

Data management and Reproducibility

DESY Summer student program

2025, September 1

Gernot Maier

“Wissenschaftler”.

“knowledge creators...”

- **create**
- **distribute, share, publish**
- **reproduce**
- **attribute**
- **re-use**

“Wissenschaftler”.

“knowledge creators...”

- **create**
- **distribute, share, publish**
- **reproduce**
- **attribute**
- **re-use**

Knowledge does not exist until you make it accessible
(publish & make it reproducible)

Knowledge as Research Objects

- Publications / Notes / Presentations
- Data (raw, derived, processed)
- Software
- Algorithms
- Simulations
- Videos
-

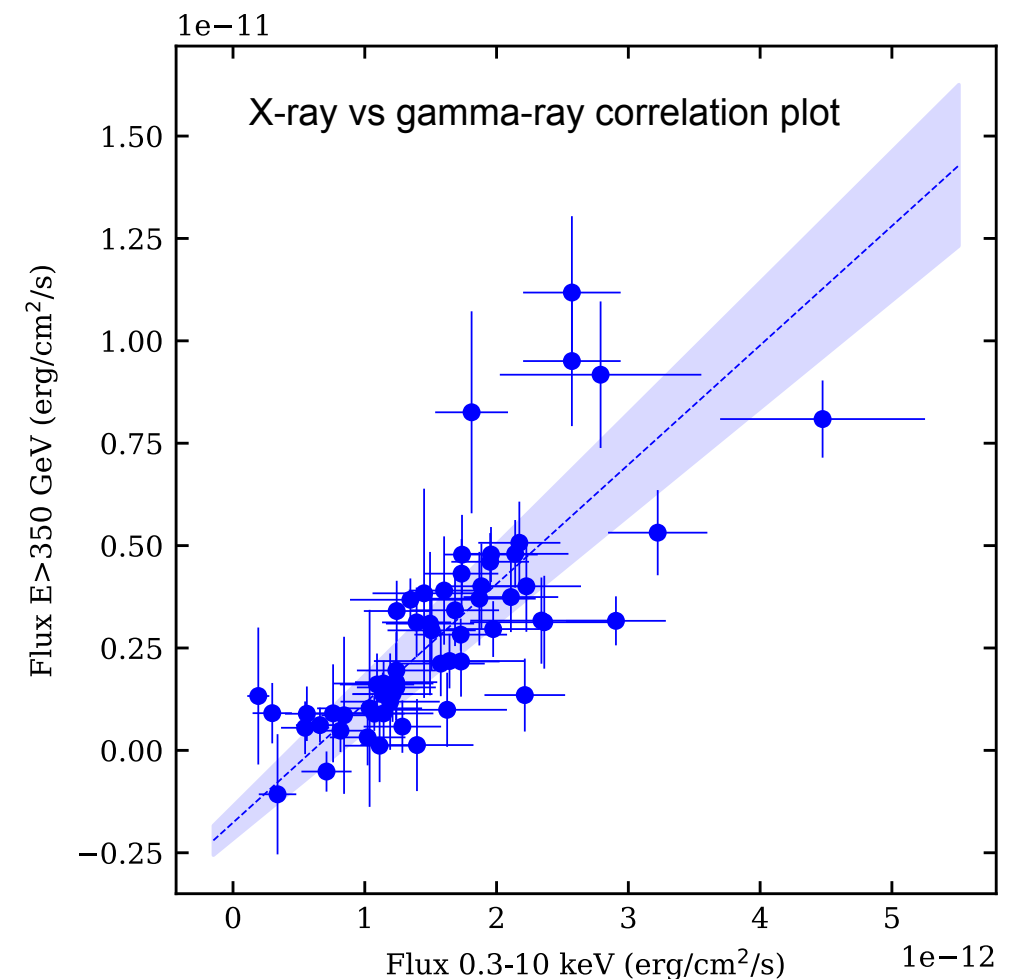
Publication consisting of
paper + data + software +
environment:
“dynamic and executable
publications”

“Research Data Management”

**Why is research data
management
important?**

Example

Would we be able to reproduce
this figure in 5 years?
5 months? Ever struggled to redo
a plot?



Data as plotted (flux points), units, high-level analysis, fit results, event lists,
instrument response functions, reconstruction code with all options, raw data files,

Re-run, Repeat, Reproduce, Reuse, Replicate

- Re-runnable (R^1): have you ever tried to re-run a program you wrote a few years ago?
- Repeatable (R^2): do you get the same result when running your code twice?
- Reproducible (R^3): can your colleague take your data and software and reproduce the result?
- Reusable (R^4): can your colleague make use of your data, algorithms, or software?
- Replicable (R^5): can your colleague take your data, writes his own software and come to the same conclusion? Are the algorithms applied documented?

see [Benureau et al \(2018\)](#)

Example (2)

Example (2)

- “Could we have the data used for figure 11 in paper X?”
“Yes of course - wait, not sure if it is in *final_data.dat*, *final_data_v1.dat*, *final_data_really_final.dat*, *final_more_final.dat*, *final_most_final.dat* ...”
“Yes - here is a link to the original data in my google drive. What, it doesn't exist anymore?”
“Yes, here are the files. It is saved as (*..*..*).”

