FAIR and open biodiversity collection data management

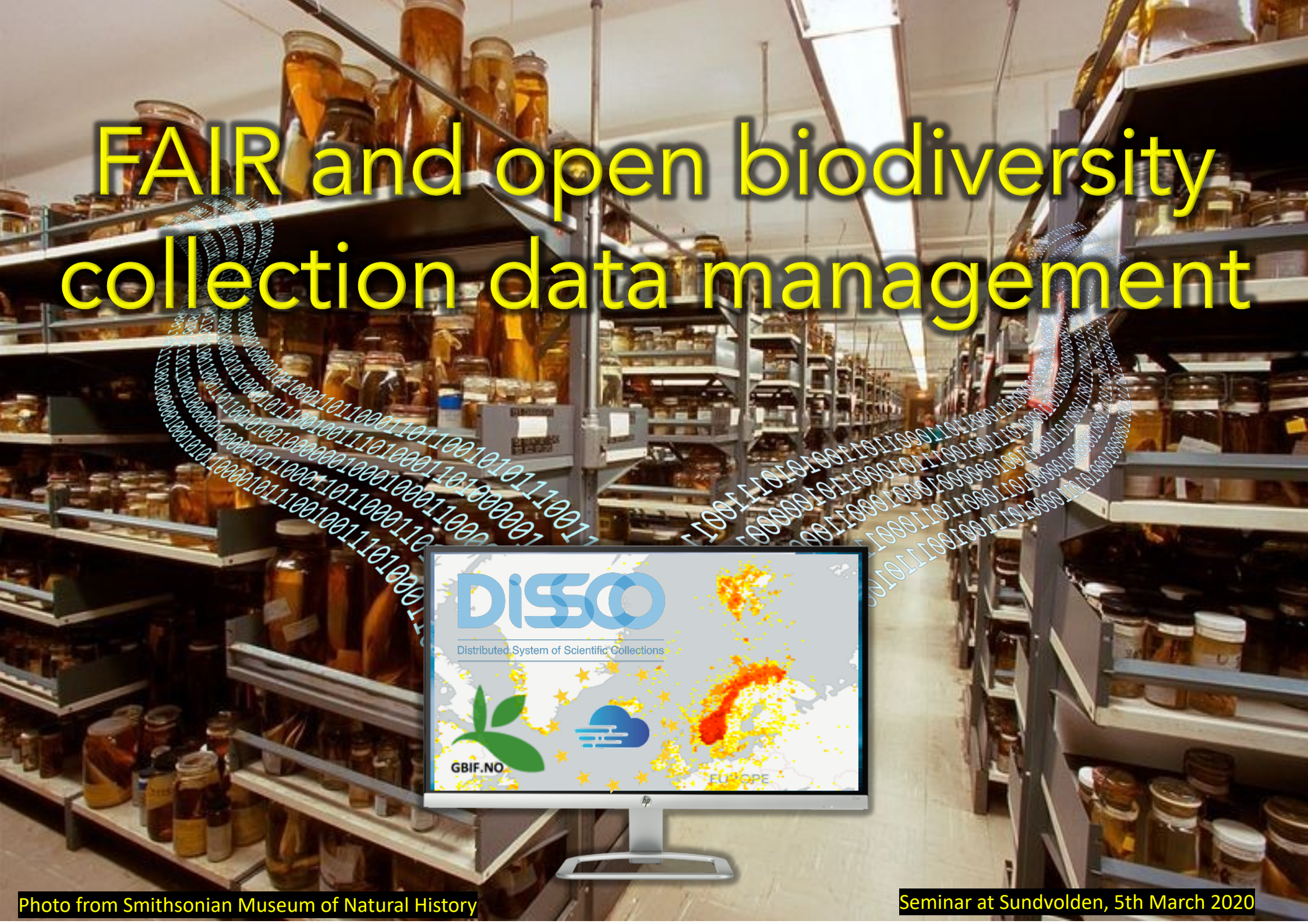


Photo from Smithsonian Museum of Natural History

Seminar at Sundvolden, 5th March 2020

WHAT IS OPEN SCIENCE?

"***Open science*** *is the movement to make scientific research (including publications,* ***data****,* ***physical samples****, and software) and its dissemination accessible to all levels of an inquiring society, amateur or professional*"

https://en.wikipedia.org/wiki/Open_science

Woelfle M, Olliaro,P, Todd MH (2011) Open science is a research accelerator. doi:10.1038/nchem.1149

Open Access (OA): Research results (often meaning research articles) distributed online and free of costs or other barriers.

Open Science: Researchers to share their methods, materials, computer code, algorithms, and research data.

Open Data: is freely available to everyone to use and re-publish as they wish, without restrictions from copyright, patents or other mechanisms of control.

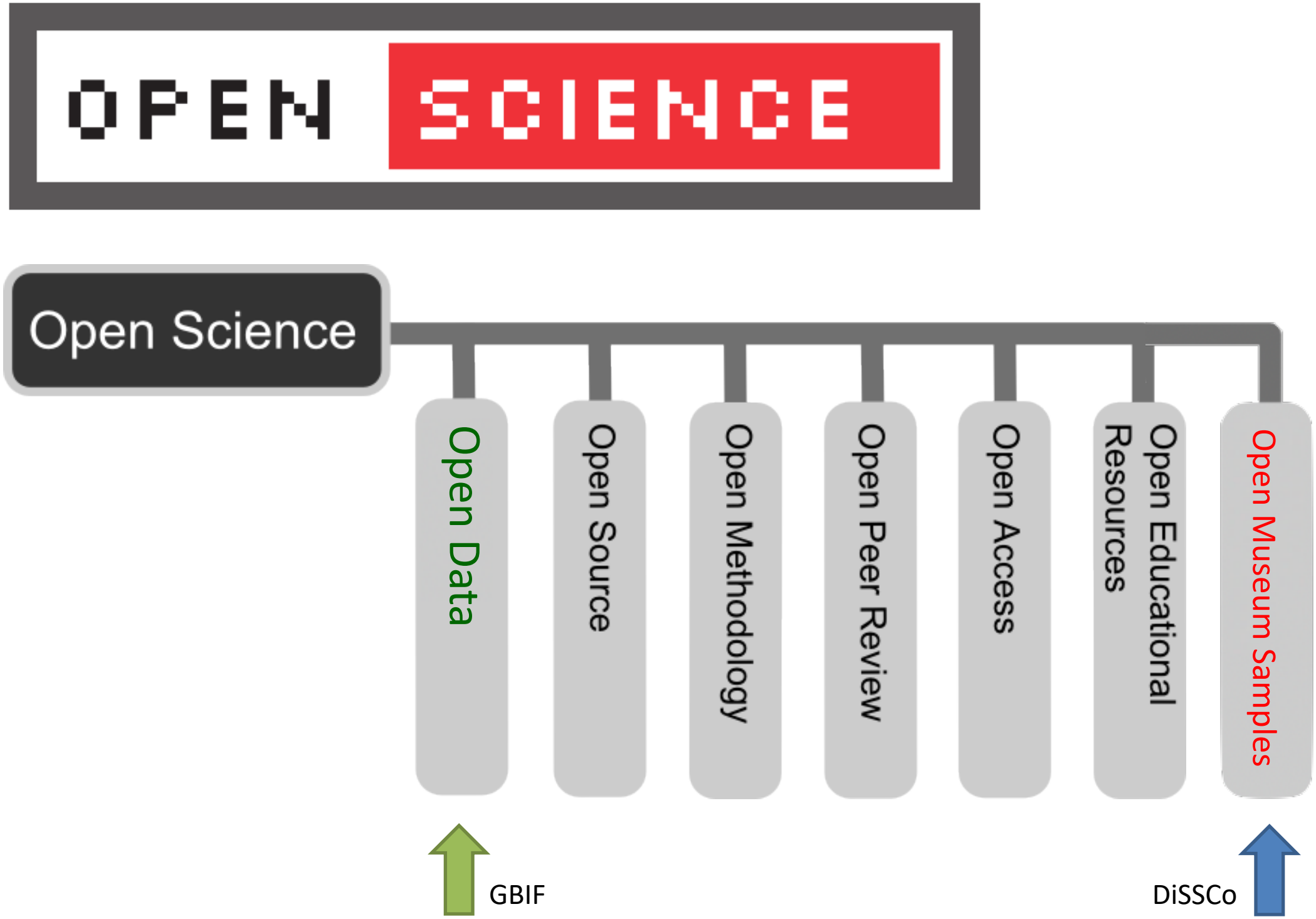
WHAT IS OPEN SCIENCE?

