Open Biodiversity Knowledge Management System (OBKMS)

The pro-iBiosphere project deals with the technical, legal and sustainability aspects of OBKMS

Funded by the 7th Framework Programme of the European Union

PRO-iBIOSPHERE
WWW.PRO-iBIOSPHERE.EU

- Consistent biodiversity information space
- Open and accessible research data
- Collaboration in data collection and re-use
- Convenient provision of high quality information
- New formats to support novel and diverse uses
- Prognostic modeling
- Linkages with other resources
- Improved markup processes and infrastructure
- Literature discovery and conversion
- Accreditation for researchers' work

Linked Open Data (LOD)

Resource Description Framework (RDF)

Unified system of persistent, stable, HTTP identifiers denoting and interlinking different biodiversity data elements (taxon names, taxon name usages, specimens, occurrences, taxon treatments, gene sequences, traits, references, images, multimedia, etc.)

GBIF | **Gen Bank** | **EOL** | **BHL** | **RefBank** | **EDIT** | **CDM** | **Scratchpads** | **Plazi** | **Other users**

Darwin Core Archives
Data
Metadata
EML description

Direct export

ABCD, BioCASE, TAPIR, DIGIR

Darwin Core Archives
Data
Metadata
EML description

Darwin Core Archives
Data
Metadata
EML description

Data export

Markup (TaxPub), data publishing

Markup (TaxonX), data extraction, re-publishing

Primary biodiversity data (collections, observations, gene sequences, images, traits, etc.)

Newly published literature

Legacy literature (Flora and Fauna series, taxonomic revisions, inventories, research articles)

Ten key outputs of the pro-iBiosphere project

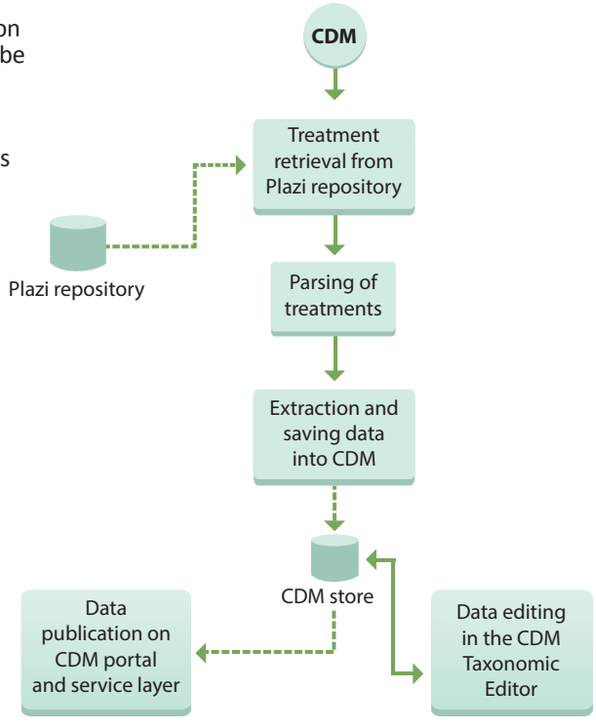
Key output 1: Improved cooperation and interoperability of e-infrastructures

Challenges related to the technical interoperability of biodiversity data present themselves in competing standards, ambiguous, poor or absent documentation, lack of stable identifier systems and the absence of semantic interoperability. For improving the interoperability between e-infrastructures, stable identifiers for biodiversity collection objects and a global service registry were identified as the two major achievables for progress. The use of state-of-the-art digitisation software & tools for literature markup is another important factor.

Step 1. Implementation of HTTP-URIs by 8 major institutions for their collection objects by October 2013, treatments, and recommendations for further topics to be explored in detail.

Step 2. Agreement on the BiodiversityCatalogue as a global registry for biodiversity related services. Improvement of the recommendations and guidelines for it to be able to fill this role even better; registration of services available now.

Step 3. Workflow improvement between the Plazi document registry and the Common Data Model (CDM)-based EDIT Platform for Cybertaxonomy (http://wiki.pro-ibiosphere.eu/wiki/Pilot_3). In the course of this a markup granularity table evolved. The pro-iBiosphere pilot portals visualize the data results at different stages and show the possibilities for scientists willing to markup their data. The markup granularity table explains in detail work load and connected output gain.



Lead partners: FUB-BGBM, Plazi.

Who is it for? Scientists, markup annotators, especially those who publish taxonomic novelties and want to gain digital content.

