ICEDIG.EU

Grant Agreement Number: 777483 / Acronym: ICEDIG
Call: H2020-INFRADEV-2017-1 / Type of Action: RIA
Start Date: 01 Jan 2018 / Duration: 27 months

REFERENCES:
DELIVERABLE D6.4

Work package WP6 / Task T6.3.1 / Lead: University of Helsinki
Deadline M20

Innovation and consolidation for large scale digitisation of natural heritage

Digitisation infrastructure design for national open science clouds

Deliverable D6.4

Author:

Zhengzhe Wu

Contributors:

Hannu Saarenmaa, Ville-Matti Riihikoski, Abraham Nieva de la Hidalga, Alex Hardisty, Mathias Dillen, Quentin Groom

Abstract

This report describes the feasibility and potential role of national data infrastructures in large-scale digitisation of natural history collections, especially in the case of Finland. The descriptions of services, capacities, costs, and data flows from digitisation facilities to these national level systems and further to European systems are studied in the context of Finland. Discussion of use and possible designs are also included. The report is structured in five parts:

1. The Introduction part describes the background of this study and the general status of national data infrastructures in Europe. The case of Finland is emphasized and the available computing and data services are introduced.

2. The Infrastructure chapter describes FinBIF (Finnish Biodiversity Information Facility), CSC (the Finnish national IT centre for science) Pouta cloud computing services, and the Finnish Fairdata services.

3. In chapter 3, the data module used in FinBIF is presented, including the indentifier, metadata, and API (Application Programming Interface).

4. The design chapter describes the overall architectural view and data flows of the digitisation process with the use of Finnish national data infrastructures.

5. In the last chapter, the feasibility and potential role of national data infrastructures for digitisation facilities are discussed. We concluded that national data infrastructures tailored for biodiversity data , such as FinBIF, are highly necessary for the digitisation facility of natural history collections.

Table of contents

1. Introduction

There are more than two billion specimens in the world's natural history collections, 55% of which are in Europe (Ariño 2010). The digitisation of natural history collections not only facilitates the long-term preservation of the valuable biodiversity information, but also boosts the easy access and sharing of information over the world for various purposes in different work. Therefore, digitisation draws increasing attention and the number of digitisation activities is growing worldwide (Blagoderov and Smith 2012, Oever and Gofferje 2014, Tegelberg et al. 2014, Tegelberg et al. 2017, Wu et al. 2019). With the fast development of emerging digitisation and imaging techniques, the speed of digitisation is increasing, resulting in growing number of digitised specimens. On the other hand, the data size of one single digitised specimen is increasing, because of the very high-resolution images and 3D data. With the fast growth of the data volume, the requirements of data storage capacities correspondingly increase rapidly, demanding more robust and reliable data infrastructures.

In DiSSCo (https://www.dissco.eu/), a new world-class Research Infrastructure for natural science collections, it is assumed that up to 40 million specimens may need to be digitised each year so that the digitisation of a significant part of important public natural history collections can be achieved in foreseeable time. It is up to hundreds of megabytes data per each digitised specimen, which will be generated at different distributed digitisation facilities across Europe. Therefore, the data storage infrastructure at national level will play an important role to handle the data up to petabyte from the local digitisation facilities. Moreover, the data need to be accessed and analysed in one pool at European level. This requires the data flow from the national data storage further to the European level systems.

In many European countries, there are national data Infrastructures existing or are under development. For example, there are

- Czech republic: CESNET
- Finland: CSC, Fairdata services, and FinBIF
- France: CINES
- Germany: GFBio
- Netherlands: DANS and SURF
- Sweden: SND
- UK: JASMIN, JISC and NERC data centers
- Annotating shared data to Settle its reliability and quality

. They mainly provide services for higher education and research institutions, national science agencies, national libraries, and national research funders. The services include data storage, repository, preservation, reporting, metadata and ontologies, analysis, exchange, and discovery. Finland have well defined and mature national data services (RDA NDS Interest Group, 2018).

In Finland, CSC (https://www.csc.fi/), the Finnish national IT centre for science, provides data services, such as computing, data management, and data analytics. Pouta is CSC's cloud computing platforms, offering high performance computing with superior flexibility and user experience via Infrastructure as a Service (IaaS) ("CSC," n.d.). Fairdata services (https://www.fairdata.fi/), offered by the Finnish Ministry of Education and Culture and produced by CSC, consist services of data storage, search,

describe and preserve. FinBIF (https://laji.fi/), Finnish Biodiversity Information Facility, is the Finnish national infrastructure for biodiversity information. It compiles Finnish biodiversity information to one single service for open access sharing.

Digitised specimens, the output of digitisation of natural history collections, contain three part, PID, metadata, and contents, e.g. images, 3D data, and DNA barcodes (Hardisty et al. 2019). The diagram of a digital specimen is shown in Figure 1. Therefore, not only the specimen contents like images have to be preserved, but also the specimen metadata need to be deposited. Specimen metadata is important in the data FAIRness of the digitised specimens. It will make the specimen easy to find, access, interpolate, and re-use. Moreover, digitised specimens have to be open-access, with the exceptions of sensitive data that have legal basis.

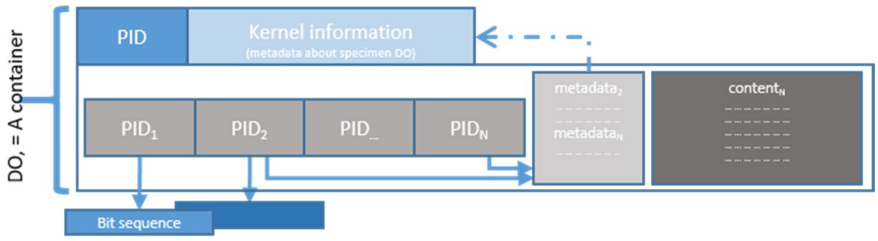


Figure 1. The diagram of a digital specimen. (Hardisty et al., 2019)

Not all data infrastructures are suitable for digitised specimens. Firstly, the PID may not be compatible. Secondly, it lacks the flexibility on the metadata terms. The specimen metadata have terms that may not be available at the data services, making the data incomplete. This will also limit the data FAIRness. Moreover, there may only support limited file formats, so the upload of un-supported data will be restricted. Generally, it lacks generic data services to handle universal digital objects, such as digital specimens. To tackle this, it requires the integration and further development of the existing services from the data infrastructures to arhive the dedicated data infrastructure for digitised natural history collections.