Open Science *can be seen as a continuation of, rather than a revolution in, practices begun in the 17th century with the advent of the academic journal*

cf. Wikipedia; David 2004

… and arguably also a continuation of the practice of natural history collections.

David, PA (2004) Understanding the emergence of 'open science' institutions: Functionalist economics in historical context. doi:10.1093/icc/dth023


OPEN SCIENCE
Open Science
Open Data
Open Source
Open Methodology
Open Peer Review
Open Access
Open Educational Resources
Open Museum Samples
GBIF
DiSSCo

WHY APPROACH OPEN SCIENCE IN MUSEUMS?

WHY APPROACH OPEN SCIENCE IN MUSEUMS?

- We are in the middle of an **ongoing paradigm shift** in scientific practice (and impact metrics).
- The open science wave is moving **fast**!
- Natural History Museums will need to develop **different approaches**, than they needed in the past – to remain relevant.
- Society is quickly gaining Big Data maturity and will **expect** new services from museum collections.


Photo: Botany Collection, Algae, Smithsonian National Museum of Natural History Museum, by Chip Clark.

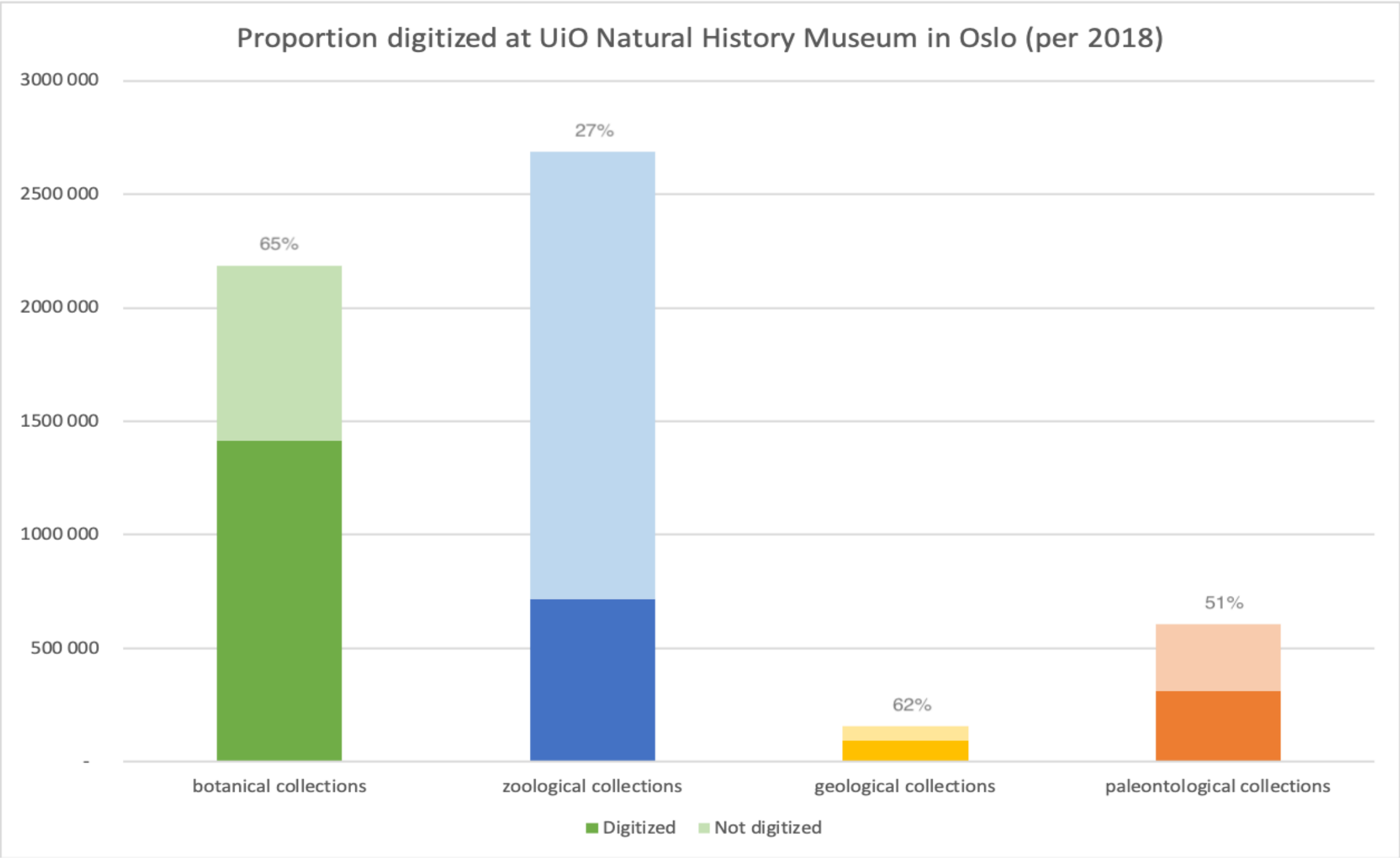
Very few specimens are digitized

The total number of specimens in natural history collections worldwide is estimated to **1.2 - 3 billion**.
(Ariño 2010; Duckworth *et al*. 1993)

GBIF publishes 1,4 billion records – including **168 million specimens**

5% to 10% coverage?

