the occurrences of a taxon.
5. Gene sequence(s) of a taxon.
6. SDD (Structured Descriptive Data) (or any) matrices, or raw data, or databases. Treatments can be relatively easily generated from databases, however, and information on a taxon can be considered as becoming a treatment when (a) it is published, and (b) corresponds to the aforementioned description of a taxon treatment.

A publication may consist of one or many treatments of different taxa of different taxonomic ranks. One taxon may have more than one treatment within a publication, although the tradition of systematics publishing usually assumes one “core” treatment per taxon within a document. One treatment can include nested treatments, e.g., a genus and its species.

3 – The diversity of biodiversity informatics landscape

3.1 – Data elements constituting taxon treatment

Biodiversity in the context of the pro-iBiosphere project focuses on the “species diversity” domain of biological diversity as defined in the Convention on Biological Diversity (CBD, 1992): "Biological diversity" means the variability among living organisms from all sources including, inter alia, terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems.

One way to visualize the diversity of species is to use hierarchical views. A phylogenetic tree is most widely used to represent the evolutionary relationship of the species (Fig. 1A). In a phylogeny, DNA sequences, genes and traits are part of an organism that is part of a species that can be grouped within more and more inclusive higher taxa. In an ecological view (Fig. 1B), DNA sequences, genes and traits are parts of an organism, which in turn is part of a species assemblage of an ecosystem. While a phylogeny represents the current knowledge of the unique evolution of the diversity (although competing phylogenetic hypotheses might exist), the species can belong to various species assemblages and ecosystems. Similarly, the classification of both species and ecosystems can be based on competing hypotheses, and thus a species might have multiple relationships to higher taxa. Specimens can also be identified differently (by different authors, or at different times) and thus might have multiple relationships.

These relationships are part of the scientific discovery process and documented in the scientific record, that is traditionally in formal publications and increasingly in electronic databases.

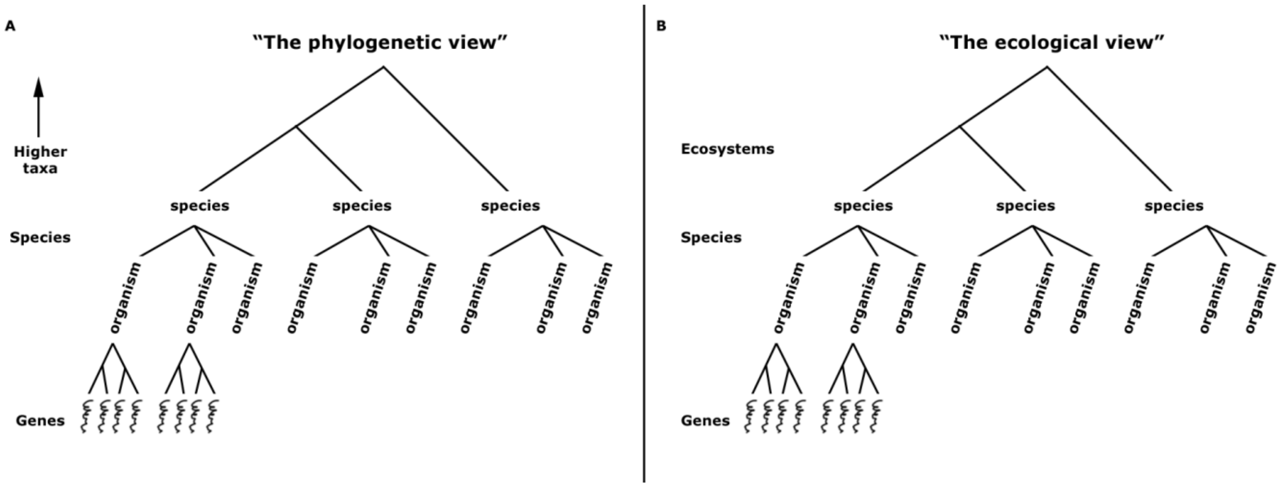


Figure 1. Two hierarchical views of biodiversity. A: Phylogenetic view. B. Ecological view.

These relationships, the elements they link (Fig. 2), their scientific documentation, and the discoverability of the elements and relationships play a crucial role in building an Open Biodiversity Knowledge Management System (OBKMS) (Fig. 3). They are the basis of an increasingly complex network of biodiversity related institutions; interactions within (Fig. 4) and outside the community; projects (e.g., LifeWatch, EU BON) and workflows (e.g., BioVel), each of which is built for a specific purpose (Triebel et al. 2012; LifeWatch 2008).

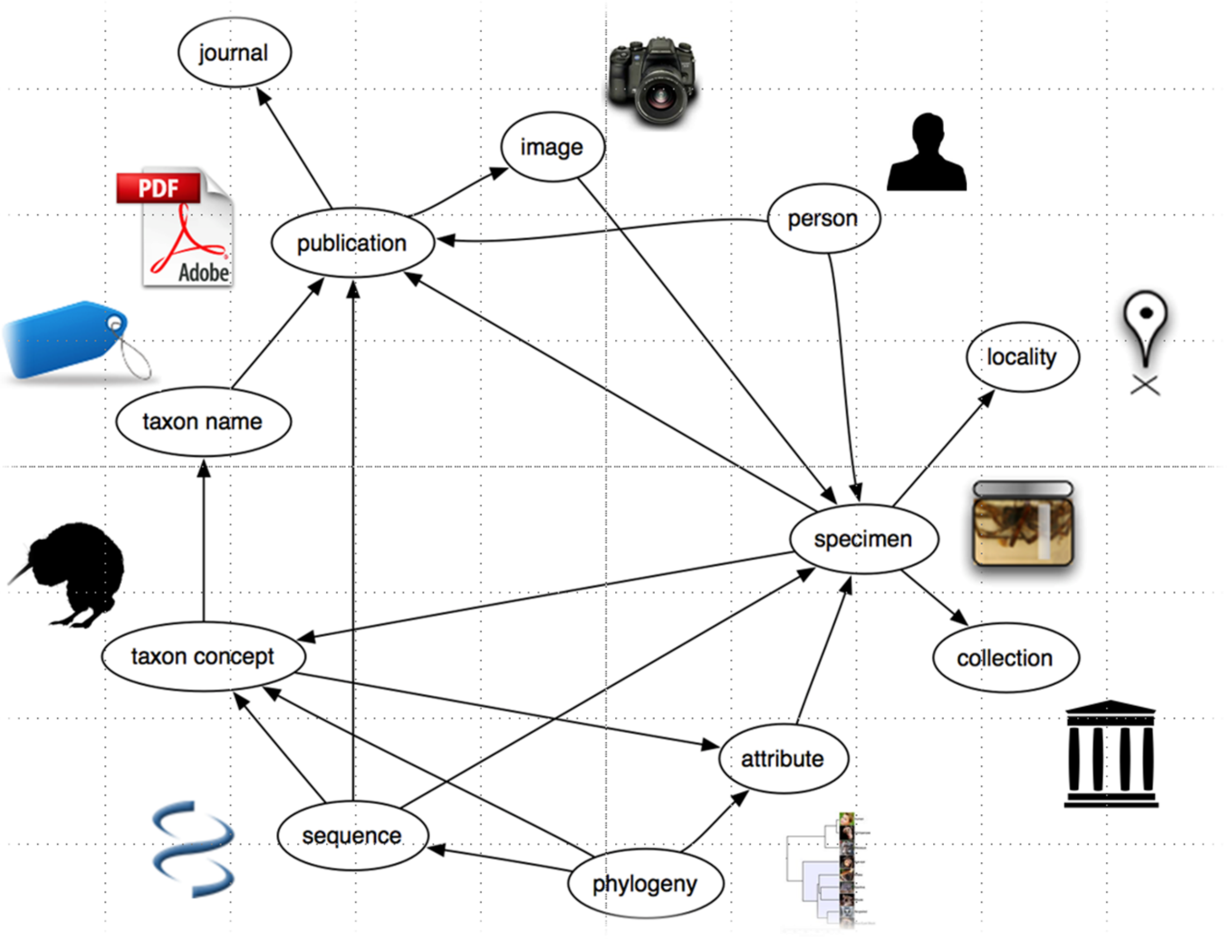


Figure 2. The main data elements connected to biological species that may constitute synthetic data (“taxon treatments”) and their relationships (based on http://iphylo.blogspot.com/2013/01/megascience-platforms-for-biodiversity.html and the presentation “Surfacing the deep data of taxonomy” given by Roderic Page at the pro-iBiosphere workshop held on the 13th of February 2013, in Leiden, the Netherlands).

An Open Biodiversity Knowledge Management System (OBKMS) that facilitates the open access of taxonomic data is essential because it will create synergies with other initiatives and projects. Taxonomic data will be linked in a wider context, for instance specimen data (e.g., morphology, anatomy, ecology, phytochemistry, etc.) linked with the genotype data generated by other communities. This will facilitate a better understanding of the animal and plant genes and simultaneously foster explanations of variation in biodiversity patterns. Such a system will also facilitate the acquisition, curation, synthesis and sustainable provision of high quality biodiversity data to partners and users, including e-science infrastructure projects as well as global initiatives on biodiversity informatics, because it will create an authoritative framework including naming of specimens, identification tools and descriptions.

Fig. 3 shows the general set up of the envisaged Open Biodiversity Knowledge Management System (OBKMS). Unlike the traditional cycle, legacy literature (1) first goes through a digitisation process (OCR) that makes further processing by computers possible (1a). The newly digitised legacy literature is used as literature data for finding specimen-related information by a scientist (2). This step is assisted by state-of-the-art tools to find additional information not present in the legacy literature (2a), which allow interaction and exchange with other, external, data resources (E1). Although steps (3)-(8) are essentially the same as in the traditional workflow, they are performed in real-time collaborative writing (e.g., Scratchpads or Pensoft Writing Tool) with other scientists around the world (E2), instead of the mostly solitary traditional work. Treatments are stored in an open access treatment repository (9). Taxonomic information dissemination and/or publication (9a) may follow two different pathways: i) The traditional way (10) that is printed media or PDF (10a), or static websites and digital documents (10b), and ii) Dissemination/publication of atomised data sets, texts that use semantic markup to make their contents available for processing by both humans and computers, is made possible by XML (11).

Ontologies are used to describe the data and its in- and external relationships in a text (12). Data come in many different categories, each of which will be made available separately (13). To the human reader, the information is presented as an integrated original document. To computers, data exchange and semantic integration and interoperability becomes possible, both within (E3) the Open Biodiversity Knowledge Management System (OBKMS), and with data resources outside it (E1).

Furthermore, converted legacy taxonomic literature (1a) can be fed directly into the dissemination and publication process (9a), again allowing the choice between non-atomised texts and atomised datasets. This avoids the need for a full taxonomic revision before re-publication and is especially useful for legacy data of uncontested quality. The curation cycle also never stops; treatments can be updated on the fly, either using new information from specialist scientists (14), information generated from within the semantic integration and interoperability process (E3), or interaction with other data resources (E1).

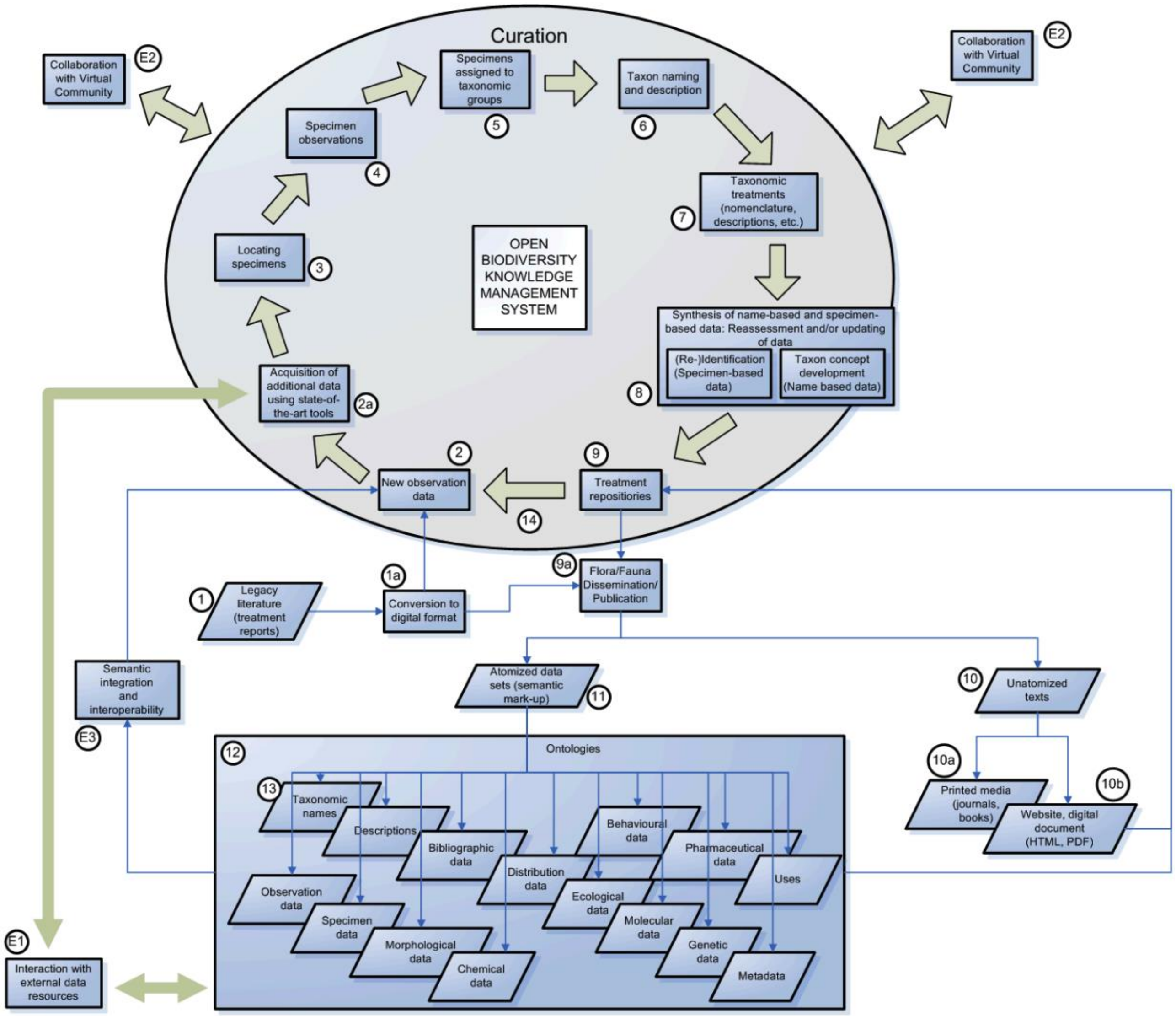

Curation
Collaboration with Virtual Community
E2
Specimens assigned to taxonomic groups
5
Taxon naming and description
6
Collaboration with Virtual Community
E2
Specimen observations
4
Taxonomic treatments (nomenclature, descriptions, etc.)
7
Locating specimens
3
OPEN BIODIVERSITY KNOWLEDGE MANAGEMENT SYSTEM
Synthesis of name-based and specimen-based data: Reassessment and/or updating of data
(Re-)Identification (Specimen-based data)
Taxon concept development (Name based data)
8
Acquisition of additional data using state-of-the-art tools
2a
2
New observation data
9
Treatment repositiories
14
1a
9a
Flora/Fauna Dissemination/ Publication
1
Legacy literature (treatment reports)
Conversion to digital format
Semantic integration and interoperability
E3
Atomized data sets (semantic mark-up)
11
10
Unatomized texts
10a
Printed media (journals, books)
10b
Website, digital document (HTML, PDF)
12
Ontologies
13
Taxonomic names
Descriptions
Bibliographic data
Behavioural data
Pharmaceutical data
Distribution data
Uses
Observation data
Ecological data
Specimen data
Molecular data
Morphological data
Genetic data
Chemical data
Metadata
E1
Interaction with external data resources

Figure 3. Open Biodiversity Knowledge Management System (pro-iBiosphere Annex I - Description of Work, 2012).

There was a consensus among the participants of the three pro-iBiosphere workshops held in Leiden that there is no up-to-date overview of all major recent biodiversity projects and initiatives focusing on treatments and handling data that may constitute treatments, as well as on technical and semantic interoperability between their data holdings, approaches and work flows.

Previous surveys suggest that all the projects have some overlap but different goals and different uses of data (e.g., plot specimens for a distribution map, create niche models, create policy recommendations for conservation). A straightforward and standardised way for cooperation between the projects is a major challenge. An analysis of the interaction of scientists participating at the e-Biosphere meeting in London in 2008 shows a very dense network of interaction between the scientists that includes both standardised exchange as well as specific person-to-person interaction and data exchange (Fig. 4).

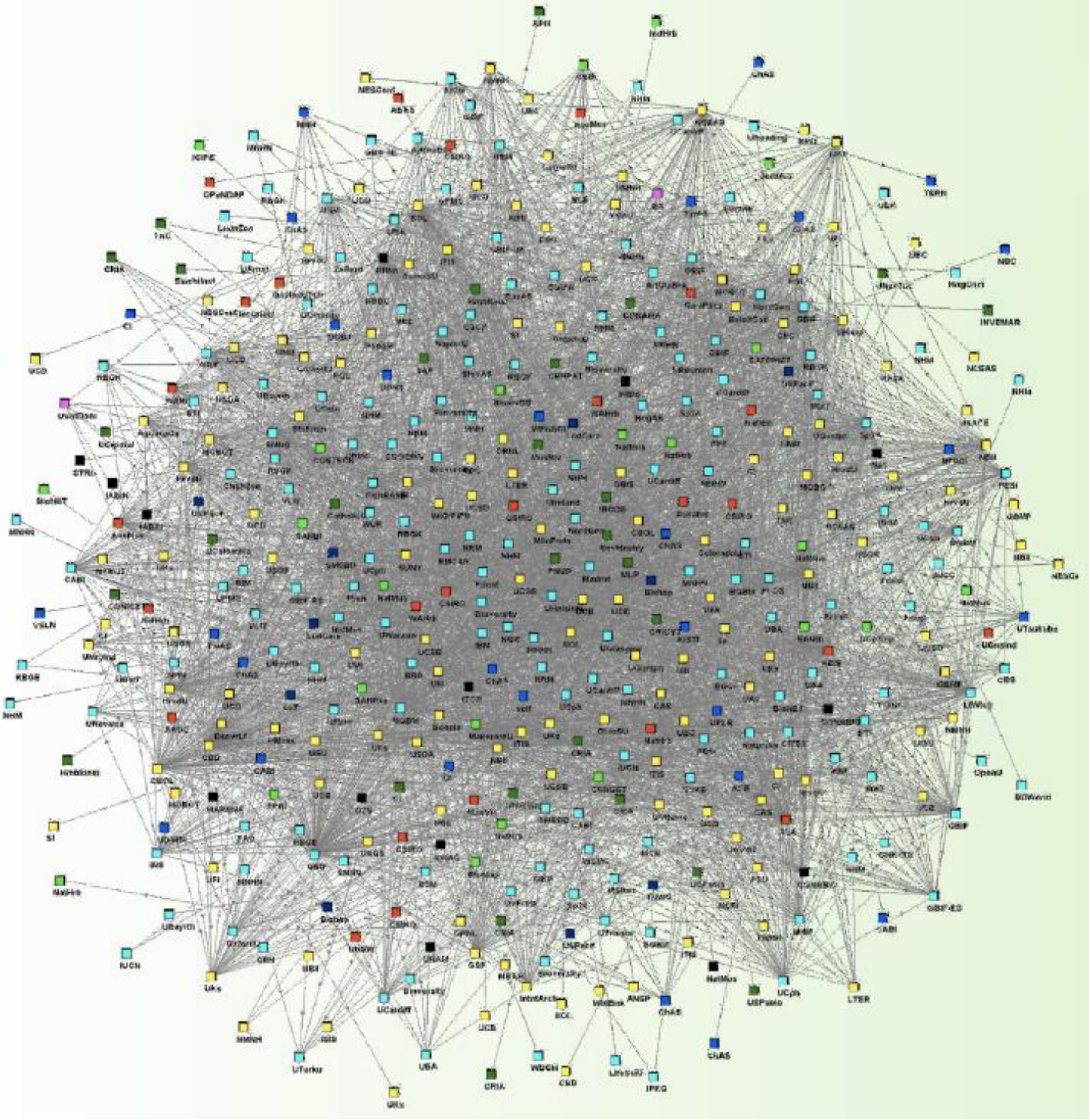


Figure 4. Network of collaborations of 446 individuals participating at the e-Biosphere meeting in London, 2008 (Meredith Lane, e-Biosphere meeting London, 2008).

