Recent Survey on RDM Practices

HELMHOLTZ Metadata Collaboration

https://oceanrep.geomar.de/id/eprint/57494/

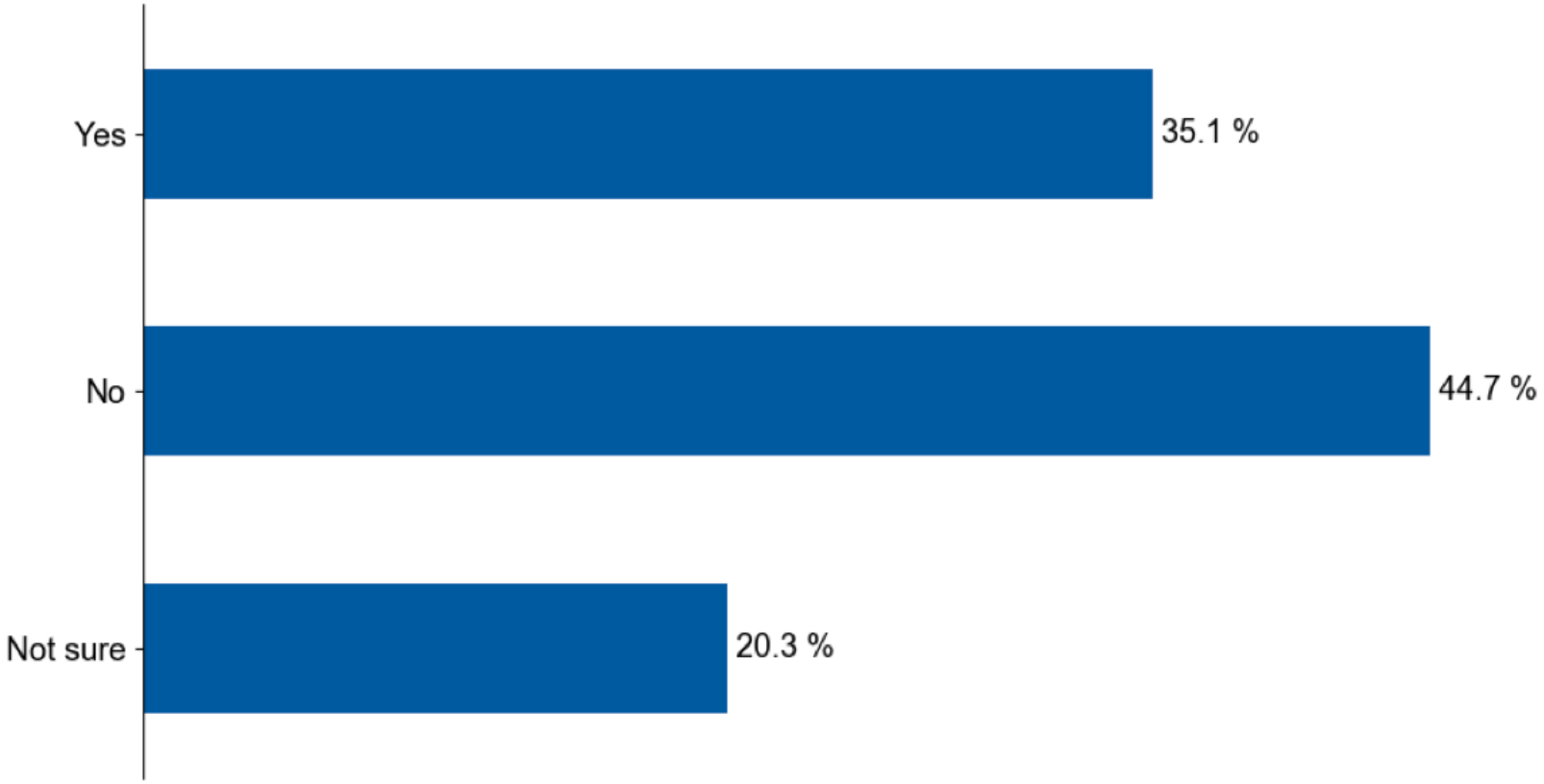


Figure 14: "Do you document your research data in a structured way? (e.g., using forms, templates or schemas)". (Single choice question, available to all respondents, number of respondents who answered this question: n = 582, relative amounts refer to n)

With massive “thanks” to M. Demleitner, H. Enke, P. Fuhrmann, A. Geiser, G. Günther, A. Haungs, M. Köhler, O. Mannix, S. Servan, C. Wissing + others!

(Thoughts on) Requirements on FAIR Data Management

Thomas Schörner (DESY)

JENA Computing Workshop

Bologna, 12-14 June 2023

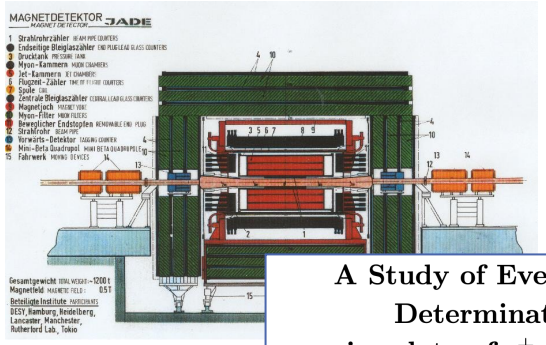
HELMHOLTZ ASSOCIATION

DFG Deutsche Forschungsgemeinschaft

DESY.

Introduction
Data Preservation,
FAIR Principles etc.

Instead of an Introduction (1)



A Study of Event Shapes and Determinations of α_s using data of e^+e^- Annihilations at $\sqrt{s} = 22$ to 44 GeV

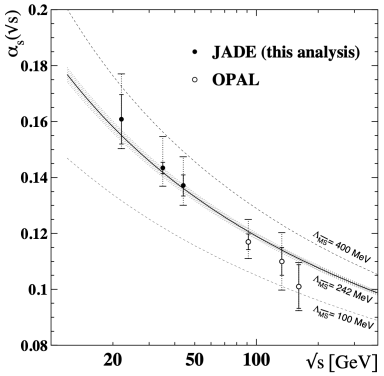
P.A. Movilla Fernández[(1)], O. Biebel[(1)], S. Bethke[(1)], S. Kluth[(2)], P. Pfeifenschneider[(1)] and the JADE Collaboration[(3)]

Abstract

ed by the JADE experiment at the PETRA e^+e^- collider were used the event shape observables thrust, heavy jet mass, wide and total jet and the differential 2-jet rate in the Durham scheme. For the latter three no experimental results have previously been presented at these energies. tions were compared with resummed QCD calulations ($\mathcal{O}(\alpha_s^2)$+NLLA), ng coupling constant $\alpha_s(Q)$ was determined at different energy scales e results,

$= 0.161^{+0.016}_{-0.011}$, $\alpha_s(35\text{ GeV}) = 0.143^{+0.011}_{-0.007}$, $\alpha_s(44\text{ GeV}) = 0.137^{+0.010}_{-0.007}$,

nent with previous combined results of PETRA albeit with smaller un- Together with corresponding data from LEP, the energy dependence of antly tested and is found to be in good agreement with the QCD expec- ilarly, mean values of the observables were compared to analytic QCD where hadronisation effects are absorbed in calculable power corrections.



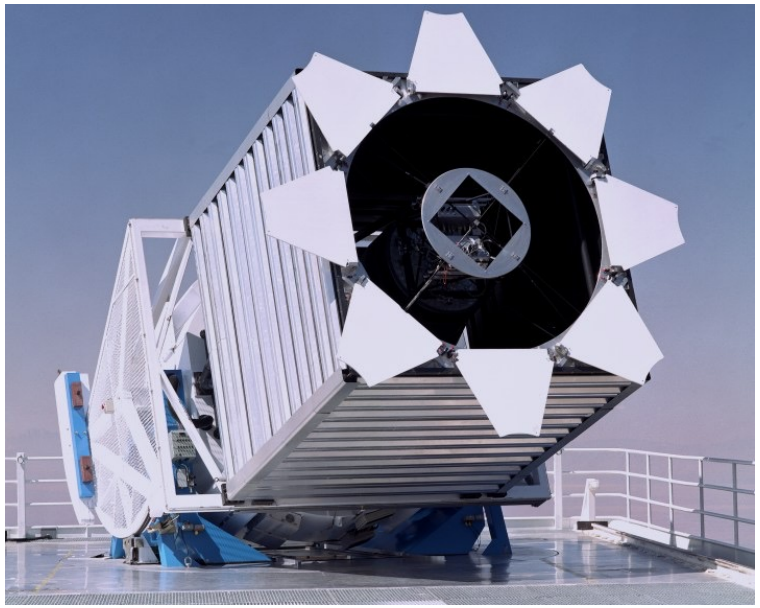
arXiv:hep-ex/9708034
Eur.Phys.J. C1 (1998) 461

JADE experiment at DESY's PETRA e+e- collider (1979-86) E.g. discovery of three-jet events → gluon discovery

Re-use and re-interpretation →new insights!

Herschel's discovery of Uranus: It is a not a star, and discovery of Neptune!

Interpretation in light of new discoveries → far-reaching conclusions.



SDSS designed to determine beta parameter, focus on galaxies. Later numerous discoveries concerning stars.

New by-product results that enter the focus only later.

Instead of an Introduction (2): ZEUS and DPHEP

International Conference on Computing in High Energy and Nuclear Physics 2012 (CHEP2012) IOP Publishing
Journal of Physics: Conference Series **396** (2012) 022033 doi:10.1088/1742-6596/396/2/022033

The ZEUS data preservation project

Janusz Malka and Katarzyna Wichmann on behalf of the ZEUS collaboration
Deutsches Elektronen-Synchrotron, Notkestr 85, 22607 Hamburg, Germany
E-mail: `janusz.malka@desy.de,katarzyna.wichmann@desy.de`

Abstract. A project to allow long term access and physics analysis of ZEUS data (ZEUS data preservation) has been established in collaboration with the DESY-IT group. In the ZEUS approach the analysis model is based on the Common Ntuple project, under development since 2006. The real data and all presently available Monte Carlo samples are being preserved in a flat ROOT ntuple format. There is ongoing work to provide the ability to simulate new, additional Monte Carlo samples also in the future. The validation framework of such a scheme using virtualisation techniques is being explored. The goal is to validate the frozen ZEUS software against future changes in hardware and operating system. A cooperation between ZEUS, DESY-IT and the library was established for document digitisation and long-term preservation of collaboration web pages. Part of the ZEUS internal documentation has already been stored within the HEP documentation system INSPIRE. Existing digital documentation, needed to perform physics analysis also in the future, is being centralised and completed.

Data Preservation in High-Energy Physics

Study Group for Data Preservation and Long Term Analysis in High Energy Physics

http://dphep.org

Table 1. The DPHEP preservation modes listed in order of increasing complexity.

Preservation Model	Use Case
1. Additional information	Publication related information
2. Provide data in simplified format	Outreach, training
3. Preserve the analysis level software and data format	Full scientific analysis possible, based on existing reconstruction
4. Preserve the full simulation and reconstruction software as well as the basic level data	Retain the full potential of the experimental data

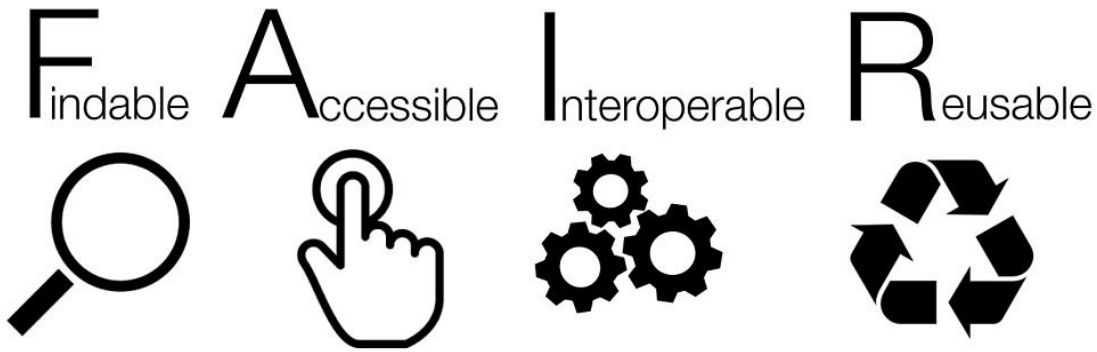
dphep.org, arXiv:0912.0255

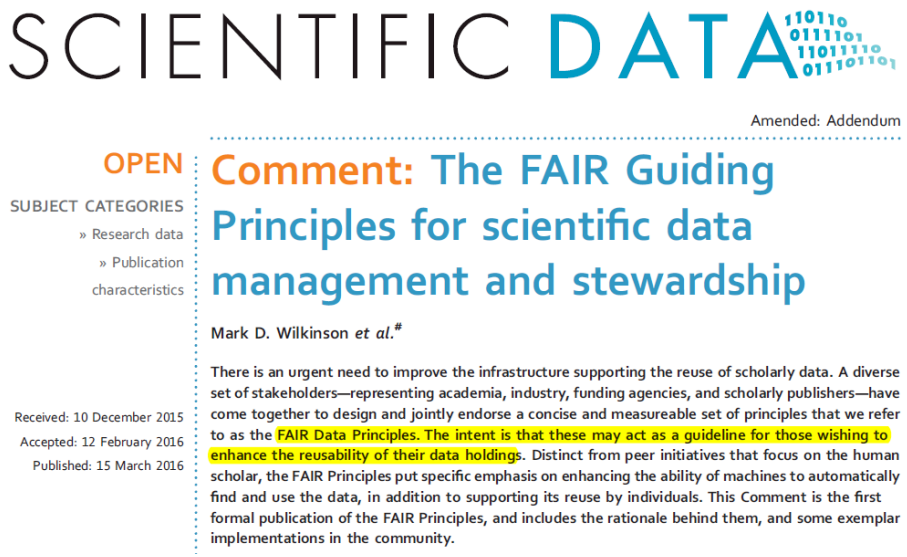
One important conclusion: Computers and storage alone (today: federated computing) are not enough; software and (meta)data management are as important.