DIGITIZED SPECIMENS AT UIO NHM OSLO (47%)



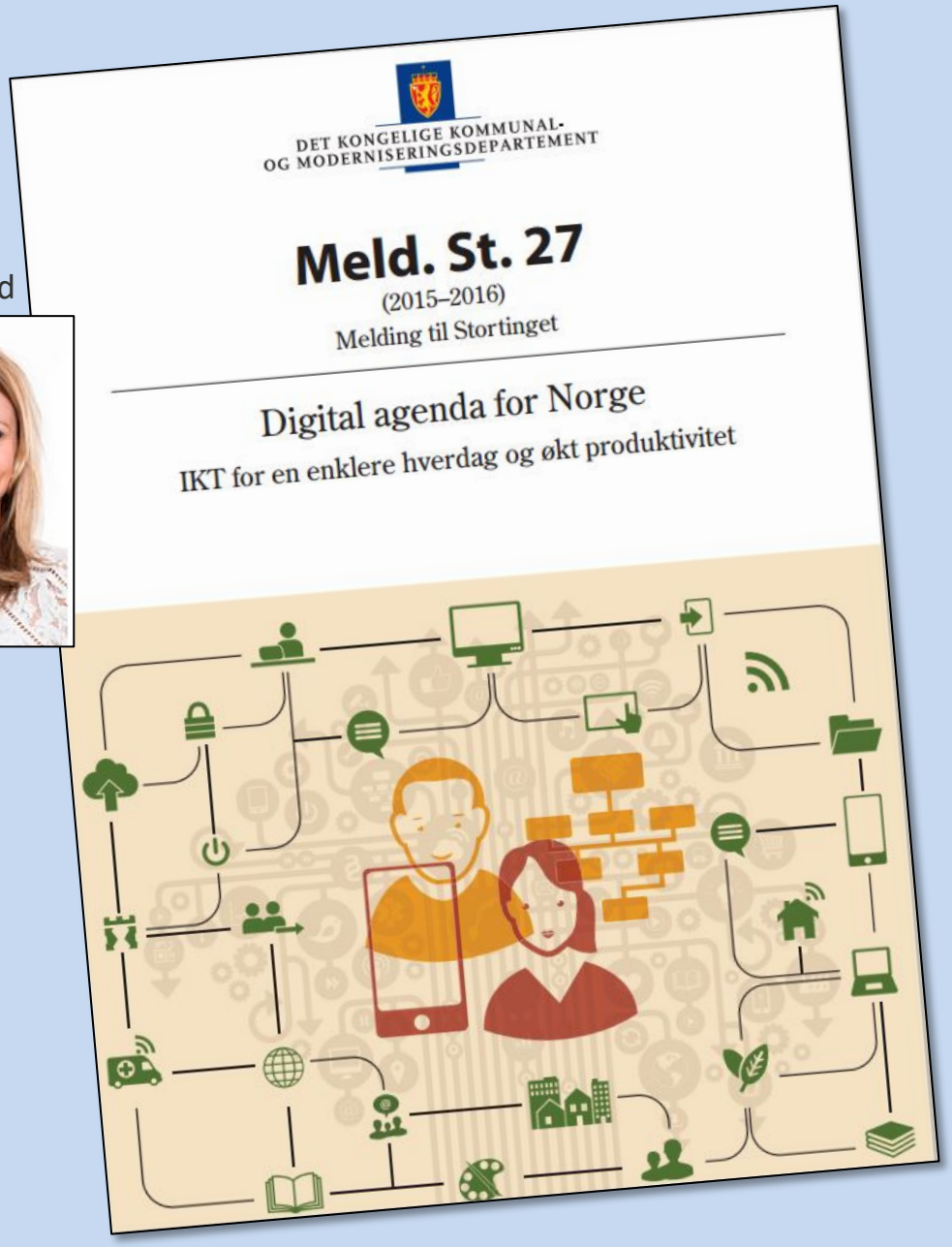
Regjeringen.no

The Government of Norway has decided that the public sector shall digitize

Astrup

Helleland

digitaliseringsminister



Research data must be as open as possible and as closed as necessary.

Research data should be managed and organized so that the values in the data can be utilized to the best possible extent.

Ministry of Education and Research, 2017

*Norway to follow the ambition of EU on **full open access to publicly funded research by 2020.***

UNIT

Direktoratet for IKT og fellestjenester i høyere utdanning og forskning

Established 1st January 2018, after combining BIBSYS, Ceres, and parts of Uninett.

Mission statement:

- *An innovative driver for digitisation.*
- *Utilize new technology to improve efficiency, improve quality, and ensure access to knowledge.*

*The **European Open Science Cloud** will uncover a "hidden treasure and untapped opportunities". A multi-billion euro project to transform research and **add value to vast stores of unused data**. Research data free to use for science including non-EU researchers.*