3.2 – Survey on the current providers/users of biodiversity data

To reveal an up-to-date picture of the current situation on the biodiversity informatics landscape, we performed an online survey through a questionnaire (data available at: pro-iBiosphere Workshop Registration Questionnaire Responses). Participants that registered to the three pro-iBiosphere workshops were asked to fill in the questionnaire:

- Participants (66 in total) of the workshop on Coordination and routes for cooperation consisted of persons interested in specimen or observation data, standards, bibliographic references, digital publications, names, morphology, genes, media, annotations, and identifiers.

- Participants (46 in total) of the workshop on Requirements of Flora, Fauna, or Mycota publications or services consisted of a broad range of specialist and non-specialist providers and consumers of Floristic and Faunistic information, including (but not confined to) the following disciplines:
  - ✓ Taxonomy
  - ✓ Conservation
  - ✓ Ecology
  - ✓ Environmental Science (environmental modelling and trait analysis)
  - ✓ Education
  - ✓ General Public
  - ✓ Users and producers of e-Floras, e-Faunas, e-Mycotas
  - ✓ Users and producers of ID tools and services

- Participants (61 in total) of the workshop on Measuring and constraining the costs for delivering services consisted of (i) Developers/host institutions providing services/tools to the community and (ii) Projects, initiatives, organisations that use taxonomic content (e.g., aggregators of marked-up legacy and prospectively published literature, disseminators of atomised content, etc.).

The survey questionnaire consisted of 5 multiple choice questions (one of them required*), aims at receiving feedback from users:

1. Please indicate whether you are planning to attend the following workshop*
2. You are a user/provider/both of Flora, Fauna or Mycota information
3. Your main interests are in:
4. The information you use is:
5. What type of biodiversity information and/or data do you offer or use

The questionnaire also allowed the respondents to reveal details on the specifics of the services they provide through 12 open-ended questions (5 of them required*) listed below:

1. Is your database open access and if yes, since when?
2. How many users do you have?
3. What is the original source of the content you provide (project name, funding)? *
4. Who are the main sources (institutions, data suppliers, partners, etc.) of the content you provide? *

5. What are the formats and standards of the information/data you provide (please specify dataset format, with a link to documentation, if any)? *
6. What methods have been applied in data collecting (e.g., manual input, Web, harvesting, ftp uploads, protocols)? *
7. Are your databases or platforms currently using the services of other providers (machine-to-machine interoperability without human intervention, e.g., processing services or regular data exchange)? (please specify which services you use).
8. Are your services currently being used by other databases or platforms (machine-to-machine interoperability without human intervention, e.g., processing services or regular data exchange)? (please specify which of your services are being used).
9. How can the data be accessed/exported (through e.g., API, web services, automated exports, etc.)?
10. Please provide links to the API or web service documentation, if available.
11. What persistent identifiers have been used in your datasets? If known, please explicitly mention their type, persistence, dereferencing, resolvability.
12. What are the obstacles and what are your actual needs in making your database fully interoperable with other databases and platforms? *

67 participants completed the survey between the 5th of March and 14th of May 2013. The main findings are structured around the questions posed in the survey. The main findings are presented below. All data are available at: pro-iBiosphere Workshop Registration Questionnaire Responses**.**

1. Please indicate whether you are planning to attend the following workshop (Fig. 5)

More than one third, 39%, of the respondents planned to attend the “Coordination and routes for cooperation across organisations, projects and e-infrastructures” workshop. 30% of the respondents planned to participate in the “Measuring and constraining the costs of delivering services workshop”, and 29% of the respondents planned to attend the “Requirements of users for floristic and faunistic information and services” workshop. 2% of the respondents expressed that they were not able to attend the workshops, but filled in the questionnaire.

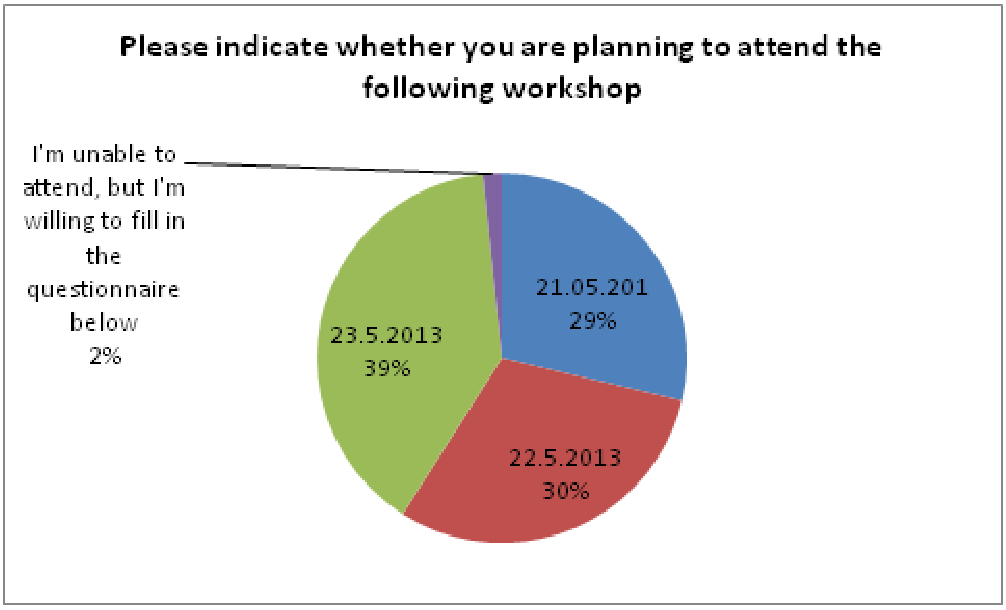


Figure 5. pro-iBiosphere Workshops attendance .

2. *You are a user/provider/both of flora, fauna or mycota information (Fig. 6)*

The majority part of the respondents (35%) is both using and providing flora, fauna and mycota information. 33% of the respondents are providers versus 20% only using this kind of information. 12% of the respondents did not answer the question.

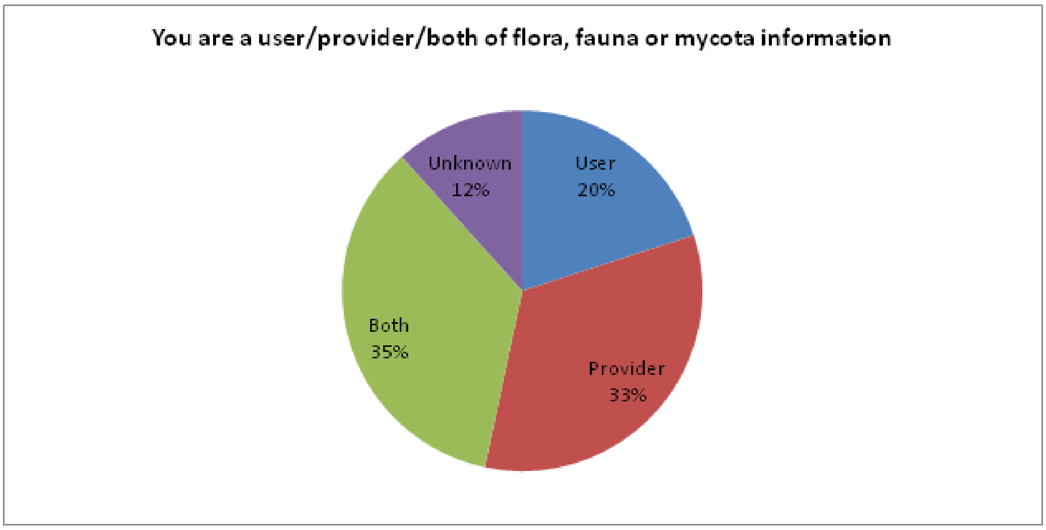


Figure 6. Users/providers/both of Flora, Fauna or Mycota information.

3. *Your main interests are in (Fig. 7):*

Half (51%) of all respondents are interested in Flora information. 21% of the respondents are mainly oriented towards Fauna information. Another 16% are looking for mycota information. 12% did not specify.

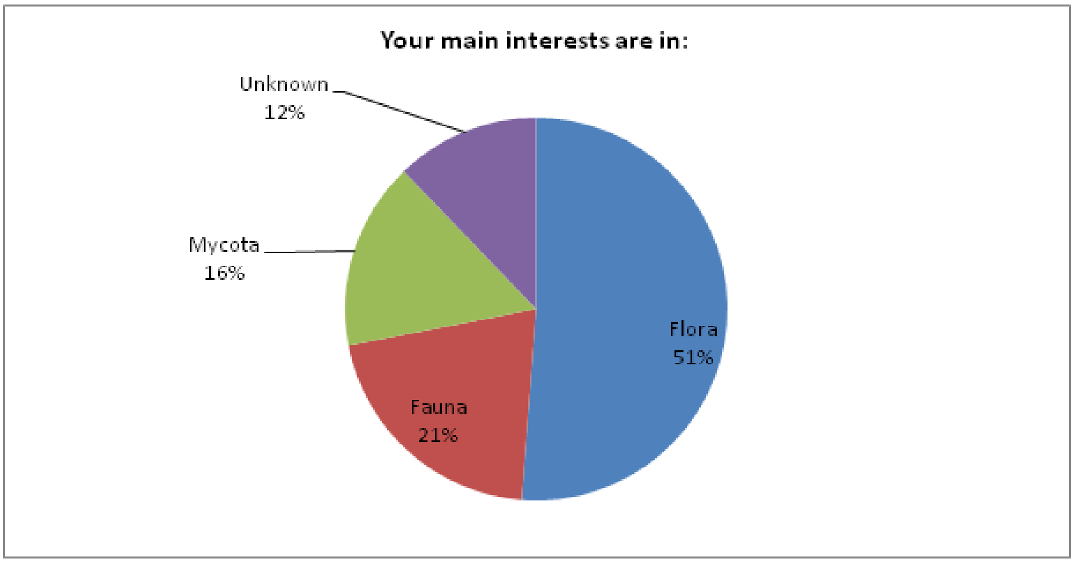


Figure 7. Main interests.

4. *The information you use is (Fig. 8):*

The prevailing part (54%) of the respondents use both digital and hardcopy information, 30% use digital information while an insignificant part (3%) of the respondents use only hardcopies. 13% did not answer the question.

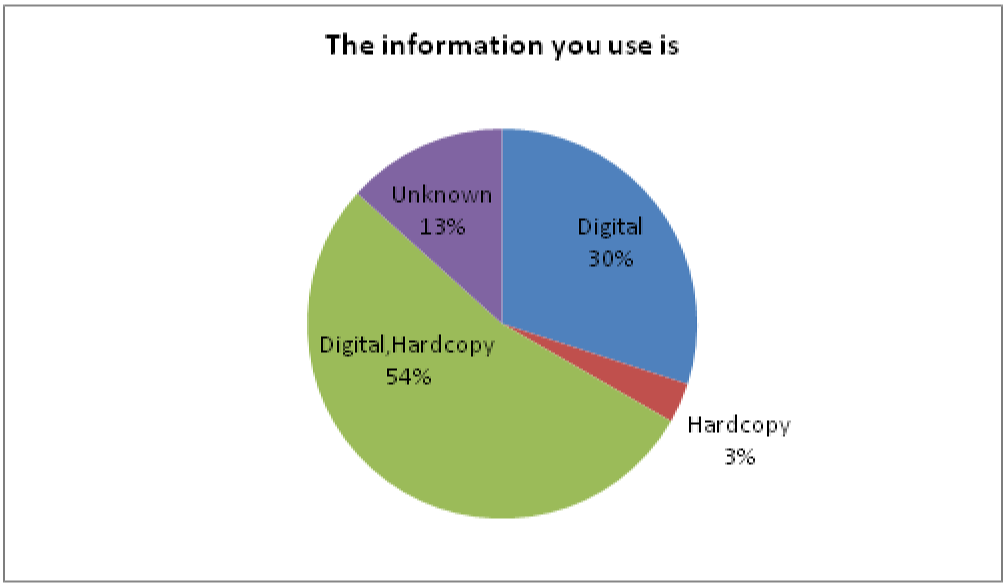
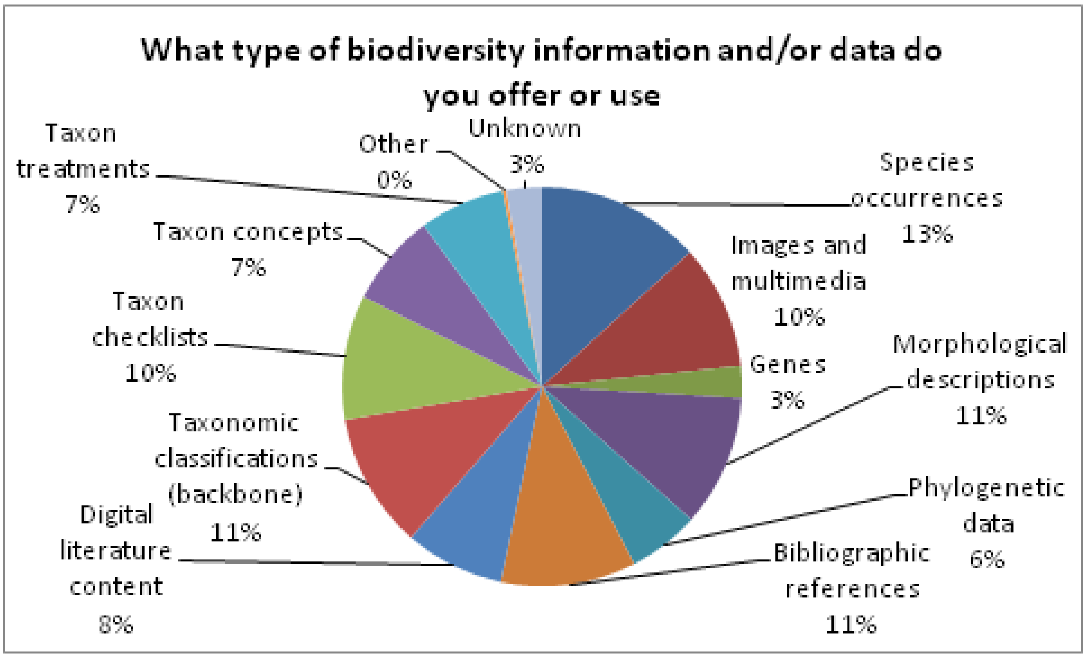


Figure 8. Type of information used.

5. *What type of biodiversity information and/or data do you offer or use (Fig. 9)*

There are no substantial differences in the percentage of the answers received. The most frequently offered/used information is the species occurrences (13%). Part of the other responses is shared between Morphological descriptions (11%), Bibliographic references (11%), Taxonomic classifications (backbone) (11%), Images and multimedia (10%) and Taxon checklists (10%). Another 28% is shared between Digital literature content (8%), Taxon treatments (7%), Taxon concepts (7%) and Phylogenetic data (6%). The less offered/used data is on the genes. 3% of the respondents did not answer the question.



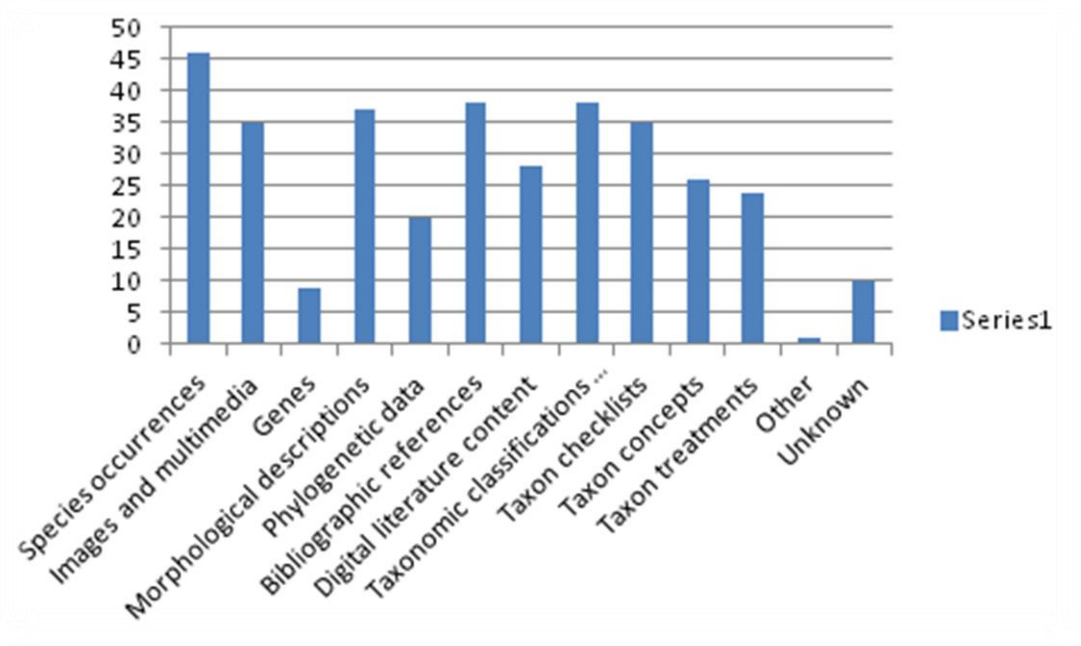


Figure 9. Type of biodiversity information and/or data offered/used.

