Sustainability of Research Data Management

Dr Özlem ÖZKAN

Sustainability in the Digital Transformation of Basic Research on Universe & Matter

30 May 2023 - 2 June 2023
Landhaus Nordhelle, Meinerzhagen

www.helmholtz-metadata.de

Özlem Özkan

Bsc: Computer Education

Msc & PhD: Medical Informatics

Protection & Privacy of Medical and Genetic Data

Worked: Researcher, Data Scientist in KPMG Berlin &

Research Data Manager in Max Delbrück Center

Now: Data Policy Officer at Hub Matter of Helmholtz Metadata
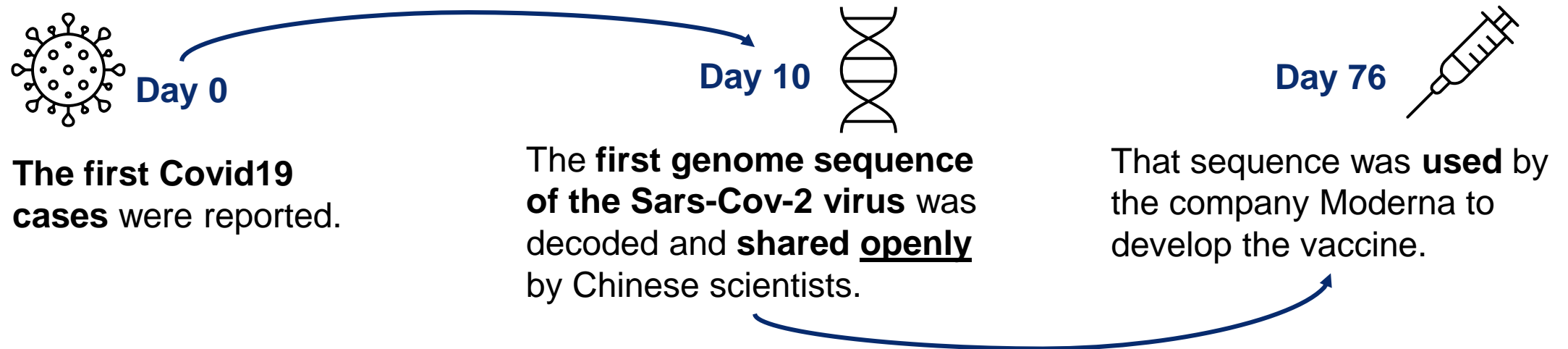Collaboration, Berlin

Provide services and tools for efficient metadata handling

Make Helmholtz Data FAIR - findable, accessible, interoperable and reusable

- Turning FAIR into reality* on all levels to enable reuse of data
- Develop HMC as a research infrastructure platform!

From: Turning FAIR into Reality, Final Report and Action Plan from the European Commission Expert Group on FAIR Data , doi: 10.2777/1524

Day 0

Day 10

Day 76

The first Covid19 cases were reported.

The **first genome sequence of the Sars-Cov-2 virus** was decoded and **shared openly** by Chinese scientists.

That sequence was **used** by the company Moderna to develop the vaccine.

We saw the **impact** and **effect of Open Science in a real life** example**!**

Recently, **the science ministers of the G7 countries** (2023) and **the US President Office of Science and Technology Policy** (2022) published their strong commitments to Open Science:

2. Learning from the Lessons of COVID-19

When federally funded research is available to the public, it can improve lives, provide policymakers with important evidence with which to make critical decisions, accelerate the rates of discovery and translation, and drive more equitable outcomes across every sector of society.

Americans were offered a window into the great benefits of immediate public access to federally funded research at the outset of the COVID-19 pandemic. In the wake of the public health crisis, government, industry, and scientists voluntarily worked together to adopt an immediate public access policy, which yielded powerful results: research and data flowed effectively, new accessible insights super-charged the rate of discovery, and translation of science soared. The shift in practice during COVID-19 demonstrated how delivering immediate public access to federally funded research publications and data can provide near real-time returns on American taxpayer investments in science and technology.

Communiqué. G7 Leaders' Summit, 2023. (2023, May 13) Retrieved from
https://www8.cao.go.jp/cstp/kokusaiteki/g7_2023/230513_g7_communique.pdf

...icy. (2022). Retrieved from
...OSTP-Public-Access-Memo.pdf

Open Science is also embraced by **Horizon Europe**, the **European Commission's** new Framework Program for Research and Innovation.

