FinBIF Data Management Plan

A Data Management Plan created using DMPTuuli

Creators: Leif Schulman, Kari Lahti, Leif Schulman

Affiliation: University of Helsinki

Template: Academy of Finland DMP

ORCID iD: 0000-0002-1990-2173

Project abstract:

DMP for the virtual research infrastructure 'Finnish Biodiversity Information Facility FinBIF'. FinBIF collects, collates, and distributes data on species and metadata on those data. The data are used for a wide range of questions in basic and applied research, for environmental and natural resources management, in education at various levels, and for business activities. In addition, they are of general public interest and, hence, used by citizens at large.

ID: 3988

Last modified: 28-04-2021

FinBIF Data Management Plan

Date of the Plan

Date of the DMP

15.05.2018

1. General Description of Data

What kinds of data are collected or reused?

FinBIF collects, collates, and distributes data on species, such as occurrence data and DNA barcodes, and also on their basic attributes (taxonomy, natural history, administrative attributes etc.). The data include geographical information, data on collectors/observers (i.e., people), information on projects etc.

The data are collected by various means: digitisation of natural history collections; long-term and fixed-term surveys, mappings, and monitoring; autonomous recordings; specialist evaluations; ad hoc observations; literature surveys; and laboratory work (DNA). Some of the data are collected as part of non-project-related data collection activities, some as part of specific research projects that use the infrastructure as a data repository.

Types of data include: tables, texts, various kinds of images, videos, audio recordings, geographic information, and amino acid sequences. Many of the data are related to physical samples in natural history collections, and reference is made to these.

What file formats will the data be in?

All data (images, permits etc) have descriptive metadata in Oracle database format.

The primary data are in Oracle database format. Secondary data (such as data from primary data sources of collaborating organisations) is in Vertica analysis database format. The data is in a triplet format (subject, predicate, object), making extending easy.

The image files are stored in Web standard formats (.jpg, .gif, .png). Original, large images are stored as .tif files.

Permits, contracts etc. are saved as PDF/A files.

In some cases, archived (i.e. very old) data may be saved in other formats. In this case the format is not strictly prescribed, but described in the metadata of the dataset. Reasonable effort is made to convert the data into the Oracle format, but this may not be practical for all data.

2. Documentation and Quality

How will the data be documented?

The data are based on ABCD and DarwinCore schemata, with some extensions based on the GGBN Data Standard. All data variables, fields and classes are documented online at http://schema.laji.fi/

How will the consistency and quality of data be controlled and documented?

Since the individual data range from nature observations made by unauthenticated individuals to monitoring schemes done with strict guidances or actual taxonomic checklists of species, the data quality will vary. As the majority of nature observations come from layman citizens (as opposed to professionals), this is to be expected. Metadata on each dataset will describe the methods used for data collection, and the expected data quality.

The quality of an individual datum is further controlled through an annotation system, which enables authenticated users to question or reinforce data from uncertain sources.

The highest quality data (taxonomic data or data on geographical boundaries) is only maintained by a limited set of authenticated users working in an expert capacity. FinBIF provides tools for appropriate data management.

Full history of changes made to primary data is saved and can be reverted if necessary.

3. Storage and Backup

How will the data be stored and backed up?

The primary data are stored permanently on database servers hosted by the University of Helsinki, with backups provided by the University IT Services. Images are stored in the IDA cloud service of the CSC.

How will you control access to keep the data secure?

Most of the data are meant to be openly accessible by anyone. An authentication system relying mostly on Haka and Virtu authentication gives access rights to data owners for write access, or others for annotation.

About 1% of the data are classified as sensitive, requiring limits to open access. Access to this data is also controlled by the authentication system.

4. Ethics and Legal Compliance

How will ethical issues be managed?

FinBIF has three major ethical concerns, as governed by applicable law: the dissemination of data on endangered species, the duty of public agencies to work transparently, and the privacy of its users.

On the first matter, FinBIF works closely with organizations that are best informed on which information on endangered species should be restricted. Data giving away sensitive information (such as the nest locations of birds

threatened by illegal egg collecting) are obscured before being made public (for instance, by only giving a rough location data, or in some rare cases, hiding them entirely). The rules for this are appended as necessary by a specialized expert working group.

On the second matter, by default all data collected by public agencies are openly available to anyone. For the data that have been restricted, a system for creating requests to give access to restricted information is implemented.

On the third matter, we provide our users the privacy protections guaranteed by European and national law (the right to access own personal information, the right to anonymity in public etc) when submitting data to the system. We do not store sensitive personal information, and the only personal information we show (when the user has not specifically requested full anonymity) is the name of the person who has provided the data.

How will ownership, copyright and Intellectual Property Right (IPR) issues be managed?

As FinBIF is an open data project, we steer people providing data towards the recommended open access licenses (the CC licenses recommended by the JHS). Voluntary data providers do, however, have the option to restrict the use rights of their data more strictly than the recommended licenses. Data restricted by the user will nevertheless be available for authorised officials for administrative purposes.

5. Data Sharing and Long-Term Preservation

How, when, where and to whom will the data be made available?

FinBIF is a public open data project: as a ground rule, all data are made available to everyone immediately. There are some exceptions: data collected as a part of an active research project has an embargo of a maximum of 4 years, and data restricted by expert opinion and the law (such as data on endangered species' nesting locations) are made less specific.

All data are available in (almost) real time on the FinBIF website https://laji.fi/ , where datasets and search results can also be downloaded by anyone.

How and where will data with long-term value be made available?

All data are made available on the laji.fi website permanently, where they can be browsed with online tools, or downloaded as datasets. The data can also be accessed with our documented API.

Have you estimated costs in time and effort for preparing the data for preservation and sharing?

Costs in time and effort for preparing the data for preservation and sharing have been estimated. However, FinBIF is a data infrastrcuture, so it is not easy to separate between costs that go exactly to preparing the data for preservation and sharing vs. time going to ICT development and digitisation of collections to be able to share and preserve the data. The total running cocts of FinBIF have been estimated at 420 000 € / year in 2018. Of this, c. 1/4 (100

000 €) could be quoted as the cost of preparing the data for preparation and sharing.