LEIBNIZ INFORMATION CENTRE
FOR SCIENCE AND TECHNOLOGY
UNIVERSITY LIBRARY

TIB

Introductions

Katrin Leinweber, Angelina Kraft

TIB, 9. July 2018 Recording: doi.org/10.5446/37824

FAIR Data & Software (Carpentries-based workshop) **#TIBFDS**

Leibniz Association

Agenda for Introductions

1. the TIB
2. the Carpentries
3. this workshop
   - its instructors
   - its participants
4. The FAIR principle
   - for research data & software
   - DMPs & SMPs

German National Library of Science and Technology (TIB.eu)

Research library for science and technology, architecture, chemistry, computer science, mathematics and physics

Member of **Leibniz Association**, 500 members of staff

Global supplier for scientific and technical information

Founding member of DataCite

- 55,345 journal subscriptions (15,967 print; 39,378 digital)
- 9.1 m items, 17.3 m patents & standards

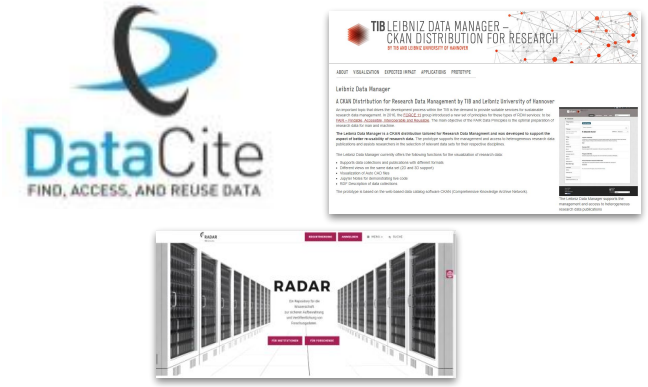
Managing digital resources

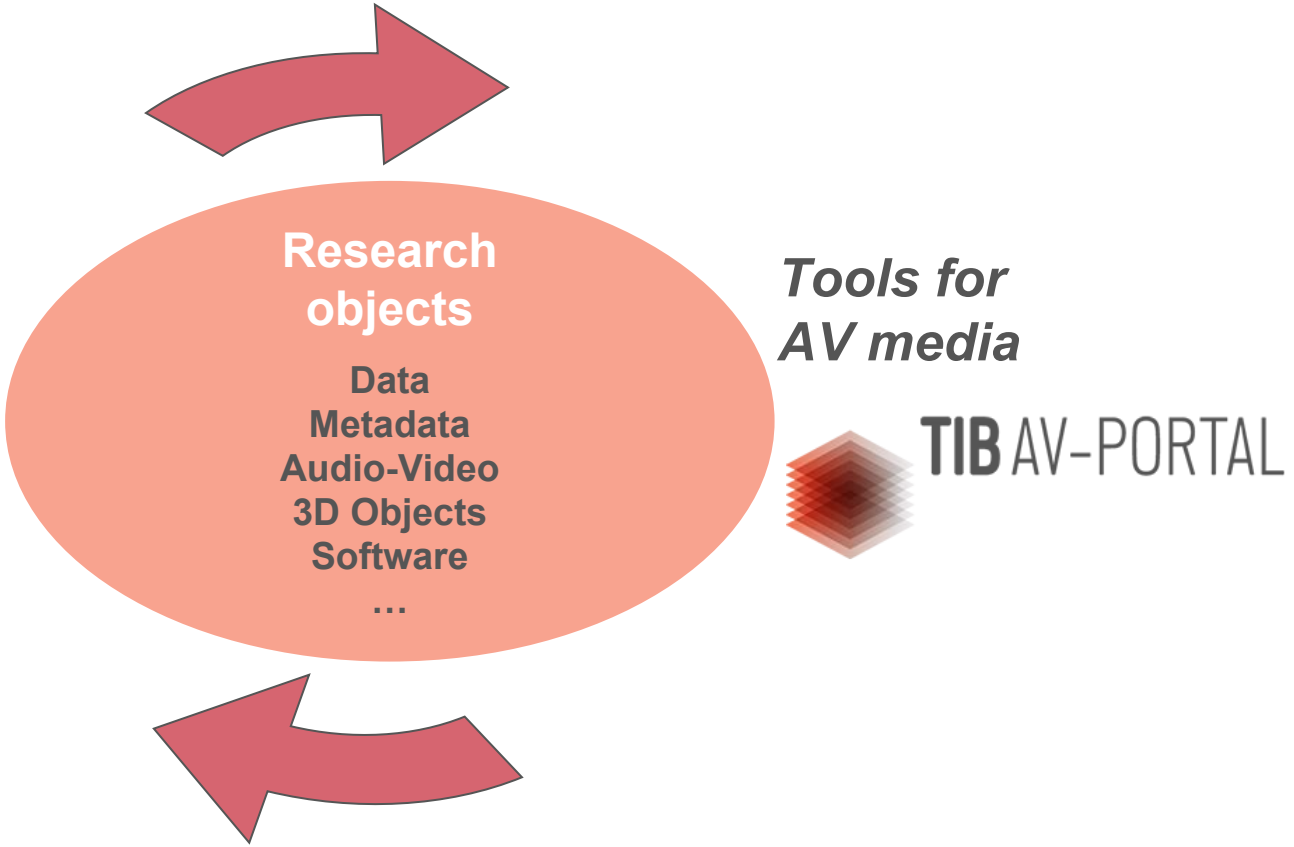
→ **Non-textual material referenced via discovery portal** (https://tib.eu)

TIB foci: Science and technology, architecture, chemistry, computer science, mathematics and physics:
Includes Research Data, 3D objects, AV media, …

Tools for data management

RDM guidelines,
DOI service, **Leibniz Data Manager + repository**





Tools for AV media

TIB AV-PORTAL

TIB as an information service and infrastructure provider

Communities: Variety of scientific & technical research objects
→ Unique characteristics & life cycle
→ Varying capability of accepting & managing new media formats
→ **Essential: Trust**

Roles of TIB:

- **Provide assurance & support for institutions** planning to submit their data & media to data centres & publishers
- Upgrade established **workflows** for indexing, cataloguing, digital preservation, DOI names, licencing
- Systematic **collection of non-textual materials**
- Develop **innovative, media-specific portals**
- **Linking non-textual materials to other research information** such as full texts & research data via the specialist portals & CRIS
- **Engage in communities**, provide training & open educational resources

→ **Trust in libraries as preservers of knowledge & research objects**

Agenda

expanding preservation into digital realm

1. the TIB
2. **the Carpentries**
3. this workshop
   - its instructors
   - its participants
4. The FAIR principles
   - for research data & software
   - DMPs & SMPs

THE CARPENTRIES

merged in 2017

software carpentry

DATA CARPENTRY

up & coming

library carpentry

∀uthorCarpentry

High-Performance Computing

- teach computing & data skills to scientists
  - things that “work too well to be worth teaching” doi.org/10.12688/f1000research.3-62
  - youtu.be/1e26rp6qPbA?t=35s
- collaborative, community-developed lessons: bio- & genomics, geospatial, social, ...
- evidence-based pedagogical methods (sources)
  - instructor training & mentoring
  - learning-by-doing & live coding

Interested in setting up a local chapter?

Agenda

1. the TIB expanding preservation into digital realm
2. the Carpentries teach the glue skills of scientific computing
3. **this workshop**
   - its instructors
   - its participants
4. The FAIR principles
   - for research data & software
   - DMPs & SMPs

2nd Conference on Non-Textual Information: Software and Services for Science (S3), av.tib.eu/series/310

Angelina's & Katrin's shared interest in the Carpentries "movement"

BMBF ideas competition „Wissenschaft im digitalen Wandel"

This "FAIR Data & Software" workshop

- Carpentries-based, due to CC-BY-licensed material
- start topics with short lectures, then exercise & discuss
- time for questions about FAIR-ifying your datasets & software

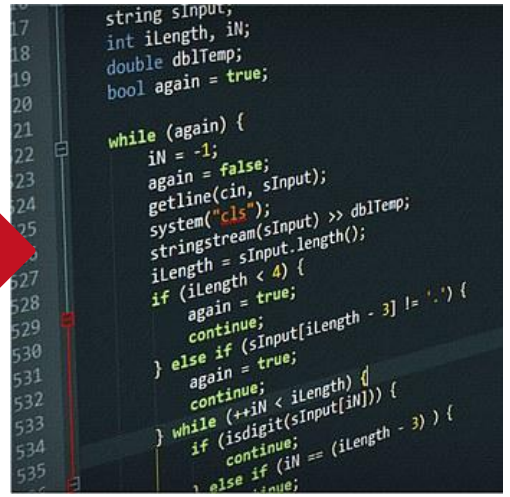
Our goals for this week are to help you:

Survey on Conference Recording Service among the Institutic

Publication year: 2018
Subject areas: Bibliothekswissenschaft
Creator/Author: Drees, Bastian

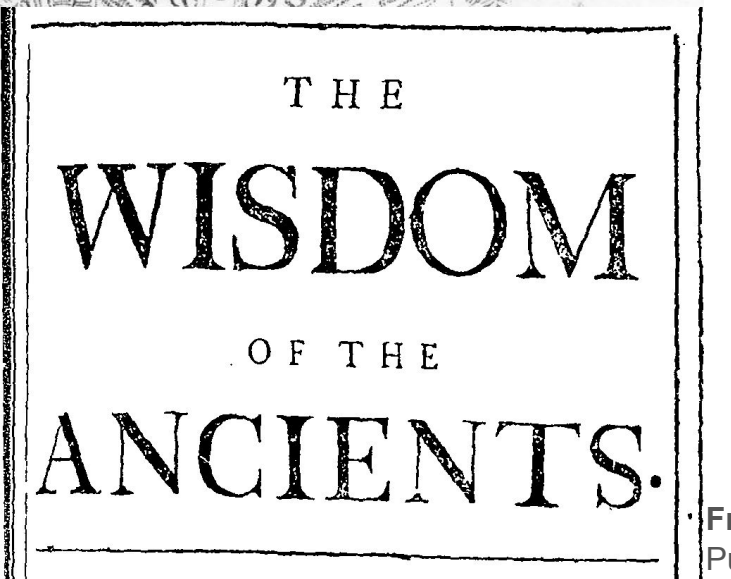
Questionnaire and Dataset of the TIB Survey 2017 on informatic

Publication year: 2017
Subject areas: Social Sciences / Bibliothekswesen / Library Science
Creator/Author: Technische Informationsbibliothek (TIB) (conducted by engage AG)



CC0 by pixabay.com/users/kuszapro

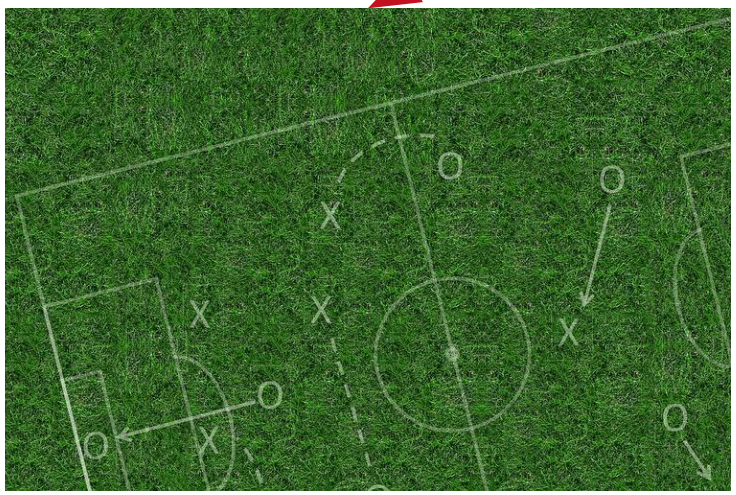
- implement FAIRer research data & software management & development practices
- focus on solution ideas for common problems, without promising “silver bullets” (there are none)



Francis Bacon
Public Domain via Wikimedia Commons

Few lessons are new, but some may need adaptation to new realities.

- 2 years in the lab can save you 2 hours in the library...
- read more old papers
- learn from colleagues' experience
- learn from the experience of other subject domains



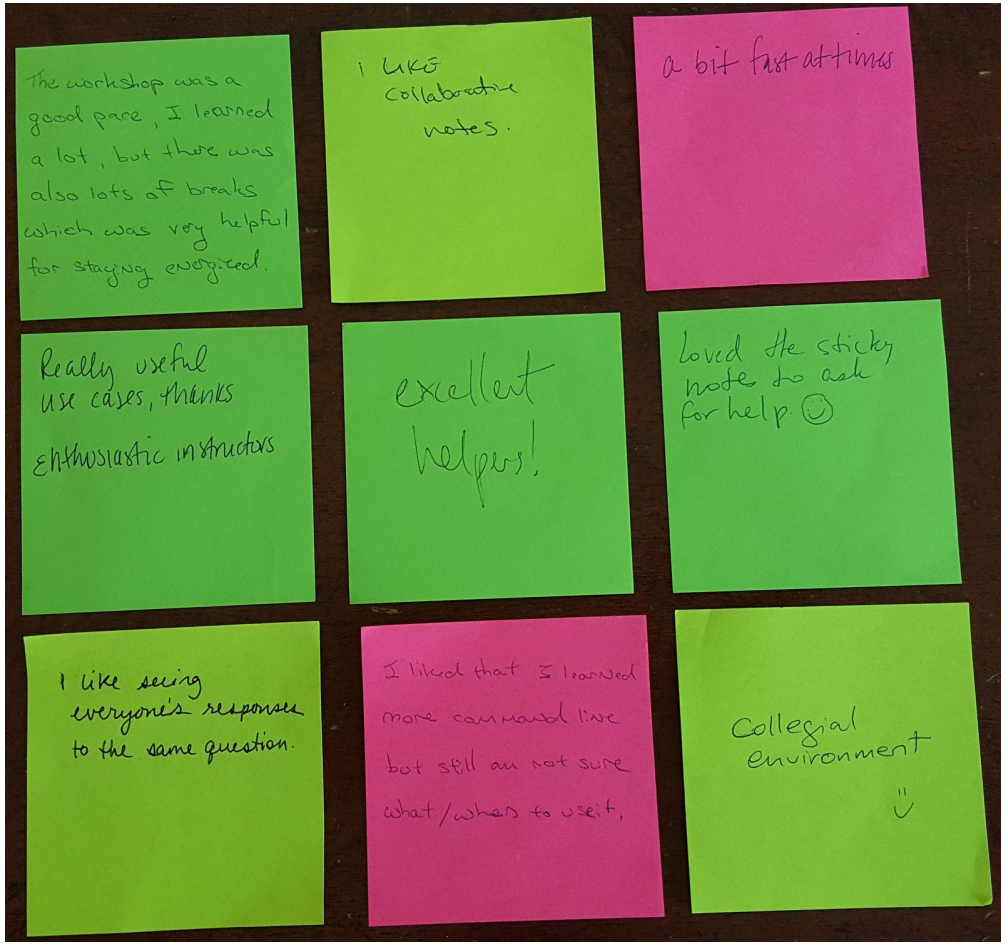
pixabay.com/de/users/annca (CC0) /Pexels

Open Educational Resources

- Carpentry-style lessons & workshop site → GitHub.com/TIBHannover
- recorded instructor presentations → av.tib.eu
  - Please write questions down for the discussion slots.
  - If you ask them immediately, the presenter will repeat them for the recording.

About the red & green sticky notes

- 1 thing about each session that you:
  - learned or liked
  - didn't understand or didn't like
- during live coding: indicate success or problem



© Tracy Teal DataCarpentry.org/blog/minute-cards

Agenda

1. the TIB expanding preservation into digital realm
2. the Carpentries teach the glue skills of scientific computing
3. this workshop experimental format, but established methods & content
   - its instructors
   - its participants
4. The FAIR principles
   - for research data & software
   - DMPs & SMPs

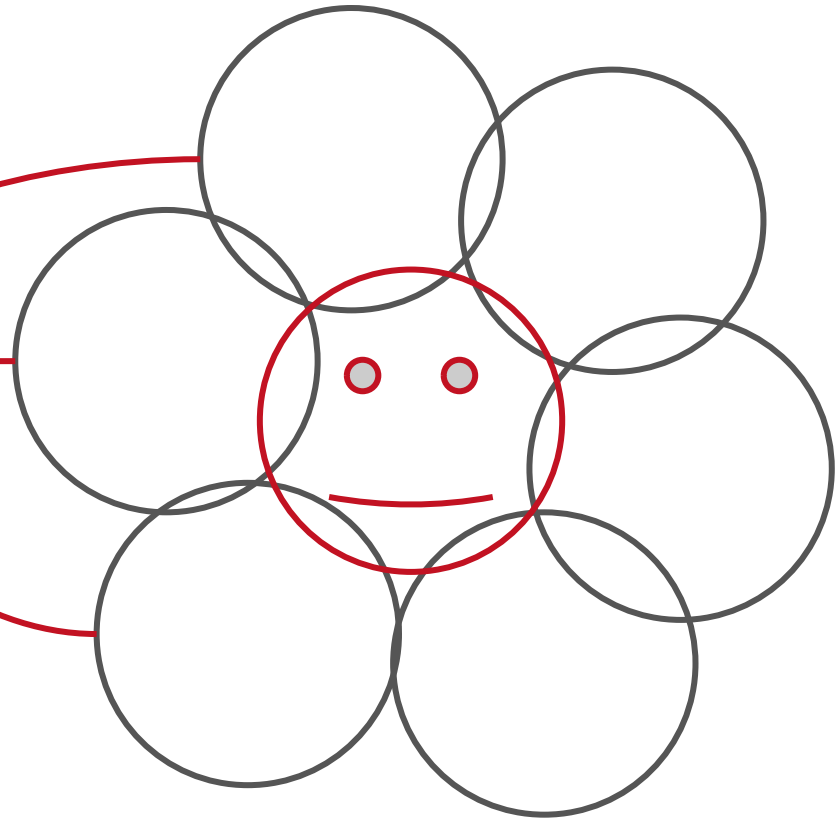
In such a diverse group, all might *feel* they know too little.