The questionnaire also contains seven optional and five required open-ended questions encouraging respondents to share more information on:

- the type of database and number of users they have
- the original source of the content provided
- the main sources of the content provided
- the formats and standards of the information/data provided
- the methods applied in data collecting
- are their databases or platforms currently using the services of other providers?
- are their services currently being used by other databases or platforms?
- how can the data be accessed/exported?
- what persistent identifiers have been used in their datasets?
- what are the obstacles and what are their actual needs in making their database fully interoperable with other databases and platforms?

Respondents were asked to provide links to the API or web service documentation, if available. Half (51%) of the respondents reported open databases. Another 7% are partially opened databases. 2% answered negatively and 40% did not answer the question. The oldest database (i.e., of the CBS-KNAW Centraalbureau voor Schimmelcultures) is functional since 1990. The most recent one (of Max Planck Institute for Biogeochemistry) started functioning in May 2012 (Fig. 10).

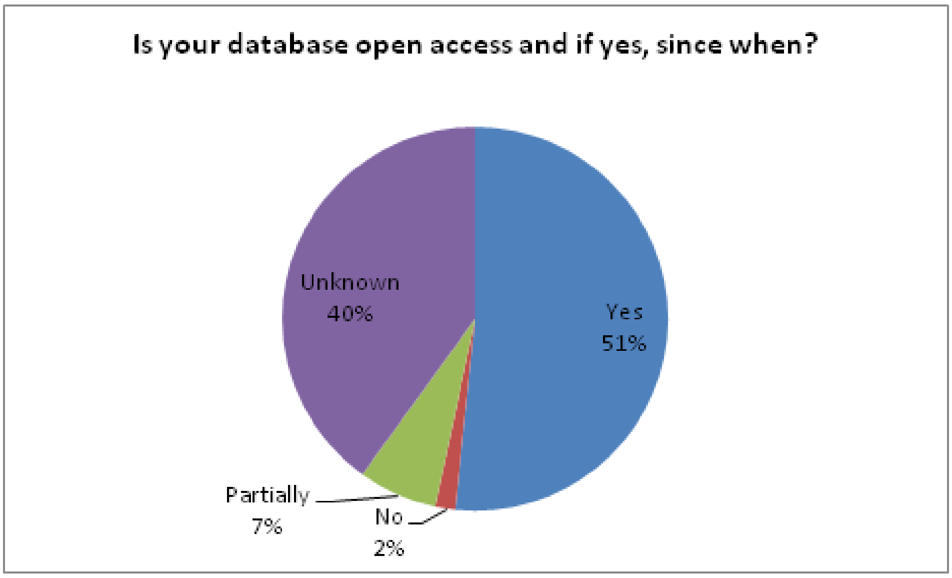


Figure 10. Type of database with regard to "openness".

The number of databases users varies considerably. The bigger number of databases users is reported by the representatives of the Missouri Botanical Garden (over 500,000 annually); department of Life Sciences, University of Trieste (ca. 4,000 page loads/day, ca. 600 unique visitors/day); Bath University (> 2,500 per month); Wikimedia community (500 million per month in total, of which perhaps some tens of millions looking up biodiversity information); Ecoflora (>2000 per month); RBG Kew (3,000 page hits per day); Museum für Naturkunde Berlin (2000/month) and University of Reading (>1.000.000). 80% of the respondents did neither respond nor specified a number.

30% of the databases/platforms are using the services of other providers (machine-to-machine interoperability without human intervention, e.g., processing services or regular data exchange) versus 22% that are not using them. Almost half (48%) of the respondents did not give an answer (Fig. 11).

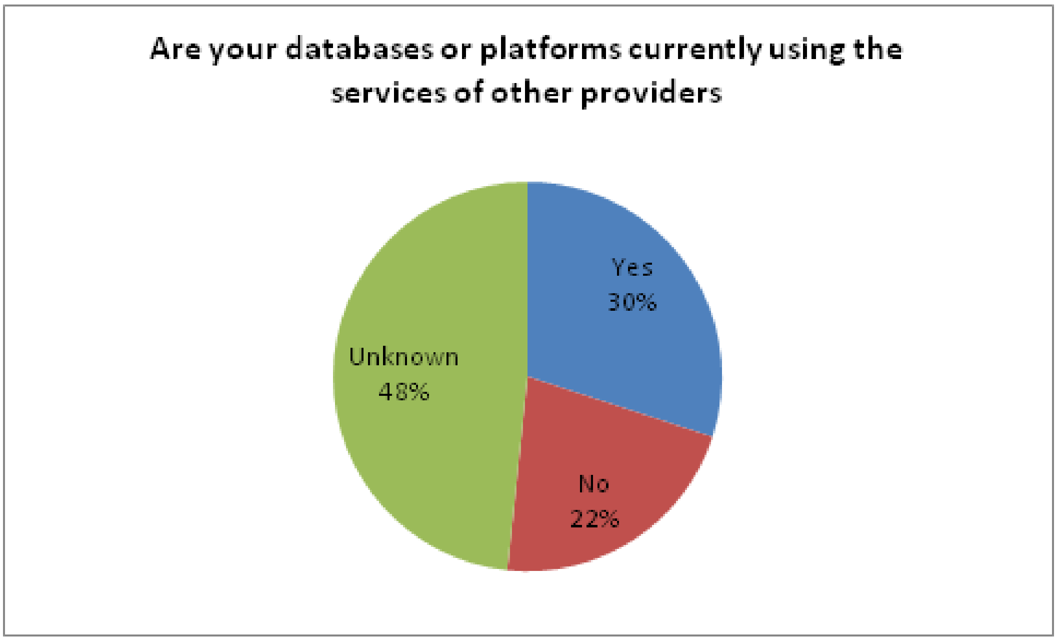


Figure 11. Use of other providers' services.

Over quarter of the respondents (28%) stated that the services they provide are currently being used by other databases or platforms. Another 20 % responded that their services are not being used by other databases or platforms. 52% remain unknown (Fig. 12).

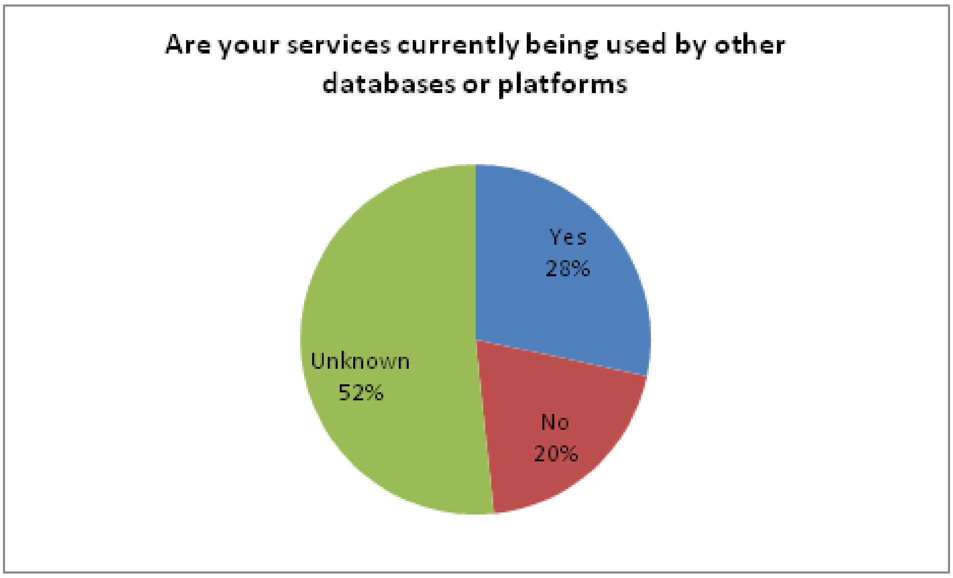


Figure 12. Services used by other databases/platforms.

44% of the respondents confirmed that the data they provide can be accessed/exported versus 3% responded negatively. 53% did either not answer the question or not sure about the response (Fig. 13).

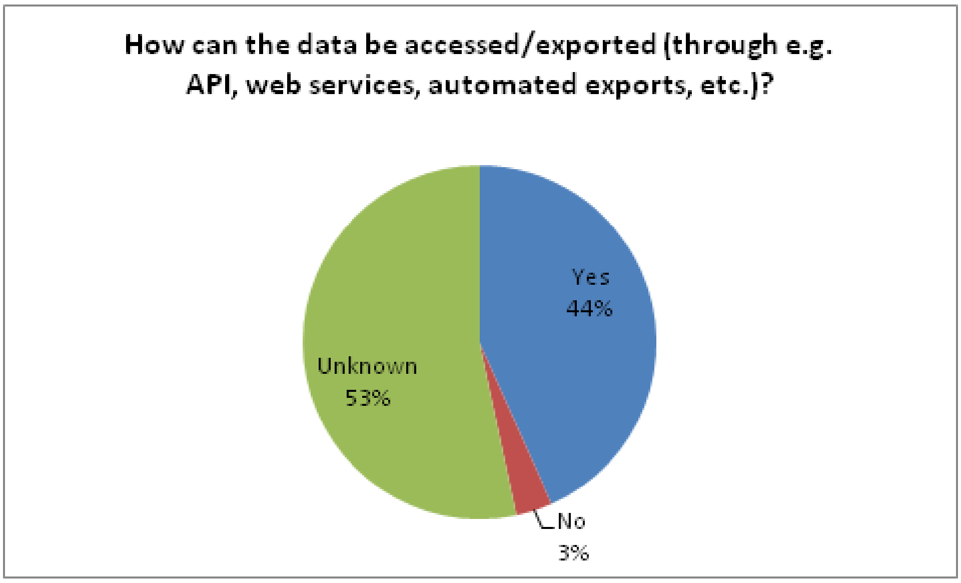


Figure 13. Accessibility/export of data.

The main persistent identifiers used in the datasets provided are as follows: LSID (8%), DOI (5%) and UUID (5%). 15% of the respondents use other identifiers such as binomials, MycoBank numbers, Internal code numbers, URLs, etc. 8% of all respondents does not use any persistent identifiers. 59% did not answer the question (Fig. 14).

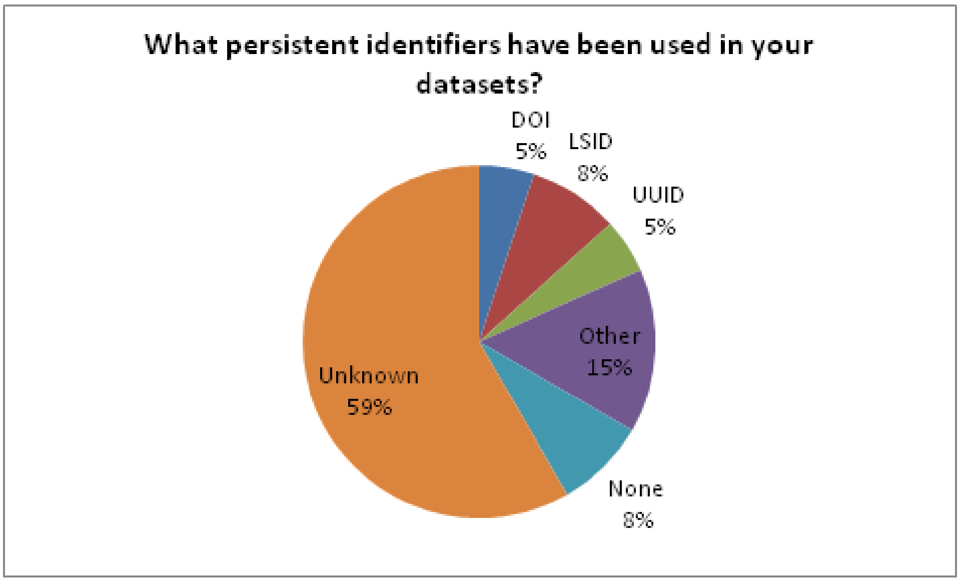


Figure 14. Persistent identifiers used in datasets.

The following main conclusions have been drawn from the survey:

1. There is an obvious trend towards the use of digital resources. The prevailing part is that the respondents use both digital and hardcopy information, 1.3 use digital information while 3% uses only hardcopies.
2. There are no substantial differences in the percentage of users of the various data elements within taxon treatments. The most frequently offered/used information is species occurrence data; most other responses are divided between Morphological descriptions, Bibliographic references, Taxonomic classifications, Images and multimedia and Taxon checklists. Another 1/3 is shared between Digital literature content, Taxon treatments, Taxon concepts and Phylogenetic data. The less offered/used data is on the genes, probably due to the fact that the sampled community consisted mostly of biodiversity specialists and few focused on genomics.
3. One third of the databases/platforms are using the services of other providers (machine-to-machine interoperability without human intervention, e.g., processing services or regular data exchange) versus 22% that are not using them.
4. The digital resources of the participants are, as is to be expected, widely used, but also at various degrees.
5. Almost one half of the respondents confirmed that the data they provide can be accessed/exported versus 3% that responded negatively.
6. The main persistent identifiers used in the datasets provided are LSID (8%), DOI (5%) and UUID (5%).

3.3 – Towards a set of Best practices to facilitate permanent digital identification for specimens

Recently, the desire to participate in the Semantic Web and Linked Open Data has caused new interest in modern alternative identifiers for natural history collection specimens. In 2012 the Royal Botanic Garden Edinburgh (RBGE) published a paper (Hyam et al.); see also Stable Citations for Herbarium Specimens on the Internet) on using the Linked Data principles (http://www.w3.org/TR/cooluris/) to issue HTTP URIs (URLs) for their specimens.

The pro-iBiosphere project was instrumental in furthering this discussion by addressing the issue in depth during both the Leiden (02-2013) and Berlin (05-2013) workshops and developing a best practices document (http://wiki.pro-ibiosphere.eu/wiki/Best_practices_for_stable_URIs). For additional information on this subject please see the pro-iBiosphere report on “Towards Best Practices Guide on Editorial Policies”.

The system proposed by the Royal Botanical Garden Edinburgh (Hyam et al. 2012) allows to implement a consistent identifier system for collections held by CETAF institutions (Consortium of European Taxonomic Facilities). It is based on HTTP URIs and allows a clear distinction between physical collection objects and their associated metadata. It can provide both human and machine-readable object representations.

In a “stable identifier hackathon” in Edinburgh (06- 2013), five CETAF institutions (Royal Botanical Garden Edinburgh, Museum für Naturkunde in Berlin, Royal Botanic Garden Kew, National Museum of National History Paris and Botanical Museum Berlin-Dahlem) committed to a rapid pilot implementation of the system. Naturalis Biodiversity Center in the Netherlands also plans to join this effort (http://stories.rbge.org.uk/archives/3846).

During the pro-iBiosphere workshop on “How to improve technical cooperation and interoperability at the e-infrastructure level”, to be held in the week of the 8th to 11th October 2013 a test application to confirm that these five institution’s systems are working will be developed. Furthermore a demonstration on applications to show how the system can be used will be organised.

At present, Best Practices (see here) for the use of stable URIs are being discussed and developed by pro-iBiosphere. As we demonstrate the value in this way of working, the pro-iBiosphere consortium aims to encourage other institutions to adopt HTTP URIs for their specimens. The annual meeting of TDWG Biodiversity Information Standards that will that will take place in Florence in November 2013 and future pro-iBiosphere workshops may be a suitable platforms for this.

The URI approach will be expanded to other domains like treatments or images within the pro-iBiosphere pilot studies (Task 4.2), and through a dialog between DataCite and our community on creating a subdomain of DOIs for observation data.

Zoobank will offer this approach as an alternative to its present LSID-based system. We expect other institutions in the biodiversity domain to follow. We hope that the system can serve as a blueprint for additional biodiversity informatics object types beyond collections.

4 – Standards used for treatments and their data elements

At present there are several different XML schemas and Document Type Definitions (DTD) (in the text, schema refers to both, unless specifically mentioned) being used for the markup of taxonomic literature (see Penev et al. 2011), of which the most widely used ones are briefly outlined in this report. The different schema designs reflect different priorities and consequently criteria for development. One distinction is whether the focus of the markup is on structure of the document as a whole (document-centric) or some part of the content of the document (content-centric). Another is the extent to which the marked-up text is potentially interoperable with (or using common elements with) other implementations. Notably, even with these distinctions, convergences are developing between different approaches.

An example of the content-centric approach is a focus on morphological descriptions (Heidorn et al. 2002; Cui and Heidorn 2007; Cui 2008a, b; 2010a, b). In their work, the publication is viewed more as metadata and the emphasis placed on the detail of morphological terms and the potential or repurposing the content. In this, the markup approaches SDD (Hagedorn et al. 2005), a schema produced explicitly for descriptive data.

A little broader is the concept of TaxonX, a content-centric schema developed to mark up all taxonomic relevant semantic elements in a publication, without keeping the structure beyond paragraphs, and at the same time importing existing schemas (Sautter et al., 2007, Agosti & Egloff, 2009). To cover all the legacy literature, it is a very loose schema. A more stringent version that includes structural elements as well is Taxpub DTD. It is a semantic extension of one of the most widely applied publishing and archiving schema (JATS) by the US National Library of Medicines. It is a self-contained schema that does not rely on other schemas like TaxonX, but the respective elements can be mapped (Catapano, 2010; Penev et al., 2012). At the other extreme, some projects have employed a very generic schema to contain the document and structural information (i.e., pages, paragraphs, lines, headings, etc.) and used particular elements of taxonomic texts to assist in markup, relying on repeatable structural components of taxonomic descriptions (for example distributions, taxon names, morphological descriptions, stratigraphic detail, etc.) (Kirkup et al. 2005; Curry and Connor 2007, 2008).

In addition, we add short descriptions of some other schemas that, despite not being designed for treatments, are widely used or are expected to become widely used for treatments data elements, such as occurrences, taxon names, taxon concepts, biodiversity-related multimedia, etc.

4.1 – TaxonX

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

Description: TaxonX is an XML schema for encoding taxonomic literature in order to:

- Create open, stable, persistent, full text digital representations of taxonomic treatments.

- Make taxonomic treatments and their major structural components identifiable to enable networked reference and citation.
- Make identifiable lower level textual data such as scientific names and localities (Darwin Core or any other relevant schema may be used), morphological characters, and bibliographic citations in order to facilitate their extraction by, and integration with, external applications and resources.
- Study and describe the structure of systematics publications by creating some typical corpora of literature, such as entire journals (e.g., AMNH Novitates, Zootaxa), taxa (e.g., all ant systematics papers after 1995), or faunistic studies (e.g., all ant systematics paper covering Madagascar ranging from 1758 to 2011).