How is it used? The improved workflows and the agreed commitments to common standards and registries contributes to the construction of a service-based infrastructure for the mobilization, management and publication of biodiversity data. In particular, workflows for extracting scientific data from legacy literature can be used as a blueprint for future large-scale data mobilization activities.

Key output 2: Recommendations for use of persistent stable HTTP identifiers

The biodiversity community had long discussions about the preferred identifier schemes. Often this discussion has been dominated by the long term perspective of research, leading to the conclusion that any identifier scheme must be independent of current internet technologies. The pro-iBiosphere project investigated the arguments for various options. It quickly became clear that the previously adopted internet-independent solution of life science identifiers (LSID) has no future. The discussion between doi, ark and Semantic Web standards (http-based) identifiers is ongoing and all have advantages.

However, a consensus emerged on:

1. We need to interlink our objects and information into a Semantic Web of Knowledge.
2. We need to break the barriers between disciplines and use standards adopted across all disciplines.
3. We need to use standard software developed in large communities.

The conclusion is that Semantic Web compatible HTTP-based identifiers are a necessity, while other identifier schemes may be used in parallel. For these, the following important points were worked out:

1. Keep mission-critical URIs (or URIs, or IRIs, or web-addresses) for resources stable, whether providing Semantic Web data or not.
2. The choice of resources that need stable URIs is a management decision. Do not aim to keep all your institutions URIs stable forever. Requiring permanent stability for every web page may become unmanageable. Management needs to decide which are the most valuable resources an instution provices and issue clear directives to keep those URIs stable for decades to come. In our opinion, URIs for institutions, collections, specimens, geolocations, taxa, publications, treatments, traits and features need to be given priority.
3. Stability is easily achieved by currently installed tools, provided the choice of resources is clearly defined and IT staff has clear instructions.
4. Any pattern of dereferencable http-URIs will work. However, it is strongly recommended to keep the pattern simple and adding some parts that transparently map to resource classes, to allow handling different resources classes with different technologies in the future. A document on some recommendations has been prepared by pro-iBiosphere: http://wiki.pro-ibiosphere.eu/wiki/Best_practices_for_stable_URIs.
5. To fully participate in Linked Open Data and the Semantic Web it is necessary to provide URIs for physical objects and abstract concepts independent of the web information resources, link the two, and return RDF data where this is requested (through content negotiation).

Lead partners: Plazi, MfN, BGBM, PENSOFT.

Who is it for? All scientists in biodiversity and related areas.

How is it used? To date, some major institutions (e.g. BGBM, Harvard Herbaria and MCZ, RBGK, MfN, MNHN, Plazi, RBGE), have started to implement persistent HTTP identifiers.

Key output 3: Interoperability of taxon treatments

In the past, taxonomic information has been published in numerous scattered outlets and in different formats. The production of a taxonomic revision or such as a Flora or Fauna series required that the appropriate text was discovered and retyped manually. The current pilot demonstrates a greatly accelerated workflow that takes advantage of the informatics developments of pro-iBiosphere. The workflow locates, identifies, and enhances data included in treatments from both legacy and newly published taxonomic literature, facilitating discovery, analysis, and re-use through the Plazi Treatment Repository (PTR).

The workflow includes the following steps:

Step 1. Convert printed taxonomic articles/monographs to digital text format.

Step 2a. Markup generic document features and domain-specific information (taxon treatments) and store the results at Plazi; and also

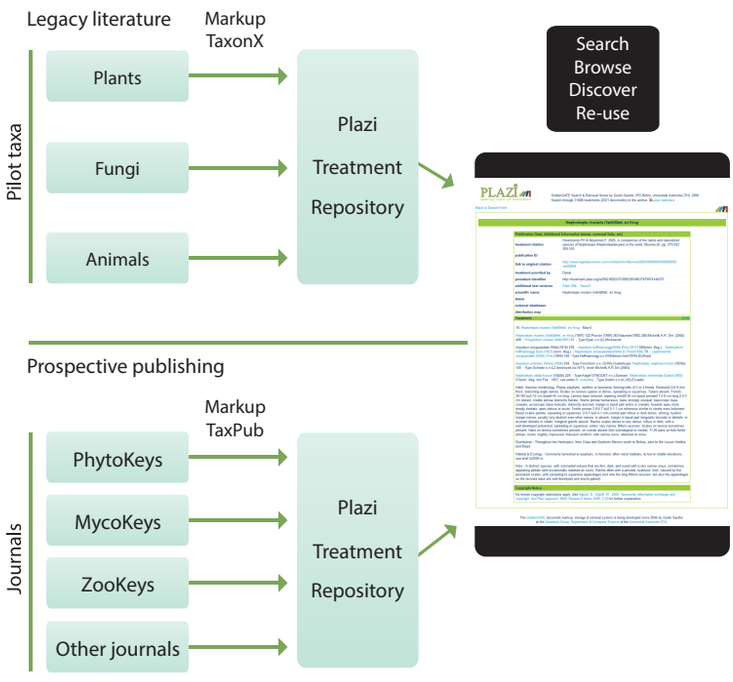
Step 2b. Export newly published treatments marked up during the editorial process (for example in the journals ZooKeys, PhytoKeys and Mycokeys).