In our work for ICEDIG (https://www.icedig.eu/) task T6.3.1, we will do the case study of Finland on the Finnish national data infrastructures. Especially, we will explore the FinBIF infrastructure on digitising specimens from natural history collections.

1.1. Certifications

FinBIF is built on the both institutional and national IT infrastructures. It has not yet applied for Core Trust Seal certification (https://www.coretrustseal.org/) for its data repository portal (https://laji.fi/), but may do it later.

1.2. Business model

FinBIF is developed and operated by Finnish Museum of Natural History (LUOMUS, http://www.luomus.fi/en). It is a Finnish national digital and Internet-based infrastructure for data management that compiles, archives and distributes biodiversity data. In addition to managing and sharing biodiversity data, FinBIF offers services for storing biodiversity observations for public use through the facility. It promotes open data and science, and citizen science.

At FinBIF, resources are particularly needed for maintaining the data system and coordinating processes for the flow of biodiversity data. In addition, resources are required for both domestic and international cooperation and partnership management. Development of FinBIF must be safeguarded by partner collaboration or separate organisational activities.

Currently, it is free to use the FinBIF portal and its API (Application Programming Interface). Open-access data are available to registered users free of charge.

2. Infrastructure

FinBIF (Finnish Biodiversity Information Facility) is the Finnish National Infrastructure of Biodiversity Information. It is a primary source of biodiversity data for different purposes, such as research, education, public authority and administration, and citizen science and general public. It compiles Finnish biodiversity information to one single service for open access and sharing by providing a modern digital and Internet based service under constant development. Biodiversity data available through the facility is primarily compiled from different sources, such as partner organisations, research institutes of Finland's environmental and natural resources administration, other government organisations managing and producing biodiversity data, collections maintained by museums of natural history, and citizen science sources. The principal duties of FinBIF are as follows (FinBIF n.d. (d)),

- Compiling species observation data and sharing the data for free use
- Promoting open data and science
- Maintaining the national taxonomic nomenclature
- Strengthening and clarifying processes related to biodiversity data flow
- Data archiving
- Maintaining and updating the data policy related to national biodiversity data
- Promoting citizen science
- Annotating shared data to Settle its reliability and quality

FinBIF maintains a data warehouse that store biodiversity data from the data sources of partner organisations based on primary observations. It includes species names, location data, observation time, and observer names. In addition to observation data, related metadata, such as observation methods and conditions, will be published, to increase the FAIRness of the data. Besides the original biodiversity data, there are also refined lists, statistics, maps and diagrams. FinBIF is capable of managing sensitive biodiversity data with the legal basis. There are quality check of observation data , which is constantly being developed by FinBIF and its partners. All verified errors will be corrected before sharing data and unreliable data will not be published. FinBIF can share biodiversity data worldwide to other data infrastructures. It provides the observation data for international open usage by the cooperation with GBIF (Global Biodiversity Information Facility). The data flow in FinBIF is shown in Figure 3.

The summary of FinBIF infrastructures is shown in Figure 4.

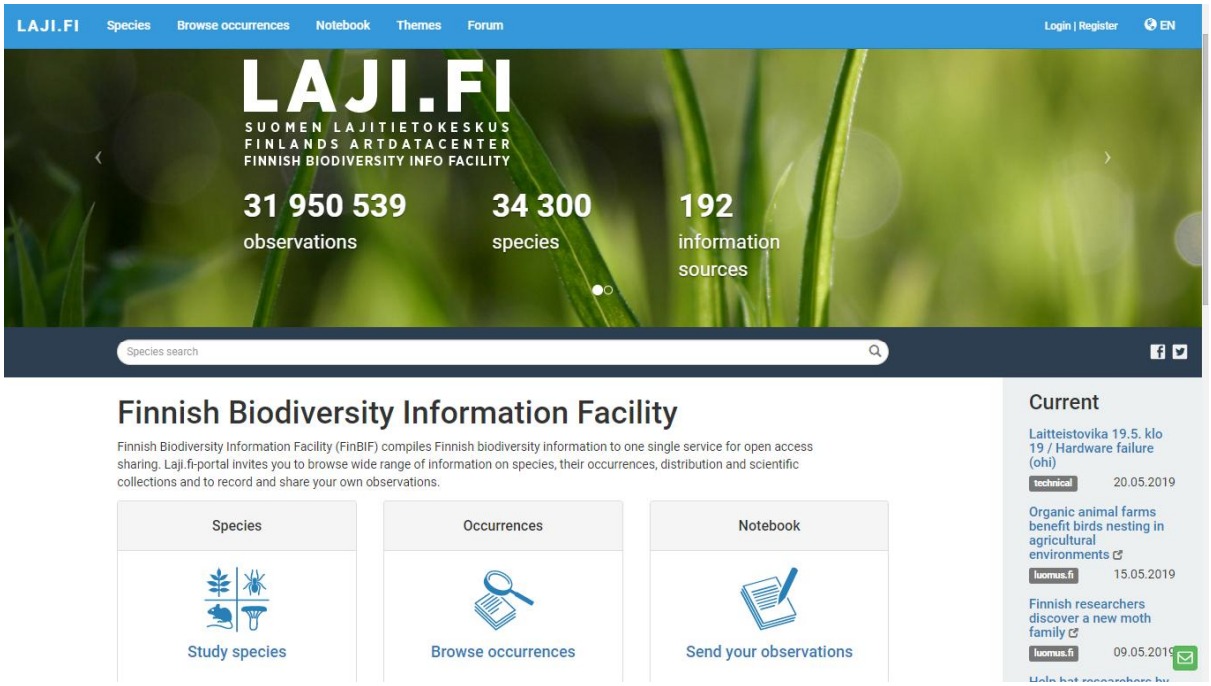


Figure 2. Screenshot of the front page of FinBIF data portal laji.fi.