Instead of an Introduction (3): The FAIR Principles

Encountering problems

to find data to access data to understand data to (re)use data





SCIENTIFIC DATA

Amended: Addendum

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*[#]

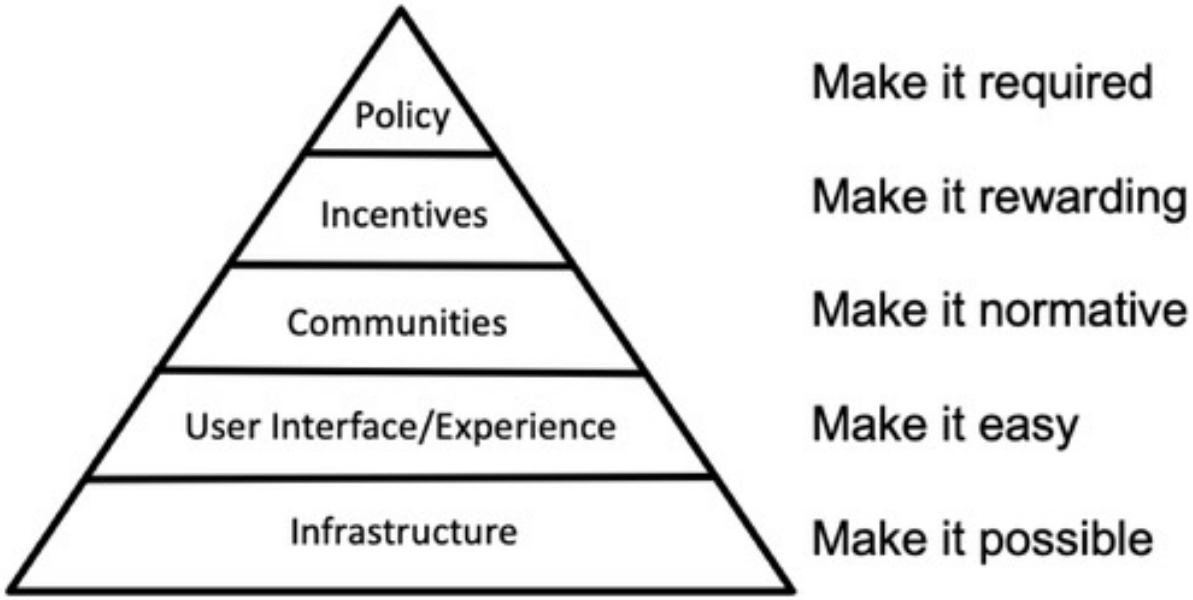
There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

FAIR Guiding Principles (2016)

M. D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data* 3.1 (2016): 1-9.

What to do?

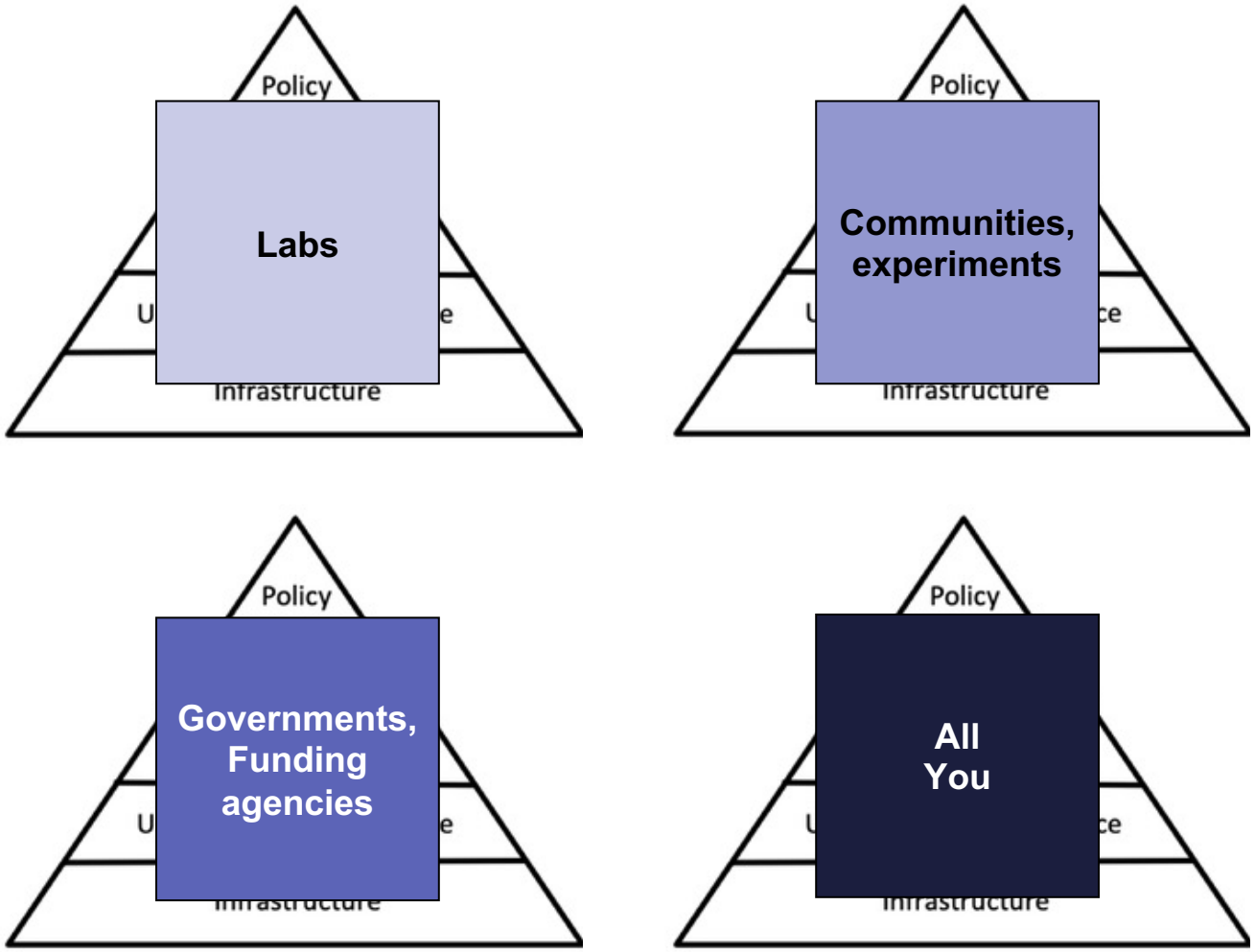
The pyramid of cultural change – different players involved



https://www.cos.io/blog/strategy-for-culture-change

What to do?

The pyramid of cultural change – different players involved



https://www.cos.io/blog/strategy-for-culture-change

NFDI & PUNCH4NFDI

Helmholtz Metadata Collaboration HMC

Gemeinsame Wissenschaftskonferenz GWK

Funded by

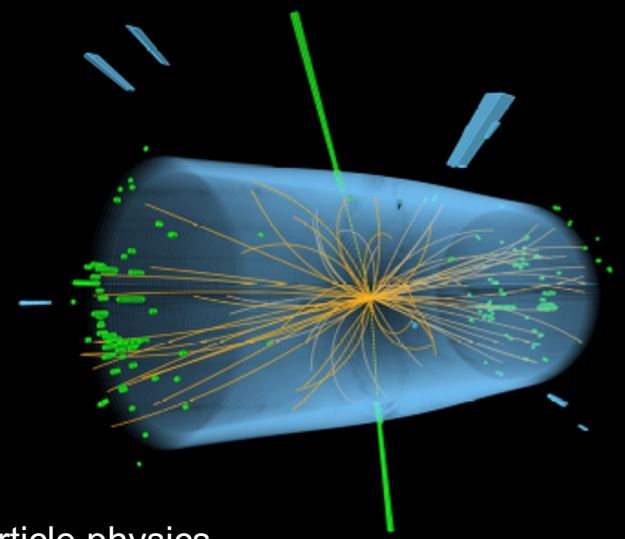
DFG Deutsche Forschungsgemeinschaft

German Research Foundation

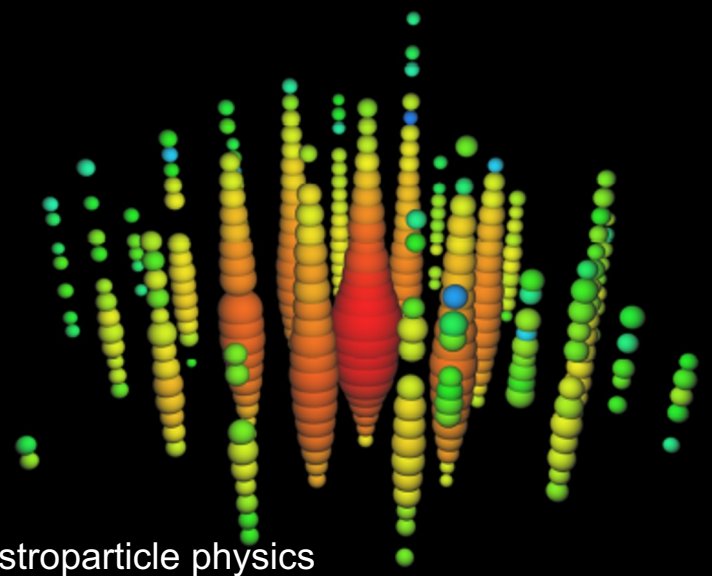
Nationale Research Data Infrastructure (NFDI)

- Sustainable utilisation of research data
- Establishment of FAIR RDM
- Connection to Europe & the world
- Bottom-up approach: 27 consortia
- Base service initiative
- 5 (+5) year funding (PUNCH4NFDI: 3,5MEUR/a)

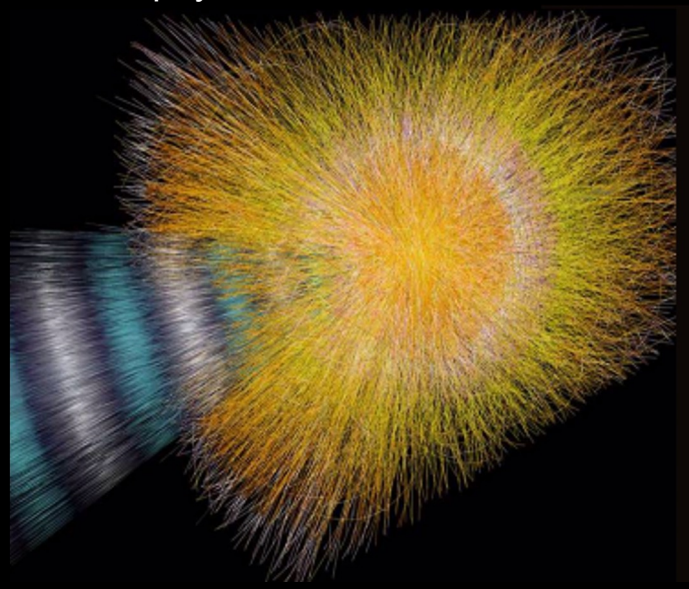
See also DFG.de/nfdi and nfdi.de



Particle physics



Astroparticle physics



Hadron & nuclear physics

Astronomy

Particles, Universe, NuClei and Hadrons for the NFDI

The prime goal of PUNCH4NFDI is the setup of a federated and "FAIR" science data platform, offering the infrastructures and interfaces necessary for the access to and (re)use of (meta)data and computing resources of the involved communities and beyond.