What you think you know

What you think others know

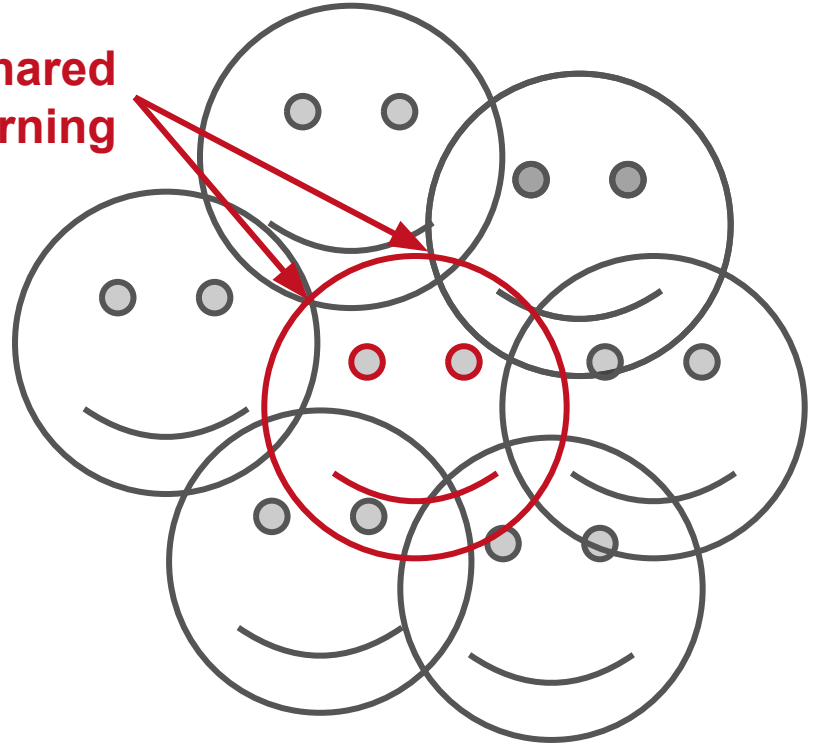
Reality

What others know



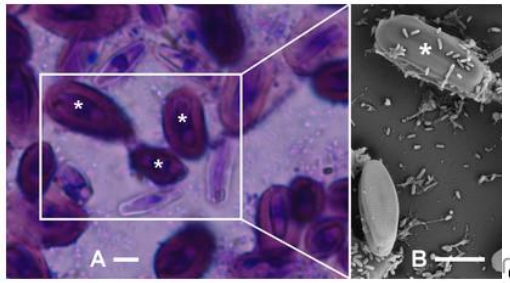
Our goal

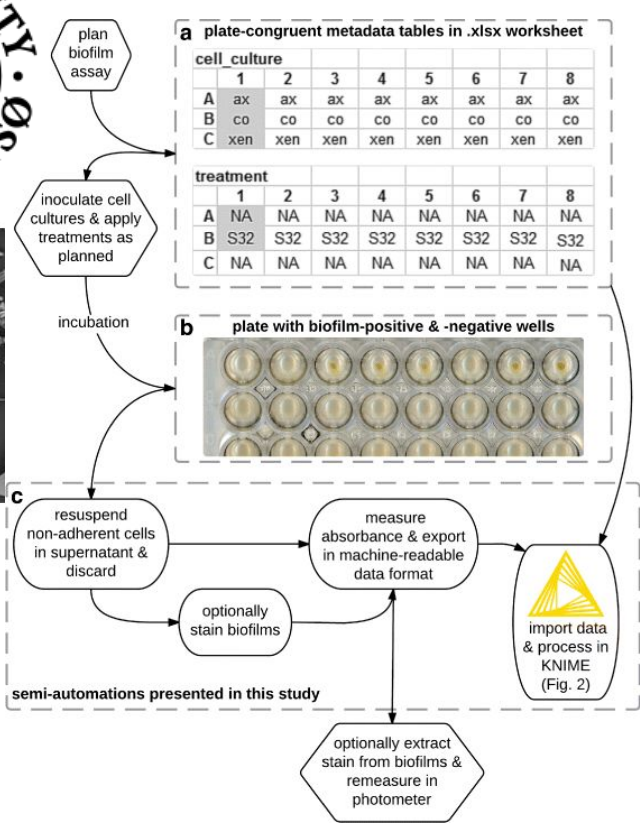
Shared learning



Katrin Leinweber (ORCiD.org/0000-0001-5135-5758)

- Life Science at Universität Konstanz and ERASMUS Arctic ecology & geology at University of Tromsø
- 3 years support specialist at Prezi
- PhD: method development for reproducible bioassays
  - coupled light & electron microscopy doi.org/10.7717/peerj.858
  - robotisation & data pipelining of a biofilm assay doi.org/10.1186/s12866-016-0676-9
  - R for data analysis & visualisation; git, LaTeX & MD for writing
- 7 quarters design & implementation of a pharma lab information & management
- at TIB since Nov'17: supporting scientific software projects
- Carpentries instructor & maintainer of r-novice-inflammation since Feb'18





Angelina Kraft (ORCiD.org/0000-0002-6454-335X)

- Data Manager, Project Manager, Researcher ...
- Background: Biology/Oceanography in Bremen/Bremerhaven
- PhD in Arctic Biology (& data management, also from research vessels); Postdoc marine sciences




- Since 09/2013 at TIB, head of Team Research Data and Scientific Software
- Research Data Management (RDM), Data Life Cycle, putting the FAIR principles into practice
- Various Research Data projects, establishing generic and discipline-specific RDM

Luke Johnston (ORCiD.org/0000-0003-4169-2616)

- Postdoctoral researcher in diabetes epidemiology at Aarhus University in Denmark
- Avid and frequent user of R, focus on packaging up analysis code into reusable R packages:
  - CRAN: prodigenr, mason, carpenter
  - GitHub: ggepi, famnet (in dev), broom (contributed), NetCoupler (collaboration)
- Coding teaching:
  - Software and Data Carpentry Instructor
  - AU Open Coders (au-oc.github.io)
  - UofTCoders (uoftcoders.github.io)
- GitHub: https://github.com/lwjohnst86

Konrad Förstner

- until recently: Head for Core Unit Systems Medicine, Würzburg
- now: Head of Information Services, ZB MED and Prof. for Information Literacy, TH Köln
- #Bioinformatics #Pythonista #OpenScience # ResearchSoftware
- GitHub.com/konrad
- @konradfoerstner

Technology
Arts Sciences
TH Köln

Introducing Mateusz Kuzak

Present:

- Scientific Community Manager at the Dutch Techcentre for Life Sciences
  - software and data training coordination in Life Science in NL
- the Carpentries instructor, trainer, mentor, EC member

Past:

- Research Software Engineer at the Netherlands eScience Center
- Bioinformatician / Data Analyst at the University of Amsterdam
- Life Scientist (Biotechnology, Biophysics, Cell Biology, Microscopy) at Jagiellonian University

  Kraków, Poland

Introducing Martin Hammitzsch

- Interests
  - Research Infrastructures and eScience platforms
  - Research Software Engineering
  - Research Data Management
- Stations
  - ICT in Research at
    Helmholtz Centre Potsdam - German Research Centre for Geosciences GFZ
    - eScience Centre (present)
    - Centre for GeoInformationTechnology - CeGIT
    - Data and Computing Centre
  - Industry (before 2008)
- Studies in Communication Systems and Software Engineering
- What else
  - de-RSE, Carpentries …

