PRO iBiosphere web site

PRO iBIOSPHERE
WWW.PRO-IBIOSPHERE.EU

Subscribe to this service

NEWSLETTER

Issue 5

(January/April 2014)

In this Issue:

Extracting trait information from digitized floras

24.04.2014

Robert Hoehndorf (Aberystwyth University), Quentin Groom (Botanic Garden Meise), George Gosline (Royal Botanic Gardens Kew), Claus Weiland (Biodiversity and Climate Research Centre / Senckenberg), Thomas Hamann (Naturalis Biodiversity Center) The aim of the Traits task group at the recent pro-iBiosphere Biodiversity Data Enrichment Hackathon was to extract plant trait data from... more

SWeDe (Scientific Web-service Description) - an XML Schema Definition for describing Web Services in the scientific domain

24.04.2014

Niall Beard (University of Manchester), Patricia Kelbert (FUB-BGBM), Bachir Balech (Institute of Biomembranes and Bioenergetics - Italian National Research Center) At the Biodiversity Data Enrichment Hackathon in Leiden we created an XML Schema Definition for describing Web services in the scientific domain called SWeDe (Scientific Web-service Description). A web service provider wishing to... more

The running of Taverna Workflows within an IPython Notebook

23.04.2014

Alan Williams (University of Manchester, Aleksandra Pawlik (Software Sustainability Institute), Youri Lammers (Naturalis), Ross Mounce (University of Bath) During the recent pro-iBiosphere Data Enrichment Hackathon, a prototype Taverna Player Client Python package was developed for IPython Notebook. The package allows the listing of workflows available on a Taverna Portal, selection of... more

Hacking OCR for pro-iBiosphere

22.04.2014

* by David P. Shorthouse, Rod Page, Kevin Richards, Marko Tähtinen Taking his own lead from a pitch he delivered to an audience of receptive biodiversity informaticians at the outset of the March 17-21, 2014 pro-iBiosphere hackathon, Rod Page (University of Glasgow) fashioned an engaging interface to edit the OCR text... more

Important principles of identification and web integration: Identifier and Resolution

22.04.2014

by Kevin Richards, email: richardsk777@gmail.com The topic of "stable unique identifiers" in the biodiversity informatics community has had quite a varied history in recent years. With the fast changing world of technology, information and the latest approaches to deal with information storage and access, several changes in direction have taken place. In... more

Data visualisation task for pro-iBiosphere

22.04.2014

by David King* (Open University), Jeremy Miller (Naturalis), Guido Sautter (Plazi), Serrano Pereira (Naturalis)
* david.king@open.ac.uk Inspired by Pensoft's development in electronic publishing workflows, in combination with marked-up texts generated using GoldenGATE, Jeremy Miller (Naturalis) devised the design for a dashboard to visualise treatment data with the aim of better understanding that data and... more

Despatch from the field: New species discovery, description and data sharing in less than 30 days

27.03.2014

Researchers and the public can now have immediate access to data underlying discovery of new species of life on Earth, under a new streamlined system linking taxonomic research with open data publication. The partnership paves the way for unlocking and preserving a wealth of 'small data' backing up research conclusions, which... more

Outcomes of the pro-iBiosphere Workshop on Sustainable Business Models

26.03.2014

Charlotte Johns, Kew Royal Botanic Gardens, Email: c.johns@kew.org A workshop dedicated to sustainable business models was held during the 5th pro-iBiosphere project meeting on the 11th and 12th of February 2014, at the Museum für Naturkunde (MfN) in Berlin, Germany. It was attended by consortium members and eight external... more

REGISTER NOW: pro-iBiosphere Final Event in Meise (Brussels) - June 10-12, 2014

18.03.2014

The pro-iBiosphere Final Event will take place on June 10-12 2014, at the Bouchout Castle – Meise in Belgium (Agentschap Plententuin Meise, also known as Botanic Garden Meise). The aim of these series of activities is to present the achievements of the project and its sustainability perspectives. The week agenda comprises: Tuesday June... more

Register now to the pro-iBiosphere Final series of events in Brussels!

Stephanie Morales 30.04.2014

by Stephanie Morales

The pro-iBiosphere project supported by the European Commission (DG CONNECT) through its FP7 research funding programme has the pleasure to invite you to join its Final Event.

During its two-year duration, pro-iBiosphere contributed to making fundamental biodiversity data digital, open and re-usable. The achievements of the project will be presented in a series of activities (workshops, trainings, demonstrations and a Final Conference) that will take place from Tuesday the 10th to Thursday the 12th of June 2014 at the Bouchout Castle in the Botanic Garden Meise, in Meise (Brussels), Belgium.

The event wiki page here has been recently updated with additional information on the different series of activities organised and the Final Conference agenda now comprises worldwide high-level speakers, including (i) officials from the European Commission DG Connect, the US National Academy of Sciences, (ii) representatives from botanic gardens, natural museums, other biodiversity initiatives and (iii) experts or (iv) researchers specialized in biodiversity informatics, environmental/natural science.

One of the key objectives of these series of events is to provide key recommendations and inputs from biodiversity experts for the preparation of the next WP 2016-2017 as specifically asked by the European Commission.

The number of registered attendees has already reached a good level of participation, in this context, if you plan to attend and have not yet registered, we can only recommend you do to it as soon as possible here (due to room capacity constraints).

For further information on this event (agenda, concept & objectives, registration) please visit the Event wiki page or contact us at final-event@pro-ibiosphere.eu.

pro-iBiosphere project demonstrates its pilots at the final conference in Meise (Brussels)

Pro- iBiosphere 29.04.2014

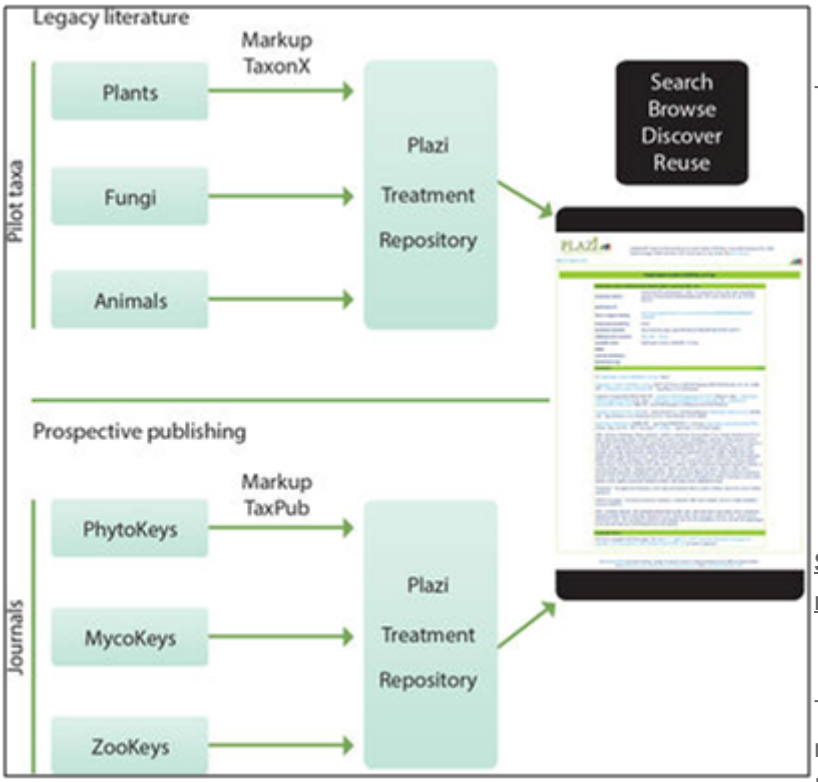
The pro-iBiosphere Final Event will take place on 9 – 13 June 2014 at the Bouchout Castle of the Botanic Garden Meise, Brussels. During the third day of the meeting a special event is designated for the demonstration of the pro-iBiosphere pilots.

