Universiteit Utrecht

FAIR & Research Software

Dr. Anna-Lena Lamprecht (@al_lamprecht)
Utrecht University, Netherlands

GO FAIR US Webinar Series, 14 January 2021

About me

Universiteit Utrecht

Dr. Anna-Lena Lamprecht

Assistant professor of Software Technology at Utrecht University, Netherlands

Areas of work: (research) software engineering and applied formal methods

Current focus: scientific workflows, FAIR software, computational science and RSE education

Application areas: life sciences and geosciences

OSCU Open Science Community Utrecht

RSE

Utrecht Bioinformatics Center

EASST

elixir

FORMAL METHODS EUROPE

Outline

Recap: The FAIR data principles

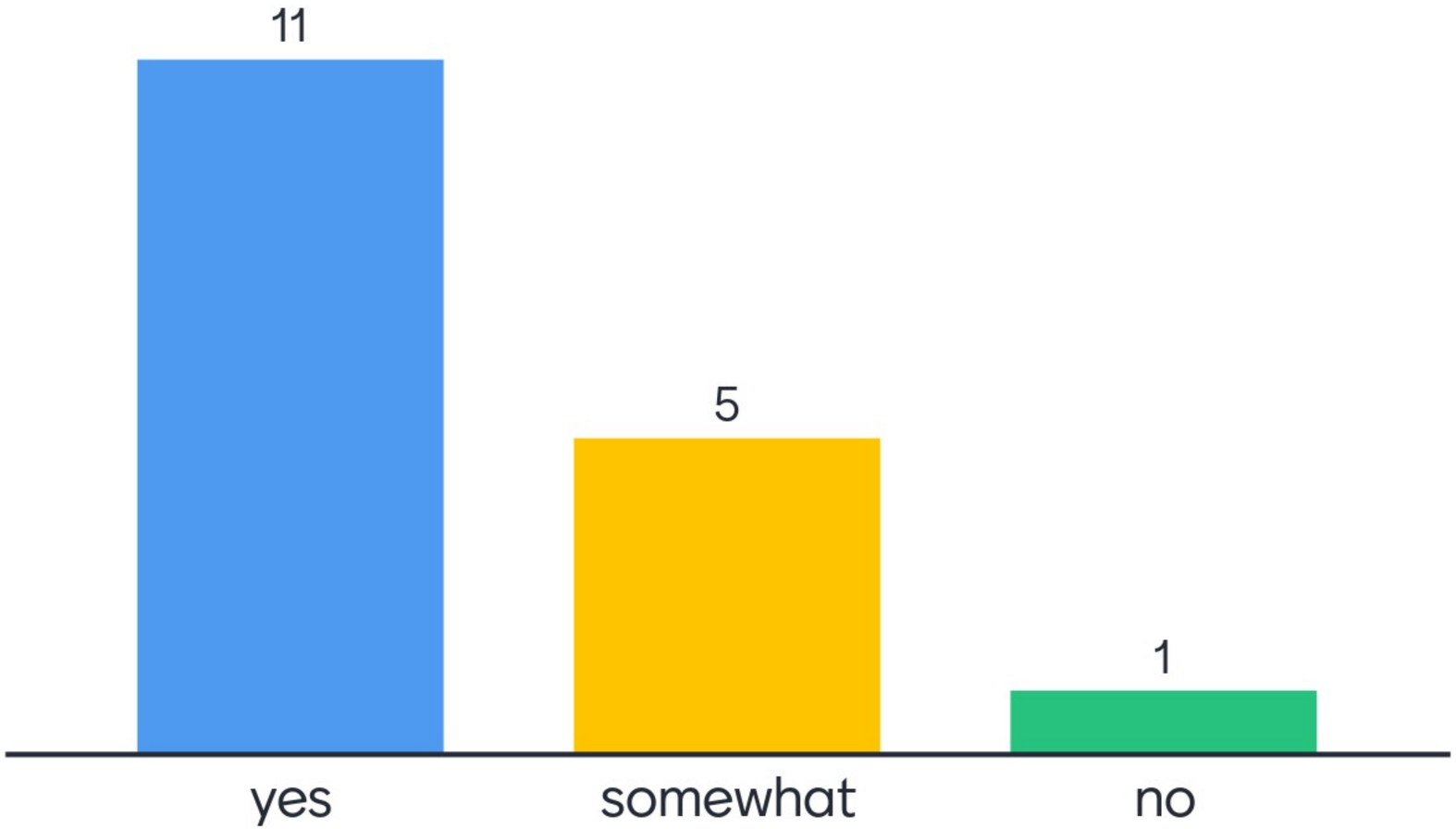
Data and software

Towards FAIR principles for research software

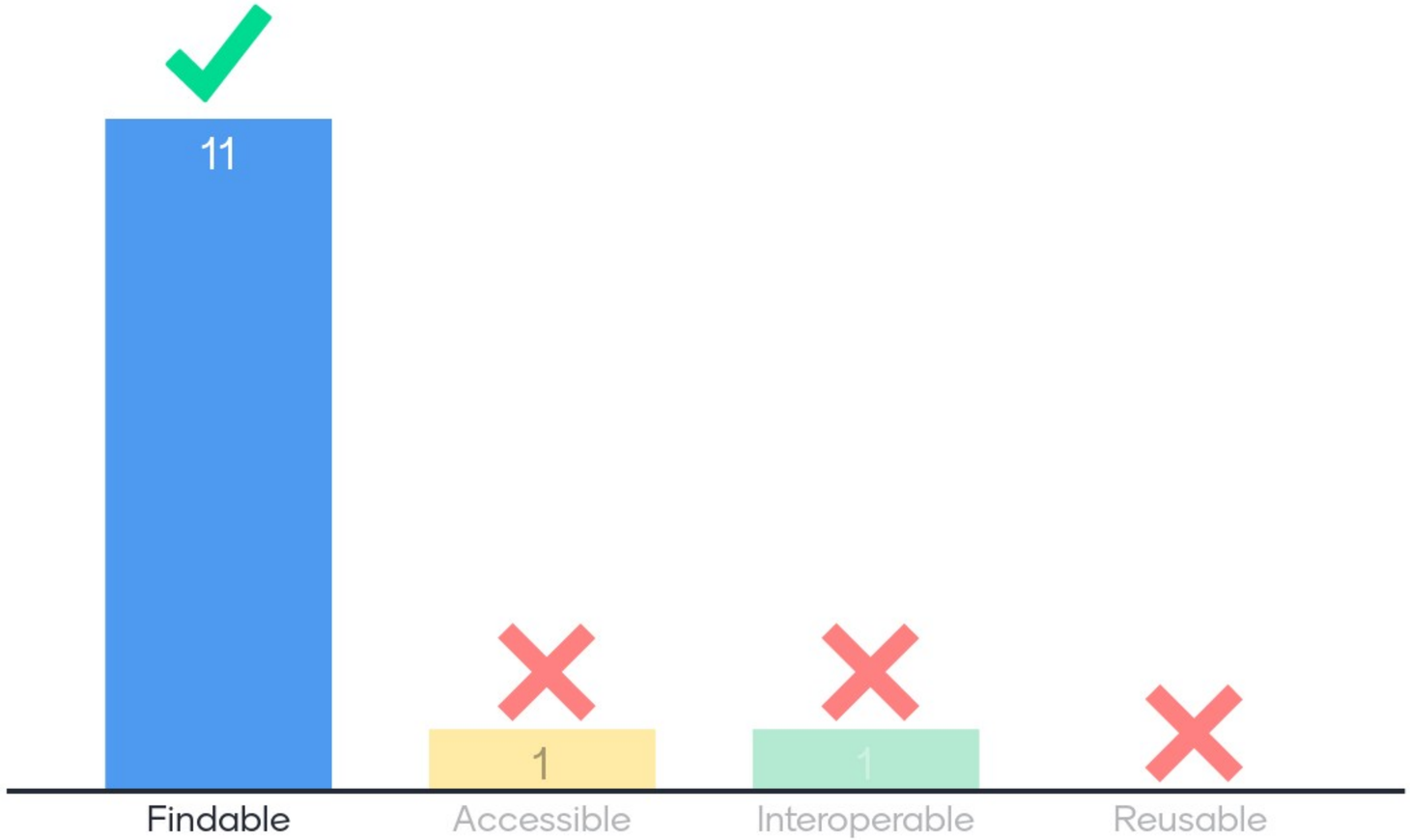
Outcomes so far

FAIR4RS community and ways to get involved

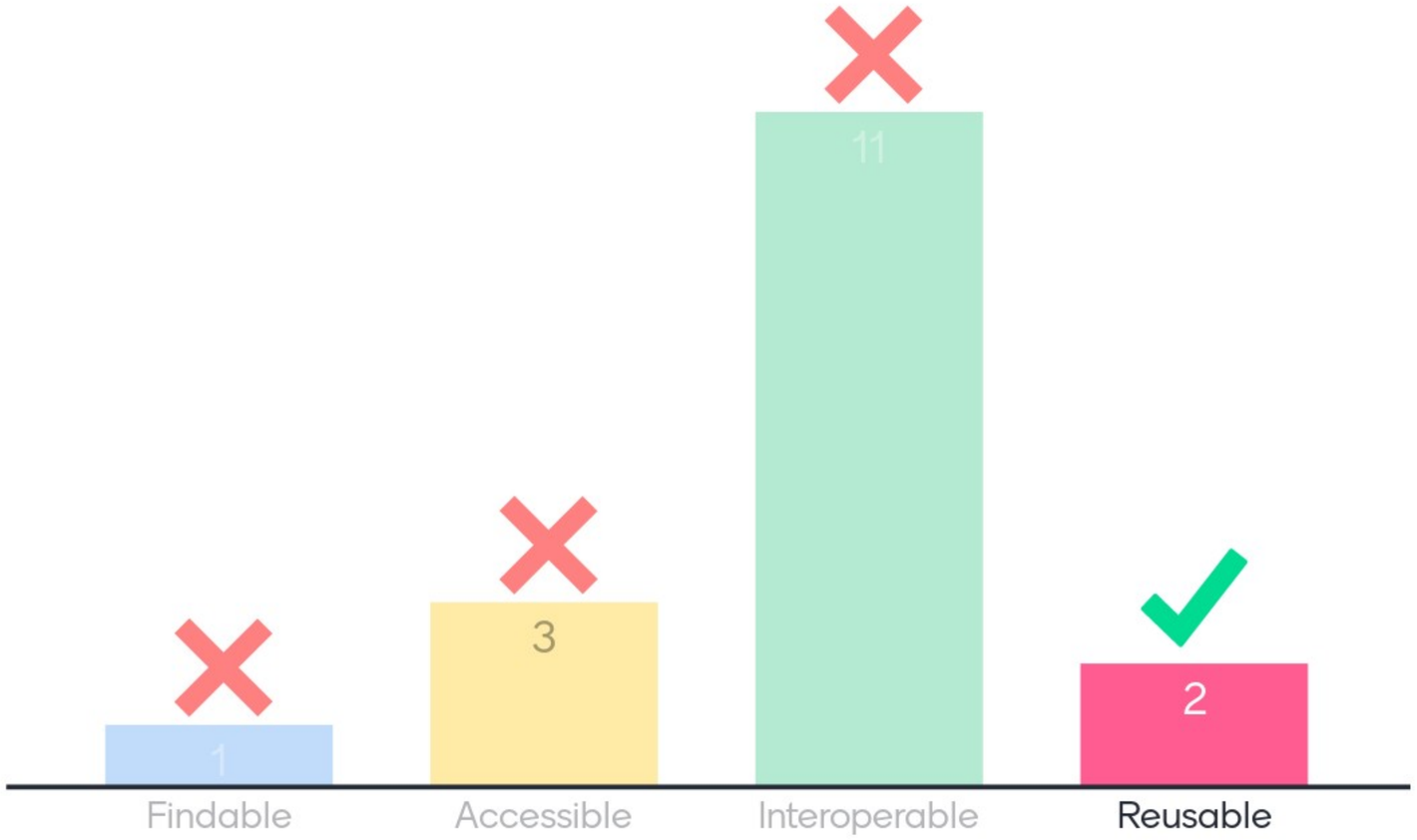
Are you familiar with the FAIR data principles?