EU President **Ursula von der Leyen**, in Davos February 2020

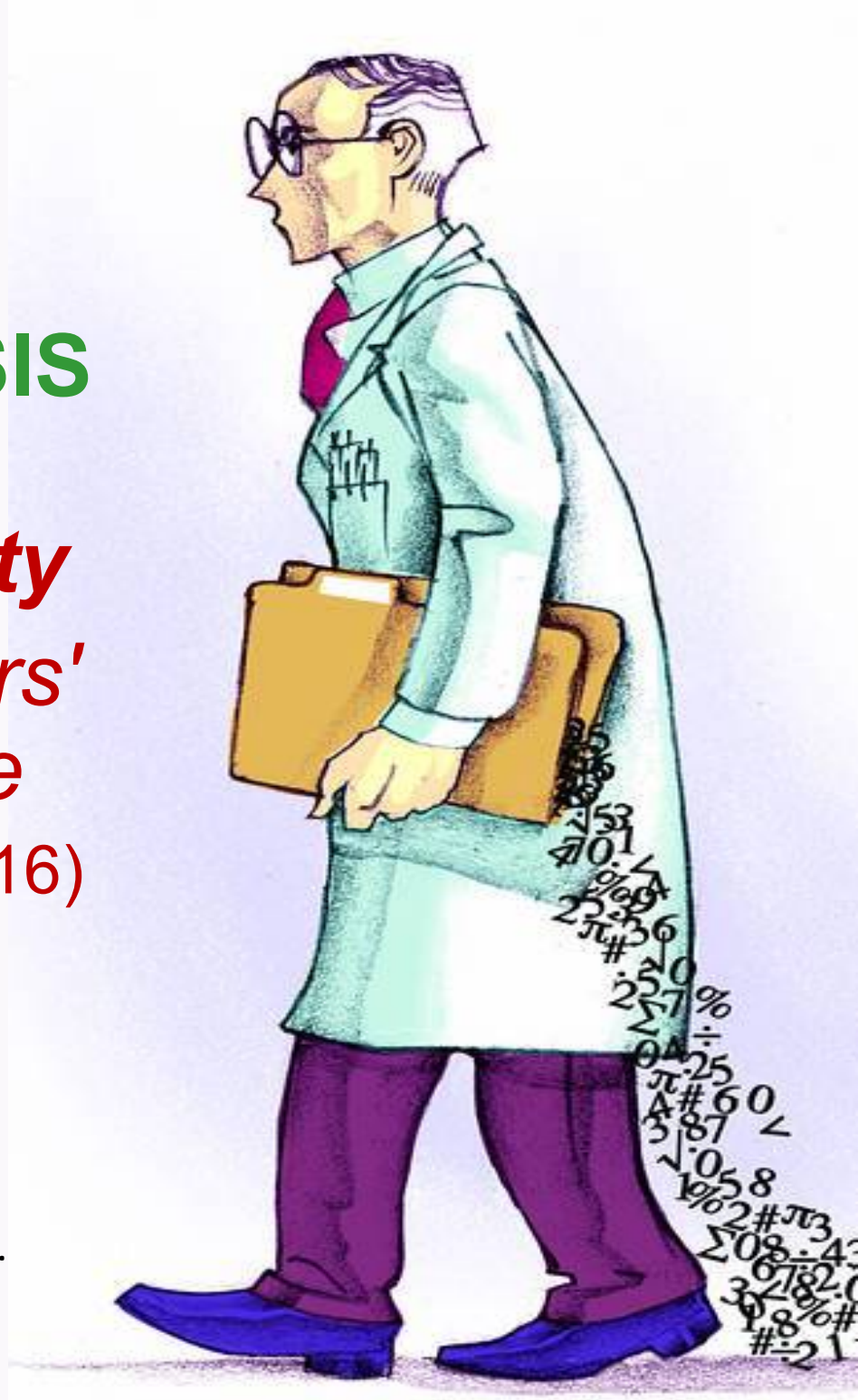
EUROPEAN OPEN SCIENCE CLOUD

The European Open Science Cloud – EOSC – a virtual environment for services for data storage, data management, data analysis, and reuse of research data across countries and scientific diciplines.

REPRODUCIBILITY CRISIS

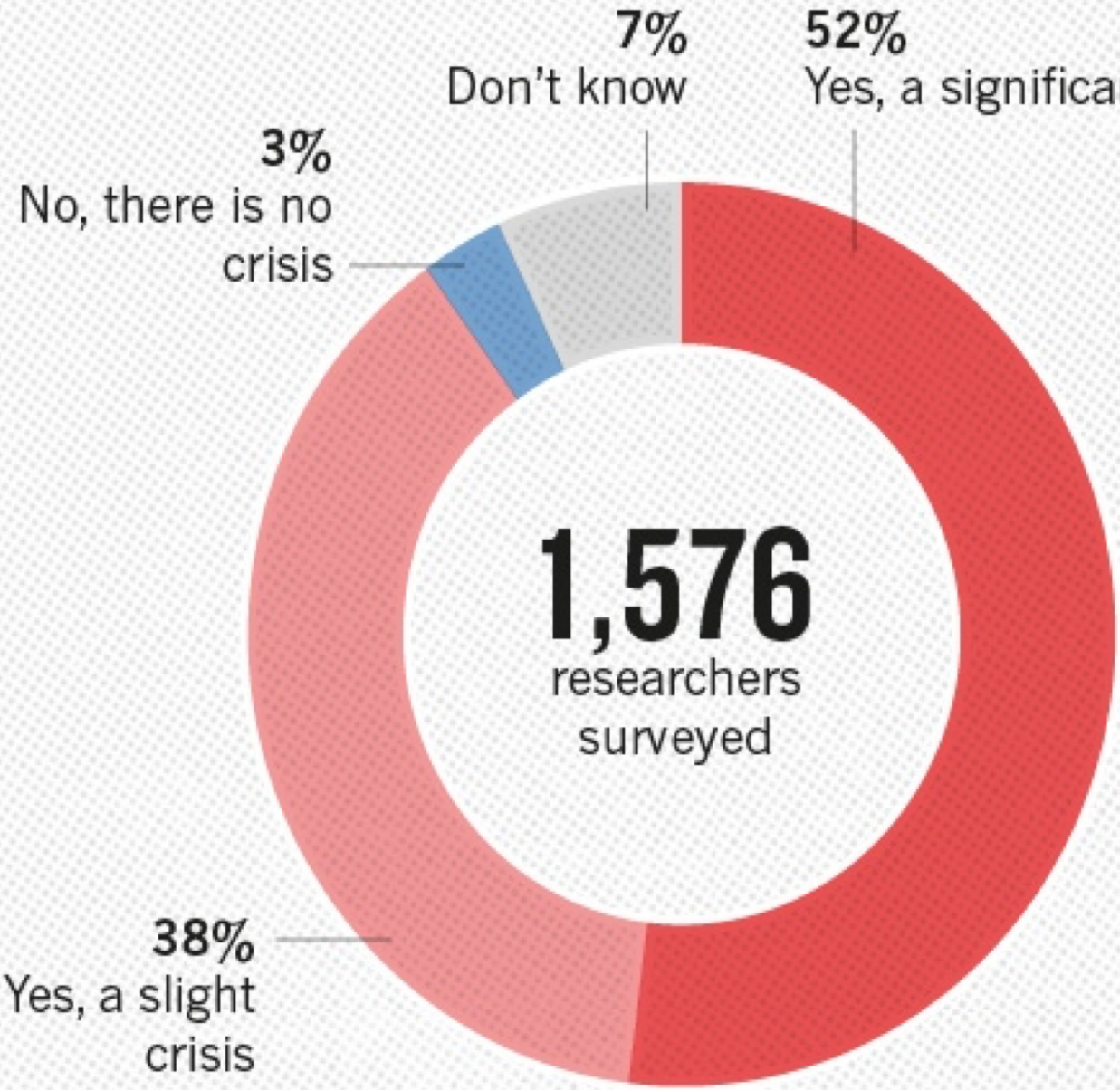
Scientific irreproducibility

the inability to repeat others' experiments and reach the same conclusion (Nature 2016)



Baker (2016) 1,500 scientists lift the lid on reproducibility. *Nature* doi:10.1038/533452a

IS THERE A REPRODUCIBILITY CRISIS?



