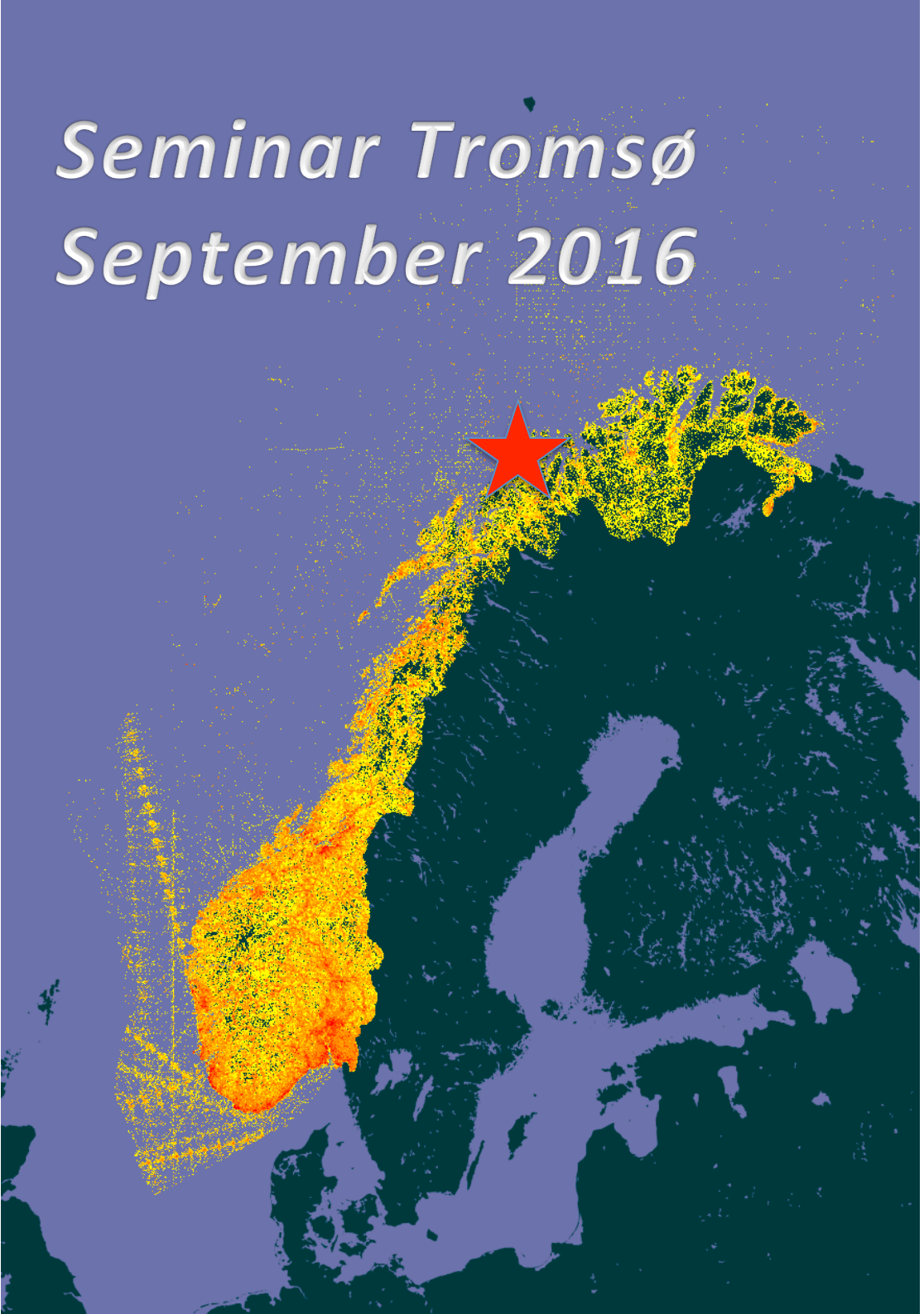
Seminar Tromsø
September 2016



ARTSDATABANKEN

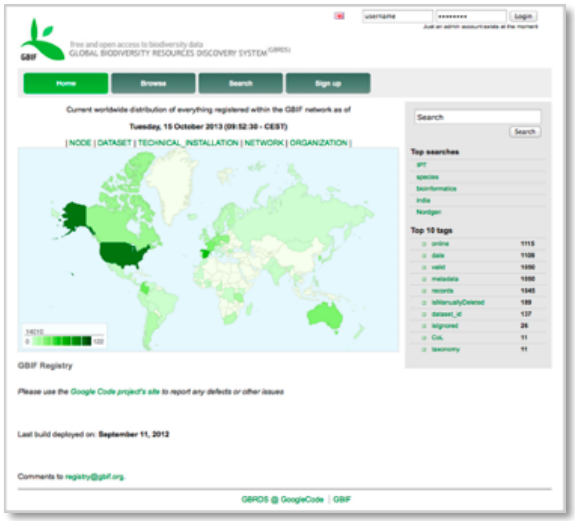
GBIF Norway

UiO : Natural History Museum
University of Oslo

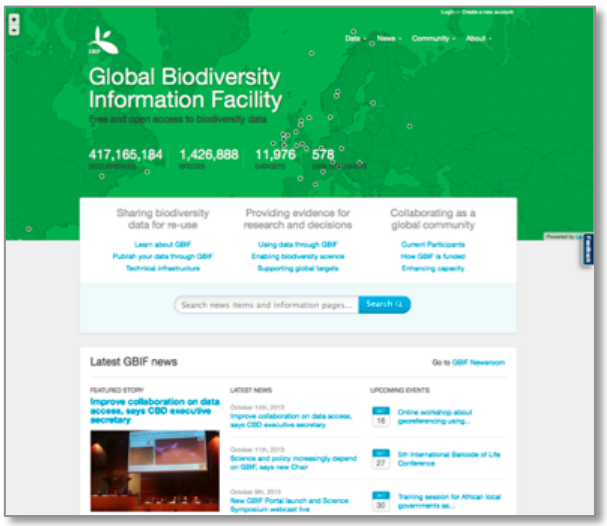
GBIF provides a data publishing infrastructure