During this session, the task and pilot leaders will demonstrate the tools and workflows developed or

improved in the course of the project. The demonstration will be interactive and will allow for discussions, real-time tests and consultations on possible implementations by the interested stakeholders. The pilots and demos planned until now are:

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 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 reuse through the Plazi Treatment Repository (PTR).

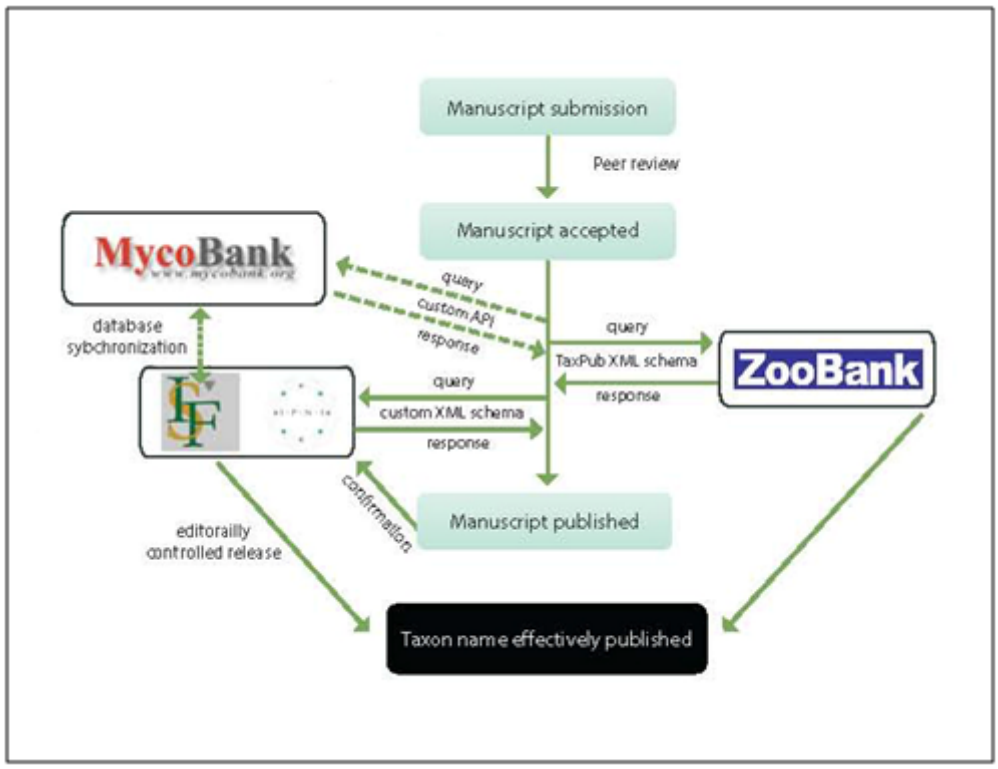


The workflow includes the following steps:

- Step 1: Convert printed taxonomic articles/monographs
- Step 2a: Mark up generic document features and doma and store the results at Plazi; and also
- Step 2b: Export of newly published treatments marked example in the journals ZooKeys, PhytoKeys and Myco
- Step 3: Browse, search, export and re-use treatments

Streamlining automated registration of taxon names between publishers and registries

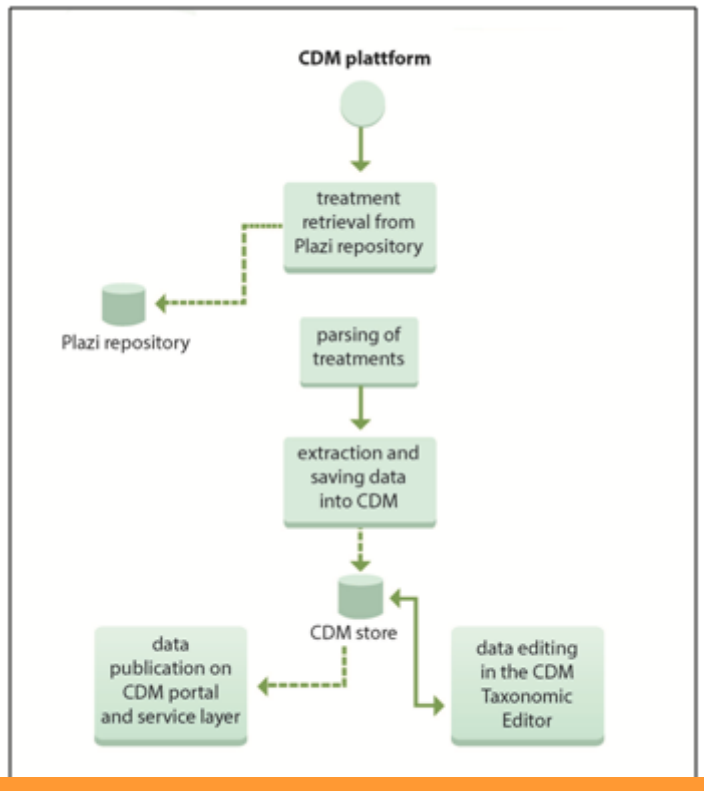
The pre-publication registration of taxonomic and nomenclatural acts with registries such as the International Plat Name Index (IPNI), Index Fungorum, MycoBank, and ZooBank involves two main classes of actors: (1) publishers, and (2) registry curators. The publisher takes the responsibility for initiating the registration of nomenclatural acts so that the work?ow can be performed following a common stepwise model:



- Step 1. XML message from the publish
containing the type of act, taxon names
store the data but not make these publi
- Step 2a. Response XML report contain
registry and/or any relevant error mess
- Step 2b. Error correction and d-duplica
registry's or publisher's side (or at both
- Step 3. Inclusion of registry supplied id
nomenclatural acts).
- Step 4. Making the information in the re
link from the registry record to the artic

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.



- Steps forward 1: Implementation of HTTP-URIs by 8 major institu
by October 2013 and recommendations for further topics to be ex
- Steps forward 2: Agreement on the BiodiversityCatalogue as a gl
related services. Improvement recommendations for it to be able
registration of services available now.
- Steps forward 3: Workflow improvement between the Plazi docum
Data Model (CDM)-based EDIT Platform for Cybertaxonomy (
http://wiki.pro-ibiosphere.eu/wiki/Pilot_3). In the course of this a
evolved. The pro-iBiosphere pilot portals visualize the data results
the possibilities for scientists willing to mark up their data. The ma
in detail work load and connected output gain.

Workshop on mark-up of biodiversity literature

Pro- iBiosphere 26.04.2014

Daniel Mietchen (Museum für Naturkunde Berlin)

For two days in February 2014, a pro-iBiosphere workshop on mark-up of biodiversity literature brought together a group of 20 participants at the Museum für Naturkunde in Berlin. In an introductory talk, Rod Page of the University of Glasgow presented the idea of a biodiversity knowledge graph that interlinks the biodiversity literature with the wider biodiversity information landscape. He then discussed a number of use cases for mark-up - namely for archiving, display and citation linking - and questioned whether it was actually necessary for identifying nodes and extracting edges of the knowledge graph, which could be achieved by simple indexing. He also discussed collaborative editing and version control with regards to mark-up.