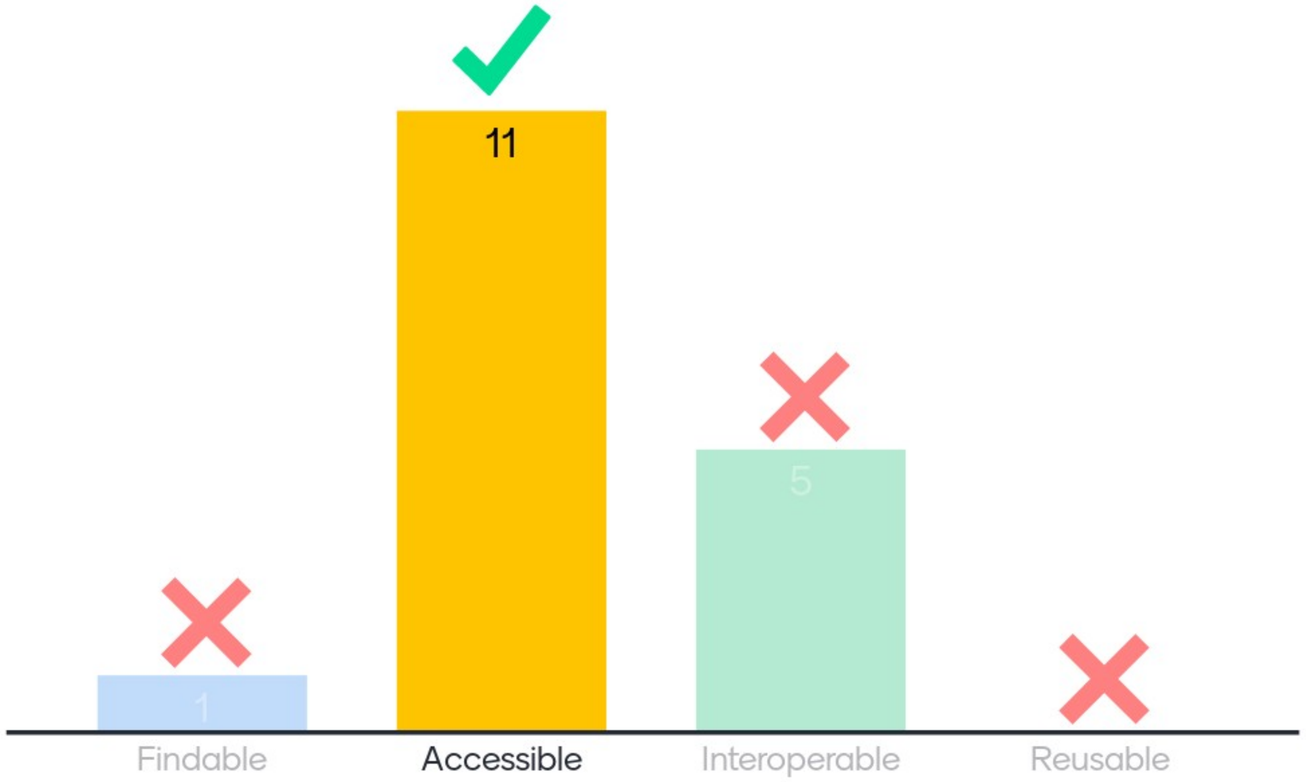
"(Meta)data are assigned a globally unique and persistent identifier." Which FAIR principle?