GBIF provides a data discovery system

that is dependent on resolvable stable identifiers for efficient functionality



global registry



data portal

GBIF and GEO

Intergovernmental group on earth observations

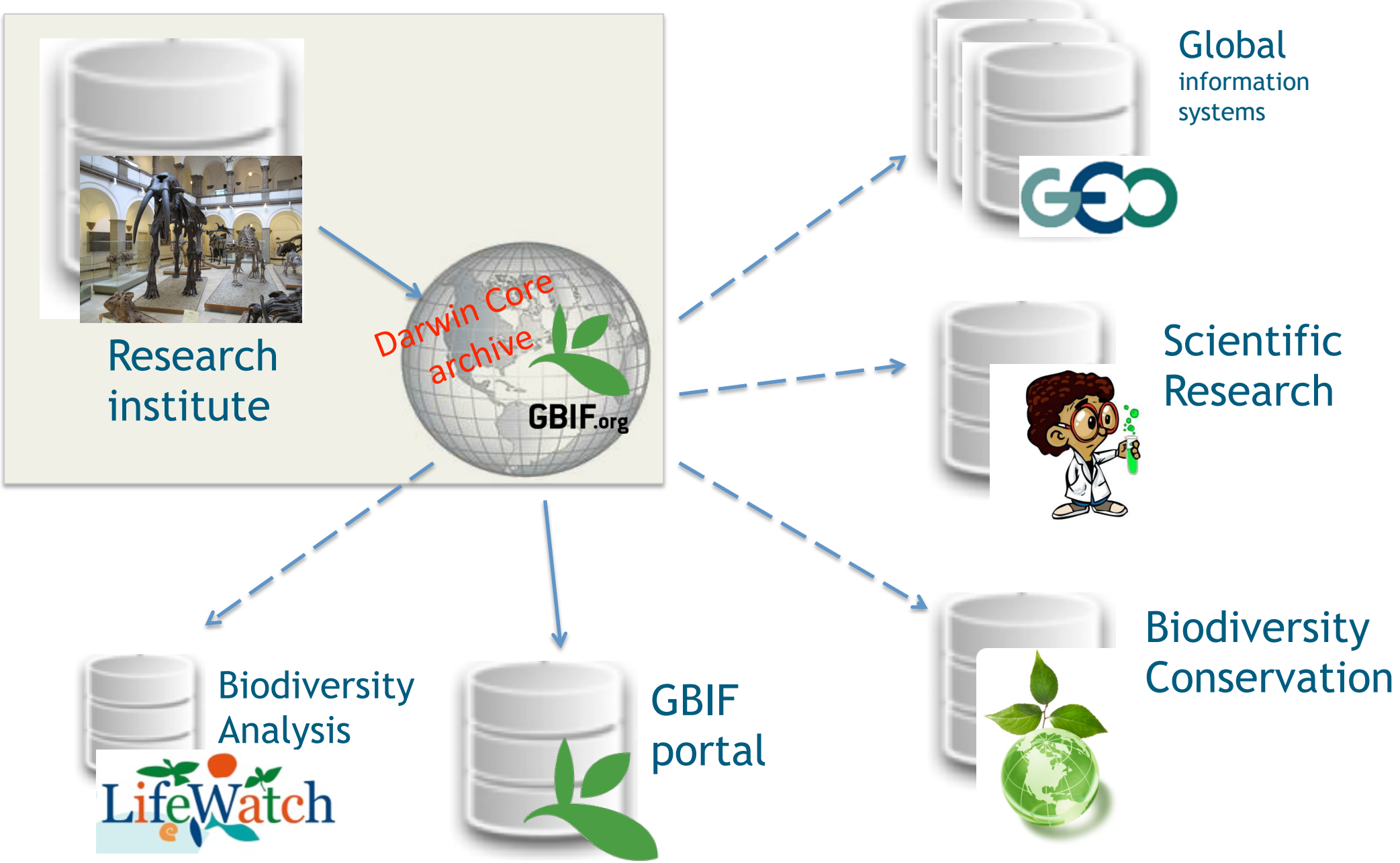
GEO BON

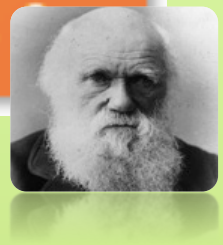
Biodiversity observation network

Data Integration & Interoperability

GBIF provides the infrastructure delivering species occurrence data in GEO.