Step 3. Browse, search, export and re-use treatments coming from different sources.

Lead partners: Plazi, PENSOFT, Naturalis, BGM, RBGK.

Who is it for? Taxonomists, aggregators of information on species (EOL, GBIF, Species-ID, etc.), bioinformaticians, ecologists, conservationists.

How is it used? To date, more than 1000 treatments have been marked up and stored at Plazi, in the course of the project. They are accessed to and linked back from newly published articles. Most of the treatments are exported to relevant aggregators, such as EOL, AntWeb, and EDIT CDM.



Key output 4: Removing legal barriers to the exchange of taxonomic information

Pro-iBiosphere addressed legal aspects that will arise as we build OBKMS. The exchange of information is hampered by claims of copyrights and database protection. We analysed taxonomic data and information and defined those areas in which copyright or database protection claims are not appropriate because of the standardized and systematic character of the information. We evaluated European copyright and database protection laws with respect to rules on the use of protected items for research purposes and came to the conclusion that the harmonisation of copyright rules within the EU, as it is envisaged in the EU Infosoc-Directive (EU-Directive 2001/29), is far from being realised. Copyright and database protection rules in the EU Member States differ not only in details, but in substance. This legal situation is a major stumbling block to the construction of a Europe-wide OBKMS and will need to be addressed.

Step 1. A list (the 'blue list', http://plazi.org/?q=blue_list) of components of taxonomic treatments to which intellectual property and database protection rights cannot apply.

Step 2. A list of specific elements of copyright legislation in European countries that show that exceptions and limitations provided for in the EU Infosoc-Directive (EU-Directive 2001/29) have not been transformed into national law at all, or have been transformed in ways that are inconsistent with practices elsewhere.

Step 3. Development of an information policy based on open accessibility of taxonomic data and information; the policy is accepted by EU-BON in form of a data sharing agreement.

Step 4. Promote the Bouchout Declaration launched at the pro-iBiosphere Final Event conference, June 12th 2014 to recruit support among the community of biodiversity specialists for open access.

Lead partners: Plazi, BGBM, PENSOFT, Naturalis, BGM, RBGK.

Who is it for? Our progress will better help scientists to know how to clarify the use of content they publish; it will inform publishers, and most importantly, those who wish to take data from other sources and incorporate them in their own researches and products.

How is it used? The pro-iBiosphere investigations into copyright established that considerable bodies of biodiversity information are being subjected, inappropriately, to copyright. Past and current practices and widespread misunderstanding of copyright law has impeded the readiness of data providers to share their content. The 'blue list' s a practical guide to the categories of information to which copyright does not apply, and so will remove uncertainty among biodiversity experts about sharing or re-using content. The Bouchout Declaration (http://bouchoutdeclaration.org) is a very visible device to allow the community to support the principle of open data; and to use this to identify, and remove, impediments to open data sharing.

Key output 5: Re-publication of biodiversity monographs in semantically enriched open access format

The bulk of the taxonomic information is closed in paper-based legacy literature, especially in fundamental regional treatises such as Flora, Fauna and Mycota series. The current pilot demonstrates a workflow that enhances the marked up content of Flora Malesiana and re-publishes it into an open access, semantically enriched HTML edition available on the newly launched, Advanced Books publishing platform (http://advancedbooks.org). With this, scientifically important historical monographs, enriched with additional information from up-to-date external sources related to taxon names, species treatments, morphological characters, etc., become freely usable for anyone at any place in the world, in addition to other benefits of the digitization and markup effort such as data extraction and collation, distribution and re-use of atomized content, archiving of different data elements in relevant repositories and so on.

The workflow includes the following steps:

Step 1. Convert printed or digitally-born monographs to digital text format.

Step 2a. Mark up generic document features and domain-specific information following the FlorML or TaxPub XML schemas.

Step 2b. In case the marked up text is available in FlorML, convert it to a TaxPub-based XML file.

Step 3. Provide additional markup, hyperlink and enhance content of the XML file.

