BioCASe IPT Workshops Report

Tim Robertson, Jörg Holetscheck, Andrea Hahn, John Wieczorek, Peter Desmet

As part of the TDWG 2020 working sessions week, two 90 minute workshops were run to capture ideas to assist in planning for the future of the BioCASe Provider Software and the GBIF Integrated Publishing Toolkit (IPT).

Table of Contents

Process

In preparation an *ideas paper* (https://doi.org/10.35035/cdps-md62) was shared beforehand capturing possible topics for discussion.

The workshop format was a series of presentations and discussion, the main outcomes of which are captured in this document. The sessions are available on YouTube

- Session 1: https://tinyurl.com/tdwg2020-bps-ipt-a
- Session 2: https://tinyurl.com/tdwg2020-bps-ipt-b

The presentations are available in the google drive folder https://drive.google.com/drive/folders/1NbIqMRnoaVI_c0fu2TI-nGvRY3JcMHHd

115 participants joined between the days; see the participants table at the end of this document.

Presentations

The following presentations were given.

Presentation	Key topics
Day 1: Introduction *Jörg, Tim, Andrea, Matt*	- Long history of use with both tools - Rapid increase in use of GBIF IPT in recent years
Day 1: OBIS Perspectives (IPT) *Pieter Provoost*	- Star schema limitations and workaround - Streaming data - Validation / data quality control - Registration of networks
Day 1: INBO Perspectives (IPT) *Peter Desmet*	- Peer review of data - Validation / Data quality control - Improving metadata capture (e.g. from data) - Using alternative repositories (zenodo)
Day 2: AVH Perspectives (BioCASe) *Niels Klazenga*	- Complex data mapping - Federated querying - Limitations of BioCASe tools - BioCASe and aggregation portals - Incremental updates
Day 2: VertNet Perspectives (IPT) *David Bloom*	- Importance of user experience (simplicity and stability) - Limitation in standards use (namespaces) - Metadata improvements
Day 2: Frictionless Data *André Heughebaert*	- An alternative for the DwC-A standard - Schemas - Structure - Data validation - Available tools

Key Discussion points

This section summarises the key points raised during the meeting; from the ongoing zoom chat and raised verbally.

The function of the IPT as a repository

There was discussion around the deployments of the IPT and the needs and perceptions of it as a web-based institutional repository.

- Participants emphasised that the IPT is seen as a “gateway” tool to shape data for sharing and not as an archival repository. In several cases participants stated that the data is archived elsewhere, often in a richer format than the IPT supports.
- There was no strong need from the participants that the IPT be run within an institution, and for many a central, or a few well-managed installations, would be preferable to simplify things; for data publishers and for technical support/upgrades etc.
    - The ability of GBIF to provide a centralised repository was raised as an option to explore, noting that an independent certification of robustness (e.g. CoreTrustSeal) may be needed
    - It was noted that several GBIF Nodes operate well-managed data hosting capabilities using the IPT with ongoing support
    - It was noted that for database-driven datasets, and in particular those on automatic schedules, may necessitate the need for local installations to safeguard access to databases.

The publication workflow

There was discussion around the publication workflow (mapping, publication, registration) for datasets within the IPT.

- Having the entire publication workflow (mapping, documenting, publication, registration) in one tool has one of the main reasons the IPT has been successful.
- The ability to review a dataset after data mapping was seen as desirable. This process may involve peer review (perhaps with user assignment), but would need to allow the manager to make changes before the dataset is made public.
    - Ability to create DwC-A before (re)publication was seen as desirable
    - Use of data quality reports (see below) would be a good addition
    - Ability to preview the data would be a good addition (possibly including using the GBIF sandbox environment www.gbif-uat.org)
    - Care should be taken to keep the IPT simple - preview and quality tools should be external services
- Registration is geared towards the GBIF network, and exploration of how this process could be improved to better support the needs of the OBIS network and perhaps the GeoCASe and OpenUp! networks should be undertaken
    - Datasets that might not be suitable for or desired to be indexed in GBIF
    - Better representation of “sub-networks” within the GBIF registry
    - Self-declaration of which aggregators should index data
    - Maybe keep this generic, so that potential future networks can jump aboard

- Provide ability to copy the data mappings from an existing resource to save manual editing
- Pay close attention to the training material and the simplicity of the publication workflow. Data managers need to be able to “self-serve” following training with minimal input from administrators.