MULTIPLE-PURPOSE DATA SERVICES





Biodiversity Information Standards
TDWG

GBIF.org

Darwin Core

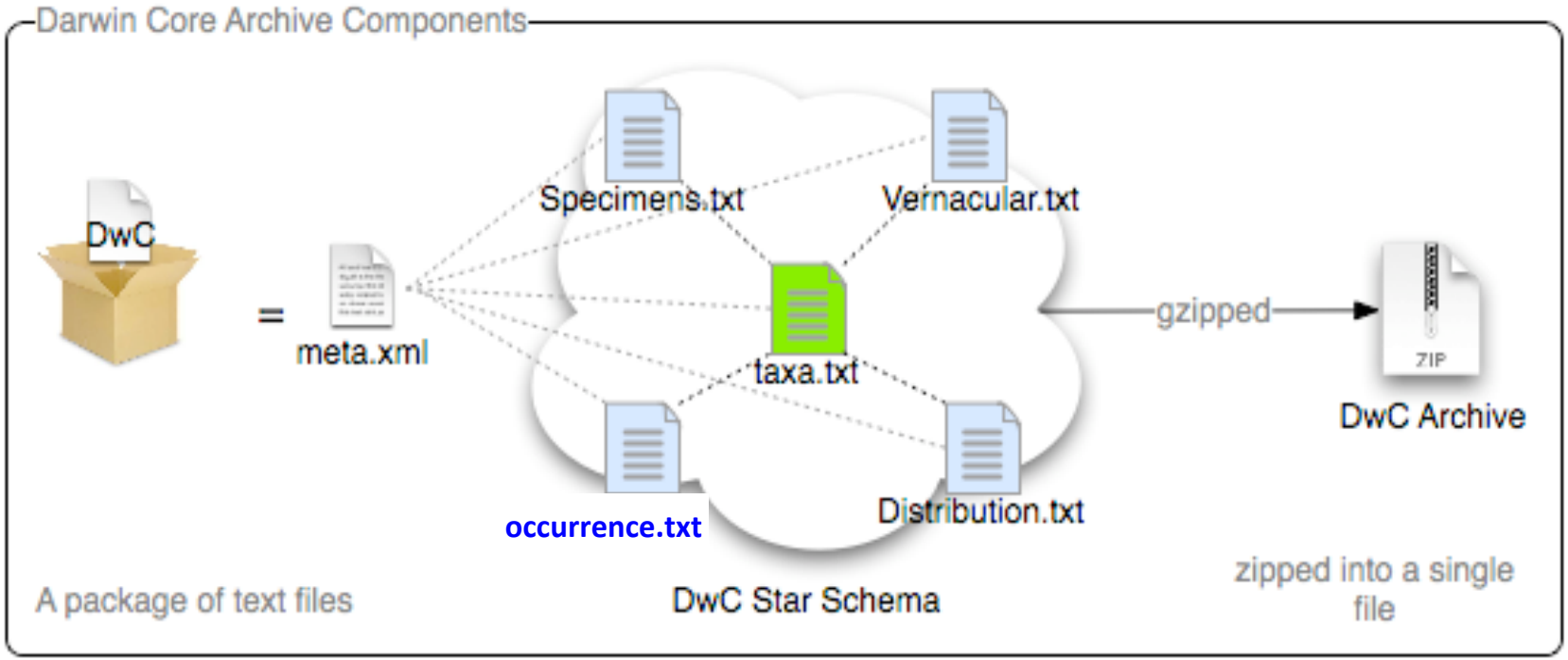
data exchange standard

Darwin Core – a vocabulary of terms

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, De Giovanni R, Robertson T, and Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE* 7(1): e29715. (doi:10.1371/journal.pone.0029715)

DARWIN CORE ARCHIVE (DWC-A)

- DwC-A publish DwC records including terms from DwC-A extensions.
- Simple text based format.
- Zipped single file archive.



DATA STANDARDS

Biodiversity Information Standards TDWG

ABCD Access to Biological Collection Data (2005)

DwC Darwin Core (2009)

AC Audubon Core Multimedia Resources Metadata Schema (2013)

NCD Natural Collection Descriptions (Draft 2008)

EML Ecological Metadata Language (Ecological Society of America)

Publish your biodiversity data with GBIF

PUBLISH DATA IN GBIF

Step 1: data holding research institutes seek endorsement as an approved data publisher.

Step 2: datasets are identified and converted to standard Darwin Core format.

Step 3: datasets can be published directly from the data node and/or with the assistance from a national GBIF node.

Citizen science data platforms also publish in GBIF.

The GBIF Integrated data Publishing Toolkit (IPT)