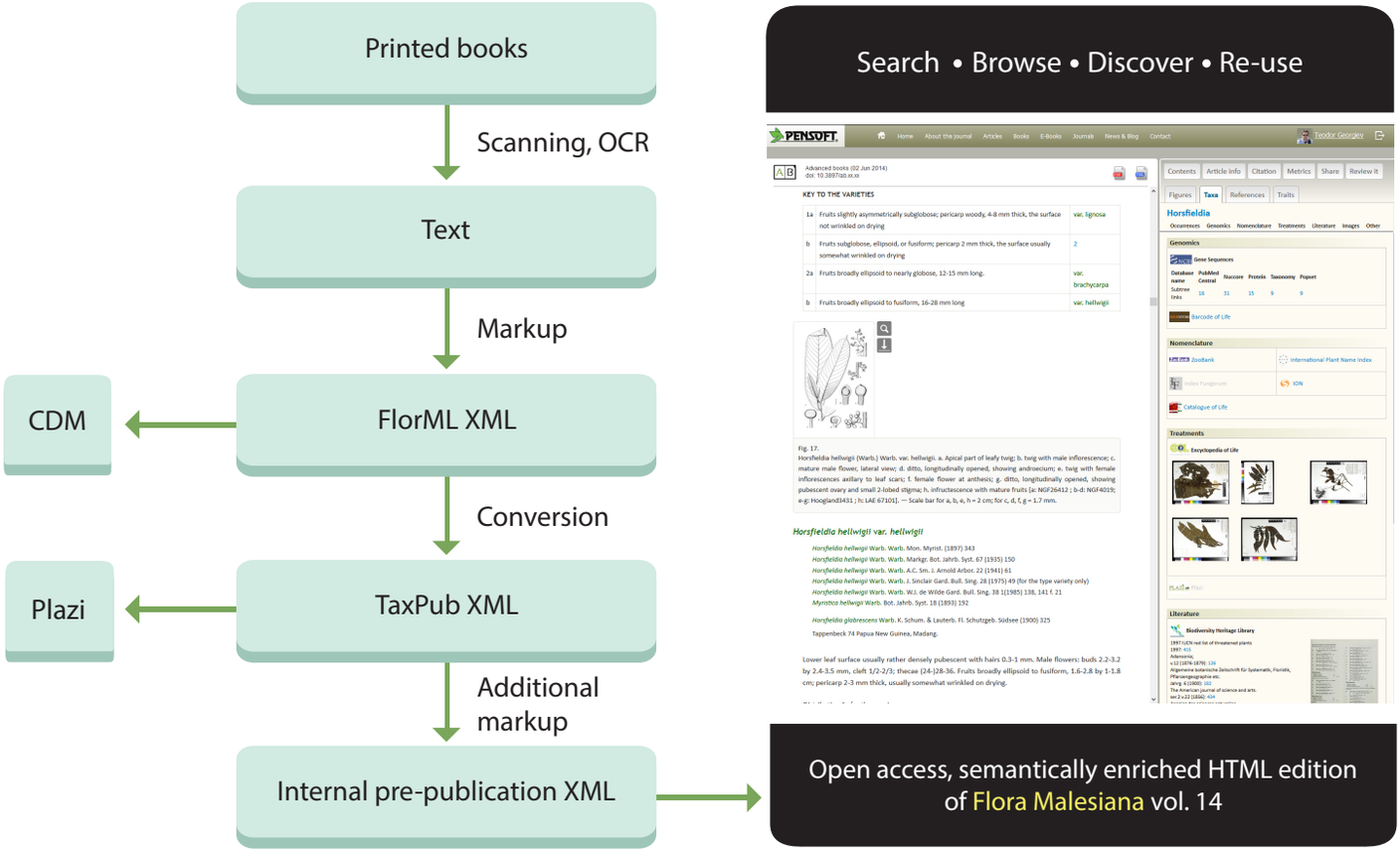
Step 4. Convert and publish the XML file into semantically enriched open access HTML format.

Step 5. Browse, search, export and re-use the atomized content (e.g., taxon treatments, images, morphological characters, etc.).

Lead partners: PENSOFT, Naturalis, Plazi, BGBM.

Who is it for? Taxonomists, ecologists, conservationists, practitioners and any other user interested in biodiversity.

How is it used? The platform advancedbooks.org is open to publish or re-publish legacy or new monographs in open access, as well as to provide OCR, markup, text conversion and semantic enhancement services. The (re-)published content is free to use for anyone.



