Advancing the IPT and BioCASe toolkits

Jörg Holetschek, Tim Robertson, Matthew Blissett

Version 27acaab, 2020-09-18 11:49:32 UTC

Table of Contents

Colophon

Jörg Holetschek, Tim Robertson, Matthew Blissett. Advancing the IPT and BioCASe toolkits. GBIF Secretariat: Copenhagen. https://doi.org/10.35035/cdps-md62

Authors

Jörg Holetschek, Tim Robertson [https://orcid.org/0000-0001-6215-3617], Matthew Blissett [https://orcid.org/0000-0003-0623-6682]

Licence

Persistent URI

https://doi.org/10.35035/cdps-md62

Cover image

Perentie (*Varanus giganteus*), Macleod, Western Australia. Photo 2012 Tim Robertson via iNaturalist research-grade observations [https://www.gbif.org/occurrence/2423018547], CC0 [https://creativecommons.org/publicdomain/zero/1.0/].

Background

This ideas paper explores needs and opportunities for the future direction of the BioCASe Provider Software [https://www.biocase.org/products/provider_software/] (BPS) and the GBIF Integrated Publishing Toolkit [https://www.gbif.org/ipt] (IPT). These tools have been fundamental components used to build open biodiversity data infrastructure. Closely following the TDWG standards [https://www.tdwg.org/standards/], they have been used for more than a decade across several data networks and have an established community of users. This experience has led to a good understanding of the limitations of the existing products which we aim to summarize in this consultation. Additionally we recognize that standards have evolved, and there have been changes in open data practices, such as an increased use of open data repositories and reliance on hosted infrastructure. We will also consider if the scope and functionalities of the products still address today's needs of the communities running open data networks.

IPT Features

The IPT is a Java-based web application that can be used to run a data repository service to publish biodiversity data. The history, key functionality and design decisions leading to the current version of the IPT are well documented.[1] A brief summary of functionality is listed:

1. Support multiple users with distinct permissions to administer the software and to manage the resources it hosts.
2. Upload spreadsheet, delimited text file (e.g. CSV) or connect to a database (requiring a network connection) to use as sources of data.
3. Map the data content (e.g. fields in a database or spreadsheet) from the source dataset to the terms in the Darwin Core standard. The IPT provides capabilities to follow changes to the standards by connecting to a central registry.
4. Enter dataset metadata that specify scope, methodology, ownership, rights, etc.
5. Produce a Darwin Core Archive and a publicly accessible web page that shows the metadata and links to the archive and other documents that were created. As versions are published, it is possible to archive historical versions.
6. Automate the registration of datasets (or notify of the availability of new versions) in the GBIF registry (https://registry.gbif.org) so they are discoverable, and indexed by GBIF and others.

BPS Comparison to IPT

The BPS provides similar capability to the IPT of enabling a structured mapping of database content to a versioned data standard. As a Python-based web application, an administrator can connect BioCASe to a relational database and map tables and views to elements of one of the XML-based data standards the tool supports (e.g. ABCD[2]).

In contrast to the IPT, typical data sources are the institutional relational databases, not delimited text files. Once configured, datasets will be published as a BioCASe web service[3], allowing a live request on the underlying data. In addition, all data published can be stored in (ABCD-)XML or Darwin Core archives

The following feature matrix summarizes the key distinctions between the BPS and the IPT:

Feature	**BPS**	**IPT**
Input data sources	Relational databases	Delimited text files (CSV), Excel, DwC-A or relational databases
Output formats	XML, DwC-A	DwC-A
Data standards supported	ABCD (+ extensions), Darwin Core	Darwin Core
Single-record retrieval	Yes	No
Live access	Yes	No. Latest or archived version export of dataset only
Data validation	Very basic (mandatory fields, data types, simple vocabularies)	Limited. Checks on `occurrenceID` and warnings of data not aligning to the recommended vocabulary values

Versioning support	No	Yes, with archival and DOI support if configured
Incremental updates	No	No
Metadata authoring	No	Yes
Automatic registration	No	Yes, to GBIF Registry
Runtime environment	Python	Java

How to respond to this Ideas Paper

The purpose of this paper is to capture ideas, topics that need further exploration and opinions for consideration for the future roadmaps of the BPS and IPT *without offering recommendations or decisions*. We welcome contributions as follows:

- Join the online discussion at the 2020 TDWG workshops [https://www.tdwg.org/conferences/2020/working-sessions/#ws01:%20capturing%20ideas%20for%20the%20future%20of%20biocase%20provider%20software%20and%20the%20gbif%20integrated%20publishing%20toolkit%20(ipt)] (see schedule [https://www.tdwg.org/conferences/2020/working-sessions-schedule/])
  - Propose to present (max 7 minutes) of your idea, limitation or vision for the product
- Open discussion topics on the GBIF open forum [https://discourse.gbif.org/]. These could be responses to topics raised here or offering new ideas
- Register feature requests on the IPT GitHub issues [https://github.com/gbif/ipt/issues] if they are not already captured
- Please contact trobertson@gbif.org [mailto:trobertson@gbif.org] (Tim) or j.holetschek@bgbm.org [mailto:j.holetschek@bgbm.org] (Jörg)

1. Purpose of the tools

This section covers questions to explore and ideas relating to the scope of the tools and the purpose they fulfil for the institutions using them.