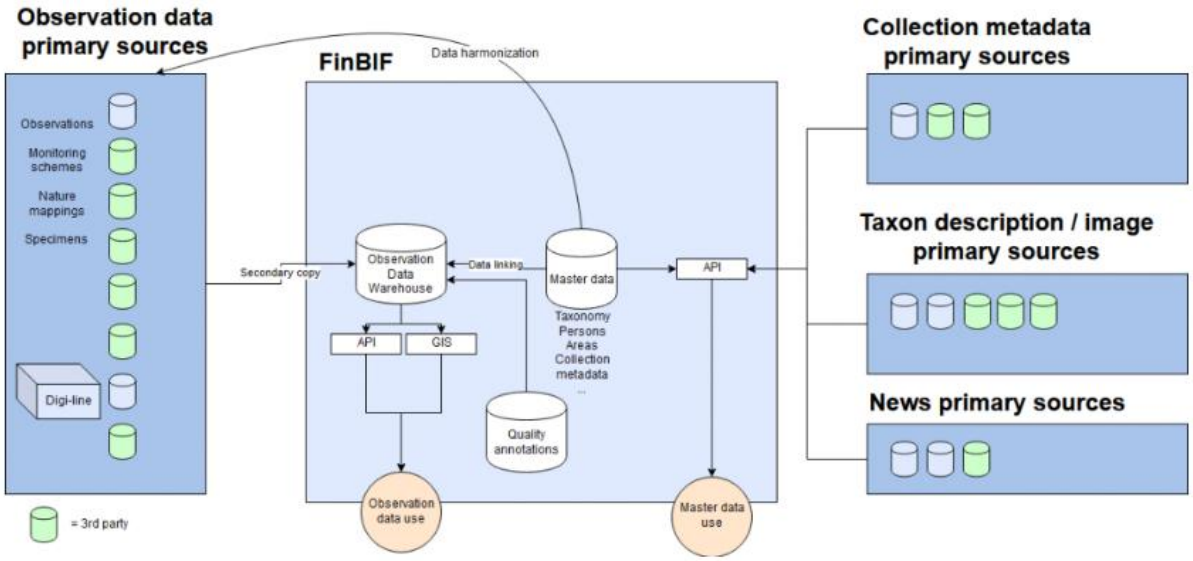


Figure 3. The data flow in FinBIF infrastructures. (FinBIF n.d. (a))

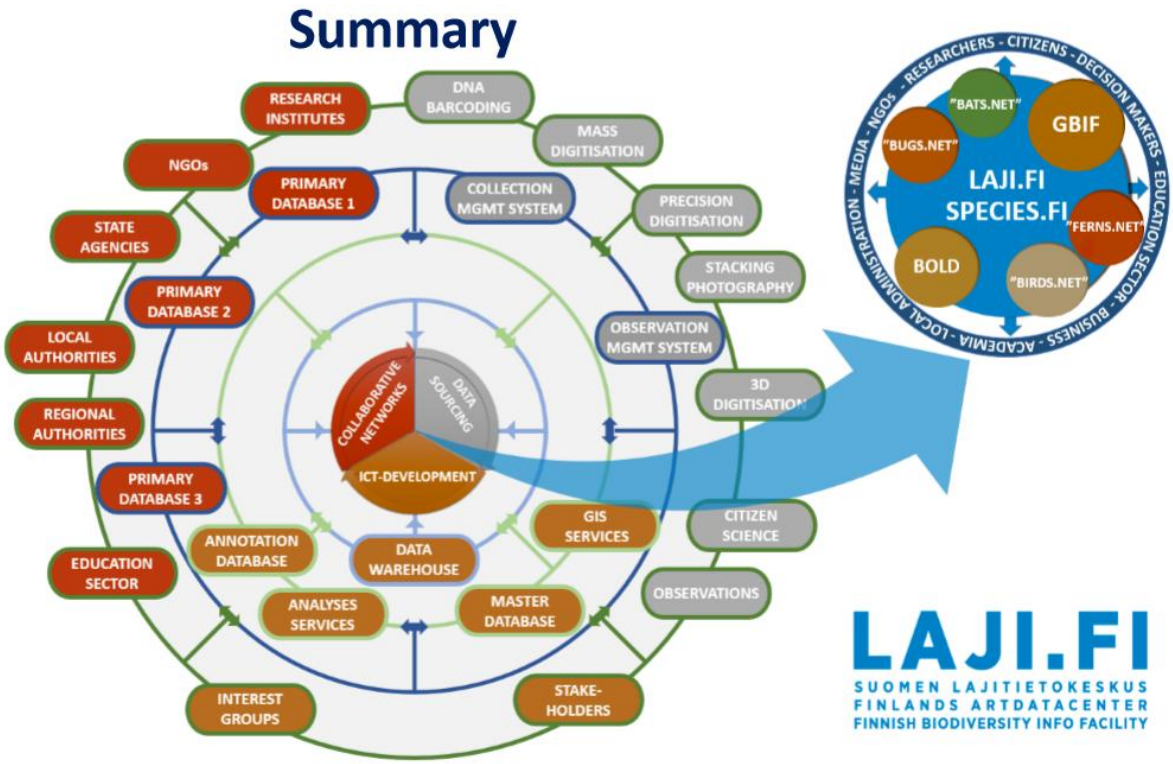


Figure 4. The summary of FinBIF. (Lahti 2017)

FinBIF is developed and operated by Finnish Museum of Natural History (LUOMUS). It is built on both institutional and national IT infrastructures. For national IT infrastructures, it mainly uses CSC services and Fairdata services.

CSC provides Pouta cloud computing services which are IaaS (Infrastructure-as-a-Service) clouds. There are two types of cloud services, cPouta and ePouta. cPouta cloud is the public cloud and ePouta cloud is the virtual private cloud for sensitive data. Both clouds are run with open source OpenStack cloud operating system (https://www.openstack.org/). Users can manage their resources using a web interface through a web browser and also through a set of APIs which allows programmatic management of resources. Using Pouta services, it is quick and easy to deploy multiple pre-defined virtual machines, which are secure and highly scalable. The cPouta virtual machines are connected to the high performance Funet network (Finnish University and Research Network), which is a backbone network providing Internet connections for Finnish higher education institutions and as well as other research facilities. There is no cost for Finnish higher education institutions and national research institutes (subsidized by the Finnish Ministry of Education and Culture) to use Pouta services for academic research and education. For the other usages, it is also available with a fee, starting from 1190 EUR (VAT 0%) for the base package. For more detail, refer to https://research.csc.fi/billing-and-monitoring.

Fairdata services provides general data services for storing, sharing, publishing, and long-term preserving research data. The services are for general and not tied to a specific domain of science. There are four component tools as shown in Figure 4.

Figure 4. Services from Fairdata services (Fairdata.fi n.d. (a))

- IDA (https://ida.fairdata.fi/) is for safe storing research data. Data in IDA can be shared among the project members to benefit the collaboration work across different organizations. There are two data areas in IDA, the staging area and the frozen area. The first one receives the new data uploaded to IDA, where users have both both read and write file permissions. The latter one is read-only for storing data in an immutable state. Only files in the frozen area are visible to other Fairdata services. There is also an option to share data by links temporally to make it open access to others. Both web interface and command line environments are available to IDA users.