Today many collaborative initiatives promote Open Science as the **'new normal'!** Like:

In 1987, the United Nations defined sustainability as "meeting the needs of the present without compromising the ability of future generations to meet their own needs."

Cracks run through the partially dried-up river bed of the Gan River, a tributary to Poyang Lake during a regional drought in Nanchang, Jiangxi province, China. (REUTERS/Thomas Peter/File Photo)

Sustainable Research Data Management

Sustainable research data management refers to

the practices and approaches employed to ensure

- the long-term availability,

- accessibility,

- usability &

- integrity of research data

while <u>minimizing negative environmental, social and</u>

<u>economic impacts</u>.

Most Scientific Research Data From the 1990s Is Lost Forever

A new study has found that as much as 80 percent of the raw scientific data collected by researchers in the early 1990s is gone forever, mostly because no one knows where to find it.

By Danielle Wiener-Bronner

https://www.theatlantic.com/national/archive/2013/12/scientific-data-lost-forever/356422/

Sustainable Research Data Management

Good data management helps to **prevent errors** and **increases the quality of the analyses, saves time** and **resources** in the long run, so it <u>minimizes negative environmental, social, and economic</u> effects.
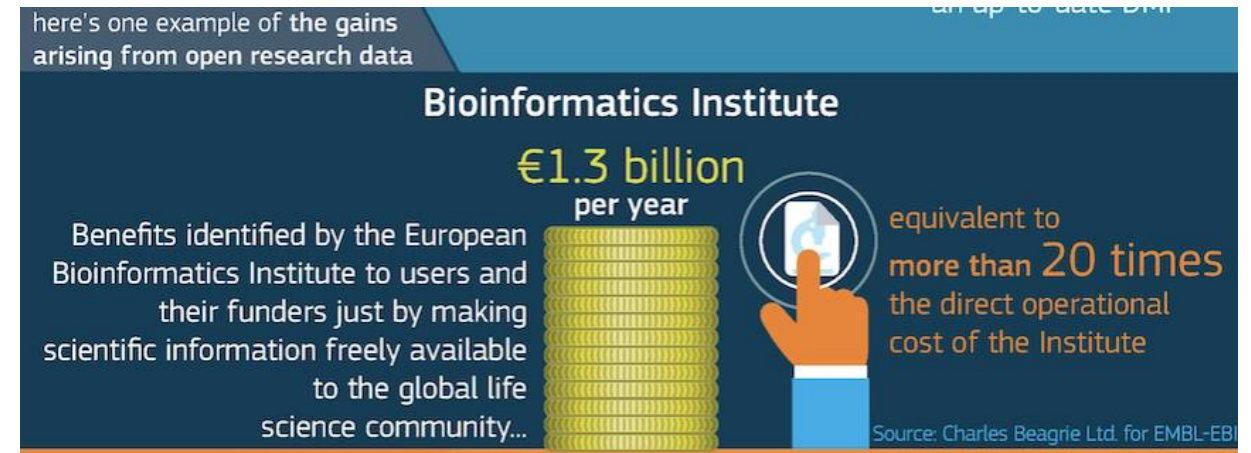
European Commission

In May 2018, the EU Commission published a report: not having FAIR research data costs the European economy at least **€10.2bn/**year

Cost of not having FAIR research data

Cost-Benefit analysis for FAIR research data

http://publications.europa.eu/resource/cellar/d375368c-1a0a-11e9-8d04-01aa75ed71a1.0001.01/DOC_1

here's one example of **the gains arising from open research data**

an up-to-date DMP

Bioinformatics Institute

€1.3 billion per year

Benefits identified by the European Bioinformatics Institute to users and their funders just by making scientific information freely available to the global life science community...

equivalent to **more than 20 times** the direct operational cost of the Institute

Source: Charles Beagrie Ltd. for EMBL-EBI

https://www.embl.org/documents/wp-content/uploads/2021/10/EMBL-EBI-impact-report-summary-2021.pdf

Data sustainability

involves ensuring that data remains **accessible, usable** & **valuable** over time.

