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)**

D3.2.1 Concept paper for involvement of individual experts, commercial vendors, and citizen scientists

Deliverable Status:	**Final**
File Name:	**pro-iBiosphere_WP3_Plazi_VFF_30052013.pdf**
Due Date:	**31 May 2013 (M9)**
Submission Date:	**31 May 2013 (M9)**
Dissemination Level:	**Public**
Task Leader:	**Donat Agosti (Plazi)**
Authors:	**D. Agosti, T. Catapano, S. Eckert, Q. Groom, A. Güntsch, G. Hagedorn, P. Hovenkamp, E. Kralt, L. Penev, S. Sierra**

European Commission

Copyright

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

Consisting of:

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

Disclaimer

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

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

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

REVISION CONTROL

Version	Author	Date	Status
1.0	Donat Agosti	14.5.2013	Initial draft
2.0	Anton Güntsch, Sabrina Eckert	29.5.2013	Comments
3.0	Terry Catapano	29.5.2013	Comments
4.0	Gregor Hagedorn, Peter Hovenkamp, Quentin Groom	30.5.2013	Comments
5.0	Donat Agosti, Gregor Hagedorn, Lyubomir Penev, Soraya Sierra, Eva Kralt	30.5.2013	Final version

Contents

Executive summary

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

The pro-iBiosphere project is divided into Work Packages (WP), each of them being sub-divided into Tasks (T). One of the project tasks (Task 3.2) consists in Task 3.2 Semantic mark-up generation, data quality, and user-participation infrastructure. This objective is namely reached by Plazi.

The present deliverable (D3.2.1), prepared by Plazi (Project Task Leader), is the final report related to this activity.

The objectives of the concept paper are to:

* identify stakeholders interested in semantic mark-up of biodiversity literature, such as authors, citizen scientists (amateur naturalists, or people generally interested in semantic mark-up) and commercial vendors of this kind of services (e.g., publishers, repositories, data mining companies),

* analyse the cross-points of collaboration between them to formulate a strategy for integration and interoperability of services and data.

* consider the required or desirable incentives and others means to get the collaboration off ground.

For this reason, a workshop has been convened in Leiden in February (See Appendix 2) including representatives from the above mentioned stakeholder groups (See Appendix 3). This report has been based on representative, specifically invited lectures covering the spectrum, and blocks of discussion.

An important source to seed and populate the proposed Biodiversity Knowledge Management System is the vast corpus of hundreds of millions of pages of legacy taxonomic publications and the integration of the content of on-going publications. This includes well over 17,000 treatments of taxa new to science and estimated 50-100,000 re-descriptions annually, scattered in over 2000 journals and books. This corpus comprises largely unstructured publications geared towards the human reader, but not to machines that could assist in the ability to manage and derive knowledge from the deluge of information.

A decisive element in providing access to the vast corpus of legacy biodiversity publications is the conversion from printed, and to a large degree digital copies in Portable Document Format (PDF), copies into semantically enhanced documents that can be read by machines. The conversion of millions of pages and articles is a complicated process that cannot be done without a massive support by many possible stakeholders, incentives to perform the work, auxiliary infrastructures like ontologies to define the terms and their relationships used in the semantic mark-up, and software, workflows, and systems to effectively perform the work.

Introduction

The virtue of the Internet is that machines do most of the data searching, analysing and visualization to support our scientific work. A rapidly increase in storage capacity, processing power, and content provides an environment in the which Semantic Web envisioned by Tim Berners Lee could emerge. The knowledge of biodiversity is an ideal candidate to benefit from Semantic Web technologies and approaches. There are hundreds of millions of published records accumulated during the last 250 years. Billions of specimens exist in collection in a global network of natural history collections many of which are referenced in publications (e.g. the Type material in every of the original description of the estimated 1.8m species), rapidly increasing genetic and visual data, and not least a tradition to publish new taxa in a highly structured way, adding about 17,000 new taxa and an unknown number of re-described taxa to this body of knowledge annually. Increasingly, links or at least a list of detailed basic data (DNA sequences, morphological data , materials observation data, additional digital media) are provided in taxonomic publications, all resulting from a curation and assessment process during the underlying scientific study. This trove of data, essentially representing the current state of knowledge, appears to be one of the best sources for research. When considering the long time series and global geographic coverage this corpus of knowledge is a treasure chest for understanding our environment.

Where do we stand? Impediments and options

The field of Biodiversity has not yet taken full advantage of the technologies available to it. Why has this community lagged behind? A wide range of issues are involved, particularly in the current capabilities and activities in the conversion of legacy literature, including the ongoing publishing of unstructured publications (both print and PDF). These span from problems such as the choice between publishing static documents vs. semantically enhanced documents, open access vs. restricted access, to technical issues hampering the conversion of legacy data into a form that can be used in a semantic web, and finally to funding.

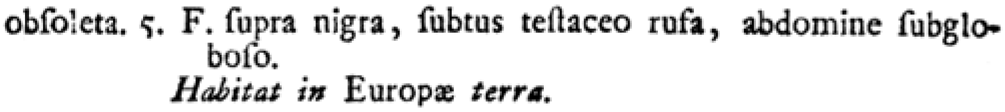
Defining the goal

Why mark-up taxonomic literature?