NFDI Consortia

LIFE SCIENCE RI

ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

panosc
photon and neutron open science cloud

SSHOpenCluster
Social Sciences & Humanities Resources

ENVRI community

Base4NFDI – a Base Service Initiative Across Consortia

base4nfdi Basic Services for NFDI

HOME | WHY B4N | SUBMISSIONS | HOW B4N WILL DECIDE | WHO CAN APPLY AND HOW | FURTHER INFO

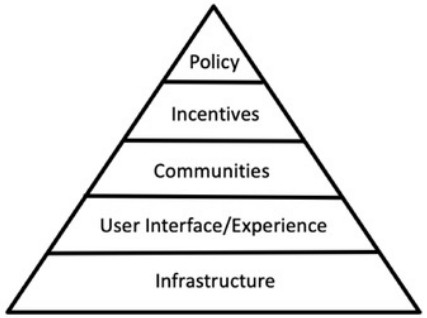
- Framework for user-driven basic service development
- Starting with IAM, PID services, terminology services
- Building on existing solutions and complementing EOSC

Basic Services for NFDI

Create NFDI-wide basic services in a world of specific domains

Helmholtz Metadata Collaboration

- Make Helmholtz data **FAIR**
- Provide services for **sustainable** and efficient metadata handling
- Develop, share and **consolidate community-expertise** in metadata across Helmholtz
- Address **all levels** of change pyramid



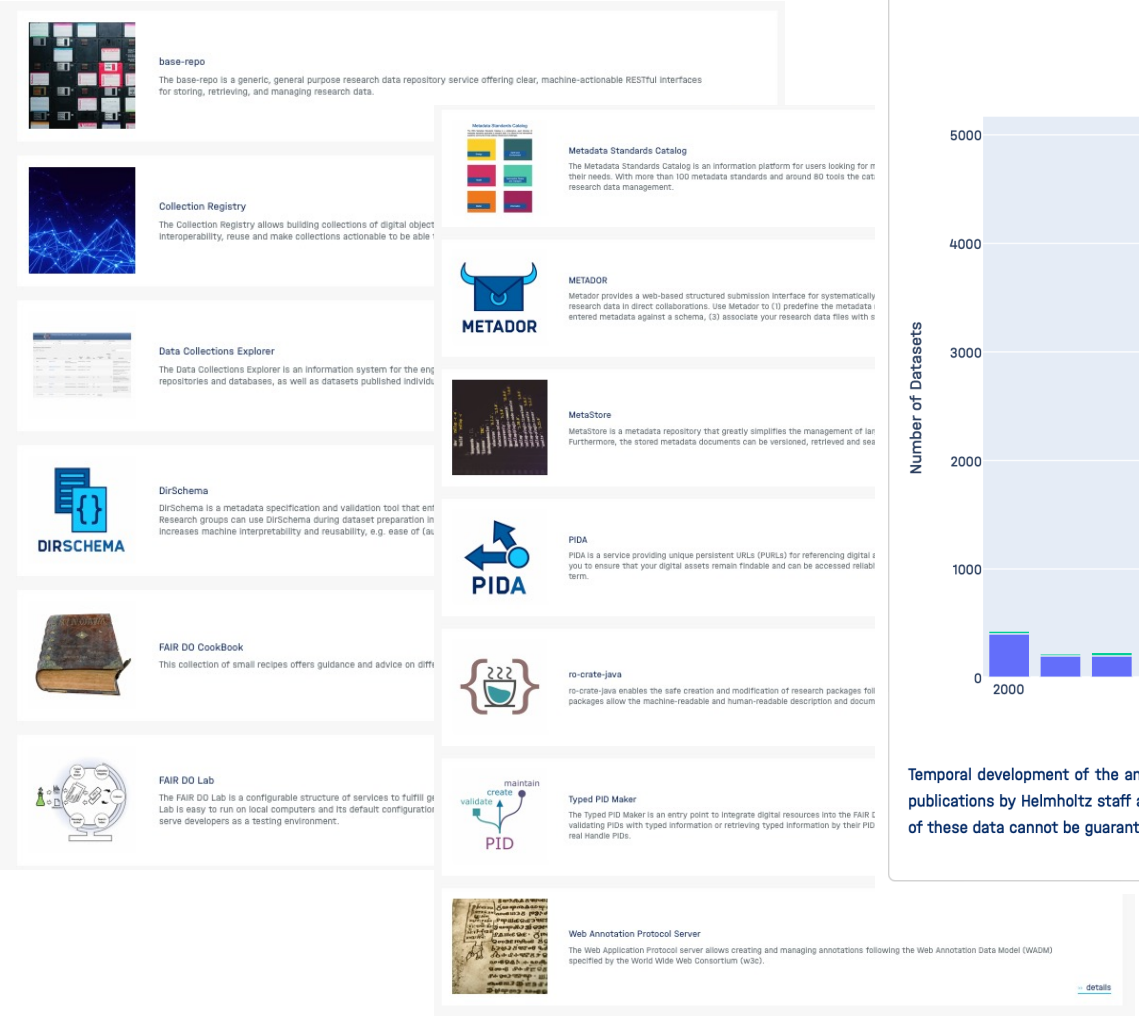
Helmholtz Association: largest German science organisation

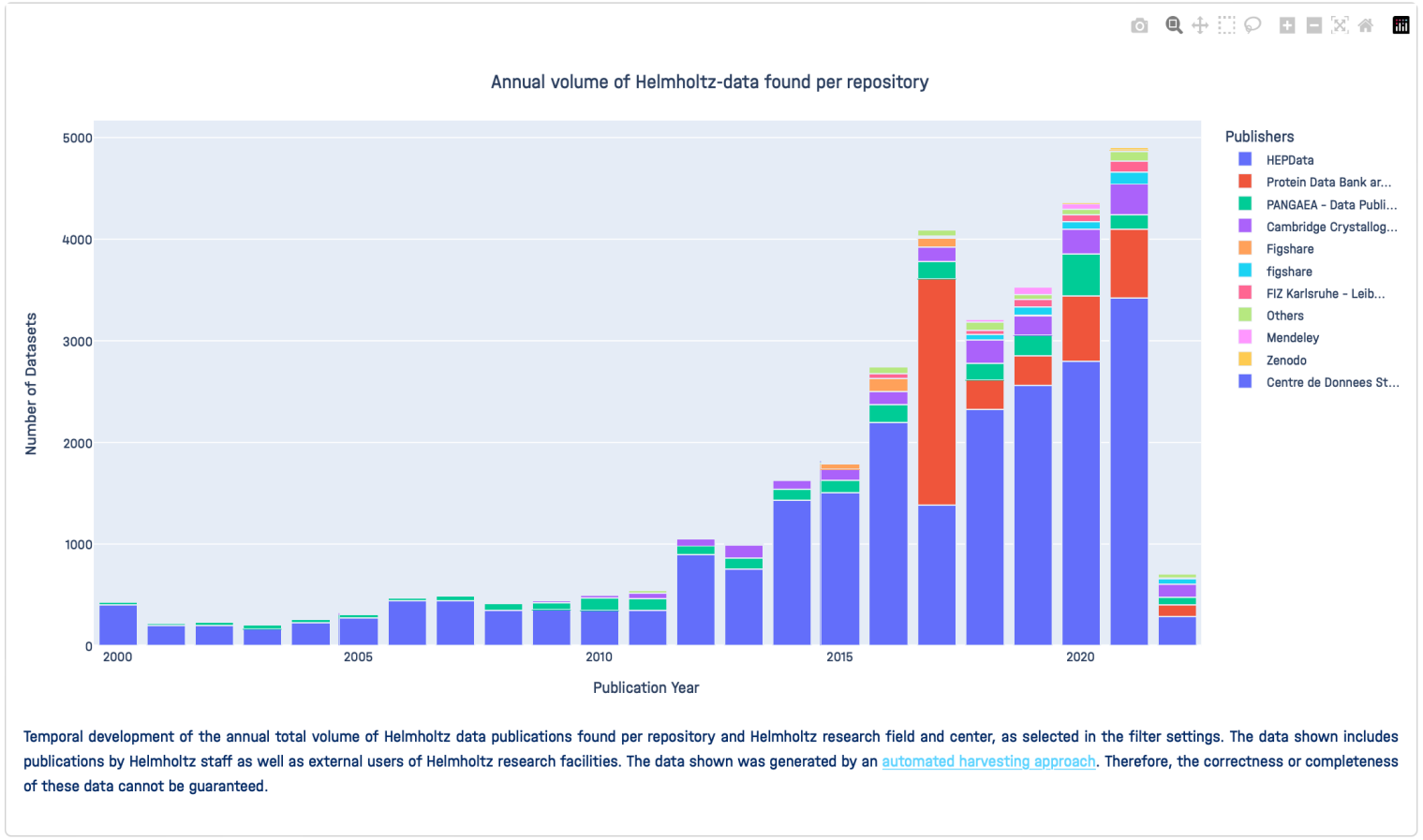
6 research fields, 19 centres, > 40.000 staff, ~ 4.5 billion Euro

Helmholtz Metadata Collaboration

HELMHOLTZ METADATA COLLABORATION

https://helmholtz-metadaten.de/en





Temporal development of the annual total volume of Helmholtz data publications found per repository and Helmholtz research field and center, as selected in the filter settings. The data shown includes publications by Helmholtz staff as well as external users of Helmholtz research facilities. The data shown was generated by an automated harvesting approach. Therefore, the correctness or completeness of these data cannot be guaranteed.

General caveat:

High # of data sets: Harvesting of HEPData

→ usefulness of FAIR indicators / KPIs?

→ What is a useful metric for data sets & software?

HMC Survey

HELMHOLTZ Metadata Collaboration

https://oceanrep.geomar.de/id/eprint/57494/

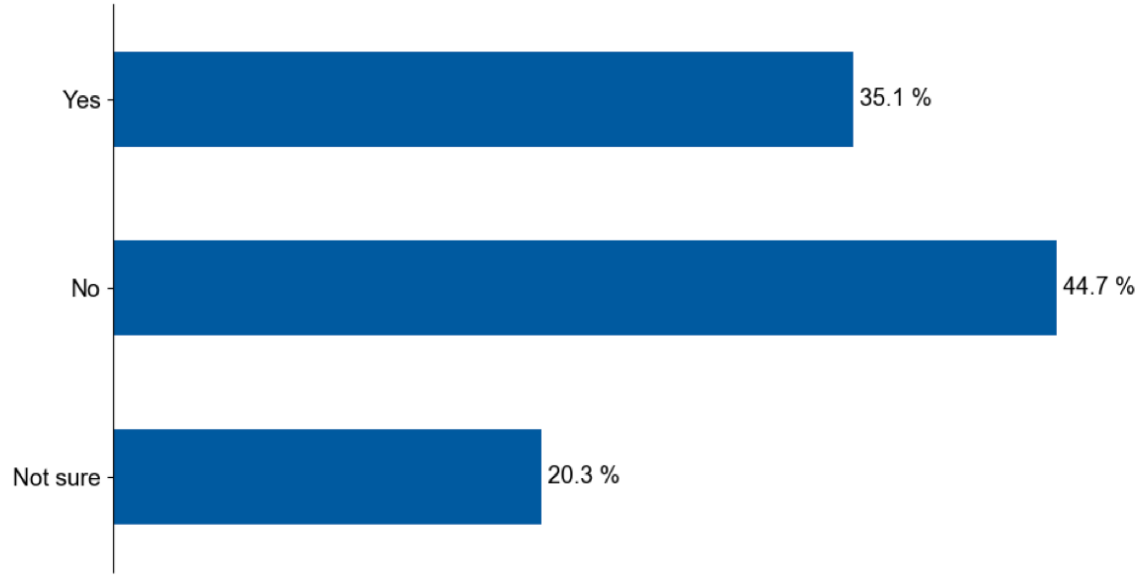


Figure 14: "Do you document your research data in a structured way? (e.g., using forms, templates or schemas)". (Single choice question, available to all respondents, number of respondents who answered this question: n = 582, relative amounts refer to n)