FAIR data principles

aim to maximize the **usefulness** and **impact of research data** by enabling it to be **easily shared** and **reused** by others.

Open Science Training Handbook https://doi.org/10.5281/zenodo.1212496
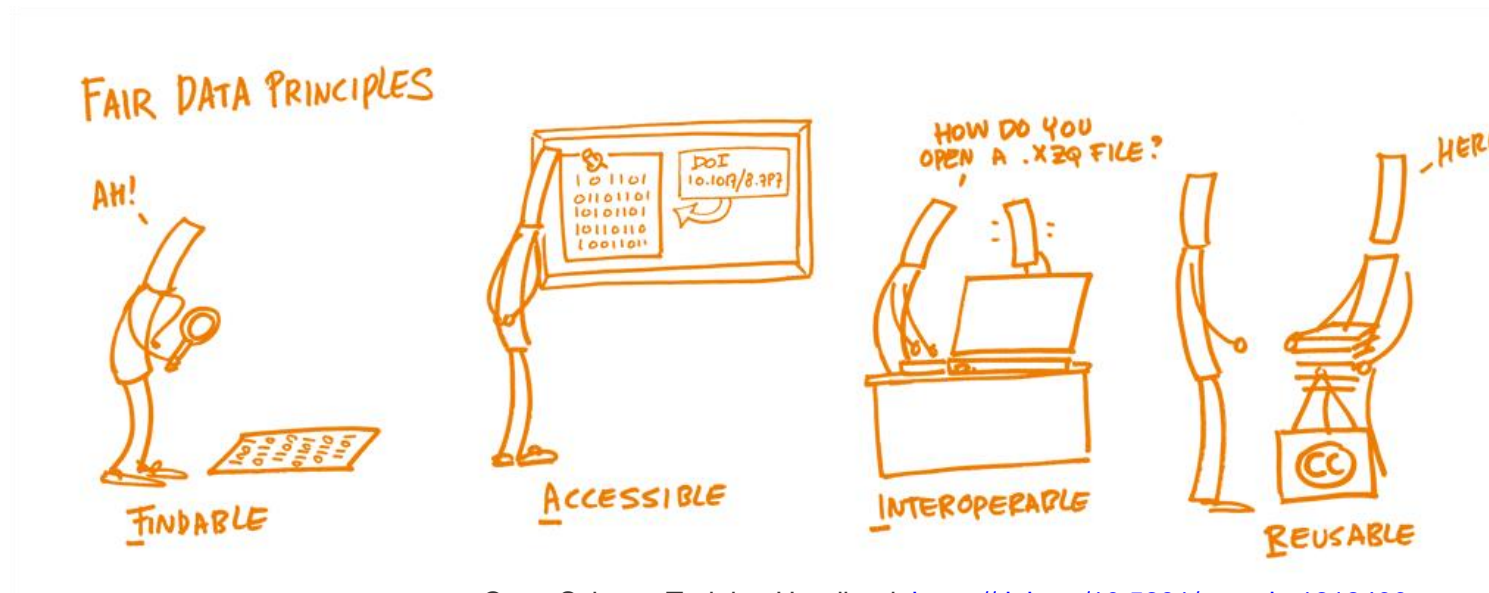
FAIR DATA properly documented + easily searchable

discoverability and accessibility
(key components of data sustainability)

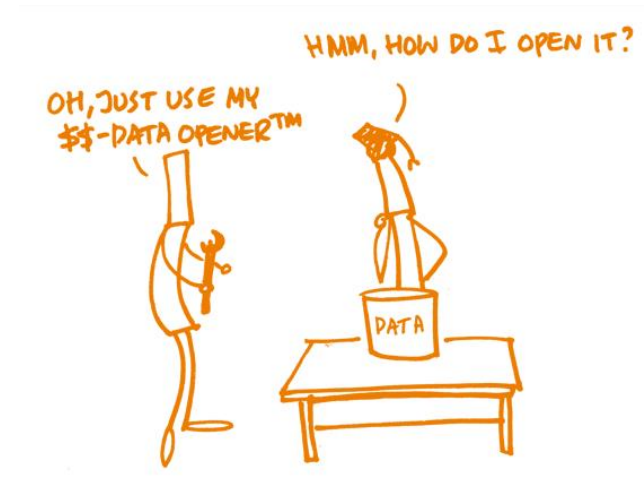FAIR DATA stored in open + interoperable formats:

easily reusable

=

likely to remain **relevant** and **valuable** over time
(usability)