(*) a file format which you have never heard off, and for which only a reader for Windows 95 exist

Example (2)

- “Could we have the data used for figure 11 in paper X?”
“Yes of course - wait, not sure if it is in *final_data.dat*, *final_data_v1.dat*, *final_data_really_final.dat*, *final_more_final.dat*, *final_most_final.dat* ...”
“Yes - here is a link to the original data in my google drive. What, it doesn't exist anymore?”
“Yes, here are the files. It is saved as (*..*..*).”

(*) a file format which you have never heard off, and for which only a reader for Windows 95 exist
- “Let's update the dark matter analysis published 5 years ago. Could we first look again at what data, software, algorithms were used?”
“Hmm, that was done by the PhD Student N, who graduated and left science. All data / software is on her laptop”
“No idea which software version was used? Does someone has the run list?”

Example (3)

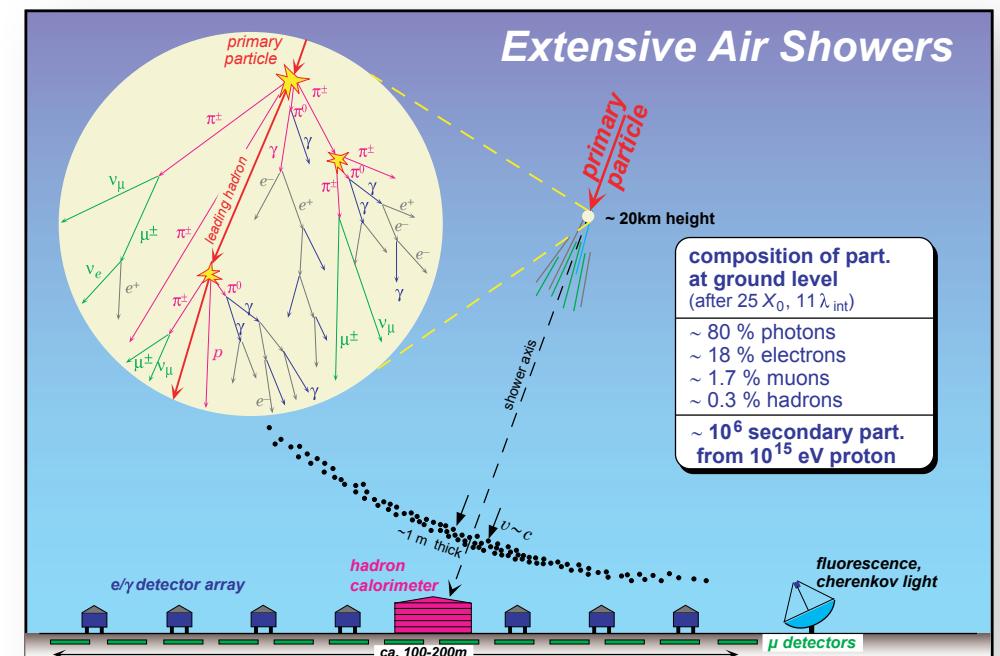
Example (3)

- “Very nice summary plot let’s use it in our publication”

“Wait - who did it?”

I found them on the web, can we use it?”

(note there is something like copyright)