With this introductory presentation having set the stage for discussing mark-up of the biodiversity literature in general terms and from a long-term perspective, the following presentations looked at specific subsets of that literature corpus, at specific use cases, at specific approaches to mark-up, and at workflows and business models around that. For example, Dimitris Koureas of the Natural History Museum in London discussed how the mark-up of specimen records in the literature could help with the digitization of specimen labels (which are often transcribed in systematic reviews of taxa or collections), and how the tracking of specimen citations in the literature could allow to assess the impact of collections on current and past research. Another perspective was provided by William Ullate of the Biodiversity Heritage Library, who described how BHL is ramping up its efforts on mark-up, including through gamification.

Throughout the workshop, there was a lively discussion, and the individual talks were given not according to a fixed schedule but when the respective topic came up in the discussion. All presentations are linked from the workshop page on the pro-iBiosphere wiki.

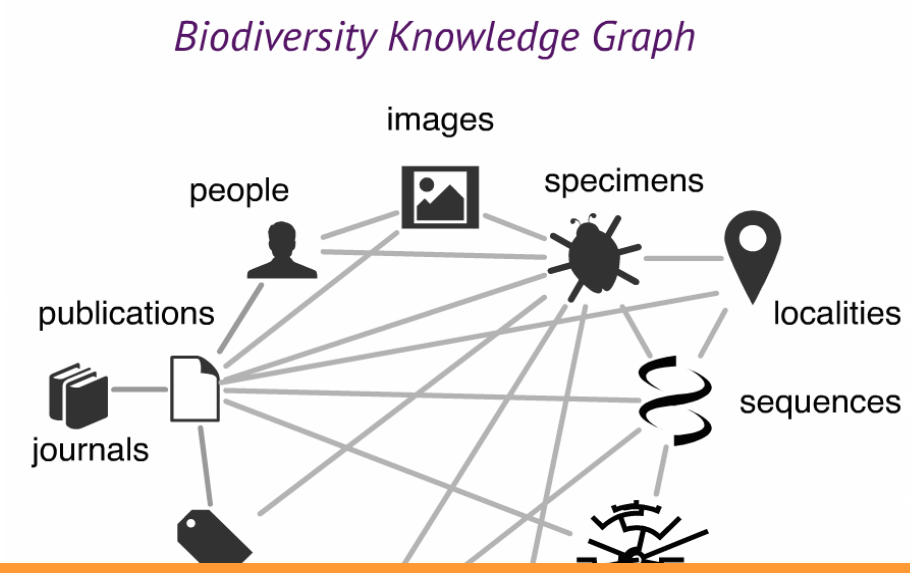


Figure 1: The Biodiversity Knowledge Graph. By Roderic Page.

The 2014 Biodiversity Data Enrichment Hackathon in a nutshell

Pro- iBiosphere 26.04.2014

Soraya Sierra, Rutger Vos* (Naturalis Biodiversity Center)*

*soraya.sierra@naturalis.nl, rutger.vos@naturalis.nl

From 17 – 21 March 2014, software developers and taxonomists came together in Leiden, the Netherlands, to address the challenges, and highlight the opportunities, in the enrichment of biodiversity data by engaging in intensive, collaborative software development: The Biodiversity Data Enrichment Hackathon. The event had two goals:

1. To facilitate re-use and enhancement of biodiversity knowledge by a broad range of stakeholders, such as ecologists and niche modelers.
2. To foster a community of experts in biodiversity informatics and to build human links between research projects and institutions.

The Hackathon brought together 37 participants (including developers and taxonomists, i.e. scientific professionals that gather, identify, name and classify species) from 10 countries: Belgium, Bulgaria, Canada, Finland, Germany, Italy, the Netherlands, New Zealand, the UK, and the US. The participants brought expertise in processing structured data, text mining, development of ontologies, digital identification keys, geographic information systems, niche modeling, natural language processing, provenance annotation, semantic integration, taxonomic name resolution, web service interfaces, workflow tools, and visualization.

The Biodiversity Data Enrichment Hackathon followed a use-case-driven model, i.e. a model where effort during the Hackathon was prioritized on the basis of compelling end user scenarios that could be enabled by the combined contributions of people that otherwise, outside of the Hackathon, do not collaborate. Most use cases and exemplar data were provided by taxonomists. The suggested use cases resulted in nine breakout groups addressing three main themes: (i) mobilizing heritage biodiversity knowledge; (ii) formalizing and linking concepts; and (iii) interoperability between service platforms.

Beyond deriving prototype solutions for each use case, areas of insufficiency were discussed and are being pursued further. It was striking how many possible applications for biodiversity data there were and how quickly solutions could be put together when the normal constraints to collaboration were broken down for a week. Conversely, mobilizing biodiversity knowledge from their silos in heritage literature and natural history collections will continue to require formalization of the concepts (and the links between them) that define our research domain as well as increased interoperability between the software platforms that operate on these concepts.

The tangible outcomes of the Hackathon are finding sustainable homes in the appropriate code bases (e.g. the code bases for CDM platform, the Plazi server, the BHL server) and registries and repositories (e.g. the BiodiversityCatalogue, the Pypi index, the NCBO BioPortal), or form the basis of proofs-of-concept for scientific publications and project proposals. The main intangible outcomes of the event are turning out to be the fostering of a community of experts in biodiversity informatics and the strengthened human links between research projects and institutions. The event also demonstrated both the ongoing need for data normalization and integration, e.g. through the application of ontologies, as well as the opportunities for innovative research such integration will afford.

Additional information of the Hackathon is available here. The outcomes of the Hackathon will be reported in the Biodiversity Data Journal (May 2014 issue) and presented during the pro-iBiosphere final event.

Pro- iBiosphere 25.04.2014

Patricia Kelbert1 and Quentin Groom2

1. *Botanischer Garten und Botanisches Museum Berlin-Dahlem, Freie Universität Berlin, Germany*
2. *Botanic Garden Meise, Belgium*

The EDIT Platform for Cybertaxonomy is a convenient tool for managing and editing details of specimens and observations. Also, the BioVel workflows for data refinement and niche modelling provide a powerful means to clean up and analyse the distributions of organisms. A way to join these seamlessly together was lacking so that, at the one end of the workflow, a researcher can manage their data in a user friendly interface, and at the other, sophisticated models of distributions can be generated. This problem was tackled by a task group at the recent pro-iBiosphere Hackathon.

One of the pro-iBiosphere pilots was to use legacy literature as a source of data on the historical changes to the distribution of Chenopodium vulvaria. Details of over 2000 observations and specimens were imported into the Common Data Model (CDM) database administered with the Taxonomic EDITor. Many of these data were extracted from legacy literature through a process of digitization and mark-up. These were imported as a whole into the CDM and are a valuable test dataset for bioclimatic niche modelling. In this way, heterogeneous data was homogenised to make it tractable to statistical analysis.