GFZ Helmholtz Centre POTSDAM

→ https://www.gfz-potsdam.de/en/staff/martin-hammitzsch/

Agenda

1. the TIB
2. the Carpentries
3. this workshop
   - its instructors (Konrad Förstner, Luke Johnston, Mateusz Kuzak & Martin Hammitzsch)
   - **its participants: You!**
4. The FAIR principles
   - for research data & software
   - DMPs & SMPs

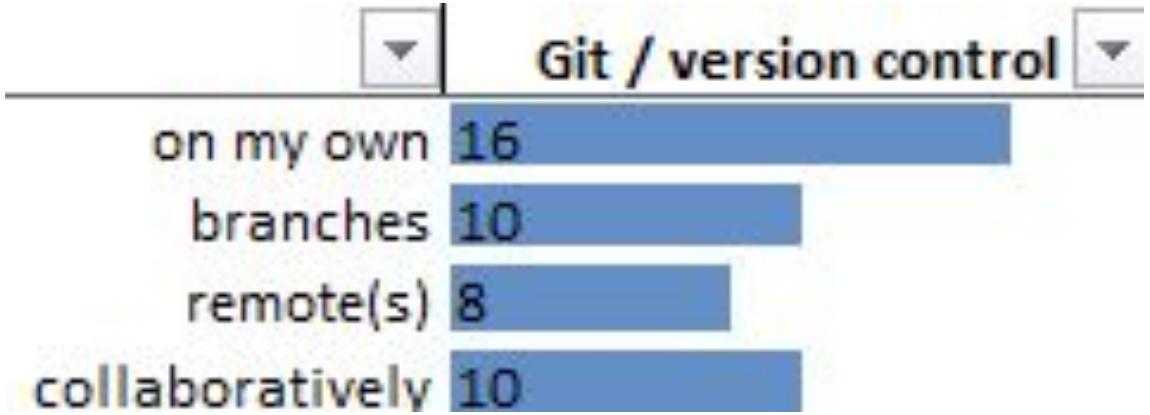
expanding preservation into digital realm

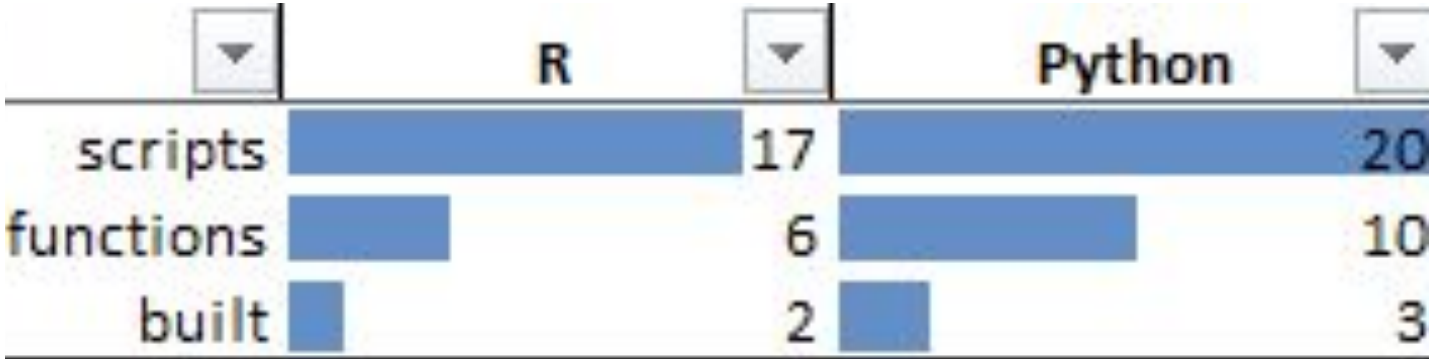
teach the glue skills of scientific computing

experimental format, but established methods & content

have natural sciences background & many years of software & data management

Summary of pre-workshop survey (N = 25, multiple-choice)

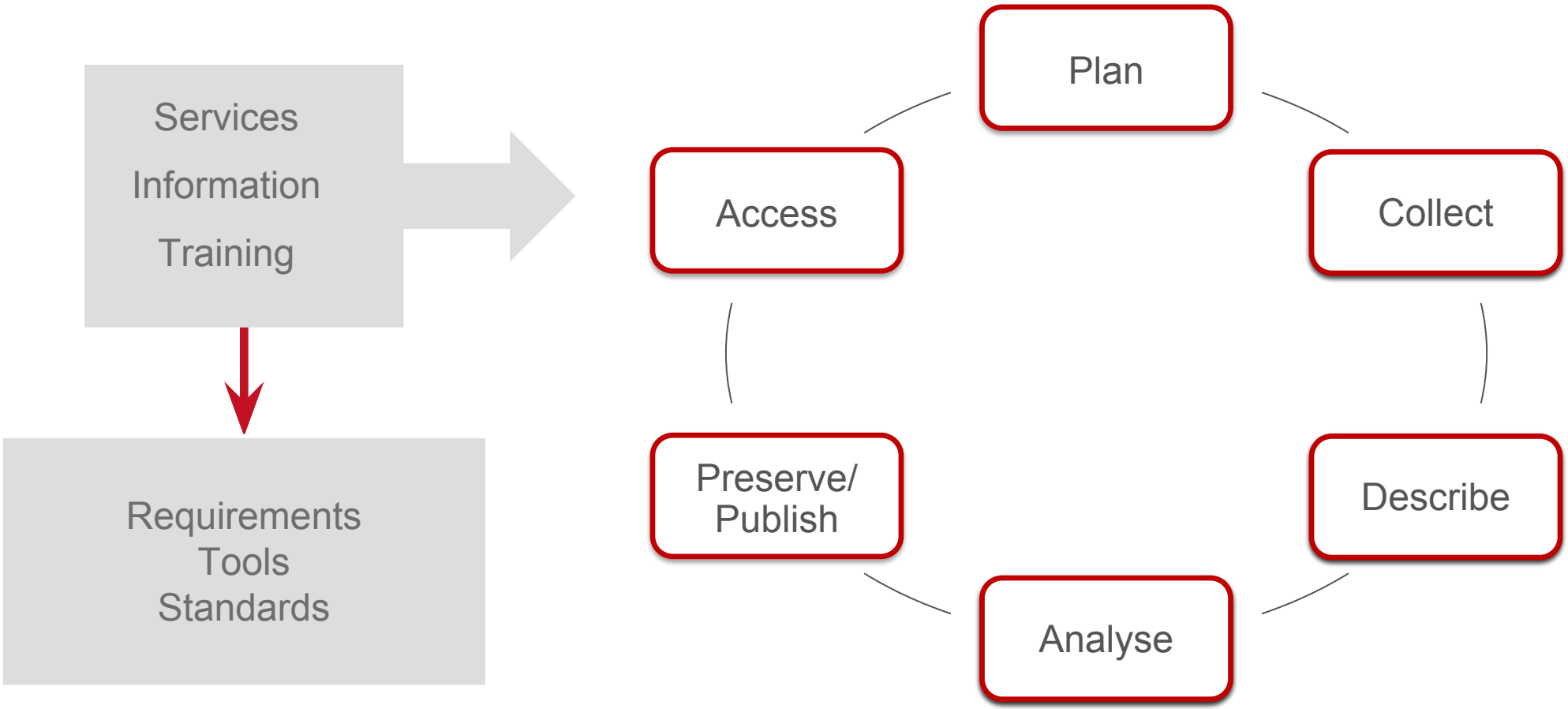




Agenda

1. the TIB — expanding preservation into digital realm
2. the Carpentries — teach the glue skills of scientific computing
3. this workshop — experimental format, but established methods & content
   - its instructors — have natural sciences background & many years of software & data management
   - its participants: You!
4. **The FAIR principles**
   - **for research data & software**
   - DMPs & SMPs

Research Data Management – Key Roles for Libraries



Modified after http://www.dcc.ac.uk/resources/curation-lifecycle-model

Requirements for Research Data Preservation and Sharing

1. Trustworthy research data repositories
2. Data policies
3. Standards for data citation, metadata, licencing
4. Intellectual property rights and proprietary data
5. Methods and Tools adapted to the scientific workflows
6. Cost recovery strategies
7. **Motivation for change**