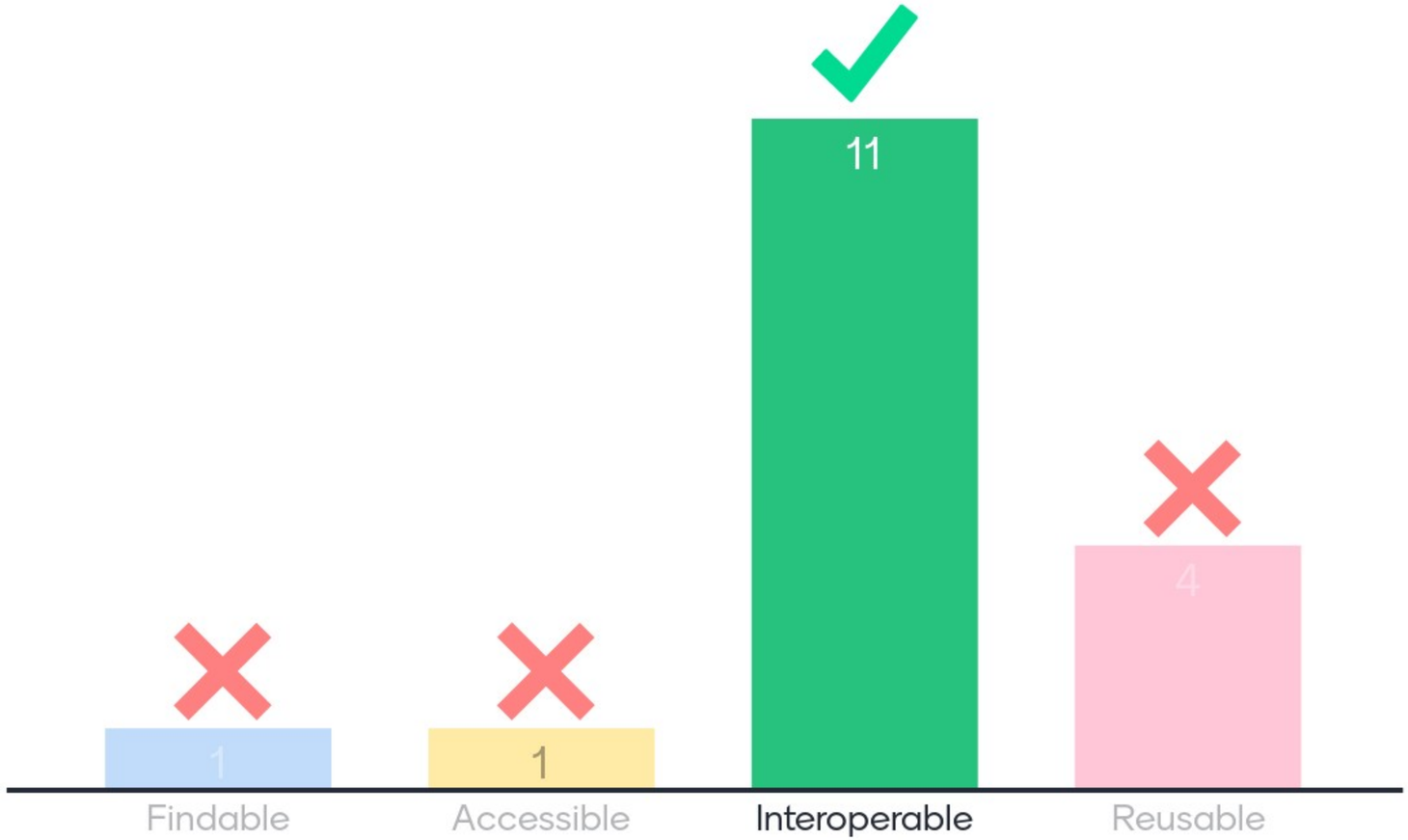
"(Meta)data meet domain-relevant community standards." Which FAIR principle?



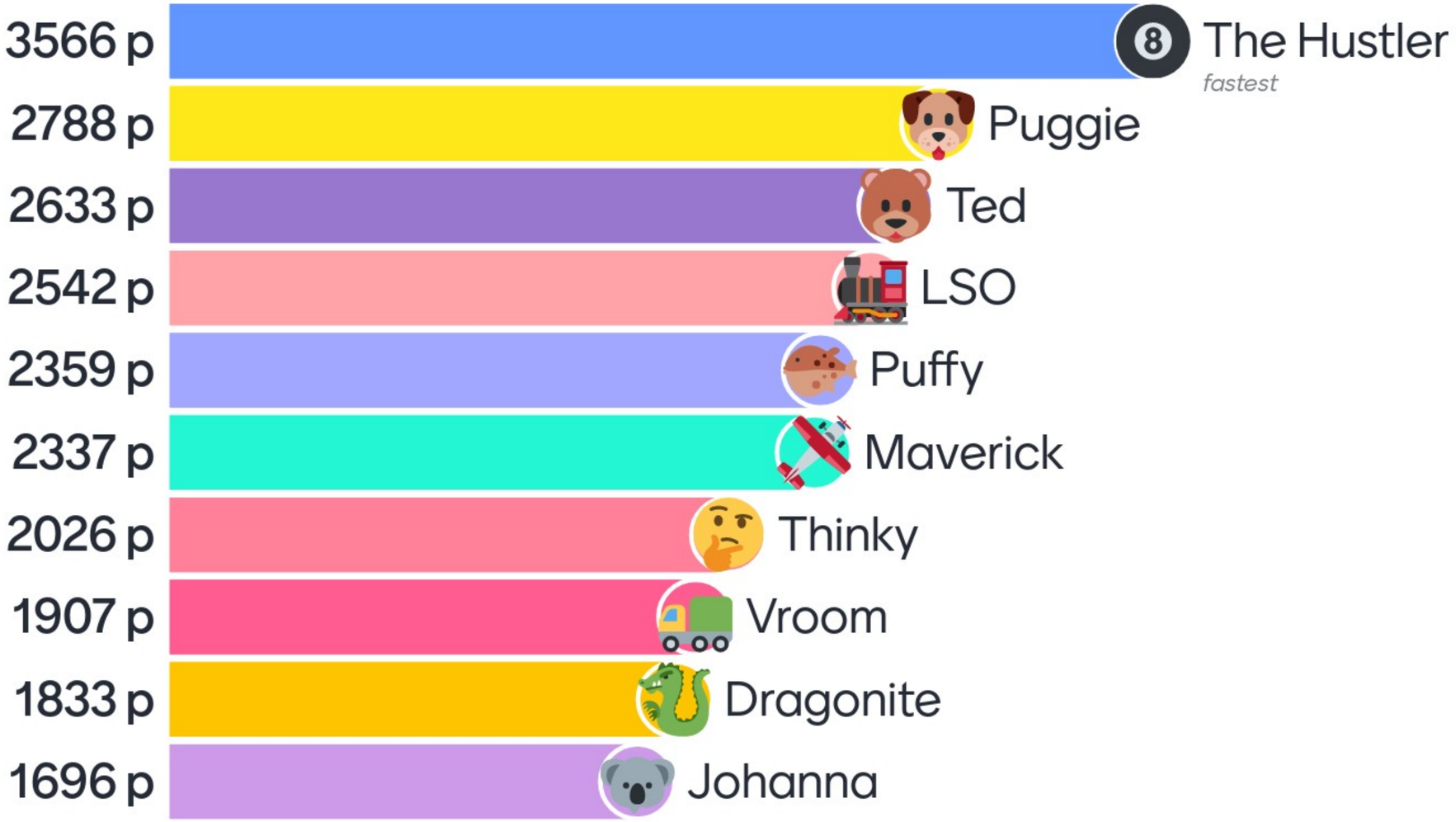
"(Meta)data are retrievable by their identifier using a standardised communications protocol." Which FAIR principle?