Data quality

Data quality was a topic mentioned at several points during the workshop.

- There is a strong desire for the tools to help users produce high quality data, and flags of issues
- The tools must be kept simple, and there is a preference to make use of external service(s) to provide this function
- Data quality assertions should be available to a data manager before they publish the data
- Possible solutions could involve
    - Using the GBIF data validator to run the quality checks GBIF ingestion runs and make the report available to the IPT
    - Letting users preview data before publishing, by indexing in the GBIF sandbox environment (www.gbif-uat.org)
    - Allow users to define data specifications such as https://github.com/inbo/whip
    - Use Frictionless Data schemas that allow for constraints to be defined
    - Support better data vocabularies for controlled values

Metadata content

There was discussion around metadata capture for the IPT.

- A proposal was made to have required/recommended metadata visually stand out from optional metadata in the editor (cf. Zenodo).
- A proposal was to make use of the data to autofill metadata fields, such as temporal coverage, geographic coverage
- The TDWG Collections Descriptions group propose a review of the GBIF Metadata profile and inclusion of a simple schema of the newly emerging CD standard
    - Possibly involving forms to fill in (i.e. an editing tool), or possibly involving the ability to upload spreadsheets
- Some IPT users deal with data already documented to a richer format of EML than the IPT supports; the IPT should let people upload this to create a dataset and keep the richer content
- It was recognised that datasets often relate to multiple projects, so the multiple project identifiers should be supported.

Usability issues

In addition to the issues known on the GBIF IPT issue tracker (https://github.com/gbif/ipt/issues) the following points were emphasised.

- Metadata authoring is the biggest area of concern; in particular the need to input data for the same person (contact) repeatedly
- The IPT double logging-in issue causes *significant* annoyance
- Having the publishing schedule reset by publishing manually causes maintenance issues
- Provide ability to copy the metadata from an existing resource to save manual editing (e.g. metadata templates)

Data structure

The star schema imposed by the Darwin Core Archive was discussed at several points.

- The rigidity of the star schema causes frustration with many and is a limiting factor for adoption; in particular for communities with sampling event and environmental trait data
    - Even if GBIF do not index it in GBIF.org users would like to prepare/map rich datasets
- The Frictionless data or W3C CSV recommendations now offer solutions to the star schema issue
    - If adopted, make sure migration is as simple as possible for users
    - Suggest to support multiple formats simultaneously
    - Some raised caution that the flexibilities in data mapping (e.g. many-to-many joins) may raise problems for consuming data reliably; use of a fixed set of shared schemas may be a route to overcome this.
    - It was noted that while Frictionless Data Packages will help in data structure, it has limited support for structured metadata as it uses a markdown format for generic descriptive text. Policies for including e.g. EML in the package would need to be considered, similar to DwC-A.

Querying and exposing data

The BioCASe Software allows querying and live access to data by using BioCASe protocol requests. This can be used to retrieve full records for subsets of the published data (*Search* request) or to inspect all distinct values for certain data items (*Scan* request).

- Even though this is a useful feature for implementing incremental updates, this is not used for on-the-fly searches anymore, availability of providers and performance issues don't allow that. Indexing/caching and portals are standard.
- Useful for getting counts of fields
    - Suggestion to have a summary of useful counts in the IPT (list distinct values from every column)

- Useful to be able to get a filtered subset of dataset as a separate downloadable archive.
    - In BioCASe, this can be done by the admin, it's called a filtered download. The same BioCASe filter that can be used in a search request can be used to create an XML archive that stores all marching ABCD records.
    - Recognition that GBIF/iDigBio/ALA/OBIS portals etc does this
- versioning system (more advanced view and control over the versioning process)
    - integrate a diff to compare versions of datasets to generate a type of changelog
- Harvest log (better integration to know when you were last hit by e.g. GBIF)
    - Strong support
    - This is particularly useful if a provider is part of several networks (e.g. GBIF and GeoCASe and OpenUp) and eliminates the need to merge logs from different sources.

Bundling of tool(s)

Some discussion around the tool bundling was held.

- There is a strong desire to keep the IPT simple. It is still a useful tool and with some relatively minor improvements (usability, quality reports, peer review step, more expressive models) could be even more usable.
- Modularisation and multiple tools are preferred over a one-tool-to-rule-them-all approach.
- While modularity and use of Python was of interest, Java was also supported by some, and recognised for its stability over long time periods.
- The idea of merging the BioCASe tool and the IPT was considered as too early to discuss and raised questions; particularly around complexities of ABCD mapping. More clarity on the proposal is needed before input could be given.