Barriers of publishing data

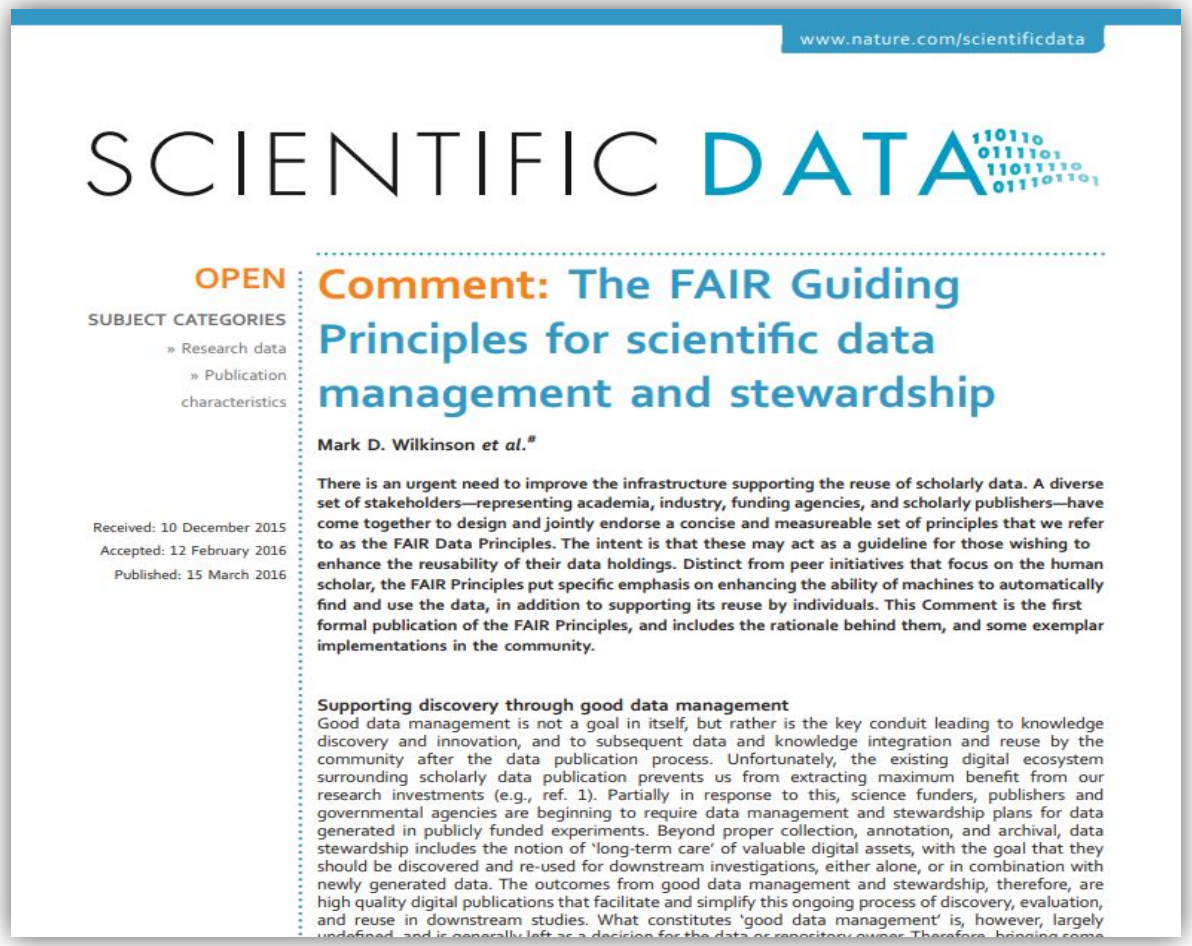
- Most research data is not accessible
  → Incentivise data sharing and publication
- Standardized metadata or documentation are not available
  → Mandatory metadata required for DOI registration
- No consistent practice of citing and referencing data
  → Providing citation standard DataCite Metadata Schema 4.1
- Researchers are unwilling or unable to share data
  → Inform researchers about benefits and possibilities

FAIR (and Software) Data Principles

Origins:

In 2016:
Findable
Accessible
Interoperable
Re-usable

Key point:
FAIR means FAIR
*for **machines***
(e.g. machine-readable metadata)
and only secondarily
for humans…



www.nature.com/scientificdata

SCIENTIFIC DATA

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.

Supporting discovery through good data management

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the notion of 'long-term care' of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. What constitutes 'good data management' is, however, largely undefined, and is generally left as a decision for the data or repository owner. Therefore, bringing some

Wilkinson et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data https://doi.org/10.1038/sdata.2016.18

The FAIR Data Principles by the FORCE11

to be Findable

1. (meta)data are assigned a globally unique and eternally persistent identifier
2. data are described with rich metadata
3. (meta)data are registered or indexed in a searchable resource
4. metadata specify the data identifier

to be Accessible

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

to be Interoperable

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

to be Reusable

1. meta(data) have a plurality of accurate and relevant attributes
   1.1. (meta)data are released with a clear and accessible data usage licence
   1.2. (meta)data are associated with their provenance
   1.3. (meta)data meet domain-relevant community standards

https://blogs.tib.eu/wp/tib/2017/09/12/the-fair-data-principles-for-research-data/

FAIR Data (and Software) Principles

In 2017, 2nd paper:

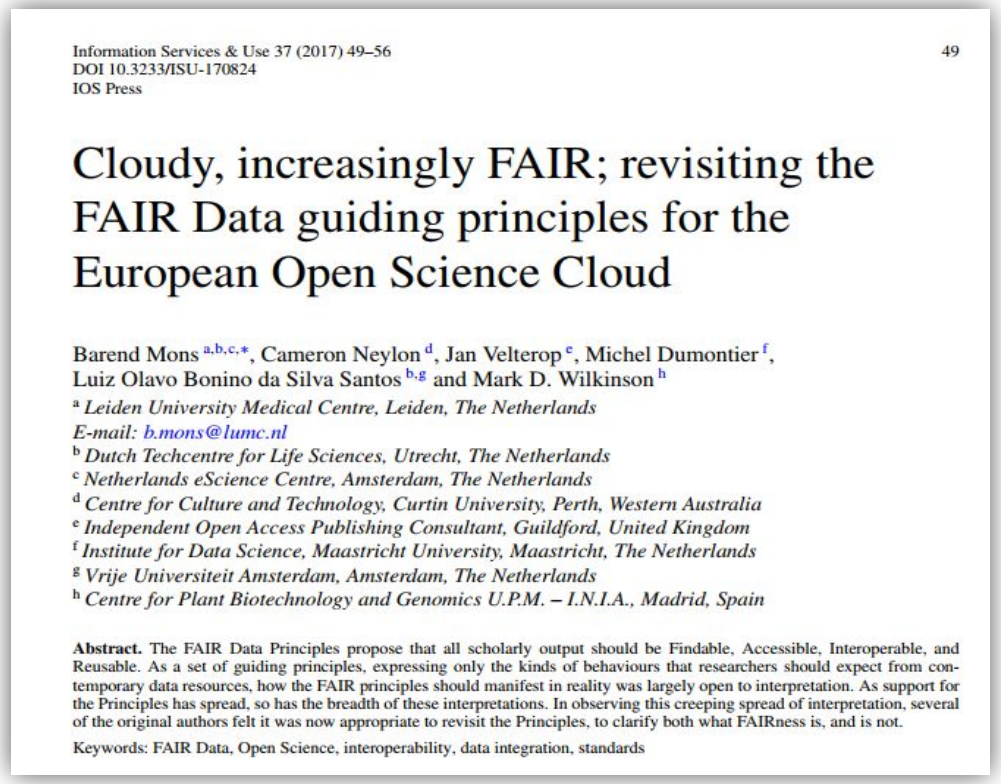
→ FAIR: not a standard

→ Different approaches

→ About data FAIRness for machines (and humans):
„*Partly FAIR may be FAIR enough*“

Mons et al. 2017, 6 categories:

1. Re-useless data
2. Findable (PID)
3. FAIR metadata
   (PID + machine readable MD)
4. FAIR data – restricted access
5. FAIR data – open access
6. FAIR data /
   open access & functionally linked
   = ‚*Internet of FAIR data and services*‘



Information Services & Use 37 (2017) 49–56
DOI 10.3233/ISU-170824
IOS Press

Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud

Barend Mons [a,b,c,*], Cameron Neylon [d], Jan Velterop [e], Michel Dumontier [f], Luiz Olavo Bonino da Silva Santos [b,g] and Mark D. Wilkinson [h]
[a] *Leiden University Medical Centre, Leiden, The Netherlands*
E-mail: b.mons@lumc.nl
[b] *Dutch Techcentre for Life Sciences, Utrecht, The Netherlands*
[c] *Netherlands eScience Centre, Amsterdam, The Netherlands*
[d] *Centre for Culture and Technology, Curtin University, Perth, Western Australia*
[e] *Independent Open Access Publishing Consultant, Guildford, United Kingdom*
[f] *Institute for Data Science, Maastricht University, Maastricht, The Netherlands*
[g] *Vrije Universiteit Amsterdam, Amsterdam, The Netherlands*
[h] *Centre for Plant Biotechnology and Genomics U.P.M. – I.N.I.A., Madrid, Spain*

Abstract. The FAIR Data Principles propose that all scholarly output should be Findable, Accessible, Interoperable, and Reusable. As a set of guiding principles, expressing only the kinds of behaviours that researchers should expect from contemporary data resources, how the FAIR principles should manifest in reality was largely open to interpretation. As support for the Principles has spread, so has the breadth of these interpretations. In observing this creeping spread of interpretation, several of the original authors felt it was now appropriate to revisit the Principles, to clarify both what FAIRness is, and is not.