Key output 6: Automated registration of taxon names for publishers and registries

The pre-publication registration of taxonomic and nomenclatural acts with registries such as the International Plant Name Index (IPNI), Index Fungorum, MycoBank, and ZooBank involves two main classes of actors: (1) publishers, and (2) registry curators. The publisher takes the responsibility for initiating the registration of nomenclatural acts following a common stepwise model and for insertion of the registries' identifiers in the published article.

The workflow includes the following steps:

Step 1. XML message from the publisher to the registry on acceptance of the manuscript containing the type of act, taxon names, and preliminary bibliographic metadata; the registry will store the data but not make these publicly available before the final publication date.

Step 2a. Response XML report containing the unique identifier of the act as supplied by the registry and/or any relevant error messages.

Step 2b. Error correction and de-duplication performed manually; human intervention, at either registry's or publisher's side (or at both).

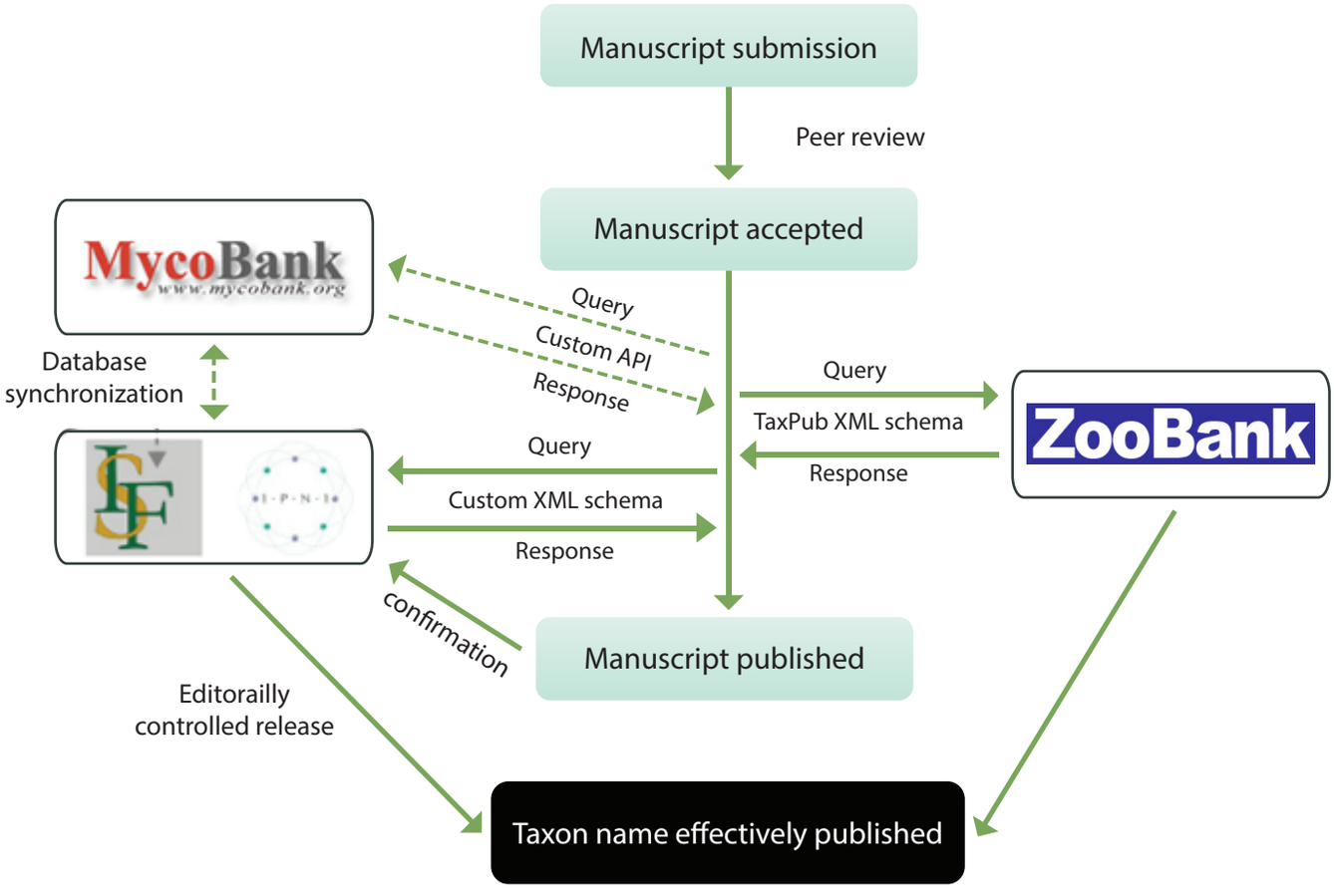
Step 3. Inclusion of registry-supplied identifiers in the published treatments (protologues, nomenclatural acts).