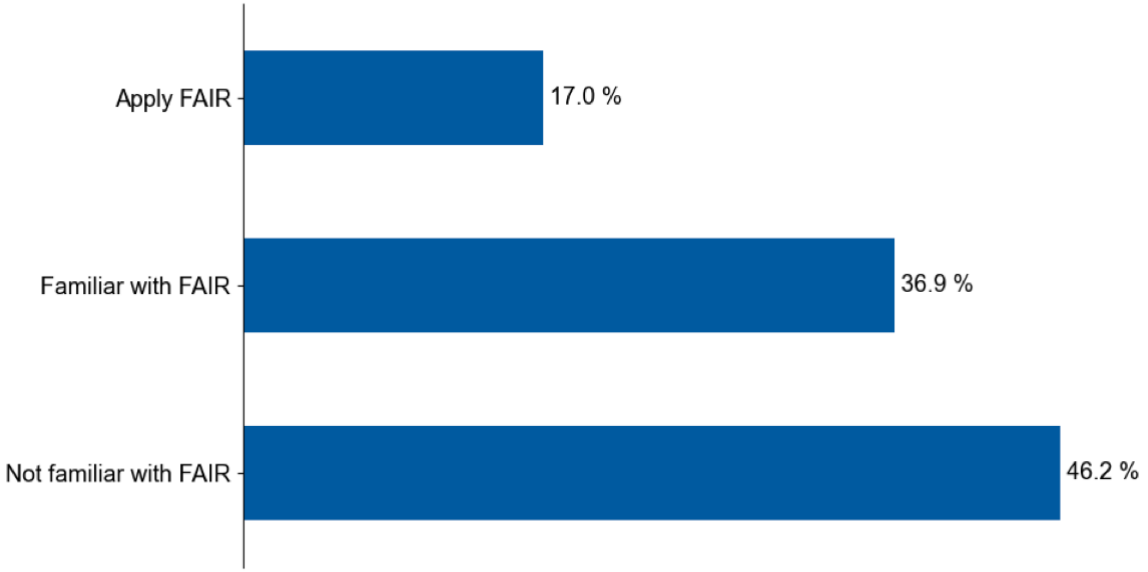
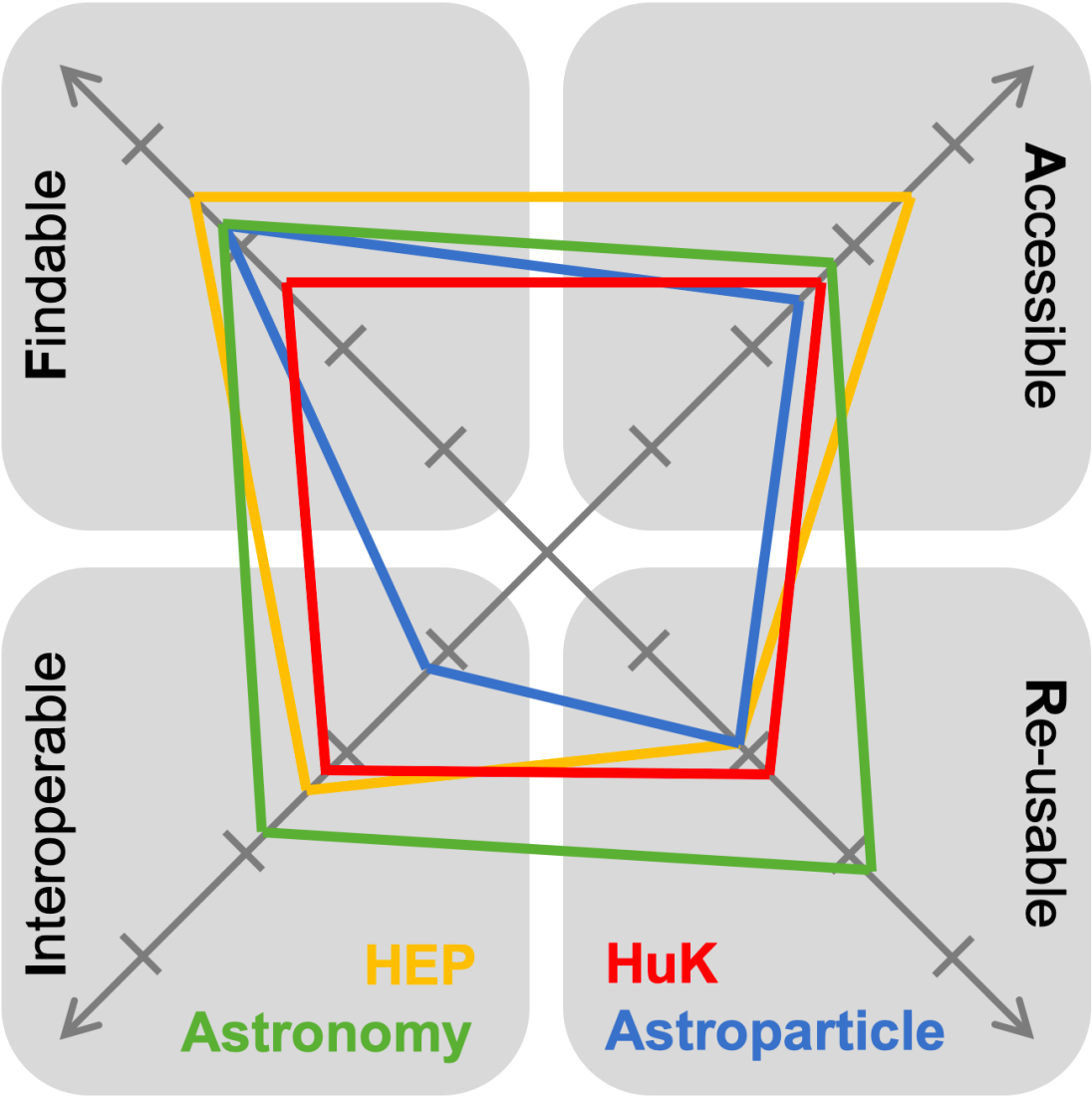


Figure 8: "How familiar are you with the FAIR data guidelines?". Question not available to respondents who previously assigned themselves to the Helmholtz research field Information. (Single-choice question, number of respondents who answered this question: n = 559, relative amounts refer to n)

FAIR Assessment

Making Things Operational: PUNCH4NFDI Proposal (2020)



A lot of HEP data are in HEPData (open, F&A), and WITHIN collaborations (e.g. at LHC) data are FAIR.

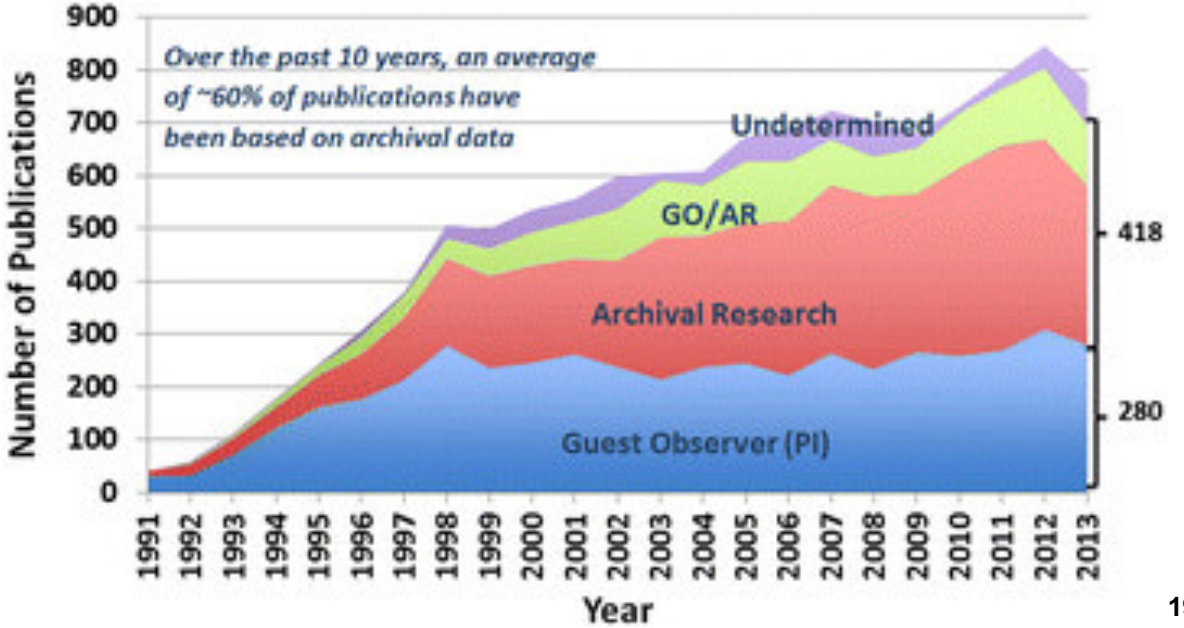
No (?) solutions for medium and small experiments.

Limited I&R.

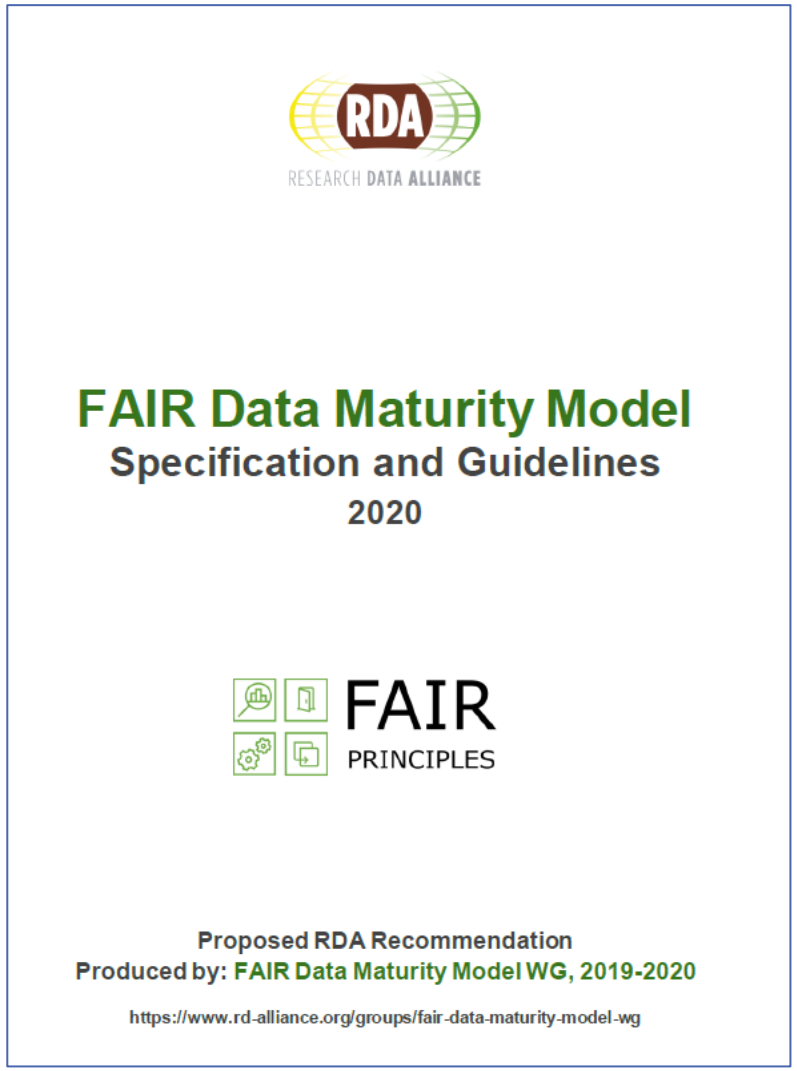
50% of all astronomy publications are based on the re-use of "old" data, and astro data are often open.

There are no attached compute resources.

DOI:10.1186/s40192-014-0022-8



Making Things Operational: FAIR Indicators



https://zenodo.org/record/3909563

From 15 FAIR Guidelines (2016) to 41 FAIR Indicators (2020)

Metric:			Score
	0 – not applicable	0	0
	1 – not being considered this yet	1	0
	2 – under consideration or in planning phase	2	0
	3 – in implementation phase	3	0
	4 – fully implemented	4	1

RDA-F1-01M Metadata identified by a persistent identifier	●●● Essential
Principle to which the indicator relates	This indicator is linked to the following principle: *F1 (meta)data are assigned a globally unique and eternally persistent identifier*.
Description of the indicator RDA1-F1-01M	This indicator evaluates whether or not the metadata is identified by a persistent identifier. A persistent identifier ensures that the metadata will remain findable over time and reduces the risk of broken links.
Assessment details	The persistence of an identifier is determined by the commitment of the organisation that assigns and manages the identifier, so the evaluation of this indicator needs to take into account the persistence policy of that organisation. Such a commitment could be expressed by a university or research institute, by a research infrastructure or by an organisation that issues formal identifiers, such as the International DOI Foundation. A possible way to evaluate this indicator is to verify that the identifier used for the metadata is listed in a registry service like the RDA-endorsed FAIRsharing.[15]

There are ways to measure FAIRness. Details are extremely controversial (scoring systems).