- Qvain (https://qvain.fairdata.fi/) is a metadata tool for describing the data. Using Qvain tools, dataset descriptions can be linked to the corresponding data stored in IDA.

- Etsin (https://etsin.fairdata.fi/) is the research data finder. It can publish dataset from Qvain to make it findable.

- Fairdata-PAS (http://www.digitalpreservation.fi/en) is the long-term digital preservation services.

The data flow in Fairdata services is shown in Figure 5.

The services are offered by Finnish Ministry of Education and Culture. The development and operation are by CSC. It is free of charge to Finnish higher education institutions, state research institutes and for those funded by the Academy of Finland. Fairdata-PAS service is based on agreement. The data and information stored in the services is stored in Finland. Finnish national research funders such as the Academy of Finland promotes the use of Fairdata services.

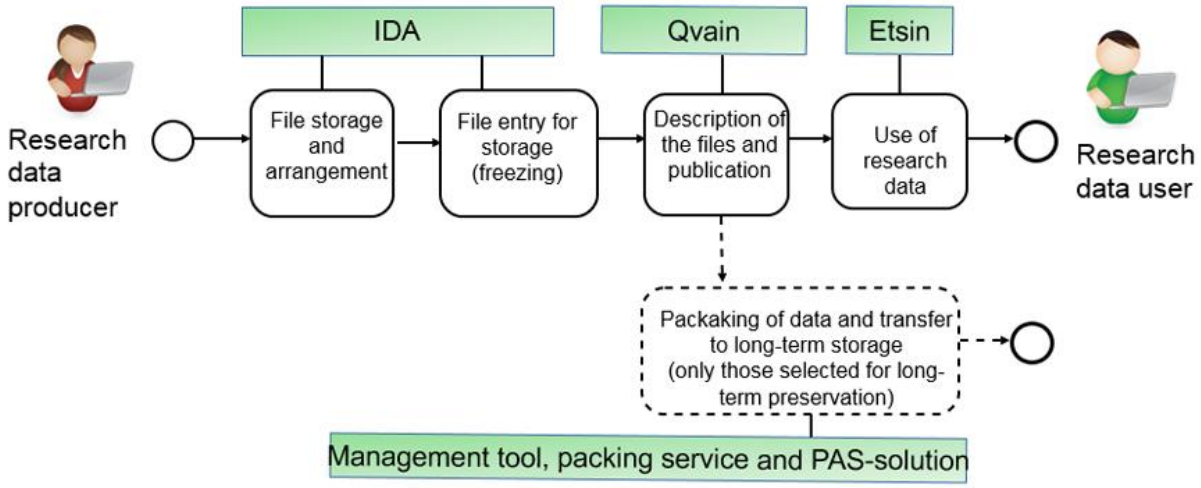


Figure 5. Data flow in Fairdata services (Fairdata.fi n.d. (b))

3. Data model

FinBIF uses a persistent HTTP URI-identifier for all objects, such as occurrences, taxa, and collection metadata, which is recommended by W3C (World Wide Web Consortium) (W3C working group, 2014). The identifiers enables the information of the objects in human- and machine-readable formats. Most objects uses the FinBIF's own defaults namespaces http://tun.fi/. There exists other name spaces, such as http://id.luomus.fi/ for collection specimens from Finnish Museum of Natural History, http://id.zmuo.oulu.fi/ and http://mus.utu.fi for collections from University of Oulu and University of Turku respectively. Instead of a full URI-identifier, it is possible to use a shorter QName, since it can be challenging to use full URI-identifiers in REST-API. (FinBIF n.d. (e))

Table 1. List of namespaces and their Qname prefixes known by FinBIF (FinBIF n.d. (e))

```
rdf         http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs        http://www.w3.org/2000/01/rdf-schema#
xsd         http://www.w3.org/2001/XMLSchema#
xml         http://www.w3.org/XML/1998/namespace#
owl         http://www.w3.org/2002/07/owl#
vcard       http://www.w3.org/2006/vcard/ns#
dc          http://purl.org/dc/terms/
dwc         http://rs.tdwg.org/dwc/terms/
dwctype     http://rs.tdwg.org/dwc/dwctype/
so          http://schema.org/
cc          https://creativecommons.org/
```

```
skos        http://www.w3.org/2004/02/skos/core#
herbo       http://id.herb.oulu.fi/
luomus      http://id.luomus.fi/
utu         http://mus.utu.fi/
zmuo        http://id.zmuo.oulu.fi/
```

In Table 1, there is a list of namespaces and corresponding QName prefixes in FinBIF. The QName prefix for the default namespace http://tun.fi/ is an empty string “”. So the request https://api.laji.fi/collections/http%3A%2F%2Ftun.fi%2FMX.1 can be made more reliably using a short QName: https://api.laji.fi/collections/MX.1. For the URI-identifier based on non-default name spaces such as http://mus.utu.fi/TBR.119974, its Qname is utu:TBR.119974. Similarly, the QName of the term http://rs.tdwg.org/dwc/terms/vernacularName is dwc:vernacularName. The schema classes in FinBIF are listed in Appendix A1.

The metadata of specimens is loosely based on ABCD standard, but uses many proprietary fields. It is relatively simple to map the fields to Darwin Core. There are only few fields are mandatory and can be extended as needed. Also it is hierarchical, allowing multiple organisms, identifications and type specimen information per one specimen. For example, a herbarium sheet can have multiple plants, each with multiple identifications. Most data is denormalized. For example, any kind of species name can be used. Moreover, both specimens and observations can use it. (FinBIF n.d. (b))

Besides the data portal (https://laji.fi/), FinBIF provides robust and easy-to-use API (Application programming interface) for the public to access its open data. To use the API, an access token is required, which is easy to acquire with an email address. There are mainly six groups of endpoints as listed below. (FinBIF n.d. (c))

- Observations and collections
    - Warehouse - Observation Data Warehouse API
    - Collection - Collection metadata
    - Source - Information sources (IT systems)
    - Annotation - Quality control
- Taxonomy
    - Taxa - Taxonomy API
    - InformalTaxonGroup - Informal taxon groups are used in taxa and warehouse endpoints
    - Publication - Scientific publications
    - Checklist - Mainly you only work with one checklits: the FinBIF master checklist. There are others.
- Other master data
    - Metadata - Variable descriptions
    - Area - Countries, municipalities and biogeographical provinces of Finland, etc.

- Person - Information about people.