TaxonX is a lightweight (with only 30+ elements) and flexible schema for markup of treatments which can be quickly learned and may be applied to the wide variety of formatting present in legacy documents as well as new publications. Plazi has, for example, used it to encode nearly 24000 treatments from over 1900 publications drawn from nearly 300 different journals and books dating as far back 1758. In many cases the TaxonX relies on use of external schemas for modelling certain kinds of information [e.g., the use of MODS (Metadata Object Description Schema: http://www.loc.gov/standards/mods/) for file level bibliographical metadata; Darwin Core for observation data: http://rs.tdwg.org/dwc/]. It has loose content requirements that allow for a wide variety of instances to be encoded over time and at many levels of granularity, while maintaining validity through iterations. Additionally, TaxonX contains mechanisms for semantic normalisation of the data contained in treatments.

4.2 – TaxPub

Sources: http://sourceforge.net/projects/taxpub/; Catapano 2010; Penev et al., 2012

Description: TaxPub was designed with the aim to enable the markup of new "born-digital" taxonomic literature that could forgo unnecessary variation in style and form and adhere to a limited set of data elements so as to lower costs of both authoring and processing. TaxPub is an extension of the Journal Publishing Tag Set of the U.S. National Library of Medicine's Journal Archiving Tag Suite (see http://dtd.nlm.nih.gov/). For more details see Catapano (2010). Starting in 2008, TaxPub was designed and developed by members of Plazi with the assistance of experts from the U.S. National Center for Biotechnology Information and Pensoft Publishers (see Penev et al. 2012 for detail). The TaxPub extension is maintained as an open source project at SourceForge (http://sourceforge.net/projects/taxpub/) inheriting from the base DTD an extensive and robust set of elements for generic textual structures while adding a small number of elements relevant to taxonomy. These include elements for markup of taxon names, citations to specimens and other material, and statements describing morphology, as well as for treatments and treatment sections. Further semantics may be applied to many elements through use of terms in external vocabularies (such as Darwin Core) as values of attributes (more details in Catapano 2010 and http://species-id.net/wiki/TaxPub).

TaxPub, being an extension of the National Library of Medicine's Journal Article Tag Suite (JATS), has the additional advantage of facilitating archiving in PubMedCentral, one of the most secure existing archives and, as a consequence, its content is cross-linked with the huge body of biomedical literature stored therein. Starting in 2010 the Pensoft has been submitting its TaxPub encoded articles from its journals ZooKeys and PhytoKeys to PubMedCentral, with nearly 1300 articles contributed to date. A description of the applications of TaxPub in the Pensoft journal production workflow is provided by Penev et al. (Penev et al., 2012).

4.3 – Plazi Internal Generic XML

Source: http://www.plazi.org/?q=GoldenGATE

The XML used in the Plazi production workflow and GoldenGate editor is explicitly loose (not based on schema or DTD) to allow markup of any elements in a text, and it allows crossover of elements. For example it the document is treated for different purposes, such as ecology and taxonomy, particular elements can be assigned different markup that overlap. XSLT transformations will then create valid XML, for example TaxonX. This generic markup is not supposed to be exposed to the human reader, nor to applications outside of the Plazi repository. However, the elements listed below are created so that they can be mapped to external schemas like Darwin Core.

Element	**Purpose**	**Parent Element(s)**	**Attributes**	**Added in Step (Removed in Step)**
document	Root element	-	docAuthor, docTitle, docDate, docOrigin, pageNumber, lastPageNumber, docSource, ID-*	Loading, attributes during Document Meta Data Import
pageBorder	Page breaks in HTML	document	-	Loading HTML (HTML Normalisation)
pageNumber	Markup of page numbers	paragraph	value, score, ambiguity, fuzziness	added temporarily during HTML Normalisation
pageTitle	Markup of page headings	document, paragraph	-	
page	PDF layout structure	document	pageId, pageNumber, box, imageName, imageDpi	Loading PDF (PDF Normalisation), added temporarily during HTML Normalisation
block		Page	box	Loading PDF (PDF Normalisation)
column		page, block		
paragraph	Markup of document paragraphs, both layout and logical	page, block, column (layout) document, subSection, treatment, subSubSection,	pageId, lastPageId, pageNumber, lastPageNumber, box, indentation, textOrientation, type	Loading (layout), Normalisation (logical)

Element	Purpose	Parent Element(s)	Attributes	Added in Step (Removed in Step)
		caption, footnote, keyStep (logical)		
line	PDF layout structure	Paragraph		Loading PDF (PDF Normalisation)
word		paragraph, line, td	box, bold, italics, fontSize, baseline, ocr, str	
subSection	Logical document structure	document	type	Treatment Markup
treatment		document, footnote	-	
subSubSection		treatment	type	Treatment Structuring
table	Representation of tables	document, paragraph	-	Loading HTML, PDF Normalisation
tr		Table	-	
td		Tr	isEmpty, colspan, rowspan	
caption	Markup of captions	document, subSection, treatment, subSubSection	imageUrl	Normalisation, bibRef attributes during Bibliography Parsing
footnote	Markup of footnotes			
bibRef	Markup of bibliographic references	paragraph	author, editor, year, title, volumeTitle, journalOrPublisher, partDesignator, pagination type	
keyStep	Markup of keys	subSection, subSubSection	-	Key Structure Markup
keyLead		keyStep, paragraph	-	
pageBreakToken	Markup of first token in a page	Any	pageId, pageNumber, start	Normalisation
normalisedToken	Markup of tokens with normalized diacritics		originalValue	
misspelling	Markup of possibly misspelled words or possible OCR errors	Any		added temporarily during Spell Checking
correctedMisspelling	Markup of corrected misspellings,		originalValue	Spell Checking

Element	Purpose	Parent Element(s)	Attributes	Added in Step (Removed in Step)
	retaining the original text			
mods:any	Document meta data in MODS XML	document	-	Document Meta Data Import
author	Markup of bibliographic reference details	bibRef	-	Bibliography Parsing
editor				
year				
title				
volumeTitle				
journalOrPublisher				
partDesignator				
pagination				
publicationUrl				
bibRefCitation	Markup of citations of bibliographic references in main text	paragraph	All attributes of cited bibRef	Citation Tagging
taxonomicName	Markup of taxonomic names	paragraph, keyLead	rank, genus, species, etc., authority, authorityName, authorityYear	Taxonomic Name Tagging
taxonomicNameLabel	Markup of type status specifiers	paragraph, keyLead	rank	
date	Markup and normalisation (to yyyy-mm-dd) of dates and date ranges	paragraph, materialsCitation	value, valueMin, valueMax	Materials Citation Markup
geoCoordinate	Markup and normalisation (to signed decimal) of geographical coordinates		value, orientation, direction	
quantity	Markup and normalisation (to metric units) of arbitrary quantities (lengths, areas,		value, unit, metricValue, metricMagnitude, metricUnit	

Element	Purpose	Parent Element(s)	Attributes	Added in Step (Removed in Step)
	volumes, weights, etc.)			
country	Markup and normalisation (to contemporary English names) of country names		name	
typeStatus	Markup of type status information			
collectionCode	Markup of collection codes		collectionName	
materialsCitation	Markup of materials citations	paragraph	country, stateProvince, municipality, location, longitude, latitude, elevation, collectingDate, collectingDateMin, collectingDateMax, collectorName, collectingMethod, specimenCount, specimenCode	Materials Citation Markup, attributes during Materials Citation Parsing
collectingCountry	Markup and normalisation of materials citation details	materialsCitation, paragraph (for global variables not given in individual materials citations, but in introduction or methodology)	name	Materials Citation Parsing
collectingRegion			-	
collectingMunicipality				
locationDeviation				
location			longitude, latitude, country	
elevation			value, unit, metricValue, metricMagnitude, metricUnit	
collectingDate			value, valueMin, valueMax	
collectorName			-	
collectingMethod				
specimenCount				
specimenCode				

4.4 – taXMLit

Source: http://wiki.tdwg.org/twiki/bin/viewfile/Literature/WebHome?rev=1;filename=taXMLit_v5-04.xsd; Weitzman and Lyal (2004)

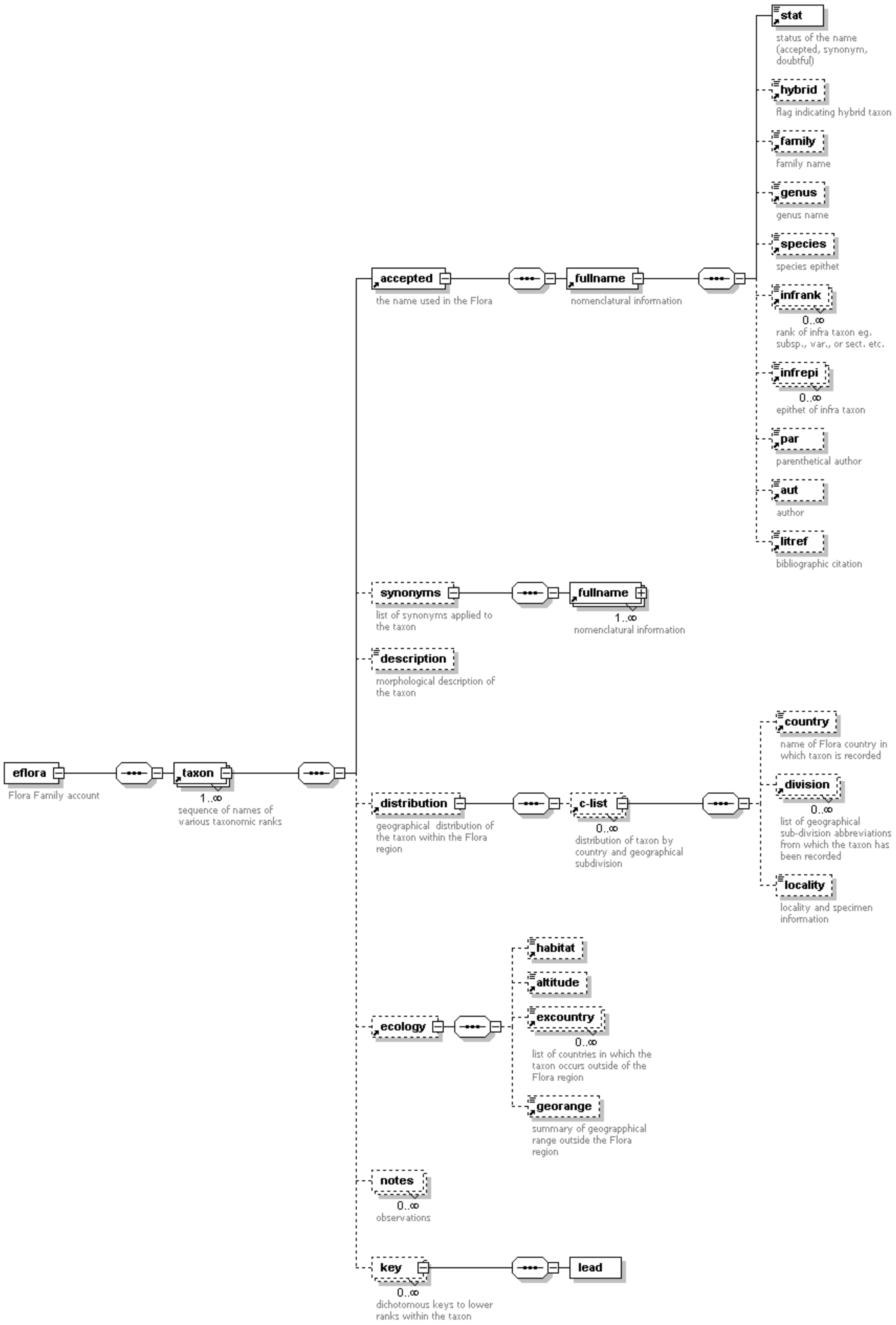
Description: The taXMLit schema is designed to accommodate taxonomic literature. It was developed particularly in the context of Zoological and Botanical publications and should also be applicable to publications on fungi and paleontology, although this has yet to be tested. The schema does not take into account the kinds of data needed for viral or bacterial publications. It covers all of the components of taxonomic publications and the taxon treatments contained within them, but it does not encode individual character statements, which are dealt with by other projects such as SDD.

The schema is highly atomised, permitting both recovery of publication components (e.g., taxon treatments, diagnostic keys, images, bibliographic entries, discussion paragraphs) and of data within those components, such as specimen data, biological associations, atomised taxonomic names, and nomenclatural and taxonomic acts. It can be applied to the entire text of a publication and not only to formal treatments as discussed above. The richness permits full application to any legacy format so far encountered. The full taXMLit contains data elements extracted from the text that permit detailed data querying, browsing, and download; a version that does not include the respective elements and is more document-centric has also been developed ('taXMLite':). This was developed to permit preliminary markup and subsequent upload access through the INOTAXA interface developed for taXMLit (see below); it is not discussed further here.

Implementation of the schema in an appropriate system ('INOTAXA' – has been designed for this purpose) allows the text of marked-up taxonomic publications to be fully humanly searchable. In INOTAXA users may choose to view and download data (e.g., taxonomic names, specimen data, citations, biological association data, persons' names) for use in analysis or other applications, or access taxon treatments, keys, images, or other content components as reference resources. In conjunction with the appropriate system, the schema would also facilitate static links from the text to other data sources (e.g., specimen databases on the web, ZooBank). Use of the schema for multiple taxonomic works allows these to be searched or browsed simultaneously, and permits links between different works that cover the same taxa or their synonyms. Moreover, this paves the way for users to create virtual compilations of taxon treatments, comprising components of more than one original work, e.g., checklists, Faunas, and Floras. These applications require that the schema should, in the appropriate parts, use elements the same as or similar to those in schemas used by other relevant systems, and be mappable to them.

4.5 – XML schema used at Royal Botanic Gardens Kew (RBGK)

In 2000 RBGK developed a simple approach to mark up of its legacy African Floras using word processor styles, macros and xslt stylesheet transformations (Kirkup et. al 2005). The final markup schema (markup table and schema diagram both shown below) resulted from a combination of the information requirements for nomenclatural and geographical based queries and a pragmatic approach to capture that required information, any other information that was easy to extract ("low hanging fruit"), but not an attempt to atomise the whole work. The resulting structure also traces its heritage to the SPICE for SPECIES 2000 project and the structure of the wrappers for RBGK's world checklist databases.


stat
status of the name (accepted, synonym, doubtful)
hybrid
flag indicating hybrid taxon
family
family name
genus
genus name
species
species epithet
accepted
the name used in the Flora
fullname
nomenclatural information
infrank
0..∞
rank of infra taxon eg. subsp., var., or sect. etc.
infrepi
0..∞
epithet of infra taxon
par
parenthetical author
aut
author
litref
bibliographic citation
synonyms
list of synonyms applied to the taxon
fullname
1..∞
nomenclatural information
description
morphological description of the taxon
country
name of Flora country in which taxon is recorded
eflora
Flora Family account
taxon
1..∞
sequence of names of various taxonomic ranks
distribution
geographical distribution of the taxon within the Flora region
c-list
0..∞
distribution of taxon by country and geographical subdivision
division
0..∞
list of geographical sub-division abbreviations from which the taxon has been recorded
locality
locality and specimen information
habitat
altitude
ecology
excountry
0..∞
list of countries in which the taxon occurs outside of the Flora region
georange
summary of geographical range outside the Flora region
notes
0..∞
observations
key
0..∞
dichotomous keys to lower ranks within the taxon
lead

This markup schema was originally applied to the previously published parts of Flora Zambesiaca, then later (with minor modifications) to the Flora of Tropical West Africa, the Flora of Tropical Africa, Flora Capensis, Flora of Somalia and the Flora of Tropical East Africa. The schema is currently used within the Flora Zambesiaca editorial process in order to provide marked-up versions of new accounts as they are published. The approach is flexible, with the schema having also been adapted for economic botany works (The Useful Plants of the Flora of Tropical West Africa) and monographs (e.g., Polhill & Weins, The Mistletoes of Africa).

Paragraph styles	**Character style within paragraph styles**	**Style description**	**Macros used?**
Accepted		The paragraph which contains the accepted name of a taxon	Yes
	sp-id	The identification number of a species as it occurs in the Flora	Yes
	genus	The genus name	Yes
	species	The species epithet	Yes
	infrank	Contains the words variety, subspecies or forma	Yes
	infrepi	Contains the infra specific epithet	Yes
	par	Parenthetical author	Yes
	aut	The taxon author	Yes
	litref	The piece of literature in which the taxon name was first given and all the relevant references	Yes
Synonym	(character styles as Accepted)	Contains synonyms of the accepted taxon	Yes
Description		General description of the taxon	Yes
Distribution		Information on the distribution of the taxon within FZ	Yes
	country	The countries within the FZ that the taxon occurs in	Yes
	division	Division of an FZ country where the taxon occurs	Yes
	locality	Locality information, collector name and date of collection	Yes
Ecol		Paragraph containing ecological information of the species	Yes
	range	Ranges of countries outside the FZ region over which taxon occurs and level of endemism	No

Paragraph styles	Character style within paragraph styles	Style description	Macros used?
	habitat	Habitat information of the taxon	Yes
	excountry	Countries outside FZ where the taxon occurs	Yes
	altitude	Altitude information	Yes
Notes		Any other information that is included	Yes
Doubtful	(character styles as Accepted)	Doubtful names of accepted species	No
Dubious	(character styles as Accepted)	Dubious names of accepted species	No
Acceptedhybrid	(character styles as Accepted)	The paragraph which contains the accepted name of a hybrid taxon	No
Synonymhybrid	(character styles as Synonym)	Hybrid which is synonym of the accepted taxon	No

The resulting markup differs slightly between the different Flora publications but these are transformed (using xslt) into a common database structure which has a web query interface http://apps.kew.org/efloras/advsearch.do?reset=true.

Continued work on schema development has focused on the atomisation of geographical and morphological information for identification tools and trait databases.

4.6 – XML schema used at Naturalis Biodiversity Center (Naturalis)

The XML schema co-developed by Naturalis and FUB-BGBM is based on the RBGKew method. The schema (still under design) uses a data-driven approach, by analysing the structure of these works, and identifying the types of contents they have and where these can occur in a taxonomic work. It divides a taxonomic work into four different content types: metadata, taxonomic content, non-taxonomic content, and errata to previous volumes, each of which is subdivided further. The most important type, taxonomic contents, can be subdivided into keys, nomenclature, references and descriptions as well as a variety of other features. A further possible subdivision, atomisation, of certain content is required for certain purposes, such as the creation of interactive multi-access keys. At present, full atomisation of many types of contents is being implemented, including nomenclature, taxonomic specimens and types, literature references and citations, distributions, and descriptions (down to character level). Furthermore, predispositions have been taken enabling even further atomisation if required, including that of text not normally atomised.