Step 4. Making the information in the registry publicly accessible upon publication, providing a link from the registry record to the article.

Lead partners: PENSOFT, RBGK (IPNI, Index Fungorum), ZooBank, MycoBank.

Who is it for? Biodiversity publishers and journals, especially those publishing taxonomic novelties.

How is it used? The registration workflow is free to use for anyone and the XML query and response formats are publicly available at http://wiki.pro-ibiosphere.eu/wiki/Pilot_2. The registry curators and Pensoft are available to advise journals that intend to implement the automated registration process.



Key output 7: Outcomes of the 2014 Biodiversity Data Enrichment Hackathon

Recent years have seen a surge in projects that produce large volumes of structured, machine-readable biodiversity data. During the Biodiversity Data Enrichment Hackathon, software developers and taxonomists came together to address the challenges and highlight the opportunities in the enrichment of such biodiversity data by engaging in intensive, collaborative software development. Detailed information of the activities and outcomes (including links to code repositories and demos) is available in the publication "Enriched biodiversity data as a resource and service" in the Biodiversity Data Journal.

The Biodiversity Data Enrichment Hackathon had two goals:

Goal 1. Facilitate re-use and enhancement of biodiversity knowledge by a broad range of stakeholders, such as taxonomists, systematists, informaticists, ecologists and niche modelers. The proposed use cases resulted in nine breakout groups addressing three main themes: (i) mobilizing heritage biodiversity knowledge; (ii) formalizing and linking concepts; and (iii) addressing interoperability between service platforms.

Goal 2. Foster a community of experts in biodiversity informatics and to build human links between research projects and institutions, in response to recent calls to further integration in our research domain.

Accomplishments

The Biodiversity Data Enrichment Hackathon showed a diversity of applications for biodiversity data and a complementary way to develop outlines of solutions in such an environment. Biodiversity data was mobilized from its silos and enriched with meaningful links to related resources, such as links from taxon names to taxon concept URIs; links from described habitats to environment ontologies; links from character traits to trait ontologies; links from species treatments to relevant images, publications and specimens. Workflow and data publishing platforms were enhanced to provide greater interoperability and data integration functionality.

MOBILIZING HERITAGE BIODIVERSITY KNOWLEDGE

Biodiversity data analytics: The biodiversity data aggregator GoldenGATE was enhanced with a search facility that allows extracting statistical data about specimens for visualisation in a dashboard.

OCR correction: A collaborative platform where authenticated users can correct OCR documents rendered on webpages.

Open Access images: A pipeline to mobilize, expose and extract images from open access publications.

FORMALIZING AND LINKING CONCEPTS

Trait ontology: A plant ontology, FLOPO, that allows extracting plant trait data from digitized Floras.

SWeDe: A standard, SWeDE, for documenting web services using XML.

Specimen links: Exploration of collection data to inform novel taxonomic concepts.

WORKFLOW AND DATA PUBLISHING PLATFORMS

EDIT Platform Common Data Model API: A web-service to extract occurrences out of Common Data Model (CDM) instances.

iPython notebook/Taverna: A more integrated, powerful research environment that exposes remote HPC resources (such as BioVeL services) to iPython notebook users.

BioVeL/NeXML services: BioVeL services to merge and query data and metadata.

Institutions involved: Aberystwyth University; Biodiversity and Climate Research Centre - Senckenberg Nature Research Society; Botanic Garden Meise; Landcare Research; Freie Universität Berlin - Botanischer Garten und Botanisches Museum; Institute of Biomembranes and Bioenergetics - Italian National Research; Museum für Naturkunde - Leibniz-Institut für Evolutions- und Biodiversitätsforschung; Open University; Pensoft; Plazi; Royal Botanic Gardens, Kew; Software Sustainability Institute, myGrid; Université de Montréal / Canadensys; University of Bath; University of Eastern Finland; BioVeL; University of Glasgow; University of Illinois; University of Manchester; Naturalis Biodiversity Center.

Key output 8: Services provided by the pro-iBiosphere consortium

The pro-iBiosphere consortium provides the following consultancy and implementation services that will help you to join OBKMS:

- Digitization of natural history collections.
- Implementation of persistent identifiers for biodiversity data.
- Digitization, markup and data extraction from historical biodiversity literature.
- Repository for taxon treatments, bibliographic references and published literature.
- Re-publication of Flora, Fauna and Mycota series in advanced open access.
- Open access semantic publication of biodiversity journals and books.
- Consultancy in legal aspects of use and re-use of biodiversity data.
- Integration of content in the community-owned big biodiversity data pool.