Biodiversity literature, specifically taxonomic literature describing the diversity of the world's species, their evolutionary relationships, and their biogeography is a vast corpus including several hundred million pages from well over 250 years when the printed record of biodiversity science officially began. This accumulation of data and information forms a huge stock of knowledge (Appendix 1), which, not least by virtue of its well-structured content, is the ideal source of data for a biodiversity knowledge management system (as proposed in the current EU-grant). This is complemented by the well over 1.5-3 billion specimens in natural history collections holding distribution records that are often the only way to study changes in distribution patterns over time or to study the evolution of the living world (Duckworth et al., 1993). Global initiatives such as the Intergovernmental Panel on Biodiversity and Ecosystem Services (IPBES) depend on the analyses of the published record to come to conclusion on the state and change of global biodiversity. Another option is to extract all the described taxa to create the still not yet existing list of the world's species, and to link each name to a treatment (a text including the description, nomenclature, distribution, behaviour and other elements of each taxon) a primary goal by 2020 of the planned World Flora Online in response to the Convention on Biological Diversity target to halt loss of species by that date.

Such initiatives define which content in the legacy record ought be marked up and made accessible for machine processing, data mining, and reuse. For example, the EU funded EU-BON project, one of the main contributor to IPBES will need observation data for their model calculations, the list of the world species needs names drawn from the literature, and the World Flora Online names and treatments.

Though names can be relatively easily extracted, as long as the text is properly digitized and not distorted by Optical Character Recognition (OCR), other more complex content, like observation data also needs extraction as well as linking to a particular taxon, or even entire treatments. Identification and mark-up of treatments, the units of taxonomic communication, helps reduce costs of further, more fine grained, mark-up, but requires some effort upfront. Semantic mark-up also allows reuse the content for other purposes if exposed properly.



obſoleta. 5. F. ſupra nigra, ſubtus teſtaceo rufa, abdomine ſubgloboſo.
Habitat in Europæ *terra.*

Fig. The two most common elements in taxonomic work: Scientific names and treatments, illustrated by Formica obsoleta, Linnaeus 1758: 580.

Increased granularity in mark-up, down to the level of observation data parsed to its basic elements, and morphological descriptions to characters and their states demands greater costs and the assistance of special purpose programs to help perform mark-up automatically or at least semi-automatically.

A final challenge faced in the conversion of the documents is that they have been written over many generations and with the human reader in mind. A consumer that can combine content that is often only implicitly provided has the advantage over a machine that has problems with incomplete data. Furthermore, often only the information relevant to make a particular case is given, which makes it often difficult to compare data across a corpus from a very different perspective. Faced with limited resources for the conversion process, a major task of pro-iBiosphere is to answer the question which part of the legacy literature has to be given priority in the conversion process, based on the need to answer particular research questions or specific scientific challenges.

Technical issues

One of the fundamental challenges to the encoding of legacy publications is that while they are relatively homogeneous in what they convey, they are very heterogeneous both in terms of content and form of presentation. For example, the taxonomic literature, probably one of the most structured forms of scientific literature in its focus on treatments each documenting a single taxon, spans more than 250 years of published record and several hundred million pages (BHL). During this time printing methods have changed, materials have changed, fonts have multiplied and changed, the use of different languages exploded after Latin was abandoned as the scientific lingua franca, and, last but not least, that fact that human readers (not machines) were the primary consumers of the information permitted a wide range of freedom for idiosyncrasies and variation in editorial standards in the way journals and books could be published and still be understood by the reader. An illustrative example are bibliographic references that can come in very different forms, ranging from highly detailed to very reduced, for instance " L.", which, while simply referring to Linnaeus, in context may have to be understood as a reference to "Linnaeus, C., 1758, Systema Naturae…", or even to a specific page in that work where the name has originally been published.

There is not only this diversity that complicates matters; there are also technical issues in the capture of data through OCR. Roman fonts from the period before about 1850 often employ the long S and ligatures which defy OCR software as do characters in scripts such as Fraktur. Bad printing quality and damage to the primary documents also present challenges. The unclean text resulting from these issues causes serious obstacles to machine processing. In order to generate sufficiently clean text it may be necessary to use double keying or more sophisticated OCR approaches which employ multiple engines and a "voting" mechanism to determine the best reading of a given character or word. Also needed is a clear understanding of how much of the original formatting needs to be retained and for what purposes: is only machine processing and data extraction required, or also accurate representation of the look of the original publication?

Mark-up issues

Issues similar to those posed in the OCR process occur during the mark-up process. To deal with hundreds of millions of pages of legacy literature, machine processing at industrial strength is needed. The *status quo* is unsatisfying.

Even clean text may provide impediments for successful mark-up. While a relatively clean OCR is sufficient for humans reading the text, for machine processing this leaves artefacts, many of which are caused by structural issues such as line, column, and section breaks, missing paragraphs or misinterpreted fonts. Thus, mark-up is not just semantically enhancing text but depends upon a high degree of need for text accuracy in order to be optimized for processing. With each step from the scanning to the OCR to mark-up, the difficulties multiply. Errors tolerated in the initial text capture later lead to problems in mark-up and require correction, at probably even larger cost, in subsequent phases of preparation of digitized text for any use beyond gross search and retrieval.

Existing tools for markup include the general purpose GoldenGATE mark-up editor and special purpose scripts and processes developed for particular journals or floras (Flora of Tropical East Africa, Flore d'Afrique Centrale, Flora Malesiana or Biologia Centrali Americana) at the other.

Mark-up workflows are generally based on insider knowledge, often retained by one single producer of the software. In other cases they are partly documented (GoldenGATE) but still retain a steep learning curve. Programs used are either in open source (GoldenGATE) or are project specific and not publicly distributed. Greater effort must be made to either to develop generic yet customizable software, or develop reusable single purpose tools in widely employed scripting languages such as Ruby, Python, or PERL for particular tasks common to most text clean-up and markup activities. Needless to say, such tools should be open source and distributed in code repositories such as GitHub in order to foster community development.