Keywords: FAIR Data, Open Science, interoperability, data integration, standards

doi:10.3233/ISU-170824

FAIR for Software?

doi:10.12688/f1000research.11407.1

FAIR for Software? Loosely coupled...

EUROPEAN COMMISSION
Directorate-General for Research & Innovation

H2020 Programme

Guidelines on

FAIR Data Management in Horizon 2020

"Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?"

Peter Doorn (2017) Does it make sense to apply the FAIR Data Principles to Software? (Software Sustainability Workshop)

commons.wikimedia.org/w/index.php?curid=52216180 (Sae1962, CC BY-SA 4.0)

- software quality guidelines existed for decades in military, industry, academia & FLOSS
  - ISO 9000-3, 9126-1, 25010:2011
  - GNU Coding Standards & Quality Code
  - ECSS Software Product Assurance
  - CLARIAH software quality guidelines
- compromise, due to simplicity, popularity & politics

We'll follow the spirit of the FAIR principles, not always their letter.

Research Data Management

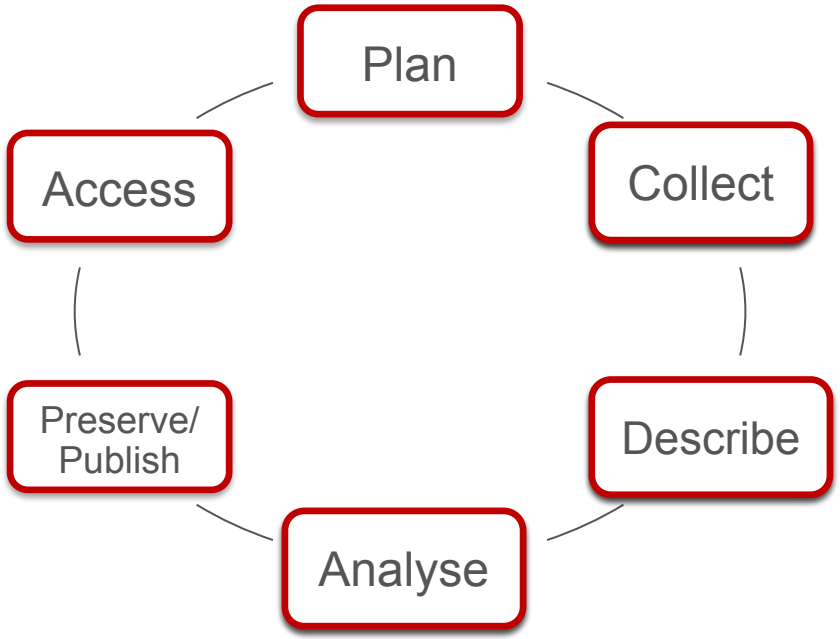
Research data management means to

- organize
- store
- preserve
- and possibly share

data from the beginning of a project (or from proposal writing) until the data is archived/published.

→ This requires **persons**, **tools**, services, etc.

→ Data Management Plan (DMP) as an instrument for curating data



Modified after http://www.dcc.ac.uk/resources/curation-lifecycle-model

Research Data Management and Software Management Services for Leibniz Universität Hannover

- RDM Service to support local researchers
- Cooperation between research department, IT-Service Center and TIB

TIB RDM Services

Infrastructures, services, training

- TIB focus:
  - Publication of research data / research software
  - Re-use and semantics, Data Science
  - DOI services
  - Legal aspects
  - Licences
- Basic Training or specialized courses
  - Data storage, structuring, data and software documentation, archiving, etc.
  - Data Management Plans / Software Management Plans

Agenda

1. the TIB — expanding preservation into digital realm
2. the Carpentries — teach the glue skills of scientific computing
3. this workshop — experimental format, but established methods & content
   - its instructors — have natural sciences background & many years of software & data management
   - its participants: You!
4. The FAIR principles — principles emphasise machine-actionability
   - for research data & software
   - **DMPs & SMPs**

What is a Data management plan?

A **Data management plan** …

- addresses **issues** related to data management
- might be required by **funding bodies** (NSF, EU H2020)
- is a (**formal) document** developed at the start of a research project which outlines **all aspects of data created/used**
- must be **updated** throughout the course of research

Common checklist (all DMPs):

- Administrative Information
- Data Collection
- Documentation and metadata
- Ethics and Legal Compliance
- Storage and Backup
- Selection and Preservation
- Data Sharing
- Responsibilities and Resources

Image: Stephanie Albert, CC0, by pixabay
https://pixabay.com/de/steine-stapeln-steinbalance-3390423/

General data approach:

as open as possible, as closed as necessary

RDM and Data Management Plans (DMPs)

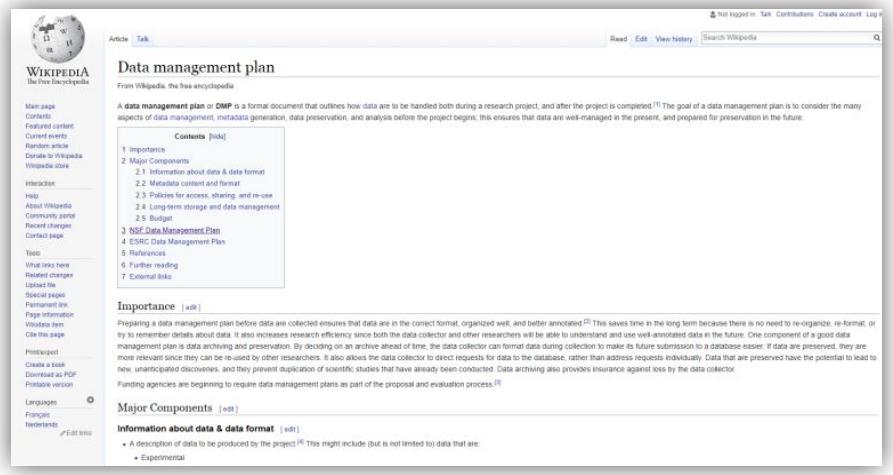
Present:

- Mixed DMP world of infinite variations …

RDMO

DMP ONLINE

DMPTool
Build your Data Management Plan



Future:

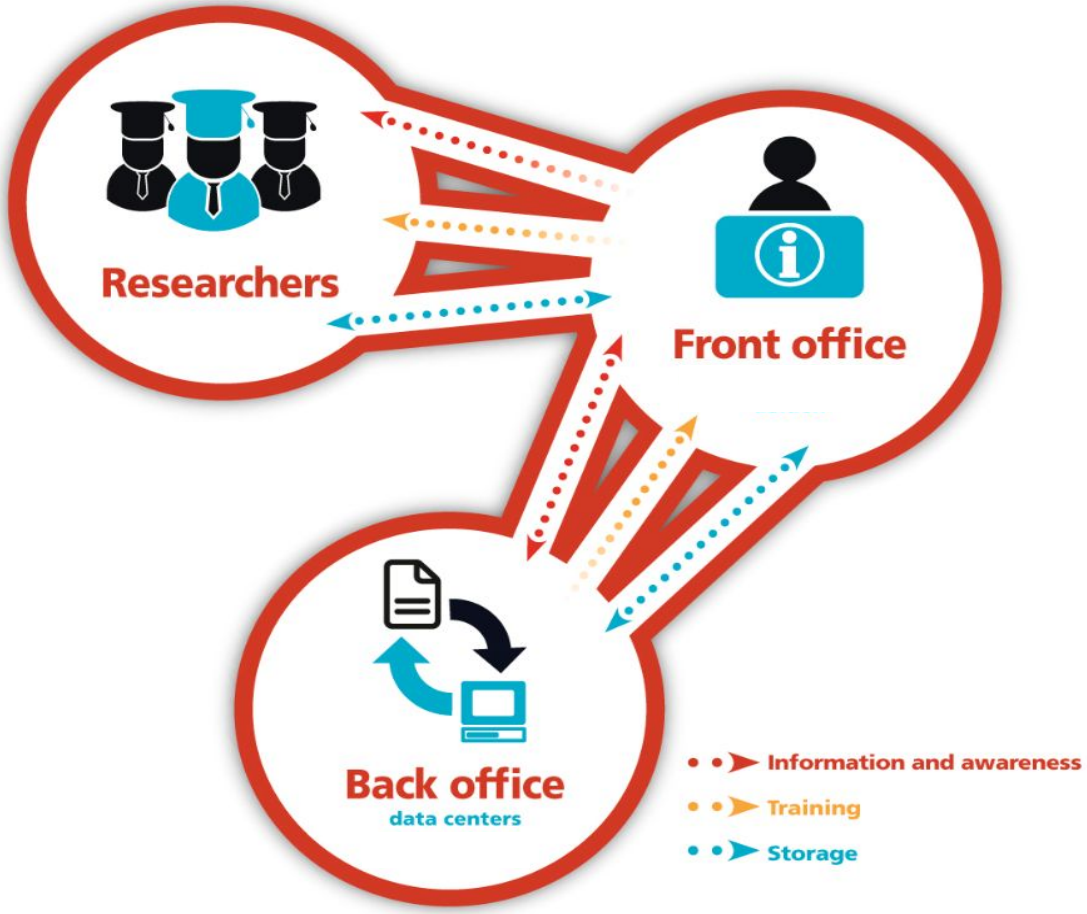
- Post-Static/Dynamic/Machine-Actionable DMPs with PIDs (DOI, ORCiDs)

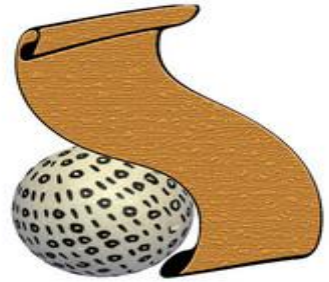
Stakeholders of a DMP:

- ➢ Researchers
- ➢ Institutions/Organizations
- ➢ Repositories/Infrastructure
- ➢ Funders
- ➢ Publishers

DMP: Include stakeholders

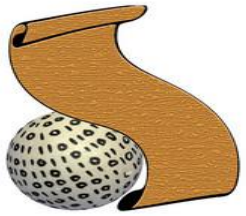
Commercial partners





Institution

RDM policy
Facilities



Publishers

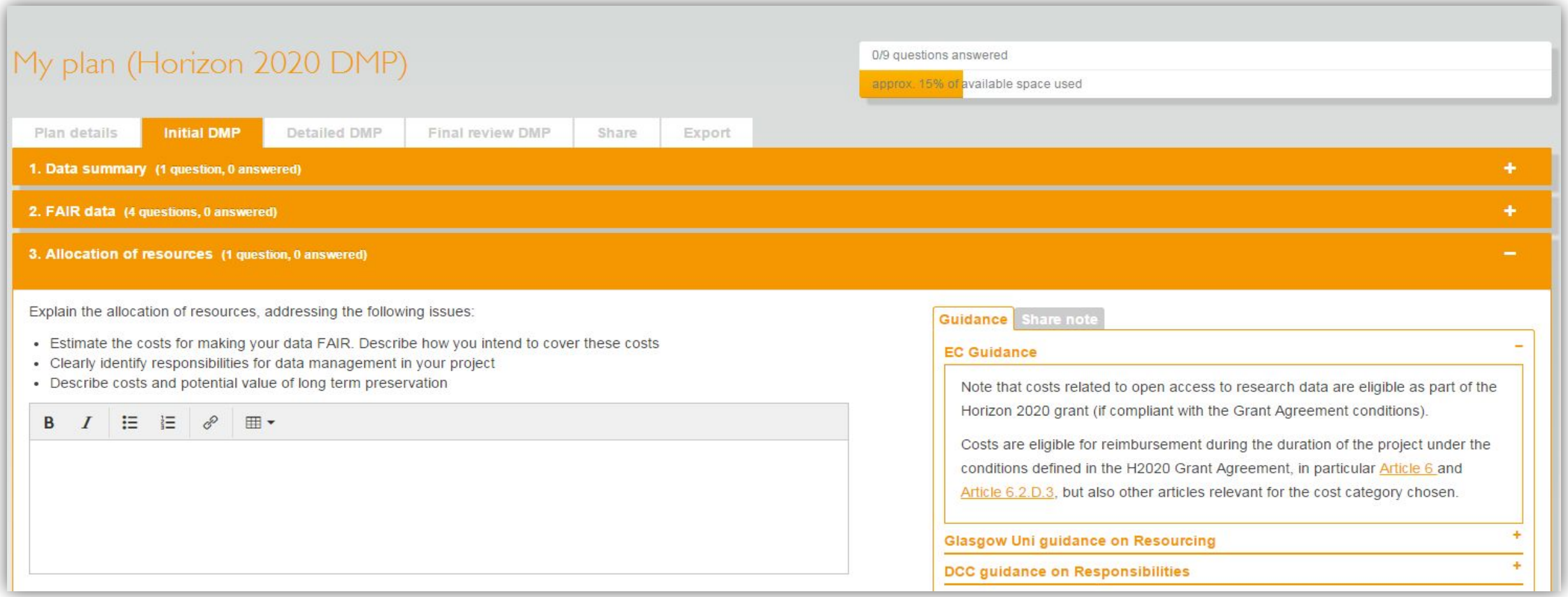
Data Availability policy

Research funders

Source: OpenAIRE Research Data Management Briefing paper: www.openaire.eu/briefpaper-rdm-infonoads, CC BY

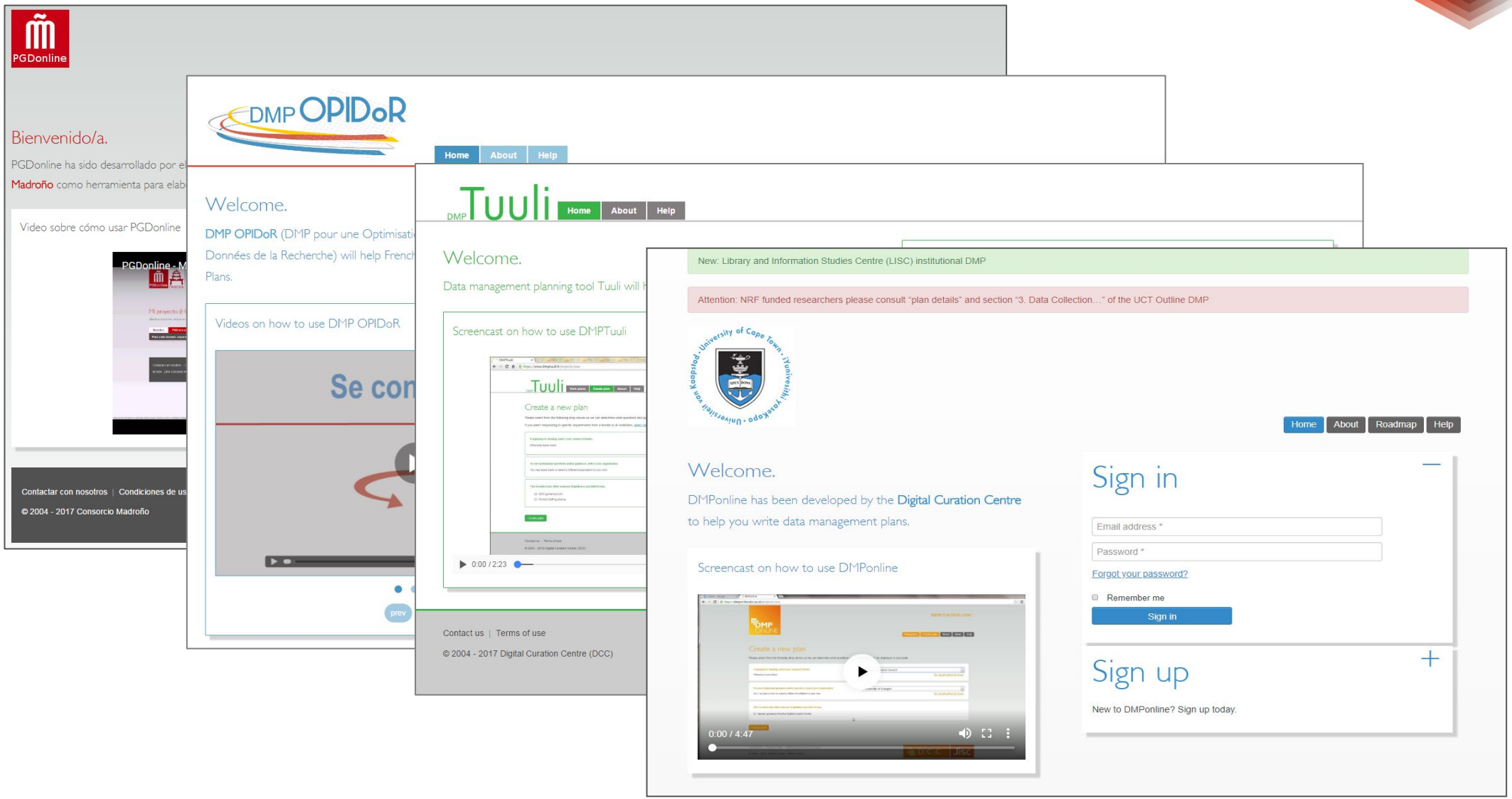
Example: UI of DMPonline

→ a web-based tool to help researchers write DMPs, which includes a template for Horizon 2020



