SDMS WG Training #4:
How to create Darwin Core Archives for biological data

NINA
Norwegian Institute for Nature Research

About

Who I am

- Employed at the Norwegian Institute for Nature Research (NINA)
- Part of the environmental data team
- Technical support for researchers
- Part of SIOS SDMS WG

What I do

- GIS development (python/web)
- Remote sensing
- Building Data infrastructures
- FAIR data management
- Technical training

Matteo De Stefano
GIS developer
matteo.destefano@nina.no

The Norwegian Institute for Nature Research (NINA)

Who we are

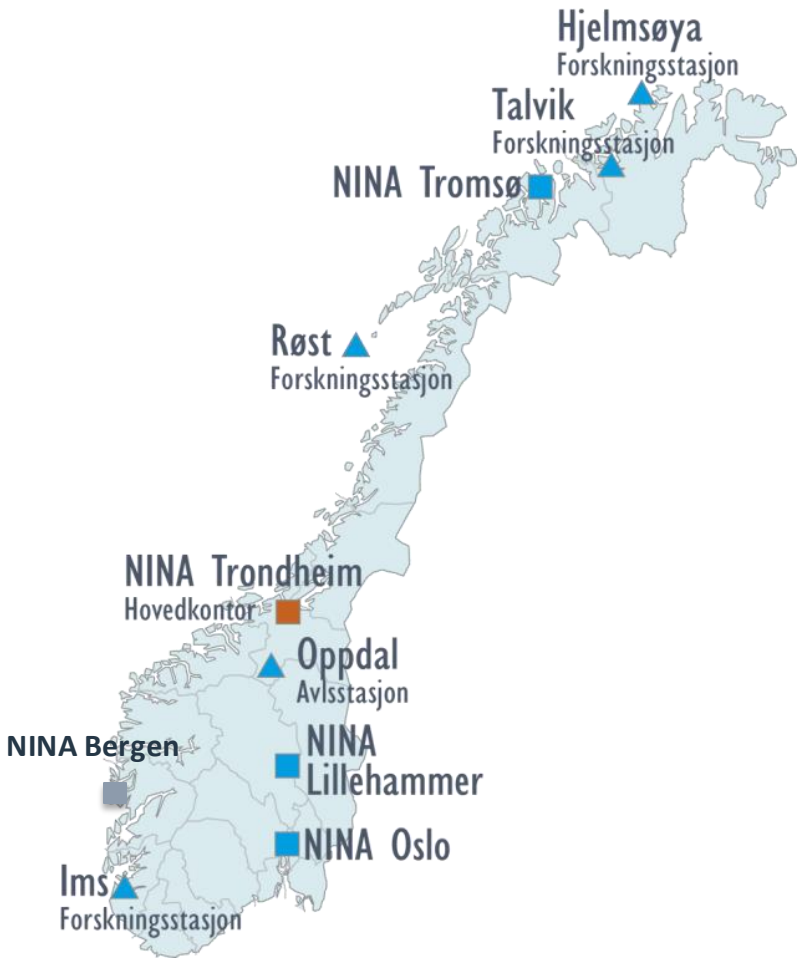
- Independent research foundation
- Norway's leading institution for applied ecological research
  -> Science <-> Policy interface
- ~260 employes (2021)
- Projects Worldwide

What we do

- Ecological and social research
- Ecological long-term monitoring
- Research in the field
- Statistical and Geospatial modelling & analysis
- Counselling and guidance
- Capacity building
- Impact assessments

NINAs motto is:

- Cooperation and expertise for a sustainable future



Training outline

- Introduction to Darwin Core (DwC)
- IIntroduction to GBIF
- ntroduction to Darwin Core Archives (DwC-A)
- Exploring the GBIF data portal
- Accessing GBIF data
- Introduction of IPT – tool for DwC-A publishing to GBIF
- Exercise publishing DwC data with IPT
- Few final comments
- QA and Wiki links

Primary Biodiversity data vs «ecological data»

- **Primary biodiversity data** can be defined as data that document the occurrence of a species (or higher taxa).
- Such data are relatively simple, consisting of a timestamp, location and the species (taxa) name
- Sometimes referred to simply as:

occurrence data

- **Ecological data** containes richer information than primary biodiversity data
- Most basic extention is information about sampling protocol and study design
- Can also include additional descriptors, such as species traits, relationship between occurence records (mother-offspring, predation event, recapture of the same individual etc)
- Also include data captured using modern tools
- Biologger data, eDNA data, cameratrap data

Why Darwin Core?