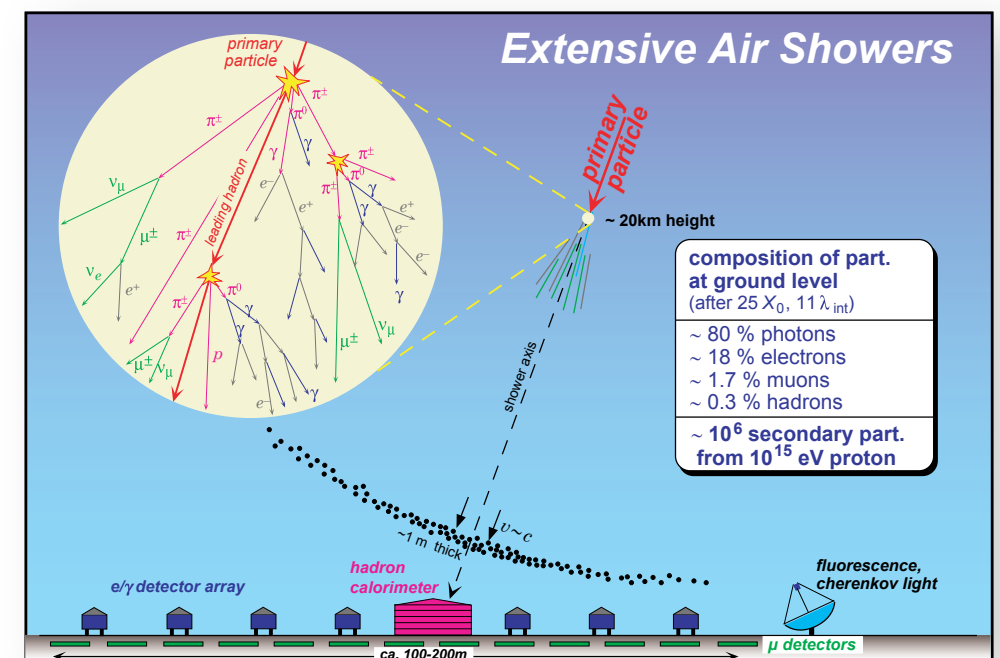
Example (3)

- “Very nice summary plot let’s use it in our publication”

“Wait - who did it?”

I found them on the web, can we use it?”

(note there is something like copyright)



- “Let’s query all published spectra of object type X and search for feature Y”

“Can’t find them, need to digitise them, ...”

Obvious (maybe good?) advices.

- document now (today) what you do (not tomorrow)
- add README files everywhere
“love letters to your future self”
- describe your data
(how was it derived; when; units; ..., Metadata)
- assume whatever you do, it will be wrong / full of mistakes
- assume that anything you do, you will have to do again
(write scripts and document it)
- assume that of any document, program, data set there will be different versions
- assume that what you do is useful for others

"FINAL".doc



FINAL.doc!



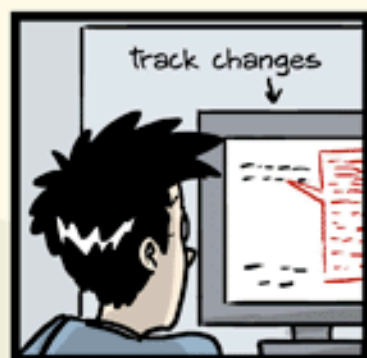
FINAL_rev.2.doc



FINAL_rev.6.COMMENTS.doc



FINAL_rev.8.comments5.
CORRECTIONS.doc



FINAL_rev.18.comments7.
corrections9.MORE.30.doc



FINAL_rev.22.comments49.
corrections.10.#@\$%WHYDID
ICOMETOGRADSCHOOL?????.doc



JORGE CHAM © 2012

Version Control - git.

✓ Update LAT-000023-2-lc.ecsv

changed the unit for LAT flux from m-2 s-1 to cm-2 s-1

main (#213)

v0.8.0

qi-feng committed 11 days ago

Verified

1 parent 97f00ec

commit bc71321a12d1b0b35df3dec1252f132be15b60d3

Showing 1 changed file with 2 additions and 2 deletions.

2022/2022ApJ...932..129A/LAT-000023-2-lc.ecsv

@@ -3,8 +3,8 @@	
3 # datatype:	3 # datatype:
4 # - {name: e_min, unit: MeV, datatype: float64}	4 # - {name: e_min, unit: MeV, datatype: float64}
5 # - {name: time, unit: MJD, datatype: float64}	5 # - {name: time, unit: MJD, datatype: float64}
6 - # - {name: flux, unit: 1e-7 m-2 s-1, datatype: float64}	6 + # - {name: flux, unit: 1e-7 cm-2 s-1, datatype: float64}
7 - # - {name: flux_err, unit: 1.e-7 m-2 s-1, datatype: float64}	7 + # - {name: flux_err, unit: 1.e-7 cm-2 s-1, datatype: float64}
8 + # meta: !!omap	8 # meta: !!omap
9 # - data_type: lc	9 # - data_type: lc
10 # - source_id: 23	10 # - source_id: 23

0 comments on commit bc71321

Lock conversation

You need to be familiar with git.

