Lecture 8
creating, distributing and obtain specimen data

PBIO 161 Biological collections

Jere Kahanpää & Kari Lahti (Luomus)

Basic plan of the lecture

- Creating digital data from biological collections (*Jere*)
  - Defining digitization & digitalization + why do we do it?
  - Digitization tools used in biological collections
- Distributing data (*Kari*)
  - Finnish Biodiversity Information Facility
  - Legal issues: licenses, restrictions (Nagoya Protocol etc)
- Finding & acquiring collection specimen data (*Jere*)
  - Some data sources in more detail
  - Data formats (Darwin core, FASTA, yms./tms.)
  - Caveats concerning specimen data

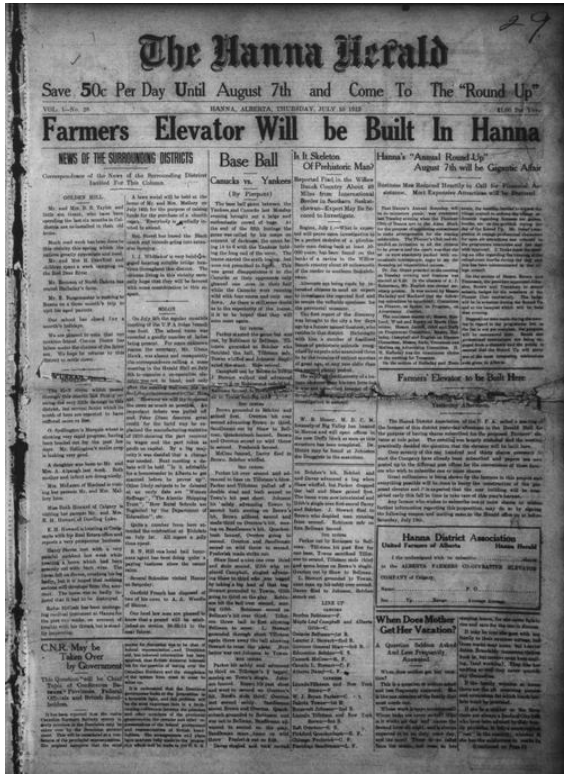
DDDDefinitions

- Data
  - ”Data is a set of values of qualitative or quantitative variables.” [Wikipedia]
  - “Definition of data
    - 1 : factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation
    - 2 : information in digital form that can be transmitted or processed
    - 3 : information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful”

      [merriam-webster.com/dictionary]

DDDDefinitions

- Digital data
  - In practice data converted to a binary (compter-accessible) format
  - May or may not be in a format easily processed by a computer
- Digitization
  - Converting into digital data
  - Often, but not always, means imaging (or otherwise recording) a specimen in our context
- Digitalization
  - Increasing digitization & use of digital data


The Hanna Herald

Save 50c Per Day Until August 7th and Come To The "Round Up"

Farmers Elevator Will be Built In Hanna

NEWS OF THE SURROUNDING DISTRICTS

Base Ball

Is It Skeleton Of Prehistoric Man?

Hanna's "Annual Round-Up" August 7th will be Gigantic Affair

Farmers' Elevator to be Built Here

Hanna District Association United Farmers of Alberta

When Does Mother Get Her Vacation?

C.N.R. May be Taken Over by Government

Why do we digitize natural history collections?

- Distribution/access to data!
- Ditto for metadata
- Backup/security
- Analysis
- (PR etc.)

(By) Felipe Milanez CC-BY-SA 4.0

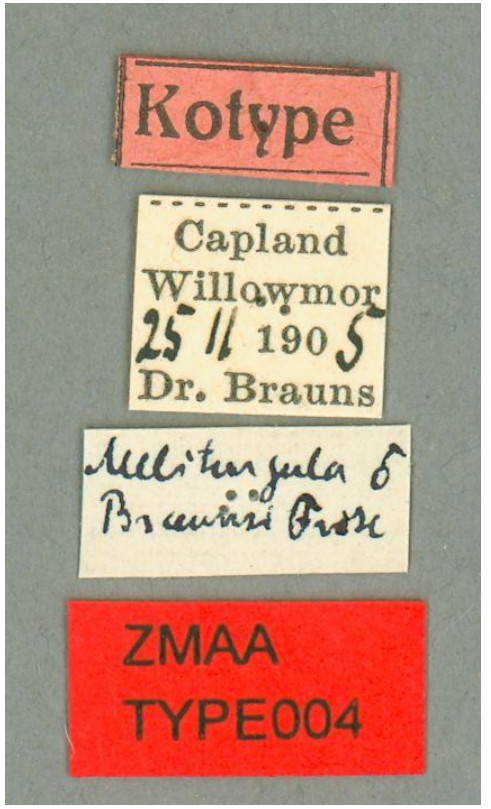
Items digitized in natural history collections

- Physical specimens (imaging, source, history etc.)
- Chemical data (esp. DNA sequences)
- Collection details
- Loans and donations
- Metadata
- Literature
- Field notes, manuscripts, other papers
- Experimental data (lab & field)
- (Observations)

Physical specimens

http://mus.utu.fi/ZMAA.TYPE004

cotype *Meliturgula braunsi* Friese, 190



(By) Pekka Malinen, Luomus

Collection details & Metadata

http://tun.fi/HR.2189

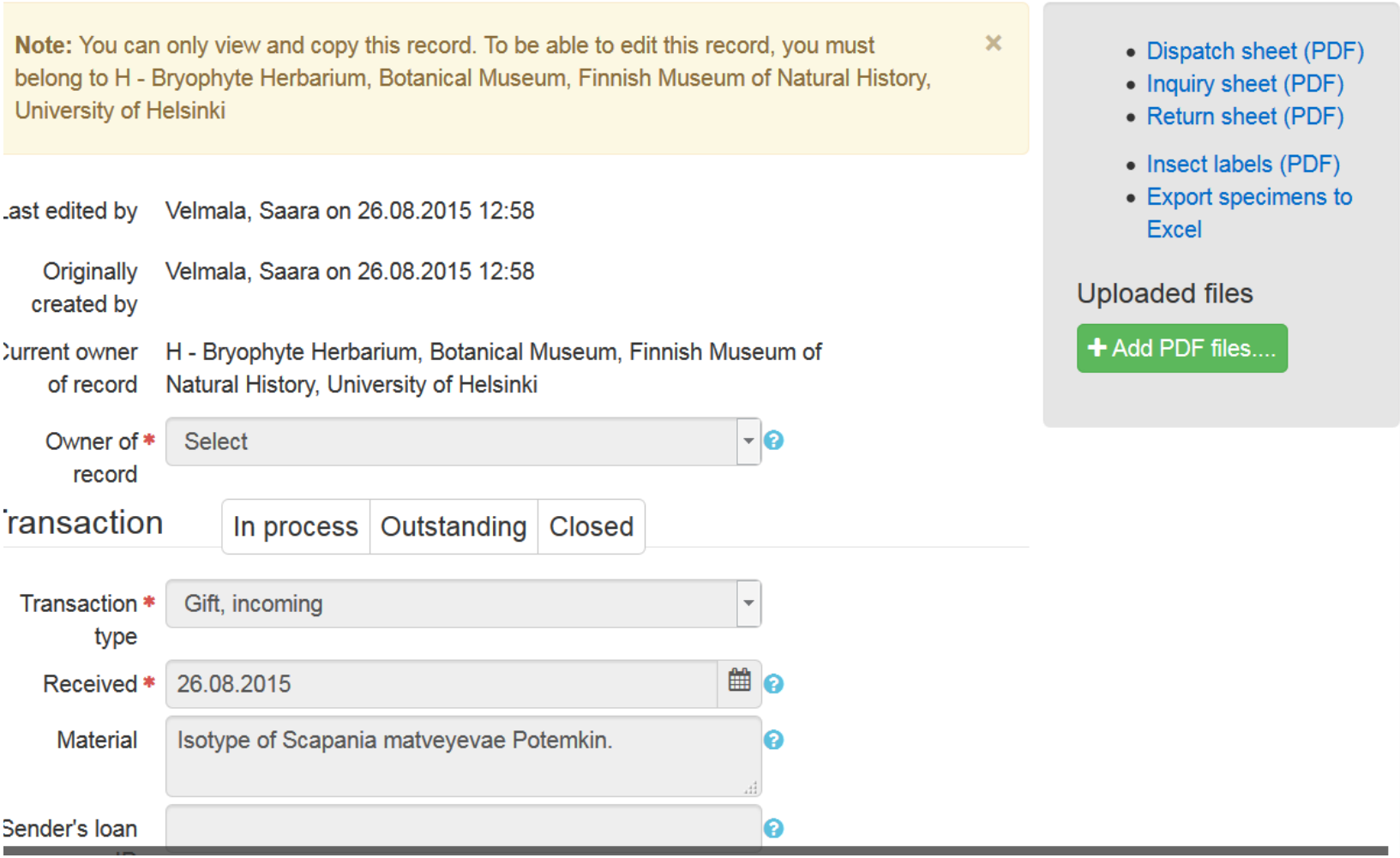
http://tun.fi/HR.2189

Accessions in this location Specimens in this collection

Name:	KUO Lepidoptera collections (world)
Name (fi):	KUO perhoskokoelmat (maailma)
Type:	Specimen collection
Description:	Scientifically arranged Lepidoptera collection of the world
Description (fi):	Tieteellisesti järjestetty maailman perhosten kokoelma
Taxonomic coverage:	Lepidoptera
Geographic coverage:	World excluding NW-Europe
Temporal coverage:	1900-
License for use:	Creative Commons Zero
Size (approx.):	100000
Is part of:	Zoological Collections of the Kuopio natural history museum
Person responsible:	Kettunen, Jukka
Contact email:	jukka.olavi.kettunen@kuopio.fi
Data quality:	4 star
Secure level:	MX.secureLevelNone

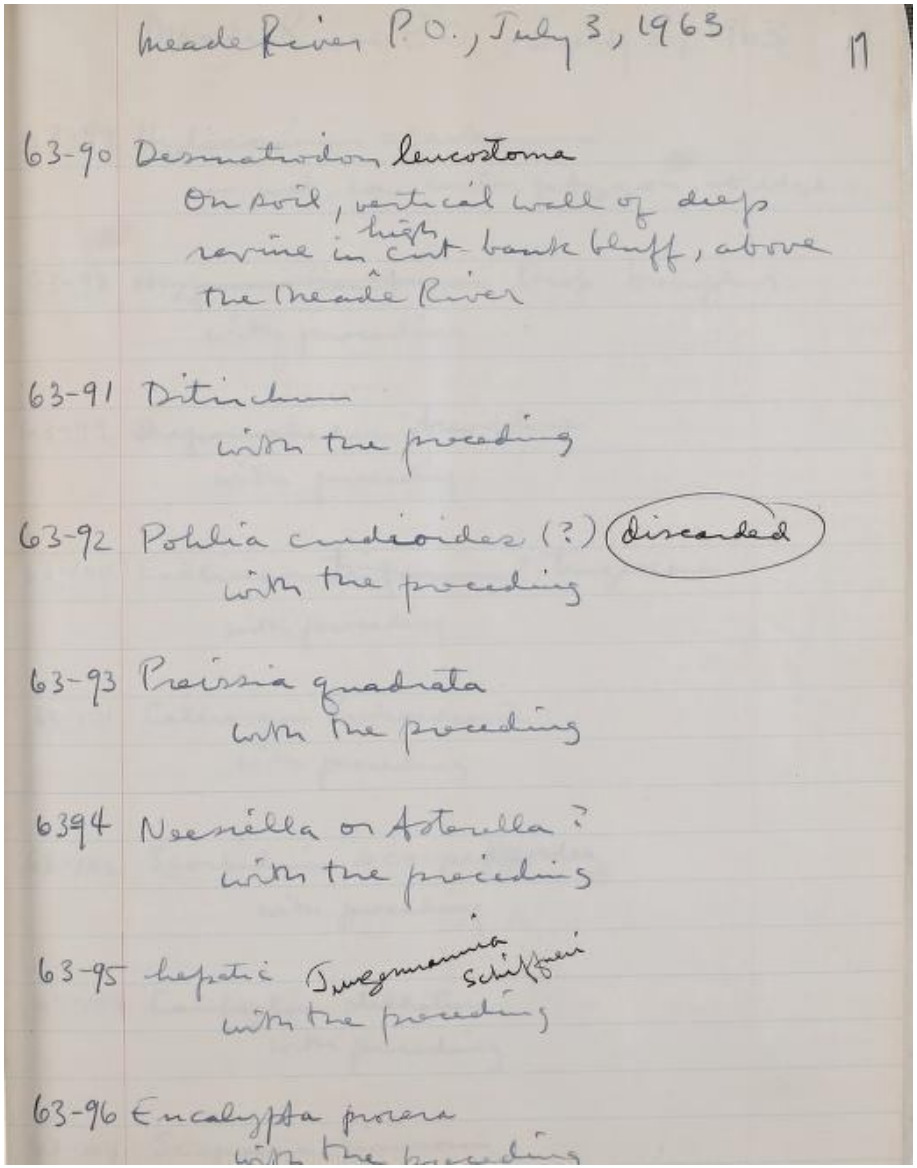
Loans and donations

- **transaction**
  http://tun.fi/HRA.3580



Field notes, manuscripts other papers

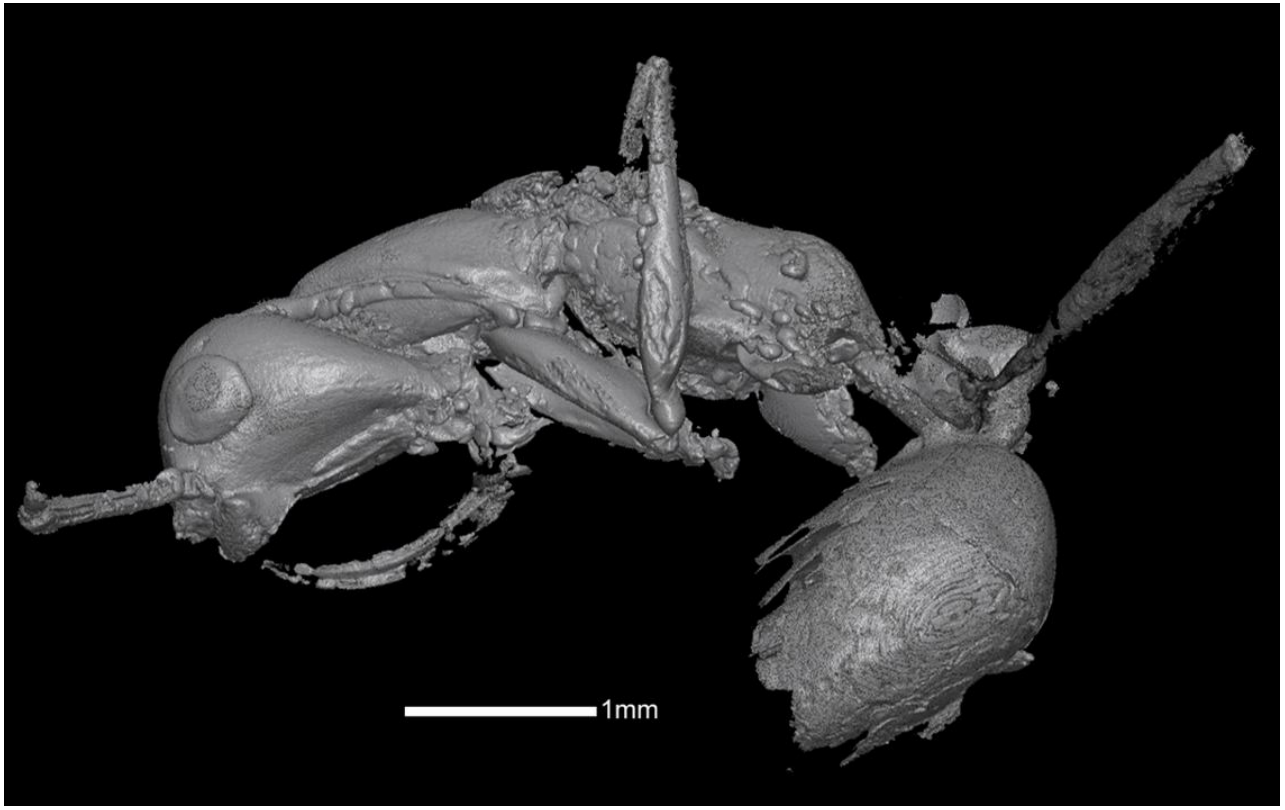
- https://www.biodiversitylibrary.org/browse/collection/FieldNotesProject