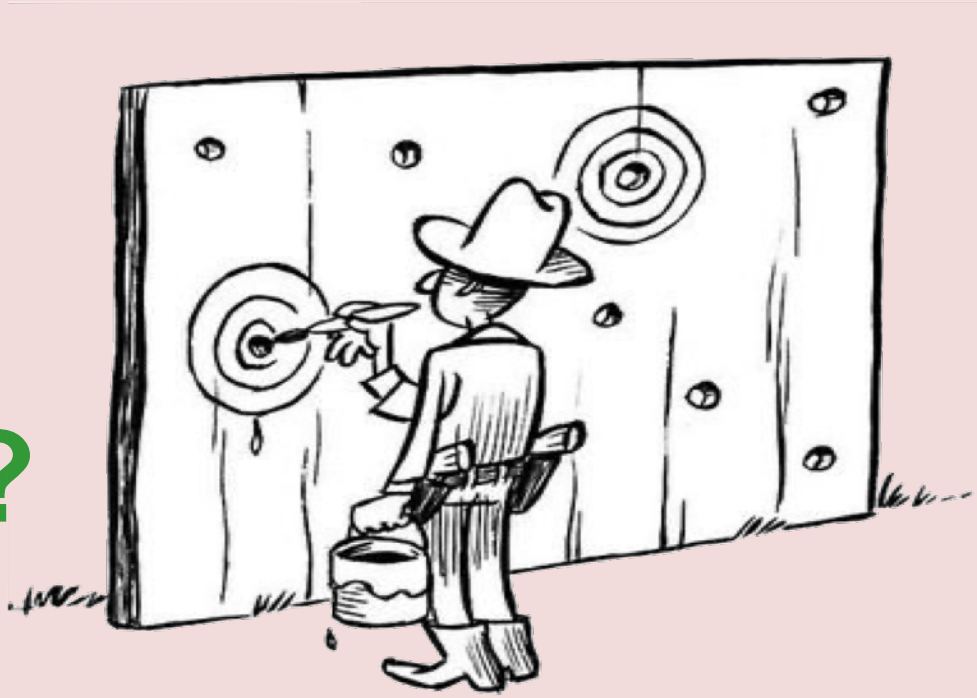
"***Scientific irreproducibility*** *— the inability to repeat others' experiments and reach the same conclusion — is a growing concern*".

Baker (2016) **Nature**
doi:10.1038/533452a

Open Science solution*:* researchers to share their **methods**, **data**, computer **code** and results in central data repositories.

Note that we also need herbarium ***specimen*** *and bio-repositories (eg. museums).*

WILL ANYBODY TRUST CLOSED SCIENCE AGAIN?



Recent studies indicates that **p-hacking** is a significant problem – sometimes without the scientist even being aware of doing so.

p-hacking, occurs when researchers collect or select data or statistical analyses until nonsignificant results become significant.

Pre-registered (open) data & material samples provides an insurance against suspicion of data dredging (and plain data falsification).

Ioannidis (2005) Why Most Published Research Findings Are False. PLoS Med. doi:10.1371/journal.pmed.0020124.
Head *et al.* (2015) The Extent and Consequences of P-Hacking in Science. PLoS Biol. doi:10.1371/journal.pbio.1002106

Reuse of research data and other research products

CREDIT FOR DATA REUSE

To incentivize the sharing of useful data, the scientific enterprise needs a well-defined ***system that links individuals with reuse of data sets they generate***

Pierce *et al*. Credit data generators for data reuse, *Nature* 6 June 2019

DATA CITATION AS A NEW CURRENCY OF SCIENCE

- **Peer-reviewed scholarly papers** in high impact journals maintain considerable weight for impact metrics.
- *A **movement** is under way to **build similar status** for open data, open metadata, open material samples, and other open scientific research products…*

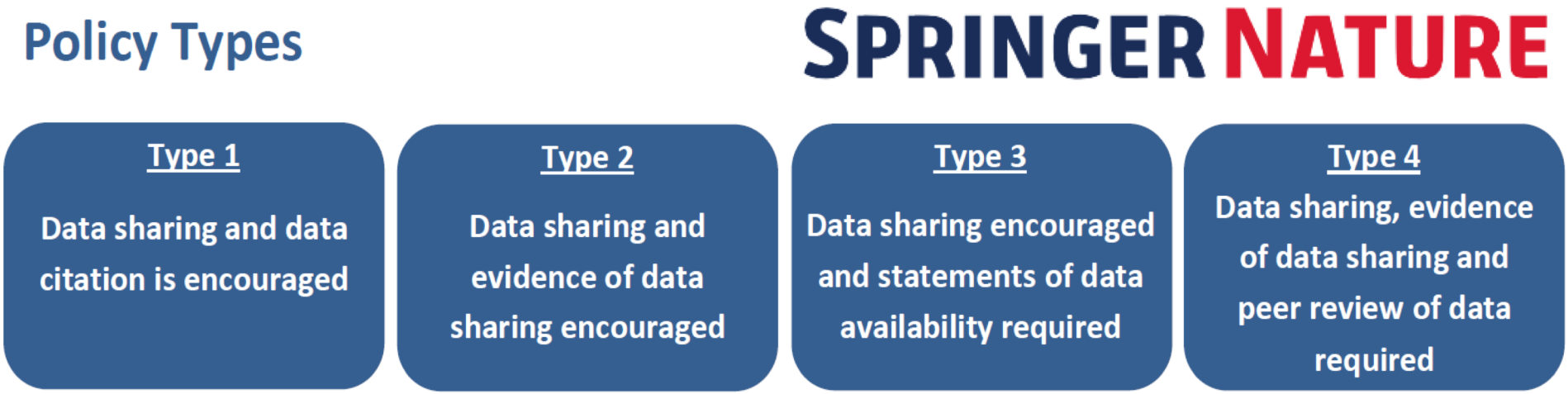
DATA CITATION PRINCIPLES

FORCE11
The Future of Research Communications and e-Scholarship

1. Data to be legitimate **citable** products of research.
2. Data citations giving **scholarly credit** and attribution.
3. In scholarly literature, whenever claims are based on data, **data should always be cited**.
4. Persistent method for **identification of data**, that is machine actionable, globally unique, universal.
5. Data citation facilitate **access to data** or at least to **metadata**.
6. **Unique identifiers** that persist even beyond the lifespan of the data.
7. Data citation identify and access the specific data that **support verification** of the claim (provenance, time-slice, version).
8. Flexible, but attention to **interoperability** of practices across communities.

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014

OPEN RESEARCH DATA POLICIES



- Scientific journals is starting to provide guidelines for availability of research data to accompany published research papers.
- GBIF is here recommended as a data repository for biology.
- *Museum collections could provide a similar function for open availability to biological material samples…*