Document repositories

The conversion of legacy literature is dealt with in some projects by extracting the data that is then hosted in dedicated databases (e.g. RBGK: Flora of Tropical East Africa; Naturalis: Flora Malesiana; NBGB: Flore d'Afrique

Centrale). The processed document is often not widely available. In other workflows (e.g., the Plazi workflow) the entire marked-up document is retained and kept in the dedicated Search and Retrieval Server (SRS). Because of copyright reasons the document is not openly accessible, but used for further mark-up steps using the GoldenGATE tool. Only the extracted treatments are displayed and accessible to download services, which among others feed into the Encyclopedia of Life or AntWeb, which is a valuable outcome of the mark-up process. A link is provided to the original publication, and it is planned that links will be made from the Biodiversity Heritage Library, the holder of the original document, to the extracted treatments on Plazi.

By adding a unique persistent identifier to each treatment, as is planned by Plazi, treatments can be directly cited and linked to from subsequent publications. If a repository provides machine readable data in formats such as RDF, JSON, or XML then additional value is gained and potentially opens up the content to semantic web applications.

Documentation and standards

All mark-up workflows are still in the early stages of their development and are often elements of projects undertaken to solve one particular task. As consequence, little effort has been made to document the tools and engage in a professional process to create widely used robust practices and procedures.

When workflows are provided with proper human-machine interfaces for the programs used, manuals that guide through the mark-up process, a help desk function, and training, mark-up projects might proliferate.

Similar to the workflows and programs, the tag sets that are used for mark-up (e.g., TaxPub for prospective, TaxonX and TaXMLit for legacy publications (Penev et al. 2011)) remain somewhat underdeveloped and immature. Although there are some definitions of mark-up elements ("tags") available, there is only a rudimentary documentation and a lack of guidelines for application. This might have great impact , for many of the concepts behind the tags are new or obscure.

Semantic enhancement depends greatly on the presence of ontologies that define the concepts of interest and the relations between them. There is increasing awareness that such ontologies should be built (Shotton, 2009). Examples are the Citation Typing Ontology (CITO) or the Hymenoptera Anatomy Ontology.

Social issues

In the traditional scientific process preservation and access to legacy publications is delegated to libraries, and individual scientists extract their knowledge directly while reading publications. The currently dominant format for publications, PDF (Portable Document Format) demonstrates how that focus of publishers on human readers, and reader's habits, has persisted. Indirectly, the sheer technical difficulty of extracting fine grained data from the content of a PDF document is another indication that machine readability not considered as a priority. As a consequence, an infrastructure is needed to incorporate the information locked inside the printed content into a seamless knowledge management system. This requires a significant amount of effort to convert printed material, and in many cases also born digital material, into semantically enhanced documents that allow linking, retrieval, and analysis necessary of scientific inquiry. Consequently, not only is more storage space needed, but, like in a chemical reaction, so is activation energy to be able to transform the dominant print/human readable information model to one that can capitalize on the development of sophisticated machine assisted analysis, retrieval, and linking in the semantic web. Since this is not seen as part of the traditional scientific process, it needs additional resources beyond what is currently contributed by very small scientific projects such as the conversion of the Flora of Tropical East Africa, or descriptions covering particular taxa like Solanaceae or for regions like the ants of Madagascar (Plazi and AntWeb).

In most cases, the conversion process is initiated by editors (e.g. the European Journal of Taxonomy; Smithsonian Institution, Flora publishers), aggregators (GBIF and Encyclopedia of Life providing respective grant money to convert journals and have access to material citations and treatments respectively) or publishers (Pensoft, upfront mark-up). The motivation is to provide the best possible access to their content by using additional dissemination channels based on an XML version of the publications. All the publishers involved have become aware of this possibility by attending conferences related to the future of taxonomic publishing.

In contrast, several training courses held by Plazi in Australia, Brazil, and Europe did not result in a movement to get authors and users involved in the mark-up process with the exception of the training course held in January in the framework of the pro-iBiosphere pilot studies (T4.2) which has so far lead to a group of users that continue to convert document, and the integration of markup into a workflow from printed material via Plazi into the EDIT platform at the

BGBM. A possible reason for this lack of uptake is that the GoldenGATE interface and manuals were not yet fully developed and stable at the time of these earlier training courses.

The lesson learned might be that despite the difficulties with the GoldenGATE interface and manuals, other aspects seem to be a more decisive factor: conversion and mark-up is not as an end to itself, but provides the opportunity to reuse content in various ways in the future. Conversion is just one of many steps. The help for GoldenGATE is presently being improved and, unlike in earlier training courses, this time a follow-up and help desk has been provided. The concept of the treatment, only recently established, and its value is increasingly being recognized, among others by the main digital biodiversity library, BHL, which will include links from the scanned objects to these extracted subparts. Interestingly, PubMed Central which archives the TaxPub based journals by Pensoft (Zookeys, Phytokeys) decided by themselves, due to their prominent role in taxonomic publications, to extract the treatments from the TaxPub-xml documents and provide them individually as supplementary materials to the archived publication.

An additional move into conversion to semantically enhanced documents might result from an idea discussed at the pro-iBiosphere workshop (Coordination and Cooperation, T2.1) in Berlin which essentially focuses on re-publishing floras and faunas as semantically enhanced documents, including links to type material, additional media, and name servers. In this concept, simple digital conversion no longer is the dominant end goal, but focus is shifted to the creation of digital publications the production of which includes mark-up that will be the basis for a wide dissemination.