- Helpers
  - APIUser - Register as an API user
  - Autocomplete - For making an autocomplete filed for taxa, collections or persons (friends)
  - PersonToken - Information about an authorized person
- Vihko observation system
  - Form - Form definition
  - Document - Document instance of a form
  - Image - Image of a document
- Laji.fi portal
  - Feedback - Feedback form API
  - Information - CMS content of information pages
  - Logger - Error logging from user's browsers to FinBIF
  - News - News

The full list of endpoints is shown in Appendix A2.

4. Design

Based on the FinBIF infrastructures, the data from digitisation facilities will be stored, published, and shared. Figure 6 shows the data flow. CSC cPouta cloud computing services and Fairdata sevices are used to provide robust and reliable computing and data storage. cPouta services is used to process the imaging data uploaded from the digitisation systems. Fiardata IDA is used to safe store the data. Fairdate-PAS is planned for the long-term preservation of the data.

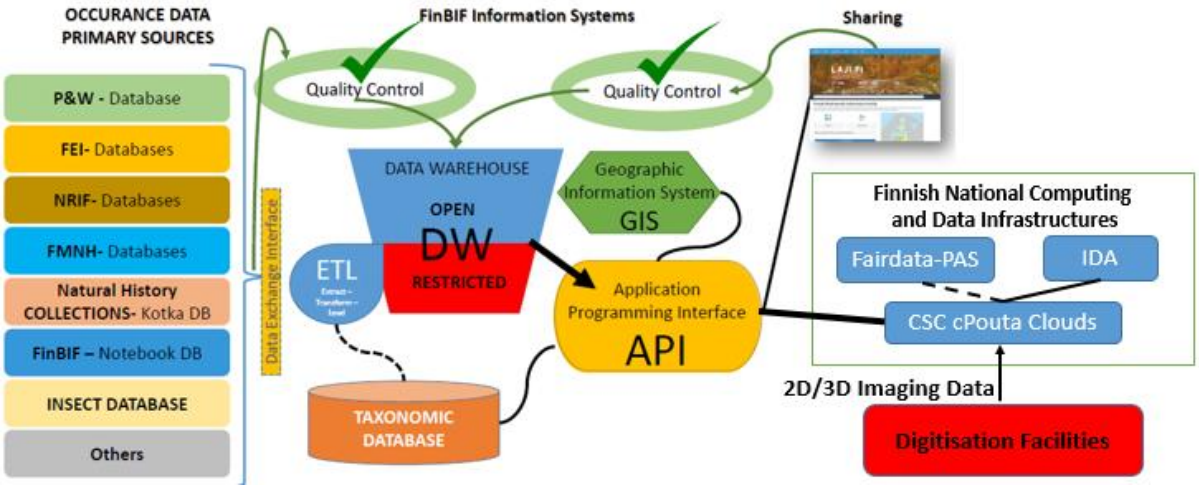


Figure 6. The data flow of the imaging data from the Digitisation facilities to FinBIF infrastructures. (Modified from (Lahti 2017))

The digitisation facilities at LUOMUS, have two digitisation systems in operation, one is for herbarium sheets and the other is for small pinned insect (Tegelberg et al. 2014, Tegelberg et al. 2017). The data flow of the mass digitisation at LUOMUS is shown in Figure 7. As introduced in chapter 1, digital specimens, the output from the digitisation process, have three parts, PID, data, and metadata. In FinBIF, HTTP URI-identifier is used for digital specimens from the LUOMUS digitisation systems. If the specimen is already existed in Kotka (LUOMUS collection manager system), the original ID from Kotka will be used for it in the following data flow. If not, the new ID will be given and the metadata will be entered in Kotka. For example, http://id.luomus.fi/C.56686 is the ID for a digital specimen from the herbarium digitisation system, and http://id.luomus.fi/F.56686 is the ID for that from the pinned insect digitisation system. The data of the digital specimen is the imaging data, such as 2D images and 3D data, which is captured from the imaging devices. Imaging data can be uploaded via SFTP to the virtual machines run at cPouta clouds in real-time after imaging or later in batch. cPouta clouds will process the data for the post-processing and then transfer it to the IDA data storage with IDA command line tools. Kotka and FinBIF can use the image API hosted on the cPouta clouds to access the specimen imaging data to show them instantly in the relevant data portals. For example, Figure 8(b) and Figure 9(b) are the screenshots of digital specimen http://id.luomus.fi/C.56686 and http://id.luomus.fi/F.56686 shown in FinBIF data portal http://www.laji.fi/ respectively. Figure 8(a) and Figure 9(a) are the images of the two specimens respectively. Figure 9(c) shows that the location information of the specimen http://id.luomus.fi/F.56686 is given in both coordinate values and also on the map at http://www.laji.fi/. Moreover, after the release of Fairdata-PAS from CSC, it is planned to use it for the long-term data preservation of digital specimens. In addition, there are HPC (high-performance computing) resources available at CSC, which enable the possibility to process and analyse the data at a large scale.