4.7 – XML schema used by the National Botanic Garden of Belgium (NBGB)

For the digitisation on the Flore d'Afrique Centrale, the National Botanic Garden of Belgium used a series of bespoke XML schemas during the markup process. XML was used as intermediate format, with the finished data stored in an fully atomised database. The advantage of using a series of XML custom schemas was that automated processes could be used in the markup procedure and then validated against the schema. Once a document was compliant to an initial schema, it could be further processed for compliance with successively more atomised schemas. This stepwise approach has many advantages for the rapid, automatic markup of text. The final document may be a standard XML schema or a database containing the same data.

4.8 – Darwin Core

Source: http://rs.tdwg.org/dwc/

Description: The Darwin Core (DwC) standard facilitates the exchange of information about the geographic location of organisms and associated collection specimens between different databases.
DwC contains elements that cover specimen information, taxonomic classification, specimen identification, locality details, collecting event information (who, why where, when, how), biological data about the specimen and references such as images.

The capabilities of DwC can be enhanced through extensions to the core schema to meet specific needs. For example, there are extensions to the schema for specimen curation, geospatial location, paleontology and interaction between specimens. All of these extensions go beyond the core to answer questions about specimens other than those made possible by the core alone. This extensibility of DwC is one of its strengths.

The TDWG Access Protocol for Information Retrieval (TAPIR) can use DwC (and/or ABCD) to enable data discovery, search and retrieval across multiple organisations and databases.

The DwC data standard was originally produced by the Species Analyst project at the University of Kansas. DwC was then redeveloped in tandem with the Distributed Generic Information Retrieval (DiGIR) protocol (now superseded by TAPIR). DwC is used across the GBIF network and projects like MaNIS, HerpNet, OrNIS, FishNet2, OBIS and the PaleoPortal standards.

4.9 – ABCD - Access to Biological Collections Data

Source: http://www.bgbm.org/tdwg/codata/schema/
Description: ABCD - Access to Biological Collections Data - Schema is a common data specification for biological collection units, including living and preserved specimens, along with field observations that did not produce voucher specimens. It is intended to support the exchange and integration of detailed primary collection and observation data.

All of the world's biological collections contain a number of data items including specimen specific (e.g., taxon, altitude, sex) and collection specific (e.g., holding institution) elements. The set of elements used varies from collection to collection. ABCD provides a reconciled set of element names and their definition for scientists and curators to use. It is not expected (or even possible) for any collection to use more than a fraction of the elements defined in the standard.

A design goal of the data specification was to be both comprehensive and general, to include a broad array of concepts that might be available in a collection database, but to mandate only the bare minimum of elements required to make the specification functional. ABCD deliberately does not cover taxonomic data, such as synonymy, other than the use of names in identifications. Likewise, taxon-related information, such as distribution range, indicator values, etc., is also not included. The elements and concepts that are used provide as much compatibility as is possible with other standards in the field of biological collection data, such as HISPID, Darwin Core, and others.

Unlike DwC, ABCD aims to cover all potential data fields and therefore facilitate the exchange of a very wide range of data between international databases. DwC is therefore simple by comparison with ABCD. All DwC elements do however map to ABCD counterparts.

The data specification is cast as an XML schema.

4.10 – CDM - EDIT's Common Data Model

Source: http://dev.e-taxonomy.eu/trac/wiki/CommonDataModel

Description: The Common Data Model (CDM) is the domain model for the core components of the EDIT platform for Cybertaxonomy. The CDM includes the data items in the TDWG ontology and builds on relational and object-relational information models developed for the taxonomic domain, especially the Berlin Taxonomic Information Model33, the BioCASE Model for Specimen and Observation Data34, the CATE35 object model and the descriptive data definitions as laid out in TDWG's SDD standard (see 2.4.5).

The CDM is a normalised object-based format developed in UML, covering the entire taxonomic information flow from fieldwork to printed and electronic publication. The CDM acts also as an information broker for existing biodiversity informatics applications such as descriptive tools, taxonomic database systems as well as specimen and observation management systems. Interfacing between external applications and CDM, data stores are implemented primarily using TDWG XML-based standards such as TCS, ABCD, and SDD as well as its own CDM/XML format. RDF exports and imports will be implemented at a later project stage. The import of TCS/RDF has already been tested successfully.

4.11 – Taxon Concept Schema (TCS)

Source: http://www.tdwg.org/standards/117/

Description: The TCS schema was conceived to allow the representation of taxonomic concepts as defined in published taxonomic classifications, revisions and databases.

TCS presents an abstract model for a taxonomic concept, which can capture the various models represented and understood by the various data providers. It is presented as an XML schema document that is proposed as a standard to allow exchange of data between different data models. It aims to capture data as understood by the data owners without distortion, and facilitate the query of different data resources according to the common schema model.

As such, it specifies the structure for XML documents to be used for the transfer of defined concepts. Valid transfer documents may either explicitly detail the defining components of taxon concepts; transfer GUIDs referring to defined taxon concepts (if and when these are available) or a mixture of the two.

The TCS schema is not designed to facilitate the exchange or documentation of information about Taxon concepts where this information is not part of a taxonomic revision creating new concepts. The amount and variety of (additional) information that can be potentially assigned to concepts is outside the scope of a taxonomic concept transfer schema, but we would encourage the development of domain specific models that use or extend this schema. XML supports this flexibility by allowing the use of different name spaces.

4.12 – Audubon Core

Source: http://terms.gbif.org/wiki/Audubon_Core_Term_List

Description: The Audubon Core is a set of vocabularies designed to represent metadata for biodiversity multimedia resources and collections. These vocabularies aim to represent information that will help to determine whether a particular resource or collection will be fit for some particular biodiversity science application before acquiring the media.

Among others, the vocabularies address such concerns as the management of the media and collections, descriptions of their content, their taxonomic, geographic, and temporal coverage, and the appropriate ways to retrieve, attribute and reproduce them

4.13 – Taxon Concept Schema (TCS)

Source: http://standard.biodinfo.org/bsbc/

Description (goals):

- Describe the taxonomic classification system and taxa as components in it
- Be used in multi-checklist environment
- Emphasize on data curation and related reference
- Support multi-lingual environment
- Support structured taxonomy data

Characters:

- Designed to meet current needs on data transferring for taxonomic tree and taxon description;
- Mixture of multiple existing standards, such as CoL Base Schema, Darwin Core, EOL Taxon Resource Transfer Schema;
- Compatible with DwC-A, EOL Transfer Schema, CoL Base Schema, be easy to transform to these standards;
- Contains elements for information on taxonomic tree, such as name, description, authorship, copyright and curation for the tree;
- Contains elements required for multilingual environment, such as formal name in a non-English language and its roman alphabetic format, language tags in all text elements and reference descriptions;
- Contains elements for taxonomic identification key;
- Supports structured taxonomy data.

5 – Projects, initiatives and associated e-infrastructures handling taxon treatment data elements and entire treatments

The following list compiles the major projects and initiatives *that deal with treatments and their data elements*, with short descriptions and links to websites.

5.1 – Biodiversity literature and references

Legacy literature is mentioned here as an indispensable source of information on taxon treatments published in the course of more than 250 years of work of generations of taxonomists, starting with Linnaeus. Classical examples of this basic part of human knowledge are the Flora, Fauna and Mycota series of volumes developed independently by different teams, organisations in different regions of the world, stored in both institutional and private libraries.

The transition of the conventional paper/PDF based publishing model to new models that allow XML-based publishing and dissemination of information is not an easy process. We shall not go into detail here but rather refer to three recent reviews on the topic, treating zoological and botanical literature, and a workflow respectively (Miller et al. 2012, Marhold and Stuessy 2013; Agosti and Egloff, 2009).

5.1.1 Biodiversity Heritage Library (BHL)

http://www.biodiversitylibrary.org/

Ten major natural history museum libraries, botanical libraries and research institutions in the US and UK have joined to form the Biodiversity Heritage Library (BHL) project. The group is developing a strategy and operational plan to digitise the published literature of biodiversity held in their respective collections. This literature will be available through a global “biodiversity commons”.

The participating libraries have over two million volumes of biodiversity literature collected over 200 years to support the work of scientists, researchers and students in their home institutions and throughout the world.

BHL will provide basic, important content for immediate research and for multiple bioinformatics initiatives. For the first time in history, the core of these natural history and herbaria library collections will be available to a global audience. Web-based access to these collections will provide a substantial benefit to people living and working in the developing world, whether scientists or policy makers. By May 2013, BHL had brought some 40 Mio pages online.

5.1.2 CrossRef

http://www.crossref.org

CrossRef is a not-for-profit association of about 2000 voting member publishers who represent 4300 societies and publishers, including both commercial and not-for-profit organisations. CrossRef includes publishers with varied business models, including those with both open access and subscription policies. CrossRef does not provide a database of full text scientific content. Rather, it facilitates the links between distributed content hosted at other sites.

CrossRef provides the technical and business infrastructure to provide for this reference linking using Digital Object Identifiers (DOIs). CrossRef provides deposit and query service for its DOIs.

In addition to the DOI technology linking scholarly references, CrossRef enables a common linking contract among its participants. Members agree to assign DOIs to their current journal content and they also agree to link from the references of their content to other publishers' content. This reciprocity is an important component of what makes the system work.

5.2 – Taxon names

5.2.1 Catalogue of Life (CoL)

http://www.catalogueoflife.org/

The Catalogue of Life (CoL) is envisaged to become a comprehensive catalogue of all known species of organisms on Earth. It contains known species with their accepted scientific name, a cited reference and its family and/or position in the hierarchical classification. Additional common names and synonyms may be provided, but these data are not complete and for some species none may exist. The CoL provides an annual and a dynamic checklist. Both enable the user to provide a potential ambiguous taxon string and verify its taxonomic status and currently accepted name. The Annual Checklist is a snapshot of the CoL, released on CD and on the web every year. It has become well-established as a cited reference used for data compilation and comparison. For instance, it is used as the principal taxonomic index in the GBIF data portal and it is recognised by the CBD. The Dynamic Checklist is updated on the web, facilitated by the SPICE Common Access System (CAS) as an AXIS / SOAP web service. It accesses the current state of provider databases exposed to the web using wrappers. At present, the Catalogue of Life delivers over one million of the estimated 1.8 million known species from 50 institutions. The web portal receives 40 million hits per year, from tens of thousands of individual users. The CoL forms the taxonomic basis of GBIF and the Encyclopedia of Life (EOL). CoL was developed by ITIS and Species2000. From April 2013 onward Naturalis has assumed the responsibility for the Species 2000 secretariat. By doing this, Naturalis wants to position itself as an organisation where knowledge of taxonomy is maintained.

5.2.2 Bibliography of Life (RefBank + ReFinder) (BoL)

http://refbank.org

Bibliography of Life is one of the core output of the FP7 project ViBRAT (www.vbrant.eu) and consists of two main component: (1) RefBank which constitutes a database and associated services for storing and handling of literature references, mostly from the field of biodiversity sciences and (2) ReFInder services layer for discovery and download of references from various sources (including RefBank). Currently BoL contains nearly 200,000 references and is populated from various sources, including automated harvesting of the reference lists from articles published in XML versions in Pensoft's journals.

5.3 – Specimens (occurrence data)

5.3.1 Global Biodiversity Information Facility (GBIF)

http://www.gbif.org

The Global Biodiversity Information Facility (GBIF) facilitates free and open access to biodiversity data worldwide via the Internet. Priorities, with an emphasis on promoting participation and working through partners, include mobilising biodiversity data, developing protocols and standards to ensure scientific integrity and interoperability, building an informatics architecture to allow the interlinking of diverse data types from disparate sources, promoting capacity building and catalysing development of analytical tools for improved decision-making.

The current GBIF informatics architecture features a distributed network of data providing nodes linked to a central data caching system. Nodes make their data available by installing "provider" software that acts as an interface to their databases. The provider software enables a mapping from the local database schema to a federation or interchange schema such as Darwin Core and ABCD and communication between distributed databases using the DiGIR, BioCASE or TAPIR access protocols. The nodes first advertise their presence by registering an access URI in a central UDDI registry.

5.3.2 International Plant Names Index (IPNI)

http://ipni.org

The International Plant Names Index (IPNI) describes itself as "a database of the names and associated basic bibliographical details of seed plants, ferns and lycophytes." Coverage of plant names is best at the rank of species and genus. It includes basic bibliographical details, associated with the names, and its goals include eliminating the need for repeated reference to primary sources for basic bibliographic information about plant names.

IPNI is the product of collaboration between The Royal Botanic Gardens, Kew (Index Kewensis), The Harvard University Herbaria (Gray Herbarium Index), and the Australian National Herbarium (APNI).

The IPNI database is a collection of the names registered by the three cooperating institutions and they work towards standardising the information. For example, The standard of author abbreviations recommended by the ICNafp (International Code of Nomenclature for algae, fungi, and plants) is Brummitt and Powell's *Authors of Plant Names*. A digital and continually updated list of authors and abbreviations can be consulted online at IPNI.

The IPNI provides names that have appeared in scholarly publications, with the objective of providing an index of published names rather than prescribing the accepted botanical nomenclature.

5.3.3 Index Fungorum

http://indexfungorum.org

Index Fungorum is an international project to index all formal names (scientific names) in the Fungi Kingdom. It is somewhat comparable to the International Plant Names Index (IPNI), but with more contributing institutions.

Another difference is that where IPNI does not indicate correct names, the *Index Fungorum* does indicate the status of a name. In the returns from the search page a currently correct name is indicated in green, while others are in blue (a few, aberrant usages of names are indicated in red). All names are linked to pages giving the correct name, with lists of synonyms.

5.3.4 MycoBank

http://www.mycobank.org

MycoBank is an online database, documenting new mycological names and combinations, eventually combined with descriptions and illustrations. It is run by the Centraalbureau voor Schimmelcultures fungal biodiversity center in Utrecht.

Each novelty, after being screened by nomenclatural experts and found in accordance with the ICNafp (International Code of Nomenclature for algae, fungi, and plants), is allocated a unique MycoBank number before the new name has been validly published. This number then can be cited by the naming author in the publication where the new name is being introduced. Only then, this unique number becomes public in the database. By doing so, this system can help solve the problem of knowing which names have been validly published and in which year.

MycoBank is linked to other important mycological databases such as Index Fungorum, Life Science Identifiers, Global Biodiversity Information Facility (GBIF) and other databases.

5.3.5 ZooBank

http://www.zoobank.org

ZooBank is an open access website intended to be the official International Commission on Zoological Nomenclature (ICZN) registry of zoological nomenclature. It was officially proposed in 2005 by the executive secretary of ICZN. The registry was live on 10 August 2006 with 1.5 million species entered.

The ZooBank prototype was seeded with data from *Index to Organism Names*, which was compiled from the scientific literature in *Zoological Record* now owned by Thomson Reuters.

Life Science Identifiers (LSIDs) are used as the globally unique identifier for ZooBank registration entries. The first ZooBank LSIDs were issued on 1 January 2008, precisely 250 years after 1 January 1758, which is the date defined by the ICZN Code as the official start of scientific zoological nomenclature.

5.4 – Genomic data

5.4.1 GenBank and INSDC

http://www.ncbi.nlm.nih.gov/Genbank/

The GenBank sequence database is an open access, annotated collection of all publicly available nucleotide sequences and their protein translations. This database is produced and maintained by the National Center for Biotechnology Information (NCBI), which is a part of a part of the National Institutes of Health in the United States. Three platforms, EMBL-Bank (http://www.ebi.ac.uk/embl/), GenBank (http://www.ncbi.nlm.nih.gov/genbank/), and DDBJ (http://www.ddbj.nig.ac.jp) emerged, which in 1992 formed the International Nucleotide Sequence Database Collaboration (INSDC; http://www.insdc.org). Today, this consortium provides access to several databases focusing on molecular data.

GenBank and its collaborators receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms. In the more than 30 years since its establishment, GenBank has become the most important and most influential database for research in almost all biological fields, whose data are accessed and cited by millions of researchers around the world. GenBank continues to grow at an exponential rate, doubling every 18 months; release 194, produced in February 2013, contained over 150 billion nucleotide bases in more than 162 million sequences (source: http://en.wikipedia.org/wiki/GenBank) GenBank is built by direct submissions from individual laboratories, as well as from bulk submissions from large-scale sequencing centers.

5.4.2 Barcode of Life Data Systems (BOLD)

http://www.barcodinglife.org/views/login.php

The Barcode of Life Data Systems (BOLD) is an online workbench for advancing species identification and discovery through the analysis of short, standardised gene regions. It aids the collection, management, analysis, and use of DNA barcodes.

BOLD supports the organisation and analysis of barcode data and provides a repository for barcode records, storing specimen data and images as well as sequences and trace files. BOLD provides an efficient interface for submitting barcode records to GenBank and an identification engine based on the current barcode library. It monitors the number of barcode sequence records and species coverage.

5.5 – Images

5.5.1 Morphbank

http://www.morphbank.net/

Morphbank is a continuously growing database of images that scientists use for international collaboration, research and education. Images deposited in Morphbank document a wide variety of research including specimen-based research in comparative anatomy, morphological phylogenetics, taxonomy and related fields focused on increasing the knowledge about biodiversity.

5.6 – Treatments

5.6.1 Encyclopedia of Life (EOL)

http://www.eol.org/content/page/who_we_are

The Encyclopedia of Life (EOL) is a project to organise and make available via the Internet virtually all information about life present on Earth. At its heart lies a series of websites – one for each of the approximately 1.8 million known species – that provide the entry points to this vast array of knowledge. The entry-point for each site is a species page suitable for the public, but with several linked pages aimed at more specialised users. The sites contain text and images as well as providing links to specific data.

The EOL dynamically synthesises biodiversity knowledge about all known species, including their taxonomy, geographic distribution, collections, genetics, evolutionary history, morphology, behaviour, ecological relationships and importance for human well-being, and distributes this information through the Internet. It serves as a primary resource for a wide audience.

5.6.2 Scratchpads

http://scratchpads.eu/

The Scratchpads project is a framework conceived in 2007 as part of the FP6 project EDIT (RI-018340) and extended in the FP7 project ViBRANT (RI-261532) to encourage communities interested in natural history to build web sites and share their data (Smith et al. 20011). Currently the Scratchpads framework serves more than 6,500 active registered users across more than 500 sites, serving academic, amateur, and citizen-science needs. Such sites are, in effect, biodiversity portals. Scratchpads is an easy-to-use social networking application that encourages contributors to structure their data for effective sharing and publication data online. Data can be uploaded to the site in a variety of formats, the most popular being Excel spreadsheets, and are made available for download in the widely used Darwin Core Archive (DwC-A) format. Data from a Scratchpad can be exported directly to Pensoft Publishers for traditional peer-reviewed publication in one of several journals (Blagoderov et al. 2010). Most sites are hosted at the Natural History Museum, London, and offered free to anyone who completes an online registration form. The software code-base is open source and is available to be installed at any other site that wishes to offer a similar service.