Good research practice

Code of Conduct

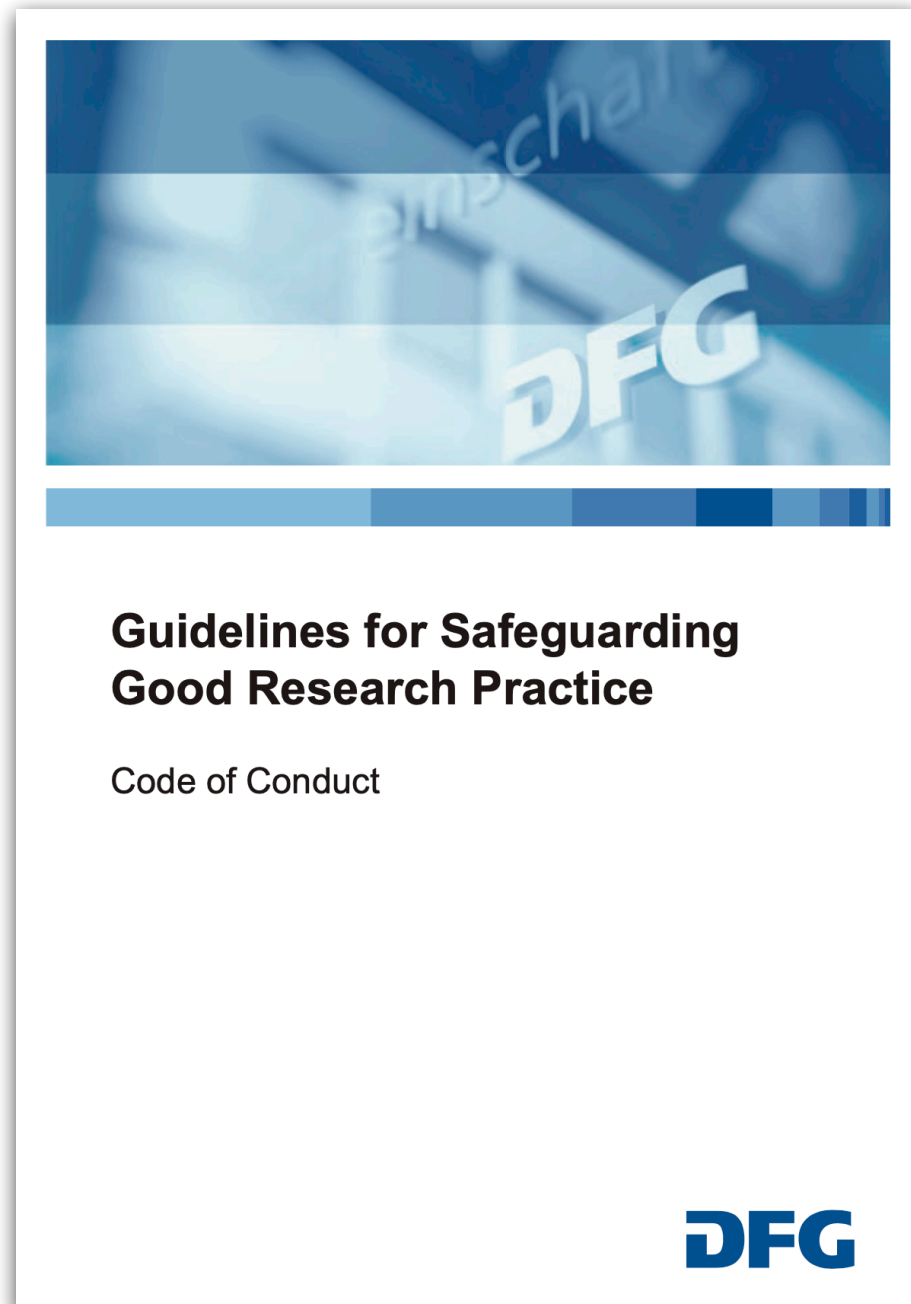
Guideline 12: Documentation

- Researchers document all information relevant to the production of a research result as clearly as is required by and is appropriate for the relevant subject area to allow the result to be reviewed and assessed. In general, this also includes documenting individual results that do not support the research hypothesis. The selection of results must be avoided. Where subject-specific recommendations exist for review and assessment, researchers create documentation in accordance with these guidelines. If the documentation does not satisfy these requirements, the constraints and the reasons for them are clearly explained. Documentation and research results must not be manipulated; they are protected as effectively as possible against manipulation.

Explanations:

An important basis for enabling replication is to make available the information necessary to understand the research (including the research data used or generated, the methodological, evaluation and analytical steps taken, and, if relevant, the development of the hypothesis), to ensure that citations are clear, and, as far as possible, to enable third parties to access this information. Where research software is being developed, the source code is documented.

There is something similar in your country.
Look for it and read it!



Scientific Misconduct

“Research misconduct is defined as fabrication, falsification, or plagiarism in proposing, performing, or reviewing research, or in reporting research results. (...)

Research misconduct does not include honest error or honest differences of opinion.”