Open Science Training Handbook
https://doi.org/10.5281/zenodo.1212496

Open Science Training Handbook
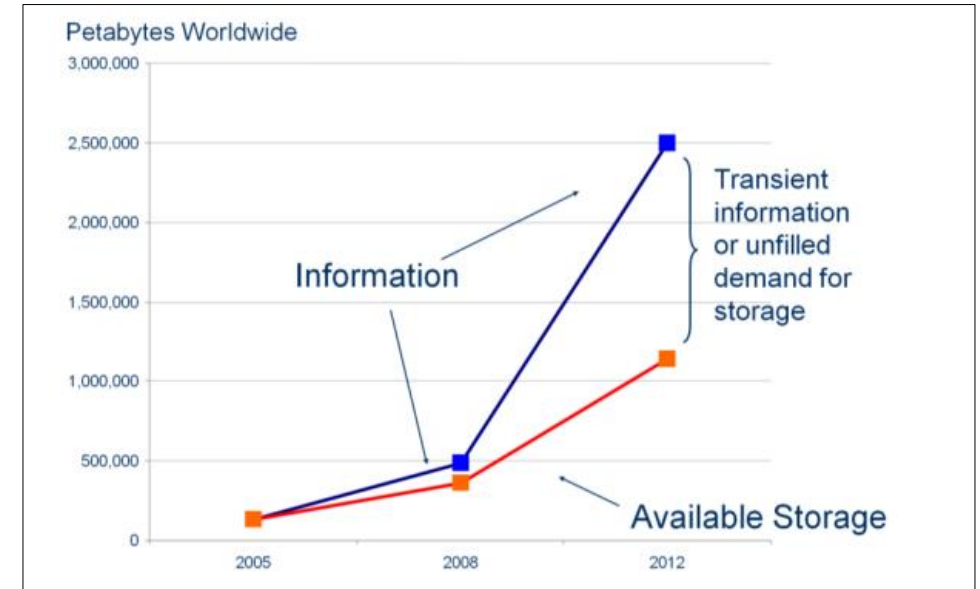https://doi.org/10.5281/zenodo.1212496

Sustainability of data was a problem in the past as well

But the main challenge was **economic**: there wasn't a balance between available storage and the volume of the data
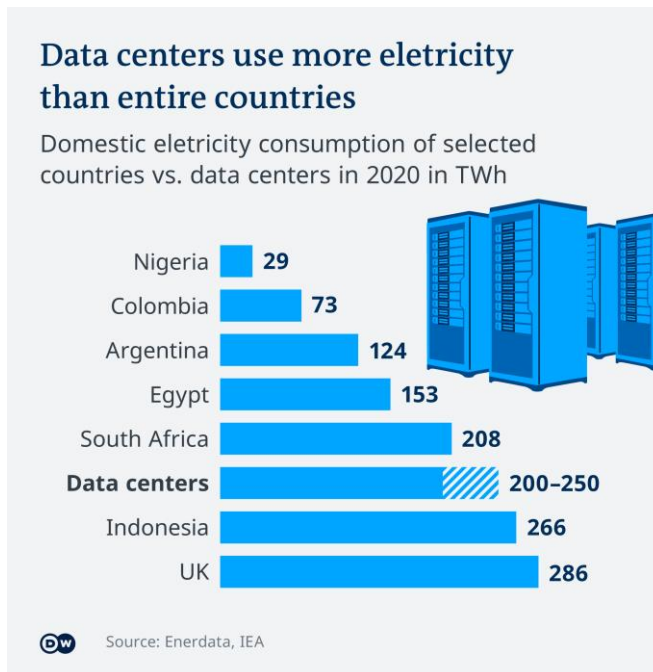
Now it is <u>not</u> an issue anymore

We have enough storages but **sustainability is still a problem**

Which digital information should we preserve?

- Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (2010). Sustainable Economics for a Digital Planet: Ensuring Long-term Access to Digital Information.

- Berman, F., & Cerf, V. (2013). Who will pay for public access to research data?. Science, 341(6146), 616-617. dsc.edu/biblio/BRTF_Final_Report.pdf

- Beagrie, N., & Houghton, J. W. (2014). The Value and Impact of Data Sharing and Curation A synthesis of three recent studies of UK research data centres.