Natural history collections, environmental monitoring programmes, recording societies, citizen scientist projects and others all hold valuable data on the world's **biodiversity**. They collect and manage their information in many different systems and environments, and vary widely, depending on what kind of details are captured and stored for any individual record.

The **Darwin Core Standard** offers a stable, straightforward and flexible framework for compiling biodiversity data from varied and variable sources. Originally developed by the Biodiversity Information Standards (TDWG) community, Darwin Core is an evolving community-developed biodiversity data standard.

What is Darwin Core?

The Darwin Core is a vocabulary **standard** (basically a glossary of terms) intended to facilitate the sharing of information about biological diversity. It is an extension of the Dublin Core standard.

Darwin Core - TDWG

What is Darwin Core, and why does it matter? (gbif.org)

Darwin Core Hour 01 - Intro to Darwin Core - Google Slides

Darwin Core terms?

Darwin Core Intro

What are Darwin Core terms?

Vocabularies

Classes

– categories, groupings, tables

- Event, Location, GeologicalContext, Identification, Taxon

Properties

– attributes, fields, columns, keys, data element

- eventDate, country, bed, identifiedBy, genus

Values

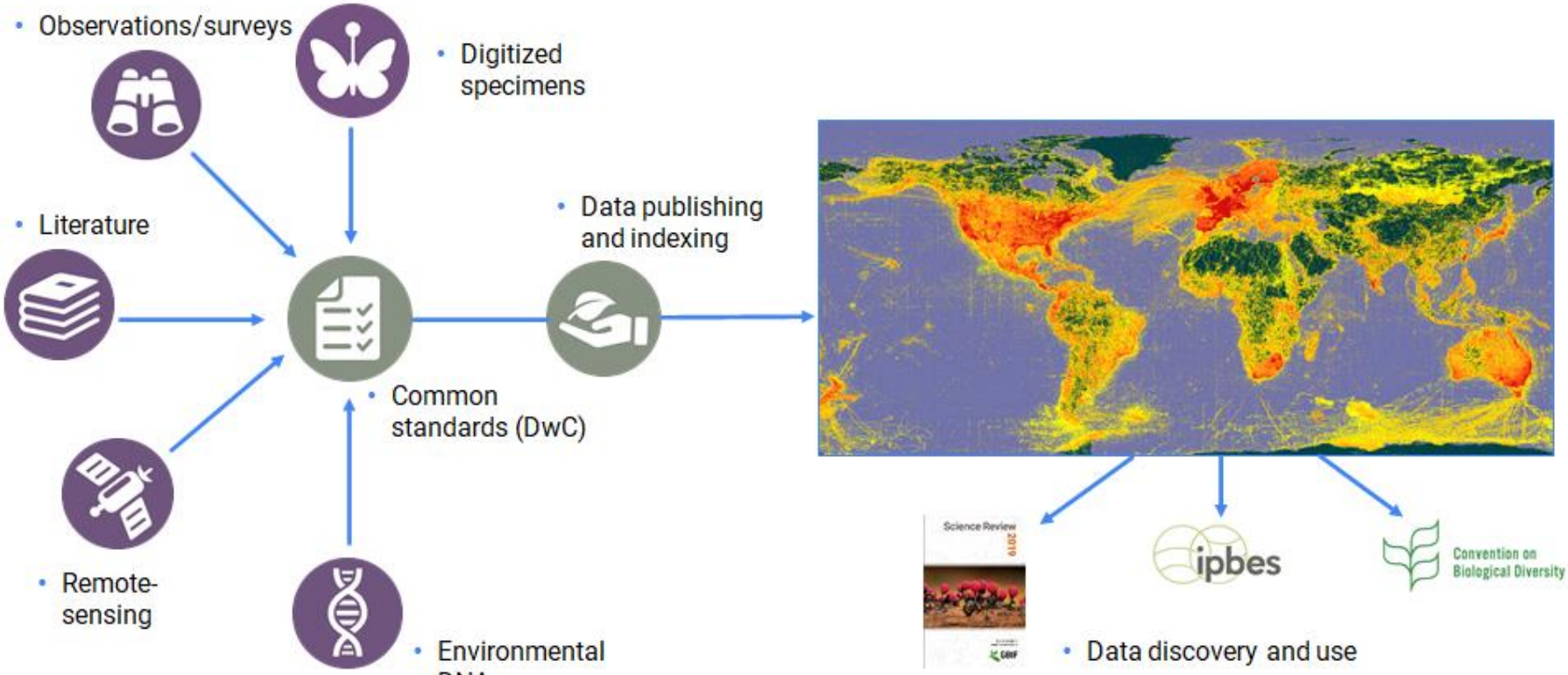
– contents of terms, data values, terms from controlled vocabularies