Experimental data

- Sonograms
- Chromosome slide preparations

Methods and tools



Profile computed tomographic scan of the *Haidomyrmex scimitarus* holotype. specimen AMNH-BUFB80 by Phillip Barden, Creative Commons Attribution 4.0 International

Imaging: 2D classic

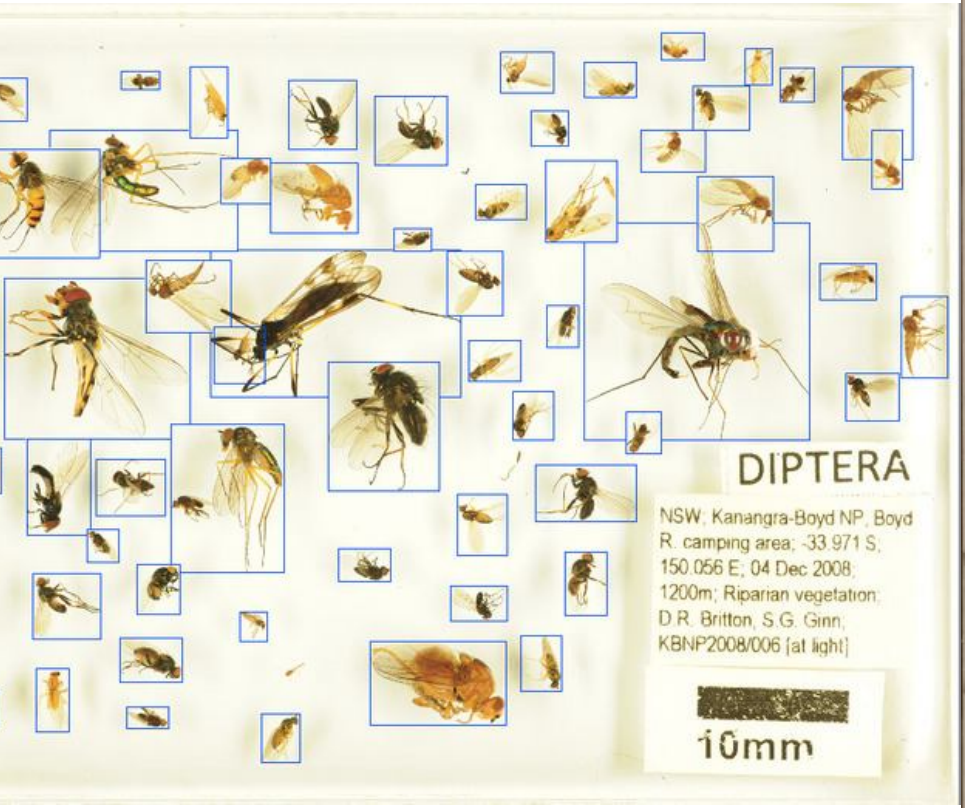
- Fast
- Cheap
- Can handle many types of material
- Limits on quality

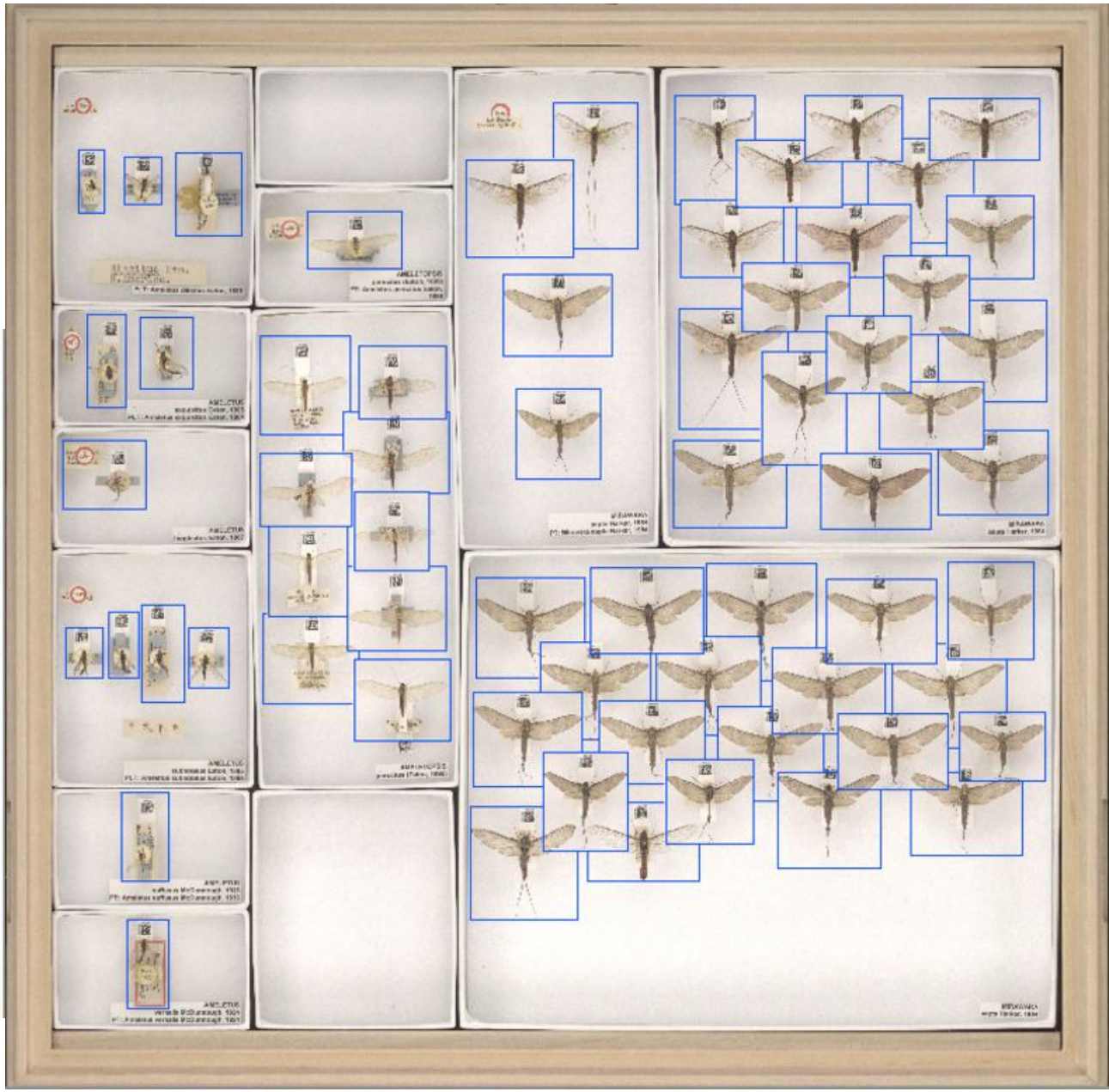
http://plants.jstor.org/stable/10.5555/al.ap.visual.ma-ajb04-d-0812

Imaging: 2D classic

- Coll. Seppo Karhula, recently aquired & documented @ Luomus

Inselect


DIPTERA
NSW; Kanangra-Boyd NP, Boyd R. camping area; -33.971 S; 150.056 E; 04 Dec 2008; 1200m; Riparian vegetation; D.R. Britton, S.G. Ginn; KBNP2008/006 [at light]
10mm



Imaging: 2D stacking

Imaging: 3D surface

http://www.zoosphere.net/sequence/177/Apiophora/paulseni

Sequences | The Project | Team | Contact | News

ZOOSPHERE

Apiophora paulseni

museum für naturkunde berlin

Museum für Naturkunde Berlin
http://www.mfn-berlin.de

Collection: Diptera
Responsible:
Email:

Vertical Perspectives: 10
Horizontal Perspectives: 10
Number of Images: 100

Send Feedback

Taxonomy

Sequence | Catalog of Life

Gender | Type Status

?

Type

Author: Philippi, 1865

Interactive View

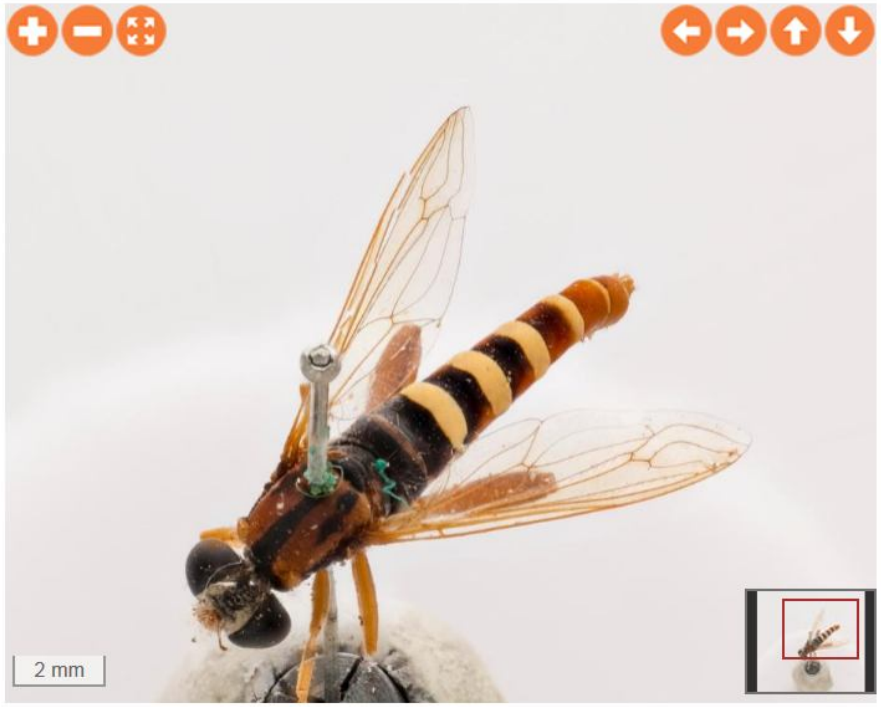
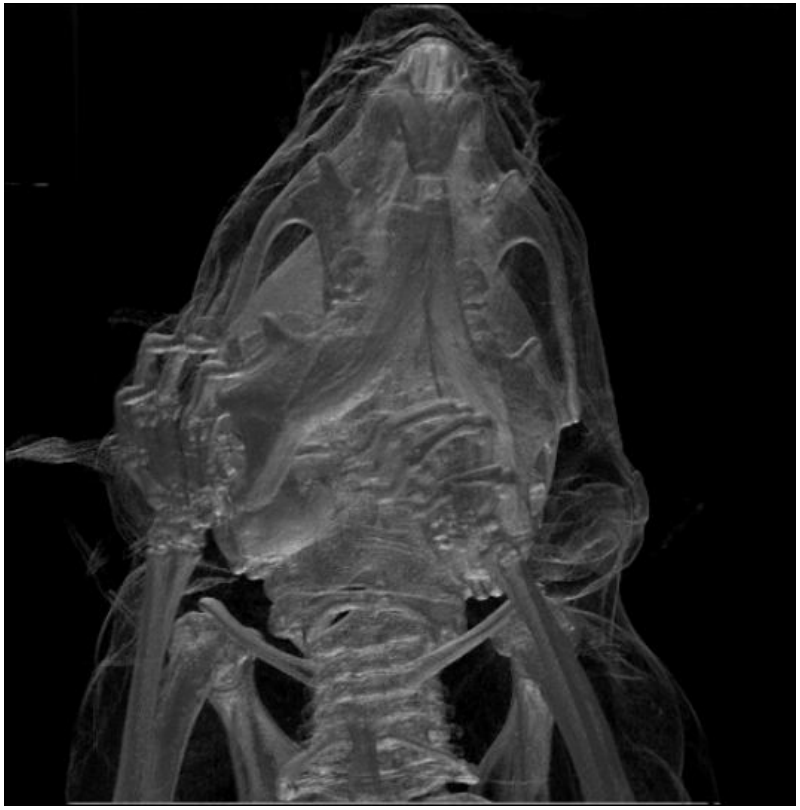


Image License: Creative Commons Zero (CC0)