Example: EMIL at BESSY (PaN Science)

HELMHOLTZ Metadata Collaboration

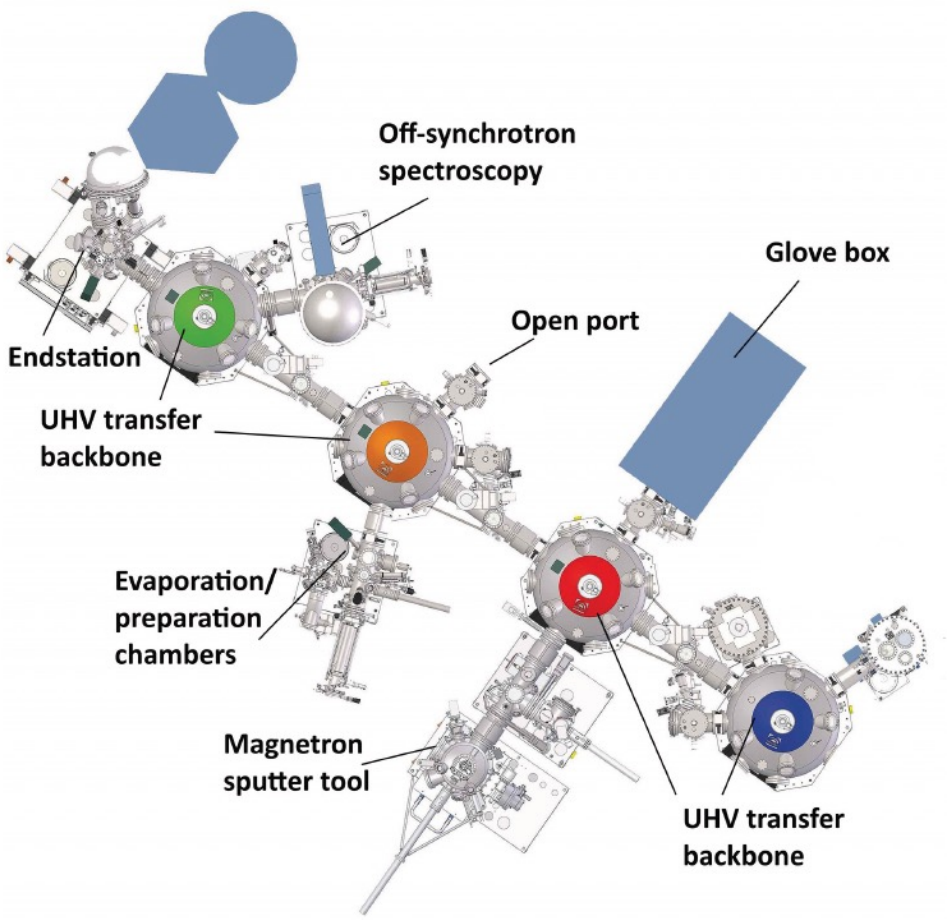


Figure 1: Schematic presentation of the SISSY@EMIL's setup connecting various processing and analysis chambers, among them the SISSY I endstation.

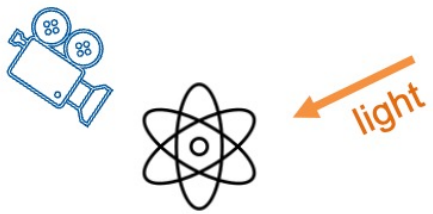
Particularly nice example:

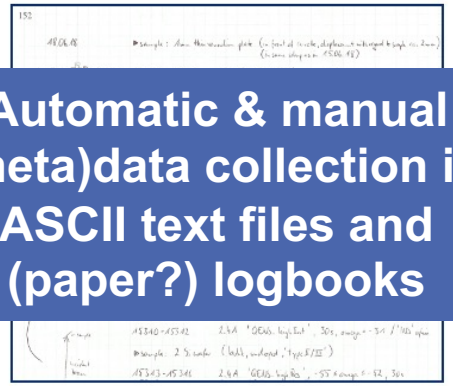
- Integrating entire data collection and processes into FAIR assessment
- Breaking FAIR down to concrete things that can be implemented at the lab level

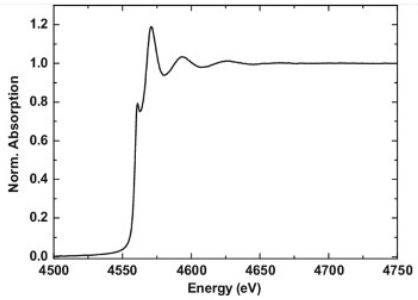
ICALEPS2021, doi:10.18429/JACoW-ICALEPS2021-WEBL05

Example from HMC: EMIL at BESSY (PaN Science)

Data collection



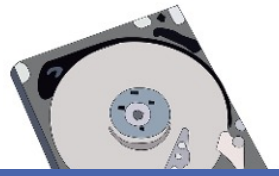




Automatic & manual (meta)data collection in ASCII text files and (paper?) logbooks

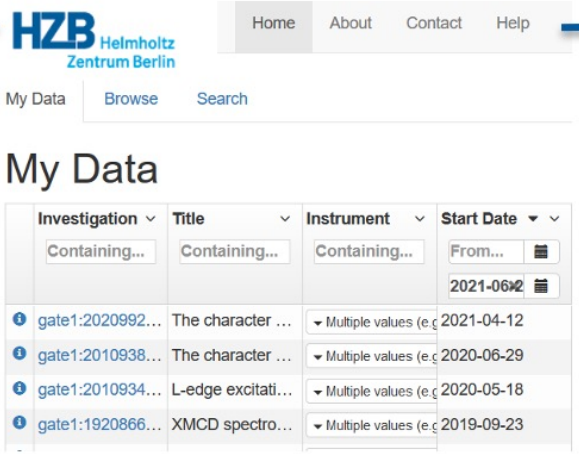
Data processing

(group's share drive)



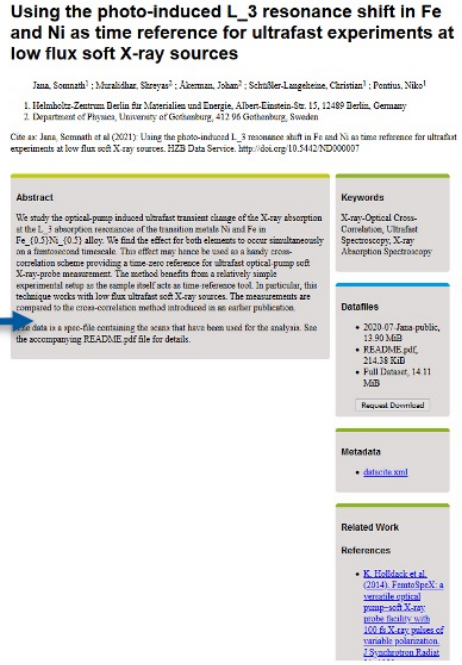
HDF5 / NeXus file format: hierachical (meta)data, NeXus definition language, persistent web description (URL) → machine readable / interoperable

Data repository: ICAT



Data publication

(manual PIDs)



http://doi.org/10.5442/ND000007 (example)

Example: M. Kubin & G. Günther (HMC)

HELMHOLTZ Metadata Collaboration

FAIR Assessment

Findability:

- Discovery metadata
- Automatic assignment of PIDs
- Connection to higher-level services

Accessibility:

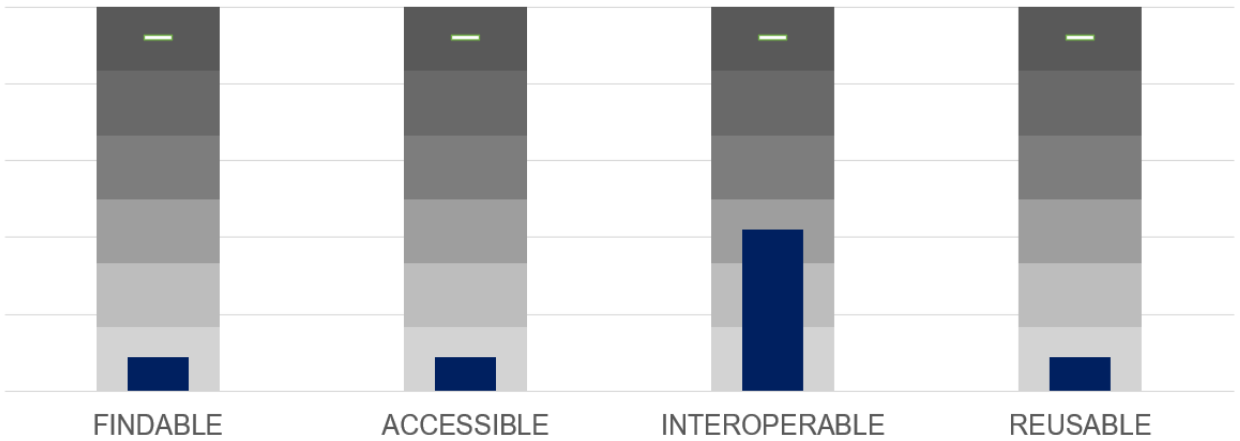
- (meta)data in ICAT repository
- Authentication / authorization

Interoperability:

- NeXus format
- ICAT schema (mappings w.i.p.)

Reusability:

- Sample + calibration (meta)data
- License metadata



Level 0	Not FAIR
Level 1	FAIR essential criteria only
Level 2	FAIR essential criteria + 50 % of important criteria
Level 3	FAIR essential criteria + 100% of important criteria
Level 4	FAIR essential criteria + 100% of important criteria + 50% of useful criteria
Level 5	FAIR essential criteria + 100% of important criteria + 100% of useful criteria

Numerous approaches; importance of process evaluation. E.g. …

F-UJI Automated FAIR Data Assessment Tool

ICALEPS2021, doi:10.18429/JACoW-ICALEPS2021-WEBL05

Some Usecases

The EAJADE Project

Europe-America-Japan Accelerator Development and Exchange Programme
A Marie Sklodowska-Curie Research and Innovation Staff Exchange (SE) action, funded by the EU under Horizon-Europe.

EAJADE
Europe-America-Japan Accelerator Development Exchange Programme

Work package title
R&D&I at currently operating state-of-the-art facilities
State-of-the-art high-gradient, high-efficiency, reduced-cost radio-frequency structures and power sources
Special technologies, devices and systems performance
Sustainable technologies for scientific facilities
Investigation of potential early applications of novel and advanced technologies for colliders
Management, dissemination, training, knowledge transfer, and communication

The EAJADE Project

EAJADE
Europe-America-Japan Accelerator Development Exchange Programme

Europe-America-Japan Accelerator Development and Exchange Programme
A Marie Sklodowska-Curie Research and Innovation Staff Exchange (SE) action,
funded by the EU under Horizon-Europe.

Questions:

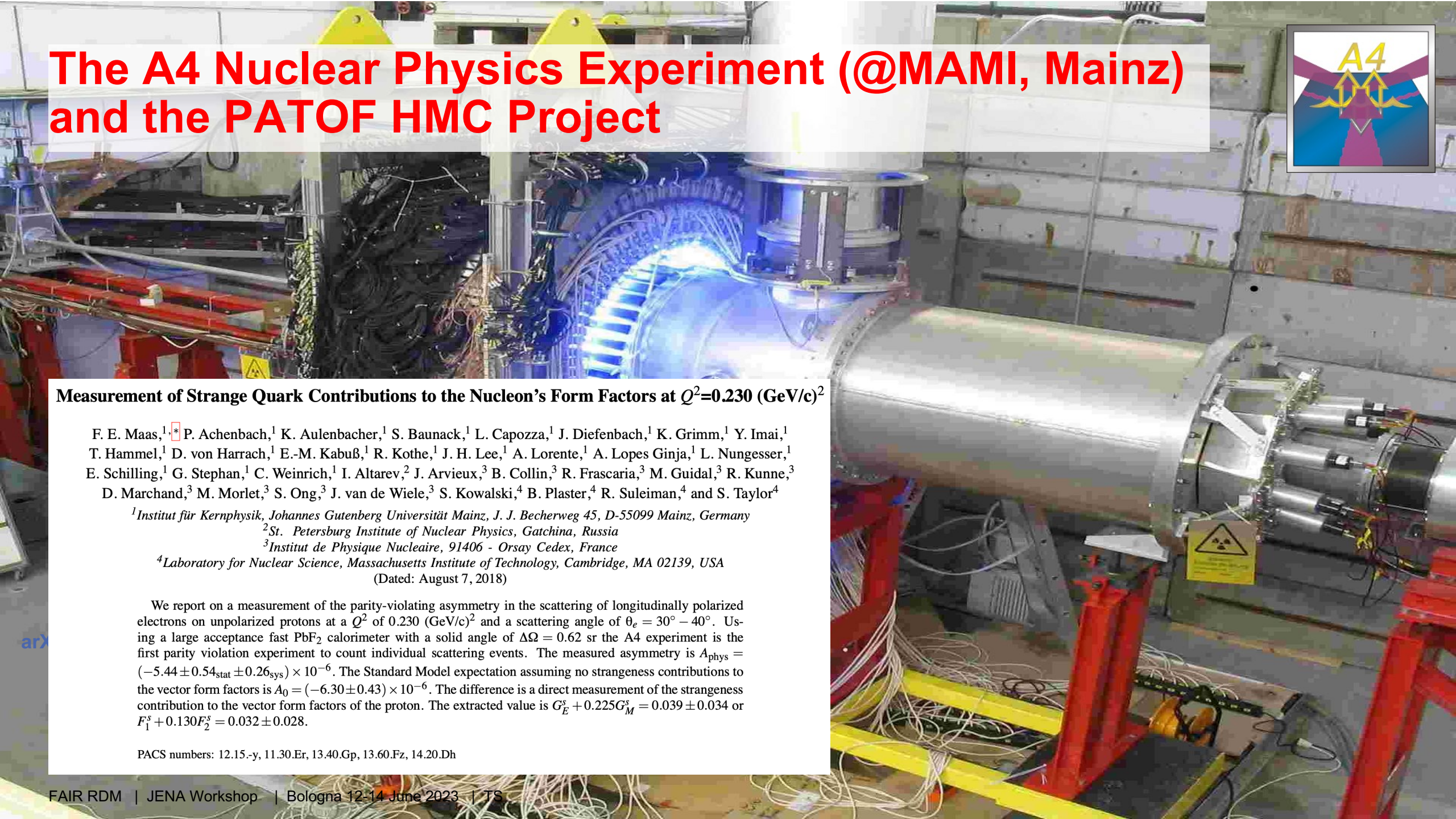
- Where and how do we store what?
- How long do we store? Just the good-practice 10 years? Persistently? What IDs do we give?
- How do we organise access to the data?
- How to create (which) metadata schema and uniform metadata for the project data?
- Who does it? Do we really have to do it? Is it worth it? …