Scientific motivation

Mark-up of legacy literature is expensive and as such needs to be targeted to the most urgent needs. Unlike the Biodiversity Heritage Library digitization, which was driven in the early phase by efficiency considerations and availability of material, smaller scale mark-up projects have generally been initiated through specific project requests, such as converting a Flora (e.g. Flora Malesiana and Tropical East Africa), marking up all the literature of a region for a particular taxon (AntWeb: Ants of Madagascar), or taxa (Solanaceae), sometimes by conversion of journals (Zootaxa part), or requests for content from a particular source (EOL project).

Motivation for larger conversion projects might come from large scale projects, like the World Flora Online, which in order to finish a description of each taxon in time by 2020, depends on access to published content, i.e. treatments, or from the quadrennial assessments by the Intergovernmental Platform for Biodiversity and Ecosystem Services, whose analyses are to a large degree built on the analyses of the printed record. At a smaller scale, each PhD study or monographic work is built on a thorough analysis of the published record, and thus might be used to build small corpora of literature that includes all the records of a particular taxon.

At the same time, mark-up of prospective publications should be given the utmost importance and it should become the standard in publishing in the biodiversity domain. Prospective mark-up that is directly embedded in the publishing process is not only an asset for future reuse of content. It semantically documents the underlying scientific question that were addressed and answered in a given publication. (Penev et al., 2010, 2012) and acts immediately to revolutionize the process of scientific discovery and the scientific discourse.

Taxonomic publications are very rich in data, and especially if they are semantically marked up, allow data mining and querying for a wide array of questions (see Appendix 1).

Linking and identifiers

Linking data to their sources in a sustainable way depends on stable identifiers. The publishing industry has one dominant identifier scheme for their publications, the DOI (Digital Object Identifier). However, a similar mechanism is not yet generally accepted in the biodiversity domain. Development and deployment of stable identifiers has stalled over the last years. One reason is that the biodiversity domain made a decision to adopt Life Science Identifier (LSID) briefly before the general and widespread adoption of the Semantic Web and Linked Open Data. However, the software and computing infrastructure of the LSIDs, originally developed by IBM, were soon abandoned by all other players except the Biodiversity Community. In the following years, the biodiversity community failed to sufficiently maintain the software and computing infrastructure for LSIDs, but the decision in favour of LSIDs was not easily reversed, since several important projects had made major investments into this.

One alternative to LSIDs is to use DOIs in the biodiversity domain. Several problem exist with this. DOIs returning publication metadata for article and data sets, but not the data about a specimen, lack of support for other use cases, and a cost model that does not easily scale to the millions of DOIs required for specimens and names in the biodiversity domain. Even in the publishing sector, the fact that only publishers can assign DOIs to work, including to out-of-copyright legacy literature is limiting the adoption of DOIs.

The result is that no comprehensive identifier infrastructure exists at the moment and the amount of material that has been assigned identifiers (and thus interlinked) is limited and has been restricted to particular projects that could create content and controlled identifiers within their own limits. Recognizing the vital importance of having stable identifiers available for semantic linking, the pro-iBiosphere project invested a considerable amount of effort in several of its workshops held so far in addressing this problem.

The recently proposed use of Semantic Web and Linked Open Data stable URIs (Hyam et. al, 2012) is now quickly being adopted as the solution of choice in the semantic Web, and is increasingly used by Natural History Institutions for their specimens (Edinburgh Botanical Garden, Berlin MfN and BGBM).

With a feasible system in place, an increasing amount of specimens and other objects may be tagged with an identifier and a critical mass might be achieved that will be a sufficient incentive to actually make linking not only a hot topic, but also a standard procedure, fostering motivation and means to mark-up legacy literature so as to expose the rich data contained with publications.

Funding

Funding agencies have so far put a substantial effort into conversion of printed materials into digital form. However, the effort to scan as much as possible and the emphasis on the image of the page (BHL; JSTOR, Elsevier,) targets the human reader by producing PDF documents. There is not a substantial amount of funding devoted to cleanup and enhancement of the text of publications so as to prepare it for more effective analysis, retrieval, and linking of the data locked inside the printed page.

Large scale funding has been focusing on journal runs and books (BHL) or journals (JSTOR) which might not have fully covered stakeholder needs. The traditional citation practice to specific pages in a taxonomic publication containing a prior description of a taxon (e.g. Linnaeus 1758: 502) is an indication that the real need for users is to access content on the level of very specific document components below the level of journal or article, such as the treatments. Thus an ideal corpus of taxonomic literature would be one that is compiled specifically to answer research questions, for example one covering all the treatments of a particular taxon, or all the taxa of a particular geographic entity (e.g. Madagascar). This of course conflicts with the scanning workflows of large scale and well-funded mass digitization efforts for which the preparation of a publication is costly and thus even scanning individual journal issues might not be feasible. However, BHL recently put in place a service to allow users to request publications for scanning so as to influence prioritization of selection for scanning.

Business opportunities

Enhancement of the large corpus of BHL literature is only now being discussed. However, commercial vendors have not been attracted to the endeavour due to a perceived lack of business opportunity (see Funding below). Since OCR might not provide an adequate degree of text accuracy for more advanced and scientifically richer applications, there is a great need to establish how much of semantic enhancement can be done routinely by a commercial vendor, and at what level of granularity domain specialists are needed.

It is perhaps worth investigating whether the inclusion of semantically enhanced text could be part of the process of republishing classic texts. This might generate a market for such products that would implicitly pay the costs of conversion (see Social issues). Similarly if data mining of published records will become a standard practice (e.g. within IPBES or the World Flora Online), funding might become available.

Community involvement

The involvement of the community has not yet been deeply investigated. The high profile of natural history in the public, illustrated, for example, by the highly active bird watching community, indicates potential for the involvement of the public at large. The question though is for what, and then how? A possible incentive to contribute might be the option to create content that can immediately be reused, or at least will show up on high profile sites like EOL.