"(Meta)data include qualified references to other (meta)data." Which FAIR principle?



Leaderboard



FAIR

Findable
Accessible
Interoperable
Reusable

2016: "**The FAIR guiding principles for scientific data management and stewardship**" (Wilkinson et al., doi:10.1038/sdata.2016.18)

2018: "Central to the realization of FAIR are **FAIR Digital Objects, which may represent data, software or other research resources**." (European Commission)

But: **How do the FAIR Principles relate to software?**

Mismatch between the broad intentions of the **4 foundational FAIR principles** and how the **15 FAIR Guiding Principles** are communicated and perceived.

FAIR and software

An ongoing discussion...

Milestone:
"Towards FAIR Principles for Research Software", Lamprecht et al., Data Science, Vol.3, Iss.1, pp. 37–59, 2020, https://doi.org/10.3233/DS-190026

Most viewed article in Data Science in the first half of 2020!

Let's get into some of the thoughts and ideas in there.

What is data?

the information you want to analyze for insight

Quantified and searchable information

Anything you can perform analyze on

information collected from participants or files of information

facts and statistics, metadata

information

input for some particular task

The factual information needed to examine and answer a question.

An artifact of an event, thing, or nugget of information.

What is data?

Any continuous or categorical information describing something.

Information artifact from observations

tables

compilation of data points

Qualitative or quantitative digital information

collected knowledge

software is a program that is used to perform a certain task

What is software?

program run by a computer

A system design for a specific task

Instructions for computers to perform specific tasks, including interacting with humans, data, and other computers.

mechanism(s) for manipulating or visualizing data

a computer program used to collect information or perform specific tasks

something that tells a computer what to do

digital instructions for machines to perform tasks

What do data and software have in common?

"Machine readability"

software collects and manipulates the data collected

Data is organizes information in a library for a software system to access the information logically

Software is an artifact of data (code) and can generated data

Software can operate on data (not a necessity) and data can be produced by software

software can be broken down into data (points)

Hopefully machine actionability. They are stored in bytes, and maybe qubits. someday

software can be used to analyze data, create data

Versioning

What do data and software have in common?

both should be verified and reproducible

What makes data and software different?

Software can modify other things, whereas data on its own does not.

One performs the actions, the other is acted upon

data is interpreted by software

Software is using some sort of data to perform some task

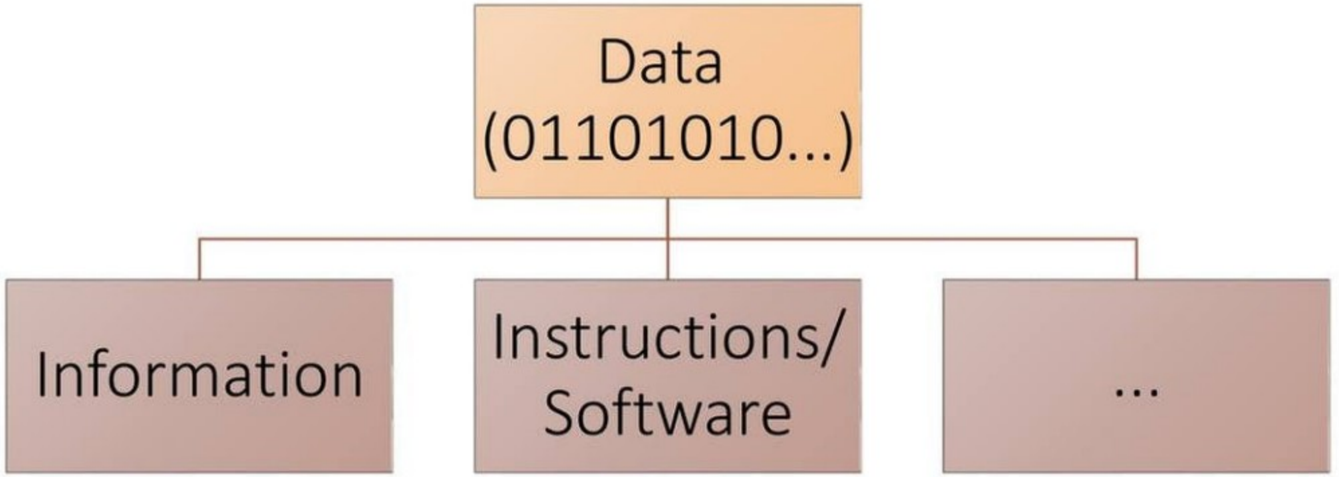
I think of software more like a verb - it is for doing and data more like a noun - what is done to