Until now, the link between database and workflow could only be performed by experienced users, who would need directly access to the database. During the hackathon the task group developed a new Java web-service within the CDM-library. This web-service takes the identifier of a taxon as input and returns a list of specimens or observation details. The precise fields returned were based on the prerequisites for reusing in the BioVel refinement workflow, but also contained other fields that might be useful in the future. In this manner we have completed the final link in a workflow that starts with 16th century botanists and ending with 21st century bioclimatic modelling.

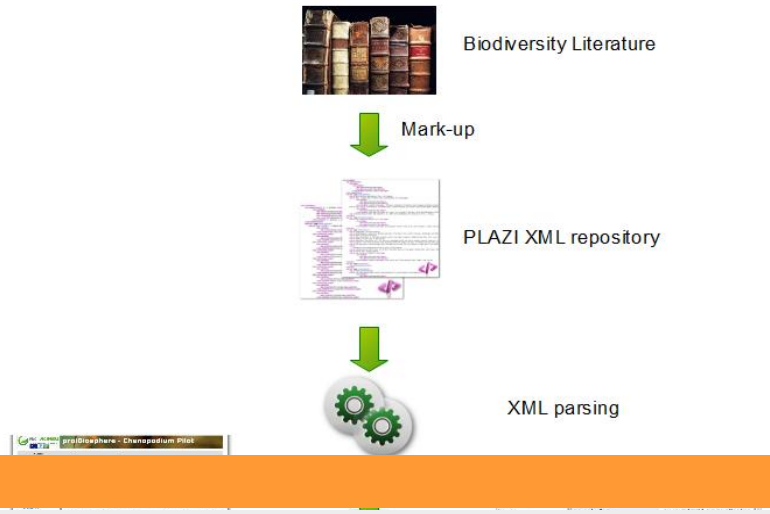
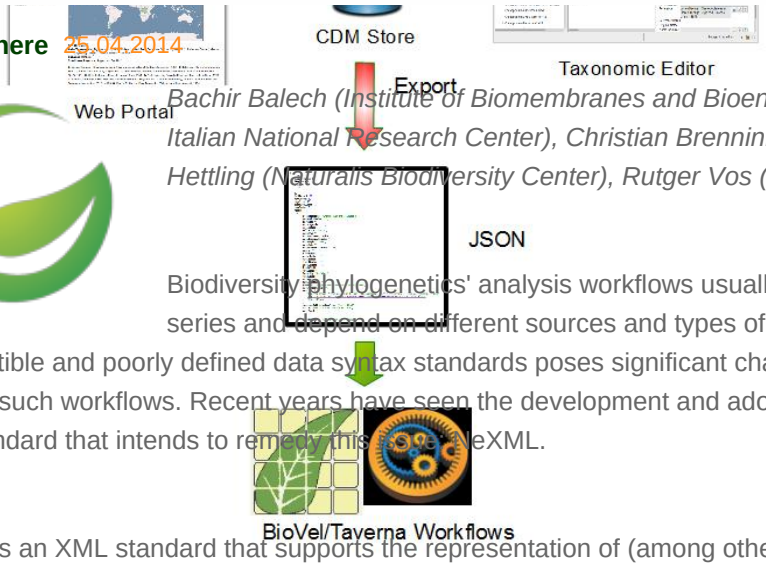


Figure 1. A schema showing the flow of data from legacy publications to modelling workflows. The red arrow shows the additional link in the chain.

Merging, extracting, and annotating biodiversity data with rich semantics using NeXML

Pro- iBiosphere 25.04.2014



Bachir Balech (Institute of Biomembranes and Bioenergetics - Italian National Research Center), Christian Brenninkmeijer (University of Manchester), Hannes Hettling (Naturalis Biodiversity Center), Rutger Vos (Naturalis Biodiversity Center)

Biodiversity phylogenetics' analysis workflows usually involve various software tools connected in series and depend on different sources and types of data. The proliferation of different, mutually incompatible and poorly defined data syntax standards poses significant challenges both for software developers and end users of such workflows. Recent years have seen the development and adoption of a new, expressive, and easy to process data standard that intends to remedy this: NeXML.

NeXML is an XML standard that supports the representation of (among others) taxa, character-state matrices and phylogenetic trees as well as semantic annotations (using RDFa) within one single document and is therefore specifically tailored to ease the interplay of different tools in evolutionary comparative and biodiversity analysis.

Since XML documents are generally intended to be handled by software rather than by users directly, a software tool to easily manipulate NeXML files appears desirable. To this end, participants of the biodiversity data enrichment hackathon (Leiden, the Netherlands, 17 – 21 March 2014) developed web services that can (i) construct NeXML documents from data encoded in commonly-used phylogenetic file formats or add metadata to an existing NeXML document, and (ii) extract information identified by the user from a given NeXML file and represent it in a variety of output formats.

To make the NeXML merger- and extractor tools easily accessible for the biodiversity research community and to enable their integration into existing workflows, they are implemented as RESTful web services, to be hosted by Naturalis Biodiversity Center and made available in the BiodiversityCatalogue. Clients that use these services can be implemented in a variety of ways; proofs-of-concept demonstrate that this is trivially done using the popular workflow management tool Taverna, such that these data merger and extractor facilities are available to the users of, inter alia, BioVeL workflows. Preliminary tests of NeXML merger and extractor have been conducted using data inputs and outputs used by the phylogenetic service set of BioVeL (https://www.biodiversitycatalogue.org/services/31/service_endpoint); while, NeXML extractor output has been tested, visualizing a phylogenetic tree with its taxa associated metadata, by implementing ITOL (http://itol.embl.de/) tool wraper within a taverna workflow.

For more information, visit the project wiki: http://wiki.pro-ibiosphere.eu/wiki/NeXML_Services

Extracting trait information from digitized floras

Pro- iBiosphere 24.04.2014

Robert Hoehndorf (Aberystwyth University), Quentin Groom (Botanic Garden Meise), George Gosline (Royal Botanic Gardens Kew), Claus Weiland (Biodiversity and Climate Research Centre / Senckenberg), Thomas Hamann (Naturalis Biodiversity Center)

The aim of the Traits task group at the recent pro-iBiosphere Biodiversity Data Enrichment Hackathon was to extract plant trait data from digitized Floras (i.e. a book that describes the plant life occurring in a particular region or time). We wanted to demonstrate the feasibility of using an ontology-based approach for extracting and integrating trait information from digitized Floras, even when the Floras are available in different languages. To tackle our main aim, we addressed two main questions: (1) Can we automatically extract trait and phenotype information from Flora descriptions written in multiple languages (English and French)?, and (2) Can we represent and integrate the extracted trait and phenotype information semantically using an ontology-based approach?

Extracting structured information about traits and phenotypes from natural language descriptions is a common problem in mobilizing legacy biodiversity data. One tool that has been developed for this purpose is the CharaParser [1], which is applied in the Phenoscape project [2] and integrated in the Phenex tool [3]. As the flora descriptions in our use cases were written in both on English and French language, and CharaParser primarily supports English language descriptions, we have chosen not to use CharaParser during the Hackathon. Instead we followed a simple text matching approach applicable to multiple languages. In particular, we identified mentions of plant anatomical entities (taken from the Plant Ontology [4]) and mentions of trait or phenotype terms (from the PATO ontology [5]) in the Flora descriptions. We used a dictionary to translate French and English terms referring to plant anatomy or plant traits. In the future, we plan to use more complex approaches such as CharaParser to provide a more complete and accurate mark-up of anatomy and phenotype terms in Flora descriptions.