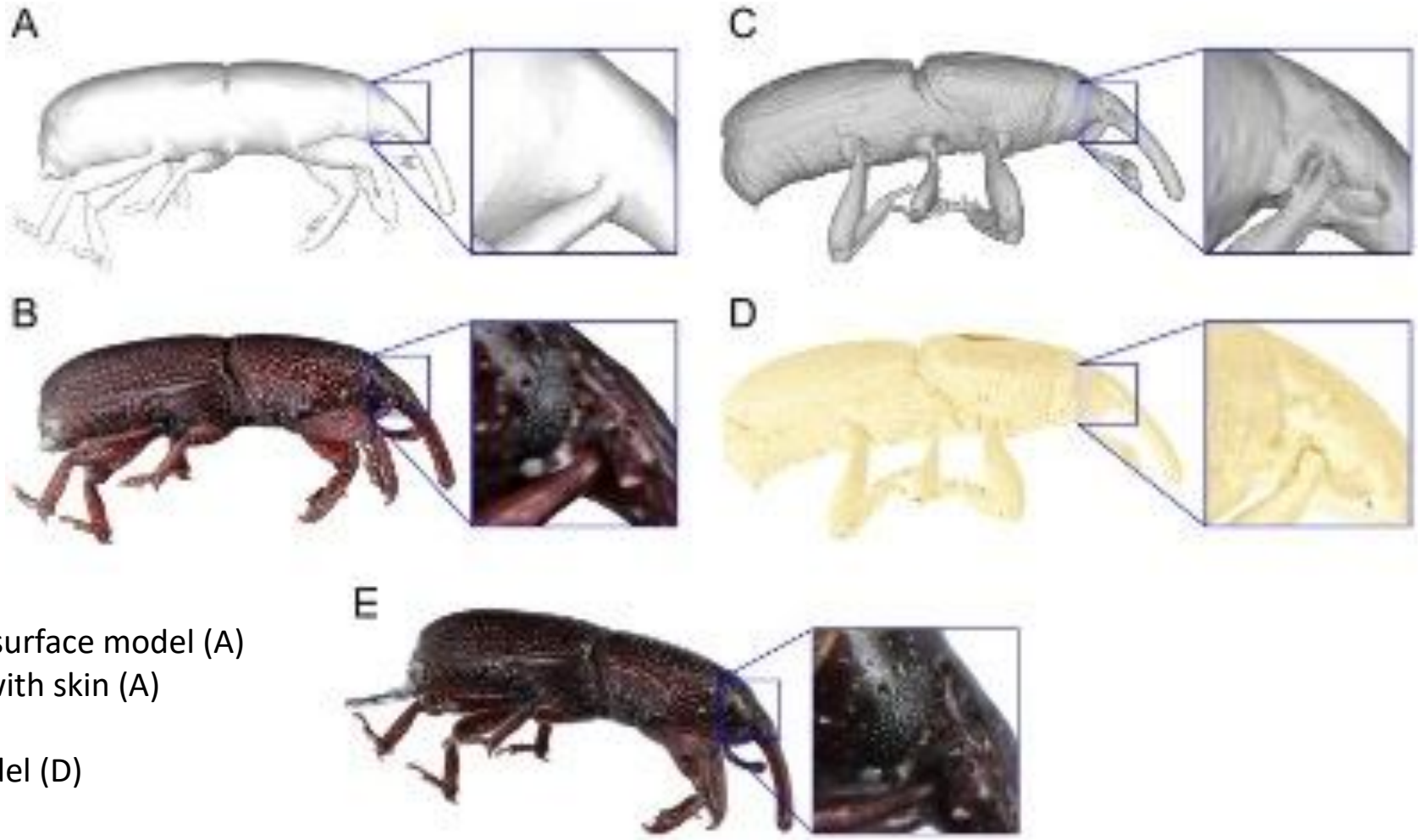
Additional Images

Imaging: 3D internal



Imaging: 3D surface

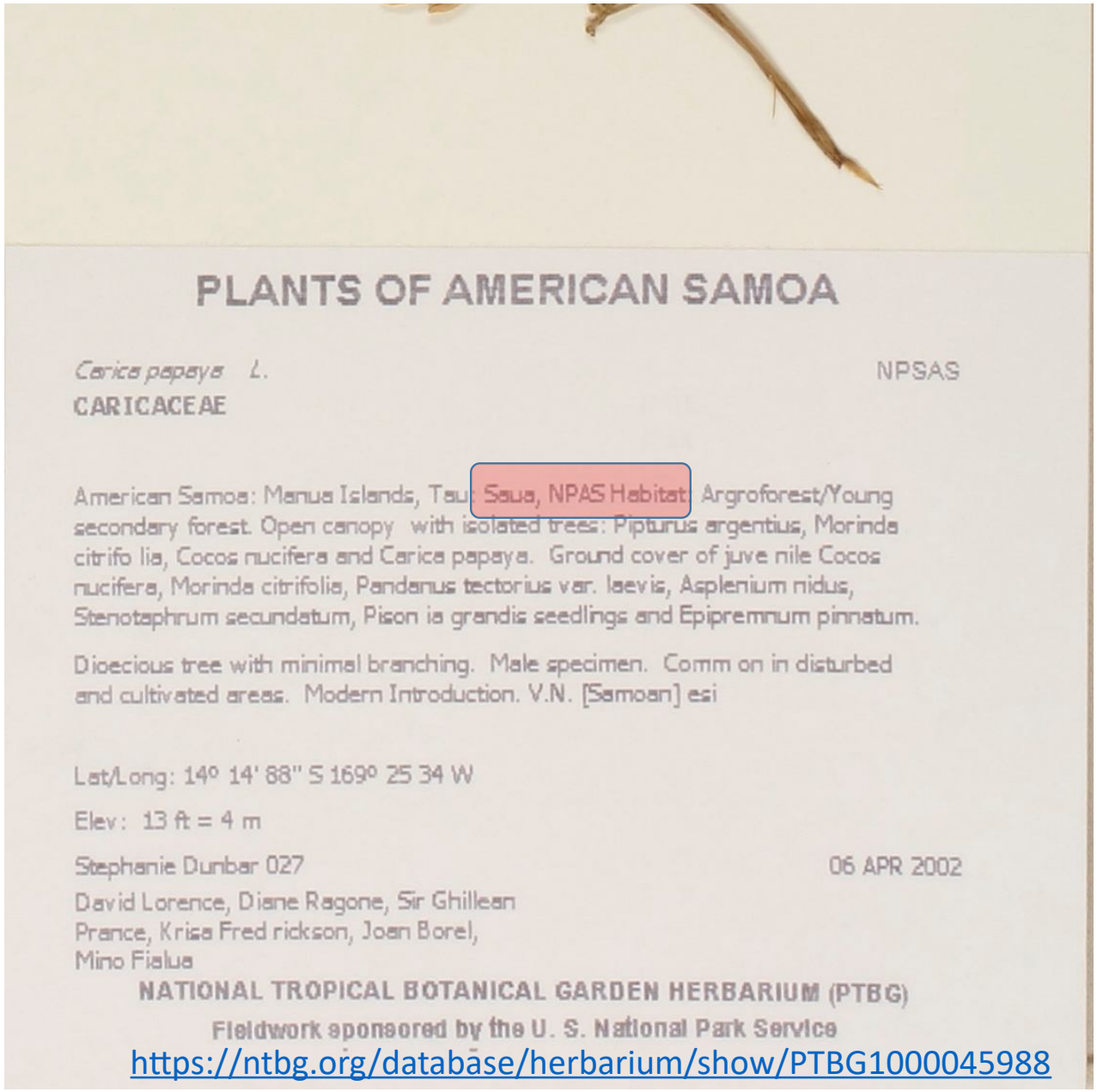
Nguyen C, Lovell D, Adcock M, La Salle J (2014). "Capturing Natural-Colour 3D Models of Insects for Species Discovery and Diagnostics". *PLOS ONE*. DOI:10.1371/journal.pone.0094346.



natural-colour 3D model, surface model (A)
natural-colour 3D model with skin (A)
Micro CT model (C)
False-colour Micro CT model (D)
2D image (E)

Data entry

- Digitized images/data need context/metadata
  - Finding
  - Verifying
  - Analyzing
- Automatic (OCR)
  - Fast, but difficult
- Manual:
  - Excel, databases (inSelect, Acess-based tools)
  - Slow, but a good op can spot mistakes & complete missing data



PLANTS OF AMERICAN SAMOA

Carica papaya L. NPSAS

CARICACEAE

American Samoa: Manua Islands, Tau: Saua, NPAS Habitat: Argroforest/Young secondary forest. Open canopy with isolated trees: Pipturus argentius, Morinda citrifo lia, Cocos nucifera and Carica papaya. Ground cover of juve nile Cocos nucifera, Morinda citrifolia, Pandanus tectorius var. laevis, Asplenium nidus, Stenotaphrum secundatum, Pison ia grandis seedlings and Epipremnum pinnatum.

Dioecious tree with minimal branching. Male specimen. Comm on in disturbed and cultivated areas. Modern Introduction. V.N. [Samoan] esi

Lat/Long: 14º 14' 88" S 169º 25 34 W

Elev: 13 ft = 4 m

Stephanie Dunbar 027 06 APR 2002

David Lorence, Diane Ragone, Sir Ghillean Prance, Krisa Fred rickson, Joan Borel, Mino Fialua

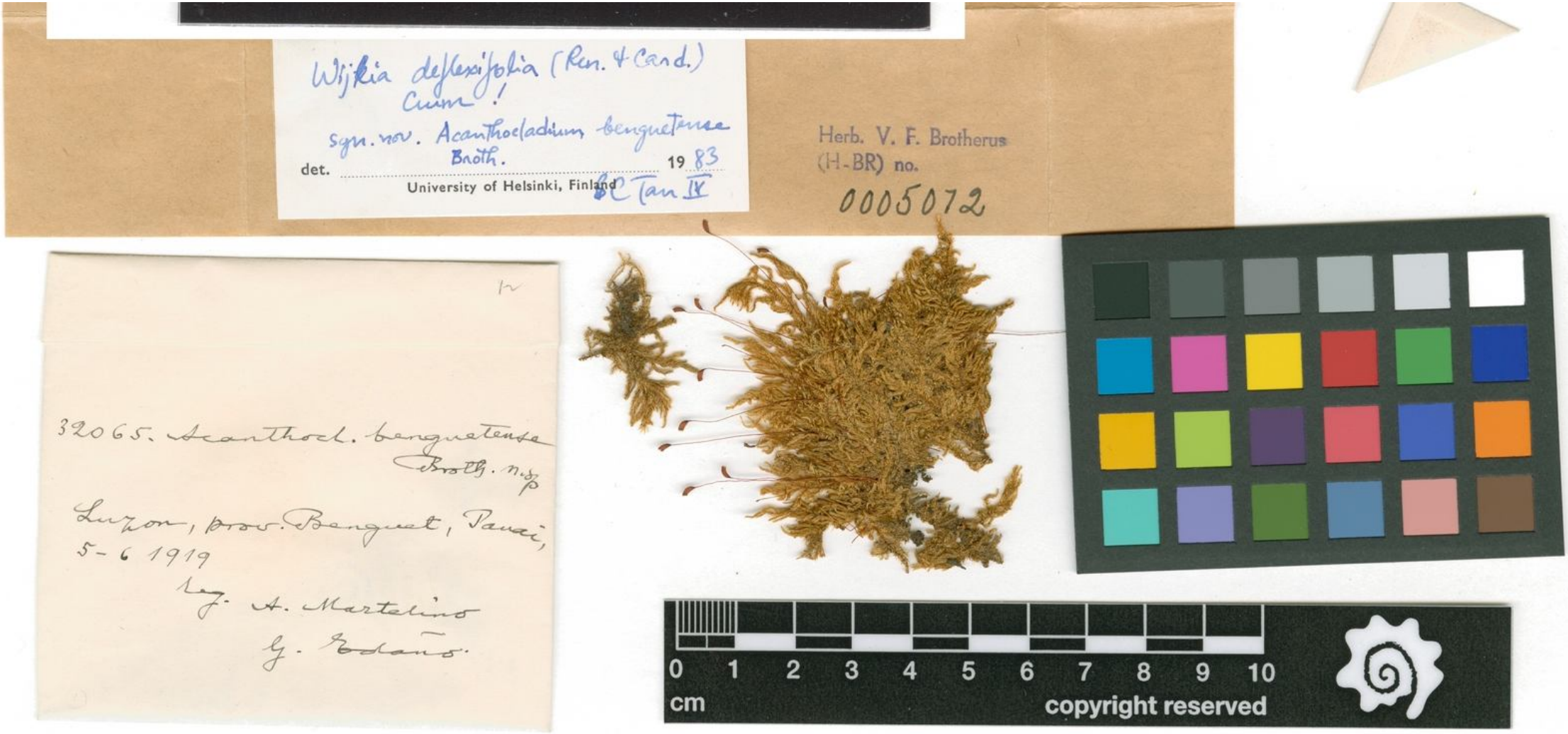
NATIONAL TROPICAL BOTANICAL GARDEN HERBARIUM (PTBG)

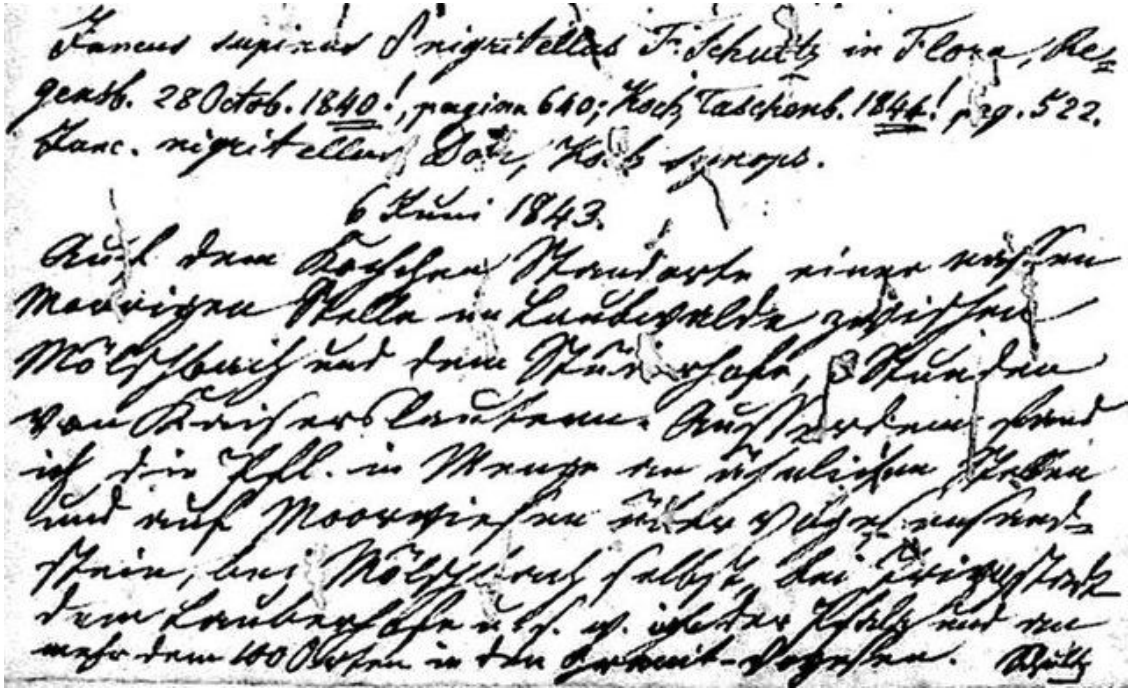
Fieldwork sponsored by the U. S. National Park Service

https://ntbg.org/database/herbarium/show/PTBG1000045988

Data entry

http://id.luomus.fi/HA.H3300003

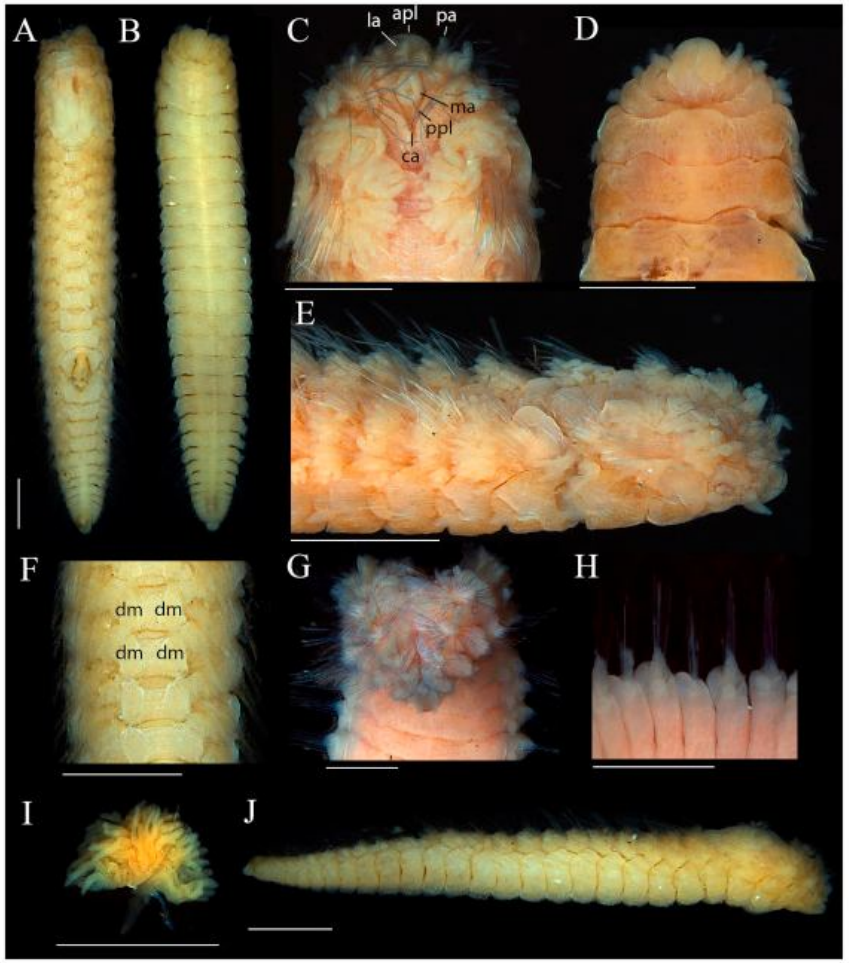




https://www.researchgate.net/publication/251231815_How_really_extensive_is_the_original_material_of_Juncus_kochii_Juncaceae_-_A_taxonomic_and_nomenclatural_revision/figures?lo=1

Keeping track of specimens & data

- Where are the specimens documented a particular record in a database or a publication?
- Which specimen was used for the illustrations in your description?
- What specimens did Dr. Krivosheina examine in 2004 at Luomus?