Uses for data are different from uses for software

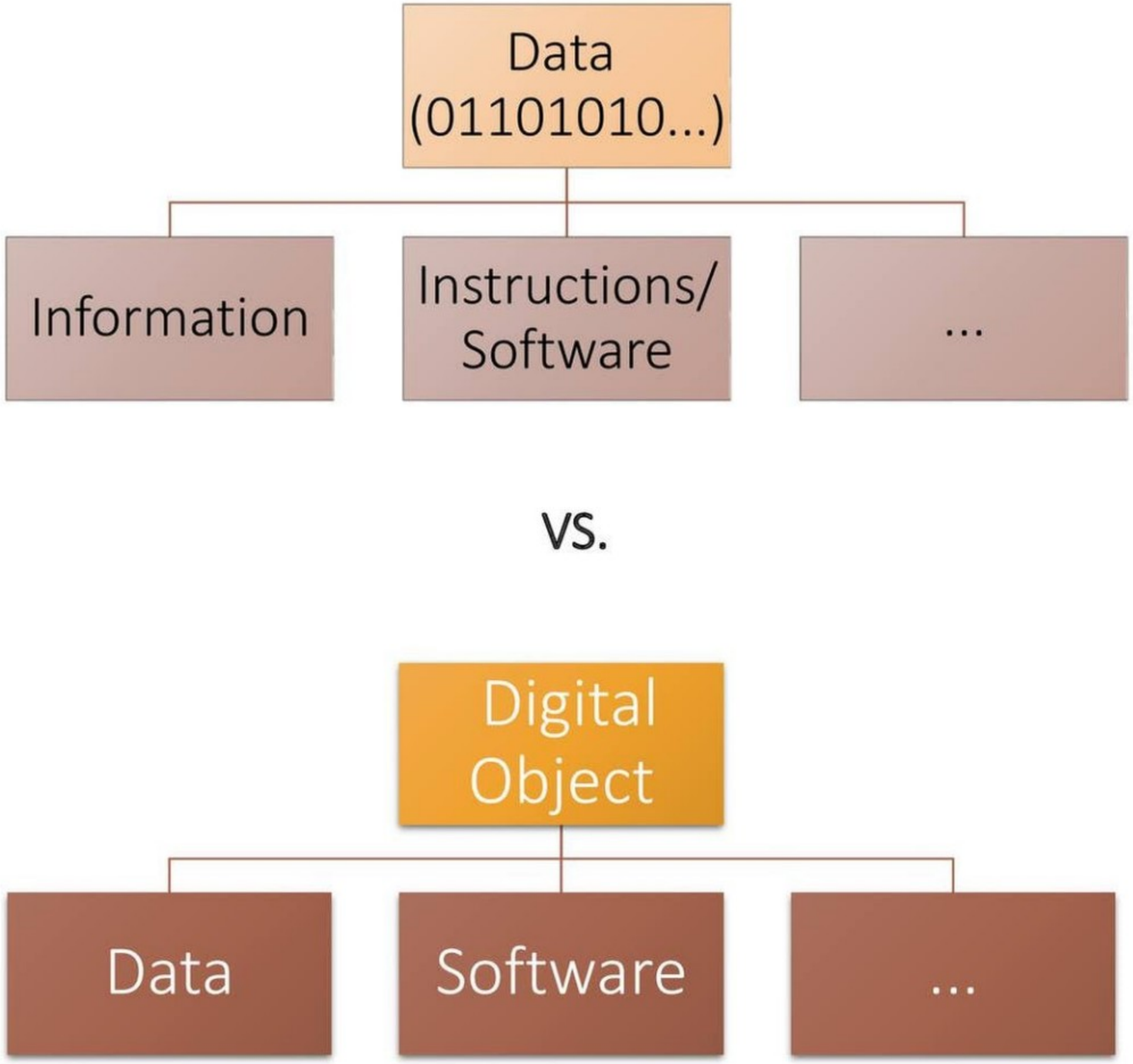
Contributions over time (data may be finite, software may be developed over time)

Software needs a lot of dependent packages however data might stay by iyself

Software is (not) data



Software is (not) data



Research software

Research software is "**software that is used to generate, process or analyze results that you intend to appear in a publication**" (Hettrick et al., 2014).

Many forms.

Many purposes.

Many distribution channels.

Traditionally, often created as Free and/or Open Source Software (FOSS).

FAIR and FOSS

Clear overlap of objectives, but not the same.

FOSS: Open source code, open licenses.

FAIR: Open data not a requirement.

- Due to, e.g., privacy and sensitivity concerns with patients' health records.
- Not in the same way valid for research software.
- There is even a demand to make methods available!

Should FAIR software require FOSS?

Ongoing discussion ...

What do you think? Should FAIR require software to have an open license? Why, or why not?