- 2017-02-07, Argentina, Olduvai Bed IV, JL Patton, Mallophora

Darwin Core Hour 01 - Intro to Darwin Core - Google Slides

Why GBIF?

A WINDOW ON EVIDENCE ABOUT WHERE SPECIES HAVE LIVED, AND WHEN



What is GBIF?

Intergovernmental network and research infrastructure

Provides anyone, anywhere, free and open access to data about all types of life on Earth

Voluntary collaboration through Memorandum of Understanding

Participant nodes, Secretariat in Copenhagen, DK

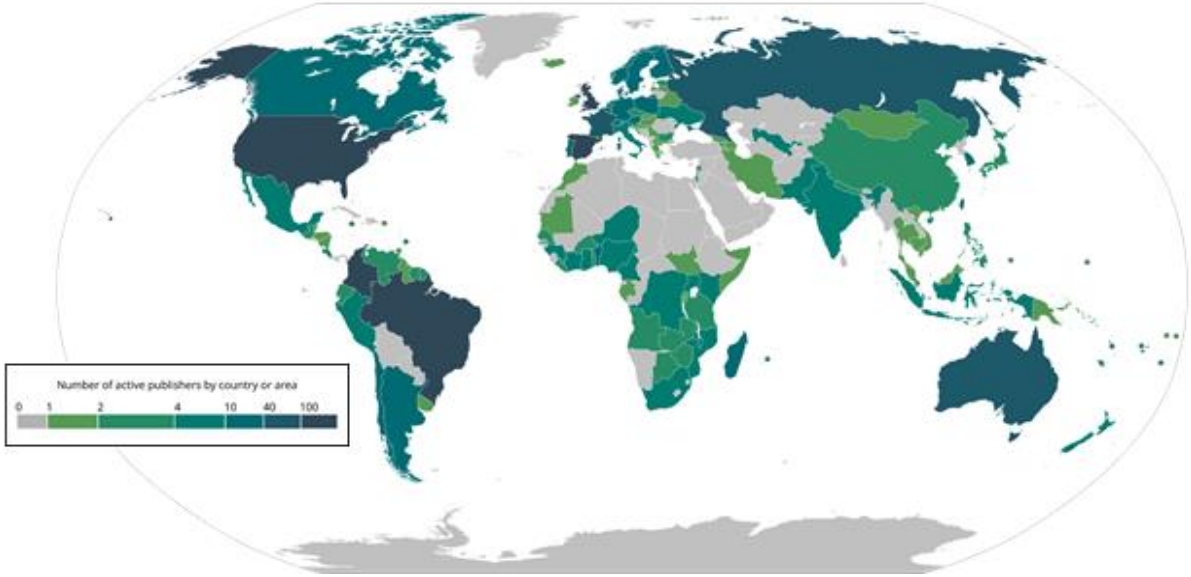
What is GBIF?

The Global Biodiversity Information Facility (GBIF) was established as a **global megascience initiative** to address one of the great challenges of the 21st century – harnessing knowledge of the Earth's biological diversity. It is an international network and data infrastructure funded by the world's governments and aimed at providing anyone, anywhere, open access to data about all types of life on Earth.

