Research Ideas and Outcomes 3: e14675
doi: 10.3897/rio.3.e14675

Data Management Plan

Data Management Plan: Brazil's Virtual Herbarium

Dora Ann Lange Canhos ‡

‡ CRIA, Campinas, Brazil

Corresponding author: Dora Ann Lange Canhos (dora@cria.org.br)

Reviewable v1

Received: 23 Jun 2017 | Published: 23 Jun 2017

Citation: Canhos D (2017) Data Management Plan: Brazil's Virtual Herbarium. Research Ideas and Outcomes 3: e14675. https://doi.org/10.3897/rio.3.e14675

Abstract

The goal of the Brazil Virtual Herbarium is to facilitate the identification of taxonomic and geographic information gaps of plants and fungi of Brazil. The system displays the status of online data for all valid species in the List of Species of the Brazilian Flora, including those without any record. The system also compares the Brazilian states where specialists indicate that the species occurs with the states that have occurrence points in Brazil's Virtual Herbarium, highlighting the gaps. This data management plan was prepared as part of a pilot project run on behalf of the International Development Research Centre (Canada) on data management policy for development funders (https://doi.org/10.3897/rio.2.e8880).

Keywords

herbaria, botany, data management plan, IDRC, infrastructure, research data management

Data Collection

What types of data will you collect, create, link to, acquire and/or record?

BVH uses the *species*Link network as its information system, an aggregator of species occurrence records. Most data providers are herbaria from Brazil and abroad that collect, preserve, and document the occurrence of specimens in nature, but some datasets refer to

observations and not collection of specimens. Herbaria from abroad are contributing with data of samples collected in Brazil.

Each textual data record may include:

- what was collected/observed (species name and who identified the specimen, date identified);
- who collected/observed (collector name and number);
- when (date: DD/MM/YYYY);
- where a specimen was collected or observed (country, state, county, geographic coordinates, precision, description of locality);
- collection code and number;
- whether it is a type specimen;
- observations (such as barcode).

Data records may be incomplete, as one of the aims of BVH is to help herbaria in improving data quality.

Images of the specimen (voucher or live) associated to the textual record may also be available as a separate file. Enter subsection text

What file formats will your data be collected in?

The data model used is Darwin Core 2 standard (see http://rs.tdwg.org/dwc). Data providers can use practically any software and data is sent as Raw Data to a PostgreSQL database. Softwares used today include Brahms, BioCase, IPT, DiGIR Provider, Firebird, MS-Access, MS-Excel, PostgreSQL, Sonnerat, and speciesBase. Data is accessed through an on-line search interface and can be viewed as an HTML file or plotted in maps, charts, or downloaded in formats compatible with MS-Excel 2007 (.xlsx), MS-Excel 2003 (.xls) or as a UTF8 tab delimited text file.

Images standard formats include TIFF, JPEG, and PNG.

Will these formats allow for data re-use, sharing and long-term access to the data?

Yes.

What conventions and procedures will you use to structure, name and version-control your files to help you and others better understand how your data are organized?

An important concept is that each data provider is responsible for his/her data. Any modification, correction of possible errors must be done by the data provider who then sends updates to the network. *species*Link indexes the contents of standard fields that are made freely and openly available by the data provider to all interested.

Each data record has (or should have) the date the specimen was collected, the date it was identified and each dataset also has the date it last sent data to the network. *species*Link does not control versions, meaning, does not store versions over time, as updating is dynamic (5 to 15 datasets per day), growing at an average rate of 30 to 40 thousand new data records per month. A data indexer retrieves data every night from updated datasets and this data is processed throughout the day.

What *species*Link offers under citation, is a clear indication as to the source, date, and time the records were retrieved from the network.

Documentation and Metadata

What documentation will be needed for the data to be read and interpreted correctly in the future?

Darwin Core, the data model used, is fully documented and follows a common structure that has been used by biological collections for more than 200 years. New data fields are added over time as a result of advancements in science.

How will you make sure that documentation is created or captured consistently throughout your project?

Darwin Core data model, established in 2003, is used to facilitate interoperability between different data sources, and has evolved over time through TWDG – Biodiversity Information Standards. Data quality is assessed through a number of tools and applications and a report is prepared indicating suspect, inconsistent, and incomplete data to help curators in identifying possible errors.

If you are using a metadata standard and/or tools to document and describe your data, please list here.