To semantically describe traits, we follow the Entity-Quality (EQ) approach [6] that has been widely applied to semantically characterize model organism [7] and disease phenotypes [8]. Using the EQ model, a trait is characterized by an entity (E) of which a trait is observed, and the quality (Q) that characterizes the trait. The characterize identity can be an anatomical entity (from the Plant Ontology), or a biological process or function (from the Gene Ontology). The Phenotypic Attribute and Trait Ontology (PATO) contains a rich classification of widely applicable traits. A phenotype is described in a similar way using the EQ pattern, but the quality has a specific value and is a subclass of the trait. For example, the trait "flower color" will be described using the entity "flower" (from Plant Ontology) and the trait "color" (from PATO). The phenotype "flower red" is described using the entity "flower" (from Plant Ontology) and the quality "red" (from PATO), where "red" is a subclass of "color" in PATO.

We then used a data-driven approach to build a flora phenotype ontology (FLOPO) from the EQ statements we identified in the Flora descriptions. FLOPO is an ontology of over 25,000 trait and phenotype terms, all of which have at least one taxon annotation in one of the Floras we processed. The draft of FLOPO is available in BioPortal (http://bioportal.bioontology.org/ontologies/FLOPO), and the source code we produced and the data we used is available from http://wiki.pro-ibiosphere.e/wiki/Traits_Task_Group.

We have also started to generate further resources that we plan to use in the future. In particular, we have started to add

environmental terms to the Environment Ontology [9] that will allow us to extract parts of the environmental conditions in which taxa are found, we collected vocabulary and glossary terms that need to be incorporated into FLOPO. We have also experimented with using an RDF store that contains the FLOPO and its taxon annotations.

[1] Cui, H. (2012). CharaParser for fine-grained semantic annotation of organism morphological descriptions. Journal of American Society of Information Science and Technology. 63(4) DOI: 10.1002/asi.22618 http://onlinelibrary.wiley.com/doi/10.1002/asi.22618/pdf
[2] http://phenoscape.org/
[3] http://phenoscape.org/wiki/Phenex
[4] http://www.plantontology.org
[5] http://obofoundry.org/wiki/index.php/PATO:Main_Page
[6] Gkoutos, G. V., Green, E. C., Mallon, A.-M. M., Hancock, J. M., and Davidson, D. (2005) Using ontologies to describe mouse phenotypes. Genome biology, 6(1).
[7] Mungall, C., Gkoutos, G., Smith, C., Haendel, M., Lewis, S., and Ashburner, M. (2010) Integrating phenotype ontologies across multiple species. Genome Biology, 11(1), R2+.
[8] Robinson, P. N. et al. (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. American journal of human genetics, 83(5), 610–615.
[9] http://www.jbiomedsem.com/content/4/1/43

For more information, please contact Robert Hoehndorf: leechuck@leechuck.de

SWeDe (Scientific Web-service Description) - an XML Schema Definition for describing Web Services in the scientific domain

Pro- iBiosphere 24.04.2014

Niall Beard (University of Manchester), Patricia Kelbert (FUB-BGBM), Bachir Balech (Institute of Biomembranes and Bioenergetics - Italian National Research Center)

At the Biodiversity Data Enrichment Hackathon in Leiden we created an XML Schema Definition for describing Web services in the scientific domain called SWeDe (Scientific Web-service Description).

A web service provider wishing to propagate their web service will upload descriptive information on catalogue sites such as the Biodiversity Catalogue, the Tools Registry or any other relevant catalogue. This information should include a textual description of how to use the service as well as usage conditions such as licensing and restrictions, and other useful annotations.

The purpose of SWeDe is to allow web service providers to maintain just one document describing their web services rather than maintaining documentation over several different catalogues.

Hence, if a provider is required to change some information about their service, they can do so once - in their SWeDe document. Participating catalogues can then both periodically or in real-time, download and parse the SWeDe file and display its contents within their site. They can then update their databases with any alterations accordingly.

The SWeDe Schema was designed by scientists and developers to cover as many aspects of scientific web services as

possible. These include attributes such as the scientific category, technological category, projects (ie. funding), contact information (ie. institutions, persons) , intellectual property rights (IPR) and citations. The SWeDe schema re-uses several components from the Access to Biological Collections Data (ABCD) Schema. It can be used to describe services of both the two most predominant service types, REST and SOAP.

In addition to the schema, a rudimentary form to create your own SWeDe document (code-named the "SWeDe farmer") was also produced which can be found at http://swede-farmer.herokuapp.com

Further steps involve collaborating with Biodiversity Catalogue to parse SWeDe schemas, to improve the SWeDe Farmer, and to disseminate SWeDe to the scientific community.

The full XSD schema can be found in its GitHub Repository and further reading about SWeDe can be found on the pro-iBioshpere wiki.

https://github.com/njall/XS-SWeDe
http://wiki.pro-ibiosphere.eu/wiki/The_SWeDe_Project
http://swede-farmer.herokuapp.com/

For more information please contact support@mygrid.org.uk

The running of Taverna Workflows within an IPython Notebook

Pro- iBiosphere 23.04.2014

Alan Williams (University of Manchester, Aleksandra Pawlik (Software Sustainability Institute), Youri Lammers (Naturalis), Ross Mounce (University of Bath)

During the recent pro-iBiosphere Data Enrichment Hackathon, a prototype Taverna Player Client Python package was developed for IPython Notebook. The package allows the listing of workflows available on a Taverna Portal, selection of a workflow and the running of the workflow within the Notebook. Data from the Python environment can be used as inputs to the workflow, and the results of the workflow run are available for further manipulation in the notebook. User can interact with the running of the workflow using the Taverna Player and interaction services.

IPython Notebook [1]provides an interactive computational environment within a web browser. Users can write and execute Python code. This code may be combined with text, mathematical and statistical calculations, production of plots and HTML display to produce shareable and re-usable notebooks. These notebooks can be shared on the IPython Notebook Viewer[2].

Taverna[3] provides a suite of tools for workflow design, editing and execution. This includes the Taverna Workbench, the main creation tool for workflows, and the Taverna Server. Taverna Server enables you to set up a dedicated server for executing workflows remotely and it can be accessed by a WSDL or a REST API.

Instances of a Taverna Portal can be used to host, share and execute Taverna Workflows. The execution takes place on a

Taverna Server and is exposed within the portal using a Taverna Player. The Taverna Player can also be accessed by a REST API.

Following discussions with the developers of IPython Notebook, the exciting potential of running Taverna Workflows from within an IPython Notebook was realized.