GBIF envisions a world in which biodiversity information is freely and universally available for science, society, and a sustainable future. GBIF's mission is to be the foremost global resource for biodiversity information, and engender smart solutions for environmental and human well-being. To achieve this mission, GBIF encourages a wide variety of data publishers across the globe to discover and publish data through its network. GBIF provides data-holding institutions around the world with common standards and open-source tools that enable them to share information about where and when species have been recorded

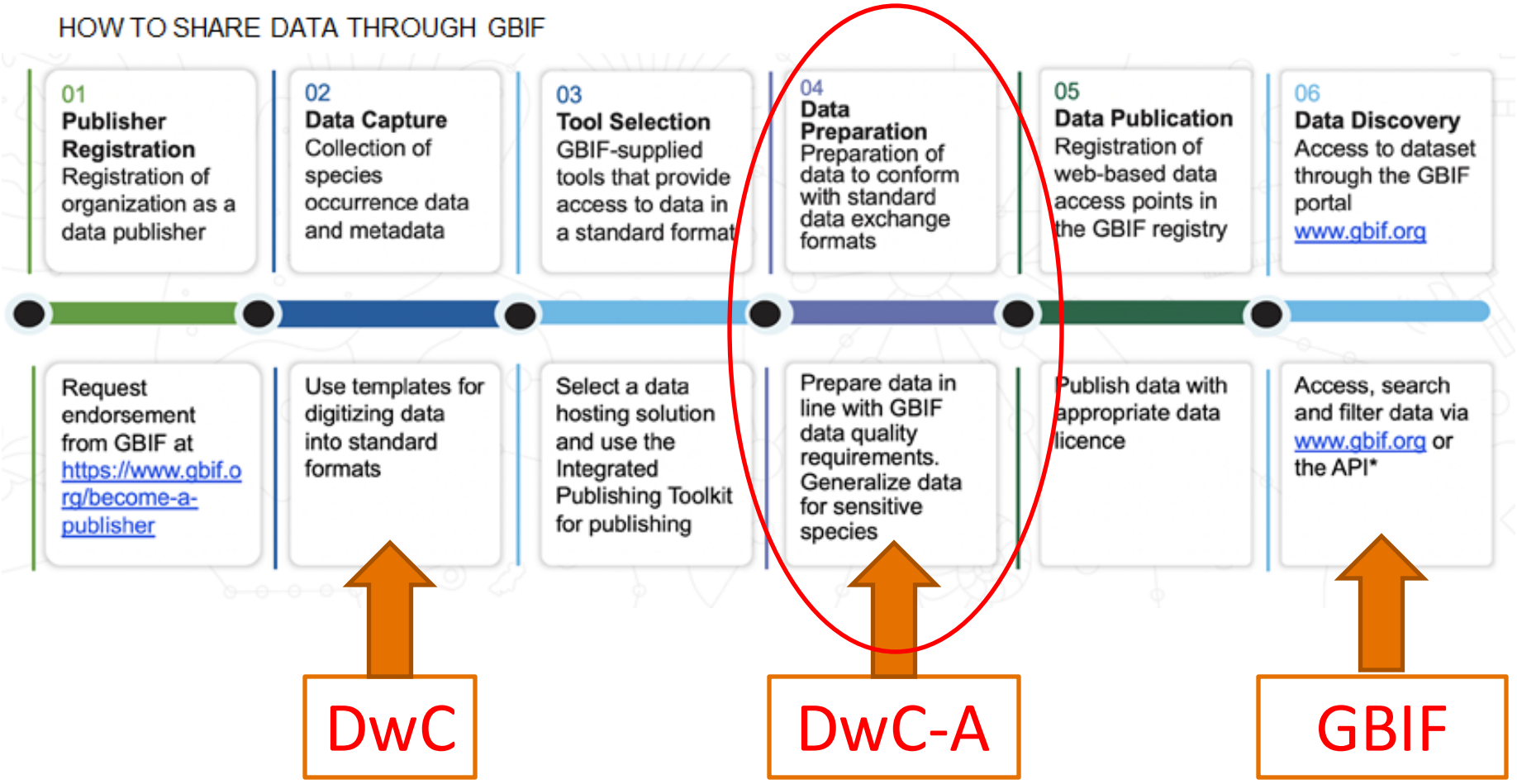
GBIF sources?

WHO SHARES DATA THROUGH GBIF?