5.6.3 EDIT Platform for Cybertaxonomy

http://wp5.e-taxonomy.eu/

In the need for workflow-based approaches for converting and integrating data and shielding the user from the complexity of the standards and data structures the European Distributed Institute of Taxonomy (EDIT) created the EDIT Platform for Cybertaxonomy. The EDIT Platform supports the entire taxonomic workflow, therefore it provides possibilities to import and export data in a standardised way (ABCD, DwC, SDD).

It provides researchers with a set of coupled tools for: full, customised access to taxonomic data; editing and management of data; collaborative work in teams; and efficient publishing to both the web and in printed form.

A central data repository and information broker application has been created to achieve interoperability with and between existing applications and web-based data providers. It allows other software to exchange data, via import and export functionality in major data formats, or via web services.

This data repository as well as the core components “EDIT Taxonomic Editor” and “EDIT DataPortal” are based on the EDIT Common Data Model (CDM), which comprehensively covers the information domain, including nomenclature, taxonomy, descriptive data, media, geographic information, literature, specimens, and persons. Wherever possible, the CDM has been made compatible with existing community data standards. The CDM library provides an import and export package for taxonomic classifications, descriptive data, specimens and observations, and media in many standardised or quasi standardised data formats.

5.6.4 Plazi Treatment Repository

http://plazi.org:8080/GgSRS/

The Plazi Treatment Repository (PTR) provides access to treatments that have been encoded in TaxonX. The repository currently contains over 23,000 treatments drawn from 1900 publications, ranging from 1758 to recent. Treatments in the repository are marked up by Plazi or its service provider Plazi GmbH, or contributed by other organisations such as Pensoft for treatments from its journals. The granularity of the markup ranges from coarse (treatment and its substructure and names), to fully atomised materials citations, and references. Treatments are linked to external sources, such as names to ZooBank or the Hyemnoptera Name Server. Users may browse the collection by taxonomic hierarchy by selecting a family and drilling down to lower levels. Search is also provided, both over the full text of the treatments and by taxa as well as several other fields including names of family and genus, the source publication's author, year, and article title, as well as by the geographic locations of the specimens cited by treatments. The full text of the treatments is presented in HTML, with links to XML versions; all taxon names mentioned in the treatment are (when available) linked to records in external name servers such as ZooBank. Treatments are also assigned unique identifiers to facilitate reference, linkage, and integration in a Linked Open Data environment.

The main purpose of the Plazi repository is to provide a source for aggregator to import treatments, like CDM in Pilot 4.2 of this project, or the Encyclopedia of Life, or parts of like materials citations, as used by the Global Biodiversity Information Facility.

5.6.5 Species-ID

http://www.species-id.net

Species-ID is dedicated to collecting and integrating open taxon descriptions and identification tools for different taxa. The audiences addressed are scientists and naturalists, both amateurs and professionals. The huge task of providing adequate documentation of the world's biota requires a collaborative approach. Several layers of contributions are welcome to Species-ID:

- Descriptions and identification tools (species treatments, dichotomous, polytomous, multi-access keys, etc.).
- Checking, editing and updating of existing Wiki pages.
- Enhancing the access and usability through restructuring, categorizing, semantic Wiki information or tools, Wiki templates or adding new software extensions.

Species-ID publishes materials under an open content policy that is compatible with other open content projects such as Wikipedia, Wikispecies or Open educational resources (OER) more generally. Unlike Wikipedia, which is dedicated to summarising information previously published elsewhere, original and authored information may be published on Species-ID. Despite this policy, for major revisions and all nomenclatural acts, a publication in a journal is recommended. The submission of raw data files for interactive identification keys (e. g. in DELTA, Xper, SDD, or other formats) is especially encouraged to provide options for a future re-use of data.

Since 2011, an automated workflow was set up that export taxon treatments from articles published by Pensoft's journals (e.g., ZooKeys, PhytoKeys and MycoKeys) on Species-ID (see Penev, Hagedorn et al. 2011 for detail). The workflow interlinks the original journal publication to the respective Species-ID Wiki page and vice versa. In addition, it allows additions and corrections to the treatment on the Wiki by external contributors, whose names are automatically added to the citation of the Wiki page. Species-ID is part of Plazi.

5.6.6 Pensoft Journal System (PJS 2.0)

http://www.pensoft.net/services-for-journals

The journal publishing system of Pensoft called PJS is a self-developed online management platform composed of several tools (Pensoft Markup Tool (PMT), Pensoft Taxon Profile (PTP), Pensoft Wiki Converter (PWC), Pensoft Writing Tool (PWT) and others that allow markup at different levels of granularity and exporting treatments to external users, such as EOL, Plazi and Species-ID. The system is integrated with data publishing platforms like Dryad and GBIF. Pensoft implements the domain specific markup based on the TaxPub NLM JATS schema since June 2010 (Penev et al. 2010). Pensoft publishes 10 open access journals, of which ZooKeys, PhytoKeys and MycoKeys have implemented several innovations for XML-based advanced open access publishing of biodiversity information.

The currently launched version PJS 2.0 is the first platform ever to support the full life cycle of a manuscript, from writing through submission, community peer review, publication and dissemination within a single, fully XML-based, online collaborative platform. The *Biodiversity Data Journal* (BDJ) (www.pensoft.net/journals/bdj) and associated *Pensoft Writing Tool* (PWT) (www.pwt.pensoft.net) are key elements of PJS 2.0 and can be used for collaborative online elaboration and publication of treatments. PJS 2.0 is supported in part by the FP7 project ViBRANT (www.vbrant.eu).

6 – Data exchange and cross-linkages

A study of the mega science platforms dealing with mycota (Triebel et al., 2012, Fig. 15) shows in an exemplar way that there is some overlap between the goals of the projects, but also a high degree of specificity.

The data life cycle and data flow starts with data production. The megascience platforms are harvesting infrastructures which are part of a 'food chain' that starts with the primary-content producers to primary and secondary harvesters and ends up with data users, consumers and digesters. Data harvesters like GBIF and CoL, which are typically fed by research data from individual scientists and institutions, may alternatively also be supplied by primary data collecting infrastructures, e.g., by the World Register of Marine Species (WoRMS; http://www.marinespecies.org/), Species Fungorum (http://www.speciesfungorum.org/), and FishBase (http://www.fishbase.org/).

Names data, taxonomy, and classifications are of essential interests for all biodiversity platforms. Thus the comprehensive and reliable species databases offered by CoL form one of the multiple taxonomic backbones of EOL, GBIF, iBOL, BHL, and the INSDC data platforms.

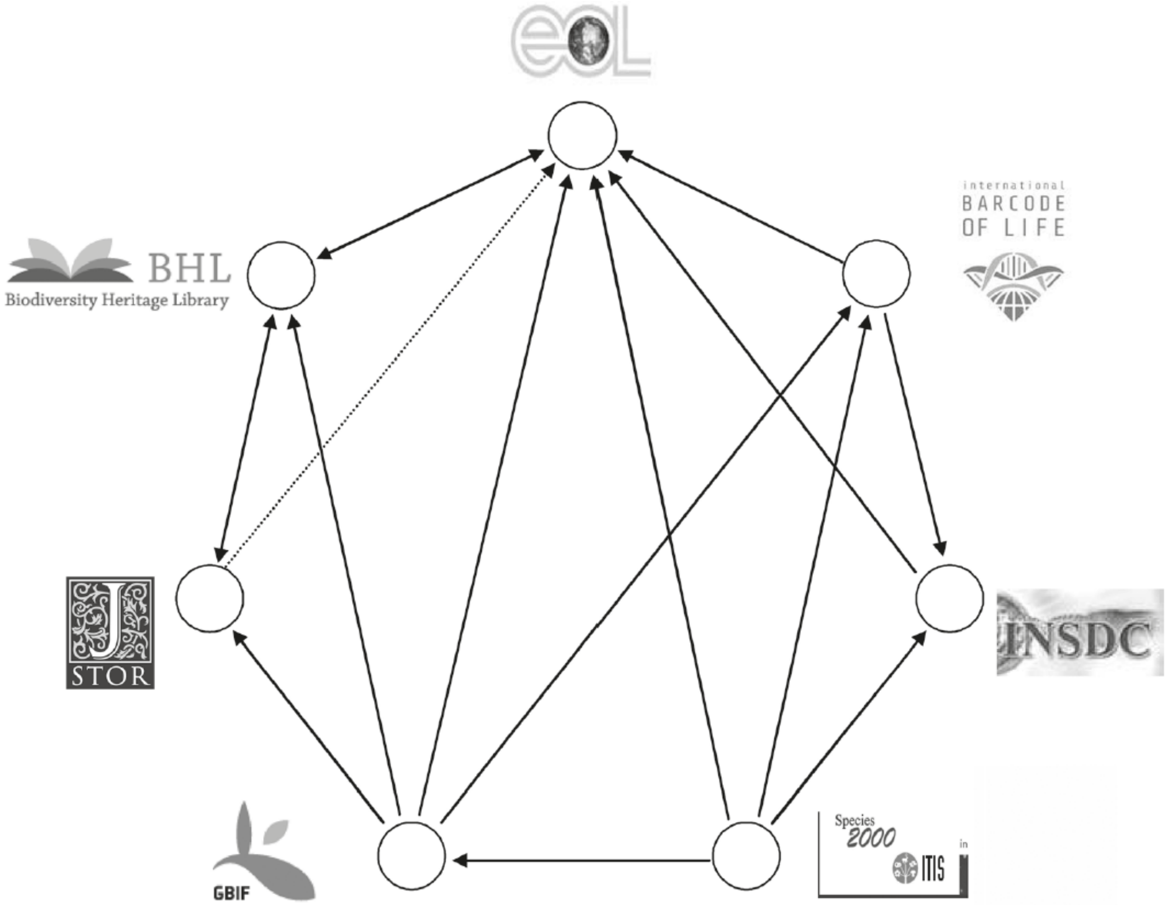


Figure 15. Biodiversity megascience platforms – cross-linkages and data exchange (Triebel et al., 2012).

Concerning taxonomic names and classifications, the data flows will be even more challenging in the future because there are overlapping and competing name thesauri for taxonomic and biological groups worldwide. As an example: Lichen names and synonym data are actually being collected by three different major sites (Index Fungorum/ Species Fungorum; http://www.indexfungorum.org, LIAS names, and MycoBank), and are either directly forwarded to several megascience platforms, or indirectly via CoL.

Another type of data flow starts with the occurrence data harvested by the megascience platform GBIF. Several initiatives or projects like EDIT and BioCASE established data flow structures with mirrors of the GBIF index database. Based on these cache databases, they forward large amounts of GBIF occurrence data to various thematic search portals (http://search.biocase.org/; http://search.biocase.org/edit/).

The cooperation and linkages between the seven megascience platforms themselves as well as between the seven initiatives and their primary data providers is assumed to be facilitated by relying on open source principles and on contents provided under creative commons or open database licenses conditions or – at least – data sharing policies on a non-exclusive basis. With growing content, the data flow and cross-linkages between the seven platforms is visible (Fig. 15). In parallel, the backtracking of multimedia data with corresponding metadata, e.g., from EOL and from thematic portals like EDIT (http://search.biocase.org/edit/: this is mirroring the GBIF index database, back to the primary providers or publishers of scientific data is possible.

Each of the seven platforms has its own profile with respect to data domains, providers and scope of contents, and user communities, but strong dependencies between the platforms (e.g., between BHL and EOL) exist. Furthermore, there is cooperation between the four platforms GBIF, iBOL, EOL and JSTOR Plant Science to visualize occurrence data and to link data from biodiversity literature. They, therefore, require a common name data backbone, provided by a jointly developed technical structure in the frame of a common project, the Global Names Architecture (GNA; http://www.globalnames.org/) project. For sequence data which is produced in the iBOL context, the INSDC consortium with NCBI GenBank has agreed to stand by as the general data repository and backup archive.

There is a remarkable congruence on some repetitively occurring elements that many of the initiatives use either directly or indirectly. Direct use deals with cornerstone elements of the natural history institutions such as collection specimens, names or literature. These basic data form a wide ground for indirect use, for example of niche models and phylogenies that are based on specimens, which, if properly documented, allows for the retrieval of linked information, such as geographic data, sequence data or images from their metadata, and at the same time can be linked to from the specimens used.

7 – Markup of treatments and integration between legacy literature and newly published data

The overall workflow of implementation of tagging of taxonomic texts, either published in legacy literature or within a prospective, XML-based editorial process, is shown in Fig. 16.

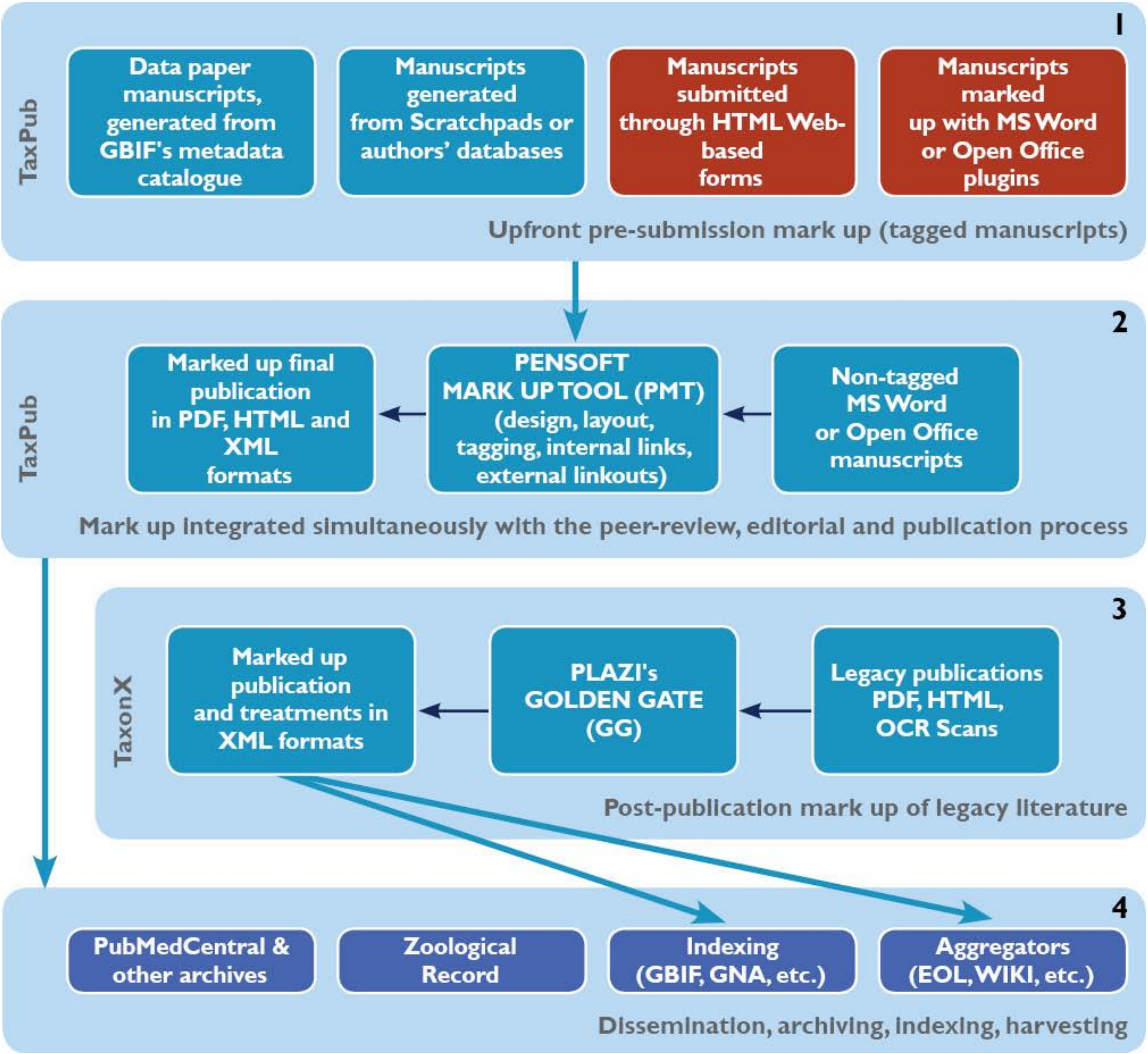

1
TaxPub
Data paper manuscripts, generated from GBIF's metadata catalogue
Manuscripts generated from Scratchpads or authors' databases
Manuscripts submitted through HTML Web-based forms
Manuscripts marked up with MS Word or Open Office plugins
Upfront pre-submission mark up (tagged manuscripts)
2
TaxPub
Marked up final publication in PDF, HTML and XML formats
PENSOFT MARK UP TOOL (PMT) (design, layout, tagging, internal links, external linkouts)
Non-tagged MS Word or Open Office manuscripts
Mark up integrated simultaneously with the peer-review, editorial and publication process
3
TaxonX
Marked up publication and treatments in XML formats
PLAZI's GOLDEN GATE (GG)
Legacy publications PDF, HTML, OCR Scans
Post-publication mark up of legacy literature
4
PubMedCentral & other archives
Zoological Record
Indexing (GBIF, GNA, etc.)
Aggregators (EOL, WIKI, etc.)
Dissemination, archiving, indexing, harvesting

Figure 16. Four stages of an XML-based editorial, publication and dissemination workflow applied in ZooKeys (stages 1, 2, 4) and/or Plazi (stages 3, 4). Forms in blue are either implemented or prototyped, forms in red are in a process of development (after Penev et al. 2010).

Tagging of taxonomic text is a quite laborious task, mostly because of the specificity of the domain, e.g., the great variety in publishing styles, taxon names (synonymy, homonymy, spelling errors, different concepts for a particular taxon name, etc.), listings of localities (long lists of terms describing a particular locality or collecting event), etc. In most cases, this is being done manually or semi-manually, which may explain why finer granularity markup has not been used by taxonomic journals thus far. Overall information on the markup process provided by Plazi through the Golden Gate tool is described below.

7.1 – Description of Plazi's Markup Process provided by the Golden Gate editor

The markup process leads from loading a scanned legacy document (directly from an image-based PDF, or from an HTML document generated by OCR software) or a contemporary born-digital PDF into the GoldenGATE Document Editor. The individual at least partially steps build upon one another. Striving maximum automation to reduce user effort, the whole markup activity is bundled into as few steps as possible. In general, each step involves one round of user interaction. The dialogs used for the latter are all pre-populated by means of natural language processing or rule based heuristics, so users can check existing markup and correct errors rather than creating the markup entirely from scratch. In addition, users never have to deal with XML syntax details, as the dialogs only offer the options required for the task at hand. This is in favor of simplicity and ease of use. The table below describes the individual markup elements in detail, including which step creates them and sets their attributes.