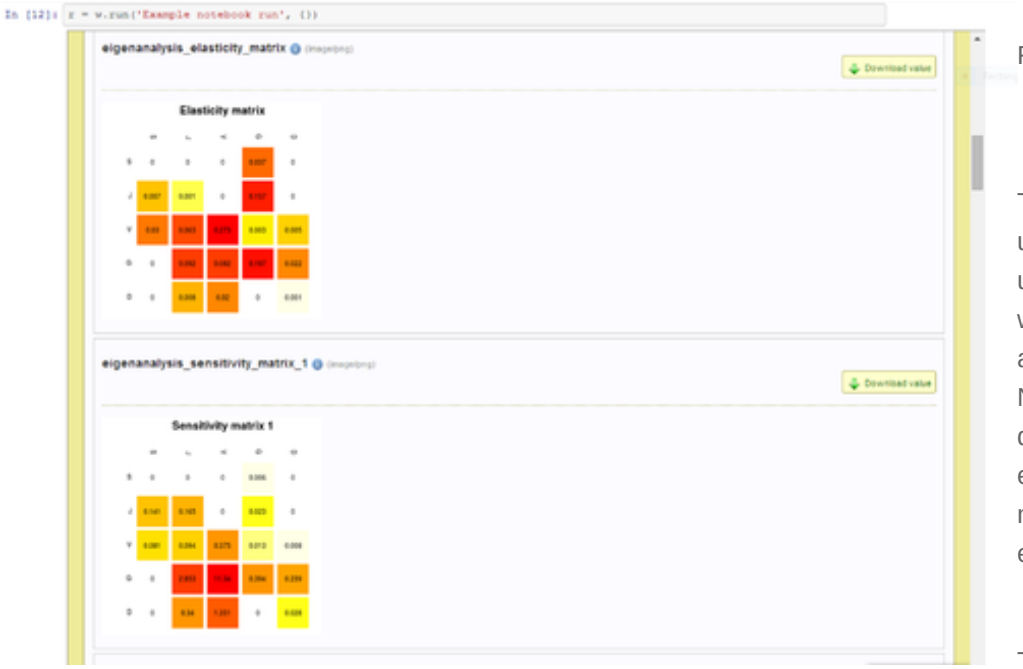


Figure 1: Running a workflow in IPython Notebook

The Taverna Player Client can be used to chain together workflows, using the outputs from one workflow run as the inputs to another. The capabilities of IPython Notebook can be used to generate documentation of the overall experiment; the templating mechanisms of jinja2 prove extremely useful for this.

The code for the Taverna Player Client is hosted on github[4] and a description of its current classes is available[5]. An example notebook has been uploaded to the Notebook Viewer [6].

Further work on the Taverna Player Client is planned, including meetings, both remote and face-to-face with the developers of IPython Notebook. The Client has been demonstrated to members of the BioVeL[7] and SCAPE[8] projects and colleagues at the University of Leiden.

We wish to thank the developers of IPython Notebook and Taverna Player, especially for their online support during the recent hackathon.

For more information, contact support@mygrid.org.uk

---

[1] http://ipython.org/notebook.html

[2] http://nbviewer.ipython.org/

[3] http://www.taverna.org.uk

[4] https://github.com/myGrid/DataHackLeiden

[5] http://dev.mygrid.org.uk/wiki/download/attachments/18972939/tavernaPlayerClient.html

[6] http://nbviewer.ipython.org/urls/raw.githubusercontent.com/myGrid/DataHackLeiden/alan/Player_example.ipynb?create=1

[7] http://www.biovel.eu

[8] http://www.scape-project.eu/

Hacking OCR for pro-iBiosphere

Pro- iBiosphere 22.04.2014

** by David P. Shorthouse, Rod Page, Kevin Richards, Marko Tähtinen*

Taking his own lead from a pitch he delivered to an audience of receptive biodiversity informaticians at the outset of the March 17-21, 2014 pro-iBiosphere hackathon, Rod Page (University of Glasgow) fashioned an engaging interface to edit the OCR text from scanned pages in the Biodiversity Heritage Library (BHL). He wooed David P. Shorthouse (Canadensys), Kevin Richards (ex Landcare Research New Zealand) and Marko Tähtinen (University of Eastern Finland, BioVeL) away from eight other competing task groups, each of which issued products in a remarkably short amount of time.

The purpose of the pro-iBiosphere hackathon was to "enrich structured biodiversity input data with semantic links to related resources and concepts". The OCR task group led by Rod had a distinctly different starting point, one that is no less important to the semantic linking of biodiversity resources. The unstructured data in the BHL is arguably the richest source of freely accessible information for taxonomists and biodiversity enthusiasts that can be mined into structured data. However, the quality of its OCR output suffers from variable typefaces, layouts, page contrasts and page bleeding, artifacts and other issues that occasionally bewilder its OCR engine. As a result, data mining and indexing routines that lift scientific names, place names, and other entities in support of semantic linking are not always successful. The browsing interface in the BHL could be made more engaging if visitors had an opportunity to rapidly correct the OCR text while viewing the original scanned image, thus enriching search and discovery for future visitors. Indeed, BHL and its partners were recently awarded a "Digging Into Data Challenge" grant (see http://blog.biodiversitylibrary.org/2014/03/first-meeting-of-mining-biodiversity.html), part of which will employ automated text-cleaning methodologies developed by its Canadian collaborators. An OCR editor might complement their funded work. Likewise, the Finnish National Library has developed its own OCR editor interface (see http://blogs.helsinki.fi/fennougrica/2014/02/21/ocr-text-editor/). Unlike the Finnish editor that uses ALTO XML as its source documents, the OCR editing interface developed during this hackathon uses BHL's DjVu XML documents as its source, rendered as HTML5.

The OCR Task Group had one aim: provide a simple interface for interactive editing of text, as well as tools to make inferences from the edits. After four solid days of hacking, the team completed this aim and integrated value-added features to engage users and to boost developer confidence in reuse of the code. The underlying document store is the cloud-based CouchDB (on Cloudant) and the team is confident that the proof-of-concept can be made to scale. The capabilities of the software are:

1. An in-place panel shows the exact line in the original scanned image while the user edits a single line of OCR text at a time (**Figure 1**)
2. Global Names scientific name-finding is integrated in real-time when a user completes a line edit, giving feedback if a scientific name is newly recognized (**Figure 2**)
3. Authentication uses the facile https://oauth.io/ such that all edits are tied to users' OAuth2-provider accounts (*eg* Google, Twitter, GitHub)

4. Frequencies of common edits are summarized in real-time and other words that may benefit from similar edits are highlighted for users
5. Batch processes collapse all user edits and text files are recreated for possible re-introduction into data mining routines
6. Unit and integration tests are included

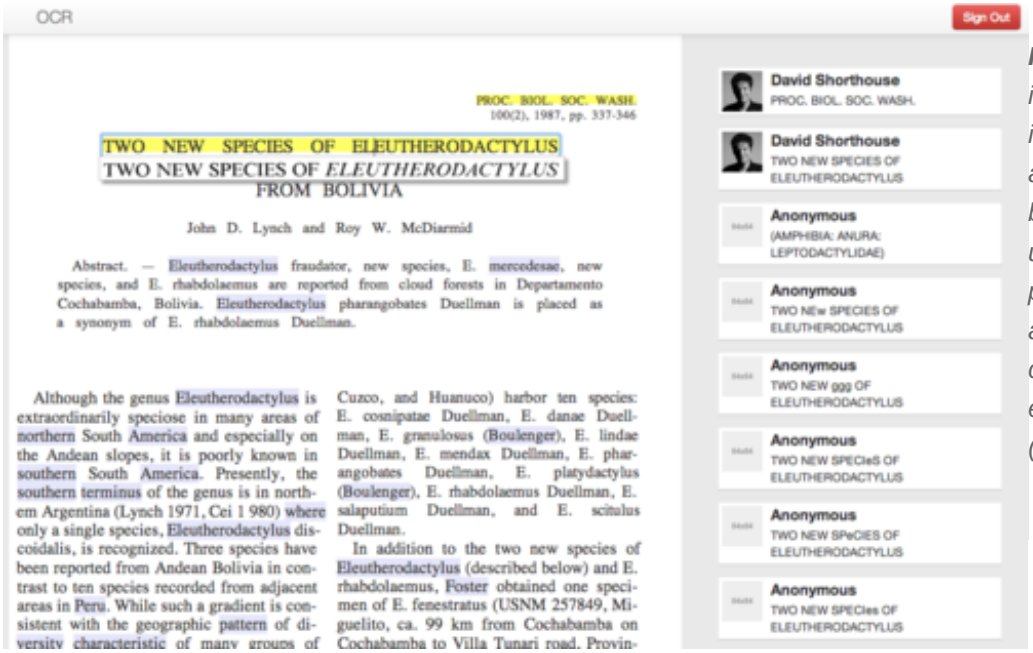


Figure 1. *The OCR Editing interface rendered as HTML5, illustrating the original line of text as a clipped image under the line being edited, a scrolling tally of user edits, lines that have been previously edited (yellow highlight) and words that share strings of characters that match previous edits elsewhere on the page (mauve highlight).*