Findings:

- There is no clear picture of the data and their future treatment.
- There is no well-specified storage location at hand.
- There is no idea about a metadata schema for this research.
- There is no consensus on the necessity.
- Many technicalities (embargo, AAI, licenses, ..) are absolutely unsolved.
- …

~~FAIR~~ **… but working on a (preliminary) solution.**

The A4 Nuclear Physics Experiment (@MAMI, Mainz) and the PATOF HMC Project



Measurement of Strange Quark Contributions to the Nucleon's Form Factors at Q^2=0.230 (GeV/c)^2

F. E. Maas,[1,*] P. Achenbach,[1] K. Aulenbacher,[1] S. Baunack,[1] L. Capozza,[1] J. Diefenbach,[1] K. Grimm,[1] Y. Imai,[1] T. Hammel,[1] D. von Harrach,[1] E.-M. Kabuß,[1] R. Kothe,[1] J. H. Lee,[1] A. Lorente,[1] A. Lopes Ginja,[1] L. Nungesser,[1] E. Schilling,[1] G. Stephan,[1] C. Weinrich,[1] I. Altarev,[2] J. Arvieux,[3] B. Collin,[3] R. Frascaria,[3] M. Guidal,[3] R. Kunne,[3] D. Marchand,[3] M. Morlet,[3] S. Ong,[3] J. van de Wiele,[3] S. Kowalski,[4] B. Plaster,[4] R. Suleiman,[4] and S. Taylor[4]

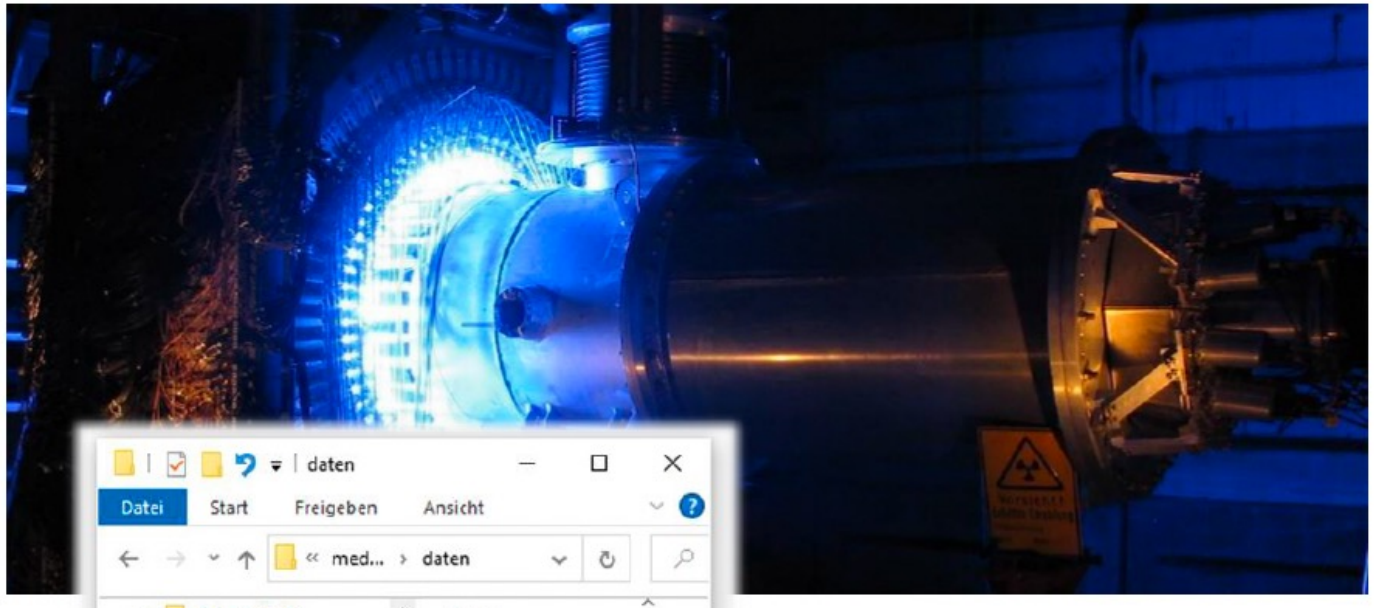
[1]*Institut für Kernphysik, Johannes Gutenberg Universität Mainz, J. J. Becherweg 45, D-55099 Mainz, Germany*
[2]*St. Petersburg Institute of Nuclear Physics, Gatchina, Russia*
[3]*Institut de Physique Nucleaire, 91406 - Orsay Cedex, France*
[4]*Laboratory for Nuclear Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

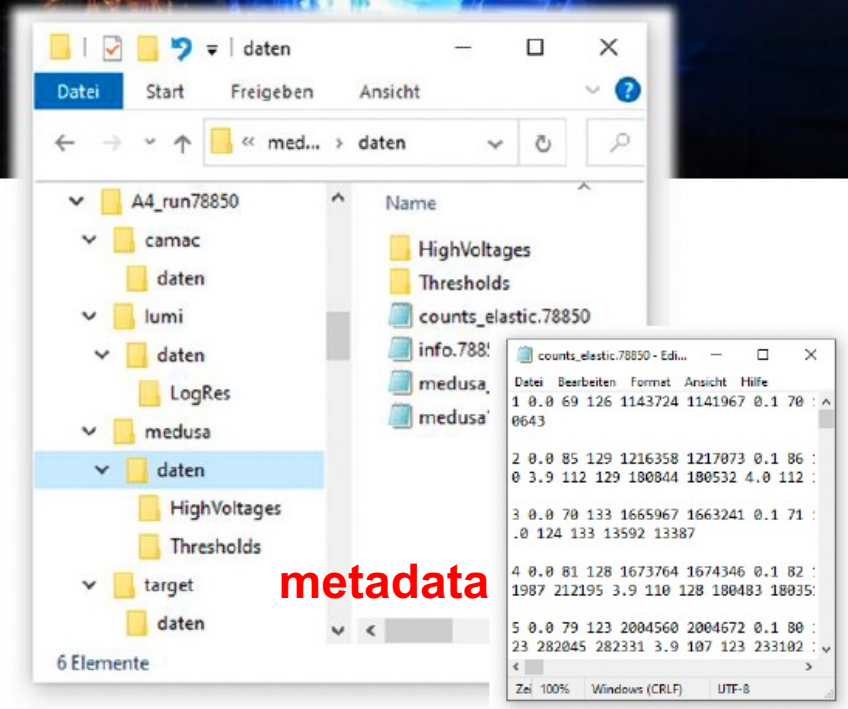
(Dated: August 7, 2018)

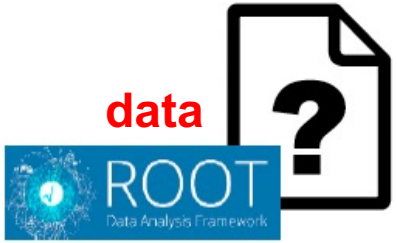
We report on a measurement of the parity-violating asymmetry in the scattering of longitudinally polarized electrons on unpolarized protons at a Q^2 of 0.230 (GeV/c)^2 and a scattering angle of $\theta_e = 30° - 40°$. Using a large acceptance fast PbF_2 calorimeter with a solid angle of $\Delta\Omega = 0.62$ sr the A4 experiment is the first parity violation experiment to count individual scattering events. The measured asymmetry is $A_{phys} = (-5.44 \pm 0.54_{stat} \pm 0.26_{sys}) \times 10^{-6}$. The Standard Model expectation assuming no strangeness contributions to the vector form factors is $A_0 = (-6.30 \pm 0.43) \times 10^{-6}$. The difference is a direct measurement of the strangeness contribution to the vector form factors of the proton. The extracted value is $G_E^s + 0.225 G_M^s = 0.039 \pm 0.034$ or $F_1^s + 0.130 F_2^s = 0.032 \pm 0.028$.

PACS numbers: 12.15.-y, 11.30.Er, 13.40.Gp, 13.60.Fz, 14.20.Dh

The A4 Nuclear Physics Experiment (@MAMI, Mainz)







metadata

A4 Experiment:

- Instrument @ Mainz Microtron
- Already dismounted
- Published papers but potential for more

Prototypical HEP Files:

- Nested folder structure (unknown context)
- Root Files: community-specific standard
- ASCII Files: Unformatted, minimum metadata
- Electronic Lab Notebook: context (xml)

Practical Goal:

- Reusable data from a dismounted experiment
- Template for future instruments

G. Günther

The A4 Nuclear Physics Experiment (@MAMI, Mainz)

Assessment

- ROOT files with customized classes: self-describing, binary, requires full software suite
  ➔ requires installation of customised classes; version control requires containerisation
  ➔ solution soon at hand

☺ RDA-I1-01D: Data uses representation in standardised format

☺ RDA-R1.3-01D: Data complies with a community standard

☹ RDA-I1-02D: Data uses machine-understandable knowledge representation

☹ RDA-R1.3-01D: Data is in machine-understandable community standard

- ASCII files: neither human nor machine readable
  ➔ require conversion to human / machine-readable XML format, adding metadata through json files (column headers, units, ..)
  RDF annotations provide machine-readable definitions e.g. using ontology QUDT or human-readable context (e.g. using weblinks)

☹ RDA-F2-01M: Rich metadata is provided to allow discovery

☹ RDA-I1-02M: Metadata uses machine-understandable knowledge representation

☹ RDA-R1-01M: Plurality of accurate and relevant attributes allow reuse

☹ RDA-R1.3-02M: Metadata is in machine-understandable community standard

- Electronic lab notebook and readme pages: May contain private / confidential entries
  ➔ make black/whitelist; transfer to XML/DataCite or html)

☺ RDA-I3-01M: Metadata includes references to other metadata

☺ RDA-I3-02M: Metadata includes references to other data

The A4 Nuclear Physics Experiment (@MAMI, Mainz)

FDMM ID	Indicator	Priority	Comment
RDA-I1-01M	Metadata uses representation in standardized format	Important	☺ (DataCite)
RDA-I1-01D	Data uses representation in standardised format	Important	☺ (root/class)
RDA-I1-02M	Metadata uses machine-understandable knowledge representation	Important	☺ (xml/DataCite)
RDA-I1-02D	Data uses **machine-understandable knowledge representation**	Important	☹ (root/class)
RDA-I2-01M	Metadata uses FAIR-compliant vocabularies	Important	☺ (DataCite)
RDA-I2-01D	Data uses **FAIR-compliant vocabularies**	Useful	☹ (root/class)
RDA-I3-01M	Metadata includes references to other metadata	Important	☺ (e.g. ORCID)
RDA-I3-01D	Data includes references to other data	Useful	☹ (root)
RDA-I3-02M	Metadata includes references to other data	Useful	☺ (via ELN)
RDA-I3-02D	Data includes **qualified references** to other data	Useful	☹ (root)
RDA-I3-03M	Metadata includes qualified references to other metadata	Important	☺ (e.g. ORCID)
RDA-I3-04M	Metadata include qualified references to other data	Useful	☺ (ELN/readme)

The A4 Nuclear Physics Experiment (@MAMI, Mainz)

FDMM ID	Indicator	Priority	Comment
RI			
RI			
RI			e)
RI			
RI			
RI			
RI			)
RI			
RI			
RI			
RI			)
RDA-I3-04M	Metadata include qualified references to other data	Useful	© (ELN/readme)

A4 lessons:

- Most measures could be taken at instrument level (i.e. writing output files in machine-readable / -understandable format).
- Some measures becoming standard – i.e. writing XML instead of ASCII etc.
- F&A on rather good level compared to I&R – repositories are quite advanced. I&R essentially requires engagement of community.