1.1. Federated (live) search

The BPS was designed to expose an institutional database on the web, so that it could participate in federated queries across institutions.

1.1.1. With a recent trend towards versioned dataset exports, is there still a demand for live, federated queries? If so, is BioCASe protocol still the preferred mechanism for this?

1.1.2. Would the support of RDF as output format create a need for a (live) detail view for individual records?

1.2. Online repository, desktop tool or traditional provider software (server installation)?

The IPT was developed to enable institutions to run a data repository on the internet, the BPS to provide an additional access point to a local database. There is a growing trend to use hosted infrastructure (e.g. Public clouds and shared IPTs) and open repositories (e.g. Zenodo).

1.2.1. How important is it to run installations of a repository within your own institution? Network access to a local database?

1.2.2. Is there a preference that the GBIF (or other network) provide a *central/hosted* tool offering data mapping and upload functionality?

1.2.3. An alternative model could be a desktop tool, allowing users to document metadata, format datasets but then push them to a hosted repository. This would remove the need to run a server. Would this be desirable?

1.3. Merging the BPS and IPT into a single product

The feature matrix above illustrates significant overlap in the IPT / BPS feature set.

1.3.1. What considerations need to be made as we investigate the possibilities of merging these into a single tool?

1.4. A single tool or a *toolkit*

Both tools are single products packaging several functions (metadata authoring, data mapping to standards, archiving in a repository and registration services).

1.4.1. Should we consider exploring a modular approach so that modules could be embedded in products beyond an institutional toolkit? An example could be a multilingual metadata authoring “wizard”, useful to embed in the GBIF registry, the IPT or network sites.

1.5. Data quality validation

Neither the BPS or IPT offer capabilities to report on, or enhance the quality of data beyond basic integrity constraints. There is potential to explore embedding validation routines, or making use of online services during the publication process.

1.5.1. Should the tool(s) bring more attention to data quality issues during the publication process, and how would people like to see this achieved – a report, a search portal, integrating with an online service?

1.5.2. Should the tool(s) include the capability for peer review of data and metadata in the publication process?

1.6. Standards and the data model

The IPT supports EML extended with some basic terms for NCD for metadata and Darwin Core Archives including versioned extensions served from a central repository. The BPS supports mapping to any XML schema but is most commonly from the ABCD family, including domain specific extensions like GGBN (Global Genome Biodiversity Network) and EFG (Extension for Geosciences).

It is well known that the star schema imposed by DwC-A has been a constraint for many uses within the wider GBIF community, but spreadsheet/tabular data has been well received.

1.6.1. Should the tools be looking to use more expressive tabular data models, and if so, what standards today would be useful to explore (e.g. Frictionless Data)?

1.6.2. Should the IPT provide better means to document collection related metadata, such as the collection descriptions proposed by the forthcoming edition of the TDWG CD standard?

2. Specific features and functionalities

This section covers ideas relating to usability issues and features lacking in the existing tools.

2.1. Single record retrieval

Based on the purpose of live federated searches across networks, the BPS allows accessing individual records or a subset of records instead of whole datasets.

2.1.1. Keeping in mind the trend towards versioned exports of entire datasets: Is this still needed?

2.2. Standards: Darwin Core/ABCD (+GGBN/EFG)/MIxS/customized sets of terms?

The BPS can be used with an arbitrary XML schema, but is centred around the ABCD standard and its extensions; the IPT is intrinsically connected with Darwin Core.

2.2.1. How close should the tool be coupled with the standard(s) supported, keeping in mind that a complete decoupling would limit certain features such as data validation?

2.2.2. Since both Darwin Core and ABCD need to be supported: Will there be two mappings if you want to provide data in both standards? Or will there be one mapping and the output will be converted into another standard upon request (as implemented now in the BPS, which transforms ABCD documents into Darwin Core archives).

2.3. Input formats: CSV, relational databases?

Typical input data for the IPT are delimited text files (with the possibility to also import data from simple database tables and views), whereas the BPS was designed to load data from (potentially complex) relational databases. This is connected with the specifics of the data standards – Darwin Core being rather flat, ABCD having a hierarchical structure with many repeatable elements.