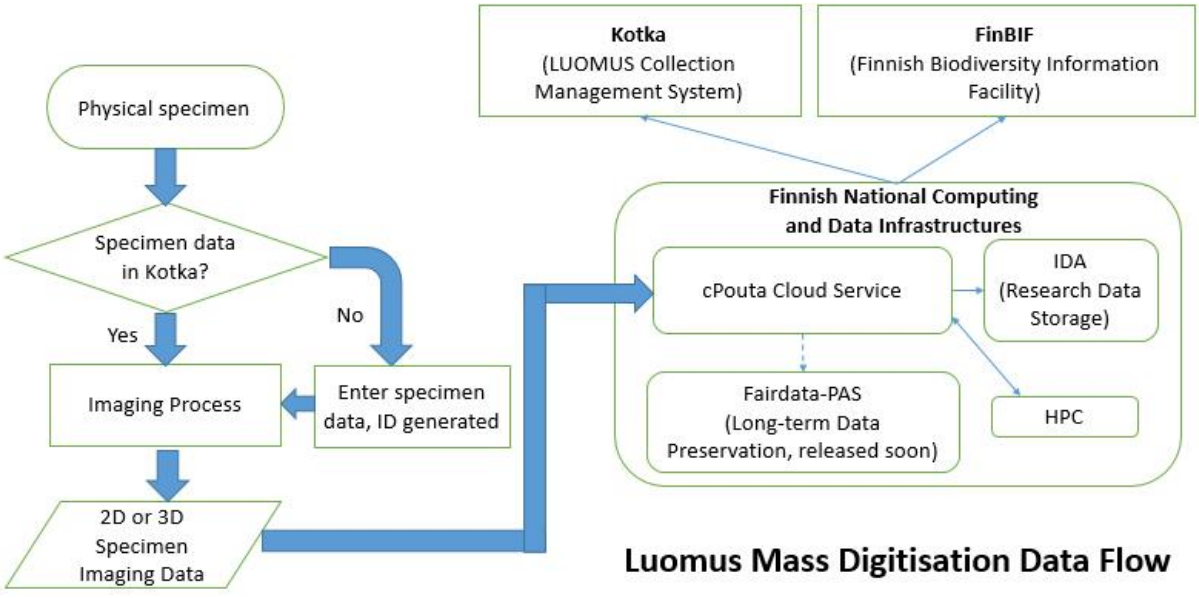
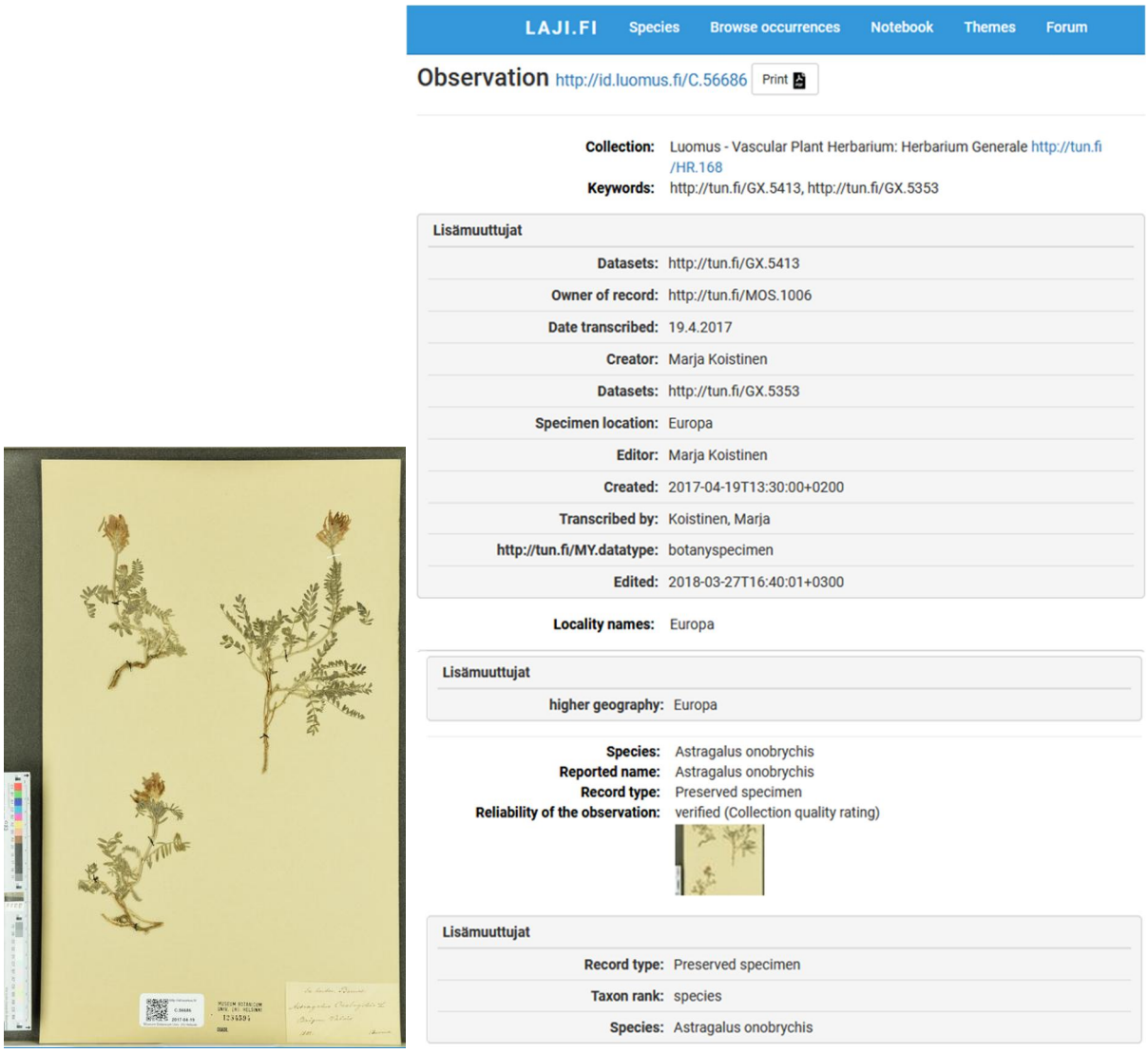


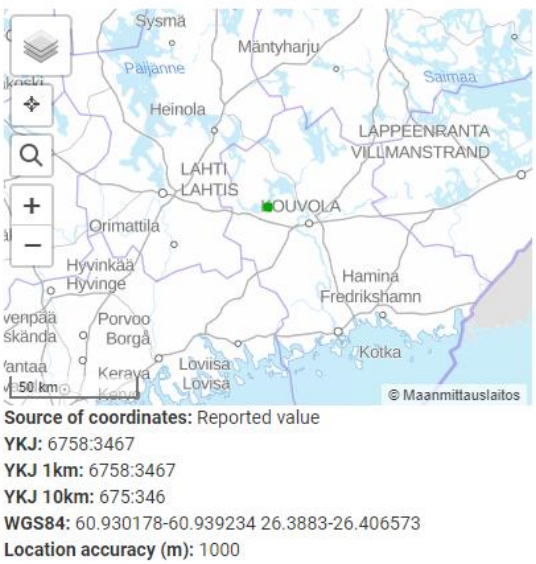
Figure 7. The data flow at LUOMUS mass digitisation facilities.



(a) (b)

Figure 8. (a)Image and (b) laji.fi screenshot of the digital specimen http://id.luomus.fi/C.56686.