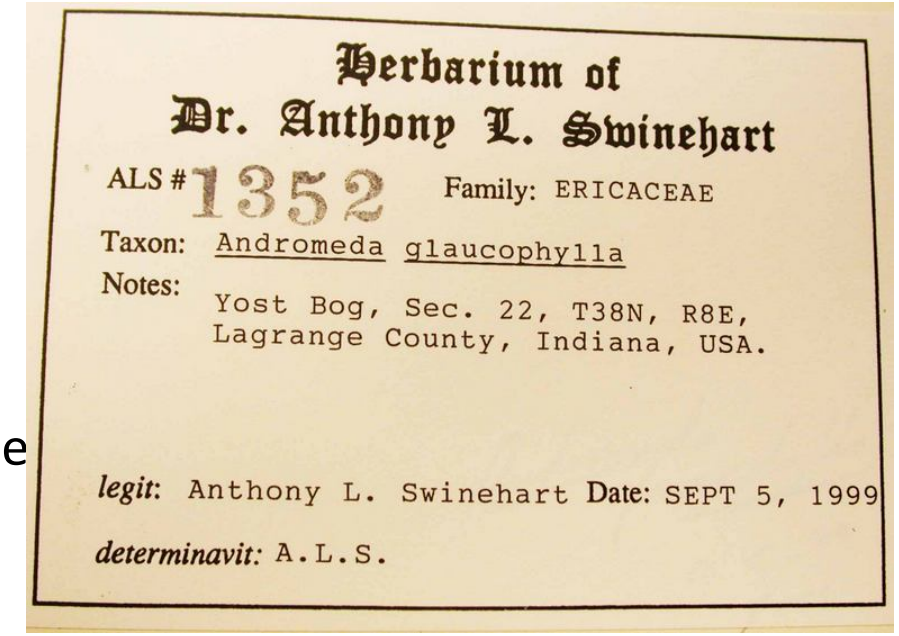
Download high-res image (2MB) Download full-size image

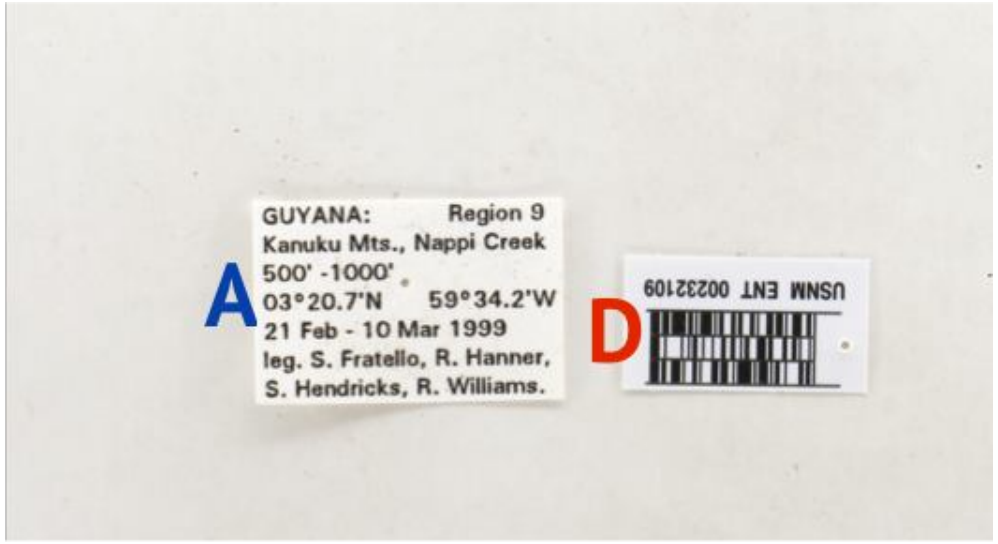
Fig. 3. *Cryptonome barbada* sp. nov. Stereoscope images. A. entire specimen, dorsal view; B. entire specimen, ventral view; C. anterior region, dorsal view; D. anterior region, ventral view; E. anterior region, lateral view; F. mid-body region, dorsal view; G. anterior region, dorsal view, H. Left side of mid-body region, ventral view; I. Branchia; J. entire specimen, lateral view. Scales bar = 1 mm. apl. anterior prostomial lobe; ca. caruncle; dm. dorsal mound; la. lateral antenna; ma. median antenna; pa. palp; ppl. posterior prostomial lobe.

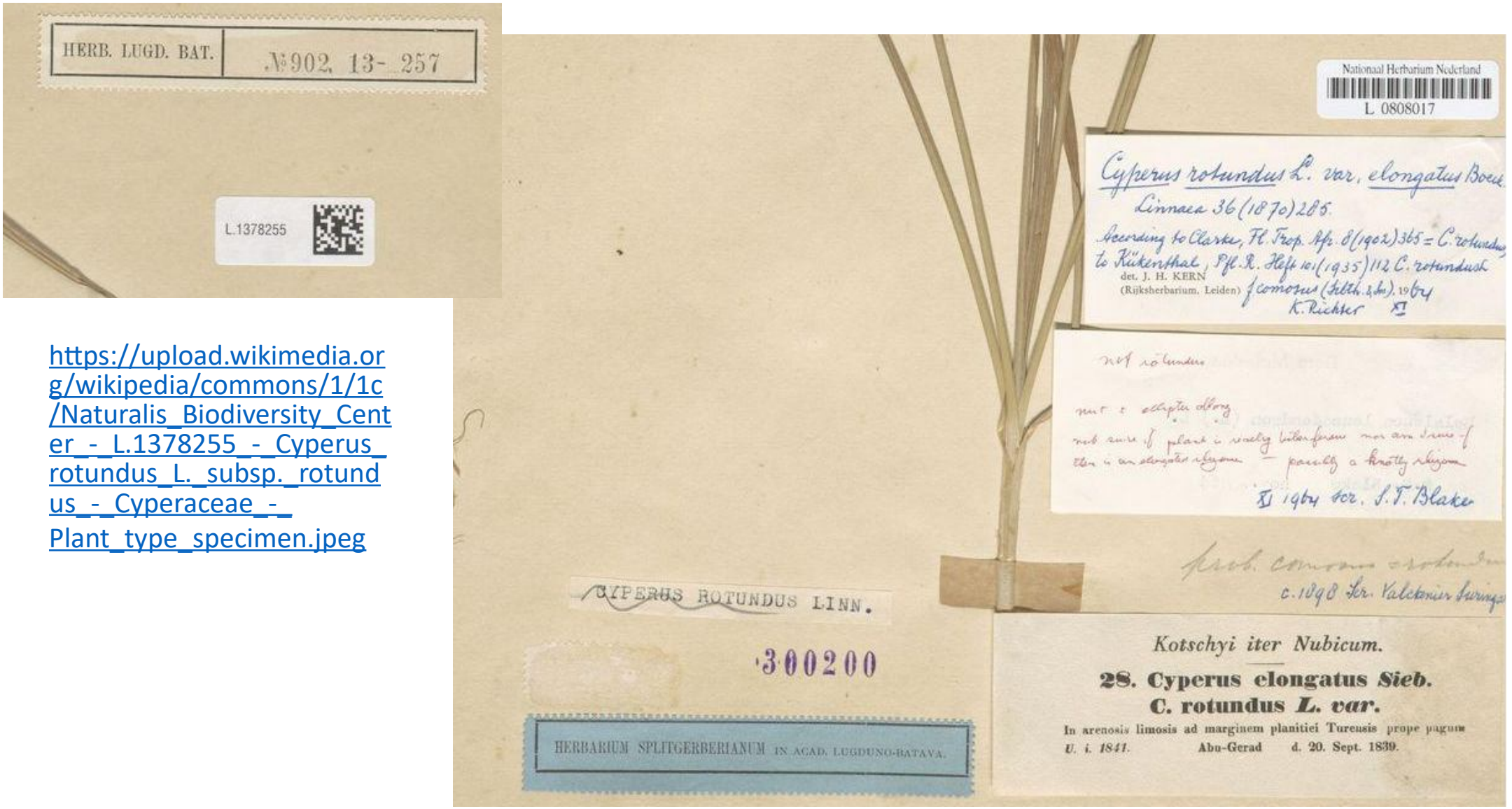
Barroso, Rômulo; Kudenov, Jerry D.; Halanych, Kenneth M.; Saeedi, Hanieh; Sumida, Paulo Y. G.; Bernardino, Angelo F. (2018). A new species of xylophilic fireworm (Annelida: Amphinomidae: Cryptonome) from deep-sea wood falls in the SW Atlantic. *Deep Sea Research Part I: Oceanographic Research Papers.* 137: 66-75

Stable identifiers

- Solves several problems:
  - Referring to a particular sample in text (e
  - Linking related data
- Stable = unique & understandable







https://upload.wikimedia.org/wikipedia/commons/1/1c/Naturalis_Biodiversity_Center_-_L.1378255_-_Cyperus_rotundus_L._subsp._rotundus_-_Cyperaceae_-_Plant_type_specimen.jpeg

Stable identifiers

- Still in flux globally
  - See DOI 10.1371/journal.pbio.2001414 : McMurry: ”Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data”
- CETAF Stable Identifiers
  - URIs: Look like web addresses
- Alternatives:
  - LSID - Life Science Identifiers (deprecated)
    - Looks like this: *urn:lsid:ncbi.nlm.nih.gov:pubmed:12571434*
  - DOI
    - Not really intended for this!

CETAF Stable identifiers

- Examples
  - **http://id.luomus.fi/GV.45118**
  - http://mus.utu.fi/ZMAA.TYPE001
- Museum für Naturkunde Berlin:
  - object at http://coll.mfn-berlin.org/u/ZMB_123
  - rdf at http://coll.mfn-berlin.org/u/ZMB_123.rdf
  - json at http://coll.mfn-berlin.org/u/ZMB_123.json
  - xml at http://coll.mfn-berlin.org/u/ZMB_123.xml
  - html at http://coll.mfn-berlin.org/u/ZMB_123.html
- Standard: https://cetaf.org/cetaf-stable-identifiers
- Best practices: https://wiki.pro-ibiosphere.eu/wiki/Best_practices_for_stable_URIs
- See 10.1093/database/bax003: Güntsch *et al.* (2017) Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. Database 2017:bax003.

The one thing to remember from this part!

- Make your work reproducible,

use stable identifiers

LUOMUS

PBIO-161 BIOLOGICAL COLLECTIONS 4/5 ECTS

Wed 26.viii.
Lecture 8 - **Jere Kahanpää**/digitization team **& Kari Lahti**

Documentation, databases (including KOTKA etc.) + DIGITALIZATION + OPEN data (FinBIF, GBIF) + biodiversity-informatics
- citing specimens by unique identifiers