Please send your inquiries to services@pro-ibiosphere.eu

Key output 9: Recommendations for the sustainability of OBKMS

To achieve its goal of delivering biodiversity data in open and re-usable forms, the iBiosphere consortium must address the sustainability of both the data and the means to access them.

Elements of sustainablity

Sustainability requires that providers can curate, supply and exploit data effectively to ensure that the benefits of participation outweigh the costs. This can be facilitated through technical innovations, but also requires a full understanding of user requirements and how these can best be met.

Benefits to the users include increased availability through a central portal as well as increases in the volume, consistency, reliability, and currency of information available through searches. The benefits to data providers include greater use of their data, accreditation, improved awareness of impact, avoidance of duplicated effort, increased opportunities for innovative collaboration and tools for curation, analysis and research.

To ensure that the benefits are realised, implementation of OBKMS requires that stronger relationships are built and sustained between the provider and user communities.

Step 1. Understanding the major costs of provision of the services and how these can be minimised.

Step 2. Understand and quantify the benefits to users and providers and how these can be maximised.

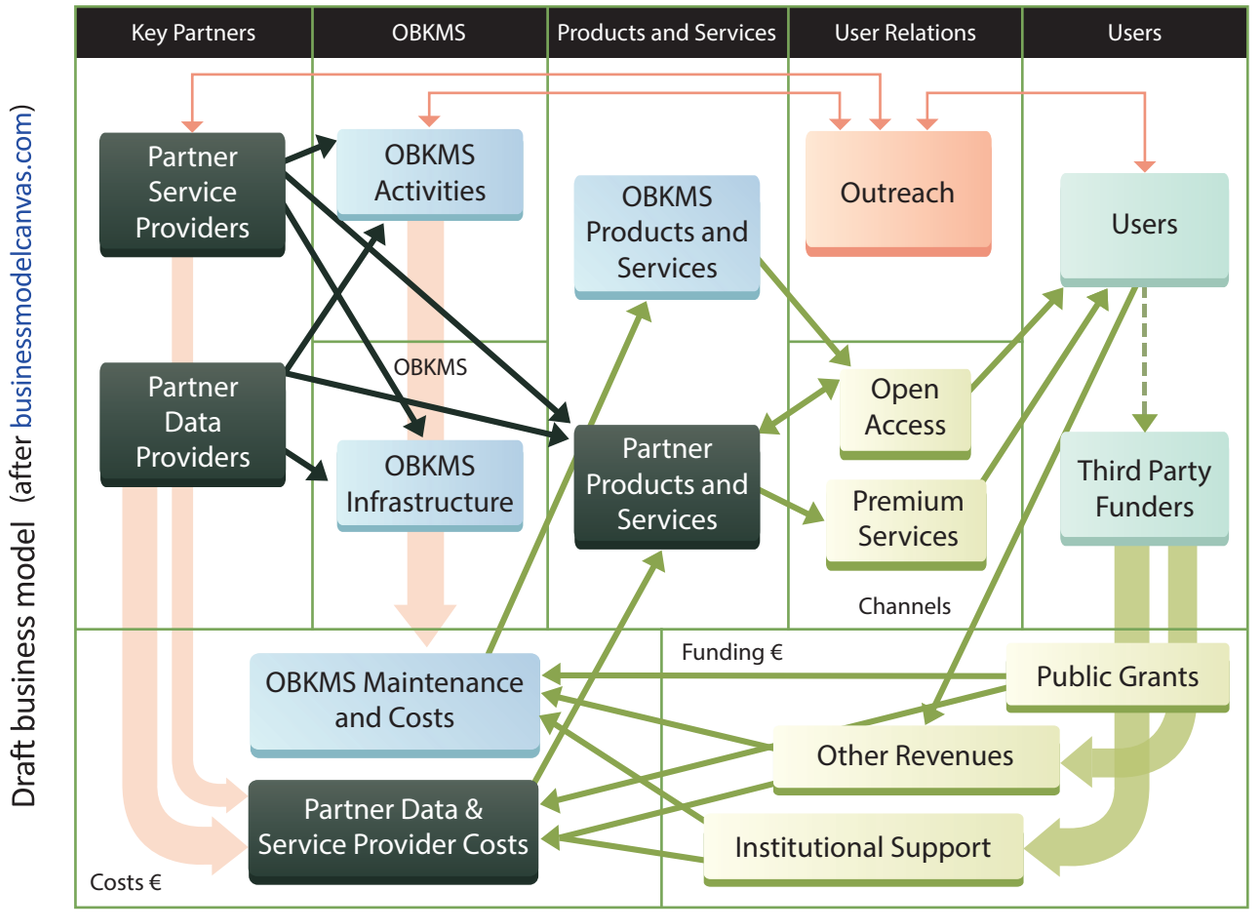
Step 3. Understanding the impact of OBKMS on partners' existing activities and the threats and opportunities that arise.