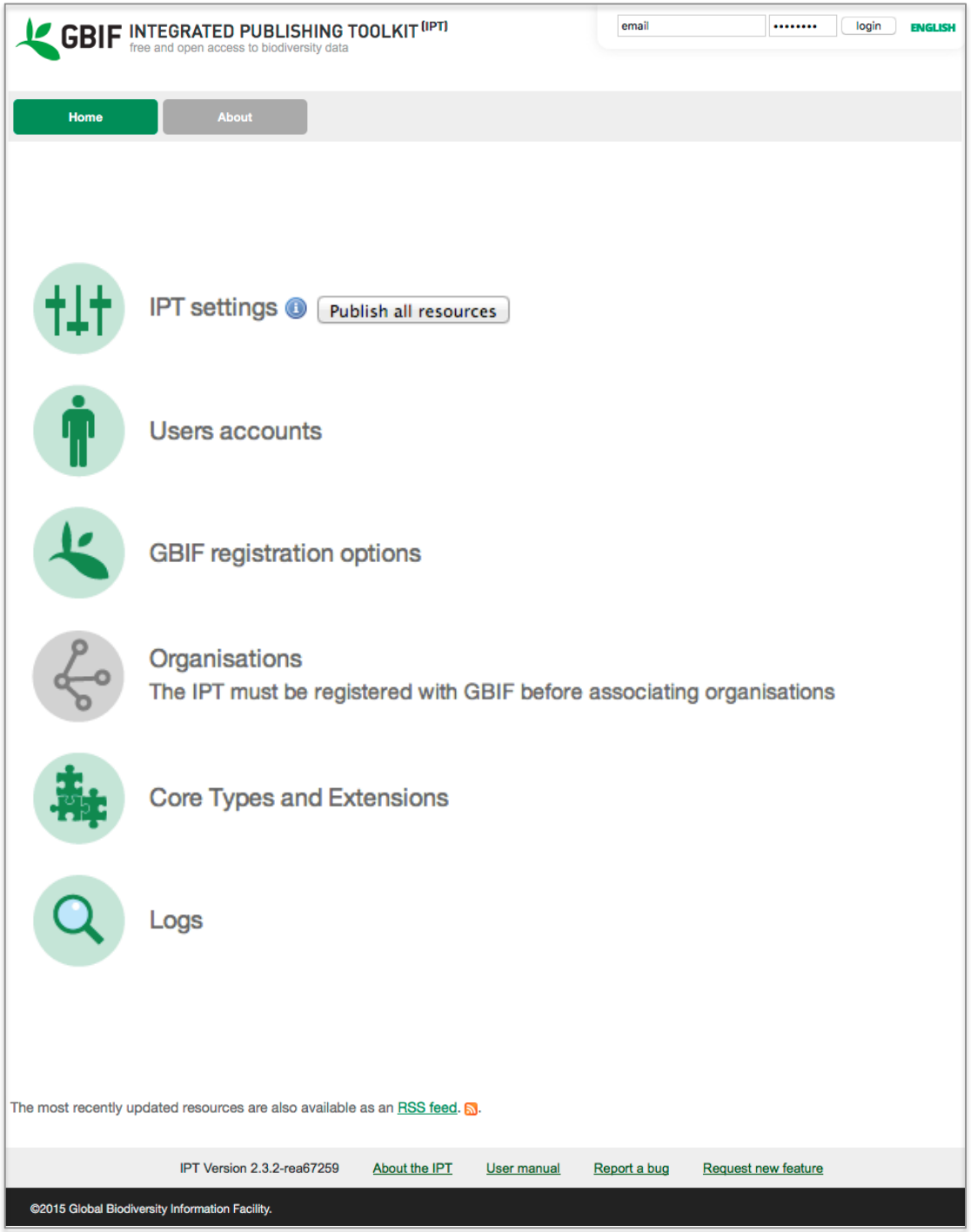
is a free open source software tool written in Java that is used to publish and share biodiversity datasets through the GBIF network.

http://www.gbif.org/ipt

IPT User Manual:

https://github.com/gbif/ipt/wiki/IPT2ManualNotes.wiki

Robertson T, Döring M, Guralnick R, Bloom D, Wieczorek J, Braak K, Otegui J, Russell L, Desmet P (2014). The GBIF integrated publishing toolkit: Facilitating the efficient publishing of biodiversity data on the internet. PLoS One 9(8). doi:10.1371/journal.pone.0102623



