WORKING GROUPS REPORT - WG 5: Development of Standards and Guidelines for Data publication (portals, publishers)

Sofia, BU, 12-13 March 2019

WG WORKSHOP:

1. Attendance:

   a. WG Leader - Stefano Martellos

   b. Convener/Rapporteur - Stefano Martellos

   c. Number of Participants - 17

   d. Inclusiveness rate of the WG participation:

      - Gender balance - 15 m / 3 f

      - Geographical distribution - Italy: 1; Portugal: 1; Spain: 1; Bulgaria: 2; Switzerland: 1; Serbia: 3; Montenegro: 1; Poland: 1; Slovakia: 1; Nederlands: 1; France: 2; Germany: 1; Iceland: 1; Denmark: 1.

      - Seniority – no data available

1. Objectives of the WG workshop - The workshop focused on identifying, and discussing the most relevant data publication pipelines. On this basis, potential strategies for enhancing data publication, especially by small institutions, are discussed.

2. Topics discussed

   - In reference to completion of Y1 objectives - Existing data publication pipelines: “homebrew” approaches, data aggregators (BioCASE, GBIF), open access journals (Pensoft platforms, CETAF European Journal of Taxonomy), the PLAZI approach for mobilizing taxonomic data. Target users of data publication platforms. The EU landscape of data publication.

   - Towards the definition of Y2 activities and actions - Limits and opportunities of current data publication pipelines. How they are affecting / enhancing / limiting data publication. Citizen Science data.

- Others - proper citation of digital objects in scholar publications. Training in data publication and citation. Metrics and incentives for institutions. Diversity of softwares and approaches for digitization.

1. Major outcomes - Dataset citation practices should be improved. Lack of proper citations biases the metrics about datasets usage, thus lowering the feedback for institutions, which are not rewarded in terms of relevance for their digitization efforts. Publishers and editors should force proper citation practices. At the same time, the publication of each dataset in the form of data paper should be preferred. Furthermore, taxonomic audit should be forced, in order to increase data usability and reusability. Given the EU Declaration on Open Science, dataset publication and citations should be used for cereer progression.

3. Challenges faced for Y2 - organising an effective collaborsative network among participants

4. Next steps - Publishing a report in the form of a review paper on an open access online journal. Arranging other two meetings (one

ANNEXES:

A. List of Participants: Participants name, affiliation, expertise, contact

Mariya Dimitrova	Pensoft	m.dimitrova@pensoft.net
Lyubomir Penev	Pensoft	penev@pensoft.net
Donat Agosti	Plazi	agosti@plazi.org
Milan Dimitrijevic	Astronomic observatory	mdimitrijevic@aob.rs
Karol Marhold	Plant Science and Biodiversity Centre	karol.marhold@savba.sk
Sreckovic Vladimir	Institute of Physics Belgrade	vladimir.sreckovic@ipb.ac.rs
Jelena Petrovic	Vinca Institute	petrovicj@vin.bg.ac.rs
Thierry Bourgoin	MNHN Paris	thierry.bourgoin@mnhn.fr
Etienne Cayeux	MNHN Paris	etienne.cayeux@mnhn.fr
Jeremy Miller	Naturalis Biodiversity Center	jeremy.miller@naturalis.nl
Rui Figueira	GBIF Portugal	ruifigueira@isa.ulisboa.pt
Arturo H. Arino	Univeristy of Navarra	artarip@unav.es
Anna Gazda	University of Agruculture	rlgazda@cyf-kr.edu.pl
Pawel Wasowicz	Icelandic Institute of Natural History	pawel@ni.is
Jorg Holetschek	Botanic Garden & Botanic Museum Berlin	j.holetschek@bgbm.org
Tim Robertson	GBIF	trobertson@gbif.org
Vladimir Pesic	University of Montenegro	vladopesic@gmail.com

Absent at the meeting, but willing to participate to the next steps of the work

Laurence Benichou	MNHN Paris	laurence.benichou@mnhn.fr
Christos Arvantidis	Hellenic Center for Marine Research	arvanitidis@hcmr.gr
Snezana Dragovic	Vinca Institute	sdragovic@vin.bg.ac.rs
Jasenka Gajdos	University of Zagreb	jasenka.gajdos@gmail.com