2.3.1. Should the new tool support both source types?

2.3.2. Are there any other input formats we should support? Frictionless data?

2.4. Output formats: DwC-A, XML, RDF, Frictionless data, JSON?

The IPT produces Darwin Core archives that store delimited text files, one core file (typically an occurrence) and potentially several extension(s) connected through the star schema. Due to its XML nature, the ABCD standard used by the BPS allows 1:n relations between entities other than the occurrence.

2.4.1. With the limitations of the star schema, should DwC archives evolve into Frictionless data packages?

2.4.2. Is XML still a desirable output format?

2.4.3. Other output formats – RDF, JSON, Catalogue of Life data packages? If yes: Should these be produced by the tool natively or could these be generated from other formats on the fly, e.g. by using Frictionless data?

2.5. Versioning

The rising importance of proper citing requires that changes in the data can be tracked through versioning.

2.6. Support of incremental updates

Citizen science has shifted dataset sizes into hundreds of millions of records, with updates putting a heavy load on providers and harvesters.

2.6.1. Are incremental updates that just hold updated and new records required or not an issue?

2.7. Dataset descriptions (Metadata)

The IPT supports metadata authoring to the EML standard, extended to support some basic collection descriptions from the (original) NCD standard.

2.7.1. Is EML still the most appropriate format for metadata, and is the profile supported by the IPT adequate. Should it describe the data files in the EML to enable wider interoperability with e.g. ILTER networks?

2.7.2. There is potential to derive metadata, such as taxonomic, temporal and geographic scope from data. To what extent should metadata be derived from the data?

2.7.3. IPT users have reported frustration in the need to re-enter contacts (people) repeatedly. Improving the usability of this section of metadata is desired (e.g. reusing existing details, or drawing information from public registries like the ORCID system). What other usability issues in metadata authoring should be addressed?

3. Integrations with networks

The BPS and IPT have supported communities to form data sharing networks by allowing for multiple repositories to be installed, providing common data schema repositories (e.g. the DwC-A extensions) and through registration with GBIF. This section discusses topics relating to the integration with other networks, infrastructure and services to support establishing a data sharing community.

3.1. Registries

The IPT can connect to GBIF's registry and automatically register published datasets.

3.1.1. Are there other registries that should be taken into account?

3.1.2. If the tool should be used in another context than GBIF – how can registration be facilitated?

3.2. Identifiers

The GBIF IPT integrates closely with the GBIF network to ensure datasets are uniquely identified allowing for updates to be tracked, and to link them to the organizations to ensure appropriate credit is given. An IPT can additionally be configured to issue DOIs through connection with DataCite for those institutions who do not wish to rely on GBIF-registry issued DOIs.

3.2.1. Are there other identifier agencies or processes around identifier management that you would like to see accommodated, such as RORs [https://ror.org] for organizations or ORCIDs [https://orcid.org/] for (living) people?

4. Technologies and process

The IPT and BPS are both developed using different software platforms (Java / Python) and have been designed, developed and maintained by a small centralized team with many external contributions. This section discusses technologies and processes for future revisions.

4.1. Familiarity and change

The tools have been used extensively, are well documented, are translated in multiple languages and included tried-and-tested training material. As tools that are infrequently used, there is benefit in consistency and familiarity for users to avoid having to “re-learn” new processes and web interfaces.

4.1.1. Balancing calls for change with a desire for consistency may be a challenge and one which could be tackled in several ways; evolving existing tools slowly, offering a major new release with significant change or even designing a new tool as an alternative. Perspectives on this are sought from the community of users.

4.1.2. Migrating existing installations is a real challenge for BioCASe users. In many cases, installations have been set up by persons that are not available any more, so instances run until they’re dead. Moving to another product will require active support, maybe on-site.

4.2. Development approach

The IPT was developed in Java and BPS in Python, each offering different benefits. There have expressions of interest to use Python by potential contributors, and there appears to be good support for Frictionless data with Python toolsets.

4.2.1. Is there a desire to foster a wider community development approach around these tools? We are interested in ideas and opinions around this, including e.g.

4.2.1.1. What language and application frameworks would be preferable?

4.2.1.2. Developments in other domains that could be used?

4.2.1.3. Development language?

[1] Robertson T, Döring M, Guralnick R, Bloom D, Wieczorek J, Braak K, et al. (2014) The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. PLoS ONE 9(8): e102623. https://doi.org/10.1371/journal.pone.0102623

[2] Holetschek J, Dröge G, Güntsch A & Berendsohn WG 2012: The ABCD of primary biodiversity data access. Plant Biosystems – An

International Journal Dealing with all Aspects of Plant Biology: Official Journal of the Societa Botanica Italiana, 146:4, 771-779. https://doi.org/10.1080/11263504.2012.740085

[3] The BioCASe protocol: https://www.biocase.org/products/protocols