Rutger Vos and Soraya Sierra (Naturalis, co-organizers) received abundant praise by all participants at the completion of the hackathon, and rightly so. The hackathon was exceptionally well organized, developer team sizes were perfect for each of the nine task groups, each participant's work ethic was remarkable, facilities were well provisioned, and nibbles and luncheons were delectable. We look forward to the reactions of pro-iBiosphere members at the final event in Meise, Brussels.

Contact:
David P. Shorthouse
Université de Montréal Biodiversity Centre / Canadensys, Montréal, QC CANADA
Canadensys Email: david.shorthouse@umontreal.ca

Important principles of identification and web integration: Identifier and Resolution

Pro- iBiosphere 22.04.2014

by Kevin Richards, email: richardsk777@gmail.com

The topic of "stable unique identifiers" in the biodiversity informatics community has had quite a varied history in recent years. With the fast changing world of technology, information and the latest approaches to deal with information storage and access, several changes in direction have taken place.

In these changing times it seems that trying to stick to basic technologies, especially those that work with standard internet protocols, is the way to go. However, it is important to emphasise the two major components of identifiers: the IDENTIFIER and the RESOLUTION. These important principles of identification and web integration were put to use at the recent Biodiversity Enrichment Hackathon that took place on 17-21 March 2014 in Leiden. The importance of identification and resolution is obvious when attempting to link various data sets and information sources in the Biodiversity domain.

IDENTIFIER for the data

The first issue for any user of data is the need to identify that particular piece of data. This has traditionally been done using fairly local identifiers such as a number counter (i.e. 1,2,3...). With the need to integrate and access data globally, other mechanisms have been required. The simplest approach to this is called Universally Unique Identifier (UUID). UUIDs are hard to read and quite unappealing to look at, for example "1696AC49-548F-404D-9DEA-8A1C4DDA37F4" but are still a good mechanism for identifying data in a computer system, and hence, work well for computer needs.

RESOLUTION of data by their identifiers

With the increasing demand to have data accessible and linked on the web other identifier mechanisms are required to allow data to be fetched via their identifiers. Within the biodiversity community several approaches have been taken. Originally LSID (Life Science IDentifiers) were promoted as they had several appealing features, namely, a degree of indirection from the domain name associated with the data host and a defined protocol for accessing the data and metadata for a particular object. Other identifier systems were also considered such as DOI, PURL and Handles. The main benefit of all these identifier systems is that the data is then accessible over the web using web technologies.

Then came along the semantic web with some really cool ideas about linking data together in a meaningful way and building a reusable, re-purposeable giant set of data. This has become really appealing to biodiversity informaticians and has consequently resulted in some interesting hurdles to jump to achieve these attractive ambitions. Firstly semantic web technologies highly depend on automation and basic web protocols for harvesting and linking data. So any identifier system that doesn't work well with basic HTTP web protocols is difficult to integrate. This meant that LSIDs have become unfavourable due to their reasonably complex resolution protocol. Instead basic stable permanent URLs have been promoted.

A good approach to using these type of identifiers is to first pick a very agnostic domain name, ie not an institution or university name, but perhaps a "project" name. A good example of this is the International Plant Names Index project – also known as IPNI (its data system is hosted by the Royal Botanic Gardens Kew, London). Then a locally unique identifier portion is attached to the chosen domain name. An example of this combination is Zoobank with their zoobank.org domain name and an identifier for a particular piece of data they host, eg http://zoobank.org/NomenclaturalActs/8BDC0735-FEA4-4298-83FA-D04F67C3FBEC is a resolvable identifier for the zoobank record for the taxon "Chromis abyssus".

The pro-iBiosphere project has created a Best Practices page for stable URIs that outlines some good approaches to creating identifiers for your data with consideration of semantic web requirements and the latest ideas on identification.

Data visualisation task for pro-iBiosphere

Pro- iBiosphere 22.04.2014

by David King (Open University), Jeremy Miller (Naturalis), Guido Sautter (Plazi), Serrano Pereira (Naturalis)*

* *david.king@open.ac.uk*

Inspired by Pensoft's development in electronic publishing workflows, in combination with marked-up texts generated using GoldenGATE, Jeremy Miller (Naturalis) devised the design for a dashboard to visualise treatment data with the aim of better understanding that data and assisting with its quality control. Ultimately, Jeremy's vision would make it be possible to offer a kind of reverse Biodiversity Data Journal, resurrecting primary data from marked-up legacy literature for aggregation and re-analysis. Our challenge in the recent pro-iBiosphere hackathon, excellently hosted by Naturalis, was to craft a prototype to extract and display the data for Jeremy's dashboard.

Working with GoldenGATE's author, Guido Sautter, enabled us to immediately refine one weakness of the original design: rather than process exported GoldenGATE marked-up text to extract statistical data, we could have GoldenGATE extract it for us and make that data available for export. Hence, GoldenGATE's functionality was extended and a new API service made freely available at http://plazi.cs.umb.edu/GgServer/srsStats for us to use, and for anyone else to use who wishes to explore this statistical data. Some solid visualisation work by Serrano Pereira, a recent recruit to Naturalis, using the established frameworks jQuery, jqPlot and jVectorMap saw the exported data rendered into the form Jeremy envisaged.

A version of the demonstrator produced during the hackathon is currently available at http://plazi.byobu.info/, courtesy of Plazi, a pro-iBiosphere partner. We look forward to refining and enhancing the existing demonstrator in-line with feedback from Jeremy and other users, and from its presentation at pro-iBiosphere's final event in June.

Jeremy's original concept for the dashboard is available from https://github.com/Dauvit/Data_enrichment/tree/master/data_visualisation/use_case.

The code for GoldenGATE can be downloaded from https://code.google.com/p/goldengate-tools/.

The documentation for GoldenGATE's statistical export service is available from https://github.com/Dauvit/Data_enrichment/blob/master/data_visualisation/Stats_queries_HOWTO.md.

The code for the demonstrator dashboard can be downloaded from https://github.com/Dauvit/Data_enrichment/tree/master/data_visualisation.