Next steps

This workshop aimed to capture ideas without making decisions.

There were some clear and concrete proposals for the IPT to consider. It is already clear that many of these could be accommodated within the framework of the current tool and would be well-received by the existing IPT community. GBIF will review the ideas, document a short roadmap capturing the achievable goals and circulate this amongst the IPT community for further comments. It is foreseeable that this could form the basis of an IPT 2.5.x development stream during 2021 and rolled out without major disruption (note the IPT is on 2.4.x).

It became clear that certain special interest networks rely on the BPS and expect the tool to evolve and be supported in the future (GeoCASe, GFBio, GGBN). However, not enough discussion or feedback from the BioCASe community was expressed to have a clear picture for a roadmap. The BioCASe team will consider a more targeted approach with existing data

publishers before committing to next steps. Porting BioCASe to Python 3 is an imminent issue, since support for Python 2 ceased in 2020. This is already under way and a Python 3 version of the BPS will be released in 2021.

All participants were reminded of ways they can contribute further ideas.

1. GBIF IPT issues (https://github.com/gbif/ipt/issues)
2. IPT mailing list (https://lists.gbif.org/mailman/listinfo/ipt)
3. GBIF Discourse forum (https://discourse.gbif.org/)
4. Mail Tim Robertson (trobertson@gbif.org) or Jörg Holetschek (J.Holetschek@bgbm.org)

Attendees

	Name	Day 1	Day 2
1	Abby Benson (GBIF-US, OBIS-USA)	x	x
2	Alex Hardisty (Cardiff University, UK) (partial-attendance)	x	
3	Alfonso Christopher Zuñiga Hartley		x
4	André Heughebaert (Belgian Biodiversity Platform)		x
5	Andrea Hahn (GBIF)	x	x
6	Andrew Doll (DMNS/Arctos)		x
7	Anke Penzlin (Senckenberg, Germany / GFBio)	x	
8	Anne-Sophie Archambeau (IRD/GBIF France)	x	x
9	Anton Güntsch (BGBM)	x	
10	Anton Van de Putte (RBINS, biodiversity.aq)	x	
11	Arnold Andreasson	x	x
12	Ben Norton (North Carolina Museum of Natural Sciences)	x	x
13	Beth Gamble (Smithsonian - National Museum of Natural History - IT - Informatics Office)	x	x
14	Birgit Klasen (ZFMK Bonn, Germany)	x	x
15	Brenda Daly (SANBI, South Africa)	x	x
16	Brenda Nyaboke (National Museums of Kenya)	x	x
17	Bushra Hussaini (American Museum of Natural History)	x	x
18	Carlos Martínez (ZMUT, Turku, Finland and Myriatrix)	x	x

19	Chihjen Ko	x	x
20	Christian Köhler	x	x
21	Christina Byrd (Harvard Museum of Comparative Zoology, USA)	x	
22	Cristina Villaverde (GBIF Spain)	x	
23	Dag Endresen (GBIF Norway, University of Oslo)	x	x
24	Damiano Oldoni	x	x
25	Dave Martin	x	
26	David Bloom (VertNet)	x	x
27	David Fichtmueller (BGBM Berlin)	x	x
28	David Shorthouse (AAFC, Ottawa)	x	x
29	Debora Arlt (SLU Swedish Species Information Centre, Sweden)	x	
30	Deborah Paul (Species File Group, INHS)	x	x
31	Dimitri Brosens (Belgian Biodiversity Platform)	x	
32	Diversidad Biológica de Guatemala -SNIBgt -CONAP-	x	
33	Falko Glöckler (MfN, Berlin, Germany)	x	x
34	Federico Mendez (GBIF)	x	x
35	Genevieve Tocci (Harvard University Herbaria, USA)		x
36	Gil Nelson (iDigBIo)	x	x
37	Guido Sautter (Plazi)	x	x
38	Hanieh Saeedi (Senckenberg Museum and OBIS)	x	
39	Holly Little (Smithsonian - National Museum of Natural History - Paleo)	x	x
40	Ian Engelbrecht		x
41	James Macklin (AAFC, Ottawa, Canada)	x	x
42	Janaki Krishna (UMNH, Utah)	x	
43	Jean Woods (DMNH, USA)	x	
44	Jean-Marc Vanel		x
45	Jeroen Creuwels (NLBIF/Naturalis Biodiversity Center NL)	x	x
46	Jiangning Wang	x	x

