



Open Research Data

Slides CC BY: Dag Endresen, GBIF Norway

INDEX

Why teach open science?

Data citation principles

FAIR data principles

Reproducible research

Open research data tools

Data management plan

Metadata, data paper

Machine-readable data



Open Science



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

Open Science: Researchers to share their methods, computer code and research data in central data repositories.

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.

FAIR data principles: **findable, accessible, interoperable** and **reusable**.

For full reproducibility we also need access to the physical biological material – deposited in **museum collections** and **biobank-repositories**.



Why should Students learn open science?



Why teach students open science?

- ❖ We are in the middle of an ongoing **paradigm shift** in scientific practice (and impact metrics).
- ❖ The open science wave is moving **fast**!
- ❖ Young scientists will need **different skills**, than was needed previously to succeed in academia.
- ❖ Researchers 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 biodiversity information and research.



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**...



Declaration on Research Assessment



- DORA recognizes **the need to improve how the outputs of scholarly research are evaluated.**
- DORA's vision is to advance practical and robust approaches to **research assessment** globally and across all scholarly disciplines.
- The Declaration on Research Assessment (DORA) was developed in San Francisco in 2012.
- It has become a **worldwide** initiative covering all scholarly disciplines and key stakeholders.
- Covering funders, publishers, professional societies, institutions, and researchers.
- To date (2021-05-14), 17 511 individuals & 2 232 organizations in **145 countries** have signed DORA.

Open research data policies

SPRINGER NATURE

Policy Types

Type 1

Data sharing and data citation is encouraged

Type 2

Data sharing and evidence of data sharing encouraged

Type 3

Data sharing encouraged and statements of data availability required

Type 4

Data sharing, evidence of data sharing and peer review of data required

The scientific journals (at Springer Nature) practices different guidelines and requirements for availability to the underlying research data for published research papers.

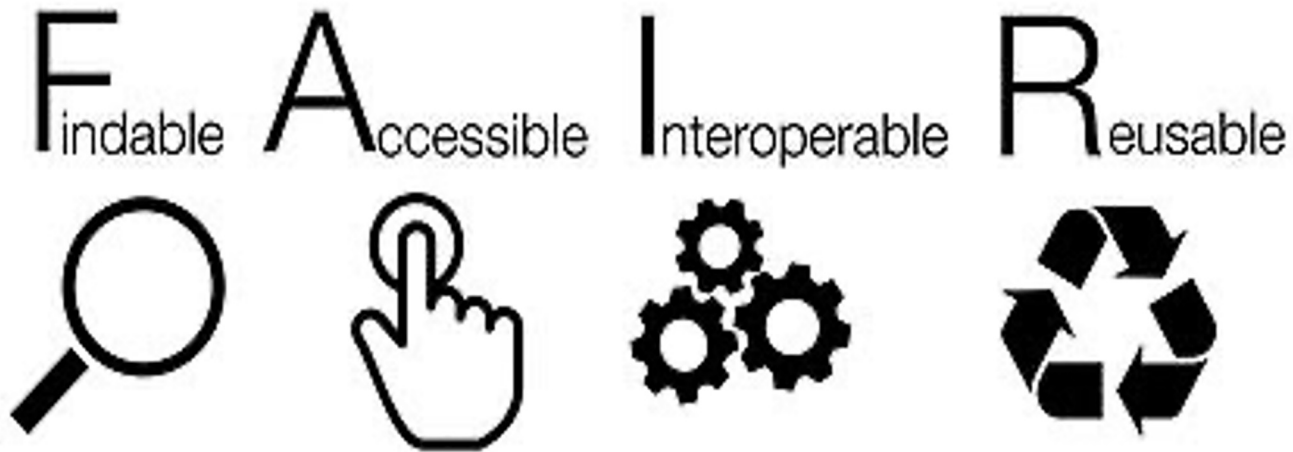
Springer Nature has made a comprehensive report on practical incentives and appropriate norms to promote open data.

GBIF listed as one of the recommended data repository [\[link\]](http://www.springernature.com/gp/group/data-policy/policy-types).

Data citation principles



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.



*... researchers need to do more than simply post their data on the web for it to be **re-usable**.*

FAIR data is about **machine-readable** data

What is FAIR data?

- **FINDABLE** → Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.
- **ACCESSIBLE** → Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.
- **INTEROPERABLE** → Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- **REUSABLE** → Data and collections have a clear usage licenses and provide accurate information on provenance.
- <https://www.go-fair.org/>