Step 1a (HTML Normalisation): Mark pages, detect page numbers and store them in paragraph attributes to ease later bibliographic referencing, check and correct paragraph boundaries, recognise and tag captions, footnotes, and bibliographic references, clean up layout artifacts like page headings, resolve hyphenation, and normalise diacritics. All of this works with a single user interaction dialog per page, in which users group lines into paragraphs and classify the paragraphs along the way.

Step 1b (PDF Normalisation): Group words into paragraphs, and paragraphs into lines, detect page numbers and store them in paragraph attributes to ease later bibliographic referencing, check and correct paragraph boundaries, recognise and tag captions, footnotes, and bibliographic references, clean up layout artifacts like page headings, resolve hyphenation, and normalise diacritics. All of this works with a single user interaction dialog per page, in which users group words into paragraphs, and paragraphs into blocks and columns, and classify the paragraphs along the way.

Step 2 (Spell Checking): Detect and correct misspellings and OCR errors. This step uses a single user interaction dialog per paragraph, and only for the paragraphs in which a possible misspelling is found. If the document was loaded from a born-digital PDF or an HTML document that was spell-checked in the OCR software that created it, this step can usually be omitted.

Step 3 (Document Meta Data Import): Import bibliographic metadata for the document, for instance from RefBank, or enter it manually in a dedicated dialog if none is found in online sources or online sources are inaccessible.

Step 4 (Bibliography Parsing): Parse details like authors and title out of bibliographic references. This works with one user interaction dialog per reference, which displays the details for checking and correction.

Step 5 (Citation Tagging): Mark citations of bibliographic references in the document main text, so the citations can be resolved even with only a small excerpt of the document at hand. This works with user interaction dialogs that display 10 citations each for manual disambiguation.

Step 6 (Taxonomic Name Tagging): Mark and preliminarily parse taxonomic names. This works with user interaction dialogs that display 10 distinct taxonomic names each for manual selection.

Step 7 (Taxonomic Name Parsing): Parse individual epithets and their ranks out of taxonomic names, filling in abbreviations and implicit epithets along the way. This works with user interaction dialogs that display 10 distinct taxonomic names each for manual completion.

Step 8 (Document Structuring): Segment the document into taxonomic treatments and other sections, like introduction, methodology, or bibliography. This works with a single user interaction dialog, which displays all paragraphs except for captions and footnotes for grouping.

Step 9 (Treatment Structuring): Segment taxonomic treatments into individual sub-sections, like nomenclature, description, or distribution. This works with one user interaction dialog per treatment, which displays all paragraphs that lie in the treatment for grouping.

Step 10 (Materials Citation Markup): Mark individual materials citations, and syntactically unambiguous details thereof, with the latter being dates, geographical coordinates, type status indicators, collection codes, and country names. This works with one user interaction dialog for each paragraph in which at least one of the aforementioned details is found.

Step 11 (Materials Citation Parsing): Parse details like collecting locations, collector names, etc. out of materials citations. This works with one user interaction dialog per materials citation, which displays the details for checking and correction.

7.2 – Description of markup work flow integrated with editorial and publishing process provided by the Pensoft Markup Tool (PMT)

There are two possible ways to solve this challenge and optimize the markup process so that it becomes economically viable. A straightforward way is to have manuscripts tagged before submission through (i) exports from databases, such as Scratchpads (http://www.scratchpads.eu), GBIF or authors' personal/institutional databases, or by using (ii) HTML submission forms, or through (iii) TaxPub or other XML schema based plugins for MS Word, Open Office or other text processors. The latter method will help authors to write extensive manuscripts of a more complicated structure than those generated from databases or submitted through HTML forms. None of these methods is widely used, to say the least. Method (ii) has been realized in Pensoft Writing Tool, and Method (iii) simply does not exist yet. There is no doubt, however, that we can anticipate a quick transformation to "automated" generation and submission of manuscripts within the coming years, and surely within the lifespan of the present-day generations of active taxonomists.

The second route to the same output is for publishers to find a way to apply XML tagging within their editorial workflows (Fig. 17). As far as it concerns the general article structure, such as title, authors, abstract, introduction, etc., this is not a problem and most major publishers do it. However, once we decide to go to for a finer markup, that is to tag taxon names, taxon treatments, sections within a taxon treatment (nomenclature, morphological description, distribution, type material, examined material with data on localities and specimens, etc.), the difficulties appear hardly surmountable, and to the best of our knowledge, there is no current working solution for them in biodiversity science.

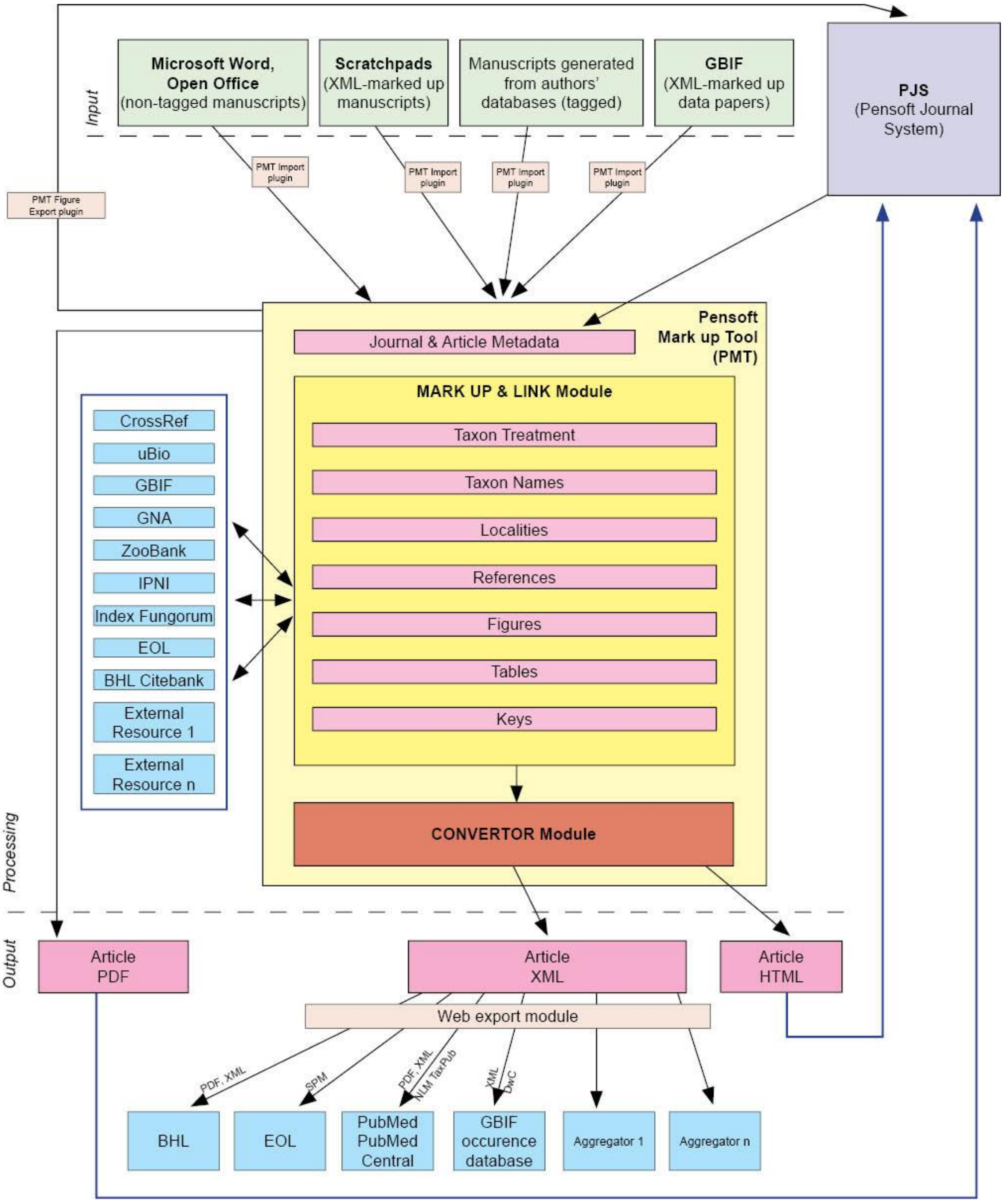


Figure 17. Flowchart of an integrated, XML-based editorial, publishing and dissemination process applied in ZooKeys through the Pensoft Markup Tool (PMT) (after Penev et al. 2010).

The Pensoft Markup Tool (PMT) (Fig. 17) was developed to integrate markup within the editorial process and provides the following functionalities:

1. Importation and retrieval of XML, HTML and InDesign files.
2. Interlinking options between PMT and InDesign allowing simultaneous markup and editorial work.
3. Tagging and "auto-tagging" at different granularity levels, according to TaxPub or any other XML schema designed for such purpose.
4. Cross-linking of citations within the text and reference lists.
5. Cross-linking of citations of figures and tables in the text.
6. Finding and linking taxon names through http://www.uBio.org and PMT's own web harvester.
7. Providing links to various external sources.
8. Exporting the text to a semantically enhanced HTML version of the paper, vizualizing some of the important tag elements, as well as the literature references cited in the text and external links to them (when available).
9. Mapping localities listed in the paper or within separate taxon treatments.
10. Generating the Taxon Pensoft Profile (http://www.pensoft.net) page for each taxon name cited in a paper, providing the reader with a quick and up-to-date summary of information on a taxon from certified external sources.
11. Offering a possibility to the reader to create their own taxon profiles for taxa of interest.
12. Export to a TaxPub XML file, validated for archiving in PubMedCentral and indexing in PubMed.
13. XML export of new species descriptions to Encyclopedia of Life, using elements drawn from Dublin Core, TDWG Darwin Core, and TDWG Species Profile Model schemas.
14. XML export of treatments or any other tagged information in various formats acceptable by aggregators and indexers, Plazi taken as an example.
15. The PMT creates profiles of any taxon name mentioned in a paper, independent of its rank or nomenclatural status. An example of a PTP page for the taxon profiles is here: Quercus suber.

Two classes of selected websites are targeted by the PTP: (1) pillars of biodiversity informatics online (e.g., GBIF, NCBI, EOL, Barcode of Life, Wikipedia, BHL, and others) have dedicated windows showing results for a particular taxon name, or reporting that no results were found (because the lack of results from key online resources could itself be an important finding), and (2) taxon-oriented websites, from which results are displayed only if a particular taxon name was found (e.g., ZooBank, International Plant Name Index, diptera.org and others).

8 – Map of ongoing biodiversity projects and data exchange involved with the production and use of taxonomic treatments

As a result from the work in Task 2.1, a map was produced to visualize the main ongoing biodiversity informatics initiatives, platforms and project that have a direct relation to production and use of taxon treatments and their data elements (Fig. 18).

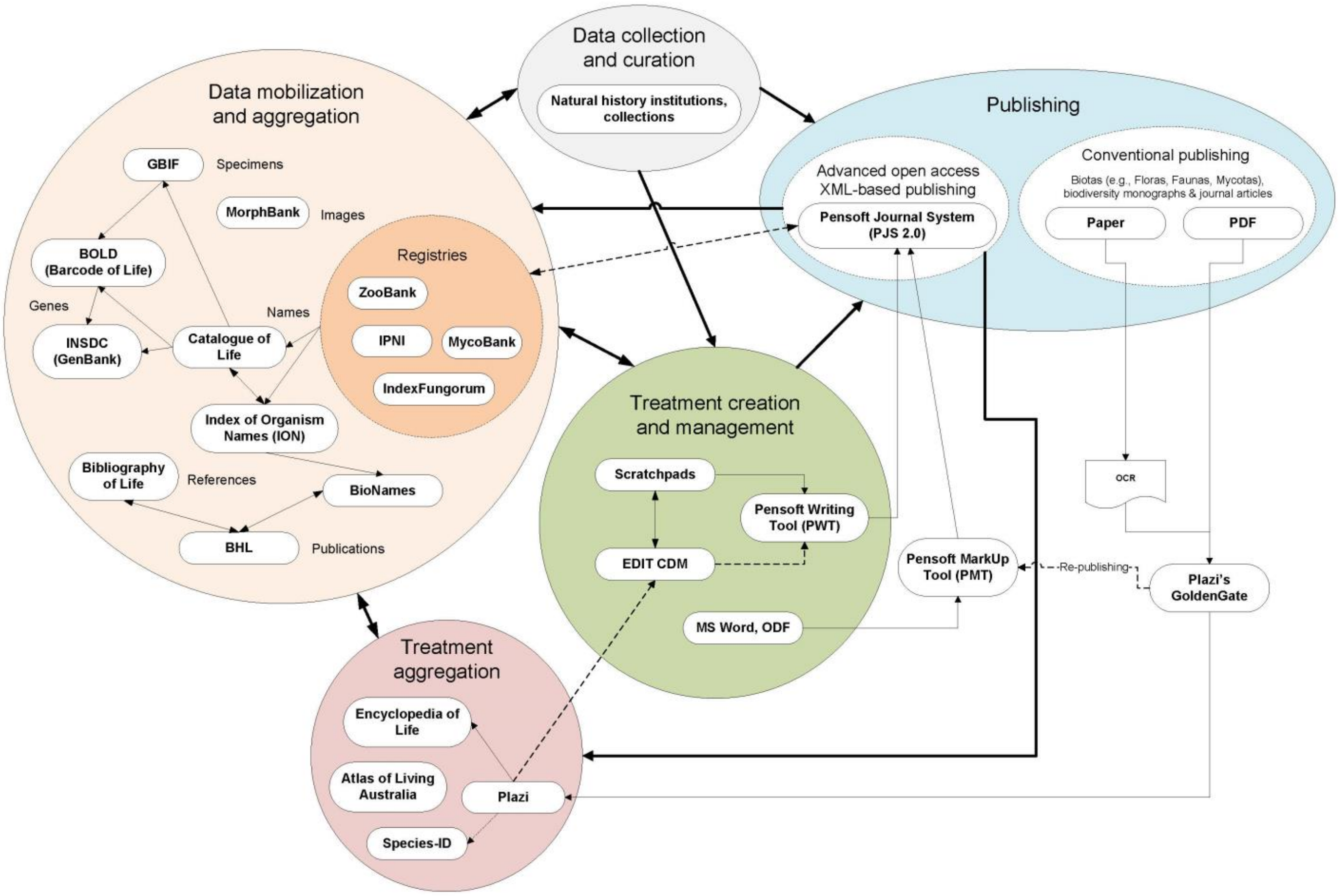

Data collection and curation
Natural history institutions, collections
Data mobilization and aggregation
GBIF
Specimens
MorphBank
Images
BOLD (Barcode of Life)
Genes
INSDC (GenBank)
Catalogue of Life
Names
Index of Organism Names (ION)
Bibliography of Life
References
BioNames
BHL
Publications
Registries
ZooBank
IPNI
MycoBank
IndexFungorum
Publishing
Advanced open access XML-based publishing
Pensoft Journal System (PJS 2.0)
Conventional publishing
Biotas (e.g., Floras, Faunas, Mycotas), biodiversity monographs & journal articles
Paper
PDF
OCR
Treatment creation and management
Scratchpads
Pensoft Writing Tool (PWT)
EDIT CDM
MS Word, ODF
Pensoft MarkUp Tool (PMT)
Re-publishing
Plazi's GoldenGate
Treatment aggregation
Encyclopedia of Life
Atlas of Living Australia
Plazi
Species-ID

Figure 18. Map of ongoing biodiversity informatics projects and e-infrastructures that forms the prototype of the future Open Biodiversity Knowledge Management System (OBKMS).

Solid arrows: established data exchange routes and links. Dashed arrows: links in progress.

9 – White paper: description of gaps, challenges and proposals for an optimized dataflow in biodiversity informatics

9.1 – Gaps in the dataflow, interoperability and cooperation between stakeholders involved in biodiversity informatics

As a result of the discussions, presentations and questionnaires at the workshops and in the process of work on the pro-iBiosphere pilots, we have identified the following most important gaps in the data flow exchange and interoperability between the various projects, platforms and initiatives. The gaps and challenges listed below do constitute a bunch of significant barriers towards the future Biodiversity Knowledge Management System (OBKMS). Definitely the barriers to data mining, aggregation and re-use of accumulated knowledge significantly decreases effectiveness of scientists' effort and science funding, at both EU and global level:

1. Copyright impediment and insufficient level of open access to published information.
2. Social barriers (e.g., "protection" of own data from external users) at personal and institutional levels hinder open data publishing and sharing.
3. Lack of sufficient coordination between EU and non-EU biodiversity informatics initiatives.
4. Inconsistent use of persistent and resolvable identifiers (or lack of use) prevents real integration between data elements and platforms.
5. Undervaluing of the importance of markup and digitisation of accumulated legacy data (e.g., Floras, Faunas and Mycotas) and their integration with newly produced data.
6. Slow adoption of advanced XML-based publishing models that will replace paper/PDF publishing allowing automated data harvesting and re-use.
7. Lack of interoperability between standards, even when they deal with the same data element (e.g., taxonX, TaxPub and taXMLit dealing with treatments).
8. Insufficient funding for digitisation processing, in first line for legacy literature and specimen collections.
9. Increasing gap between organismal biology and biodiversity science in particular and next-generation genome sequencing.
10. Lack of sufficient data repositories and mechanisms for data storage, curation and exchange.
11. The need for further development of OCR procedures and automated markup systems that will reduce the requirement for manual markup.

In order to facilitate adequate access to the rapidly developing cyberspace potential, the main data objects of the taxonomic domain need to be described, and technical and semantic interoperability barriers need to be addressed, both within as well as beyond the biodiversity domain.

9.2 – Challenges

According to the white paper recently published by Hardisty, Roberts and Bioinformatic Community (2013) ":the grand challenge for biodiversity informatics is to develop an infrastructure to allow the available data to be brought into a coordinated coupled modelling environment, able to address questions relating to our use of the natural environment that captures the '*variety, distinctiveness and complexity of all life on Earth*".