DMPOnline.DCC.ac.uk

There are many national and local DMP Tools available …



→Overview: github.com/DMPRoadmap/roadmap/wiki/Local-installations-inventory

Since 2017: A single platform for all things DMP - *“roadmap”*

- “roadmap” is a single codebase, based on DMPonline with additional features from DMPTool
- Joining features of both tools
- Co-manage, co-develop and issue joint roadmap

DMPRoadmap: github.com/DMPRoadmap

Examples of funding bodies which require DMPs…

Software Management Plans (SMPs)

- overlap with general FLOSS advice:
  OpenSource.guide/starting-a-project & /best-practices
- be pragmatic about SMP's format: use funder template
  - include SMP in DMPOnline.DCC.ac.uk
  - if not required: Maybe frame your Git(Hub/Lab) issue tracker as a "living, public plan"?

To do	Doing	Done
#4 licence	#3 Dev plan	#1 Hypoth...
#5 Publish...		#2 Literatu...
#6 ...		

```
section: Managing your software development
intro:
- Managing who develops your software, what they do, how they will deve
questions:
- question: What effort will be available to develop your software?
  consider:
  - What funded effort will you have?
  - What unfunded, or additional, effort do you have available?
  - Will you accept contributions from your users?
  - Will you encourage your users to contribute to your software?
```

?

```
question: What licence will you choose?
consider:
- Will you be releasing your software?
- Does your funder, project or employer have a licence you are required to
```

The Software Sustainability Institute. (2016). Checklist for a Software Management Plan. v0.1. Available online: software.ac.uk/software-management-plans

Software Management Plans (SMPs)

- overlap with general FLOSS advice:
  OpenSource.guide/starting-a-project & /best-practices
- be pragmatic about SMP's format: use funder template
  - include SMP in DMPOnline.DCC.ac.uk
  - if not required: Maybe frame your Git(Hub/Lab) issue tracker as a "living, public plan"?

	To do	Doing	Done
	#4 licence	#3 Dev plan	#1 Hypoth...
	#5 Publish...		#2 Literatu...
	#6 ...		

```
section: Managing your software development
intro:
- Managing who develops your software, what they do, how they will deve
questions:
- question: What effort will be available to develop your software?
  consider:
  - What funded effort will you have?
  - What unfunded, or additional, effort do you have available?
  - Will you accept contributions from your users?
  - Will you encourage your users to contribute to your software?

?

question: What licence will you choose?
consider:
- Will you be releasing your software?
- Does your funder, project or employer have a licence you are required to
```

The Software Sustainability Institute. (2016). Checklist for a Software Management Plan. v0.1. Available online: software.ac.uk/software-management-plans

Agenda

1. the TIB — expanding preservation into digital realm
2. the Carpentries — teach the glue skills of scientific computing
3. this workshop — experimental format, but established methods & content
   - its instructors — have natural sciences background & many years of software & data management
   - its participants: You!
4. The FAIR principles — principles emphasise machine-actionability
   - for research data & software
   - DMPs & SMPs — help you think backwards from desired outcomes to necessary actions

Exercise & discussion in groups (over lunch)
Summary into HackMD.io/ERXxduPVTPSc3LvSjL2nfw please!

What does “FAIR”

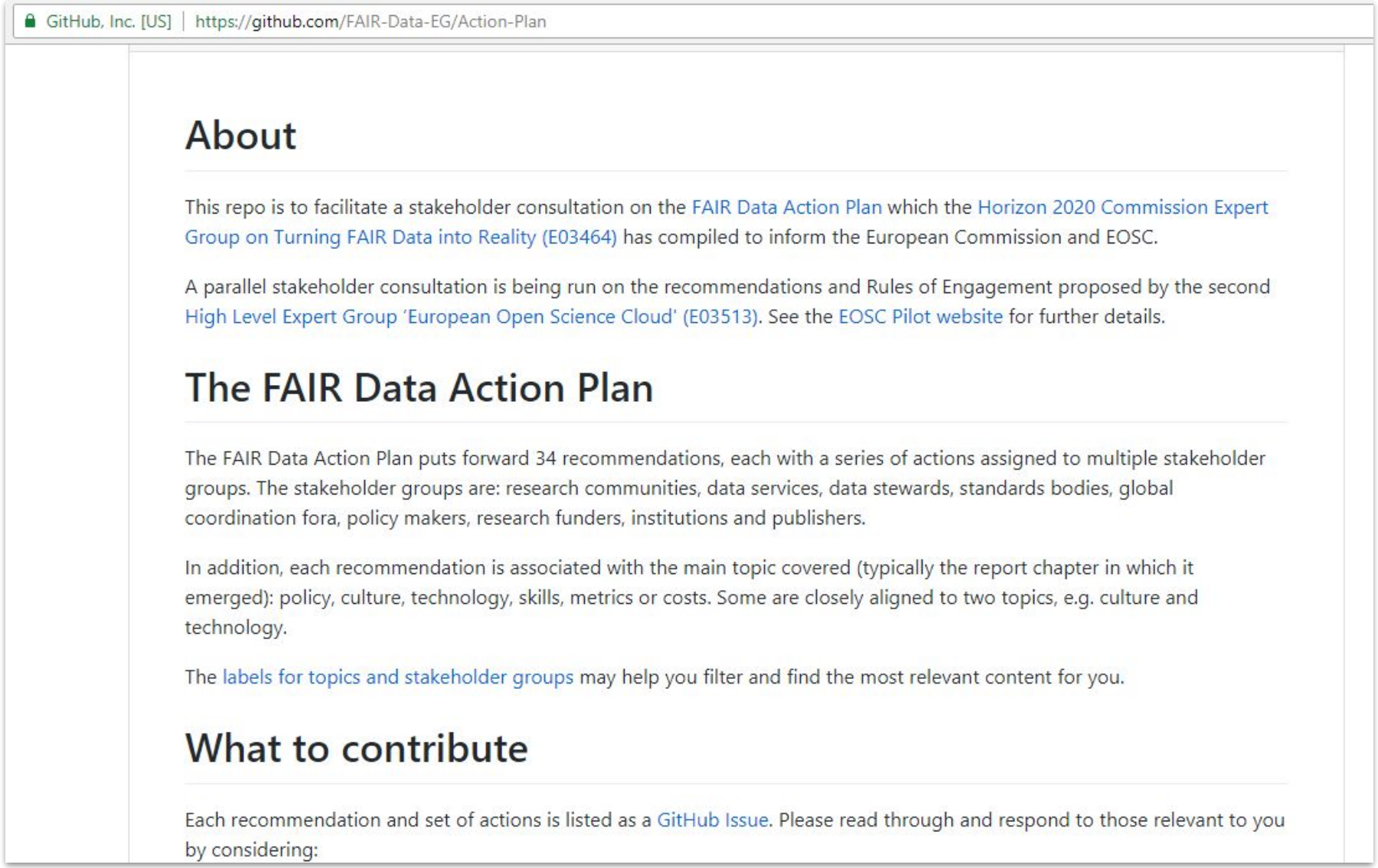
mean for your domain

of study / research /

work or field of

expertise?

FAIR Data Action Plan



https://github.com/FAIR-Data-EG/Action-Plan

LEIBNIZ INFORMATION CENTRE
FOR SCIENCE AND TECHNOLOGY
UNIVERSITY LIBRARY

TIB

Which questions do you have for us?

Contact information:
Katrin.Leinweber@TIB.eu & Angelina.Kraft@TIB.eu
T +49 511 762-14693 & -14238

Leibniz Association