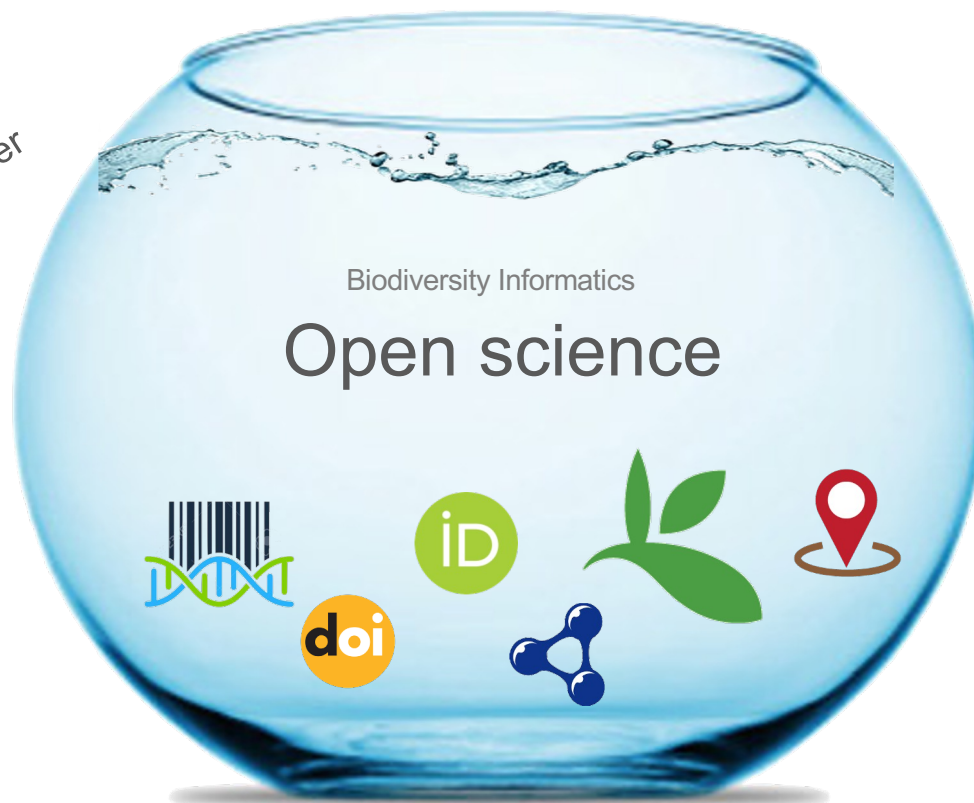
Is GBIF FAIR compliant?



- **FINDABLE** → GBIF requires a minimum required metadata. All datasets are identified by Digital Object Identifiers (DOIs).
- **ACCESSIBLE** → The GBIF Portal API provides a machine readable interface (REST + JSON) and use the IPT as trusted data repository.
- **INTEROPERABLE** → GBIF recommends using the Ecological Metadata Language (EML) for datasets and Darwin Core for occurrence data.
- **REUSABLE** → GBIF require creative common data licenses (CC0, CC BY, or CC BY-NC). Provenance available from the GBIF portal with further details available through the respective IPT.
- [FAIRsharing.org](https://fairsharing.org/biodbcore-000635/) | <https://fairsharing.org/biodbcore-000635/>



new possibilities for novel curiosity-driven research

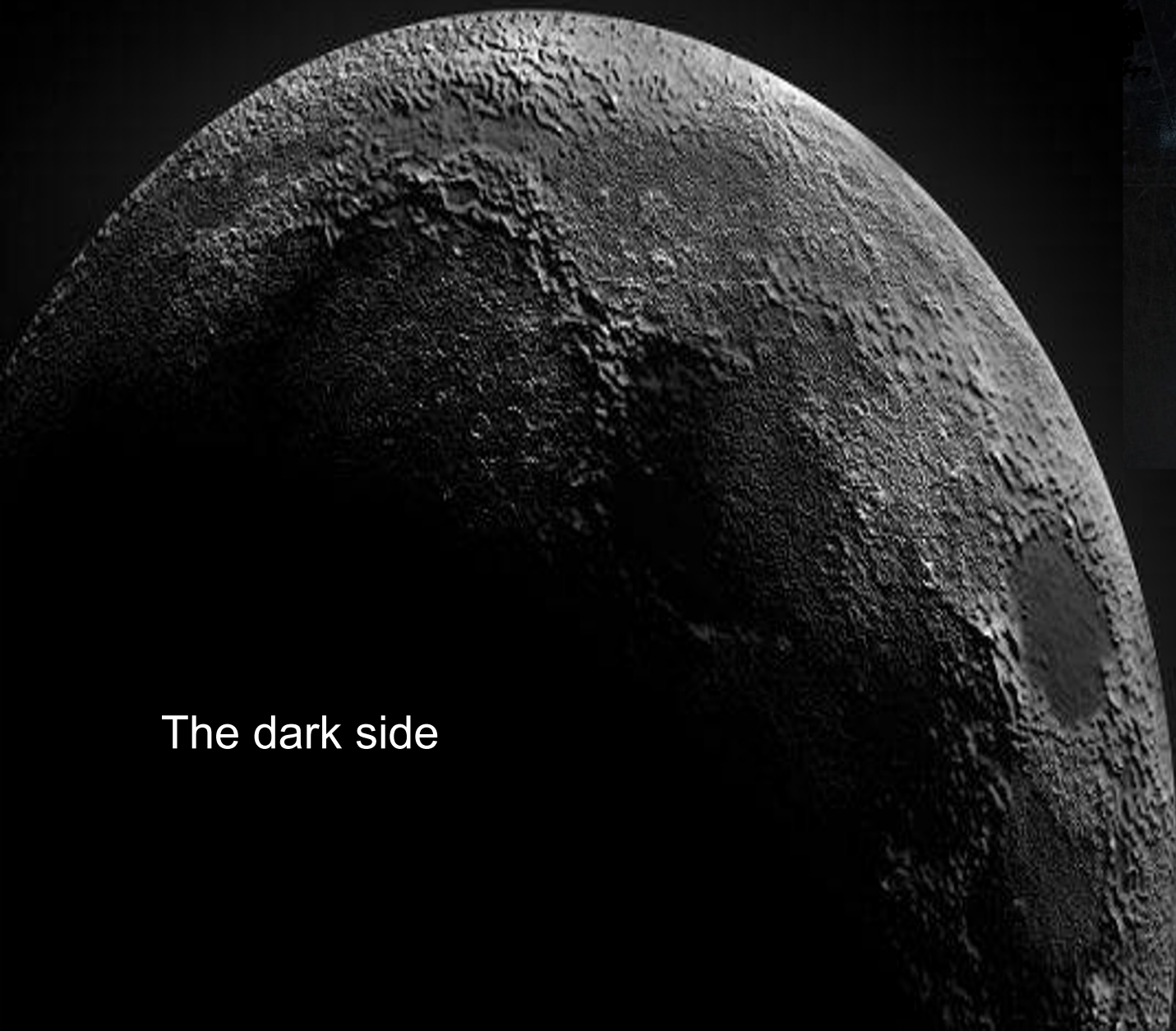


Global Biodiversity Information Facility



Career opportunities

- Skills for open research and open data are in **increasing demand!**
- Enables new research methodologies that were **not possible before.**
- Bring **benefits for your career** as a (young) researcher.
- Add **research datasets** to your CV and online researcher profiles.