47	Jodi Shippee (Vermont Natural Heritage Inventory, USA)	x	
48	Joe Miller (GBIF)	x	x
49	John Torgersen (Canadian Museum of Nature)	x	x
50	John Wieczorek (VertNet)	x	x
51	Jon Pye (Ocean Tracking Network)	x	x
52	Jose Martin Carrasco Montoya	x	
53	Joseph Chipperfield (NINA)	x	
54	Josh Humphries (NHM, London)	x	x
55	Judith Weber (Uni Bremen, GFBio)	x	
56	Jörg Holetschek (BGBM Berlin)	x	x
57	Kate Webbink (Field Museum of Natural History)		x
58	Korbinian Bösl		x
59	Laura Anne Russell (GBIF)	x	x
60	Lenore Bajona (Ocean Tracking Network)	x	x
61	Leonardo Buitrago (GBIF)	x	
62	Ls Zmvc	x	x
63	Lutz Suhrbier (BGBM Berlin)	x	x
64	Maren Gleisberg (BGBM Berlin, Germany / GFBio Project)	x	x
65	Markus Weiss	x	x
66	Martin Käck (Swedish Species Information Centre)	x	
67	Mary Kennedy		x
68	Mathias Dillen (MeiseBG)	x	x
69	Matt Woodburn (NHM London)	x	x
70	Matthew Blissett (GBIF)	x	x
71	Melisa Ojeda (Sistema Nacional de Información sobre	x	
72	Michelle Kennedy (Harvard Museum of Comparative Zoology, USA)	x	x
73	Nacho Felpete (Uni León, Spain. AIMJB IPT admin)	x	
74	Naomi Tress	x	x

75	Nicolas Noé (Open science lab for biodiversity, INBO, Belgium)	x	x
76	Niels Klazenga (Royal Botanic Gardens Victoria)	x	x
77	Niels Raes (Naturalis Biodiversity Center)	x	
78	Niki Kyriakopoulou		x
79	Olle Hints (CETAF ESG)	x	x
80	Patrícia Madeira (CIBIO-Açores, University of the Azores)	x	x
81	Patricia Mergen (Meise Botanic Garden/ Royal Museum for Central Africa)	x	x
82	Patrick Cox (RBG Kew, London, UK)	x	
83	Paul Morris (MCZ and Harvard Herbarium)		x
84	Paula Zermoglio (VertNet)	x	x
85	Peggy Newman (Atlas of Living Australia)		x
86	Peter Desmet (INBO)	x	x
87	Peter Grobe (ZFMK Bonn)	x	x
88	Pieter Provoost (OBIS)	x	
89	Rebecca Snyder (NMNH, Smithsonian)		x
90	Ricardo Ortiz (SiB Colombia)	x	x
91	Richard Pyle (Bishop Museum)	x	x
92	Rob Turner	x	
93	Rosa Bolaños (Instituto Nacional de Biodiversidad, Quito-Ecuador)	x	
94	Ruben Perez Perez	x	
95	Rui Andrade (AzoresBioPortal, University of Azores)	x	x
96	Rui Figueira (GBIF Portugal)	x	x
97	Rukaya Johaadien (GBIF Norway, University of Oslo)	x	x
98	Saara Suominen	x	
99	Sharon Grant (Field Museum of Natural History)		x
100	Sophie Pamerlon (GBIF France - UMS PatriNat)	x	x
101	Steve Baskauf (Vanderbilt Libraries)	x	x
102	Sylvain Morin (GBIF France)	x	

103	Takeru Nakazato	x	
104	Talia Karim (U. of Colorado)	x	x
105	Tanja Weibulat	x	
106	Thad Wilson	x	
107	Thomas Orrell (NMNH, Smithsonian)	x	x
108	Tim Robertson (GBIF)	x	x
109	Tomer Gueta (bdverse)	x	x
110	Utsugi Jinbo (GBIF Japan, National Museum of Nature and Science)	x	x
111	Walter Berendsohn (BGBM Berlin)	x	x
112	Wataru Ohnishi		x
113	Wiebke Walbaum	x	
114	William Ulate, Missouri Botanical Garden / CRBio	x	x
115	Yi-Ming Gan (RBINS, biodiversity.aq)	x	x