B. Extended Report

Day 1 - 12/03/2019

The meeting started with a short analysis of the current European situation, as far as Natural History Museums (NHMs) and collections are concerned. There exists a relevant variability, with two extremes. On one side, countries as Italy, in which there are hundreds of NHMs, ranging from the big ones, to several very small ones, without a National NHM. Similar situations are present in other countries, like Spain, Serbia, Switzerland, Slovakia, etc.. On the other side, the example of the Netherlands, which aggregated practically every collection in the Naturalis institutions. The same happens in Denmark, or in Iceland, with a single big institution. In the middle, countries with a National NHM, plus several other regional and municipal NHMs (e.g. France, Germany).

Thus, the landscape of potential targets of WG5 activities is quite wide and diverse, and the outputs we aim at generating have to take into consideration this diversity, in order to be effective.

The meeting continued with several presentations from different well established data publication pipelines. Presentations were not standard conference presentations, but interactive, and discussion arose frequently during them.

At the beginning, it was highlighted the importance of distinguishing between stand alone data publishing (e.g., data published by one or more persons, out of the scholar channels), and scholarly data publishing (in which there are proper citation instruments, and some peer-review process).

The first presentation was from Tim Robertson (GBIF). It aimed at describing the GBIF data publication pipeline.

The presentation highlighted some relevant points:

a) what does it mean to publish data?

- “simply” putting them on the internet,
- or a process which starts from taking the data, to properly exposing them to the users?

b) tracking of data use

This could be quite a relevant issue as far as data publishing is concerned. Since digitizing collections is a long and expensive process, it has to provide some “reward” for the effort. NHMs and other institutions which host and preserve collections need to demonstrate that their efforts for digitizing and publishing the data have a certain value for money. Thus, demonstrating that data are actually used by the scientific community (and not only) could be a proper way to stimulate the continuation of digitization and publication efforts. The GBIF permits data tracking since it provide DOIs to datasets, and maintain a registry of datasets. Thus, they can be properly cited in scientific publications. Up to date, GBIF datasets are cited twice daily in peer-reviewed journals.

c) map data to common data models

Aggregating data in the GBIF, by using of GBIF-IPT, or other pipelines, such as Pensoft, or BioCASE, means mapping them against a common data standard (Darwin Core, DwC). This clearly strongly enhance re-usability of data.

d) data refinement pipelines

The GBIF has several data refinement pipelines, which assure a certain level of automated quality control to each dataset which is aggregated in the platform. These are the taxon name pipeline (evolved in COL+ project), the occurrence pipeline (which uses some country specific rules), the multimedia object pipeline, the sequence pipeline, etc. Furthermore, the GBIF support annotations to allow flow back to original data provider.

e) the GBIF is dataset focused, not record focused (as it is Dissco)

While there exist still some taxonomic gaps in GBIF, especially in fungi, it is still growing at an incredible rate. 20 year review on GBIF is coming soon.

A second presentation come from Jorg Holetscheck, and focused on one of the most relevant data providers to the GBIF, the BioCASe (Biological Collections Access Service)
Currently, the BioCASe hosts 40 million records, in several installations around the world. Because of its structure of true federated database system, it is better suitable for networks of a size of ca. 10 million records. BioCASE publish data as a web service, and permits a full dataset retrieval, as well as allowing to query for single records. Thus, potentially, each record could be referenced. It is a little bit trickier to install and configure, and for this reason it could be better download and install as a Docker container. While it is used comonly with the ABCD data standard, BioCASe protocol can support any data standard, and even permit the use of arbitrary XML schemas. The Access to Biological Collections Data standard (ABCD) is an XML schema intended for sharing primary biodiversity data. Even though it was initially designed for storing information of natural history collections, it can be used for all types of occurrence data, both specimens and observations. ABCD is hierarchical and allows the repetition of certain sub-sections, and its multitude of data elements (988 leaf elements and attributes in the current version 2.06) enables the accommodation of rich information. For items that don't fit in, ABCD features a slot for incorporating schema extensions with additional, network-specific data elements. Currently, two extensions are in active use: The GGBN extension for the Global Genome Biodiversity Network, and the extension for Geosciences (EFG). ABCD is now undergoing a development, which will lead to the third version of the standard (ABCD3)
BioCASe is not integrated with the GBIF registry, and data validation pipelines are not provided, while here are several controlled vocabularies in ABCD. It can feed the GBIF by exporting data in DwC Archive format. BioCASe also allows harvesting through exporting data in ABCD archives.
The OpenUp! network uses BioCASe to connect 8.7m multimedia objects (images, videos and sound files) to Europeana, the European Digital Library.