http://www.springernature.com/gp/group/data-policy/policy-types

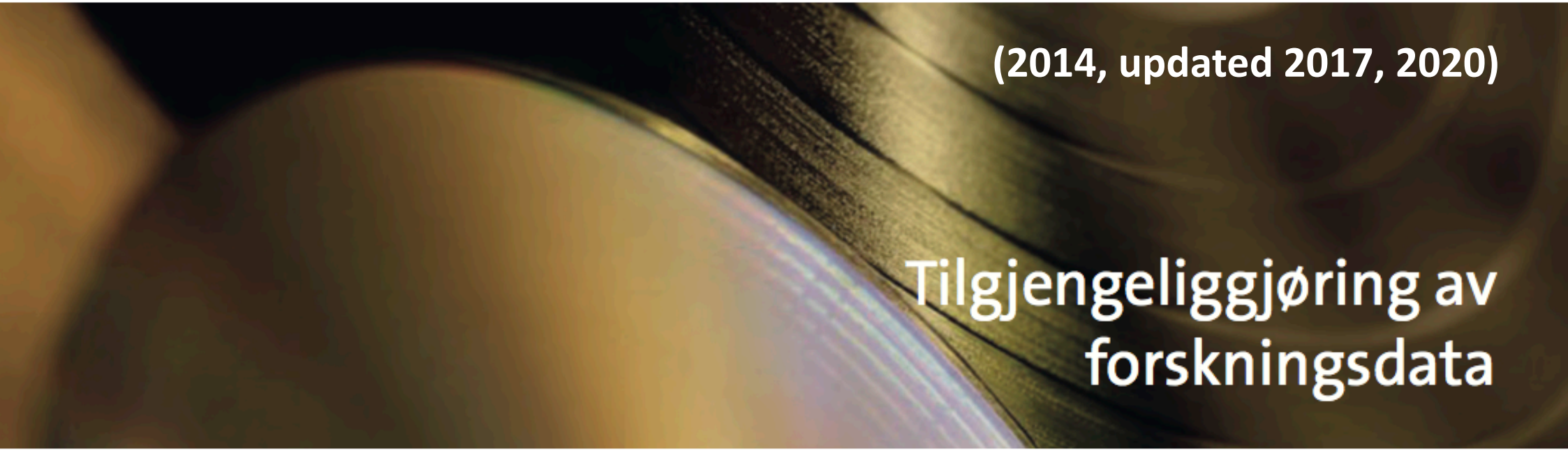
"Funders hold the stick: they should disburse no further funding without a data-stewardship plan."

nature 27 February 2020

Barend Mons (2020) Invest 5% of research funds in ensuring data are reusable. Nature 578: 491 doi:10.1038/d41586-020-00505-7

The Research Council of Norway ***expects all research data*** *from projects funded by the Research Council* ***to be made freely available as open data****.*
(For some situations there can be valid and justified reasons for exceptions).

Forskningsrådet



(2014, updated 2017, 2020)

Tilgjengeliggjøring av forskningsdata

Policy for Norges forskningsråd

INVEST 5% OF RESEARCH FUNDS IN ENSURING DATA ARE REUSABLE

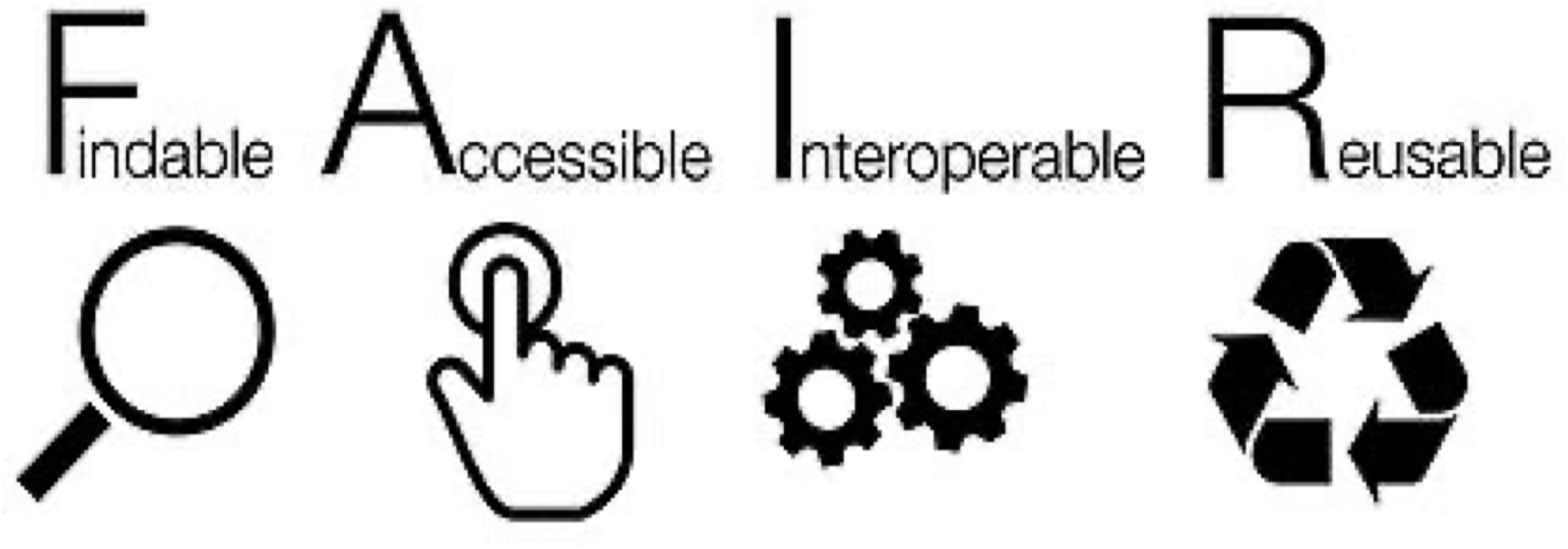
It is irresponsible to support research but not data stewardship. Huge sums of taxpayer funds go to waste because such data cannot be reused.

On average, 5% of overall research costs should go towards data stewardship.

If data are treated properly, researchers will have significantly more time to do research.

Barend Mons, Nature 578: 491, 27 February 2020

Barend Mons co-leads GO FAIR and is the co-author of the FAIR principles (2016)



Promotes maximum (re)use of research data

*Researchers need to do **more** than simply post their **data on the web** for it to be useful*

Wilkinson *et al*. (2016) The FAIR guiding principles for data management and stewardship.
doi:10.1038/sdata.2016.18