Kari LAHTI 26.8.2020

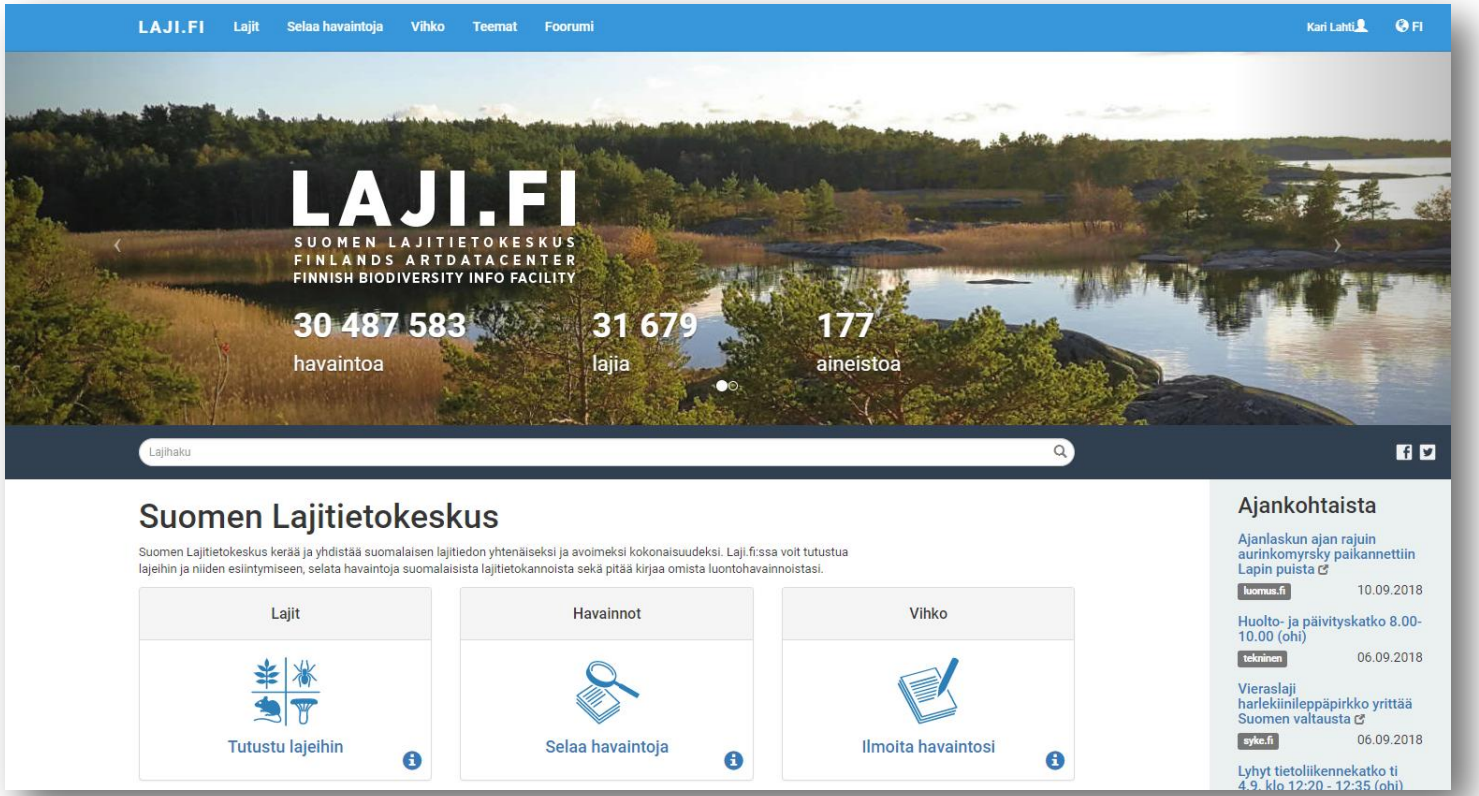
kari.lahti@helsinki.fi

LUOMUS

FinBIF – Finnish Biodiversity Information Facility

- National Infrastructure of Species Information -

LAJI.FI
SUOMEN LAJITIETOKESKUS
FINLANDS ARTDATACENTER
FINNISH BIODIVERSITY INFO FACILITY
SPECIES.FI

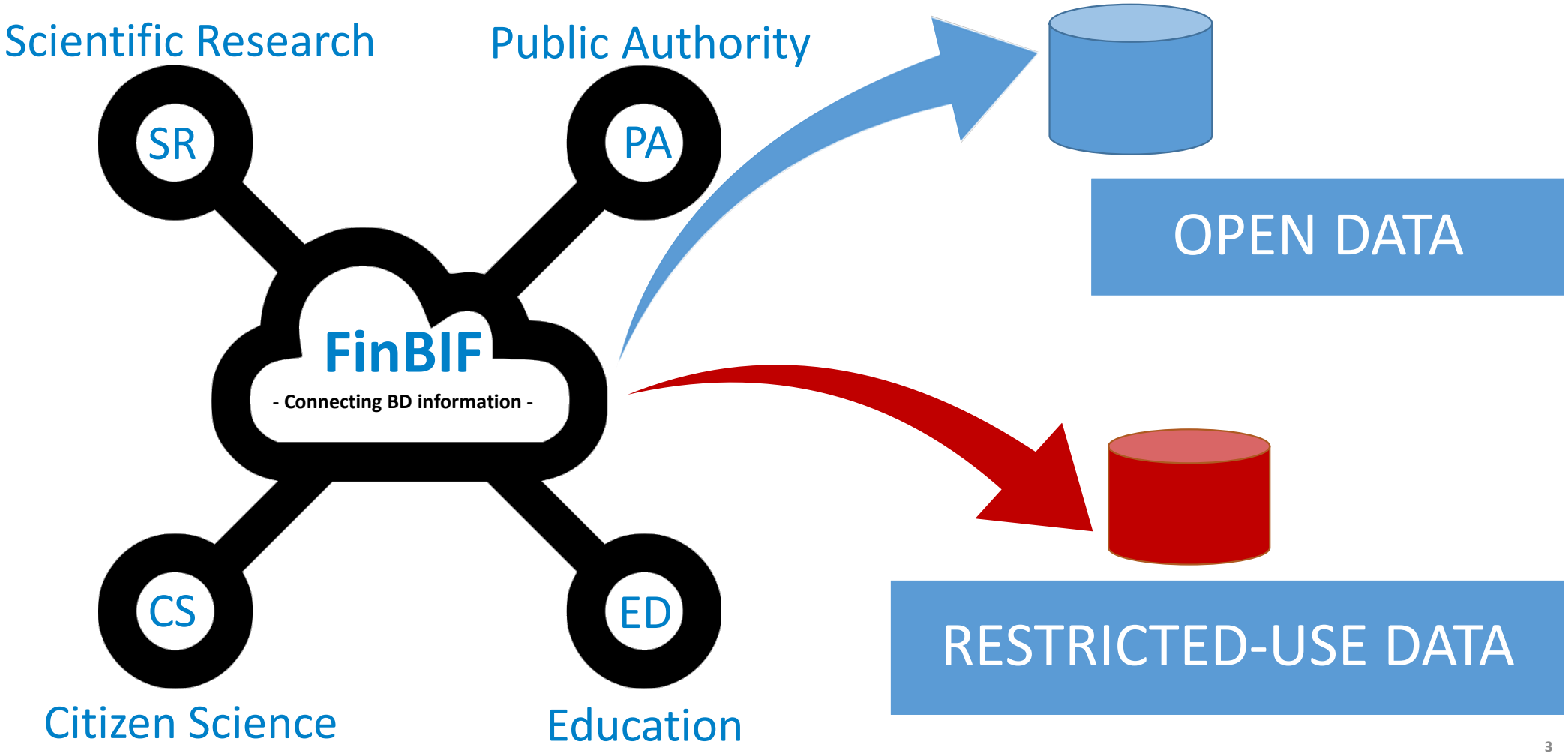


Kari LAHTI 26.8.2020

kari.lahti@helsinki.fi

LUOMUS FinBIF
INNOVATING PRACTICAL SOLUTIONS FOR FAIR BIODIVERSITY INFORMATICS

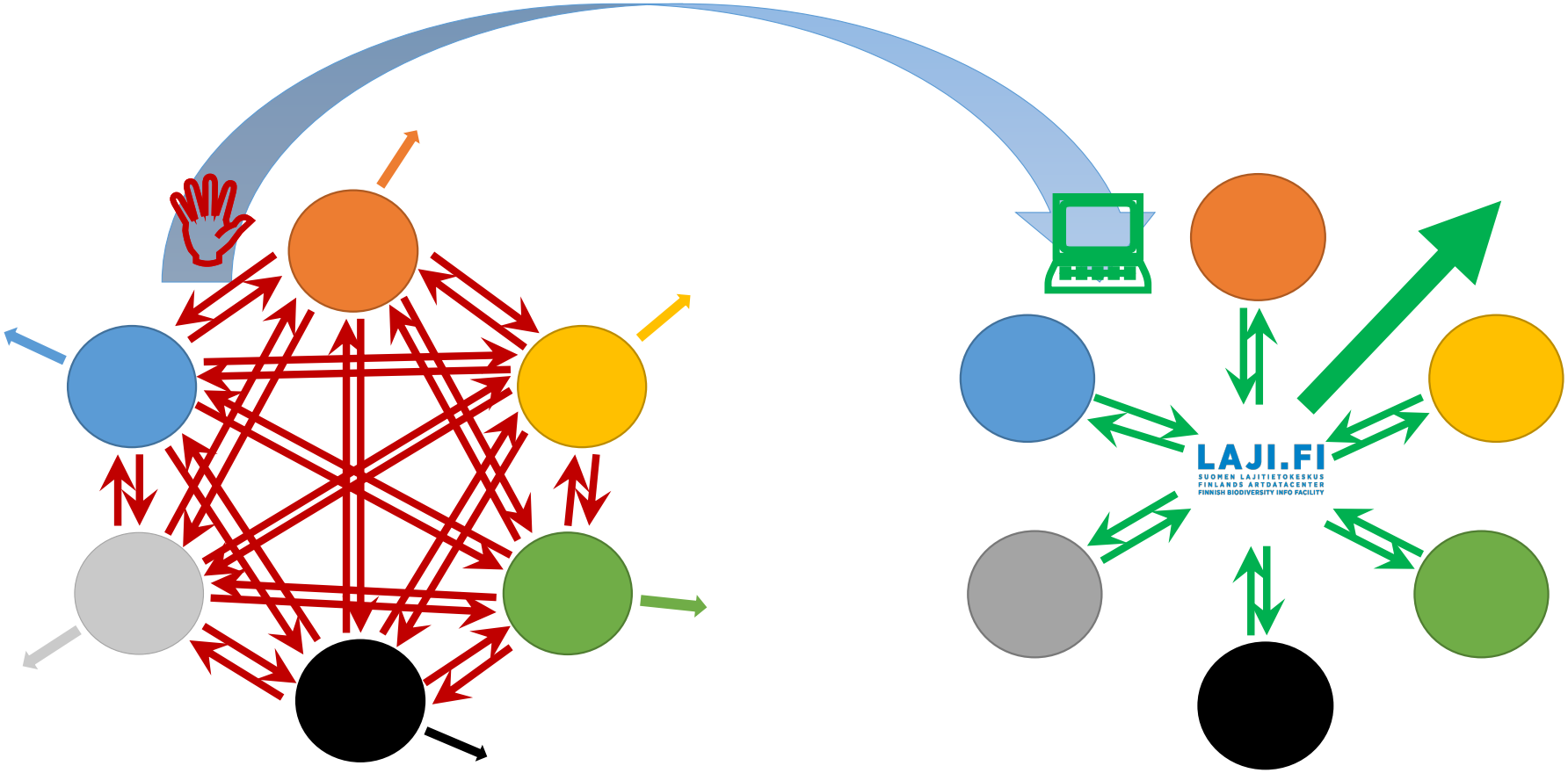
FinBIF Data Sources and Users



FinBIF – From Chaos towards Harmony

LUOMUS LAJI.FI SUOMEN LAJITIETOKESKUS FINLANDS ARTDATACENTER FINNISH BIODIVERSITY INFO FACILITY

FinBIF – From Chaos towards Harmony



FinBIF – From Chaos towards Harmony

Natural Science Collections

LUOMUS
LUONNONTIETEELLINEN KESKUSMUSEO

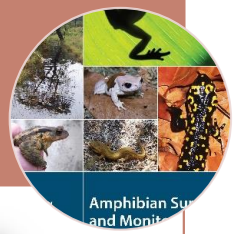
OULUN YLIOPISTO

TURUN YLIOPISTO | Biodiversiteetti
LUONNONTIETEELLINEN MUSEO

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

KUOPION LUONNONTIETEELLINEN MUSEO

Research Monitoring Mapping



SYKE

Luke
LUONNONVARAKESKUS

METSÄHALLITUS

LUOMUS
TALVILINTULASKENTA

TURUN YLIOPISTO

Valtakunnallinen päiväperhosseuranta (NAFI)

Citizen Science Education sector Associations Enthusiasts

LUKIOLAISET LEPAKKOTUTKIJOINA -TIEDEPROJEKTI
Lukiolaiset lepakkotutki

Sieniatlas

Pinkka
Species learning environment

KOKOELMA-KILPAILU

Suomen luonnonsuojeluliitto

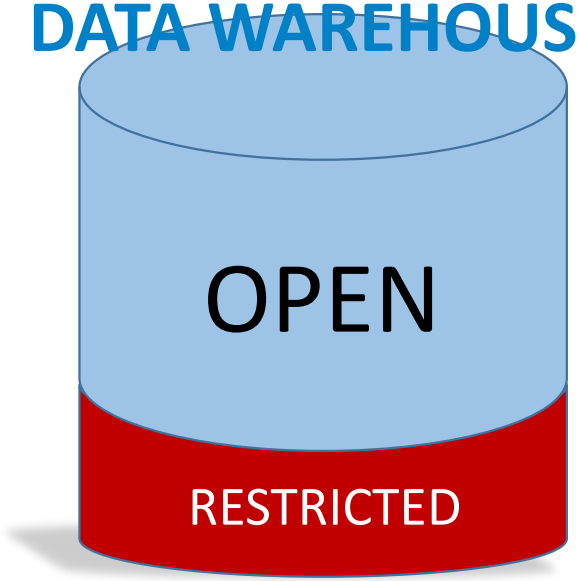
LAJI.FI

LUOMUS LAJI.FI SUOMEN LAJITIETOKESKUS FINLANDS ARTDATACENTER FINNISH BIODIVERSITY INFO FACILITY

FinBIF – Data sharing and usage

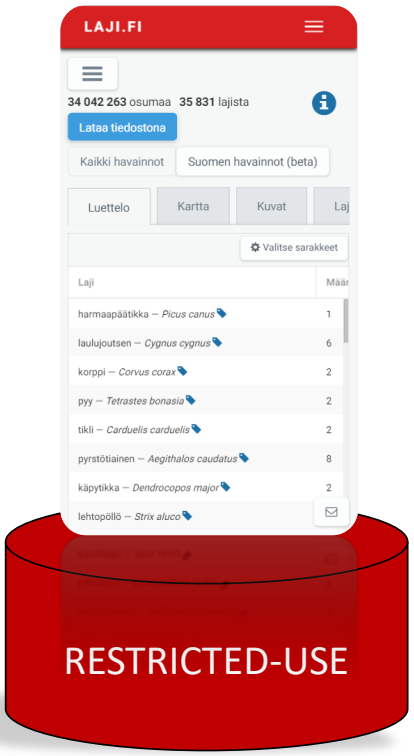
LAJI.FI

DATA WAREHOUSE



FinBIF – Data sharing and usage

publicauthorities.laji.fi



RESTRICTED-USE

DATA WAREHOUSE

LAJI.FI

LAJI.FI

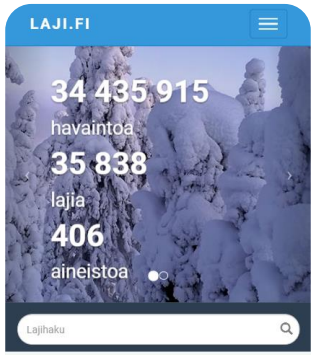
DATA WAREHOUSE

OPEN

RESTRICTED

GENERALISED

Laji.fi



Suomen Lajitietokeskus

Suomen Lajitietokeskus kerää ja yhdistää suomalaisen lajitiedon yhtenäiseksi ja avoimeksi kokonaisuudeksi. Laji.fi:ssä voit tutustua lajeihin ja niiden esiintymiseen, selata havaintoja suomalaisista lajitietokannoista sekä pitää kirjaa omista luontohavainnoistasi.

AVOIN

TIETOVARASTO

LAJI.FI

FinBIF – The usage of restricted-use data

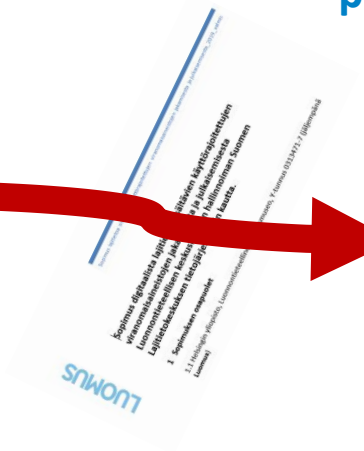
Public authorities

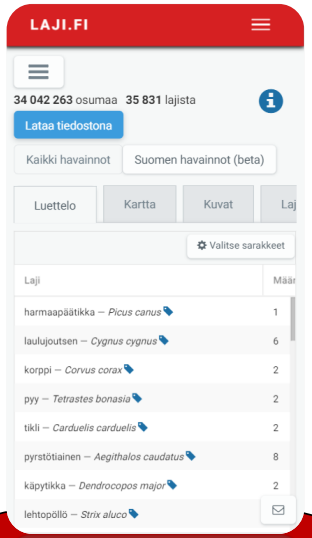
ELY’s (Regional Env. Centres)
Municipalities
Ministries
Research institutes

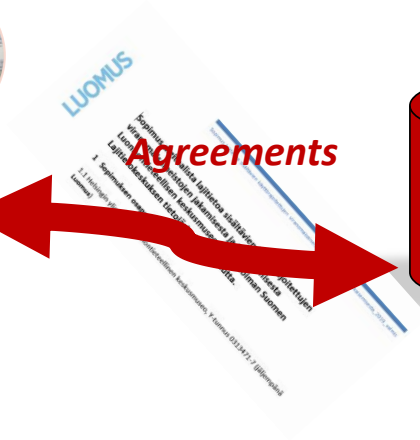
The Information systems of public authorities