The dark side



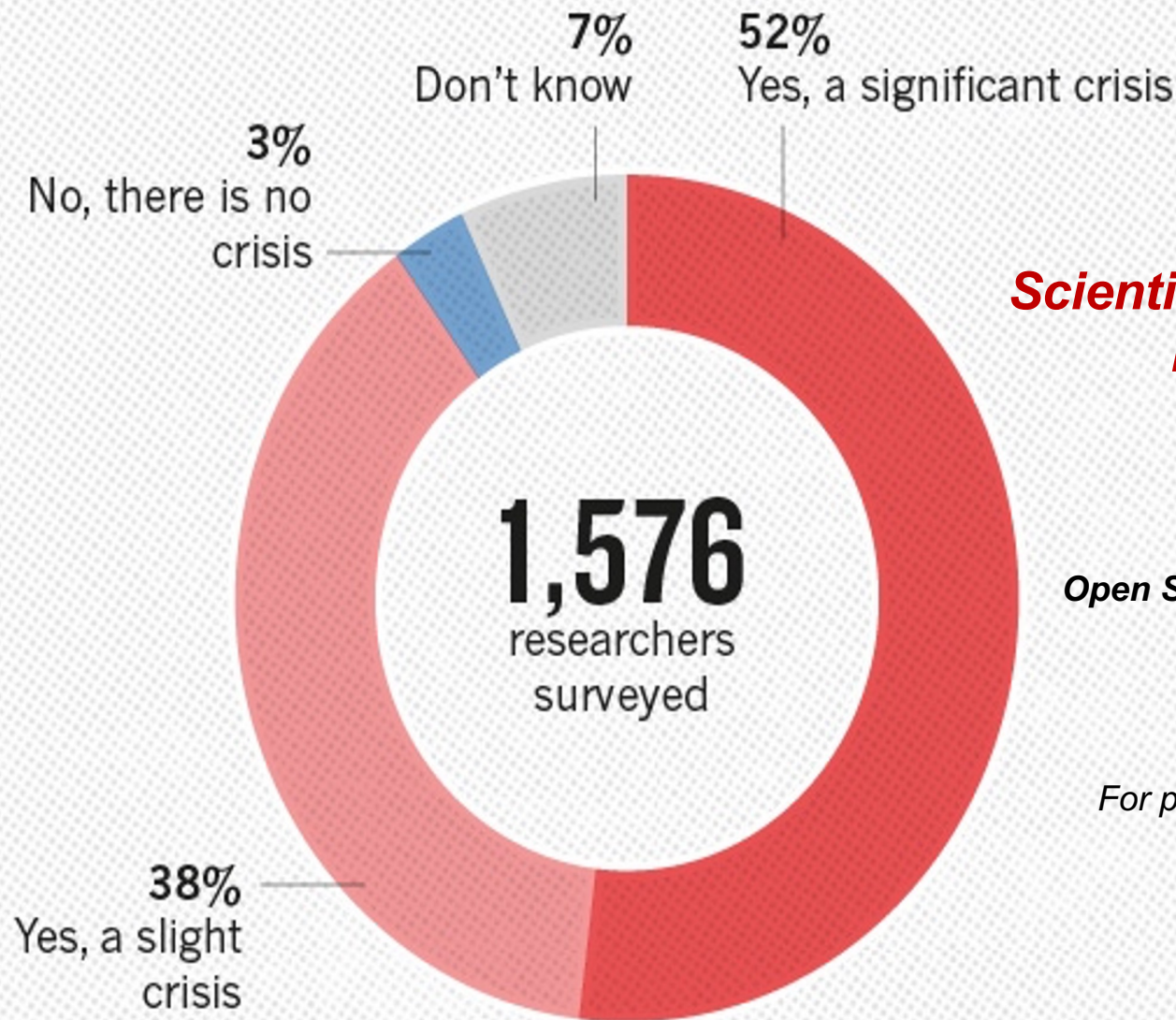
Reproducibility crisis

"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
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.

*For physical samples we also need
bio-repositories (museums).*

Will anybody trust closed science again?

- Studies (1,2) indicates that **p-hacking** is a significant problem – sometimes even without the scientist even being aware of doing so.
- **Pre-registered (open) data** provides a good insurance against suspicion of both data dredging (and plain data falsification).
- “p-hacking” = occurs when researchers collect or select data or statistical analyses until nonsignificant results become significant (*data fishing*, ...)



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

TOOLS FOR REUSE OF RESEARCH DATA & DATA CITATION



Data Management Plan (DMP)

- A **formal document** that describes how data are to be handled during a research project, and after the project is completed.
- The goal is to **plan data management** BEFORE the project begins.
- Including a **plan for the costs** of data management and archiving.
- This **saves time** in the long run and promotes data fitness for reuse.
- **Reduce duplication** of existing scientific studies.
- Reduce the **loss of data**.



Illustration CC BY Jørgen Stamp

Why write Data Management Plans?