(a)



Source of coordinates: Reported value
YKJ: 6758:3467
YKJ 1km: 6758:3467
YKJ 10km: 675:346
WGS84: 60.930178-60.939234 26.3883-26.406573
Location accuracy (m): 1000

(c)

LAJI.FI Species Browse occurrences Notebook Themes Forum

Observation http://id.luomus.fi/F.56686 Print

Collection:	Luomus - Lepidoptera Eastern Fennoscandia (Luomus) (LEPFEN) http://tun.fi/HR.527
Keywords:	http://tun.fi/GX.450, http://tun.fi/GX.750

Lisämuuttujat

Edited:	2016-10-28T11:58:58+0300
Creator:	Jere Kahanpää
Editor:	Jere Kahanpää
Datasets:	http://tun.fi/GX.450
Datasets:	http://tun.fi/GX.750
Status:	Ok
Owner of record:	http://tun.fi/MOS.1007
Created:	2016-10-28T11:58:58+0300
http://tun.fi/MY.datatype:	zoospecimen

Observer:	Jalas, Ilkka
Time:	1971-06-15
Locality names:	Finland, South Häme (EH), Iitti, kirkonkylä

Lisämuuttujat

Lat (N):	6758
Coord source:	Unknown
Municipality:	Iitti
WGS84 Latitude:	60.934708
WGS84 Longitude:	26.397438
Biogeographical province:	EH
Coordinate system:	Uniform grid (YKJ)
Start date:	15.6.1971
Leg:	Jalas, Ilkka
Locality names:	kirkonkylä
Country:	Finland
Lon (E):	3467

Species:	jänöyökkönen (fi) – *Acronicta leporina*
Reported name:	Acronicta leporina
Record type:	Preserved specimen
Reliability of the observation:	verified (Collection quality rating)
Count:	1
Life stage:	adult

Lisämuuttujat

Record type:	Preserved specimen
Life stage:	adult
Count:	1
Taxon rank:	species
Species:	Acronicta leporina
Taxon author:	(Linnaeus, 1758)

(b)

Figure 9. (a)Image, (b) laji.fi screenshot, and (c) geo information of the digital specimen http://id.luomus.fi/C.56686.

Table 2. Data Volume from LUOMUS digitisation facilities at FinBIF

Herbarium Digitisation	Small Pinned Insect Digitisation
· c.a. 435,000 digital specimens · One specimen with one image · 50-60 MB for one TIF image · 2-3 MB for one JPG image · 24TB in total · 13,600 specimen / month in average · 720GB data / month in average	· c.a. 276,000 Specimens · One specimen with two images · 40-50 MB for one TIF image · 1-3 MB for one JPG image · 22.5TB in total · 300-400 specimens / day · 25-35 GB data / day

The digitisation facilities at LUOMUS have been operating for several years, the herbarium one from November 2016 and the pinned insect one from December 2015. Both of them are designed for mass digitation with the conveyor belt driven imaging system. They can capture up to 600 very high resolution images (in TIF and JPG formats) per hour. For herbarium system, it capture one image per specimen with the file size of around 50-60 MB for the TIF image and 2-3MB for the JPG image. For pinned-insect system, there are two or more images per specimen, with around 40-50MB for one TIF image and 1-3MB for one JPG image. In total, LUOMUS digitisation facilities contribute more than 0.7 million digital specimens to FinBIF. The detailed information of the data volume from LUOMUS digitisation systems at FinBIF is described in Table 2.

The data transfer in the data flow shown in Figure 7 is very reliable with a very low error rate of the data transfer. For the data upload from the imaging station at the digitisation facilities to the remote cPouta virtual machines hosted at CSC, the transfer error rate is less than 0.1%. This low error rate may be due to the robust computing and data infrastructures, and also the reliable network services of Funet network (Finnish University and Research Network) that both digitisation facilities and cPouta clouds are using. Usually the upload is carried out in the working time or early evening in the weekday. The local network bandwidth at digitisation facilities is 100 MB/s and the network may be shared with other users.

For the data flow from the cPouta virtual machines to the IDA data storage, the associated files of each specimen are packaged into a zip file before the transfer, which means that each digital specimen has a zip file for the data transfer. After the transfer finished, the data is downloaded from IDA and then unzipped to compare the checksum values of each file in order to guarantee the file integrities. There is no transfer error detected for the around 11.5 TB data transferred to IDA from cPouta clouds from last 10 months after the release of new IDA services. Both cPouta cloud computing service and IDA data storage services are hosted by CSC.

5. Discussion

From the case study of Finland national infrastructures in this work, it shows the high potentials and the important role of national computing and data infrastructures in the digitisation of natural history collections. FinBIF (Finnish Biodiversity Information Facility) is a good example of such infrastructures that is dedicated to biodiversity data with open access free of charge. It supports the digitisation work by providing the services of storing, accessing, analysing, annotating, sharing, and preserving the digital specimens generated from the digitisation facilities. It is built on the services from other Finnish computing and data infrastructures, such CSC cPouta clouds and Fairdata services.

To use national level computing and data infrastructures for the data from digitisation facilities, the data will be highly probably hosted in the same countries. This is quite beneficial and even critical in some cases. Firstly, for some data, it may not be allowed to be stored in second countries. In this case, it is mandatory to use its own national services. Secondly, use the national services may have a faster and more reliable network connections to transfer data when compared to use services from second countries. Thirdly, the national level infrastructures may provide services free of charge to the users from the same countries, and there may be more training opportunities on using the services and more instant support.

When choosing the data infrastructures and designing the data flow from the digitisation facilities to the national level data infrastructures and further to European level and then worldwide level, the three key components, PID, data, and metadata, of the digital specimen generated from the digitisation facilities have to be taken into considerations.

- Firstly, most data infrastructures, including those at national level, are not domain specific. Usually those general data services cannot fit the cases of domain specific usages, such for the biodiversity data digital specimens. For example, the general data repositories may not have all the key terms of the metadata, which will make the data to be difficult to access, find, and re-use.

- Secondly, metadata may not be transcribed or only partially transcribed from the specimen at the digitisation facilities, especially in the mass digitisation process. This requires the data repositories only need minimum mandatory metadata terms, and also provide data versioning and annotation functions for the later transcription work to complete the metadata.

- Thirdly, there may be limitations on the supported data format, prohibiting the uploading of unsupported data.

- Moreover, on the data portals, the visualization of the data may be missing, such as 3D data. Similarly, in the metadata of digital specimens, there are usually location information. To visualize it with map services may not be available in the general data portals.

- In addition, the data flow from the general national data repositories to higher level like EU may not be well defined.

- Also, there should be robust and easy-to-use API (Application programming interface) available to access and manage the data. Furthermore, the management of sensitive data may be available.

Therefore, the data infrastructures for biodiversity data at the national level are needed, so that tailored services for digitisation facilities can be provided. The efforts of establishing the tailored services can be alleviated by using existing services from other national IT infrastructures to do integration and further development.

In the case of FinBIF, HTTP URI-identifiers are used for digital specimens from the LUOMUS digitisation facilities. It provides the support for a quite wide range of biodiversity metadata terms. There are only few mandatory terms, which is useful for the mass digitisation and enable the data available instantly to facilitate the open access and the later transcription process. However, sensitive data has strict accesses and not open to the public. In the data portal of FinBIF, the imaging data can be visualized and map services are available for the displaying the location. In addition, besides its online data portal, FinBIF provide API for the public to access its open data free of charge. Moreover, FinBIF can be as a gateway to distribute and share biodiversity data to other data infrastructures globally. For example, FinBIF works in cooperation with GBIF (Global Biodiversity Information Facility) to provide all observation data for global usage.