Artsobservasjoner
Rapportsystem for arter

More than 17 million Norwegian occurrence records published in GBIF

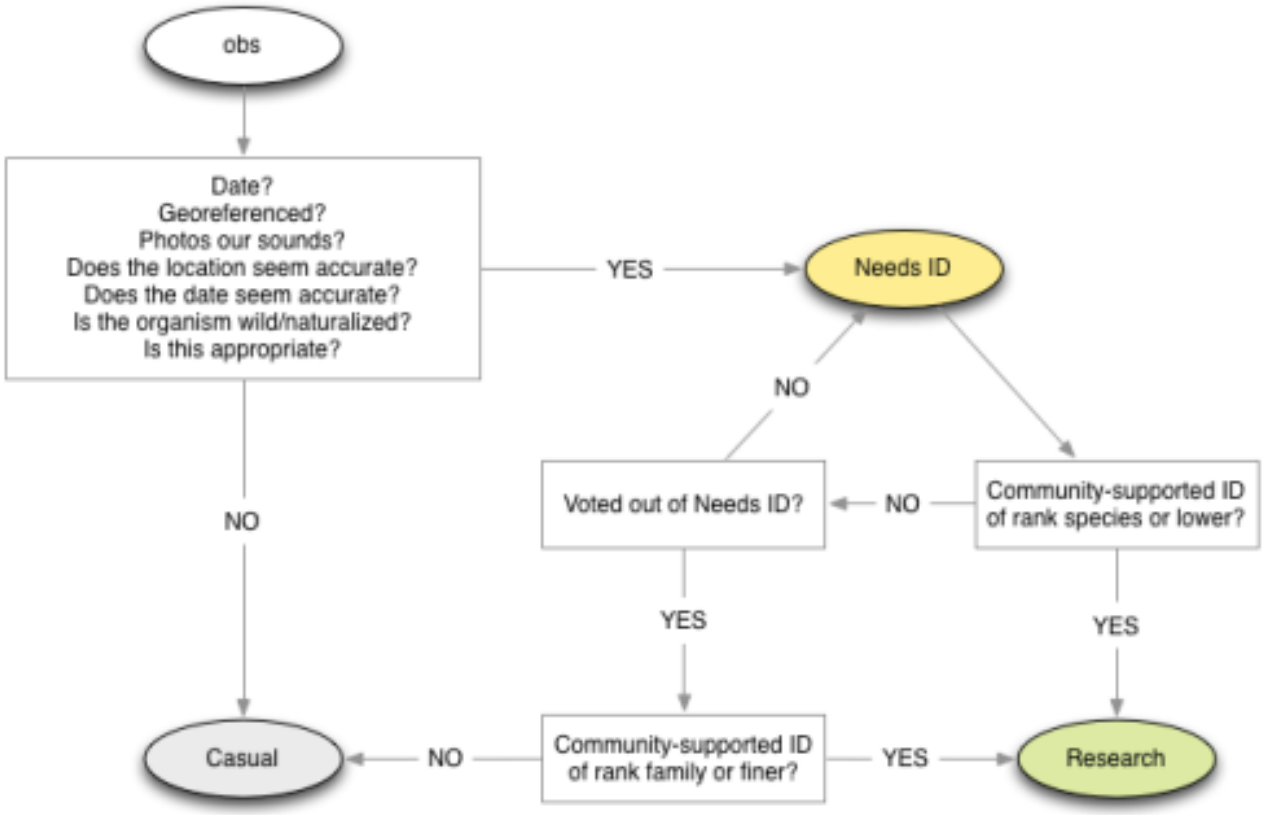
iNaturalist

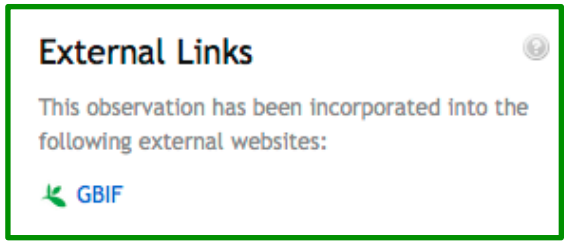
CALIFORNIA ACADEMY OF SCIENCES

Research grade: an observation must have **media**, **coordinates**, a **date**, and **pass quality metrics**, but now the **community ID** must be finer than family.

Needs ID: Any observation that could become “Research” grade but *needs more identifications*.

Casual: Any observation that cannot become “Research” grade.



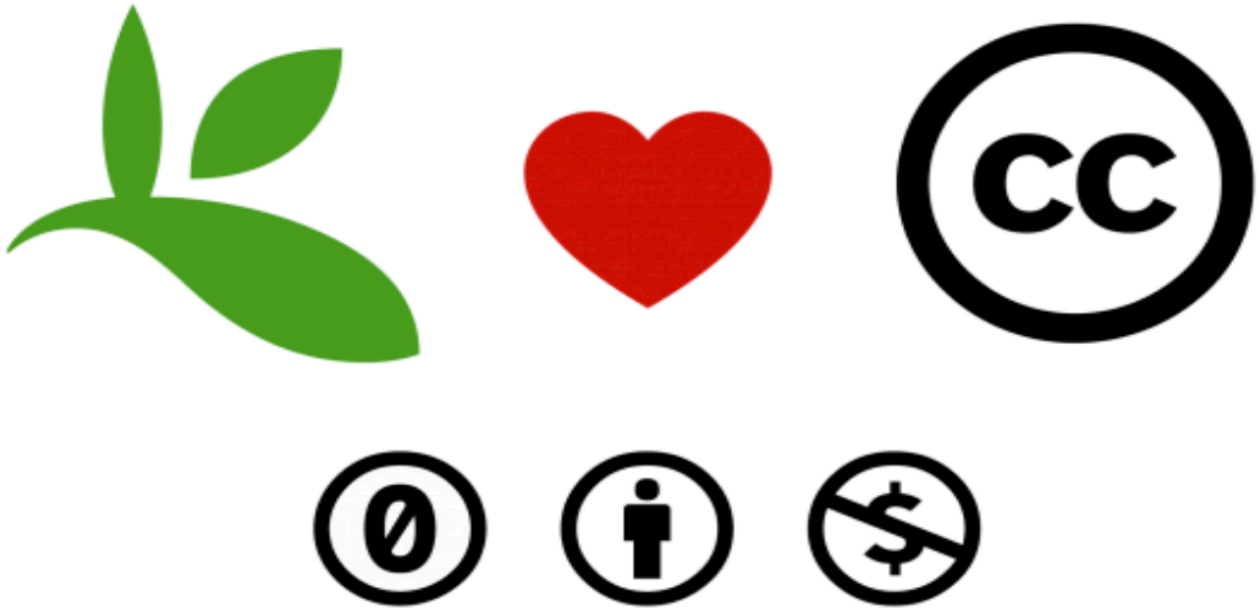


Data publishing guidelines

http://www.gbif.org/resources?f[0]=gr_purpose%3A955

Data license

Licensing milestone for data access in GBIF.org



Share Share Tweet G+1

PUBLICATION DATE
August 18th, 2016

LAST UPDATED
August 18th, 2016

TAGS

- GBIF.org
- GBIF network
- Data holders
- Data users
- GBIF.org
- Standards & protocols
- Occurrence data
- Data curation & quality
- Data publishing
- Data access