The involvement of stakeholders in the conversion process

The conversion of legacy literature is a slow and minimally coordinated effort. However, the results of this conversion become increasingly valuable for scientific research and for the general public. The following strategies to involve individual experts, commercial vendors, and citizen scientists, try to outline a way ahead.

Individual scientists and authors

Motivation: Individual scientists are motivated to make their own content as widely available and published as possible. They also need to get recognition for it in a way that advances their career. Open access is clearly the undisputed solution, and many scientists copy and paste content from taxonomic publications onto their research Web Sites (e.g. Solanaceae, AntWeb). The rapid growth of Zootaxa (even if it only provides limited access to PDF copies) and more recently Zookeys and Phytokeys covering more than 25% of the annual output for treatments of taxa worldwide are very indicative.

Why does it not (yet) work: So far there is a lack of awareness of methods that can help create semantically enhanced content (see the Pensoft model (Penev et al., 2010) as well as a lack of awareness of the service that such a semantically marked up content would provide to the future of science. Furthermore there are concerns about the sustainability of the OpenAccess business model. Other problems are that tools to convert legacy publications are cumbersome and not well documented, that there is a lack of knowledge about repositories for marked up publications and treatments and that a culture exists of separating the publication itself from auxiliary materials like supporting data. Furthermore, while substantial incentives exist for people to mark-up small parts of documents that are relevant to their research, there is too little incentive for them to systematically mark up complete works at fine detail.

Possible Incentives: Easy access to embedded or appended media that is possible in digital publishing (such as, images, audio, video, interactive data visualisation, etc.) at the moment of publications. Infrastructure for converting, linking and storing coupled with awareness campaigns and training opportunities. Persistent identifiers can be used to link content. For prospective publishing (“avoidance strategy”) make it as widely known as possible. Awareness that mark-up is removing an impediment and contributes beyond one’s own work to a much wider communities. Links from large repositories (e.g. Biodiversity Heritage Library) to converted mark-up and being credited for the work.

Commercial vendors

As **Commercial vendors** we here consider commercial service providers that produce digitized content, either as simple or semantically marked-up texts, from the printed record, irrespective of the method of digitization, such as Optical Character Recognition (OCR) or re-keying.

Motivation: The motivation is to generate an income from providing conversion and mark-up services.

Why does it not work: There is a very limited market and few sources of funding. Most of the efforts to convert printed content aim at capturing images and making them accessible, with little emphasis on text capture. Mark-up in this domain is a cottage industry driven by funding for particular tasks (e.g. Flora of Tropical East Africa through the Royal Botanical Garden Kew; Flora Malesiana by Naturalis) or by small scale grants (eg Encyclopedia of Life to Plazi to convert Zootaxa content) or by individual scientists (e.g. Pilot projects within the Pro-iBiosphere Project). Mark-up is very domain specific and needs a combination of vendor and domain scientists, a combination that is not yet properly established.

Possible incentives: Funding and clear specifications of the tasks. An established production workflow with a vendor involvement will most likely not be feasible if too much mark-up becomes dependent on domain knowledge. Mark up at treatment level might work, since this is highly structured text. See also the Publisher section below.

Citizen scientists and crowdsourcing

Motivation: There is a great potential to involve citizen scientists in the scientific process. Contributing and getting credit, both directly and indirectly by being a member of a visible project with which the participant can identify may provide a motivation for participants. Another possibility might be getting some very limited financial compensations for each successful task. Crowdsourcing has been shown to be very effective in transcribing field books (e.g. Australian Museum Volunteer Portal)

Why does it not work: Stakeholders are not properly addressed. The tasks are not offered in a way that engages the stakeholders (e.g. too many steps that do not yield a result, or credits that are not recognized in their value, either within a community or financially). Stakeholders cannot be located and addressed. Work to address stakeholders is done by non-specialists.

Possible incentives: Collaborate with professional crowdsourcing specialists (eg. Chamberlain et al., 2012). Make use of volunteers whenever they are accessible. Understand and translating the stakeholders mind-set and value system. Get feedback from the project and get a system of credit for work done; create a competition in the community with clear communication of the winner and prizes for the winner.

Funders

Motivation: Science foundations, other funders as well as natural history museums want to have the best possible communication of the scientific results they support. A shift from printed or otherwise isolated publications to other communication vehicles will allow improved quantification of the use of their assets, like specimens in their collections, images, treatments, which digitization and assignment of web resolvable identifiers are increasingly make possible. Extracting such entities from the published record is a shortcut to the digitization of ultimately billions of specimens.

Why does it not work: The full impact of new technologies and approaches that are now being discussed need time to be fully understood.

Possible incentives: Appropriate metrics, such as altmetrics allow measuring the usage of content. Funding allocated to the conversion of legacy publications, installing and maintaining auxiliary databases like ontologies needed in the Semantic Web, and business plans that allow planning of sustainability of the investment into the infrastructure (e.g. Shotton, 2009)

Publishers and editors

Motivation: In addition to using an Open Access business model to finance sustainably the creation of semantically enhanced content, publishers might be interested in republishing major taxonomic works in a semantically enhanced format.

Why does it not work: A new idea that has not been tested.

Possible incentives: The conversion of legacy literature is motivated by providing access to the content in a form that is widely used, such as a flora, fauna or monograph. These may be enhanced and provide the basis for the development of life floras and faunas. Funders might be interested to support this work, since it serves their purpose of providing the widest possible access to a community that is increasingly based on the Internet.

Conversion tool specialists

Motivation: Building tools to convert and maintain specific databases. Funding and scientific credit.