A data management plan is a tool for making your research **reproducible** and thus **trustworthy**.

Good data curation saves you research time, because you, your collaborators, and others, will **find**, **understand**, and get **access** to your research data.

Efficient data sharing provides broader distribution and **impact** for your research results.

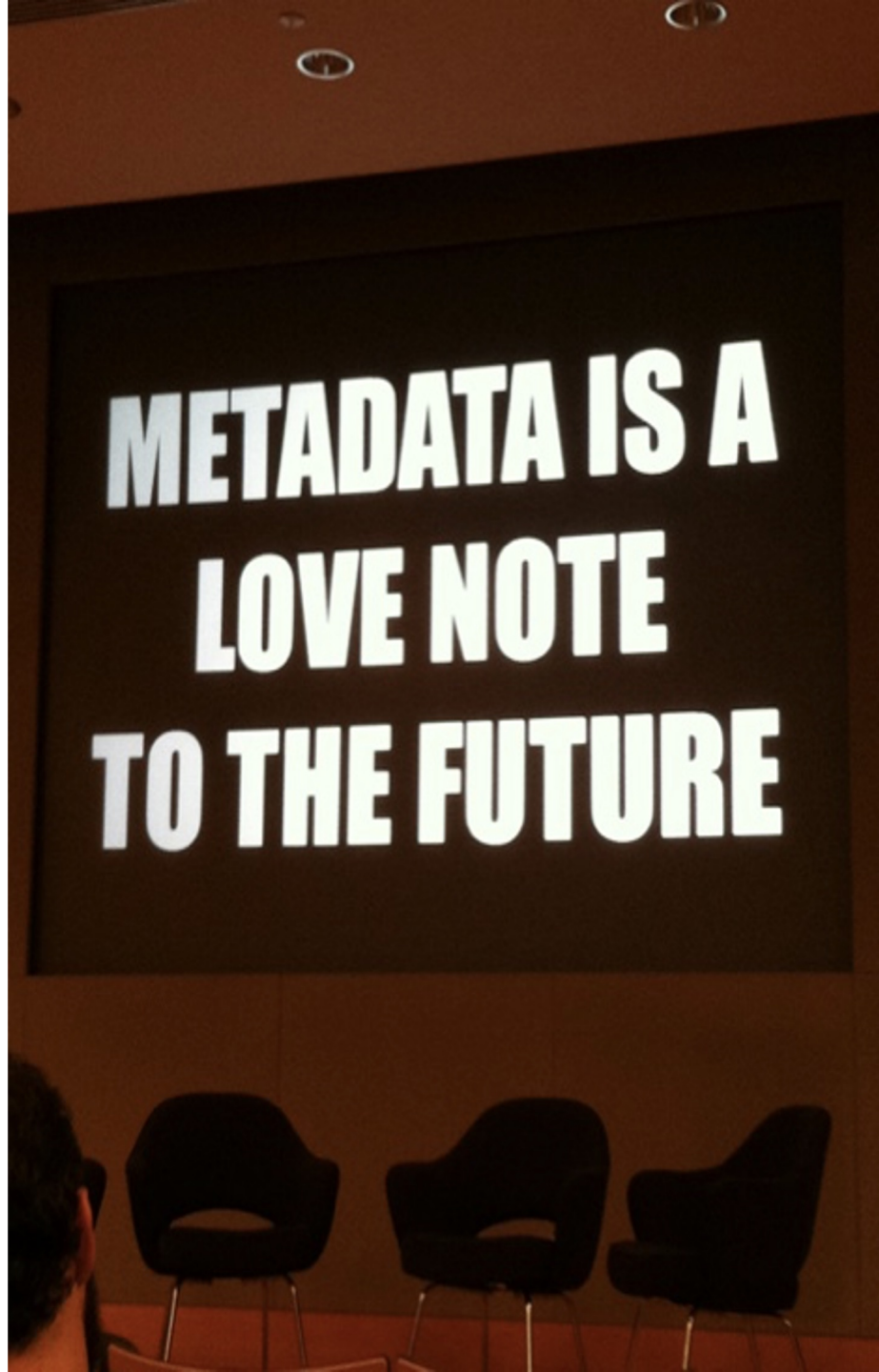
Open research data, available for reuse, strengthens open and **curiosity-driven research**, and scientific breakthrough not originally foreseen by the original data producer.



Illustration CC BY [Jørgen Stamp](#)

What is metadata?

Metadata, literally means **“data about data”** are an essential component of a data management system, describing such aspects as the “what, where, when, who and how” pertaining to a resource.



**METADATA IS A
LOVE NOTE
TO THE FUTURE**

Metadata should inform the user of the data:

- **Identify** & discover existence of data.
- How to **access** or acquire data.
- Understand its **fitness-for-use**.
- Learn how to obtain a copy of the data.
- Learn how the data should be used.