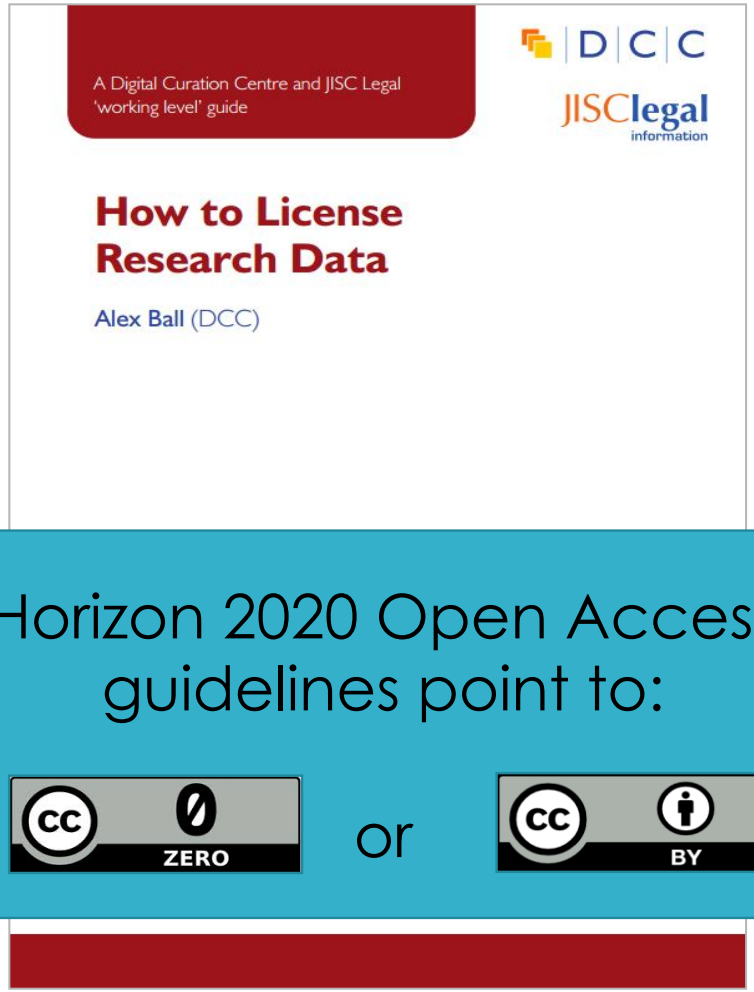
All species occurrence datasets on GBIF.org now carry standardized licences, giving data users and publishers greater clarity and enabling them to proceed with confidence about the terms of use for data accessed through the GBIF network.

LICENSING FOR DATA PUBLISHED THROUGH GBIF

GBIF Governing Board established in 2014 GBIF support for three licenses:

- **CC0**, under which data are made available for any use without restriction or particular requirements on the part of users *CC0 is strongly recommended whenever possible*
- **CC-BY**, under which data are made available for any use provided that attribution is appropriately given for the sources of data used, in the manner specified by the owner
- **CC-BY-NC**, under which data are made available for any use provided that attribution is appropriately given and provided the use is not for commercial purposes

LICENSING RESEARCH DATA



Horizon 2020 Open Access guidelines point to:

CC 0 ZERO or CC BY

This DCC guide outlines the pros and cons of each approach and gives practical advice on how to implement your licence

CREATIVE COMMONS LIMITATIONS

NC Non-Commercial
What counts as commercial?

ND No Derivatives
Severely restricts use

These clauses are not open licenses

www.dcc.ac.uk/resources/how-guides/license-research-data

Slide source: OpenAIRE & EUDAT, CC-BY-4.0, 2013

H2020 - OPEN DATA BY DEFAULT FROM 2017

Open Access EC
@OpenAccessEC

Official! From 2017 no more pilot: #open #research #data will become the rule (with opt out) ec.europa.eu/digital-single …

The Commission will make open research data the default option, while ensuring opt-outs, for all new projects of the Horizon 2020 programme.	As of 2017

RETWEETS 157 LIKES 73

Slide source: OpenAIRE & EUDAT, CC-BY-4.0, 2013

Why publish data & why make a data management plan

DATA MANAGEMENT PLAN

- Making your data available to others ensures that your research is truly **reproducible**.
- Managing your research data **saves you time** because it ensures that you and others in your collaboration will be able to **find, understand, and use the data**.
- Sharing your research data enables wider **dissemination** of your work.
- Enabling others to use your data **reinforces open scientific inquiry** and can lead to new and unanticipated discoveries.

Graphics by Jørgen Stamp CC-BY

Source: Georgia Tech Library

GBIF

DATA LOSS

Digital data are fragile and susceptible to loss for a wide variety of reasons:

- Natural disaster
- Facilities infrastructure failure
- Storage failure
- Server hardware/software failure
- Application software failure
- Format obsolescence
- Legal encumbrance
- Human error
- Malicious attack
- Loss of staffing competencies
- Loss of institutional commitment
- Loss of financial stability
- Changes in user expectations

Slide source: OpenAIRE & EUDAT, CC-BY-4.0, 2013

Image CC BY-NC-SA 2.0 by Dave Hill https://www.flickr.com/photos/dmh650/4031607067

"FAIR" DATA

Findable

- *assign persistent IDs, provide rich metadata, register in a searchable resource... (such as GBIF)*

Accessible

- *Retrievable by their ID using a standard protocol, metadata remain accessible even if data aren't...*

Interoperable

- *Use formal, broadly applicable languages, use standard vocabularies, qualified references... (e.g. Darwin Core, …)*

Reusable

- *Rich, accurate metadata, clear licences, provenance, use of community standards... (e.g. Dublin Core, EML, …)*

www.force11.org/group/fairgroup/fairprinciples

Slide source: OpenAIRE & EUDAT, CC-BY-4.0, 2013

BACKUP AND PRESERVATION – NOT THE SAME THING!

Backups

- Periodic snapshots of data in case the current version is destroyed or lost
- Backups are copies of files stored for short or near-long-term
- Often performed on a somewhat frequent schedule

Archiving

- Preserve data for historical reference
- Usually the final version, stored for long-term, and generally not copied over
- Often performed at the end of a project or during major milestones

Slide source: OpenAIRE & EUDAT, CC-BY-4.0, 2013

GBIF Newsroom

News and events from around the GBIF community

New partnership targets data persistence

Share Share Tweet G+1

PUBLICATION DATE

September 4th, 2014

LAST UPDATED

October 23rd, 2015

TAGS

- GBIF partnerships
- Ecological data
- Data publishing

A new collaboration between GBIF and the Data Observation Network for Earth (DataONE) aims to support long-term persistence of the biodiversity data shared through the network.