Why does it not work: There are few incentives to create tools. Plazi created with GoldenGATE a generic tool for mark-up of treatment and included elements, but so far it has remained a science project, not a commercially viable one. Despite training courses, the complexity of legacy publications (highly variable printed structure, OCR errors, journal specific special layouts, variation of ways to present content) coupled with difficult to use human interface of the software show little progress in getting trained users. The development of the GoldenGATE human machine interface has never been properly funded. Only recently funding could be obtained to improve the situation (EU funded EU-BON to Plazi).

Possible incentives: Funding of projects and demand will be the drivers. If research- or conservation needs start to increase the demand for access to the printed record (e.g. IPBES to assess trends in biodiversity), this might stimulate a wave of technical developments that will make it easier for the above stakeholders to develop mark-up tools.

Incentives and support actions for involved stakeholders

In order to stimulate a wave of technical developments and encourage involved stakeholders to invest into mark-up tools, suggestions of incentives and support actions are listed below:

- Support for semantically enhanced publishing: Semantically enhanced publications and their integration in the (semantic) Web will set an example for others to follow
- Promote and foster the development and use of persistent identifiers for treatments, specimen citations, scientific names, and morphological characters and states
- Sponsor projects that make use of semantically enhanced publications, e. g. data mining
- Sponsor projects that are linked to projects calling for published content (e.g. EU-BON)
- Sponsor the production of best practices for production workflows
- Sponsor the development of best practices for mark-up. Define criteria for levels of data description.
- Sponsor the development of tools, procedures, and best practices to clean up and prepare data submitted to or harvested by a consumer, accepting that only rarely will data be readily interchangeable from one project or provider to another
- Fund and foster wide and open development of tools by individuals as well as institutions, both large scale software development projects, and small scale task specific tools which can be combined with other tools.
- Establish priorities for enhancement of digitized taxonomic literature, whether organized around a taxa, a geographic region, etc... so as to gain experience, derive generalizable solutions, and stimulate analyses taking advantage of the enhanced publications.
- Sponsor community mark-up tools that will foster the conversion publications into semantically enhanced publications.

Literature references and links

Antweb. http://antweb.org

Australian Museum Volunteer Portal, http://volunteer.ala.org.au/

Chamberlain, J., Kruschwitz, U., & Poesio, M. 2012. Motivations for participation in socially networked collective intelligence systems. Proceedings CI http://arxiv.org/pdf/1204.4071.pdf

CiTO, the Citation Typing Ontology. http://www.essepuntato.it/lode/http:/purl.org/spar/cito

Duckworth, W.D., Genoways,H.H., and Rose, C.L. 1993. Preserving Natural Science Collections: *Chronicle of our environmental heritage.* National Institute for Conservation of Cultural Property, Washington,D.C, USA.

EU-BON, http:// http://www.eubon.eu/

GoldenGATE, http://plazi.org/?q=GoldenGATE

GitHUB, https://github.com/

Hyam, R.D., Drinkwater, R.E. & Harris, D.J. Stable citations for herbarium specimens on the internet: an illustration from a taxonomic revision of Duboscia (Malvaceae) Phytotaxa 73: 17–30 (2012).

IPBES, Intergovernmental Platform on Biodiversity and Ecosystem Services, http://www.ipbes.net/

Marhold, K., & Stuessy, T. 2013. The future of botanical monography. Taxon 62(1): 4-20 http://www.ingentaconnect.com/content/iapt/tax/2013/00000062/00000001/art00003

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

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 for 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 mark-up in taxonomy, from the experience of a biodiversity publisher. Journal Article Tag Suite Conference (JATS-Con) Proceedings 2012.Bethesda (MD). http://www.ncbi.nlm.nih.gov/books/NBK100351/

Phytokeys, http://www.pensoft.net/journals/phytokeys

Plazi. http://plazi.org

Solanaceae Source, http://www.nhm.ac.uk/research-curation/research/projects/solanaceaesource/

Shotton, D.2009. Semantic Publishing: The coming revolution in scientific journal publishing.

Learned Publishing 22: 85–94.

TaxonX, http://www.taxonx.org/

Taxpub, http://sourceforge.net/projects/taxpub/

The Hymenoptera Anatomy Ontology. http://hymao.org/index.php/Main_Page

World Flora Online, http://www.kew.org/about-kew/press-media/press-releases-kew/plans-to-create-first-online-world-flora/

Zookeys, http://www.pensoft.net/journals/zookeys/

Appendix 1

Use cases "scientific questions" that stakeholder might want to be answered from a publication as guide to mark-up at what kind of granularity (see also Marhold & Stuessy, 2013).

Question	Answer	Example of content	Element type
What species is living where?	materials citation (single record)	Holotype worker, MADAGASCAR: Antsiranana, Parc National de Marojejy, Manantenina River, 28.0 km 38° NE Andapa, 8.2 km 333° NNW Manantenina, 14°26'12"S, 049°46'30"E, 450 m, sifted litter, rainforest, 12–15 Nov 2003 (coll. B. L. Fisher et al.), comma collection code: BLF08985 pin code: CASENT0104542 (CASC).	record
	distribution (summary)	The distribution is limited to collections made between 450 m and 750 m in rainforest in Parc National de Marojejy and 240 m from Ambanitaza near Antalaha	section
At what elevation is a species living?	materials citation (single record)	450 m	record
Which species is living in a specific stratum?	materials citation and ecology and section of various treatments		section
With what other species is x interacting?	biology and ecology sections of a treatment	taxa mentioned in a treatment	section
What are the parasitoids of taxon x?	biology and ecology sections of a treatment		section
What correclation exists between body size and geographic latitude?	comparison of data from description and materials citation section		section
What are the distinctive characters of a species?	diagnosis section of a treatment	Blade of mandible with five teeth and denticles located along distal two thirds of blade's length. Propodeum with short teeth (Fig. 5a).	section
What collections have been used to describe taxon x?	materials citation (single record)	CASC	record
What are the gene sequences derived	materials citation (single record) or		record, auxiliary