Metadata reduces “*Data entropy*”

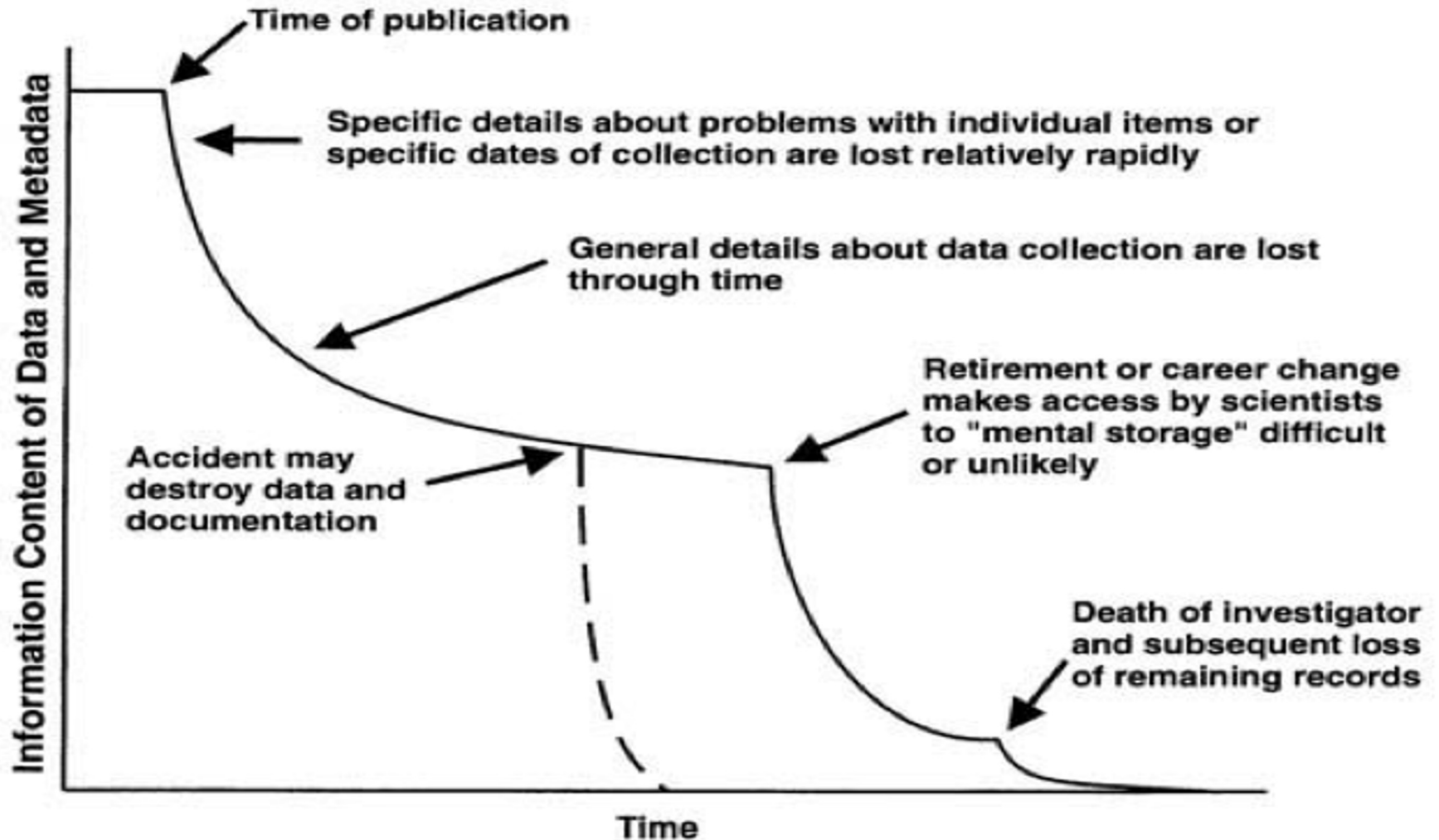
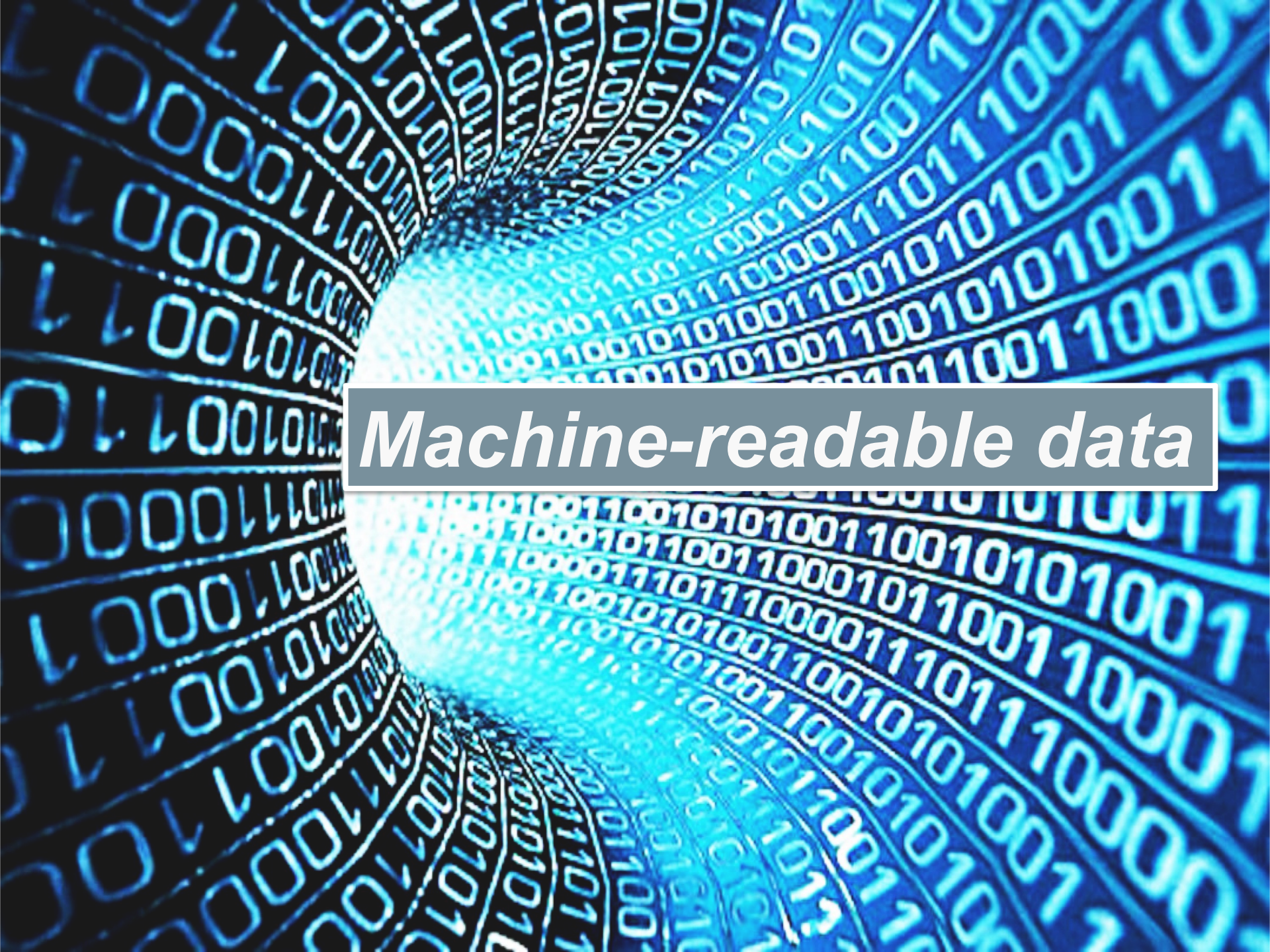