LajiGIS
Hertta
Aarni
ArcGIS
Mapinfo
…

publicauthorities.laji.fi







Agreements

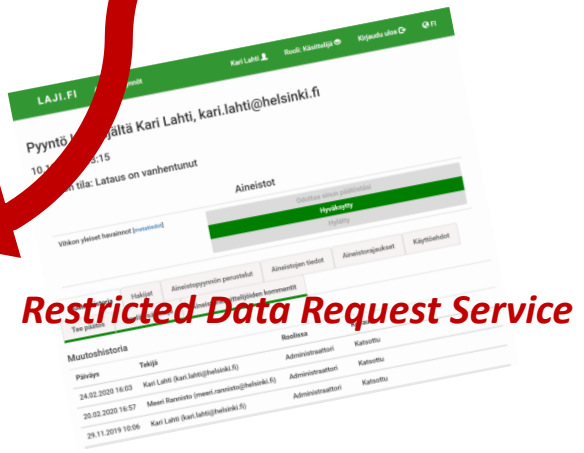
RESTRICTED-USE

DATA WAREHOUSE

LAJI.FI

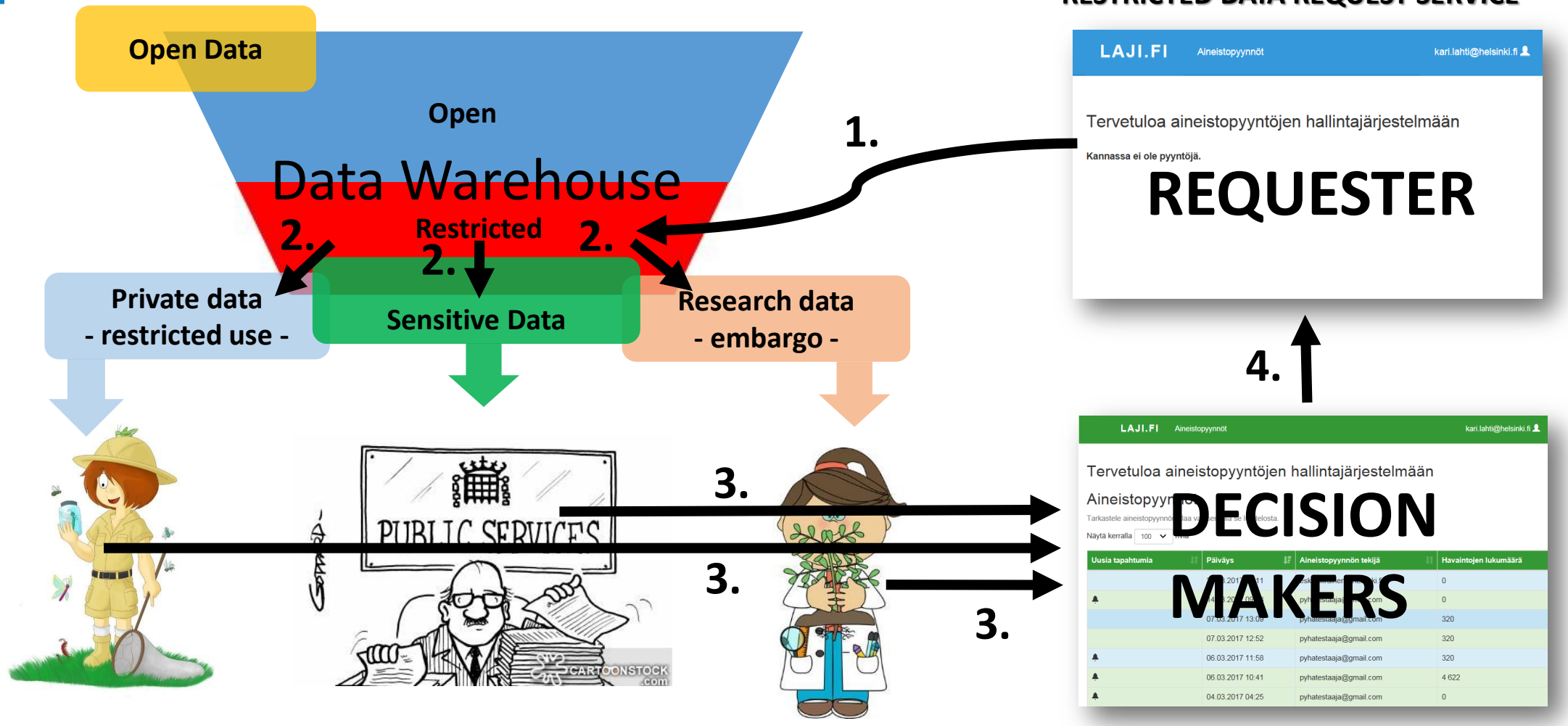
Research Assignments Consultation

Universities
Private sector
Research institutes



Restricted Data Request Service

FinBIF – The usage of restricted-use data



http://schreibaby-zephyr.deviantart.com/art/Entomologist-Kitaro-221196716

https://s-media-cache-ak0.pinimg.com/564x/2c/ec/30/2cec3023b71a55109838d3c92c932abe.jpg

FAIR principles as a “pressure test”

Published 2016*

Adopted widely

- EC European Open Science Cloud (EOSC) ”As Open as Possible, as Closed as Necessary”
- Horizon 2020
  - Turning FAIR into reality **(EUROPA>Publications Office of the EU>Publication detail> Turning FAIR into reality)**

Aim is to make the data:

- **Findable**
- **Accessible**
- **Interoperable**
- **Re-usable**

1. The elements of the FAIR Principles are related, but independent and separable.
2. The Principles assist discovery and reuse by third-parties.
3. The barrier-to-entry is maintained as low as possible.
4. The Principles function in any combination and incrementally increase degrees of ‘FAIRness’.

*https://www.nature.com/articles/sdata201618, https://www.go-fair.org/fair-principles/

FAIR & Finnish Biodiversity Information Facility

TEST results of FinBIF (self assessment) *** = max score

Findable

1. Persistent identifier, PID
2. Rich metadata
3. Registered in a searchable resource
4. PID is specified at the metadata

**→

Accessible

1. Data should be retrievable by identifier according to principle ”As Open as Possible, as Closed as Necessary”
2. Protocol is open, free, and universally implementable
3. Registration and authorisation supported, where necessary
4. Metadata still available even when the data is no longer available.

***

Interoperable

1. Data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
2. Data use vocabularies that follow FAIR principles.
3. Data include qualified references to other data

**

Re-usable

1. Meta(data) have a plurality of accurate and relevant attributes.
2. Data are released with a clear and accessible data usage license.
3. Data are associated with their provenance.
4. Data meet domain-relevant community standards

**→

FinBIF – Unique PIDs, Unique Peristent Identifiers

Unique PIDs

FinBIF uses a ***persistent HTTP-URI identifier*** for all types of real-life and digital objects (specimens, occurrences, taxa, metadata, persons, organisations, information systems, etc.), as recommended by the World Wide Web Consortium (Best practices for publishing linked data; https://www.w3.org/TR/ld-bp/).

The identifier takes the user to an ID redirect service, which redirects the user to a page that shows information about the object in human-readable format. For example, specimen identifiers redirect to information about the specimen and taxon identifiers to a page describing the taxon.

The redirect service can also provide machine-readable data about the object, if the user (client software) requests that using Accept headers.

If partner organisations do not provide HTTP-URI identifiers for their occurrences, FinBIF will use the persistent internal IDs of the data source to generate globally unique URI identifiers.

DOI (Digital Object Identifier) identifiers for data downloads and dataset metadata will be created in the near future. (https://www.doi.org/driven_by_doi/DOI_Marketing_Brochure.pdf).

FinBIF – Unique PIDs, Unique Peristent Identifiers

LAJI.FI

på svenska in English

Tiedostolataus http://tun.fi/HBF.5167

Latauspäivä: 26.8.2020

Osumien lkm: 965

Rajaukset:

Kohde (laji): Orchidaceae (MX.40029)
Eliömaakunta: Ahvenanmaa (A)
Lataus tietovarastoon, päivänä tai ennen: 2020-08-26

Viittausohje

Voit viitata tähän lataukseen seuraavasti:

Suomen Lajitietokeskus/FinBIF. http://tun.fi/HBF.5167 (haettu 26.8.2020).

Jos käytät vain osaa aineistoista, on suositeltavaa, että viittaat vain niihin aineistoihin. Latauksen osajoukkoon voi viitata seuraavasti (poista käyttämätön aineisto):

Suomen Lajitietokeskus/FinBIF. http://tun.fi/HBF.5167, http://tun.fi/HR.447, http://tun.fi/HR.169, http://tun.fi/HR.3

Viitataksesi latauksen yksittäiseen riviin voit käyttää [Document.DocumentID]-kenttää, esimerkiksi:

Suomen Lajitietokeskus/FinBIF. http://tun.fi/EXMP.1234, http://some.org/9876 (haettu 26.8.2020).

Aineistot

Lataus koostuu seuraavista aineistoista joille on määritelty käyttöoikeuslisenssi:

Kuopio Natural History Museum - KUO Putkilokasvikokoelmat (KUO) - http://tun.fi/HR.430 [metadata]
Creative Commons Nimeä
Lisätietoja tämän aineiston käytöstä antaa outi.vainio@kuopio.fi

LajiGIS: Lajin seurantakohteet - http://tun.fi/HR.3553 [metadata]
Creative Commons Nimeä
Lisätietoja tämän aineiston käytöstä antaa lajigis@metsa.fi

Luomus - Hatikka.fi:n havainnot - http://tun.fi/HR.447 [metadata]
Creative Commons Nimeä
Lisätietoja tämän aineiston käytöstä antaa info@laji.fi

Luomus - Putkilokasvikokoelmat - http://tun.fi/HR.169 [metadata]
Creative Commons Nimeä
Lisätietoja tämän aineiston käytöstä antaa henry.vare@helsinki.fi

Lataa tiedosto

Lataamalla tiedoston sitoudut noudattamaan yllä mainittuja käyttöoikeuslisenssejä. Lisenssit löytyvät myös latauksen `readme.txt` tiedostosta.

HBF.5167.zip (0.1 Mt)

Tiedoston vienti Exceliin Tiedoston vienti ArcGIS -paikkatieto-ohjelmaan

Kotka CMS – collection management system

Kotka CMS

- Kotka is one of the two **primary data** management systems of FinBIF
- Kotka applies simple and **pragmatic** approaches. This has helped it grow into a nationally used system.
- The aim is to improve **collection management efficiency** by providing practical tools.
- Kotka **emphasises the quantity** of digitised specimens over completeness of the data. It harmonises practices by bringing all types of collections under one system; the types currently covered include zoological, botanical, mycological and palaeontological museum collections, tissue and DNA samples, and botanic garden and microbial living collections.
- Kotka stores data mostly in a denormalised free text format using a triplestore and a simple hierarchical data model. This allows greater flexibility of use and faster development compared to a normalized relational database.
- Kotka does some data validation, but **quality control** is seen as a continuous process and is **mostly done after the data have been recorded** into the system.
- Kotka is a **web application**. Data can be entered, edited, searched and exported through a browser-based user interface (UI). However, most users prefer to enter new data in customizable MS-Excel templates, which support the hierarchical data model, and upload these to Kotka. Batch updates can also be done using Excel.
- Kotka **stores all revisions** of the data to avoid any data loss due to technical or human error.
- Kotka supports **designing and printing specimen labels** (Heikkinen et al. 2019b), annotations by external users, and handling accessions, loan transactions, and the Nagoya protocol (Kuusijärvi et al. 2019).
- https://biss.pensoft.net/article/37181/list/19/

FinBIF – RELEVANCE AND EFFECTIVENESS

Decision making

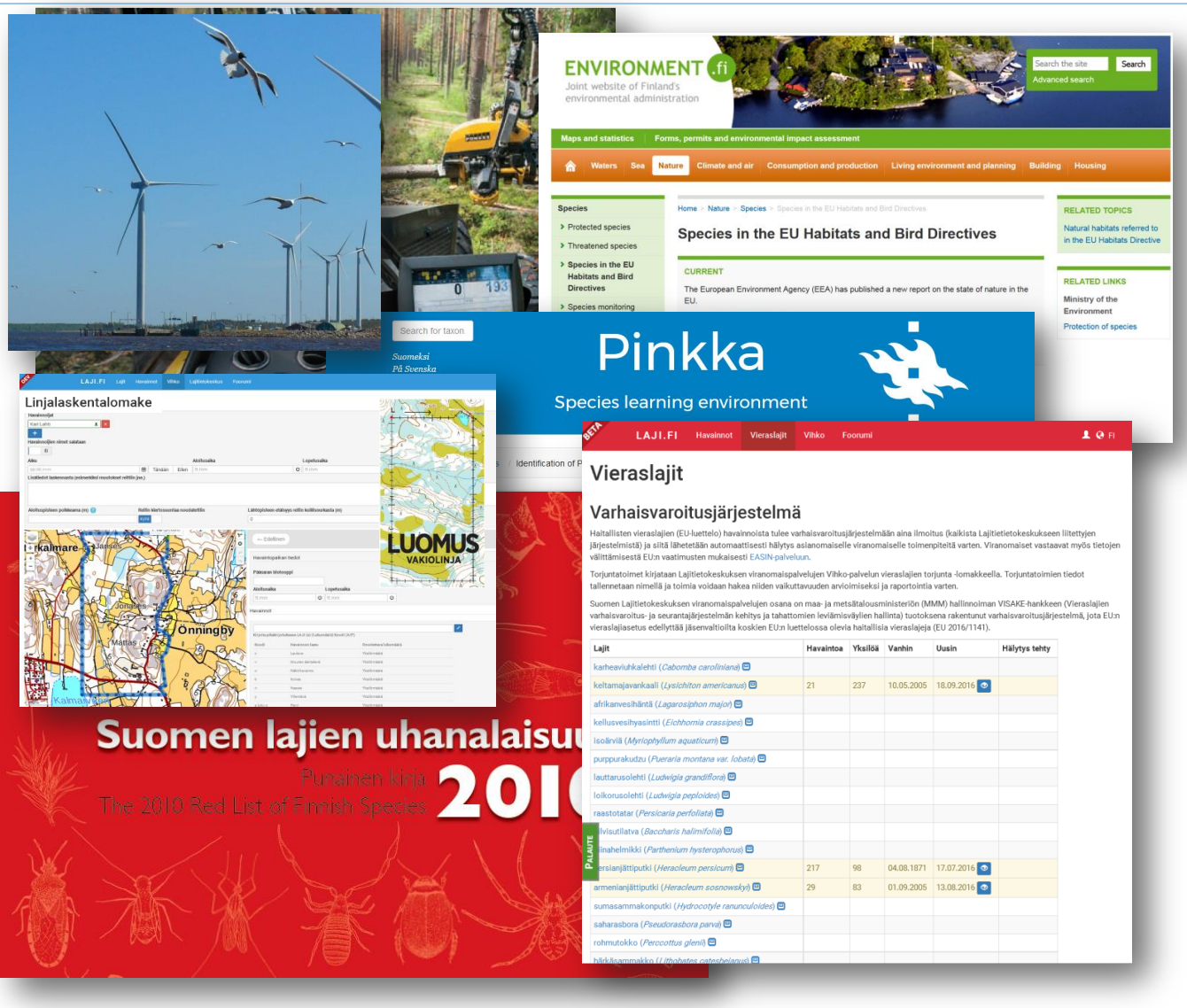
- Sustainable use of Natural Resources
- Land use practices and planning
- Nature Conservation, species protection, Red Data Books
- EU and National Reporting
- Invasive Alien Species; early warning and eradication

Research

- Species surveys and censuses
- Climate Change indications

Education

- Schools – Species identification and digital herbaria
- University – Learning environment



Challenges encountered with database and data

1. Biggest challenge is to convince data owners to **share** their data especially **as Open Data**
2. The diverse use of **different taxonomies**, taxonomic backbones and scientific names in **defining the same taxon concepts** creates a huge challenge, which we are trying to tackle by applying **Linked Data principles with taxon concept URI-identifiers.** Harmonising the used taxonomies to be linked or redefined with the national taxonomy of FinBIF is the ultimate national goal. Nordic-Baltic pilot to link the regional taxonomies is under way through NeIC led project DeepDive.
3. Endless need to **provide tools** to assist users in the **process of sharing** their data **and using** FinBIF data (Excel imports-exports, E-forms, GIS-application support, API interfaces…). Data is stored in such a huge variety of forms – standards enormously needed.
4. **Data flow issues** from a content standpoint are mainly concerning how to deal with the **data quality**, how to handle **data sensitivity**, how to **manage scientific research data** to allow enough time for analysing and publishing and at the same time share the e.g. raw species occurrence data asap for needed use (land use planning and practices, EIA etc.)
5. Data policies are often institutional and quite often protect the institution’s internal potential benefits instead of supporting open data. Licencing, sensitive information and use-restrictions are most difficult issues to solve when designing the data policy. To cover the legal aspects is another challenge.