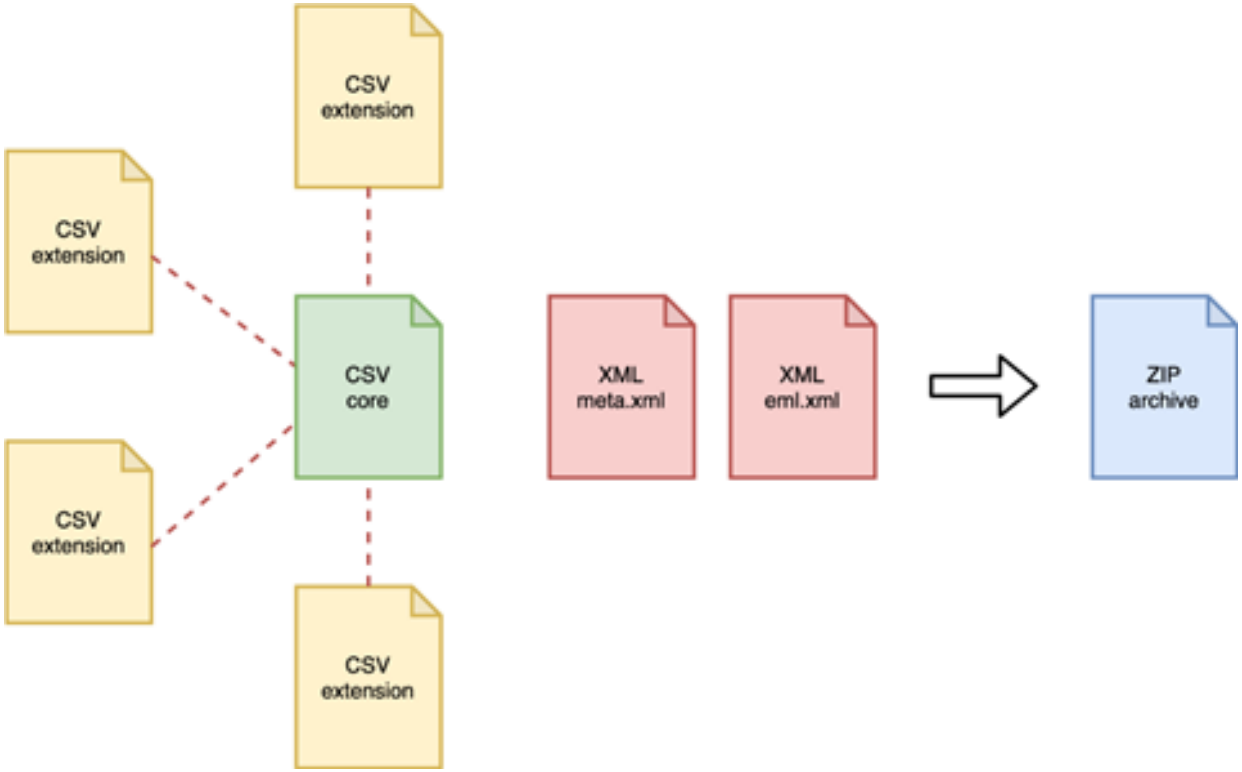


- Museums, research institutions
- Government agencies
- Citizen science networks
- NGOs
- Thematic data networks (e.g. marine, agrobiodiversity, DNA)
- Private companies, consultants

Why Darwin Core Archive?



Darwin Core Archive – Star model



The DwC-A is the main format used to exchange biodiversity data, for example sharing to GBIF

What is a Darwin Core Archive?

In practice, using Darwin Core revolves around a standard file format, the Darwin Core Archive (DwC-A). This compact package (a ZIP file) contains interconnected text files and enables data publishers to share their data using a common terminology (DwC).

When preparing a Darwin Core Archive version from their source data, publishers restructure and streamline information into a small but structured collection of text files. One of these files is the ‘core’ file and holds a separate record for each of the items included in the archive. Other ‘extension’ files may also be included. These contain additional information linked to the records in the core file. Extension files allow the archive to model many-to-one relationships.

Which cores at least?

Depending on how much information the source data contains—and how much they wish to share—publishers can create a Darwin Core Archive with one of three cores:

- a Taxon core, which lists a set of species, typically coming from the same region or sharing common characteristics
- an Occurrence core, which lists a set of times and locations at which particular species have been recorded
- an Event core, which lists field studies (including the protocols used, the sample size, and the location for each).