from taxon x?	auxiliary files of a publication		files
How can I identify a species?	diagnosis, key and description section of a treatment	Inner mandibular blade without preapical teeth and denticles (Figs 3a, 4a). . .2 Inner mandibular blade with at least four preapical teeth and denticles (Figs 2a,e). . .4	section
Is there a key to identify taxon x?	key section	Present / absent	section
What red species is living in country x?	combination of description and materials citation and distribution sections	Myrmic rubra Manica rubida	sections
What other work has been cited in the description of taxon x?	nomenclature section for treatment citations; anyhwere in the treatment	grandidieri Forel, 1891 = madecassus Santschi, 1928	section, record
What are the characters of taxon x with genome y	Characters and character state in description section	Mandidble with a row of declining teeth	section, record
What is scientifically known about taxon x?	treatment	[an entire treatment}	treatment
Who described the first time taxonx x?	nomenclature section	Formica rufa, nov. sp.	section
What taxa have been described in publication x?	publication	Formica pratensis Formica sanguinea	Marked up name
What has been the taxonomic hierarchy author x used in publication y?	publication	Formica - pratensis - rufa	publication

Appendix 2

Agenda

Legacy literature – Semantic mark-up generation, data quality and user-participation infrastructure

Leiden, February 13, 2013

PART I: Biosystematic literature: Where are we and where do we want to go?
Chair: Donat Agosti

8:30 – 9:00 What is a "flora"? – Peter Hovenkamp, Naturalis
9:00 – 9:30 Surfacing the deep data of taxonomy – Rod Page, University of Glasgow
9:30 – 9:45 Incentives for editors – Laurence Bénichou, EJT, Muséum Nationale d'Histoire Naturelle
9:45 – 10:00 Discussion

PART II: Mark-up and semantic enhancement of biosystematics literature Chair: Rod Page

10:00 – 10:15 Zootaxa / European Journal of Taxonomy pathway – Terence H. Catapano, Plazi
10:15 – 10:30 Pensoft Mark-up Tool (PMT) – Teodor Georgiev, Pensoft
10:30 – 10:45 Tea/coffee break
10:45 – 11:00 Semantic publishing: the revolution is here – David Shotton, Oxford University
11:00 – 11:15 A proposal for the use of a semantic vocabulary, taxpub as a starting point – Terence H. Catapano, Plazi
11:15 – 11:30 Questions & Discussion

Chair: Terry Catapano

11:30 – 11:50 Mark-up and parsing with Perl scripts – Thomas Hamann, Naturalis
11:50 – 12:10 Mark-up tools / GoldenGATE – Guido Sautter, Plazi
12:10 – 12:30 Charaparse – Hong Cui, The University of Arizona
12:30 – 12:45 Questions & Discussion
12:45 – 13:30 Lunch break

PART II (ctd): Mark-up and semantic enhancement of biosystematics literature
Chair: Donat Agosti

13:30 - 13:40 The BHL way to content - William Ulate & Chuck Miller, BHL, Missouri Botanical Garden
13:40 – 13:50 BioStor - Rod Page, University of Glasgow
13:50 - 14:00 Ontologies – Bob Morris, Harvard
14:00 - 14:10 Annotations – James Macklin, Agriculture and Agrifood Canada
14:10 - 14:30 What is and how to measure quality? – Christiana Klingenberg, Plazi
14:30 - 14:50 Discussion and way ahead

Potential topics:

- Where do we want to go?
- Lessons learned
- The issue of scaling
- How to get more content or how to get the crowd involved

14:50 – 15:05 Tea/coffee break
15:05 - 17:00 Discussion and way ahead (ctd.)

Discussion and small presentation to clarify points

17:00 - 17:15 Summary – Donat Agosti, Plazi

Appendix 3

The list of confirmed participants at the pro-iBiosphere workshops, Leiden 11th-14th February 2013.