LUOMUS LAJI.FI
SUOMEN LAJITIETOKESKUS
FINLANDS ARTDATACENTER
FINNISH BIODIVERSITY INFO FACILITY

Summary

FinBIF

LAJI.FI
SUOMEN LAJITIETOKESKUS
FINLANDS ARTDATACENTER
FINNISH BIODIVERSITY INFO FACILITY

Basic plan of the lecture

- Part 3: Finding & acquiring collection specimen data (Jere)
  - Some data sources in more detail
  - Data formats
  - Caveats concerning specimen data

Finding data

- Can be really tricky for most organisms!
- Too obscure/too much
- Finding data != finding good or original data
- Choose your requirements *before* searching

Sources of data: collection specimen data

- Occurence data (plentiful) vs. other data (sparse)
- **Primary & non-primary sources**

BOLD

GBIF

MorphoBank

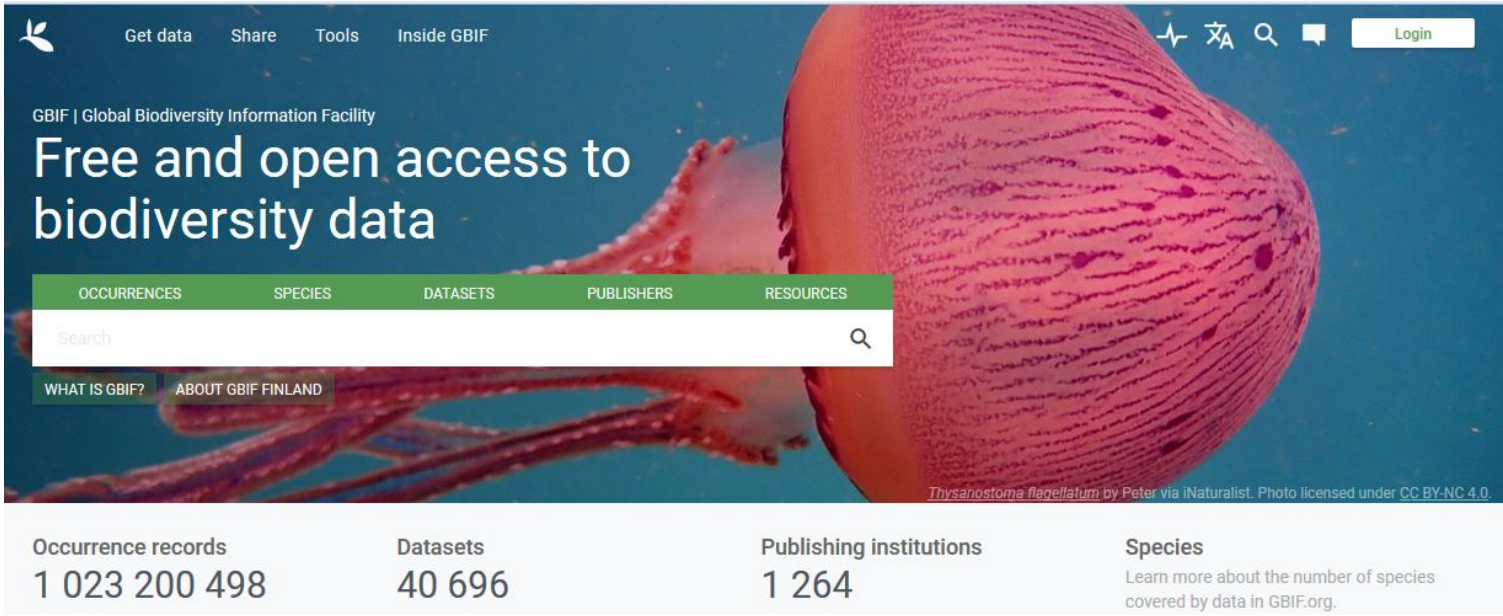
There are 630 publicly accessible projects as of September 17, 2018 in MorphoBank. Publicly available projects contain 106,451 images and 414 matrices. MorphoBank also has an additional 1,124 projects that are in progress. These contain an additional 148,110 images and 887 matrices. These will become available as scientists complete their research and release these data. 2,423 scientists and students are content builders on MorphoBank. 11183 site visitors viewed or downloaded data in the last thirty days.

Collection sample data sources: Finland

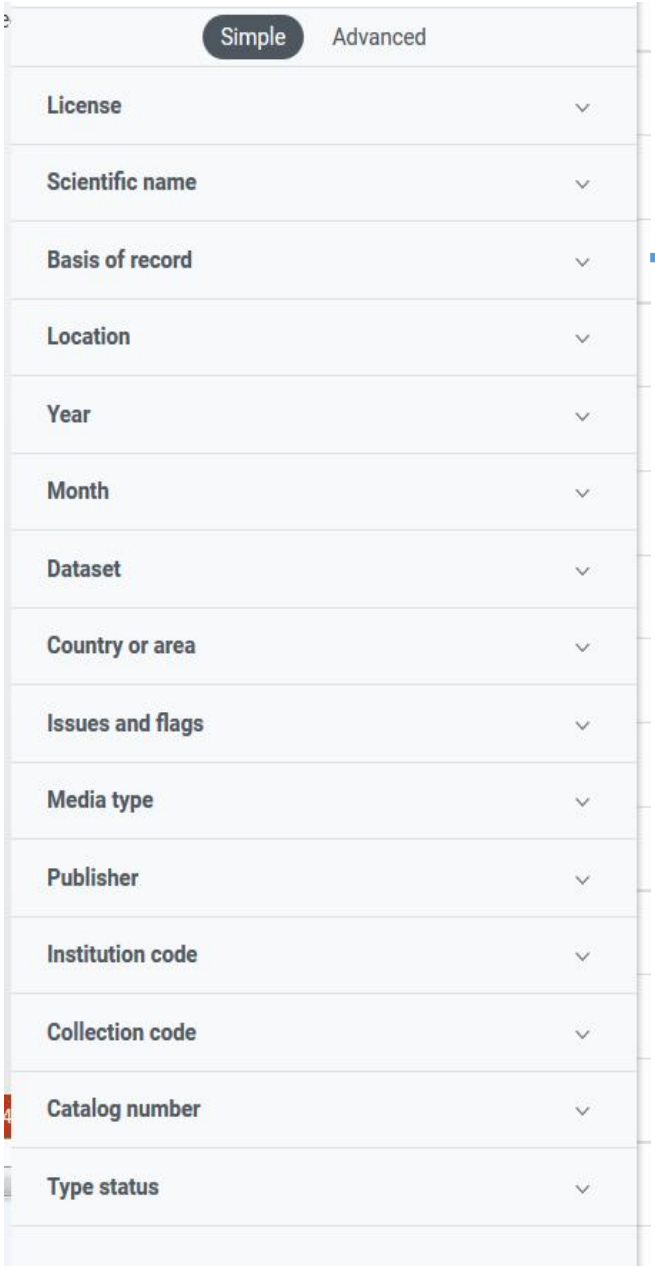
- Laji.fi (portal & partial primary source)
- Literature
- Private databases of researchers
- [Governmental databases, mostly not open data (Hertta @ SYKE, LajiGIS @ Metsähallitus, municipalities etc)]
- [Third sector databases like Tiira @ Birdlife Finland, consulting firms etc.]

Collection sample data sources: global

- GBIF (portal)
  - Original focused mostly on occurences from observations
  - Species
  - Curated datasets
- Other databases
  - www.ornisnet.org/ (NA bird specimens)
  - iobis.org/ (global marine biodiv.)
  - Many others by field or region
- Literature
  - Biodiversity Heritage Library (www.biodiversitylibrary.org/)
  - JSTOR (www.jstor.org/)

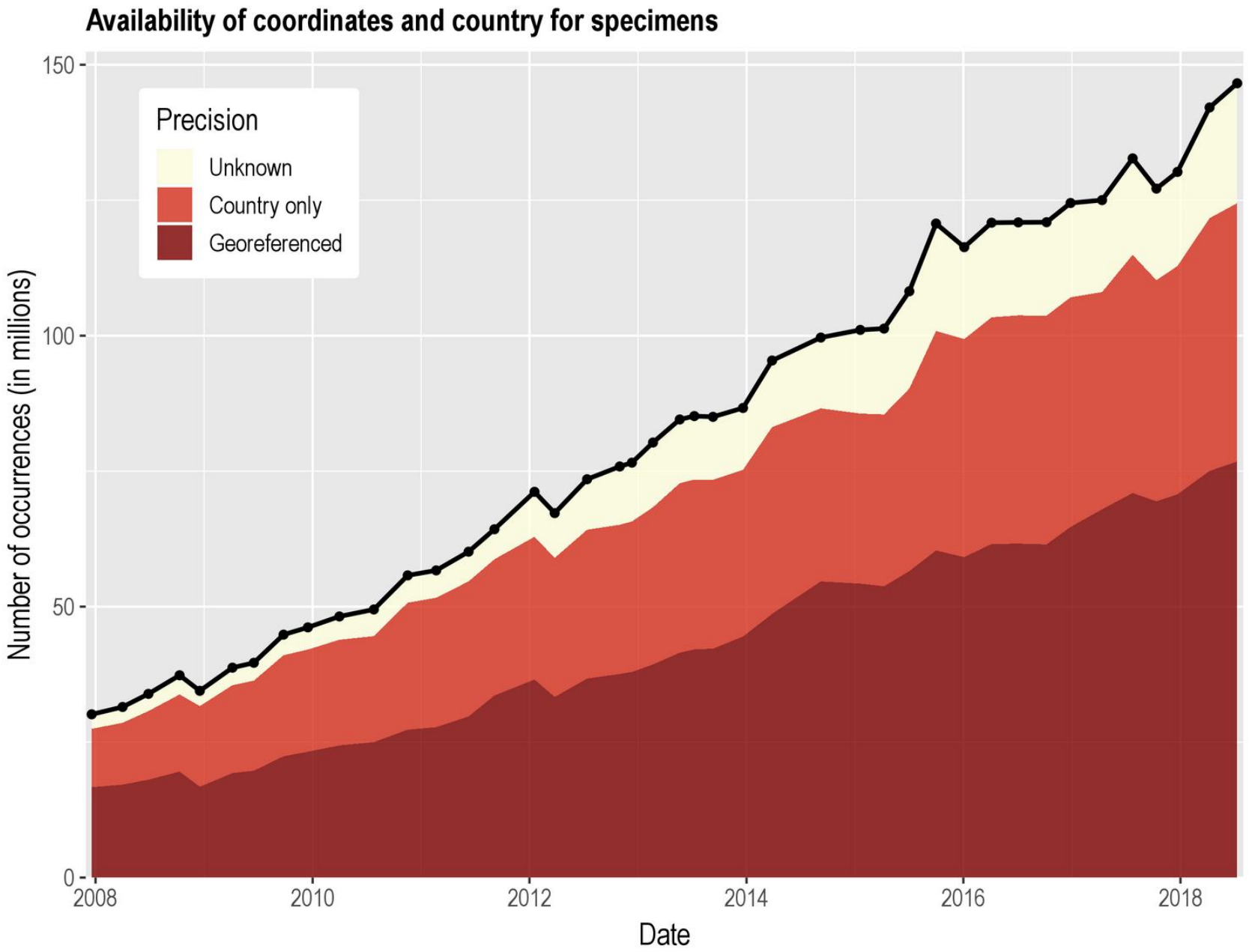


GBIF


Simple
Advanced
License
Scientific name
Basis of record
Location
Year
Month
Dataset
Country or area
Issues and flags
Media type
Publisher
Institution code
Collection code
Catalog number
Type status


Basis of record
Observation 21 737 696
Machine observation 11 004 946
Human observation 805 370 753
Material sample 554 308
Literature 234 405
Preserved specimen 149 053 992
Fossil specimen 10 053 234
Living specimen 1 409 721
Unknown 24 079 355

GBIF


Availability of coordinates and country for specimens
Precision
Unknown
Country only
Georeferenced
Number of occurrences (in millions)
150
100
50
0
2008
2010
2012
2014
2016
2018
Date

Sources of data: morpho/gene data (global)

- DNA
  - BOLD – best traceability back to specimens (v4.boldsystems.org/)
  - Genbank/European Nucleotide Archive – large, but has quality issues (www.ebi.ac.uk/ena)
- Character libraries:
  - Morphobank *morphobank.org/*
- 2D/3D image/model libraries:
  - No major archive, very scattered

Data formats: occurence data

- Comma-separated values (CSV)
- Excel files (.xls/.xlsx)
- Darwin Core

Simple Darwin Core

	A	CF	CG	CH	CI
1	Unit.UnitID	Gathering.MunicipalityVerbatim	Gathering.BioProvinceVerbatim	Gathering.ProvinceVerbatim	Gathering.LocalityVerbatim
46	http://tun.fi/MY.452398			Murmansk	Kuzomen
47	http://tun.fi/MY.452932	Silvaplana		GraubÃ¼nden	
48	http://tun.fi/MY.460127	Parainen	Varsinais-Suomi		Lofsdal
49	http://tun.fi/MY.460130	KolatselkÃ¤		Karelian Republic	
50	http://tun.fi/MY.460133	Kuusamo	Koillismaa		Juuma, PetÃ¤jikkÃ¶puro
51	http://tun.fi/MY.460136	EnontekiÃ¶	EnontekiÃ¶n Lappi		between Vittanki and Mukkav
52	http://tun.fi/MY.460140	Enontekis	EnontekiÃ¶n Lappi		between Naimakka and Vittar
53	http://tun.fi/MY.460144	Muonio	KittilÃ¤n Lappi		on the way to Olostunturi
54	http://tun.fi/MY.460148	Utsjoki	Inarin Lappi		MantojÃ¤rvi
55	http://tun.fi/MY.460152	Enontekis	EnontekiÃ¶n Lappi		KilpisjÃ¤rvitrakten
56	http://tun.fi/MY.460156	Espoo	Uusimaa		
57	http://tun.fi/MY.460160	Vihti	Varsinais-Suomi		PÃ¤ivÃ¶lÃ¤

Darwin Core

- http://rs.tdwg.org/dwc/

```
<dcterms:Location>
    <dwc:locationID>http://guid.mvz.org/sites/arg/127</dwc:locationID>
    <dwc:country>Argentina</dwc:country>
    <dwc:countryCode>AR</dwc:countryCode>
    <dwc:stateProvince>Neuquén</dwc:stateProvince>
    <dwc:locality>Valle Limay, Estancia Rincon Grande, 48 ha area with centroid at this point</dwc:locality>
    <dwc:decimalLatitude>-40.97467</dwc:decimalLatitude>
    <dwc:decimalLongitude>-71.0734</dwc:decimalLongitude>
    <dwc:geodeticDatum>WGS84</dwc:geodeticDatum>
    <dwc:coordinateUncertaintyInMeters>200</dwc:coordinateUncertaintyInMeters>
  </dcterms:Location>
```

Data formats: DNA data

- FASTA

```
>FIDIP1814-12|Oxyna parietina
---------------------------------------ACATTATATTTTATTTTTGGAGCTTGAGCAGGAATAATT
GGTACTTCTTTA---AGAATTTTAATTCGTACTGAATTAGGCCATCCAGGNTCATTAATTGGAAAT---GACCAAATT
TATAATGTTATTGTAACATCTCATGCATTTGTAATAATCTTTTTTATAGTTATACCAATTATAATTGGAGGATTCGGA
AATTGATTAGTTCCTCTTATA---TTAGGAGCCCCTGATATAGCTTTTCCACGAATAAACAATATAAGTTTTTGATTA
CTACCTCCTTCTCTTATCTTATTATTAGCCAGAAGAATAGTGGAAAATGGATCTGGAACAGGATGAACAATTTACCCT
CCCCTTTCATCTATTTCAGCTCATGCAGGATCATCTGTTGATTTA---ACAATTTTTTCATTACATTTAGCAGGAATT
TCTTCAATTTTAGGAGCAGTAAATTTTATTACAACAATTATTAACATACGATCAACAGGAATCACTTTTGATCGAATA
TCATTATTTATTTGAGCAGTTATTTTAACAGCTTTTTTACTTTTATTATCTCTTCCAGTTCTAGCAGGT---GCAATT
ACTATATTATTAACTGACCGAAATTTTAATACTTCATTTTTTGATCCTGCAGGTGGAGGAGATCCTATTTTATACCAA
CATTTA----------------------------------------------------------------------
```