OECD Global Science Forum

Fraudulent scientific papers are booming

A subset of journal editors may be partly responsible



IMAGE: BEN DENZER

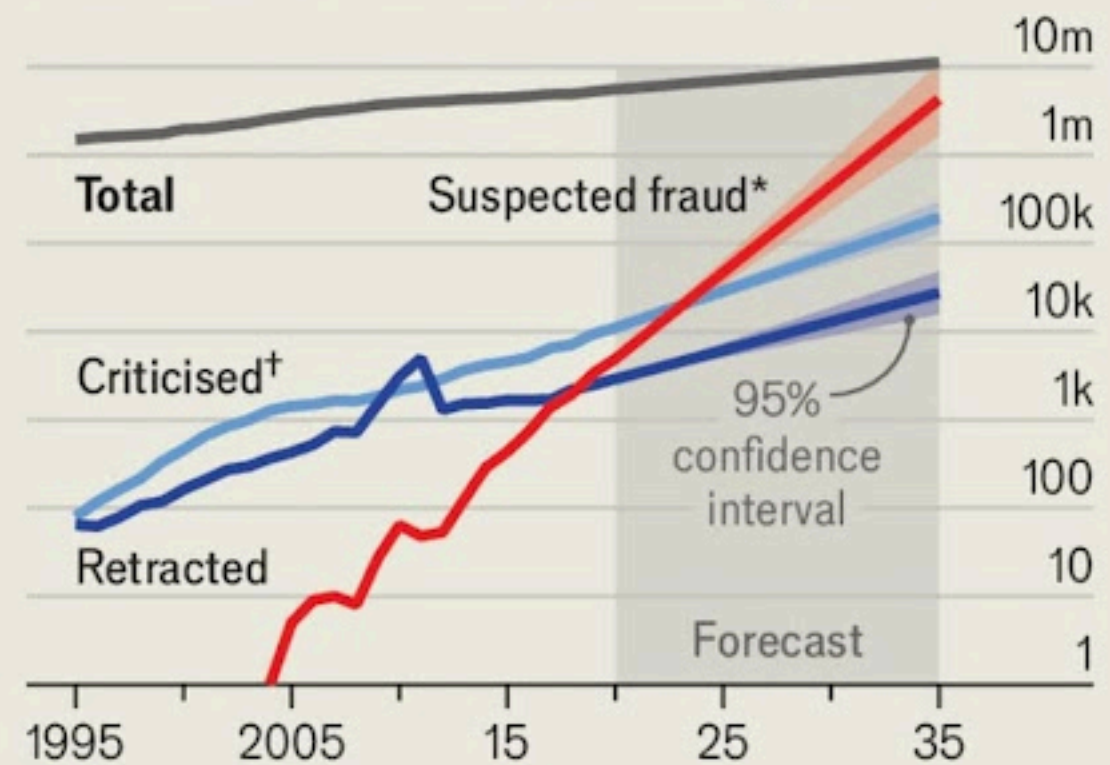
Aug 6th 2025 (updated Aug 7th 2025) · 5 min read

Listen 7:02

SCIENTIFIC JOURNALS exist to do one thing: provide accurate, peer-reviewed reports of new research to an interested audience. But

Paper trail

Global, annual scientific articles, log scale



*Produced by paper mills †On PubPeer.com

Source: "The entities enabling scientific fraud at scale are large, resilient and growing rapidly", by Luís A.N. Amaral et al., 2025

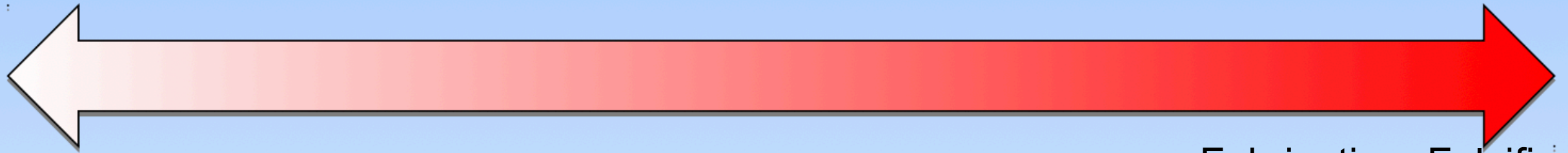
THE ECONOMIST

Degrees of Scientific Misconduct

Sloppy
work

Questionable
practice

Severe
misconduct



Carelessness
Mislabelling
Bad lab book

Bad statistics
Salami slicing
Intransparency
Using expired chemicals
Hiding “negative” results

Fabrication, Falsification,
and Plagiarism
FFP
Sabotage
Destroying data
Data theft
Ethics violation
Fake authors
Bad lab book