PATOF Helmholtz project to finish FAIRification of A4 and carry expertise to other nuclear physics experiments and axion / DM search experiments at DESY & Mainz (ALPS II, LUXE, P2, …)

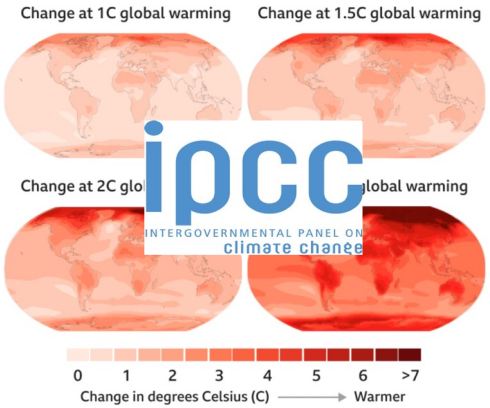
Goal: establishing of a FAIR metadata factory and a cookbook with numerous recipees for the involved communities.

Irreversibility Challenge

FAIR???

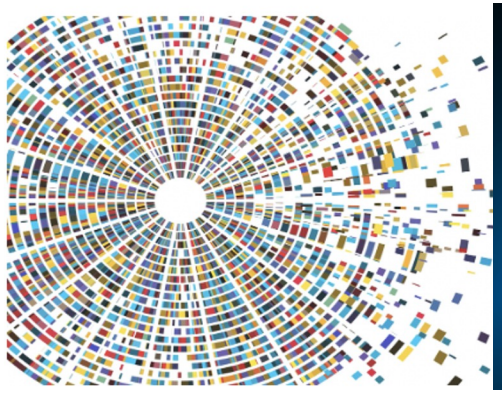
Increasing data volumes require more and other metadata

Advancement of knowledge and solutions to societal challenges depend on ICT resources & software. Data increase requires new approaches to scientific computing and data management, changing today's paradigms. This requires long-term R&D plan with coordinated efforts from different communities & large R&D investments.



Climate modelling: CMIP6 with 22 PB and 7M data sets

SKAO (> 2027): 600PB/a (archived); **MeerKAT** (now): 3 PB/d (produced + analysed + cancelled)



Genomics / bio-medicine: complex long time series

Light sources: 35k users. TB/s per facility. Factor 10^4-10^6 data rate increase in this decade

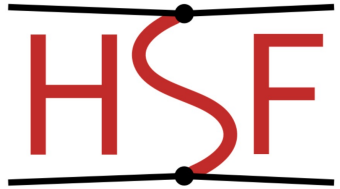
The "smart city" produces massive amounts of data

Requires in-flight analytics, data reduction / compression / loss / removal, federated & heterogeneous resources, green IT, … To tackle **irreversibility challenge and democratise analysis**: Massive investment in metadata annotation and FAIR & open data (1/3 data, 1/3 metadata, 1/3 simulation) → role of EU funding!

(Infra)Structures
Culture
Boundary Conditions

Make it possible, easy (normative,rewarding)

Technical and conceptual solutions, individual services, single experiences …



HEP Software Foundation

The HEP Software Foundation facilitates cooperation and common efforts in High Energy Physics software and computing internationally.

EUROPEAN OPEN SCIENCE CLOUD

EAJADE

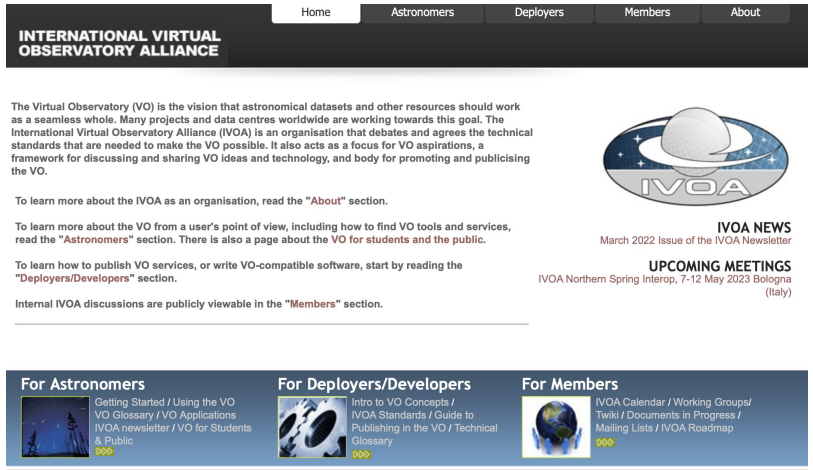
Europe-America-Japan Accelerator Development Exchange Programme

Nationale Forschungsdaten Infrastruktur

HELMHOLTZ Metadata Collaboration

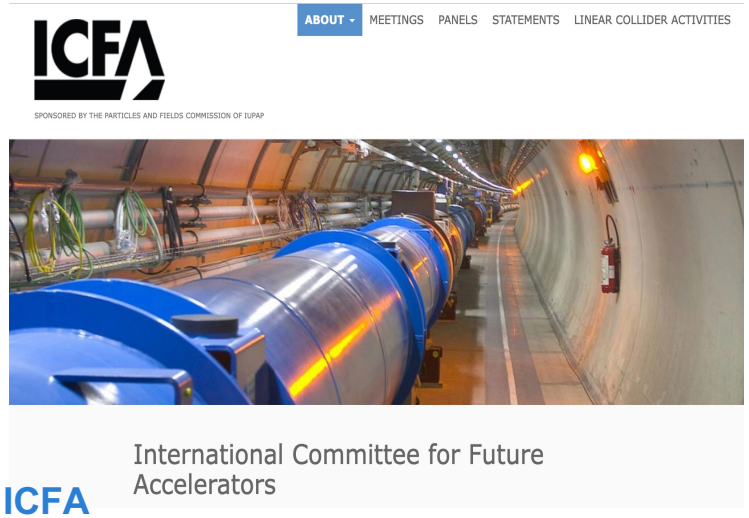
Make it possible, easy (normative, rewarding)

Labs, collaborations, communities, …



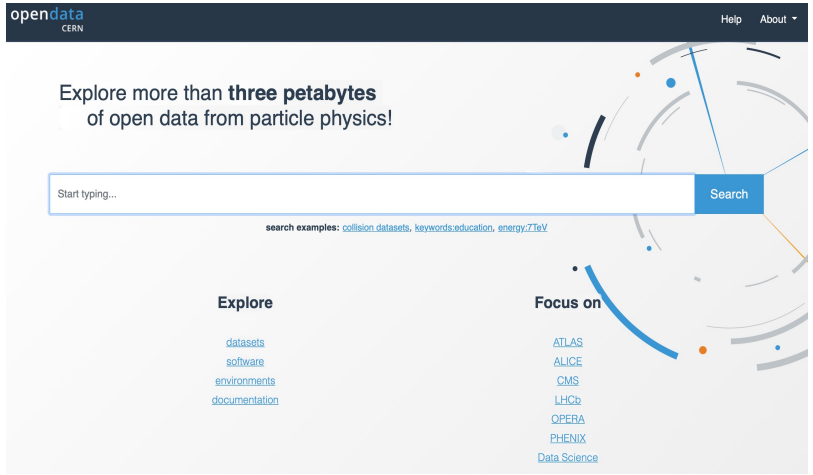
Virtual Observatory

European structures and services



ICFA

CERN Open Data



German ErUM-Data (erumdatahub.de)

ERUM DATA HUB

GEFÖRDERT VOM

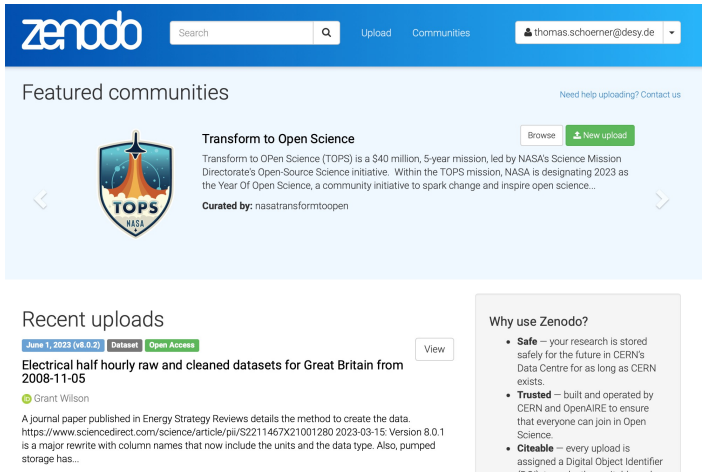
Bundesministerium für Bildung und Forschung

We are a Digital Knowledge Agent

The **ErUM-Data_Hub** , funded by the **Federal Ministry of Education and Research (BMBF),** is a central networking and transfer office for the digital transformation in the exploration of universe and matter (German abbreviation: ErUM). The main aims of the ErUm-Data-Hub are: The networking of ErUM-communities, identification and exploitation of synergies in ErUm-Data-projects in the field of digitalization, the communication of research results geared to the target groups as well as knowledge and innovation transfer. Furthermore, the ErUM-Data-Hub provides a diversified further education offer in the field of digital competence.

As part of the **ErUM-Data Plan of Action,** the ErUM Data Hub pursues the overarching goal of digital transformation in fundamental ErUM-Pro research. Read more about who we are and what we do.

ZENODO



Make it normative (possible, easy): Policies

All Helmholtz centers should have a data policy in place

According to the position paper on the handling of research data in the Helmholtz Association (Mitgliederversammlung der Helmholtz-Gemeinschaft, 2016), all member centers are expected to have a policy in place by the end of 2017. *"All member Centres need to have established guidelines **by the end of 2017**. Formulation of the discipline-specific details is expected to take some years."*

Helmholtz Open Science Policy (2022) also emphases having publicly available policies for all the Helmholtz Centers: "***All Centers will establish detailed procedures for managing research data in publicly available policies**, and will regularly examine and if necessary adapt these procedures."*

Digital research data generated should be managed in accordance with the FAIR principles

According to the new Helmholtz Open Science Policy (2022), the employees shall ensure that the digital research data that they generate shall be managed responsibly and in accordance with the **FAIR Principles**.

Retention of research data should be guaranteed for 10 years

DFG Guidelines for Safeguarding Good Research Practice (2022) says that *"When scientific and academic findings are made publicly available, the research data (generally raw data) on which they are based are generally archived in an accessible and identifiable manner for a period of **ten years** at the institution where the data were produced or in cross-location repositories"*.

- FAs and institutions must issue RDM policies – creating pressure in two directions: users must comply, institutions must make it possible (funding).
- Example: Helmholtz now preparing for including FAIRness of produced data sets in its KPIs for the next funding period.

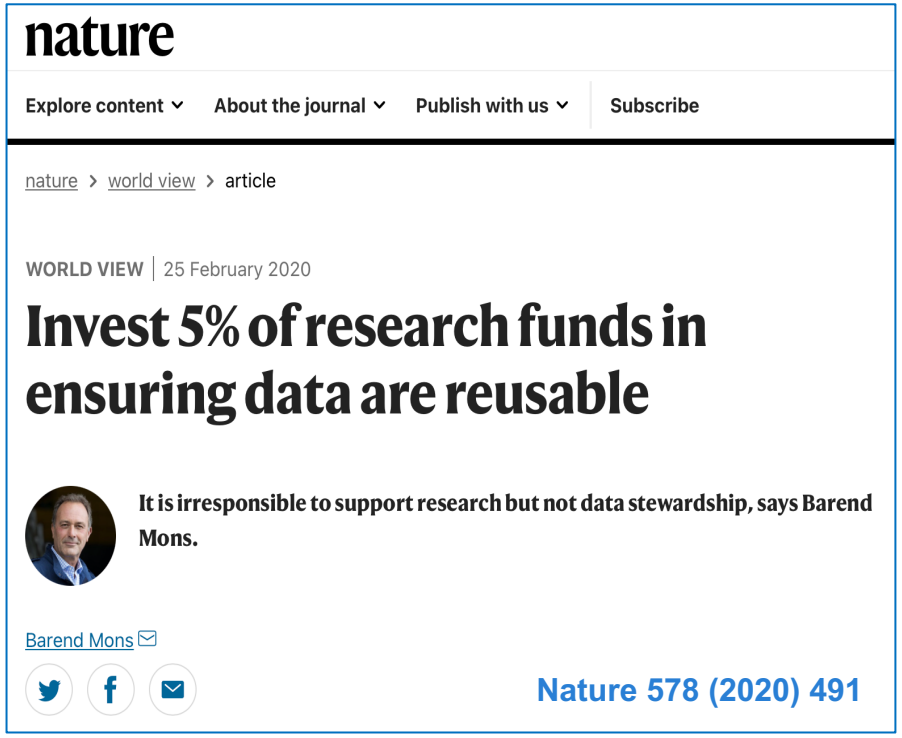
Helmholtz Open Science policy
https://gfzpublic.gfz-potsdam.de/rest/items/item_5013535_1/component/file_5013536/content

Funding Agency Level

Policies, requirements, funding, …

EU, German Research Association DFG (and certainly others) (need to)issue policies, and they require us to provide FAIR, sustainable and open data management.