In the case of an Event core, one extension file usually contains the elements displayed in an Occurrence core, which enables the inclusion of many observation records as part of a single planned field study.

Meta information?

Finally, each archive contains two more pieces that help both machines and humans interpreting the data. The first, a descriptor file (meta.xml), defines the precise structure and relationships between the core and any extensions. The second, a complementary metadata file, describes the datasets contained in the archive, typically in Ecological Metadata Language (EML.xml)—though the GBIF's Integrated Publishing Toolkit produces these files automatically for its users.

About EML: Ecological Metadata Language(EML)(ecoinformatics.org)

Exploring GBIF

- link to GBIF

- Global Biodiversity Information Facility (github.com)

Accessing GBIF data

- GBIF REST API
- rgbif
- pygbif Documentation
- BelgianBiodiversityPlatform/qgis-gbif-api: GBIF Occurrences is a QGIS plugin to directly import occurrences from the GBIF API (github.com)

How to publish data to GBIF?

- Quick guide to publishing data through GBIF.org

- There are some options to start publishing:
  - Decide that **your institution** wants to be a **data publisher**: get a GBIF endorsement
  - Refer to an **existing publishing institution:** dataHostingCentres · gbif/ipt Wiki (github.com)
  - in Norway:
    - National GBIF Node - GBIF Norway - GBIF Norway - Global Biodiversity Information Facility
    - Artsdatabanken - Kunnskapsbank for naturmangfold
    - Promoting FAIR data management (livingnorway.no)

Intergrated publishing toolkit - IPT

IPT2ManualNotes.wiki · gbif/ipt Wiki (github.com)

- If choose to use DwC-A, Prepare and map data to Darwin Core:
  - Use Excel templates based on DwC fields **OR** Set up data in a supported database (DwC column names)
  - Use IPT to create or upload a DwC-A, and publish it

See also: howToPublish · gbif/ipt Wiki (github.com)

Some examples of IPT

- IPT (nina.no) - NINA IPT
- IPT (gbif.no) - IPT of the Norwegian GBIF node
- IPT (artsdatabanken.no) - Norwegian Biodiversity Information Centre

Exercise – publish a Dwc-A in IPT

- IPT (gbif.no) - DEMO IPT

- GBIF (gbif-uat.org) - DEMO GBIF

Find Sample data here:
occurrenceData · gbif/ipt Wiki (github.com)

Some comments

Is DwC a metadata standard? It derives from Dublin Core, a metadata standard used for describing many types of resources. However, it is more a glossary of terms for structuring a subset of your data to be shared publicly. It is not easy to separate clearly the concepts of data and metadata, due to the nature of Occurrence biodiversity data.

There is always some data loss. The purpose is to select part of the data suitable for sharing according to the standard, not to be structured properly and completely.

GBIF is not storing the DwC-A, just reading what is required, and mirroring. IPT instances are considered the true data repositories, and should be managed as permanent infrastructures

DwC is not the only standard in biodiversity. Tdwg .

Resources

- DwC Questions and Answers:
- tdwg/dwc-qa: Public question and answer site for discussions about Darwin Core (github.com)
- DwC Wiki:
- Home · tdwg/dwc-qa Wiki (github.com)
- GBIF data mobilization course:
- Biodiversity Data Mobilization Course (gbif-uat.org)

Norwegian GBIF Statistics

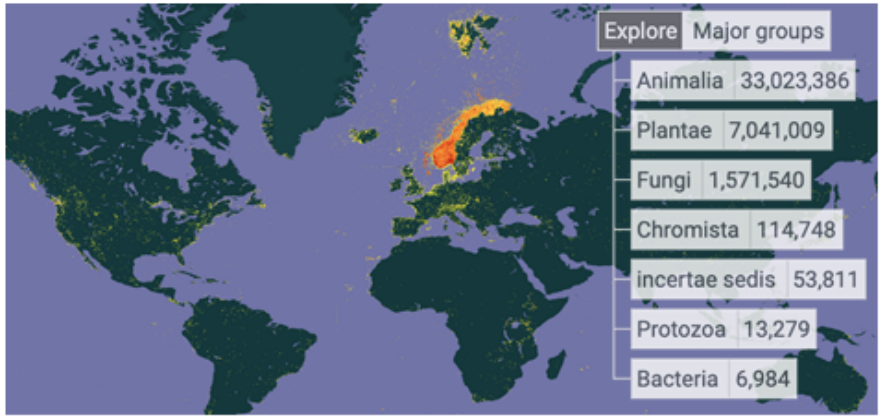
BY THE NUMBERS | 15 MARCH 2021 -- NORWAY