6. References

Ariño AH (2010) Approaches to estimating the universe of natural history collections data. BiodiversityInformatics 7: 81-92. https://doi.org/10.17161/bi.v7i2.3991

Blagoderov V, Smith V (2012) No specimen left behind: mass digitization of natural history collections. ZooKeys 209. https://zookeys.pensoft.net/browse_journal_issue_documents.php?issue_id=361

CSC (n.d.) CSC Cloud computing services. Retrieved from https://research.csc.fi/cloud-computing

Fairdata.fi (n.d.) Fairdata services. Retrieved from https://www.fairdata.fi/en/

Fairdata.fi (n.d.) Fairdata services components. Retrieved from https://www.fairdata.fi/en/services/components/

FinBIF (n.d.) FinBIF Architecture overview. Retrieved from https://laji.fi/about/3135

FinBIF (n.d.) Collection Management System Kotka. Retrieved from https://laji.fi/about/716

FinBIF (n.d.) FinBIF API. Retrieved from https://laji.fi/about/3120

FinBIF (n.d.) FinBIF Mission. Retrieved from https://laji.fi/about/2954

FinBIF (n.d.) FinBIF persistent globally unique identifiers. Retrieved from https://laji.fi/about/3118

Hardisty A, Addink W, Raes N (2019) MS37 High-level specifications of functionalities and data flow. ICEDIG Project working paper

Lahti K (2017) FinBIF – Finnish Biodiversity Information Facility. Retrieved from http://koivu.luomus.fi/gbif/02_Lahti_FinBIF-Suomen_Lajitietokeskus.pdf

Oever JP, Gofferje M (2014) From pilot to production: Large scale digitization project at Naturalis Biodiversity center. ZooKeys 209, 87-92. https://zookeys.pensoft.net/article/2921/

RDA NDS Interest Group (2018) Country Reports: National Data Services Summary Report Ver 1.1. https://docs.google.com/document/d/17iUyJ2icY0gFzMZGPWJyY5E0tUoukAtI4BFeRonefv4

Tegelberg R, Mononen T, Saarenmaa H (2014) High-performance digitization of natural history collections: Automated imaging lines for herbarium and insect specimens. Taxon 63 (6) 1307-1313. http://www.ingentaconnect.com/content/iapt/tax/2014/00000063/00000006/art00010

Tegelberg R, Kahanpää J, Karppinen J, Mononen T, Wu Z, Saarenmaa H 2017. Mass digitization of individual pinned insects using conveyor-driven imaging. In: Hereld M (Editor) High throughput digitization for natural history collections. 2017 IEEE 13th International Conference on e-Science (e-Science 2017). Auckland, New Zealand, 24-27 October 2017. 5 p. https://ieeexplore.ieee.org/document/8109190/

W3C Working Group (2014) Best Practices for Publishing Linked Data. Retrieved from https://www.w3.org/TR/ld-bp/

Wu Z, Kahanpää J, Sihvonen P, Koivunen A, Saarenmaa H (2019). Automated Methods in Digitisation of Pinned Insects. Biodiversity Information Science and Standards, 3, e38260. https://doi.org/10.3897/biss.3.38260

7. Appendix

A1. Schema Classes in FinBIF

The full URI of each class is defined by a prefix http://tun.fi/ followed by the class name. For example, the URI of the MY.document is http://tun.fi/MY.document. For more detail, refer to http://schema.laji.fi/class.

NS	Class name	Description
MX	administrativeStatus	Administrative status
MY	agent	

NS	Class name	Description
MAN	annotation	
ML	area	Area
ML	areaClass	
MM	audio	Audio recording
PUU	branch	Branch
MR	checklist	Checklist
MR	checklistVersion	Checklist version
MY	collection	Collection
MX	contentContextDescription	Instances of this describe one taxon description context
GX	dataset	Dataset
MNP	dateRange	
	dc:BibliographicResource	Bibliographic Resource
MXC	device	
MXC	deviceIndividual	
MY	document	Submissions
HBF	downloadRequest	Download request
MKV	endangermentObject	Endagerment object

NS	Class name	Description
PUU	event	
MHL	field	
MHL	fieldset	
MHL	form	
MFP	formPermission	
MFP	formPermissionPerson	
MFP	formPermissionSingle	
MY	gathering	Gathering event
MZ	gatheringEvent	Global gathering event
MY	gatheringFactClass	Keruutapahtuman faktat
MZ	geometry	
MKV	habitatObject	IUCN Red List Evaluation Habitat
MY	identification	Identification
MM	image	Image
MXA	individual	
MVL	informalTaxonGroup	Informal Taxon Group
KE	informationSystem	Information System

NS	Class name	Description
MKV	iucnRedListEvaluation	IUCN Red List Evaluation
MKV	iucnRedListEvaluationYear	IUCN Evaluation Year
MVL	iucnRedListTaxonGroup	IUCN Red List Evaluation Informal Taxon Group
MZ	keyAny	
MZ	keyValue	
MY	measurementClass	
MM	multimediaObject	
MNP	namedPlace	
MPO	news	News
MHN	notification	
MO	occurrence	Occurrence
MOS	organization	Organization
MZ	pagedResult	
HRA	permitClass	
MA	person	Person
LA	Pinkka	Instances of this describe one taxon description context (Pinkka eLearning enviroment)
MF	preparationClass	

NS	Class name	Description
MA	profile	
MP	publication	Publication
MKV	regionalStatus	Instances of this class tell endangerment of a certain area and notes about the area
MNP	reserveClass	
MF	sample	Specimen sample
HRB	specimen	TransactionSpecimen
MY	subUnit	
MMAN	tagClass	
MX	taxon	Taxon
MY	taxonCensusClass	Completeness of census
MC	taxonConcept	Taxon concept
MKV	taxonGroupIucnEditors	IUCN Editors
MI	taxonInteraction	Taxon interaction
HRA	transaction	Transaction
MHL	translation	
MY	typeSpecimen	Identification
MY	unit	Specimen

NS	Class name	Description
MY	unitFactClass	Näytteen / havainnon faktat
MZ	unitGathering	Havainnon keruutiedot
MHL	validator	
MM	video	Video

A2. Endpoints of FinBIF API

For more detail, refer https://api.laji.fi/explorer.