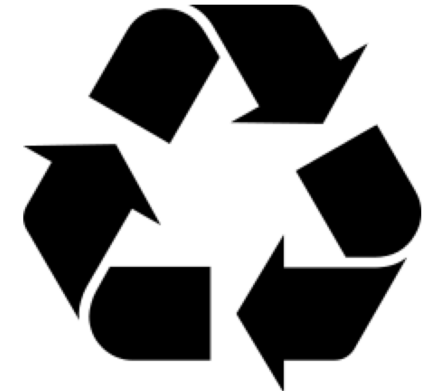
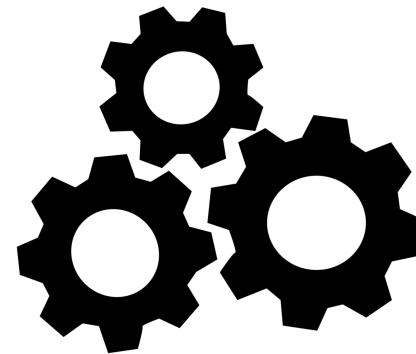
Scientific Misconduct

“The published reports on scientific misconduct are full of accounts of vanished original data and of the circumstances under which they had reputedly been lost. This, if nothing else, shows the importance of the following statement:

The disappearance of primary data from a laboratory is an infraction of basic principles of careful scientific practice and justifies a *prima facie* assumption of dishonesty or gross negligence.”

DFG Recommendations 2013, p. 75f

F_{indable} A_{ccessible} I_{nteroperable} R_{eusable}



A set of principles in data management that ensures that data is made available in a way that enables and stimulated reuse by humans and machines

Comes from life science - but today adopted for all data

Mark D. Wilkinson *et al.*[#] <https://www.nature.com/articles/sdata201618>

Sharing.



SPRINGER NATURE

F_{indable} A_{ccessible} I_{nteroperable} R_{eusable}

Box 2 | The FAIR Guiding Principles

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

assigned a persistent
identifier and
described by detailed
metadata

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
 - A1.1 the protocol is open, free, and universally implementable
 - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

Clear access
conditions;
machine access

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

Standardised formats
and vocabulary

To be Reusable:

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
 - R1.1. (meta)data are released with a clear and accessible data usage license
 - R1.2. (meta)data are associated with detailed provenance
 - R1.3. (meta)data meet domain-relevant community standards

Well-defined
licence; clear
provenance

Wilkinson et al 2015

Persistent Identifiers

an organisation made a promise to
keep it alive

globally unique string of characters

- DOI - digital object identifier
- ORCID - identify a person
- ROR - research organisation
registry
- (similar: ISBNs from books)

- **Open Researcher and Contributor ID**

C. B. ADAMS,¹ W. BENBOW², A. BRILL,¹ J. H. BUCKLEY,³ M. CAPASSO⁴,
A. FALCONE,⁶ K. A. FARRELL,⁷ Q. FENG,⁴ J. P. FINLEY,⁸ G. M. FOOTE,⁹
A. GENT,¹² G. H. GILLANDERS,¹³ C. GIURI,¹⁴ O. GUETA¹⁴, D. HANN¹⁵,
J. HOLDER,⁹ B. HONA¹⁷, T. B. HUMENSKY,¹ W. JIN¹⁸, P. KAARET¹⁹,
T. K. KLEINER¹⁴, F. KRENNRICH,⁵ S. KUMAR,¹⁵ M. J. LANG¹³, M. LUN²⁰,
P. MORIARTY¹³, R. MUKHERJEE⁴, D. NIETO²¹, M. NIEVAS-ROSILLO²²,
A. N. OTTE¹², N. PARK²³, S. PATEL¹⁹, K. PFRANG¹⁴, A. PICHE²⁴,
J. QUINN⁷, K. RAGAN¹⁵, P. T. REYNOLDS,²⁶ D. RIBEIRO,¹ E. ROACH²⁵,
M. SANTANDER¹⁸, S. SCHLENSTEDT,²⁷ G. H. SEMBROSKI,⁸ R. SHANG²⁸,
A. WEINSTEIN,⁵ D. A. WILLIAMS¹⁶, T. J. WILSON²⁹

(THE VERITAS COLLABORATION)

orcid.org

ORCID

Connecting research and researchers

Sign in / Register

English

Search the ORCID registry...

id

https://orcid.org/0000-0001-9868-4700

Show record summary

Personal information

Emails & domains

Verified email addresses

Verified email domains

Websites & social links

Keywords

Countries

gernot.maier@cta-observatory.org

gernot.maier@desy.de

cta-observatory.org

desy.de

DESY Astroparticle Section

Deutsches Elektronen-Synchrotron DESY

Cherenkov Telescope Array Observatory CTAO

VERITAS Gamma-ray Observatory

Astronomy, Astroparticle Physics

Germany

Biography

I am an astrophysicist at Deutsches Elektronen-Synchrotron DESY and the Cherenkov Telescope Array Observatory (CTAO). My main working fields are gamma-ray astronomy, high-energy emission from the direction of binary systems, black holes, and neutron stars, transient astrophysical sources, and the nature of dark matter.