Darwin Core Biodiversity Information Standards - TDWG 2015 standard is used as the data model. Different communication protocols are used, such as DiGIR (Distributed Generic Information Retrieval) DiGIR 2005, TAPIR (TDWG Access Protocol for Information Retrieval) Biodiversity Information Standards - TDWG 2007, and IPT (Integrated Publishing Toolkit) Global Biodiversity Information Facility 2016, and CRIA adapts to what each data provider uses.

Storage and Backup

What are the anticipated storage requirements for your project, in terms of storage space (in megabytes, gigabytes, terabytes, etc.) and the length of time you will be storing it?

*species*Link currently uses about 15 TB of total storage, mostly used by images. Our current available disk space is of about 22TB that we envisage to be sufficient for the next three years of the project. It is important to note that storage space is only one of the requirements among others like servers, disks lifetime, disks speed, servers' memory, database technologies, etc.

How and where will your data be stored and backed up during your research project?

All public information systems developed and maintained by CRIA, including *species*Link, are stored at the Internet Data Center (IDC) maintained by the Brazilian National Research and Educational Network (RNP) in Brasília. The systems are managed by CRIA in Campinas through a Virtual Private Network. A diagram of the architecture is shown in Fig. 1, highlighting the backup servers in both sites, CRIA and IDC.

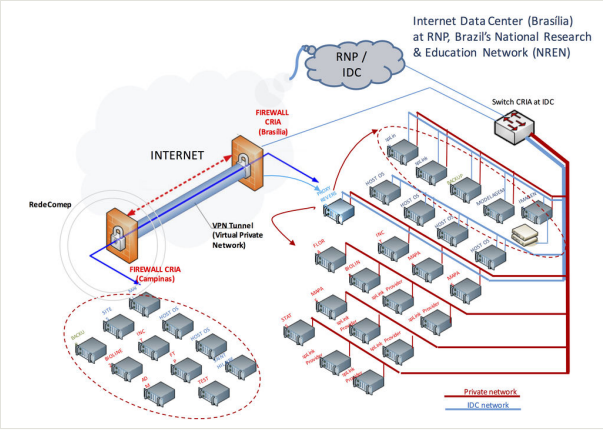

Figure 1. doi

System architecture of the BVH and supporting systems.

As can be seen in the diagram, there is a server in Brasília and another in Campinas responsible for backup. Backups of the databases and systems are carried out daily in Brasília and stored in disks. Once a week the backups are transferred to Campinas. As an additional safety measure, every month, the backups are transcribed on a tape and once every 6 months one copy is physically stored at *Embrapa Informática Agropecuária*, an institution based at the State University of Campinas.

The images are stored in a SAN (storage area network) in Brasília, managed by a specialized image software and backed up every day to our backup server in Campinas.

How will the research team and other collaborators access, modify, and contribute data throughout the project?

Each data provider (herbaria) is responsible for modifying, correcting and sending data to the network. In the case of national herbaria, a software developed by CRIA named spLinker CRIA 2009 is installed and maps data fields (in accordance to Darwin Core) and enables updates that are sent to regional servers. An indexer goes out every night to look for updates and sends the data to the network manager and to the central repository. Most international collections use IPT (GBIF's Internet Publishing Toolkit), but CRIA adapts to the system used by the data providers. Fig. 2

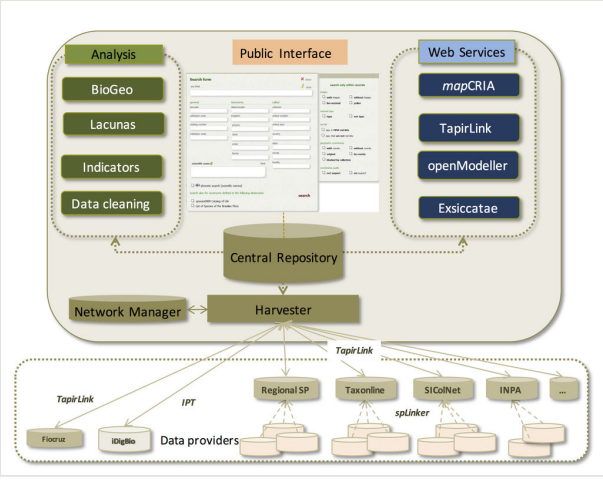

Figure 2. doi

Information and data flows in the BVH system.

Preservation

Where will you deposit your data for long-term preservation and access at the end of your research project?