Despatch from the field: New species discovery, description and data sharing in less than 30 days

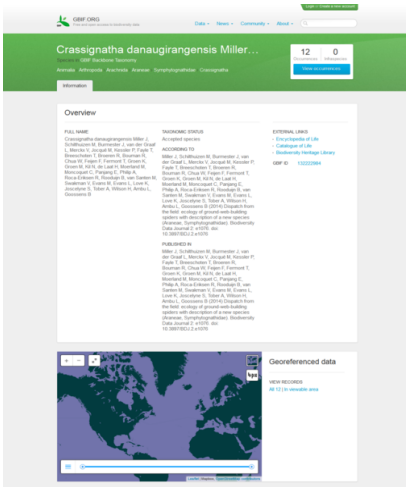
Pro- iBiosphere 27.03.2014

Researchers and the public can now have immediate access to data underlying discovery of new species of life on Earth, under a new streamlined system linking taxonomic research with open data publication.

The partnership paves the way for unlocking and preserving a wealth of 'small data' backing up research conclusions, which often become lost within a few years of an article's publication in an academic journal.

In the first example of the new collaboration in action, the *Biodiversity Data Journal* carries a peer-reviewed description of a new species of spider discovered during a field course in Borneo just one month ago. At the same time, the data showing location of the spider's occurrence in nature are automatically harvested by the Global Biodiversity Information Facility (GBIF), and richer data such as images and the species description are exported to the Encyclopedia of Life (EOL).

This contrasts with an average 'shelf life' of twenty-one years between field discovery of a new species and its formal description and naming, according to a recent study in *Current Biology*.



A group of scientists and students discovered the new species of spider during a field course in Borneo, supervised by Jeremy Miller and Menno Schilthuizen from the Naturalis Biodiversity Center, based in Leiden, the Netherlands. The species was described and submitted online from the field to the *Biodiversity Data Journal* through a satellite internet connection, along with the underlying data . The manuscript was peer-reviewed and published within two weeks of submission. On the day of publication, GBIF and EOL have harvested and included the data in their respective platforms.

The new workflow established between GBIF, EOL and Pensoft Publishers' *Biodiversity Data Journal*, with the support of the Swiss NGO Plazi, automatically exports treatment and occurrence data into a Darwin Core Archive, a standard format used by GBIF and other networks to share data from many different sources. This means GBIF can extract these data on the day of the article's publication, making them immediately available to science and the public through its portal and web services, further enriching the biodiversity data already freely accessible through the GBIF network. Similarly, the information and multimedia resources become accessible via EOL's species pages.

One of the main purposes of the partnership is to ensure that such data remain accessible for future use in research. A recent study published in *Current Biology* found that 80 % of scientific data are lost in less than 10 years following their creation.

Donald Hobern, GBIF's Executive Secretary, commented: "A great volume of extremely important information about the world's species is effectively inaccessible, scattered across thousands of small datasets carefully curated by taxonomic researchers. I find it very exciting that this new workflow will help preserve these 'small data' and make them immediately available for re-use through our networks."

"Re-use of data published on paper or in PDF format is a huge challenge in all branches of science", said Prof. Lyubomir Penev, managing director of Pensoft and founder of the *Biodiversity Data Journal*. "This problem has been tackled firstly by our partners from Plazi who created a workflow to extract data from legacy literature and submit it to GBIF. The workflow currently launched by GBIF, EOL and the *Biodiversity Data Journal* radically shortens the way from publication of data to

their sharing and re-use and makes the whole process cost efficient", added Prof. Penev.

The elaboration of the workflow from BDJ and Plazi to GBIF through Darwin Core Archive was supported by the EU-funded project EU BON (Building the European Biodiversity Observation Network, grant No 308454). The basic concept has been initially discussed and outlined in the course of the pro-iBiosphere project (Coordination and policy development in preparation for a European Open Biodiversity Knowledge Management System, addressing Acquisition, Curation, Synthesis, Interoperability and Dissemination, grant No 312848).

Original source:

Miller J, Schilthuizen M, Burmester J, van der Graaf L, Merckx V, Jocqué M, Kessler P, Fayle T, Breeschoten T, Broeren R, Bouman R, Chua W, Feijen F, Fermont T, Groen K, Groen M, Kil N, de Laat H, Moerland M, Moncoquet C, Panjang E, Philip A, Roca-Eriksen R, Rooduijn B, van Santen M, Swakman V, Evans M, Evans L, Love K, Joscelyne S, Tober A, Wilson H, Ambu L, Goossens B (2014) Dispatch from the field: ecology of micro web-building spiders with description of a new species. *Biodiversity Data Journal* 2: e1076. DOI: 10.3897/BDJ.2.e1076

Outcomes of the pro-iBiosphere Workshop on Sustainable Business Models

Pro- iBiosphere 26.03.2014

Charlotte Johns, Kew Royal Botanic Gardens, Email: c.johns@kew.org

A workshop dedicated to sustainable business models was held during the 5th pro-iBiosphere project meeting on the 11th and 12th of February 2014, at the Museum für Naturkunde (MfN) in Berlin, Germany. It was attended by consortium members and eight external participants with experience in strategic business and finance.

The workshop was planned to split into 4 sessions. The first session looked to agree the scope of a future "iBiosphere", to decide which products and services will be included as part of the Open Biodiversity Knowledge Management System (OBKMS). This session was followed by a number of talks given by the external participants, who shared their experience on the sustainability of their projects. Session three looked at enabling factors contributing towards the OBKMS, including open access, data and technology and communication. The final session concentrated upon sustainability and governance and how the management of iBiosphere should be structured.

The main outcome of the workshop was agreement on a list of core products and services which the OBKMS will provide, and an agreement on the core functionality. Information gathered through this milestone also helped to create the draft sustainability model for the OBKMS, which highlights gaps in our present knowledge and helps to decide upon future work which needs to be completed. These workshop outcomes, along with suggestions as to how the OBKMS will be governed and a list of challenges and solutions for a number of enabling functions, can be found within D6.4.2 the 'Draft Sustainability Report'. The information collected through the workshop will also aid future reports including D6.1.2 'Report on Costs', D6.4.3 'Summary of model evaluations' and D6.4.4 'Sustainability recommendations', to be made available here.

We would like to again thank all the participants for the success of the workshop and who contributed valuable information

that will help shape our future pro-iBiosphere sustainability deliverables.

REGISTER NOW: pro-iBiosphere Final Event in Meise (Brussels) - June 10-12, 2014

Camille Torrenti 18.03.2014

The **pro-iBiosphere Final Event** will take place on **June 10-12 2014**, at the Bouchout Castle – Meise in Belgium (Agentschap Plententuin Meise, also known as Botanic Garden Meise).

The aim of these series of activities is to present the achievements of the project and its sustainability perspectives.

The week agenda comprises:

Tuesday June 10 (PM)
Workshop on Model Evaluation

Wednesday June 11 (all day)
Demonstrations on pro-iBiosphere pilots
Demonstrations on outcomes of pro-iBiosphere Data Enrichment Hackathon
Workshop on Biodiversity Catalogue
Training on WikiMedia
Poster session

Thursday June 12 (all day)
Final Conference
Networking Cocktail

Do not miss this unique opportunity and join us in Meise (Brussels)!.

Registration is free of charge but compulsory due to room capacity constraints. You can register by filling out the online registration form at http://tiny.cc/pib-final-event.

For complementary information on the Final Event (background, registration, logistics), please visit the dedicated wiki page at http://tiny.cc/wiki_pib_final_event or contact us at final-event@pro-ibiosphere.eu.


PRO-iBIOSPHERE
WWW.PRO-iBIOSPHERE.EU
pro-iBiosphere Final Event
June 10-12, 2014
Bouchout Castle - Meise (Brussels), Belgium