GBIF Secretariat and DataONE have signed a memorandum of cooperation that includes plans for a pilot repository to archive datasets shared through GBIF, using the services built by DataONE.

Source: GBIF News story, September 2014, DataONE: http://www.gbif.org/page/8199

Metadata

WHAT IS METADATA?

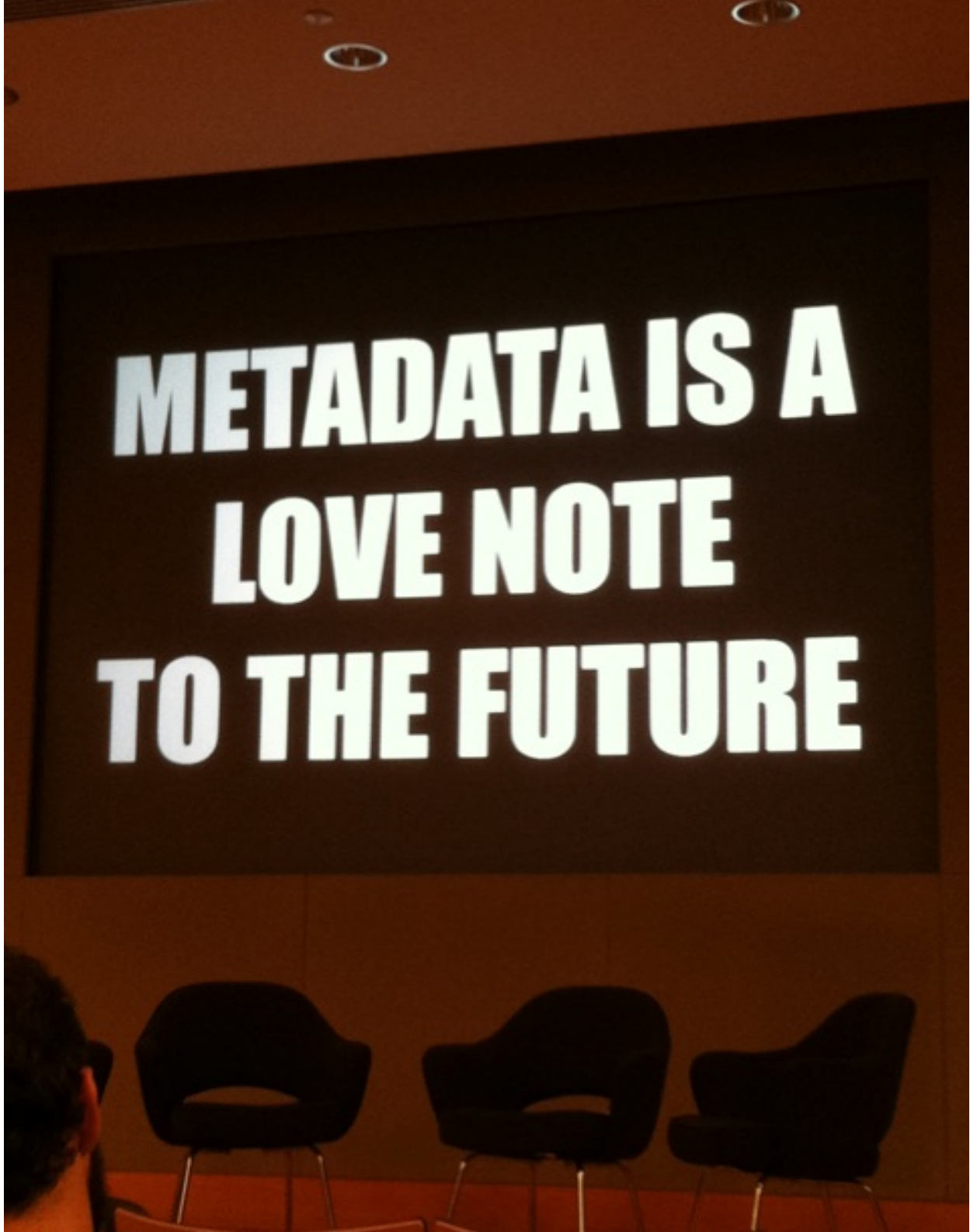


Image CC-BY 'Metadata is a love note to the future' by Cea+ www.flickr.com/photos/ centralasian/8071729256

WHAT IS METADATA?

Commonly defined as ‘data about data’, metadata helps to make data findable and understandable

Metadata can be:

Descriptive: information about the content and context of the data

Structural: information about the structure of the data

Administrative: information about the file type, rights management and preservation processes

Slide source: CC-BY EUDAT, 2015

WHY USE METADATA?

Comprehensive metadata will:

Facilitate data discovery

Help users determine the applicability of the data

Enable interpretation and reuse

Allow any limitations to be understood

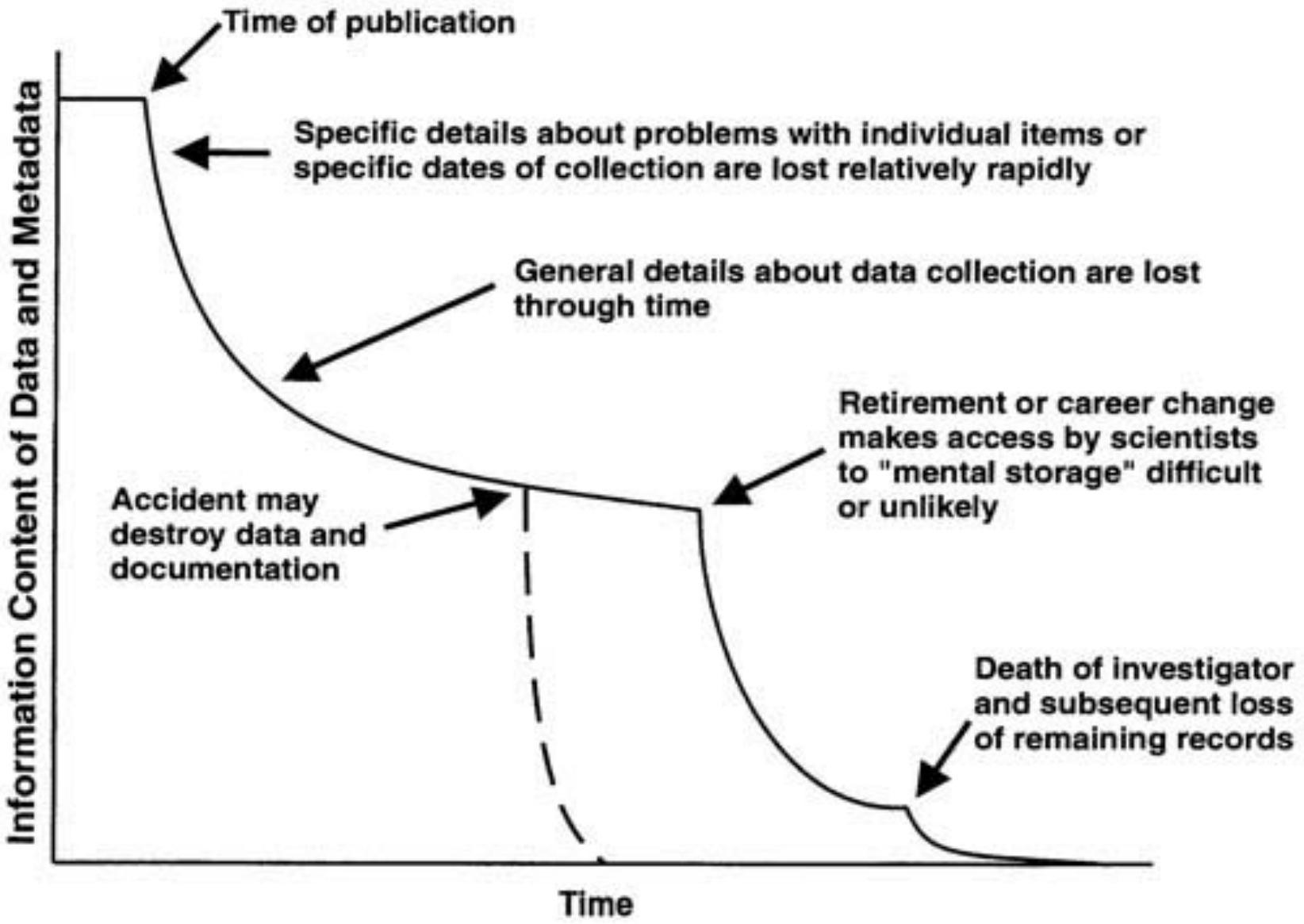
Clarify ownership and restrictions on reuse

Offer permanence as it transcends people and time

Provide interoperability

Slide source: CC-BY EUDAT, 2015

INFORMATION ENTROPY



The Loss of Information about Data (Metadata) Over Time, Michener et al, 1997

Slide source: CC-BY EUDAT, 2015

DATASET TITLES

Titles are critical in helping readers find your data

- While individuals are searching for the most appropriate data sets, they are most likely going to use the title as the first criteria to determine if a dataset meets their needs.
- Treat the title as the opportunity to sell your dataset.

A complete title includes: **What**, **Where**, **When**, **Who**, and **Scale**

An informative title includes: **topic**, **timeliness** of the data, specific information about **place and geography**

Slide source: CC-BY EUDAT, 2015

Data paper

METADATA TOPICS / HEADLINES

Dataset description
Project description
People and Organizations (including roles)
Coverage
- Taxonomic coverage
- Geographic coverage
- Temporal coverage

Methods
Intellectual property rights, licensing
Keywords

A peer-reviewed open-access journal

Biodiversity Data Journal

Making your data count!

Peer review option for biodiversity datasets.

Authors get scientific **credit** for data publication.

Meeting concerns over **data quality**.

Meeting concerns over **data citation** mechanism.

http://www.gbif.org/publishingdata/datapapers

SCIENTIFIC DATA

nature publishing group npg

RATIONALE FOR DATA PUBLISHING

Data Paper
A scholarly publication of searchable metadata document describing a dataset, or a group of datasets.

- Promote and publicize the existence of the data
- Provide scholarly credit to data publishers through citable journal publications
- Describe the data in a structured human-readable form

Slide source: Alberto González-Talavá, 2015

GBIF

WHY MANAGE AND PUBLISH DATA?

Because well-managed data opens up opportunities for re-use, sharing and makes for better science!

Make your research easier

Stop yourself drowning in irrelevant stuff

Save data for later

Avoid accusations of fraud or bad science

Share your data for re-use

Get credit for it

Meet funder/institution requirements

Slide source: OpenAIRE & EUDAT, CC-BY-4.0, 2013

GBIF Norway

Node team at NHM, University of Oslo

Dag Endresen, *Node manager*
Christian Svindseth, *Database manager*

Fridtjof Mehlum, *Research director*
Einar Timdal, *Associate professor*
Geir Søli, *Associate professor*

Artsdatabanken, Trondheim

Wouter Koch
Nils Valland

NTNU University Museum

Anders Finstad, *GBIF Science committee*

Research Council of Norway

Per Backe-Hansen, *Head of delegation*

Contact us at: gbif-drift@nhm.uio.no