Each herbaria maintains its own data together with the voucher and only sends a subset of this data to the network. So each herbaria must be responsible for the long-term preservation of the data and of the associated voucher. All data publicly shared through *species*Link is managed and maintained by CRIA. A threat is the discontinuity of support to CRIA and to the herbaria.

Indicate how you will ensure your data is preservation ready. Consider preservation-friendly file formats, ensuring file integrity, anonymization and de-identification, inclusion of supporting documentation.

The *species*Link network uses international standards that are also used by similar e-infrastructures worldwide. A user interface for searching and analyzing data is in place as are web services.

Sharing and Reuse

What data will you be sharing and in what form? (e.g. raw, processed, analyzed, final).

All data sent to the *species*Link network can be downloaded as xls, xlsx, and text delimited using tab files. Data is also served through web services and IPT DwC-A archives.

Have you considered what type of end-user license to include with your data?

All on-line data is available through a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0) license. Users are free to share (copy and redistribute the material in any medium or format) and adapt (remix, transform, and build upon the material) under the following terms:

- Attribution — users must give appropriate credit, provide a link to the license, and indicate if changes were made. This may be done in any reasonable manner, but not in any way that suggests the licensor endorses it or its use.
- NonCommercial — Users may not use the material for commercial purposes.
- ShareAlike — If the user remixes, transforms, or builds upon the material, he/she must distribute their contributions under the same license as the original.

What steps will be taken to help the research community know that your data exists?

*species*Link is a well-known e-infrastructure among the scientific community in Brazil as it has been on-line for 16 years. 95% of the herbaria that are part of BVH are associated to graduate courses, an important community that routinely uses *species*Link. Over 400 million plant records were used on-line in 2015 and 212 million in 2016 (June 21, 11:22). This represents 41 times the data available.

To disseminate new developments and increase *species*Link's visibility, CRIA maintains a Blog and a Facebook account.

Responsibilities and Resources

Identify who will be responsible for managing this project's data during and after the project and the major data management tasks for which they will be responsible.

Each data provider (herbaria) is responsible for managing its own data and all data sent to *species*Link is managed by CRIA.

How will responsibilities for managing data activities be handled if substantive changes happen in the personnel overseeing the project's data, including a change of Principal Investigator?

Small herbaria depend largely on projects for digitization. Without project support the entry of new data and further work on its quality may suffer discontinuity or may slow down. CRIA, responsible for *species*Link, also depends on projects to maintain its personnel. Lack of support may also lead to discontinuity. As to the change of the Principal Investigator, it certainly represents a loss, but the project's coordinator organized a steering committee to establish strategies and evaluate results. So a change of the Principal Investigator would not represent discontinuity.

What resources will you require to implement your data management plan? What do you estimate the overall cost for data management to be?

The overall cost to maintain *species*Link is of about 400 thousand US Dollars a year.

Ethics and Legal Compliance

If your research project includes sensitive data, how will you ensure that it is securely managed and accessible only to approved members of the project?

Sensitive or confidential data, determined as such by each herbaria, is not sent to the network. They are excluded or specific data fields are marked as blocked at the origin.

If applicable, what strategies will you undertake to address secondary uses of sensitive data?

All data sent to the network is necessarily open. But if any data field of specific records are blocked (such as geographic coordinates), users receive the information that it was blocked. This way, the system distinguishes blocked data from no data. Users are able to identify the existence of data and can request it directly from the herbaria.

How will you manage legal, ethical, and intellectual property issues?

All data sent to the network is shared under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0) license. A non-binding memorandum of understanding is signed between each herbaria and CRIA that states the obligations and responsibilities of CRIA and the Data Provider.

References

- Biodiversity Information Standards - TDWG (2007) TAPIR. http://www.tdwg.org/activities/tapir. Accessed on: 2017-6-02.
- Biodiversity Information Standards - TDWG (2015) Darwin Core. http://rs.tdwg.org/dwc. Accessed on: 2017-6-02.
- CRIA (2009) spLinker. http://splink.cria.org.br/splinker?criaLANG=en. Accessed on: 2017-7-02.
- DiGIR (2005) Distributed Generic Information Retrieval (DiGIR). http://digir.sourceforge.net/. Accessed on: 2017-6-02.
- Global Biodiversity Information Facility (2016) Integrated Publishing Toolkit. http://www.gbif.org/ipt. Accessed on: 2017-6-02.