But (today mostly) no dedicated funding for data management, data curation in projects.



nature

Explore content | About the journal | Publish with us | Subscribe

nature > world view > article

WORLD VIEW | 25 February 2020

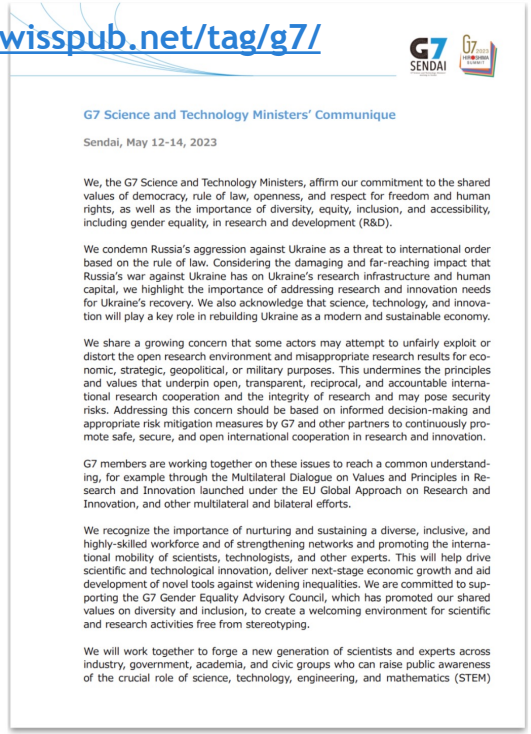
Invest 5% of research funds in ensuring data are reusable

It is irresponsible to support research but not data stewardship, says Barend Mons.

Barend Mons

Nature 578 (2020) 491

https://wisspub.net/tag/g7/



G7 Science and Technology Ministers' Communique

Sendai, May 12-14, 2023

We, the G7 Science and Technology Ministers, affirm our commitment to the shared values of democracy, rule of law, openness, and respect for freedom and human rights, as well as the importance of diversity, equity, inclusion, and accessibility, including gender equality, in research and development (R&D).

We condemn Russia's aggression against Ukraine as a threat to international order based on the rule of law. Considering the damaging and far-reaching impact that Russia's war against Ukraine has on Ukraine's research infrastructure and human capital, we highlight the importance of addressing research and innovation needs for Ukraine's recovery. We also acknowledge that science, technology, and innovation will play a key role in rebuilding Ukraine as a modern and sustainable economy.

We share a growing concern that some actors may attempt to unfairly exploit or distort the open research environment and misappropriate research results for economic, strategic, geopolitical, or military purposes. This undermines the principles and values that underpin open, transparent, reciprocal, and accountable international research cooperation and the integrity of research and may pose security risks. Addressing this concern should be based on informed decision-making and appropriate risk mitigation measures by G7 and other partners to continuously promote safe, secure, and open international cooperation in research and innovation.

G7 members are working together on these issues to reach a common understanding, for example through the Multilateral Dialogue on Values and Principles in Research and Innovation launched under the EU Global Approach on Research and Innovation, and other multilateral and bilateral efforts.

We recognize the importance of nurturing and sustaining a diverse, inclusive, and highly-skilled workforce and of strengthening networks and promoting the international mobility of scientists, technologists, and other experts. This will help drive scientific and technological innovation, deliver next-stage economic growth and aid development of novel tools against widening inequalities. We are committed to supporting the G7 Gender Equality Advisory Council, which has promoted our shared values on diversity and inclusion, to create a welcoming environment for scientific and research activities free from stereotyping.

We will work together to forge a new generation of scientists and experts across industry, government, academia, and civic groups who can raise public awareness of the crucial role of science, technology, engineering, and mathematics (STEM)

Bundesministerium für Bildung und Forschung

KULTUSMINISTER KONFERENZ

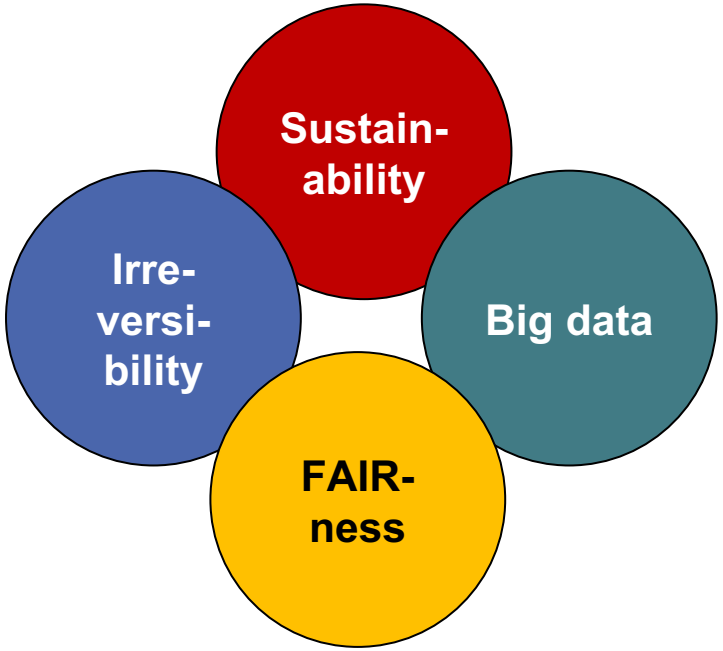
Open Access in Deutschland

Gemeinsame Leitlinien von Bund und Ländern

https://www.bmbf.de/SharedDocs/Publikationen/de/bmbf/1/772960_Open_Access_in_Deutschland

Conclusions

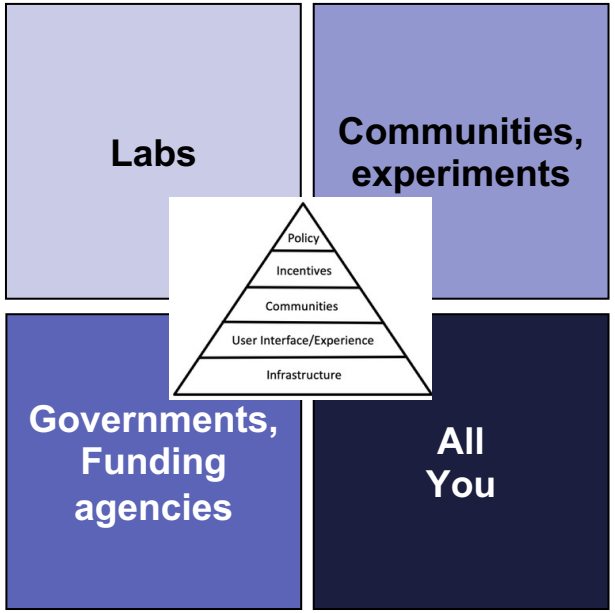
The Large Picture - Requirements



- Challenges require massive metadata harvesting – with tools that are not yet there.
- ML → need to harvest as many as possible machine-actionable metadata!
- Necessity of long-term investment plan!

- data policy and FAIR definition
- data steward
- Regular FAIR assessment
- Training, education, awareness

- Policies
- Funding
- Check implementation
- Liaisons



- Funding applications
- Standards and ter-minologies
- Recognition
- Communication

- Global AAI
- Change of mind
- Think metadata (1/3,1/3,1/3!)

Essential Tools for FAIR PUNCH Data

Federated infra-structures

Data

Digital research products

RP

Workflows Algorithms

Platform Protocols

Portal

AAI

FAIR

Software

Metadata (schemas)

1/3 data
1/3 metadata
1/3 simulation

Money

Policy

Data stewards

Assessment

Training
Education
Communication

Awareness

On Open Data

I Can't Publish My Data! Lame Excuses and Some Answers

Markus Demleitner
Universität Heidelberg, Astronomisches Rechen-Institut, Mönchhofstraße 12-14, Germany

Abstract

You already know why publishing data is the right thing to do, don't you? It's just that exactly your data is an exception, right? It cannot be published because...

Here's a collection of reasons we've heard. And some replies we've always wanted to give (but mostly swallowed).

Acknowledgement

The good parts of this were blatantly snarfed from Charly Strasser's brilliant web page `http://datapub.cdlib.org/closed-data-excuses-excuses/`

People will misinterpret my data

Good documentation and standards mitigate this.

As for what remains – well, if you're publishing prose (i.e., in a journal), how many of your readers do you think actually get what you're writing?

I might yet want to use it

...in the great seminal research paper I've always wanted to write.

If you've not done so so far, will you at all? When? Too much data obtained for uncounted kilodollars (or megadollars, for that matter) is gathering dust, waiting for the "real soon now". Be fair to the world and to the people funding you and your research and publish the data. If you're really worried, put a one-year embargo on the material.

Procraste's Law: If over a year you don't get to do it, chances are overwhelming you'll never do it.

My data is too complicated

If it's too complicated to explain: are you sure you've understood it yourself?

Be that as it may: Try explaining anyway, the improved understanding you'll get will reward you plenty.

I'm not sure I own the data

That's amazingly common. So: Are you sure you cannot find out who does? If you made a reasonable effort to figure that out but failed, the likelihood is high you've orphaned data on your hands, obtained by people who've long left science for greener pastures.

To avoid similar uncertainties with your data later on, please consider assigning explicit licenses to it – ideally CC-0. Do not worry that people will not give credit just because of a Free license. We're in science, and so this is a matter of scientific conduct rather than the law.

I'm busy

...and it's not a priority.

A-ha! Here we're talking. True, the current system of rewards in science doesn't actually encourage data publishing. But publishing is the right thing to do, anyway, even before the system gets back on a path of recovery. And: more and more funding agencies at least sound requirements for data publishing and preservation.

My data is boring

...or at least not very interesting.

Leave that decision to others. You'd be surprised how much "boring data" people point-and-click out of printed graphs or tables in the sweat of their mice. Every day.

People will contact me

...and ask about stuff.

Well, science is about exchange. Think how much *you* learned by asking other people.

Plus, you'll notice that quite a few of those questions are actually quite clever, so answering them is a good use of your time.

As to the stupid questions – well, they *are* annoying, but at least for us even those were eye-openers now and then.

It's too much work

No, it's not. For example, the GAVO data center is there to help you. Unless you data is particularly funky, you'll not have to spend more than half a day from the half-documented, messy goo that's on your disk to a shiny, blinking, proper, VO-registred, be-proud-of data service. And we'll take care of it henceforth.

Ask around at the booth for more information

My data is embarrassingly bad

Everyone's is. Good data is just bad data that more eyes have seen and more hands have improved.

http://docs.g-vo.org/talks/2013-tuebingen-lameex.pdf
(Markus Demleitner)

Thank you!

The PUNCH4NFDI Consortium

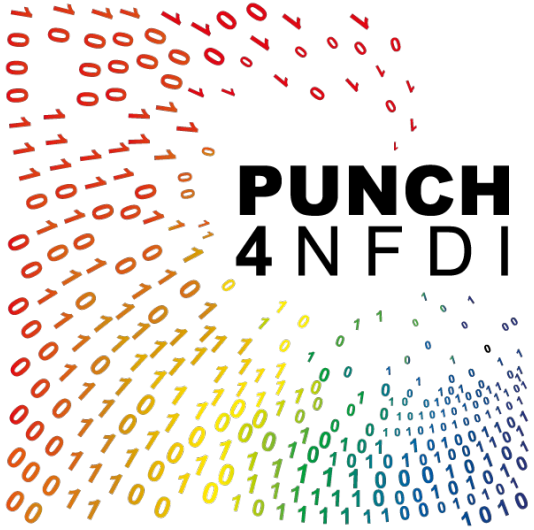
Spokesperson:
PD Dr. Thomas Schörner (DESY, thomas.schoerner@desy.de)
DESY, Notkestr. 85, D-22607 Hamburg

Contact:

Mail: punch4nfdi@desy.de

Web: www.punch4nfdi.de

Twitter: @punch4nfdi



Acknowledgements

DESY

PUNCH 4 NFDI

DFG Deutsche Forschungsgemeinschaft

HELMHOLTZ ASSOCIATION