On the way to this ultimate goal, however, there are several challenges and associated recommendations identified by the pro-iBiosphere partners and workshop participants that have to be addressed:

1. Methods should be developed to mobilise the rapidly increasing amount of primary data of biodiversity information, such as specimen occurrences, genome sequences, images etc. into "synthetic" datatypes such as taxon treatments and to make these discoverable, known and usable for the global community.
2. Biodiversity-related content should be published in open access and in a way that makes it accessible with as few restrictions as possible (e.g., free use with only attribution required) so that it can be used, reused and collated with other content (see for example the Panton Principles).
3. All new treatments need to become available in machine-readable format, and the references such as bibliographic references, citations of treatments, and material citations need to be linked to the original resources.
4. Each individual element of biodiversity data (e.g., specimens, genomes, images, treatments, etc.) should be identified using a widely adopted system of persistent, universally unique and resolvable identifiers that will ensure easy discoverability and interlinking between the various knowledge elements.
5. The data included in the publications has to be semantically enhanced to make full use of its potential.
6. Registries of biodiversity relevant information and data must be available to facilitate discovery of existing data, including names and treatment citations, bibliographic references.
7. Specific ontologies need to be developed to facilitate machine reasoning and data mining within the biodiversity domain.
8. Metadata has to follow elaborated and widely accepted standards to facilitate persistence and re-use of the data they describe.
9. Legal analysis and recommendations needs to provide the necessary space for a free flow and re-use of scientific data.
10. Policies have to be in developed and applied to stimulate and guarantee a free flow of data and its re-use.
11. Trust must be built to overcome social barriers for open sharing of data and to ensure their provenance and quality. Methods must be developed to make the contribution of people acknowledged, citable and quantifiable.
12. Mechanisms must be developed to ensure easy use and preservation of annotations, comments and other forms of feedback.
13. Adequate funding must be available to maintain the infrastructure and workflow to generate new data that is needed for external use, e.g., biodiversity monitoring (in synergy with EO BON, EU BON and IPBES).
14. Incentives should be developed to encourage journal publishers to switch from PDF-only publishing models to advanced open access publishing of semantically enhanced content and machine-readable formats, such as XML and RDF.

15. Data curation workflows and tools need to be elaborated and implemented, to allow synchronization and updates to databases, and to avoid orphaned databases or loss of central databases.
16. The level of centralization and decentralization has to be optimized to yield adequate functionality at minimized costs.
17. The huge amount of legacy data from historical literature needs to be marked up and exported to databases so that they can be easily collated with newly published data to form the basis of the forthcoming "big data" pool.
18. The vast amount of legacy data accumulated in collections must be digitised and mobilised to provide straight and easy access to this cornerstone of biodiversity research.
19. Make use of universally adapted trusted authentication measures (e.g., ORCID for authors) so that users can easily work with multiple resources.
20. Last but not least, the question why this entire infrastructure should be built and filled with content should direct future development and optimization of workflows.

9.3 – Basic principles of and recommendations towards the Open Biodiversity Knowledge Management System (OBKMS)

In order to implement the European Biodiversity Knowledge Management System envisioned in the pro-iBiosphere proposal the following elements need be addressed, and consequential actions are proposed. A Memorandum of Understanding (MoU) between platforms and initiatives will be developed by Plazi within pro-iBiosphere. The purpose of the MoU will be to express the commitment of the participants to the OBKMS and the associated minimal set of measures as described below.

Trust. The strength of the OBKMS is that content from all the participating institutions can be accessed both from inside as well as from outside at once by agreed, well documented principles, standards and vocabularies. This will minimize duplication of content such as scientific names or bibliographic references and infrastructure and thus costs are reduced.

However, this shift of paradigm from self-contained institutions to institutions becoming part of an overall infrastructure needs a considerable trust in the system to ensure that vital parts are delivered continuously and in a persistent way.

To build trust, and to build and maintain the infrastructure and its content, the following elements are needed:

- Commitment of institutions to co-fund and maintain assigned parts of infrastructures together with an EU commitment to support its maintenance as long as there is a documented need for it.
- Clearly documented goals and deliverables of the system.
- A minimal level of redundancies to guarantee the stability of services.
- Governing boards to overlook the services with clearly communicated policies to define targets, development, maintenance and services as well as application procedures to request modifications.
- A clearly documented procedure for the election and membership of the governing board.

- A stepwise procedure to add increasingly more content to the system, starting with elements that allow rapidly building up trust, such as policies on identifiers, open access, a list of names (see below), and resulting in measurable and quantifiable services.

Open Access. Open Access is vital to achieve the goal of the OBKMS, especially to allow it to play a vital role in creating new opportunities for innovation and solving problems of the civil society. The sheer number of data will need new approaches to link, mine, extract and maintain the data and its derivative products.

Copyright law has been created prior to the digital age, and is based on a business model where a work should be protected in order to recover the costs that accrued in its creation. However, science funding is typically paying upfront with the explicit goal to disseminate scientific results as widely as possible in order to ensure innovation and have results at hand for the solution of unforeseen challenges. This conflicting situation should be resolved with exemptions in the language of the copyright legislation for scientific use.

The EU could enforce an open access policy for all the scientific results and data generated by EU-funded projects. It should also involve legal experts to study the scientific process, its restrictions due to copyright legislation and propose legal language that will provide the necessary legal framework for the unrestricted functioning of the OBKMS.

One World. A considerable amount of scientific data has been produced locally but is of global importance. Local scattered data acquire scientific value with the integration into a species- or similar concept and ultimately into the published record and/or database. In this way, local data may ultimately acquire global relevance and local investments should ultimately be incorporated in a global cyberinfrastructure.

The EU could develop programs that allow building infrastructure between major funding institutions irrespective of political boundaries or strive for agreements to take global responsibility for parts of the infrastructure in exchange for reciprocal actions. The participating institutions in the OBKMS should act similarly.

Linked Data. Data is meaningless without context and scientific analysis and publication is impossible without underlying data. To understand the value of data, its usage has to be measurable and visualizable. For that purpose, all the data has to be linked with adequate identifiers and resolution services.

The OBKMS should be based on the principle of linked data.

Identifiers. To locate and link data, for each data category one or multiple (but cross-linked) identifiers have to be defined and implemented, resolvers created and maintained and at least metadata if not all the content has to be returned in machine-readable form. Since the content of the OBKMS is envisaged to be at the base for life sciences, conservation and beyond, it has to conform to globally used standards and if possible to allow the content be visible in the semantic web.

The EU could promote the use of identifiers as part of their funding requirement, and support the development, building resolving mechanisms and their resolution if necessary.

OBKMS partners should adopt identifiers, and apply them to core data like specimen data, literature, treatments, or digital imaging and other digital objects.

Biodiversity Knowledge Directory. Services and tools used to develop, maintain and use the OBKMS have to be discoverable, described, open source and accompanied with adequate tutorials. An efficient way to document tools and services is to register them as a standard procedure to a Biodiversity Knowledge Directory.

The EU could foster the use of existing directories for registration of tools and services derived from EU-funded projects, for example by making this a funding requirement. A good candidate for a services directory would be for example the Biodiversity Catalogue (https://www.biodiversitycatalogue.org/) developed by BioVeL.

The principle of machine readability. To be an actor in the cyberinfrastructure, most, if not all, the content of the OBKMS has to be machine readable, and preferably also linked and semantically enhanced, and stored in dedicated databases or similar infrastructure that are adequately documented and accessible.

At EU funding level, promote and enforce the principle of machine readability through conditional funding calls and incentives for institutions that adhere to this principle.

Publishing of scientific results. Scientific publications serve to communicate scientific results, are part of a scientific debate, a summary of current knowledge on a given topic, a quality control instrument through peer review, and more recently and increasingly the access point to all the underlying data used in the analysis. A format for scientific publication should be created that serves the needs of their users, that can and will be properly archived in standardised ways, and permits open access.

The participating institutions should request open access to all the scientific results their staff is producing and offer access and archiving facilities of this content.

The EU hold request open access to all the scientific results coming from EU-funded projects. The EU could invest into converting current journals into semantically enhanced journals by supporting the development of journal production workflows and authoring tools. In the short term, tools need to be developed that help to convert traditionally published data into semantically enhanced content. The participating institutions together with the EU, and if possible in a much wider context, should develop a repository and archive for scientific results.

Access to legacy literature. Legacy literature is an integral part of the scientific endeavor and maintains a vast amount of data that can be reused. Because of the sheer size of it, this corpus cannot be opened up at once and needs triage. Principles and mechanisms to make this content machine-readable as well as some degree of funding needs to be in place to allow conversion of specific sub-corpora for specific well defined requests. At the same time, mechanisms have to be in place that allow storing converted content and maintaining links to its source or original digital object. A straightforward way to maximise the effort spent on digitisation would be to re-publish important biodiversity monographic series, such as Floras, Faunas and Mycotas in an open access, semantically enhanced, online publishing model.

EU should fund the development of conversion tools including training and create incentives to build and maintain respective semantically enhanced content that mobilises and opens the legacy biodiversity literature.

Measuring citations. Individual contributions to science increasingly take the form of contributions to databases instead of to peer-reviewed publications, thus allowing third parties to build phylogenies, compile land-use maps or construct niche models. These contributions need to be adequately exposed, archived and given identifiers so that usage and relevance can be measured.

The EU could measure the compliance within projects and value of projects by including metrics that explicitly incorporate contributions to science other than publications.

The OBKMS partners should implement identifiers so that their contribution can be measured.

Data standards. Data standards play a central role in modern electronic communication and are a basic requirement for their interoperability, collation and re-use. Data standards concern both data themselves and associated metadata. An integral part of data standardisation concerns measuring their impact and use, for which data usage and data citation indexes should be elaborated and implemented.

The EU could enforce standardisation in all aspects of data creation, curation, management, and use. The standards should be extended beyond data content and should cover also all associated infrastructures (e.g., repositories), tools and workflows.

The OBKMS partners should use the currently accepted data standards and develop/implement new ones in a strong collaboration, to avoid duplication of effort and loss of data quality and usability.

Applications Programming Interface (API). Big and small data is dealt with by machines. For that purpose all the databases and repositories need to have instructions and standards for accessing the Web-based applications. Projects have to register service interfaces in a publicly accessible service registry such as the Biodiversity Catalogue but they do not necessarily have to be open themselves as there might be good reasons for access control (e.g., for controlling computational resources or for restricting access on localities of protected species).

The EU could request that funded projects offer APIs to access content and applications that have been funded through their awards.

OBKMS members will offer APIs as a standard procedure.

10 – Advisory Board (AB)

To successfully elaborate and implement pro-iBiosphere recommendations towards OBKMS, an Advisory Board was set up after the start of the project. At present, the pro-iBiosphere AB includes the following *members:*

Thomas Janssen (Humboldt Universität)

Laurence Benichou Bénichou (National Museum of Natural History (MNHN, France))

Hong Cui (University of Arizona)

Suzanne Sharrock (Botanic Gardens Conservation International (BGCI)

The Advisory Board will meet once a year. It will act as an advising body on strategic issues while providing recommendations for the overall success of the project. It will also enable the project to foster liaisons and synergies with other related (EU and non-EU funded) projects within the taxonomic and biodiversity informatics landscape.

New AB members may join the Advisory Board later in the project depending on opportunities and needs that might arise.

The first meeting with the pro-iBiosphere Advisory Board took place on the 15th of February 2013. The minutes of the meeting are available here.

11 – Acknowledgments

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.

12 – References

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

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

Agosti D, Klingenberg C, Sautter G, Johnson N, Stephenson C, Catapano T (2007) Why not let the computer save you time by reading the taxonomic papers for you? Biológico, São Paulo 69 (suplemento 2): 545–548. http://hdl.handle.net/10199/15441

Berendsohn W, Güntsch A, Hoffmann N, Kohlbecker A, Luther K, Müller A (2011) Biodiversity information platforms: From standards to interoperability. ZooKeys 150: 71-87. doi: 10.3897/zookeys.150.2166

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

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

Catapano T, Weitzman AL (2007) Progress in making literature easily accessible: schemas and marking up TaxonX / Goldengate & taXMLit / INOTAXA. TDWG Annual Meeting 2007. http://wiki.tdwg.org/twiki/pub/Literature/WebHome/Catapano_Weitzman_Markup_Final.pdf

Catapano, T., Hobern, D., Lapp, H., Morris, R.A., Morrison, N., Noy, N., Schildhauer, M., and Thau, D., 2011. Recommendations for the use of Knowledge Organisation Systems by GBIF. 29pp. Version 1. http://links.gbif.org/gbif_kos_whitepaper_v1.pdf

CBD, (1992) Convention on Biological Diversity. http://www.cbd.int/convention/text/

Cukier, K.N. and Mayer-Schoenberger (2013) The rise of big data. Foreign affairs 92 (3): 28-40.

Cui H (2008a) Converting Taxonomic Descriptions To New Digital Formats. Biodiversity Informatics 5: 20–40 https://journals.ku.edu/index.php/jbi/article/view/46/1551

Cui H (2008b) Approaches to Semantic Mark-up for Natural Heritage Literature. Proceedings of the iConference 2008. http://ischools.org/conference08/pc/PA5-2_iconf08.doc

Cui H, Heidorn PB (2007) The reusability of induced knowledge for the automatic semantic mark-up of taxonomic descriptions. Journal of the American Society for Information Science and Technology. 58(1): 133–149. http://www3.interscience.wiley.com/cgi-bin/fulltext/113466052/PDFSTART

Cui H, Jiang Y, Sanyal PP (2010a) From Text to RDF Triple Store: An Application for Biodiversity Literature[Demo]. Proceedings of the 73rd ASIS&T Annual Meeting v. 47. Oct 22–27, 2010. Pittsburg, PA.

http://www.asis.org/asist2010/proceedings/proceedings/ASIST_AM10/submissions/415_Final_Submission.pdf

Cui H, Boufford D, Selden P (2010b) Semantic Annotation of Biosystematics Literature without Training Examples. Journal of American Society of Information Science and Technology. 61 (3): 522–542. http://harvard.academia.edu/DavidBoufford/Papers/740926/Semantic_annotation_of_biosystematics_literature_without_training_examples

Curry GB, Connor, RJ (2007) Automated extraction of biodiversity data from taxonomic descriptions. In: Curry GB, Humphries CJ (Eds) Biodiversity databases: Techniques, politics, and applications: Systematics Association Special Volume 73: Boca Raton, Florida, CRC Press, Chapter 6: 63–81.

Curry GB, Connor RCH (2008) Automated extraction of data from text using an XML parser: An earth science example using fossil descriptions. Geosphere 4 (1): 159-169.

Darwin Core (2008) http://rs.tdwg.org/dwc/

Hagedorn G, Thiele K, Morris R, Heidorn PB (2005) The Structured Descriptive Data (SDD) w3c-xml-schema, version 1.0. http://www.tdwg.org/standards/116/.

Hagedorn G, Mietchen D, Morris R, Agosti D, Penev L, Berendsohn W, Hobern D (2011) Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information. ZooKeys 150: 127-149. doi: 10.3897/zookeys.150.2189

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

Heidorn PB, Cui H, Yu B, Wu J, Zhang H (2002) Taxonomic description creation, search and display in XML. Abstract. Botany 2002. http://www.isrl.illinois.edu/~pheidorn/papers/Botany2002Abstract.pdf

Kirkup D, Malcolm P, Christian G, Paton A (2005) Towards a digital African Flora. Taxon 5 (2): 457-466.

LifeWatch (2008) Data & Modelling Tool Structures - Status Report on Infrastructures for Biodiversity Research - Deliverable 5.1.2 https://drive.google.com/a/plazi.org/#folders/0B_yrQwn4yBySVWVqb2g1ZlNQc2M

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

Marhold K. & Stuessy T (eds.) in collaboration with Agababian M, Agosti D, Alford MH, Crespo A, Crisci JV, Dorr LJ, Ferencová Z, Frodin D, Geltman DV, Kilian N, Linder HP, Lohmann LG, Oberprieler C, Penev L, Smith GF, Thomas W, Tulig M, Turland N, 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, 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

Murray-Rust P, Rzepa HS (2002) Scientific publications in XML - towards a global knowledge base. Data Science 1: 84-98.

Page, R. (2013) Something about links. Lecture at pro-iBiosphere Leiden workshop, February 13, 2013. http://www.pro-ibiosphere.eu/presentations/Surfacing%20the%20deep%20data%20of%20taxonomy.ppt

Parr CS, Lyal CHC (2007) Use cases for online taxonomic literature from taxonomists, conservationists, and others. TDWG Annual Conference, Slovakia. http://www.tdwg.org/proceedings/article/view/269.

Penev L, Agosti D, Georgiev T, Catapano T, Miller J, Blagoderov V, Roberts D, Smith V, Brake I, Ryrcroft S, Scott B, Johnson N, Morris R, Sautter G, Chavan V, Robertson T, Remsen D, Stoev P, Parr C, Knapp S, Kress W, Thompson C, Erwin T (2010) Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. ZooKeys 50: 1-16. doi: 10.3897/zookeys.50.538

Penev L, Roberts D, Smith V, Agosti D, Erwin T (2010) Taxonomy shifts up a gear: New publishing tools to accelerate biodiversity research. ZooKeys 50: 1-4. doi: 10.3897/zookeys.50.543

Penev L, Hagedorn G, Mietchen D, Georgiev T, Stoev P, Sautter G, Agosti D, Plank A, Balke M, Hendrich L, Erwin T (2011) Interlinking journal and wiki publications through joint citation: Working examples from ZooKeys and Plazi on Species-ID. ZooKeys 90: 1-12. doi: 10.3897/zookeys.90.1369

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, Georgiev T, Sautter G, 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/

pro-iBiosphere (2013) List of other biodiversity projects and initiatives. http://wiki.pro-ibiosphere.eu/wiki/List_of_other_biodiversity_projects_and_initiatives

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

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

Sautter G, Agosti D, Böhm K (2007a) Semi-Automated XML Markup 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

Smith V, Penev L (2011) e-Infrastructures for data publishing in biodiversity science. ZooKyes 150: 1-417. http://www.pensoft.net/journals/zookeys/issue/150/

Smith V, Rycroft S, Brake I, Scott B, Baker E, Livermore L, Blagoderov V, Roberts D (2011) Scratchpads 2.0: a Virtual Research Environment supporting scholarly collaboration, communication and data publication in biodiversity science. ZooKeys 150: 53-70. doi: 10.3897/zookeys.150.2193

TDWG (2007 onwards) TDWG: standards. Biodiversity Information Standards. http://www.tdwg.org/standards/

Triebel, D., Hagedorn, G., & Rambold, G. (2012) An appraisal of megascience platforms for biodiversity information. MycoKeys 5 (2012) : 45-63. doi: 10.3897/mycokeys.5.4302

Weitzman AL, Lyal CHC (2004) An XML schema for taxonomic literature – taXMLit - http://www.sil.si.edu/digitalcollections/bca/documentation/taXMLitv1-3Intro.pdf

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

Willis A, King D, Morse D, Dil A, Lyal C, Roberts D (2010) From XML to XML: The Why and How of Making the Biodiversity Literature Accessible to Researchers. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Valletta, Malta: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/787_Paper.pdf

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