It's not always possible. So no, it's unrealistic.

Could be a best practice vs a requirement

yes, ideally i think it should require this

I think it should be an ideal, but not required

Open in the sense that the license is machine readable. But yes, the license machine readability will allow us to understand how to use the software.

No, just as sensitive data is not open to all. But it is discoverable and described.

No; there are often legitimate reasons for "closed" licenses.

No. Some software may be used for classified research or involve data that may have PHI (i.e. security issues).

Yes, it allows others to make improvements.

What do you think? Should FAIR require software to have an open license? Why, or why not?

Yes, whenever possible if the software was used for publicly funded science. It makes the science reproducible.

We might need a new acronym. FAIR also doesn't include reproducibility. Data quality is a huge oversight you raise.

if it's not open then you can't make the underlining data FAIR

Yes - though we would need to have lots of different criteria

Yes, ideally people can use for free

FAIR has other things it misses like reproducibility. Quality is a big oversight you bring up.

FAIR and software quality

Software quality is a major concern in RSE.

Can FAIR meet the expectations?

Distinguish between **form** and **function** of software:

- Quality of the **form** of software can be covered by FAIR (code quality, maintainability).
- Quality of the **functionality** of software goes beyond FAIR (functional correctness, software security, computational efficiency).

Should FAIR also take content quality into account? Why (not)?

Some researchers seem to think that FAIR and data quality (or S/W quality) must be considered together. I don't think so

I would like there to be a central way for people to comment/ rate quality as part of communities of practice to determine if it's something to review for lessons learned vs something to reuse

Quality will be determined by the publications.

No. Quality is determined by context.

Yein, potentially connected with other FAIR letters, R, or extended letters.

Yes, though I wonder how to make all the criteria for different disciplines or data types

Yes it should be as if data is of not quality then results produced from such data is also questionable

Quality is important, but it is something different, and not measurable in the way that FAIR is.

To the extent that quality can be judged / described objectively, would be helpful, but not necessarily required.

Should FAIR also take content quality into account? Why (not)?

FAIR misses reproducibility also. You are right quality is another big oversight or is out of scope. One can only pack so much into an acronym.

Will there be peer review for quality?

FAIR applied to research software

	FAIR for data	FAIR for software	Operation
F1	(Meta)data are assigned a globally unique and persistent identifier.	Software and its associated metadata have a global, unique and persistent identifier for each released version.	Rephrased
F2	Data are described with rich metadata.	Software is described with rich metadata.	Rephrased
F3	Metadata clearly and explicitly include the identifier of the data it describes.	Metadata clearly and explicitly include identifiers for all the versions of the software it describes.	Rephrased and extended
F4	(Meta)data are registered or indexed in a searchable resource.	Software and its associated metadata are included in a searchable software registry.	Rephrased

	FAIR for data	FAIR for software	Operation
A1	(Meta)data are retrievable by their identifier using a standardized communications protocol.	Software and its associated metadata are accessible by their identifier using a standardized communications protocol.	Rephrased
A1.1	The protocol is open, free, and universally implementable.	The protocol is open, free, and universally implementable.	Remain the same
A1.2	The protocol allows for an authentication and authorization procedure, where necessary.	The protocol allows for an authentication and authorization procedure, where necessary.	Remain the same
A2	Metadata are accessible, even when the data are no longer available.	Software metadata are accessible, even when the software is no longer available.	Rephrased

	FAIR for data	FAIR for software	Operation
I1	(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	Software and its associated metadata use a formal, accessible, shared and broadly applicable language to facilitate machine readability and data exchange.	Rephrased and extended
I2	(Meta)data use vocabularies that follow FAIR principles.	–	Reinterpreted, extended and split
I2S.1	–	Software and its associated metadata are formally described using controlled vocabularies that follow the FAIR principles.	Reinterpreted, extended and split
I2S.2	–	Software use and produce data in types and formats that are formally described using controlled vocabularies that follow the FAIR principles.	Reinterpreted, extended and split
I3	(Meta)data include qualified references to other (meta)data.	–	Discarded
I4S	–	Software dependencies are documented and mechanisms to access them exist.	Newly proposed

	FAIR for data	FAIR for software	Operation
R1	(Meta)data are richly described with a plurality of accurate and relevant attributes.	Software and its associated metadata are richly described with a plurality of accurate and relevant attributes.	Rephrased
R1.1	(Meta)data are released with a clear and accessible data usage license.	Software and its associated metadata have independent, clear and accessible usage licenses compatible with the software dependencies.	Rephrased and extended
R1.2	(Meta)data are associated with detailed provenance.	Software metadata include detailed provenance, detail level should be community agreed.	Rephrased
R1.3	(Meta)data meet domain-relevant community standards.	Software metadata and documentation meet domain-relevant community standards.	Rephrased

Incomplete FAIR software reading list

"Top 10 FAIR Data & Software Things", Martinez et al., Zenodo, 2019, https://doi.org/10.5281/zenodo.3409968

"Towards FAIR Principles for Research Software", Lamprecht et al., Data Science, 2020, https://doi.org/10.3233/DS-190026

"FAIR Computational Workflows", Goble et al., Data Intelligence, 2020, https://doi.org/10.1162/dint_a_00033

"From FAIR research data toward FAIR and open research software", Hasselbring et al., Information Technology, 2020, https://doi.org/10.1515/itit-2019-0040

FAIRsFAIR "Assessment report on 'FAIRness of software'", Gruenpeter et al., Zenodo, 2020, https://doi.org/10.5281/zenodo.4095092