Day 2 – 13/03/2019

Day 2 opened with three presentations, focusing on scholarly publishing pipelines.

The first one was by Thierry Burgoin, and focused on the European Journal of Taxonomy and the CETAF e-publishing working goup.
The EJT is one of the few (5%) journals published by CETAF member which is listed in the Directory of Open Access Journals (DOAJ), even if ca. 41% of them are freely available online. Thus, "open access" does not mean "freely available online", but it is something more complicated, and being listed among OAJ means to fulfill several different requirements.
Most of the new species are described on journals without an Impact Factor (IF). As far as molluscs are concernes, this is true for the 78% of new names. Since papers describing new species are crytical, beinc legal papers that document the legitimacy of names, access to publications which contain these papers is crytical as well.
The European Journal of Taxonomy (*EJT*) is a peer-reviewed, international, and fully electronic diamond Open Access journal in descriptive taxonomy in zoology, entomology, botany (in its broadest sense), and palaeontology. The journal is endorsed by the CETAF and owned by the EJT consortium, which finances the copy-editing, layout and online publication of the papers: neither authors nor readers have to pay fees.
The taxonomy-publishing landscape faces different important challenges both structural (diversity and plurality of institutions, journals, and digital initiatives dealing with taxonomy), and operational (lose of taxonomy publishing expertise) to target the need to move taxonomy (a complex science) in the digital world where numerous initiatives occur, not necessarily interoperating.
To address these challenges, to support scientific online edition of the CETAF National History Institutions, and to tackle CETAF Strategic Objectives in relation to publishing, the CETAF set-up the taxonomic e-publishing working group, in order to adderss five issues dealing with the CETAF institutional journals focussing in Taxonomy: 1) journal con-

tent inventory, 2) e-publishing impact, 3) inventory and use of existing identifiers relevant to e-publishing, 4) categorising Taxonomy within Clarivate analytics, and 5) moving toward a publishing platform for taxonomic journals (data publication and dissemination).
While fulfilling its role of being primarily a journal publishing taxonomic results, the European Journal of Taxonomy also serves as a model to test further developments for the taxonomic publishing process to meet these challenges. EJT aims at setting up a new production workflow including XMLisation process at the desk-editing level, prior to PDF publication, for producing more confident and pertinent statistics, more accurate and confident data both human and machine readable.

The second presentation was given by Donat Agosti, and focused on the limited accessibility ot taxonomic data published in scientific journals, and the PLAZI approach for mobilising them.
Taxonomic data stored in scientific journals are impressive. Over 1.9 million species have been described in scientific papers, and ca. 20 million species treatments do exist. Ca. 18.000 species are described each year, and the paper are published on scientific journals.
However, most of this knowledge is almost inaccessible. This because of:

A) Incomplete digitization
B) Texts are not ready for data mining
C) Publications are not semantically enhanced
D) Published data are not linked to the cited specimen(s)
E) Most data are not open
F) History of names is not in CoL