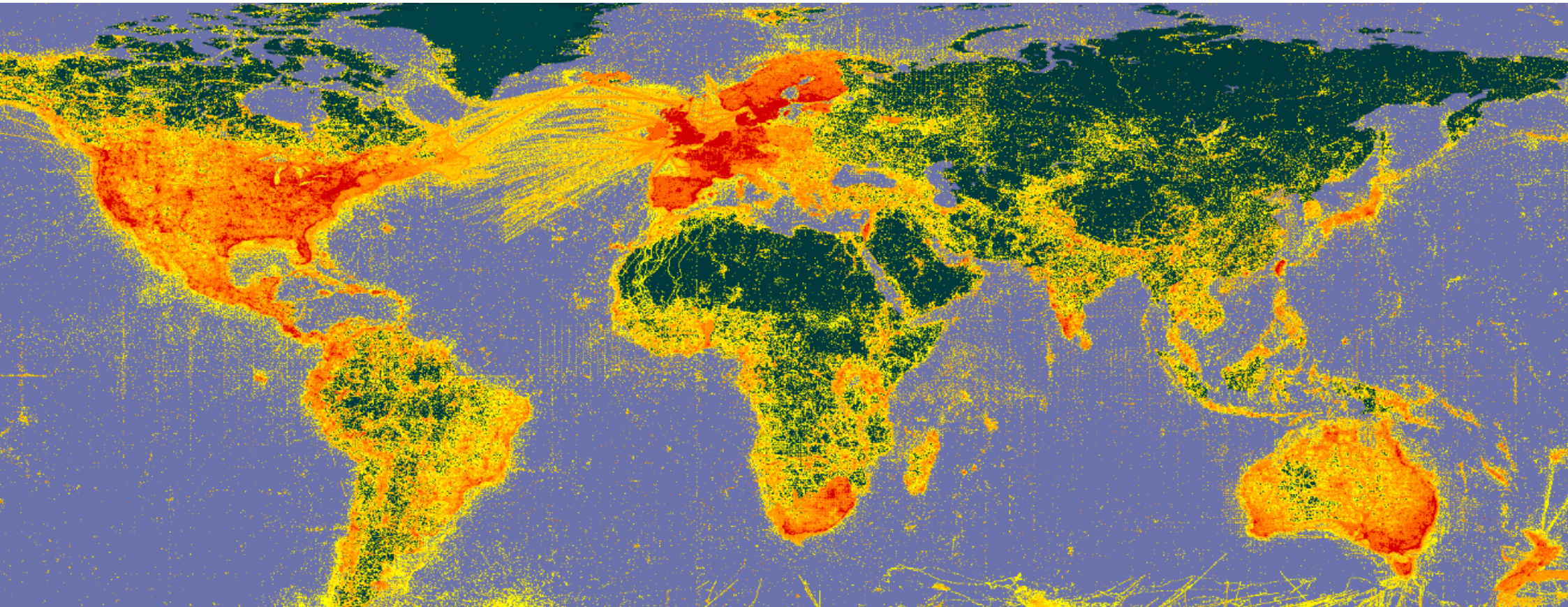
GBIF | CERTIFIED FAIR DATA

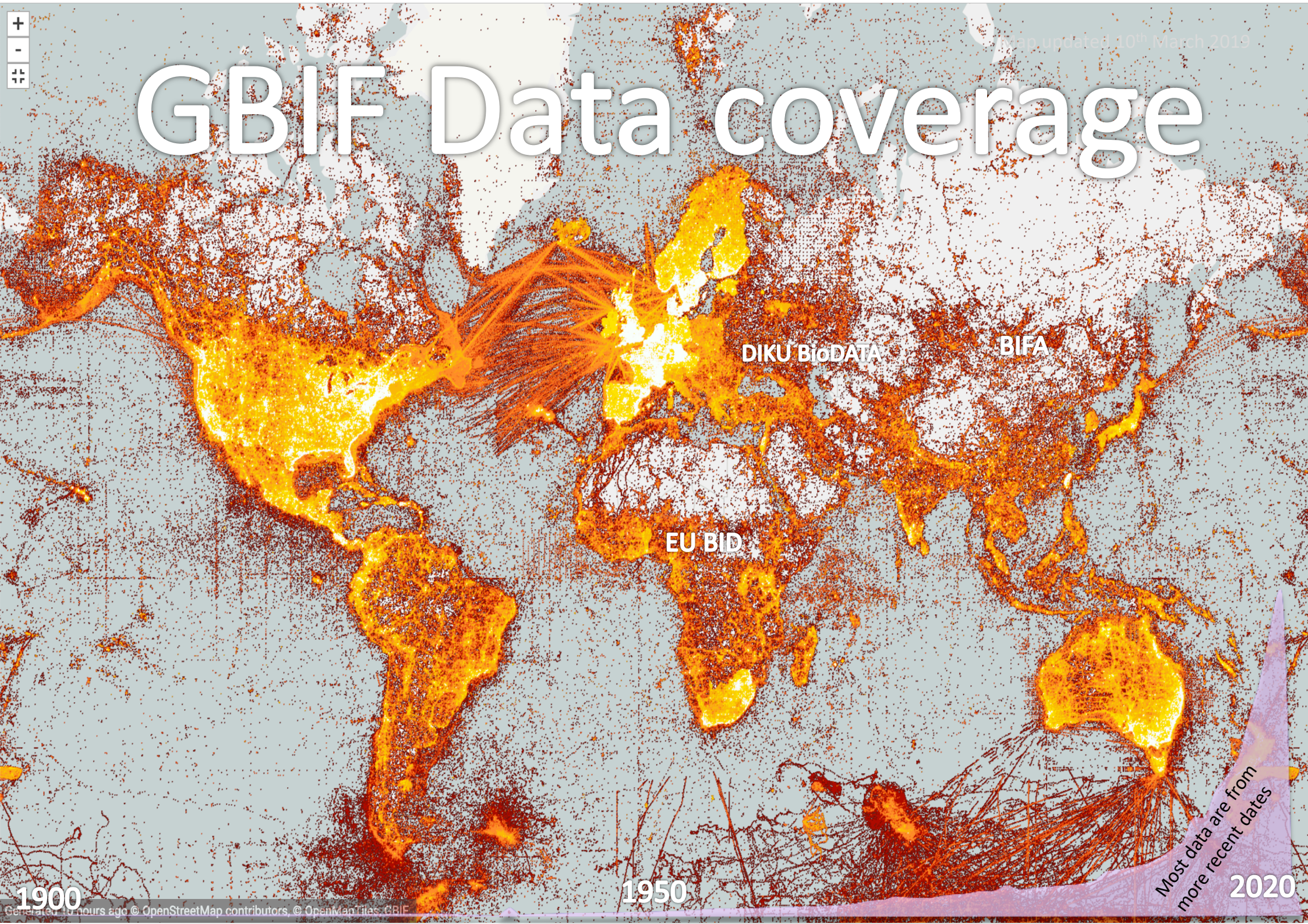
Global Biodiversity Information Facility (**GBIF**) provides certified FAIR data sharing for all biological datasets

Distributed System of Scientific Collections (**DiSSCo**) will implement FAIR data principles for natural history collections