- Dillo, I., Treloar, A., Lusoli, W., Kupiainen, I., Asch, M., Diepenbroek, M., ... & Harrower, N. (2017). Business models for sustainable research data repositories. OECD SCIENCE, TECHNOLOGY AND INNOVATION POLICY PAPERS.

According to the UK Data Service Guide on <u>Research data management and costs</u> :

the storage of an average project's data costs **60 Euros/hour**.

Data centers use more eletricity than entire countries

Domestic eletricity consumption of selected countries vs. data centers in 2020 in TWh

Nigeria 29
Colombia 73
Argentina 124
Egypt 153
South Africa 208
Data centers 200–250
Indonesia 266
UK 286

Source: Enerdata, IEA

According to the International Energy Agency:

data centers globally accounted for about **1% of global electricity consumption** in 2020

- Science Europe has been **particularly focusing on sustainability of research data** since ~2019.

- They published a guide for research funders, research organisations and research infrastructures in June 2021: „**Practical Guide to Sustainable Research Data"**
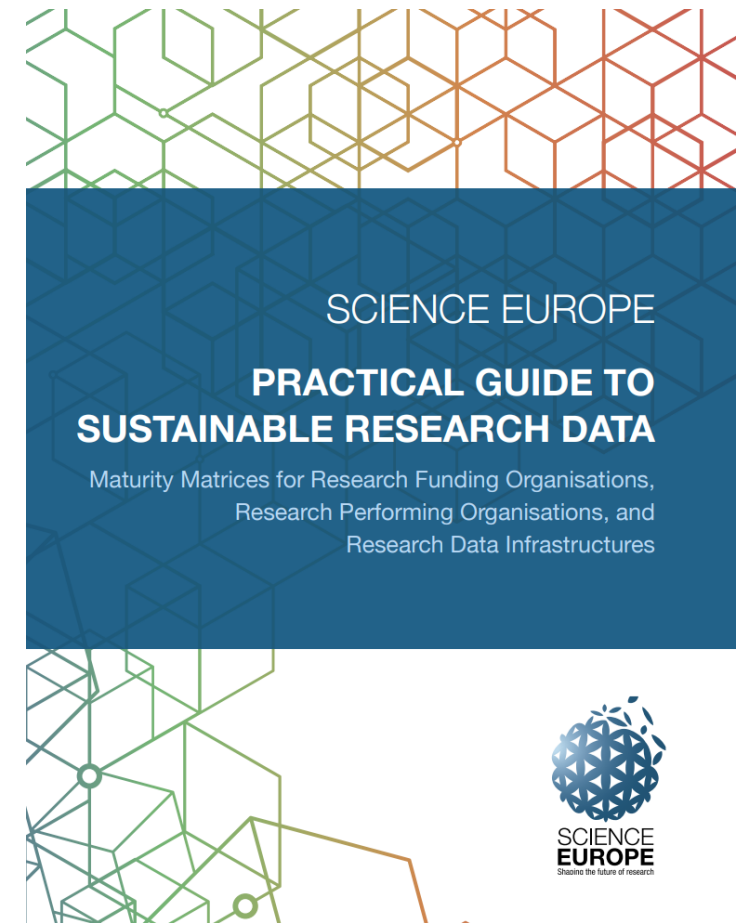
What should organisations do?

- Data Policies & Guidelines

- Training

- Roadmap for RDM

- Community engagement

Sustainable Research Data in Research Organizations (ROs):

Organizations should have **data policies** and

guidelines to

- clarify its objectives for **data sustainability**

- commit **giving the necessary support** to its

  researchers

Policy GEOMAR

Policy AWI

Policy Hereon

Policy MDC

Policy HZB

Policy HZI

Policy GFZ

Policy UFZ

Policy FZJ

Policy DLR

Policy HZDR

Policy DZNE

Policy GSI

Policy KIT

Research data policies of the Helmholtz Centers

Organizations should **provide training** on RDM and:

- define **core Research Data Management competency levels**

- have a complete **training** and **support** system with support personnel*: data managers, data stewards, librarians, etc.*

Sustainable Research Data in Research Organizations (ROs):

Planning **budget** and **technical necessities**:

- define an initial budget allocation for RDM and areas for spending.