I am a specialist in simulations and reconstruction methods for ground-based gamma-ray astronomy. I am also working in the framework of the National Research Data Infrastructure Germany (NFDI) on the implementation of FAIR data management workflows and archives in high-energy astronomy.

Much of my work is carried out as a member of the Cherenkov Telescope Array (CTA) Observatory and of the VERITAS Collaboration.

Selected publications and results:

- Long-term observation of the gamma-ray binary HESS J0632+057 (<https://doi.org/10.3847/1538-4357/ac29b7>) and LS I +61 303 (e.g. <https://arxiv.org/abs/2108.09235>)

- Gamma-ray observation of flaring binaries (e.g. <http://adsabs.harvard.edu/abs/2016ApJ...831..113A> and <https://ui.adsabs.harvard.edu/abs/2019ICRC...36..696H/abstract>)

- Dark matter searches, e.g. using by searching for anisotropies in the diffuse gamma-ray background (<https://ui.adsabs.harvard.edu/abs/2018JCAP...08..032H/abstract>)

- Characterization of the response of the future gamma-ray observatory CTA (<https://doi.org/10.5281/zenodo.5499839>)

- Lead Developer for the VERITAS and CTA reconstruction package Eventdisplay (<https://github.com/Eventdisplay/Eventdisplay>)

- VTSCat - The VERITAS Catalog of Gamma Ray Observations (<http://adsabs.harvard.edu/abs/2021arXiv210806424P>)

- Throughput calibration of the VERITAS observatory (<https://doi.org/10.1051/0004-6361/202142275>)

Activities

Employment (3)

Education and qualifications (1)

Professional activities (1)

Funding (5)

Works (50 of 496)

Expand all

Sort

Sort

Sort

Sort

Findable data

Findable **A**ccessible **I**nteroperable **R**eusable

- Deposit your data in a repository with metadata and a persistent identifier
 - e.g., online archives like zenodo+github
- Machine readable metadata that describes the datasets
 - Contextual information, title, author, keywords, when, what purpose, size, standards, ...
- **DigitalObjectIdentifiers**

Data is *not*** findable on your laptop**

The screenshot shows the Zenodo interface for a dataset. At the top is the Zenodo logo and navigation links. The dataset title is 'The VMC Survey - XXXVII. Pulsation periods of dust enshrouded AGB stars in the Magellanic Clouds' by Groenewegen, M.A.T. It shows 24 views and 1 download. The dataset is indexed in OpenAIRE. The publication date is March 18, 2020, with DOI 10.5281/zenodo.3714889. It is published in 'astronomy and astrophysics' and has a Creative Commons Attribution 4.0 International license. The files section lists two files: FigA1.tar (68.7 MB) and FigC2.tar (3.9 MB), both with download links. The citations section shows no citations.

Name	Size	Download
FigA1.tar	68.7 MB	Download
FigC2.tar	3.9 MB	Download

Accessible data

F_{indable} **A**_{ccessible} I_{nteroperable} R_{eusable}

- Can be open (but not necessarily)
- If not open: clear authentication and authorization
- Human and machine accessibility
 - (Your laptop is not machine accessible; neither tapes on a shelf; or some printouts)
- Open and free protocol

Not Found

The requested URL /oldpage.html was not found on this server.