- 1. Alan Paton — Royal Botanic Garden Kew, UK
- 2. Alexander Krings — North Carolina State University, USA
- 3. Andru Vallance — PracticalPlants.org
- 4. Ann Bogaerts — National Botanic Garden, Belgium
- 5. Anton Güntsch — Freie Universität Berlin - Botanischer Garten und Botanisches Museum, Germany
- 6. Atik Retnowati — The Indonesian Institute of Sciences, Indonesia
- 7. Bart Wursten — National Botanic Garden, Belgium
- 8. Benny Bytebier — University of KwaZulu-Natal, South Africa
- 9. Berry van der Hoorn — Naturalis, the Netherlands
- 10. Bonaventure Sonké — University of Yaounde, Cameroon
- 11. Bruce Hoffman — Flora of the Guianas
- 12. Camille Torrenti — Sigma, France
- 13. Chequita Bhikhi — Flora of the Guianas, the Netherlands
- 14. Christiana Klingenberg — Plazi
- 15. Chuck Miller — Missouri Botanical Garden, USA
- 16. Cyrille Chatelain — Les Conservatoire et Jardin Botaniques de la Ville de Genève, Switzerland
- 17. Daniel Thomas — Naturalis, the Netherlands
- 18. Dave Roberts — National History Museum of London, UK
- 19. David J. Patterson — Global Names, USA
- 20. David King — Open University, UK
- 21. David Shotton — University of Oxford, Department of Zoology, UK
- 22. Dimitris Koureas — National History Museum of London, UK
- 23. Don Kirkup — Royal Botanic Garden Kew, UK
- 24. Donat Agosti — Plazi, Switzerland
- 25. Eckhard von Raab-Straube — Freie Universität Berlin - Botanischer Garten und Botanisches Museum, Germany
- 26. Elisabeth Paymal — BioVel
- 27. Ensermu Kelbessa — Addis Ababa University, Ethiopia
- 28. Eric Pauwels — Centrum Wiskunde & Informatica
- 29. Erik Smets — Naturalis, the Netherlands
- 30. Erik van Nieuwkerken — Naturalis, the Netherlands
- 31. Eugenia Barnett — Royal Botanic Garden Kew, UK
- 32. Greg Riccardi — Morphbank, USA
- 33. Gregor Hagedorn — Julius Kühn-Institute, Berlin, Germany
- 34. Guido Sautter — Plazi, Germany
- 35. Hans Kruijer — Naturalis, the Netherlands
- 36. Heimo Rainer — Naturhistorisches Museum Wien, Austria
- 37. Helmut Knüpffer — Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany
- 38. Henry Engledow — National Botanic Garden, Belgium
- 39. Herman de Jong — Naturalis, the Netherlands
- 40. Hong Cui — University of Arizona, USA
- 41. Hui Yang — Open University, UK
- 42. Isa van der Velde — Royal Belgian Institute of Natural Sciences, Belgium
- 43. Jacqueline Henrot — Botanical consultant
- 44. James Macklin — Agriculture Canada, Flora of North America, Canada
- 45. Jana Hoffman — Museum für Naturkunde, Berlin, Germany
- 46. Jan van Tol — Naturalis, the Netherlands
- 47. Stefan Dressler — Senckenberg Gesellschaft für Naturforschung (SGN), Germany

- 48. Jeroen Snijders — Naturalis, the Netherlands
- 49. Jon Chamberlain — University of Essex, UK
- 50. Jonathan Timberlake — Royal Botanic Garden Kew, UK
- 51. Jordan Biserkov — Pensoft, Bulgaria
- 52. Jozsef Geml — Naturalis, the Netherlands
- 53. Laurence Bénichou — Muséum National d'Histoire Naturelle, France
- 54. Marc Sosef — Naturalis, the Netherlands
- 55. Marco Roos — Naturalis, the Netherlands
- 56. Marco Schmidt — Senckenberg Gesellschaft für Naturforschung (SGN), Germany
- 57. Mark Hyde — Flora of Zimbabwe
- 58. Mart Vogel — Naturalis, the Netherlands
- 59. Mike Gilbert — Royal Botanic Garden Kew, UK
- 60. Mohan Prasad Devkota — Amrit Science Campus, Nepal
- 61. Natacha Beau — National Botanic Garden, Belgium
- 62. Nicola Nicolson — Royal Botanic Garden Kew, UK
- 63. Patricia Kelbert — Freie Universität Berlin - Botanischer Garten und Botanisches Museum, Germany
- 64. Paul Kirk — CABI, UK
- 65. Pavel Stoev — Pensoft, Bulgaria
- 66. Peter Hovenkamp — Naturalis, the Netherlands
- 67. Peter Schalk — ETI, the Netherlands
- 68. Pulcherie Bissiengou — Flore du Gabon
- 69. Quentin Groom — National Botanic Garden, Belgium
- 70. Régine Vignes-Lebbe — LIS (+MNHN,GBIF-France)
- 71. Rich Pyle — Bishop Museum, Hawaii, USA
- 72. Rob Whitton — Bishop Museum, Hawaii, USA
- 73. Robert Morris - Harvard University Herbaria and Plazi, USA
- 74. Rod Page — University of Glasgow, UK
- 75. Ross Mounce — University of Bath, UK
- 76. Rusea Go — Universiti Putra Malaysia, Malaysia
- 77. Sabitrie Ramharakh — Flora of Surinam,Surinam
- 78. Sabrina Eckert — Freie Universität Berlin - Botanischer Garten und Botanisches Museum, Germany
- 79. Sander Pieterse — Naturalis, the Netherlands
- 80. Sandrine Ulenberg — Naturalis, the Netherlands
- 81. Siti Mumirah Bt Mat Yunoh — Forest Research Institute Malaysia, Malaysia
- 82. Soraya Sierra — Naturalis, the Netherlands
- 83. Stefan Dressler — Senckenberg Gesellschaft für Naturforschung (SGN), Germany
- 84. Stéphanie Morales — SIGMA, France
- 85. Steven Dessein — National Botanic Garden, Belgium
- 86. Susana Arias Guerrero — Naturalis, the Netherlands
- 87. Suzanne Sharrock — Botanic Gardens Conservation International, UK
- 88. Sylvia Mota de Oliveira — Naturalis, the Netherlands
- 89. Teodor Georgiev — Pensoft, Bulgaria
- 90. Terry Capatano — Plazi, USA
- 91. Thibaut DeMeulemeester — University of Mons, Belgium
- 92. Thomas Hamann — Naturalis, the Netherlands
- 93. Thomas Janssen — Humboldt-Universität zu Berlin, Germany
- 94. Tom Gilissen — Naturalis, the Netherlands
- 95. Tom van Dooren — Naturalis, the Netherlands
- 96. Tony Walduck — Royal Botanic Garden Kew, UK
- 97. Vincent Roberts — MycoBank, the Netherlands
- 98. Visotheary Riviere-Ung — LIS (+MNHN,GBIF-France), France

- 99. Walter Berendsohn — Freie Universität Berlin - Botanischer Garten und Botanisches Museum, Germany
- 100. Yde de Yong — Naturalis, the Netherlands
- 101. Ellinor Michel — International Commission on Zoological Nomenclature