Illustration from: The Loss of Information about Data (Metadata) Over Time, Michener et al, 1997

What is a data paper?

- A **data paper** is a peer reviewed document describing a dataset, published in a peer reviewed journal.
- **It takes effort to prepare, curate and describe data.**
- Data papers provide **recognition** for this effort by means of a scholarly article.
- Getting scholarly recognition for improving the fitness for reuse for **your own datasets**.
- Getting scholarly recognition for preparing and making **legacy research data** available for reuse.



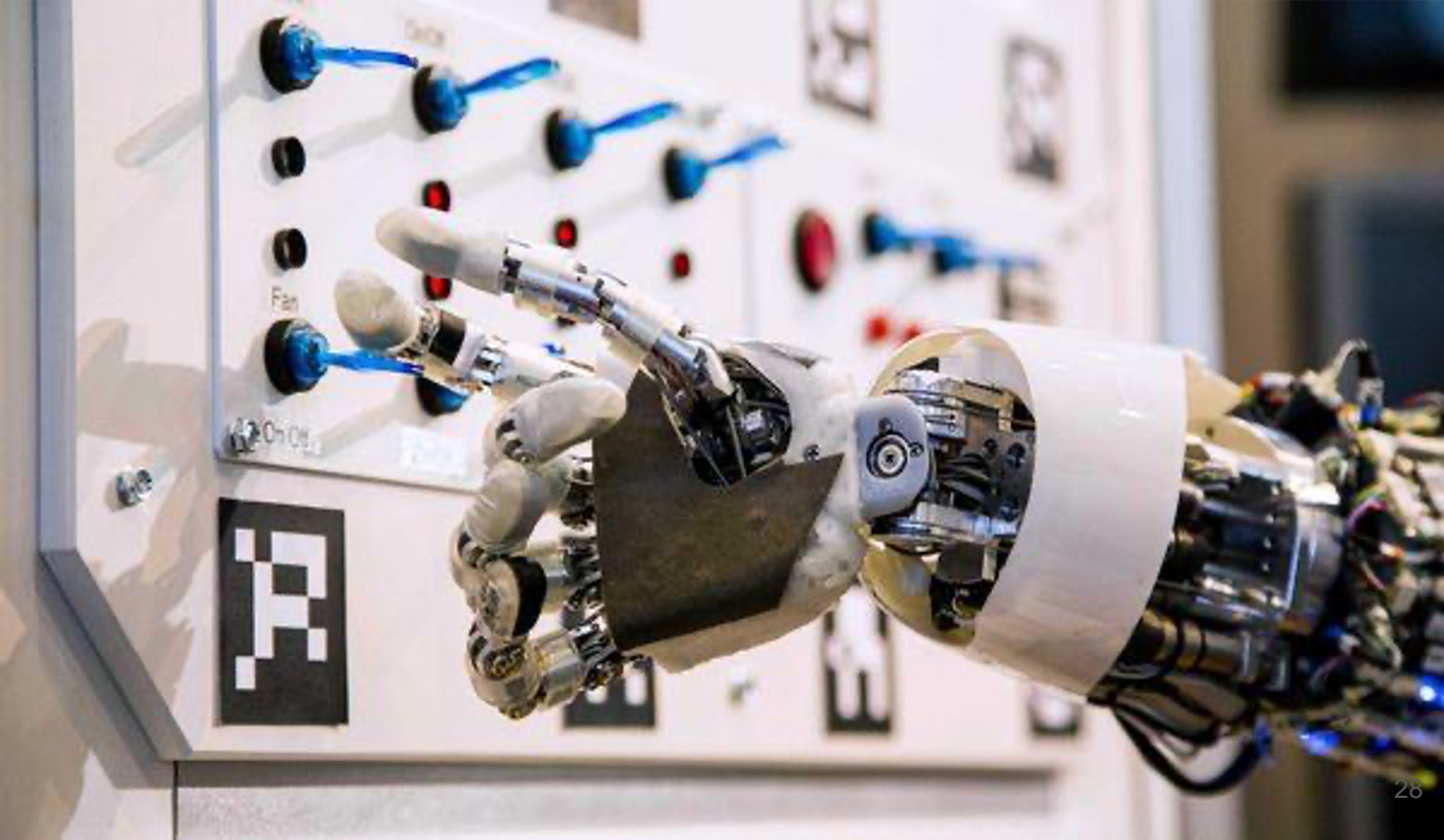
The background of the image is a vibrant blue digital tunnel. The walls of the tunnel are composed of numerous curved lines of binary code (0s and 1s) that recede into the distance, creating a strong sense of perspective and depth. The lighting is bright and ethereal, emanating from the center of the tunnel.