Digital publishing provides a mean to overcome some of these limitations.
Taxonomic publications are part of a landscape of links to different resources. These links can be extracted from the artciles, thus strongli enhancing them, and transforming them into a source of true Digital Acccessible Knowledge (DAK). Furthermore, once articles are transformed into digital objects, they can be cited properly as a whole, of as single parts, e.g., an author can properly cite a figure from an article, referring to it with a DOI. Thus, it is possible to have metrics on everything.
Taxonomic treatments can be considered the building blocks or basic data elements of taxonomic publications. They are very rich in detail. All the other elements of a publications are inferred from the analysis and synthesis of taxon descriptions. The descriptions are also the ‚legal‘ element of the publication in compliance with the ICZN. Descriptions can be further resolved into the basic units, characters in the description sensu str., and the specimen records. They could be enhanced by shared ontologies and gazetteers.
Liberating high quality biodiversity data from scholarly publications can give access to data provided by specialist, allow quality control by scientist, link to taxa and taxonomic treatments, and to related data. Each treatment has its own persistent identifier, and thus can be properly addressed.
The PLAZI pipeline allows to enhance taxonomic data, liberating them, and organising them into tretments. They are also feed to Zenodo, where each part of the article has a DOI, and can be cited properly, thus making data FAIR.
PLAZI aims at getting >50 % of newly described species into the system by 2021.
The PLAZI workflow foresees the creation of FAIR data by taking figures and other parts out, uploading them to Zenodo, and making them citable, each with a persistent identifier. Treatments have a DOI, and can be published as XML files. Furthermore, linking to external resources, e.g. names to GBIF, is allowed.
The pipeline is incredibly effective, and works well also with scanned articles in pdf format, and not only on natively digital publications. This could allow retroconversion of old journls, and of those journals which are still published only in paper-printed format.

Last presentation was made by Lyubomir Penev, and focused on scholarly publication of FAIR biodiversity data.
Being one of the first proponents of open access and open data, Pensoft has adopted

from the very start a multiple data publishing approach, resulting in a toolbox (ARPHA-BioDiv) of several novel workflows:
1. Data underlying certain research results, deposited in an external repository or as supplementary file(s), and linked/cited in the article where these results are described; supplementary files are published under own DOIs and bear own citation details.
2. Data deposited in trusted repositories and/or as supplementary files and described in data papers.
3. Integrated narrative and data publishing realised by the Biodiversity Data Journal, where structured data are imported into the article text and downloaded/distributed from there in structured format.
4. Linked Open Data (LOD) publishing of biodiversity data extracted from various literature sources and converted into interoperable RDF triples in accordance with the Open-Biodiv-O ontology.
The above mentioned approaches are supported by a whole ecosystem of additional workflows and tools, for example:
• pre-publication data auditing, involving both human and machine data quality checks (workflows 2,3);
• web-service integration with data repositories and data centres, such as Dryad, GBIF, BOLD, DataONE, LTER, PlutoF, and others (workflows 2, 3);
• semantic markup of the article texts in the TaxPub format facilitating further extraction, distribution and re-use of sub-article elements and data (workflows 3,4);
• server-to-server import of specimen data from GBIF, BOLD and PlutoR into the manuscript text (workflow 3);
• automated conversion of EML metadata into data paper manuscripts (workflow 2);
• export of Darwin Core Archive and automated deposition in GBIF (workflow 3);
• submission of individual images and supplementary data under own DOIs to the Biodiversity Literature Repository, BLR (workflows 1-3)
• conversion of key data elements from TaxPub articles and taxonomic treatments extracted by Plazi into RDF stored in and handled by the OpenBiodiv Knowledge Management System.
These approaches represent different aspects of the prospective scholarly publisging of biodiversity data, which in a combination with text and data mining (TDM) technologies for legacy literature (PDF) developed by Plazi, lay the ground of an entire data publishing ecosystem for biodiversity, supplying FAIR and re-usable data to several interoperable overarching infrastructures, such as GBIF, BLR, Plazi TreatmentBank, OpenBiodiv and others.