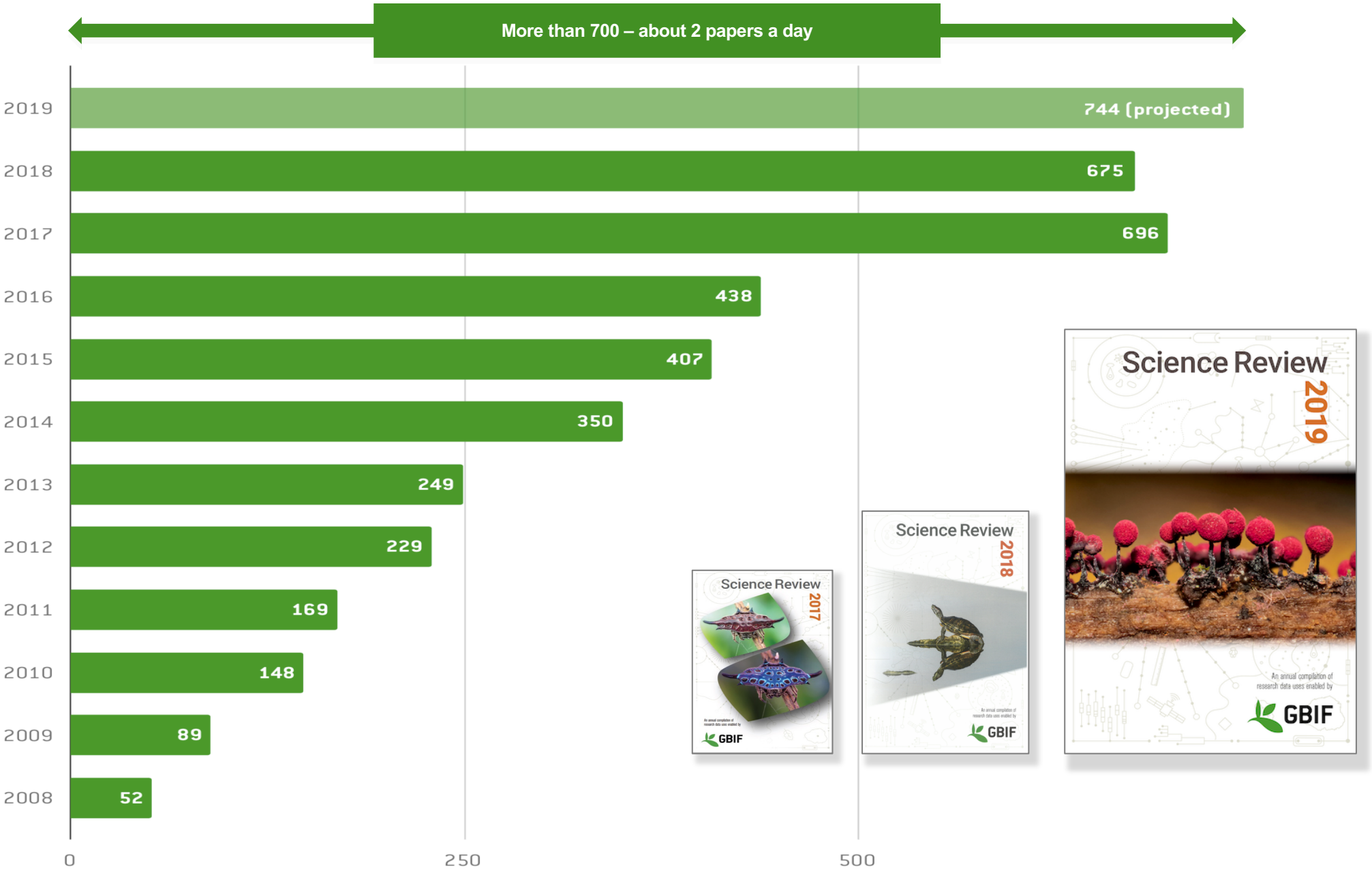
GBIF enables free & open access to biodiversity data online

We are an international government-initiated and -funded initiative focused on making biodiversity data available to all and anyone, for scientific research, conservation and sustainable development




GBIF Data coverage
Map update: 10th March 2019
DIKU BioDATA
BIFA
EU BID
1900
1950
2020
Most data are from more recent dates
Generated 0 hours ago © OpenStreetMap contributors, © OpenMapTiles, GBIF

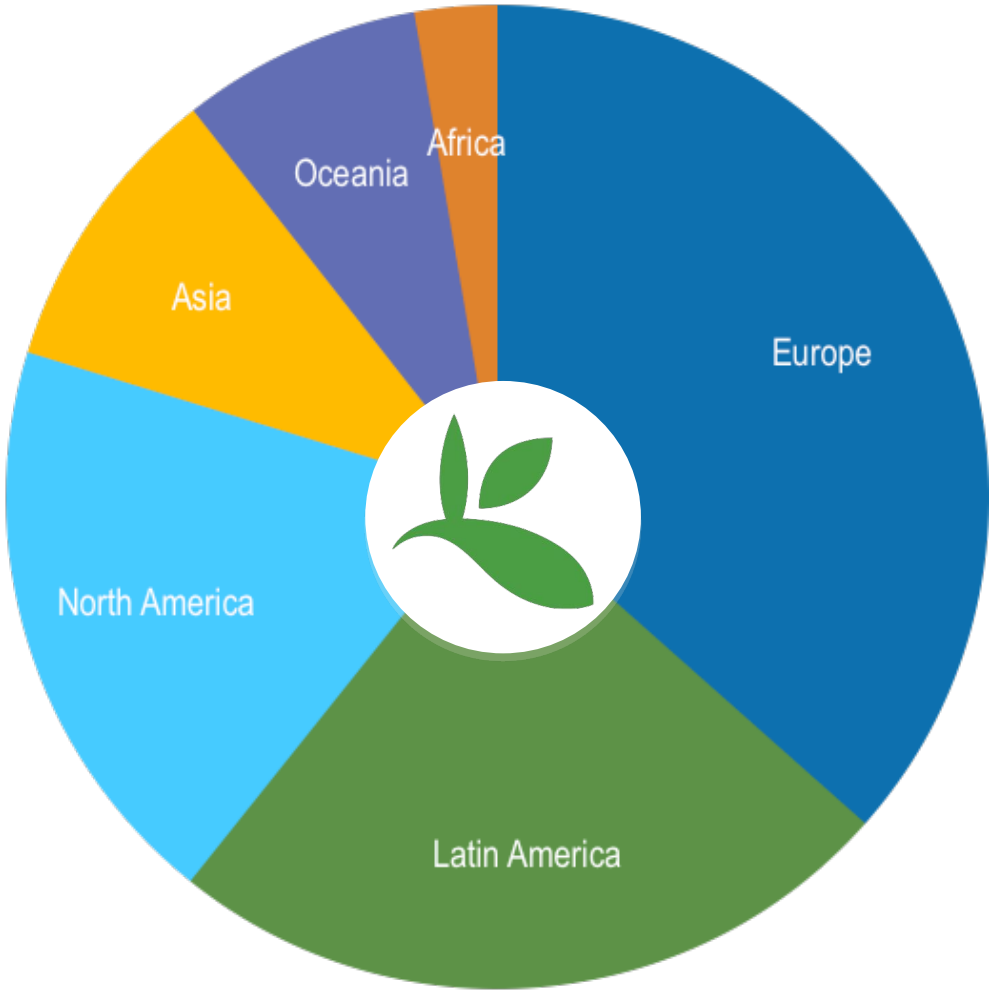
Peer-reviewed publications using GBIF-mediated data



PEER-REVIEWED USES BY COUNTRY AND REGION

STATUS *JAN-FEB 2020 – TWO MONTHS (281 CITATIONS)*

Total # of peer-reviewed papers by country		
1	United States	70
2	United Kingdom	35
3	China	31
4	Brazil	28
5	Spain	24
6	Mexico	23
7	Canada	21
8	Germany	18
9	Italy	15
10	Australia	14
19	Sweden	9
...	***Norway***	***8***



Open science provides possibilities for new & curiosity-driven research

5th March 2020, NHMO Science seminar, Sundvolden

FAIR and open biodiversity collection data management

UiO : Natural History Museum
University of Oslo