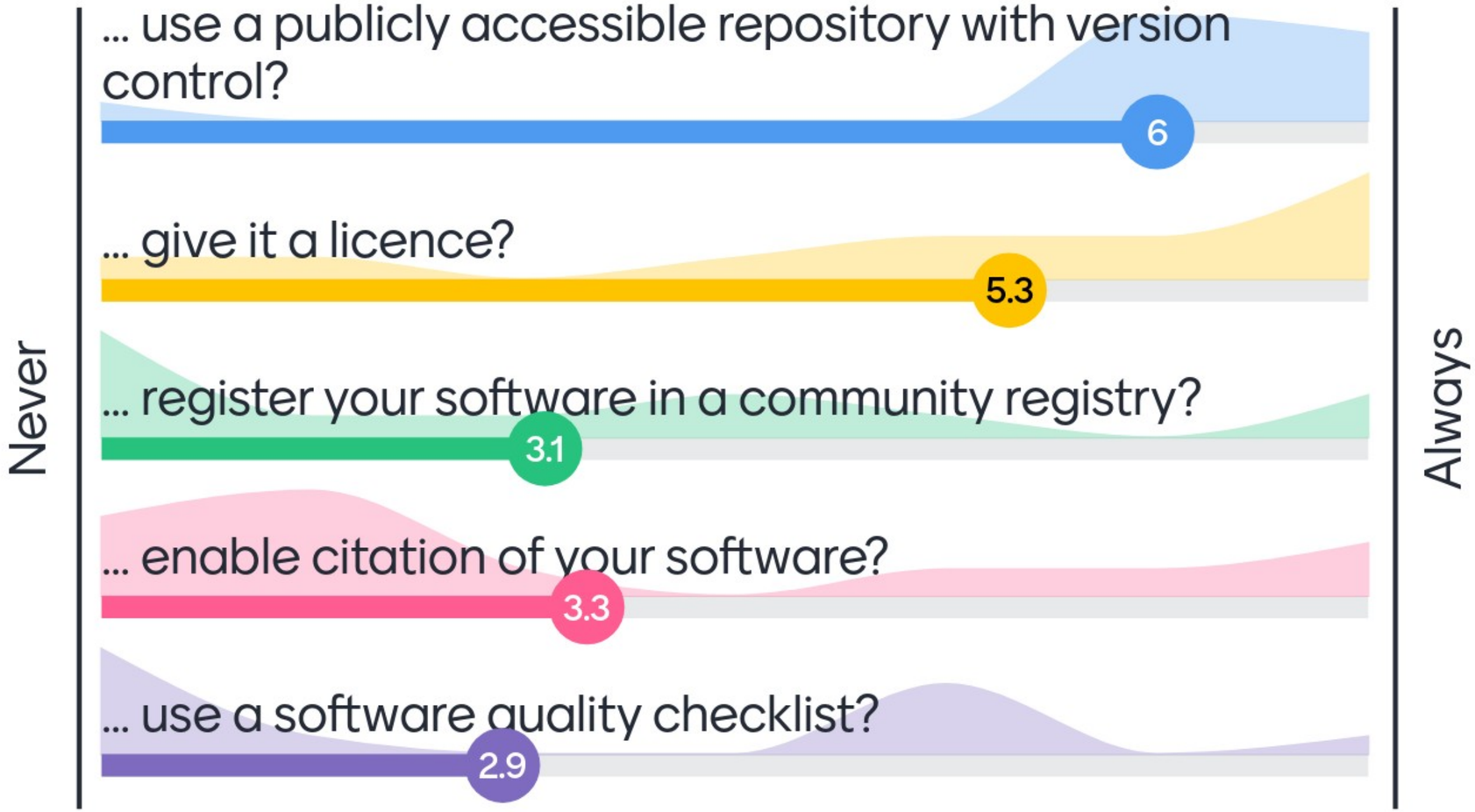
Practical FAIR

"Five Recommendations for FAIR Software",
Netherlands eScience Center, https://fair-software.eu/

1. Use a publicly accessible repository with version control.
2. Add a license.
3. Register your code in a community registry.
4. Enable citation of the software.
5. Use a software quality checklist.

(Your organization can endorse this!)

How about your software? Do you...



Get involved

RDA FAIR for Research Software (FAIR4RS) WG
https://www.rd-alliance.org/groups/fair-research-software-fair4rs-wg

Jointly convened as an RDA Working Group, FORCE11 Working Group, and Research Software Alliance (ReSA) Taskforce.

Four subgroups:

1. **A fresh look at FAIR for Research Software** (lead: Dan Katz)
2. **FAIR work in other contexts** (lead: Mateusz Kuzak)
3. **Definition of research software** (lead: Morane Gruenpeter)
4. **Review of new research related to FAIR software** (lead: Neil Chue Hong)

Next events

Public FAIR4RS Town Hall Meeting
2 February 2021, 20:00 to 21:00 UTC
https://www.rd-alliance.org/group/working-and-interest-group-chairs-fair-research-software-fair4rs-wg/event/public-fair4rs-town

“How FAIR are you” Webinar Series and Hackathon
21 January to 30 April, 2021
https://www.cineca-project.eu/news-events-all/how-fair-are-you-webinar-series-and-hackathon

SSI Collaborations Workshop 2021
FAIR Research Software as one of three themes
30 March to 1 April 2021
https://www.software.ac.uk/cw21

Acknowledgments

To the numerous people who contributed to the discussions around FAIR research software at different occasions and keep the work going!

Thank you!