At the end of the presentations, which were intermixed with comments and questions, discussion followed, addressing several relevant aspects.
1. Metrics as incentives for data publishing and demonstrating the usefulness of data publication effort (by institutions). It is obviuous that institution especially are willing to understand whether their efforts fro publishing data are effective, and whether they increase the use of their collections. Now the attention is shifting from the simple owning of collections, which granted prestige to an institution, to publishing, and now to being cited. And this is the proper way to demonstrate the usefulness of an institution hosting collections. In many countries, eg. Portugal or Spain, if an institution cannot demontrate its usefulness by showinghta tits collections are use, it is closed. Thus, institutiona are craving for metrics. Mobilising data can help in increasing their impact on the scientific world.
2. Proper citation of GBIF datasets in scientific literature. Given the craving for citations which is arising fro institutions holding collections, there is a poor practice in citing datasets, and especially those extracted from the GBIF. Even if each dataset downloaded from the GBIF has a unique identifier, and thus can be easily cited, most papers cite the aggregator, and not the dataset, which is properly addressed only in 3-5% of the cases. This is clearly a drawback, which strongly biases the metrics, thus not providing to collection holders what they need for surviving, and to continue the expesive digitization effort. Possibly, the WG could provide guidelins to Editors, pushing them to force proper citation of digital objects in the paper they publish on scholar journals. Clearly, data pa-

pers are more simple to cite than datasets. Thus, all institutions should be pushed to publish their datasets not only as they are in the GBIF, or in other relevan repositories, but also as data papers in open access scholar journals.
3. Training for publication? Training can be a key for improving both data publication and data citation. Often poor citations are due to lack of knowledge on ho data are published, and how datasets and other digital objects can be properly cited in a scientific paper. Thus, training courses for researchers could be a key to improvce these practices, and to decrease the above cited bias in the metrics o digital objects usage. Maybe, even introducing a new standard for how to cite collections (and digital object) and institutions could be created. In any case, a good knowledge of digitization pipelines could be a good thing, and move towards an increase in digital object, and to better citation practices. Each institution should have a person in charge for acquiring such knowledge from training courses, guidelines and best practices.
4. Software diversity. Each country has its own software/platform for digitizing collections. This creates a diverse landscape, especially when these collections are not feed into the GBIF, or any other relevant aggregator, thus making proper citation even more difficult. Often, these software are more fit for administrative data, than for biological data, thus resulting in digitization efforts which limits the usability of the data. Software as Specify can be a solution, but are not widely adopted, and plus do not fit all data (e.g., geological collections).
5. an issue can be that the GBIF do not work with collections, but with datasets. Thus, datasets are cited. However, institutions want credit for collections, while datasets are a little bit confusing concept. Thus, it could be necessary to enhance collection representation in the GBIF infrastructure, and to provide a proper standard to cite collections and institutions.
6. value of dataset/collections publication/citation for careeers. While publishing datasets/collections in the form of data papers can solve the issue, there are still al lot of datasets already published, and/or researchers which are not willing/able of preparing a data paper for each dataset they publish. Thus, it could be interesting to explore the possibility of making datasets publication/citation usable for career progression, both for collection keeper and acabpdemics
7. taxonomic impediment. since 20 years ago little is changed. taxonomy is seen as a service. Analysing is pushed, especially in terms of career advancement, while describing is not. Thus, taxonomists are disadvantaged. However, the need for taxonomists is still relevant. Proper taxonomic quality control in scientific journals is often missing, and this badly limits the exploitation of data stored in scientific palers and/or datasets. Proper taxonomic audit should be forced, at least in data paper, even if it is a cost.

Wrapping up

- citation practices: datasets should be properly cited, and editors/publishers should force proper citation practices
- publication/citation relevance: dataset piblication and citation shouls be relevant for career progression
- all data are relevant: all data should be published, not only those which are seen as relevant by keepers/researchers
- data paper do it better: the publication in the form of data papers should be preferred, since allows easier citation, increase reusability and lowers the risk of data loss.
- audit for taxonomy: this should be a best practice in all biodiversity data publishing pipelines
- communication with IT personnel: researchers/collection keepers should be trained to communicat with IT personnel, so that the two can properly under-

stand each other. Thus, IT personnel can provide software tools which can actually be useful. Plus, they could support data publishing properly.

- DOI must be always used in citations