Machine-readable data

Scientific synthesis can be defined as “The inferential process whereby new models are developed from analysis of multiple data sets to explain observed patterns across a range of time and space scales” (Kemp and Boynton, 2011)



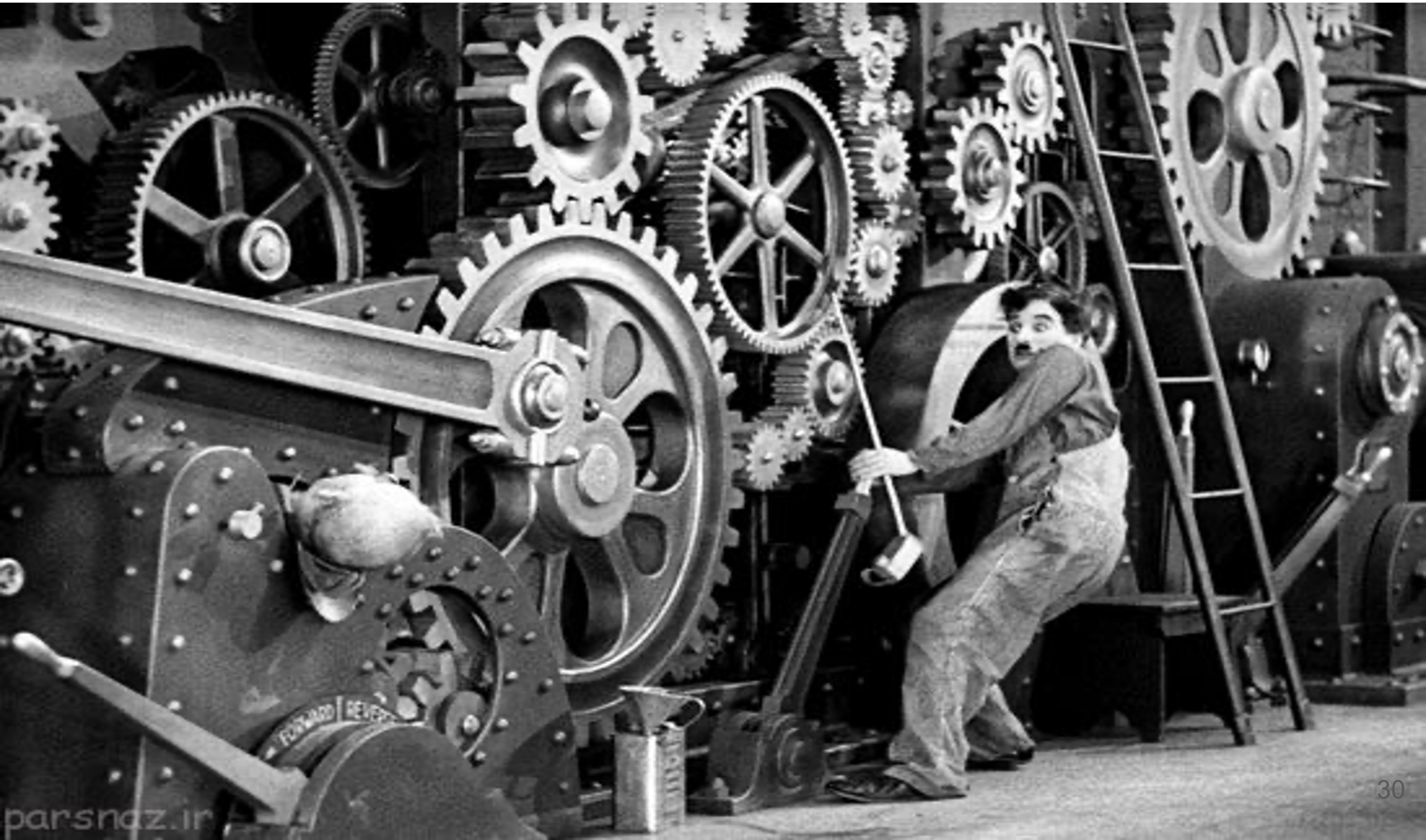
*"We are increasingly relying on **machines that derive conclusions** from models that they themselves have created, models that are often beyond human comprehension, models **that “think” about the world differently than we do**" ([David Weinberger 2017](#)).*