Step 4. Recommendations for sustainability to partners and to policy makers (due July and August 2014).

Lead partners: RBGK, Naturalis, Sigma.

Who is it for? Providers and prospective partners who provide biodiversity information; Policy makers who wish to encourage broader access and use of such information; Potential funders; Users who seek improved, wider, more effective and more efficient access.

How is it used? Make the business case for investment. To build bridges between user and provider communities and provide evidence on how best to support the business practices necessary for OBKMS outlined on the diagram.



Key output 10: The Bouchout Declaration for Open Biodiversity Knowledge Management

bouchoutdeclaration.org

Our natural world is a source of food, water, resources, protection and enjoyment that our society needs. The richness and complexity of nature, and the speed of new discoveries made possible by genomic and digital technologies, challenge us to find new ways to benefit from and be better custodians of the natural world. Digital information management systems can bring together the wealth of information now dispersed in a myriad of different documents, institutions, and locations. With such systems, we can harness the benefits of rapid discovery and open up our legacy of over 270 years of biological observations.

Intelligent information management provides mechanisms to link our understanding of biodiversity to the biomedical research that seeks new solutions to healthcare, to track change as it affects agricultural activities and food security, to support modelling of life on Earth, and to enable new discoveries. To take advantage of these opportunities, information must be made easily discoverable and openly and freely available.

The mission of the signatories is to promote free and open access to data and information about biodiversity by people and computers and to bring about an inclusive and shared knowledge management infrastructure that will allow our society to respond more effectively to the challenges of the present and future.

Collaborative Open Biodiversity Knowledge Management can bring together the achievements of many independent biodiversity projects, yet will allow them to retain their identity and missions. The resulting virtual pool of information will allow new services to emerge for everyone who relies on information about life on Earth. Awareness of, access to, preservation, and curation of information will be enhanced by a shared and seamless network of infrastructures. By enabling tracking data linking and citations, all who create, organise, or mobilise data will be fully credited for their contributions.

Open Biodiversity Knowledge Management will improve availability to information, increase the role and relevance of its participants, increase their impact, and reduce costs. As a society, we will understand our natural world better, manage it better, enable new types of discovery, return greater benefits to biomedical and agricultural endeavours, and increase food security.

As signatories, we encourage an overarching approach to Open Biodiversity Knowledge Management which is based on the following fundamental principles:

- The free and open use of digital resources about biodiversity and associated access services;
- Licenses or waivers that grant or allow all users a free, irrevocable, world-wide, right to copy, use, distribute, transmit and display the work publicly as well as to build on the work and to make derivative works, subject to proper attribution consistent with community practices, while recognizing that providers may develop commercial products with more restrictive licensing.
- Policy developments that will foster free and open access to biodiversity data;
- Tracking the use of identifiers in links and citations to ensure that sources and suppliers of data are assigned credit for their contributions;
- An agreed infrastructure, standards and protocols to improve access to and use of open data;
- Registers for content and services to allow discovery, access and use of open data;
- Persistent identifiers for data objects and physical objects such as specimens, images and taxonomic treatments with standard mechanisms to take users directly to content and data;
- Linking data using agreed vocabularies, both within and beyond biodiversity, that enable participation in the Linked Open Data Cloud;
- Dialogue to refine the concept, priorities and technical requirements of Open Biodiversity Knowledge Management;
- A sustainable Open Biodiversity Knowledge Management that is attentive to scientific, sociological, legal, and financial aspects.

Biodiversity-related institutions and individuals who share the vision expressed in the Bouchout Declaration are warmly encouraged to sign the Declaration.

If you intend to sign the declaration click at http://bouchoutdeclaration.org/sign/. If you have further queries please use the form on the web page top right.

PRO-iBIOSPHERE
WWW.PRO-iBIOSPHERE.EU

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

services@pro-ibiosphere.eu

Designed by PENSOFT

Building up a pro-iBiosphere community

The community offers real time access to the latest news, events and project activities information on open access, re-usability of biodiversity data & information, linking biodiversity data, a.o. To find out more on pro-iBiosphere community, join the different social media groups:

@proibiosphere | Pro-iBiosphere | pro-iBiosphere | Pro-iBiosphere