Apache/2.2.3 (CentOS) Server at www.example.com Port 80

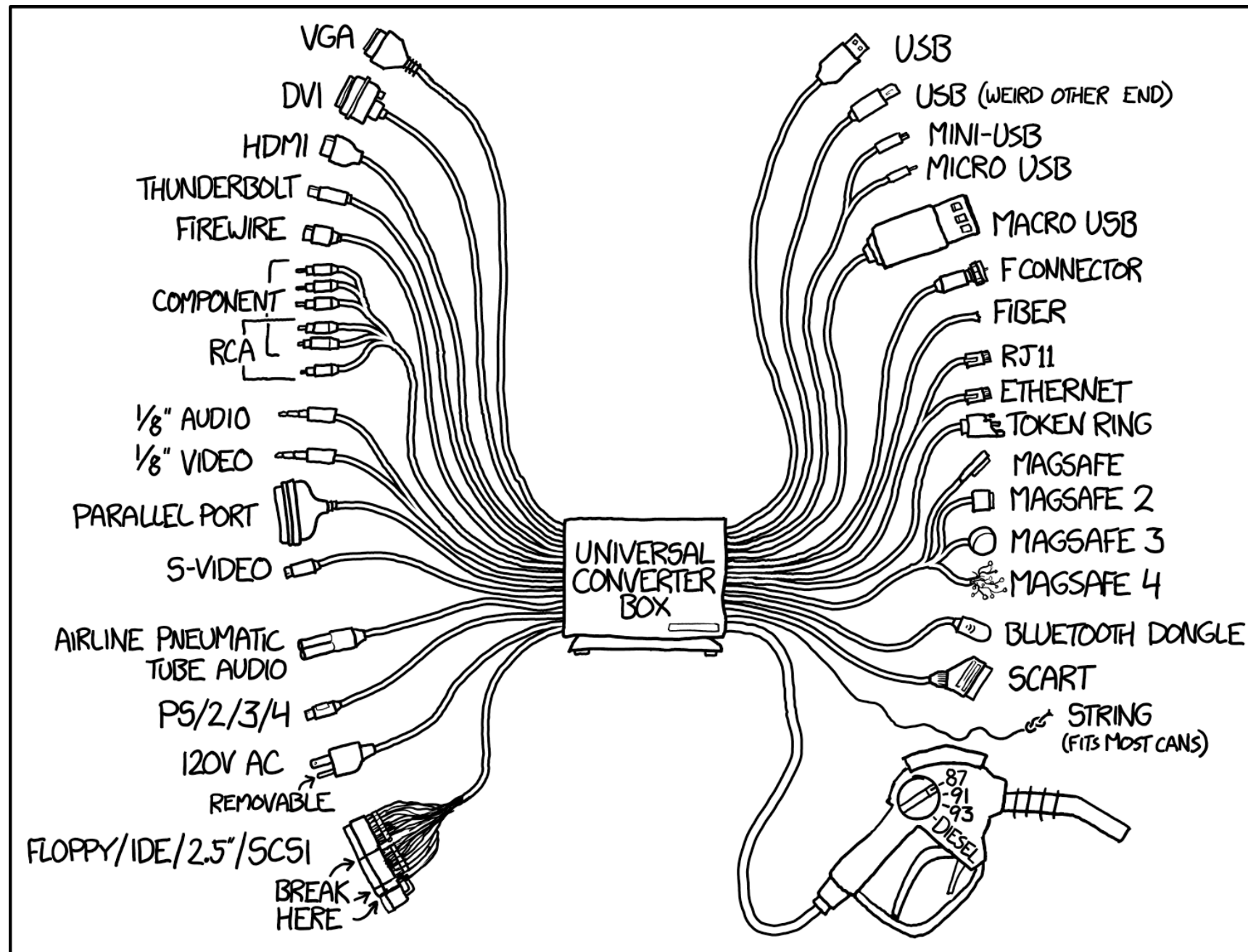


<https://enabofaisal.wordpress.com/2011/08/05/licensing-for-this-product-has-expired-cs4/>

Interoperable data

F_{indable} A_{ccessible} I_{nteroperable} R_{eusable}

Use a standard that can be mapped to others



<https://xkcd.com/1406/>

Interoperable data

F_{indable} A_{ccessible} **I_{nteroperable}** R_{eusable}

- Datasets are Interoperable if they are

- machine readable (metadata)
- specific formats (open/common)
- specific language and vocabularies

Example: FITS
data format +
FITS headers

- Formats should be:

- Community agreed, open, suitable for long-term preservation

- Metadata should use community agreed standards and vocabularies

- e.g., if the whole field talks about telescopes, don't use the expression 'antenna'

- Does your data include relevant provenance data?
 - Data is only reusable if it is known how it was obtained (e.g., software versions, IRF versions, etc)
 - Documentation, documentation, documentation
- Licensing - clearly stated re-use rights
 - (this is a tough topic for scientists)
- (You might want to get familiar with containers (e.g. run by Podman or docker))

AI tools

- use them - especially for routine task
 - text, code, ...
- learn how to write good prompts
- challenge for reproducibility
(no clear path on how this is solved)
- supports the importance of sharing data, software, knowledge (as public sources are used for training)

Conclusions - simple rules

(after Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, et al. (2014) Ten Simple Rules for the Care and Feeding of Scientific Data. PLoS Comput Biol 10(4): e1003542. doi:10.1371/journal.pcbi.1003542)

- **love your data, help others to love it**
- **share your data with a permanent identifier**
- **conduct science with reuse in mind**
- **publish workflows, methods, context**
- **link your data to your publications**
- **publish your code**
- **give credit and state how you want to get credit**
- **use data repositories**

Websites

- www.orcid.org
- zenodo.org
- github.com gitlab.desy.de ...
 - <https://docs.github.com/en/repositories/archiving-a-github-repository/referencing-and-citing-content>