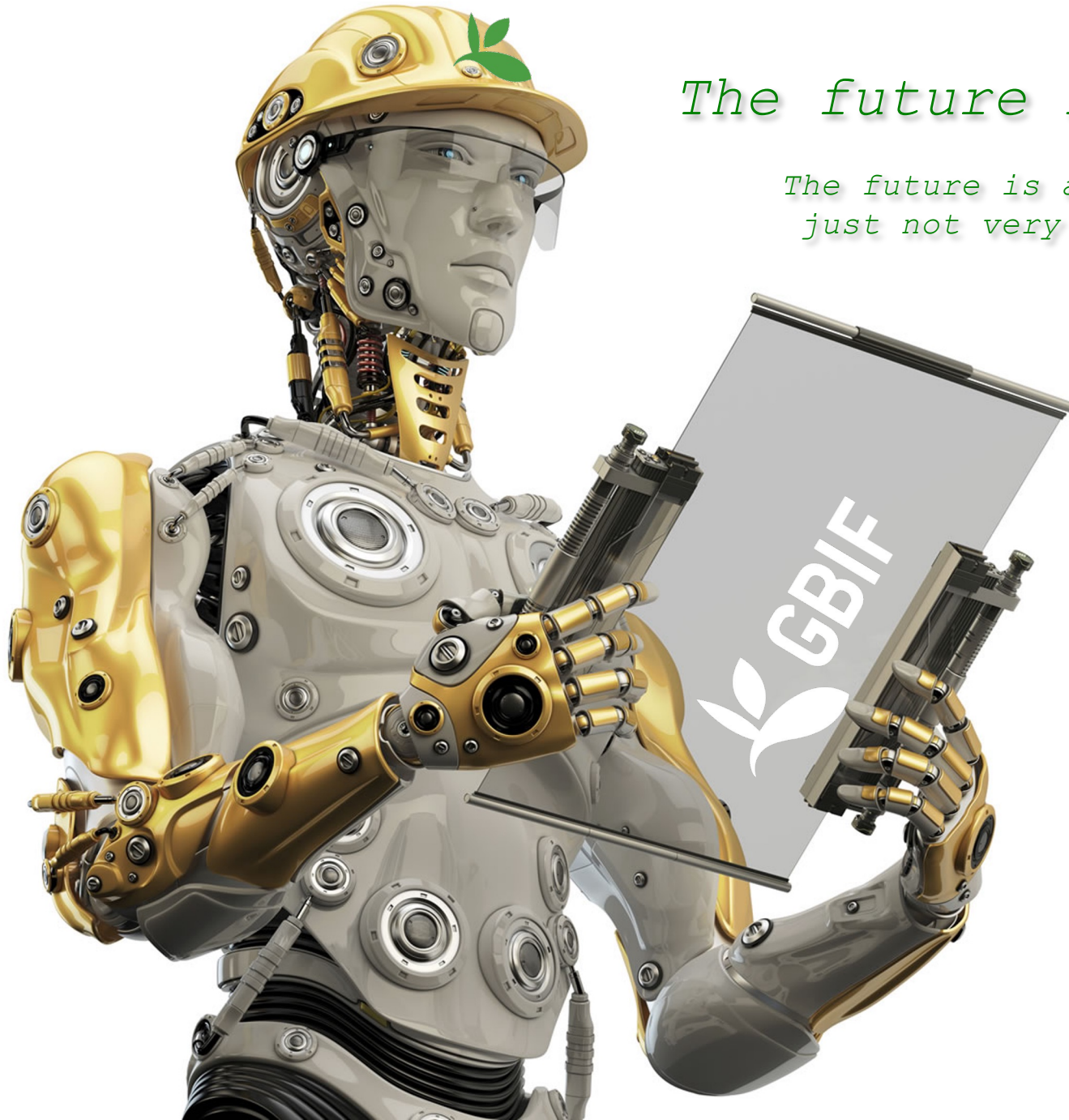




*"The new availability of **huge amounts of data**, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. **Correlation supersedes causation**, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all." (Chris Anderson, Wired Magazine, 2008)*

Semantic technologies for **integrating different types of biodiversity information** (museum collections, specimens, phenotypic trait data, & molecular data) are well advanced. **More efforts are needed to advance the uptake** for these information infrastructures to realize their potential.

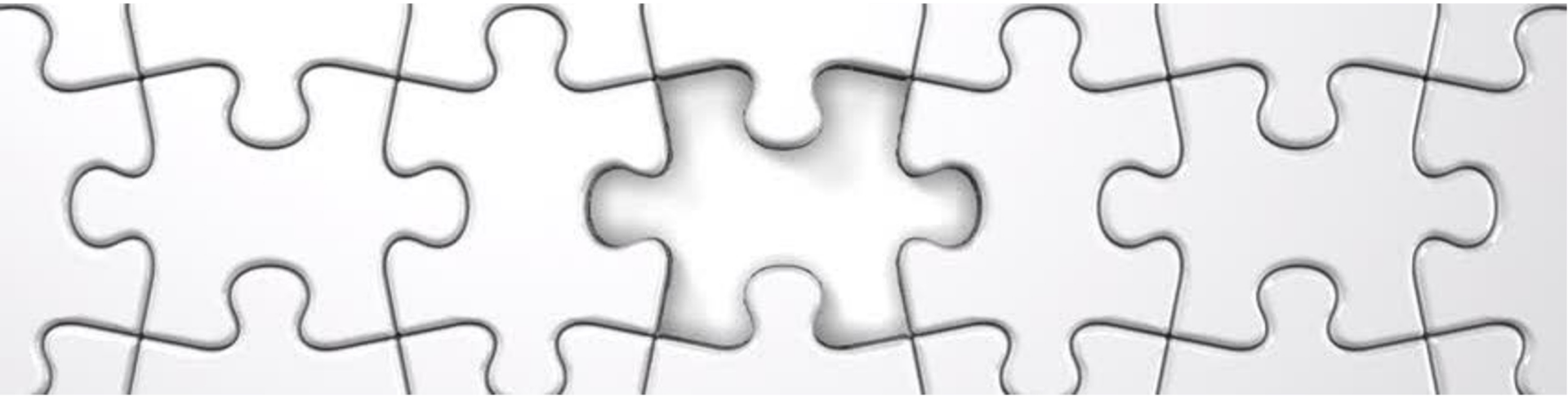




The future has arrived

*The future is already here – it's
just not very evenly distributed*

William Gibson



Open Research Data

Slides CC BY: Dag Endresen