- FASTQ (FASTA-with-quality)

```
@ FIDIP1814-12|Oxyna parietina
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

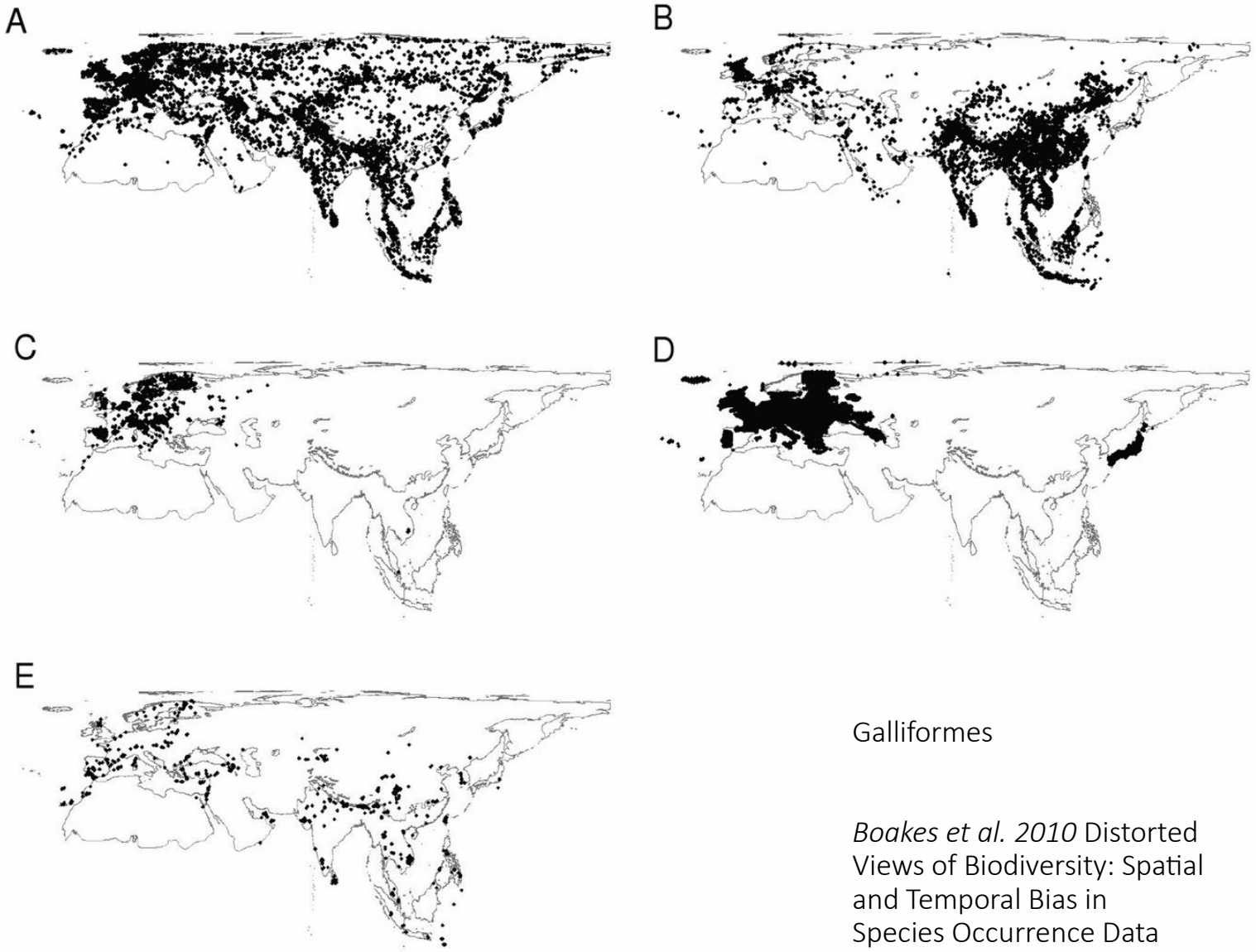
International Nucleotide Sequence Databases

```
XX
DR   MD5; 233dc548930beee1467d999bae8917e6.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..356
FT                   /organism="Minettia lupulina"
FT                   /organelle="mitochondrion"
FT                   /mol_type="genomic DNA"
FT                   /specimen_voucher="JSM2746"
FT                   /db_xref="taxon:768769"
FT   rRNA            <1..>356
FT                   /product="12S ribosomal RNA"
XX
SQ   Sequence 356 BP; 145 A; 27 C; 50 G; 134 T; 0 other;
     ttaaaatgta aaataaaaaa tttgagtagt attagatatg atcttgaaac ttaaaaaatt        60
     tggcggtatt ttagtctatt cagaggaacc tgttctataa tcgataatcc acgatggacc       120
     ttacttaaat ttgttaatca gtttatatac cgtcgttatt agaatatttt gtaaaaataa       180
     taattttcta taattttaat taaaatatat atcagatcaa ggtgtagctt atatttaaga       240
     agaaatgggt tacaataaat ttatttaaat ggatataaaa atgaaaaagt tattgaaagt       300
     ggatttgata gtaaaattat aaagattaat aatttgattt tagctctaaa atatgc           356
//
```

Issues with specimen data

- Always know your original source!
- Occurence data is extremely biased
  - Work with uniform(cough cough) subsets
  - Normalize as far as possible
- Big data is full of small mistakes

The spatial distribution of records from different sources. **A) museums**, B) literature, C) ringing, D) atlas, and E) website trip reports. https://doi.org/10.1371/journal.pbio.1000385.g002



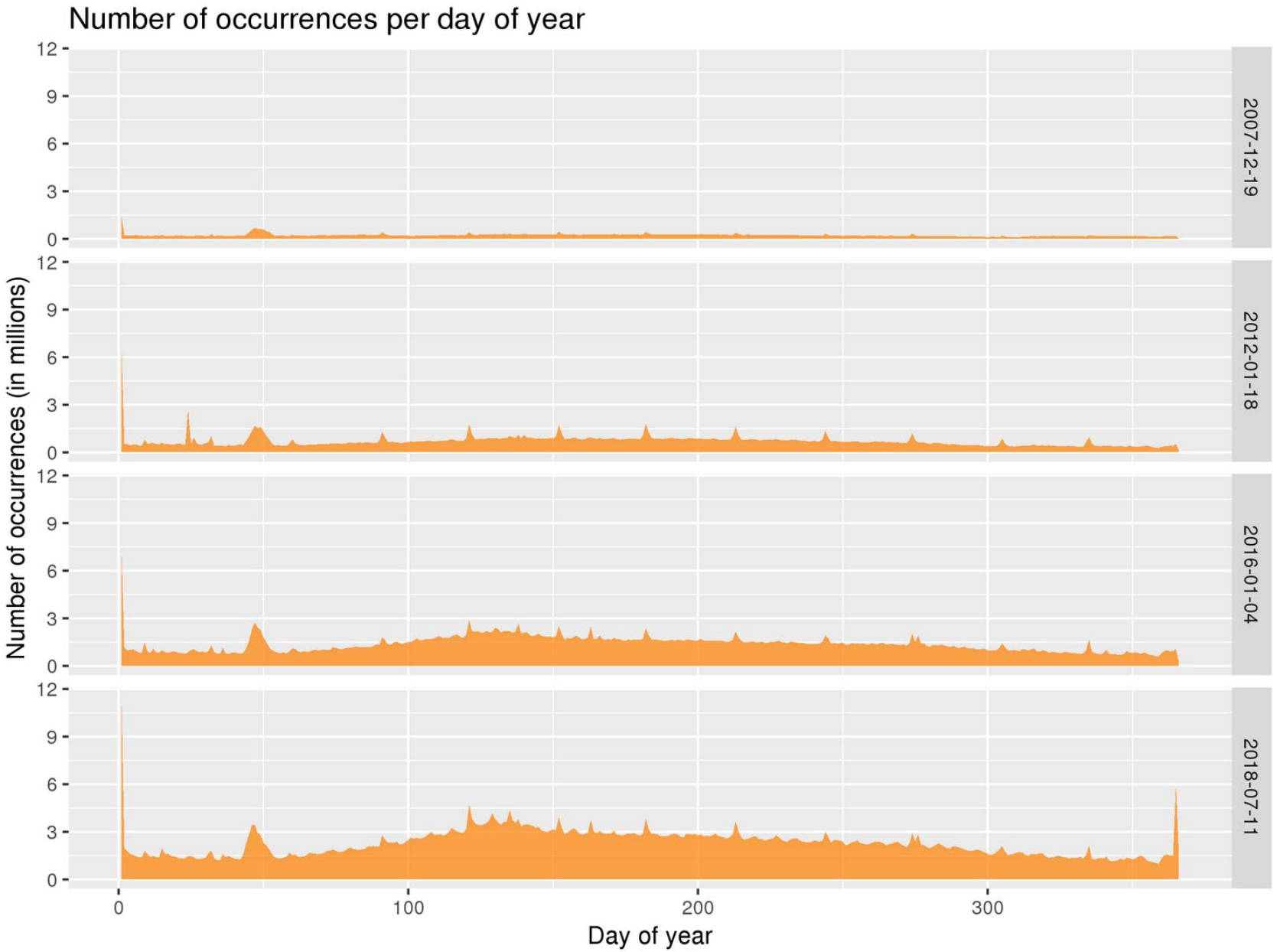
Galliformes

Boakes et al. 2010 Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data

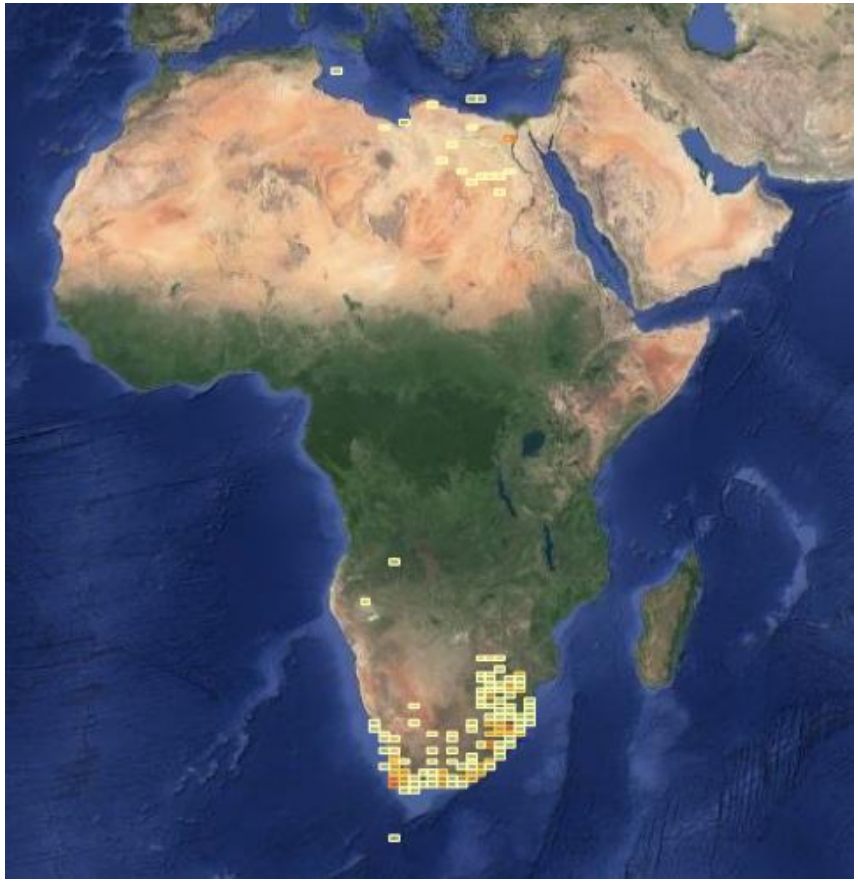
Boakes et al. 2010 Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data

- By collating a large historical database of ~170,000 records of species in the avian order **Galliformes, dating back over two centuries and covering Europe and Asia, we investigate patterns of spatial and temporal bias** in five sources of species distribution data: museum collections, scientific literature, ringing records, ornithological atlases, and website reports from "citizen scientists." **Museum data were found to provide the most comprehensive historical coverage of species' ranges but often proved extremely time-intensive to collect**. Literature records have increased in their number and coverage through time, whereas ringing, atlas, and website data are almost exclusively restricted to the last few decades. Geographically, our data were biased towards Western Europe and Southeast Asia. **Museums were the only data source to provide reasonably even spatial coverage across the entire study region**. In the last three decades, literature data have become increasingly focussed towards threatened species and protected areas, and currently no source is providing reliable baseline information—a role once filled by museum collections.

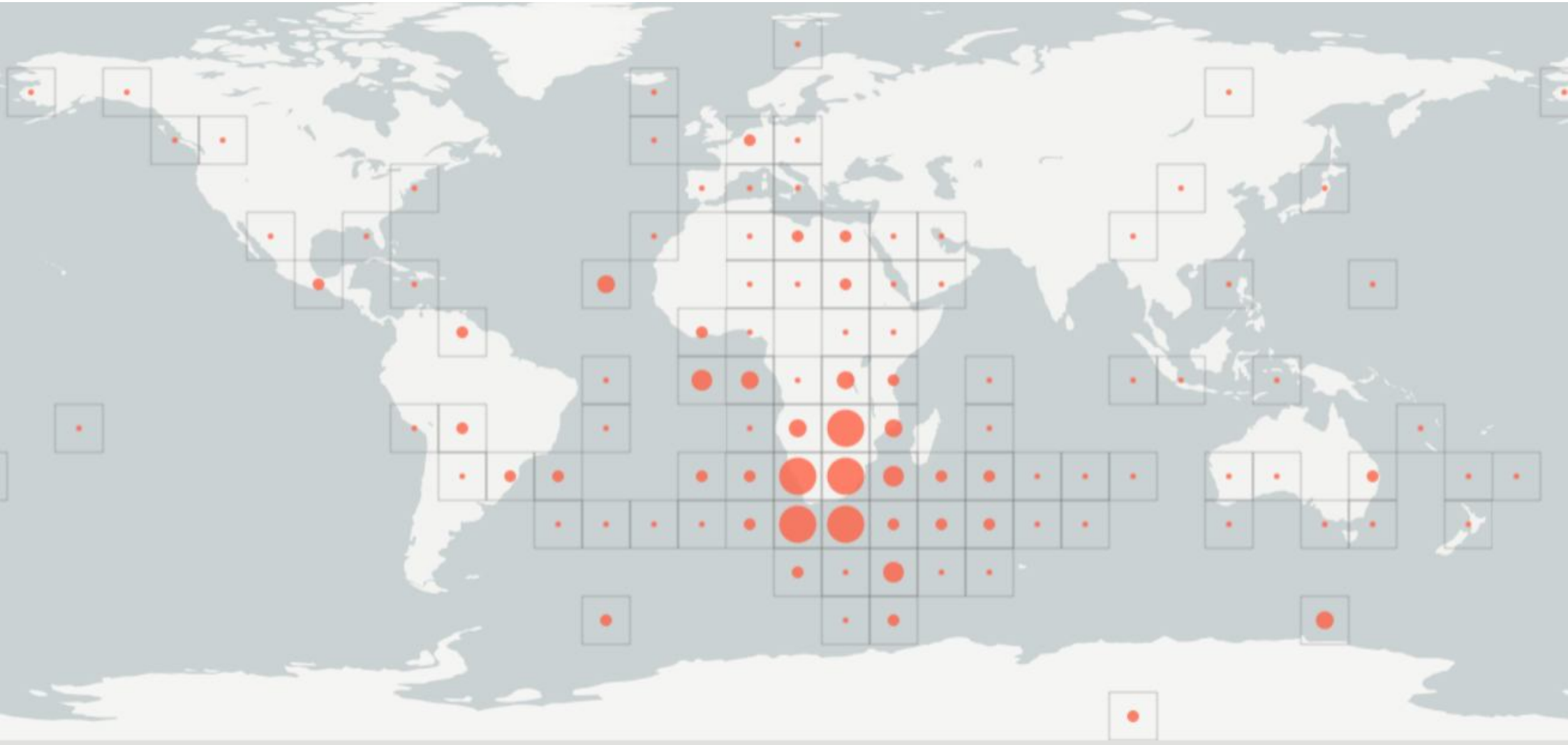
GBIF


Number of occurrences per day of year
Number of occurrences (in millions)
12
9
6
3
0
2007-12-19
2012-01-18
2016-01-04
2018-07-11
0
100
200
300
Day of year

Laji.fi country:”South Africa”



GBIF ”South Africa”



The one thing to remember from this part!

All data is biased. Especially occurence data.