Species occurrence records *(published from)*

41 824 982

Datasets *(published from)*

307



Publishers (from Norway)

38

Peer-review papers using data (co-author from Norway

131

Norwegian GBIF Node

GBIF Norway Node Staff

Vidar Bakken
Data Outreach Officer
Senior Engineer (30%)
vidar.bakken@usit.uio.no

Rukaya Johaadien
Data Manager
Head Engineer
rukayasj@uio.no

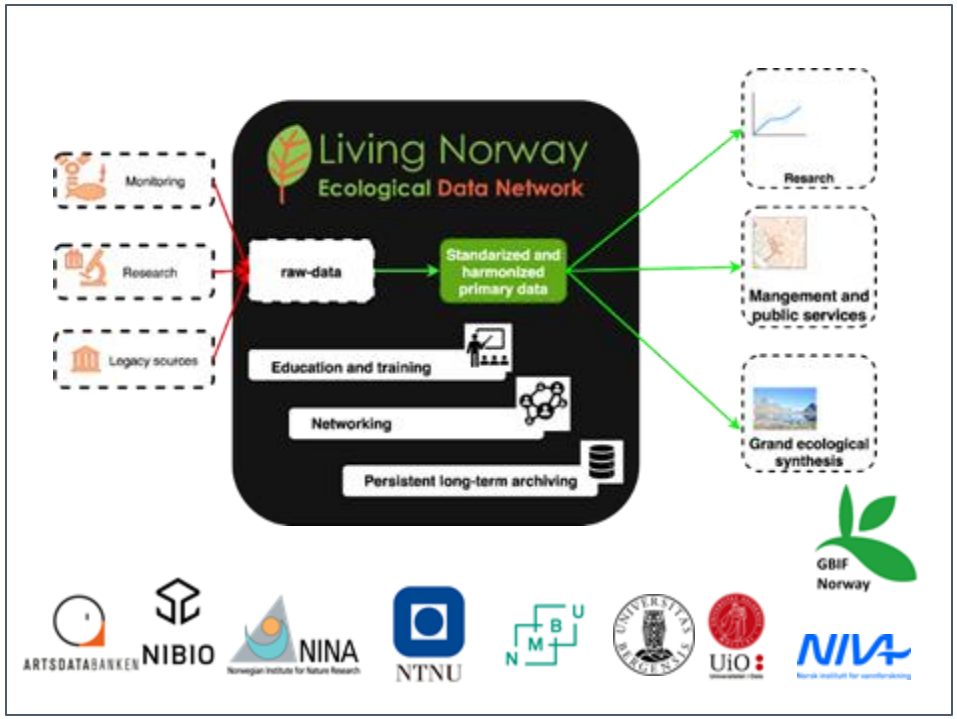
Dag Endresen
Node Manager
Chief Engineer
dag.endresen@nhm.uio.no

Questions or need advice? General email: helpdesk@gbif.no

Next training

- *In the Nansen Legacy project, a template generator, based on Darwin Core, is used to generate standardised templates that can more easily be understood and converted to DwCA. It was developed following discussions with Dag Endresen at GBIF Norway and is a development of a tool one of their employees started on. This is available for use by the broader scientific community through SIOS. Luke Marsden will talk about this on* 22nd April

Living Norway – Ecological Data Network



[Promoting FAIR data management (livingnorway.no)]

- Provide the society with documented and harmonised data from ecological sciences
- Mobilize data from Norwegian research institutions
- Contribute to development of open standards
- Operate e-infrastructure
- Educate and train students and researchers in modern data management, including research ethics related to open data
- National and international networking

Thank you!

- Contact: matteo.destefano@nina.no