- have a RDM technical plan to ensure FAIR data: i.e University of Edinburgh has detailed Roadmaps for RDM

Data Management Planning
Support and services for planning activities that are typically performed before research data is collected or created.

Addresses RDM policy clauses 3, 4.

Objective	Outcomes	Actions	Deliverables	Target date
1. Tailored DMP assistance for PIs submitting research proposals	Better alignment between IS, ERI and schools PIs have access to appropriate, accurate information about IS RDM services for their grant proposals	Analyse recent DMPs in research proposals Trial fast turnaround advisory service with nominated PIs	Set of successful examples of submitted DMPs gathered from ERI and schools DMP 'response team' established within IS if deemed necessary	Phase 3 Phase 4
2. Customise DMP Online for optimal UoE use	Clear online service provision for those PIs needing a data management plan	Evaluate DMPonline and analyse user requirements Refine tool or fully adopt the DCC version, adding UoE specific questions and guidance Advertise to schools that customised DMP templates can be created for their researchers as required Create customised DMPonline templates for schools on request Metrics and reporting of DMP creation and use	University branded DMP Online tool Boilerplate text about IS services for use in DMPs Research support staff and research committees are aware of the potential to customise it for their school. Respond to demand from schools as it arises Monthly reports on number of new DMP's created	Phase 1 (Completed) Phase 3 (Underway) Phase 3 (Underway) Ongoing Ongoing

RDM Roadmap University of Edinburgh http://www.ed.ac.uk/schools-departments/information-services/about/strategy-planning/rdm-roadmap

Having active **community engagement,**

- Researchers are involved in process: Data managers do <u>not</u> know **what scientists need.**
  Only researchers know which data should be kept for reusing.

- Organizations work with research communities to **identify appropriate metadata standards** and **required infrastructure** addressing the disciplinary needs.

- NeXus is an effort by an international group of scientists to define a **common data exchange** and **archival format** for neutron, X-ray and muon experiments.

- It is developed as **an international standard** by scientists and programmers **without any significate funding.**

Helmholtz Zentrum Berlin (HZB) uses NeXus standard:

- The center provides **one staff** to work in the advisory committee in the NeXus project.

- This person communicates with **HZB's scientists** and identifies the needs of the researchers.

- The committee examines all proposed amendments and extensions in the NeXus format.

- They decide to extend the format accordingly.

NeXus International Advisory Committee

Current Members

- Aaron Brewster, Lawrence Berkeley Laboratory, USA (*Chair*)
- Sandor Brockhauser, Center for Materials Science Data (HU Berlin), Germany (*Executive Secretary*)
- Peter Chang, Diamond Light Source, UK (*Documentation Release Manager*)
- Benjamin Watts, Swiss Light Source, Switzerland
- Mark Könnecke, PSI, Switzerland
- Herbert Bernstein, CIF (*non-facility member*)
- Thomas Caswell, Brookhaven National Laboratory/NSLS-II, USA
- Bjørn Clausen, Los Alamos National Laboratory, USA
- Heike Görzig, Helmholtz Zentrum Berlin, Germany
- Pete Jemian, Advanced Photon Source, USA
- Raymond Osborn, Argonne National Laboratory, USA (*non-facility member*)
- Wout de Nolf, European Synchrotron Radiation Facility, France
- Takahiro Matsumoto, Spring8, Japan
- Balázs Bagó, the Extreme Light Infrastructure, Czech Republic, Hungary and Romania
- Russ Berg, Canadian Light Source, Canada
- Majid Ounsy, Synchrotron Soleil, France
- Chen Zhang, Oak Ridge National Laboratory (SNS/HFIR), USA
- Luca Geliso, European XFEL, Germany
- Paul Millar, DESY, Germany

https://www.nexusformat.org/NIAC.html

So the scientists all over the world, who are working in the same area, are using and will probably be using the same format for their data.

This means,

- They guarantee that data will be **accessible** by the next generations in the same field: Once researchers have found the data, they should be able to access it by **using standardized mechanisms.**

- The data becomes **reusable**: Keeping **in line with these standards** will lead new data to be reused easily and suitably.

Thank you very much!

HMC