Monitoring Open Science Activities

Evgeny Bobrov

Berlin Institute of Health at Charité, QUEST Center

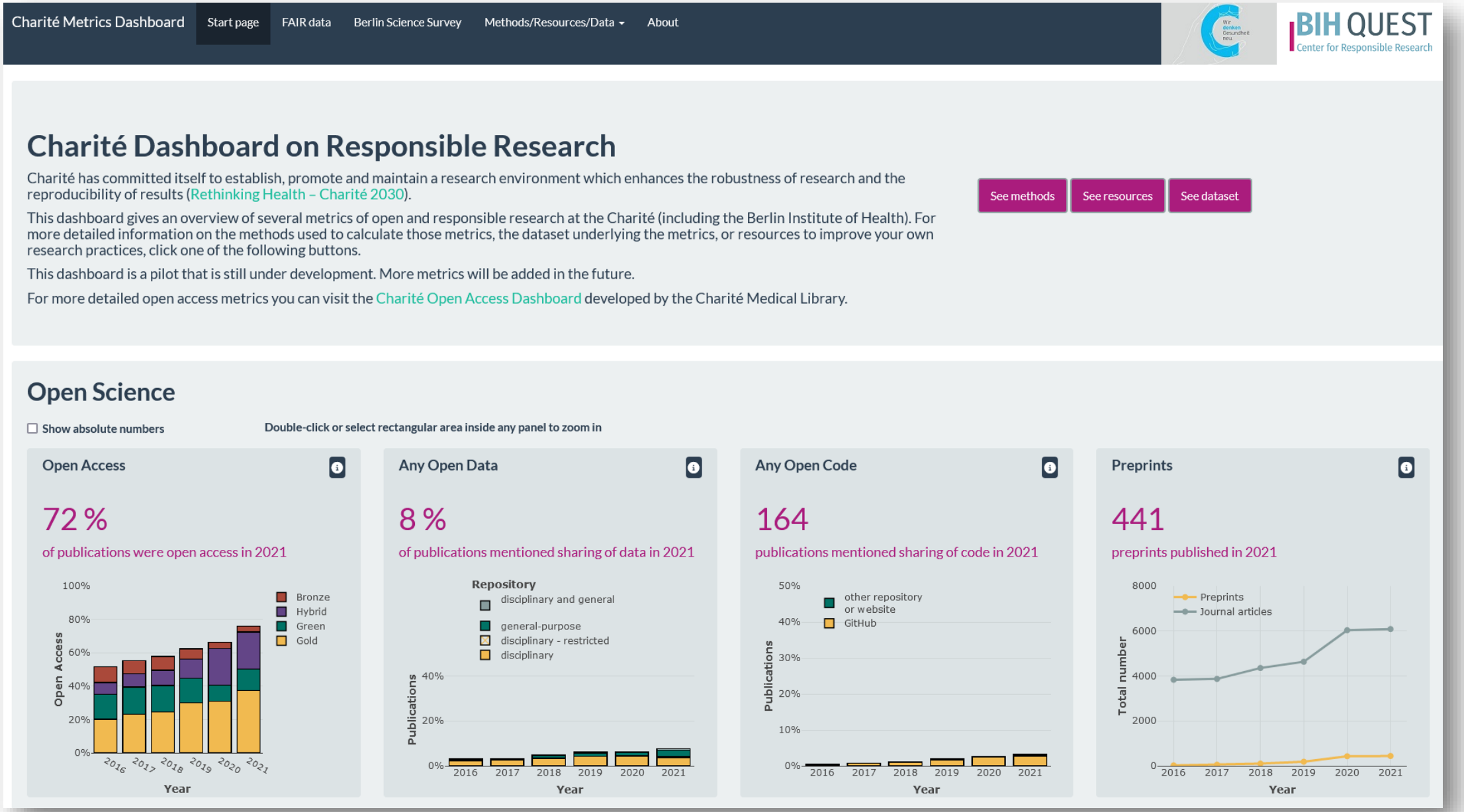
QUEST Seminar on Responsible Research

23.05.2023

BIH QUEST
Center for Responsible Research

BIH Berlin Institute of Health @Charité

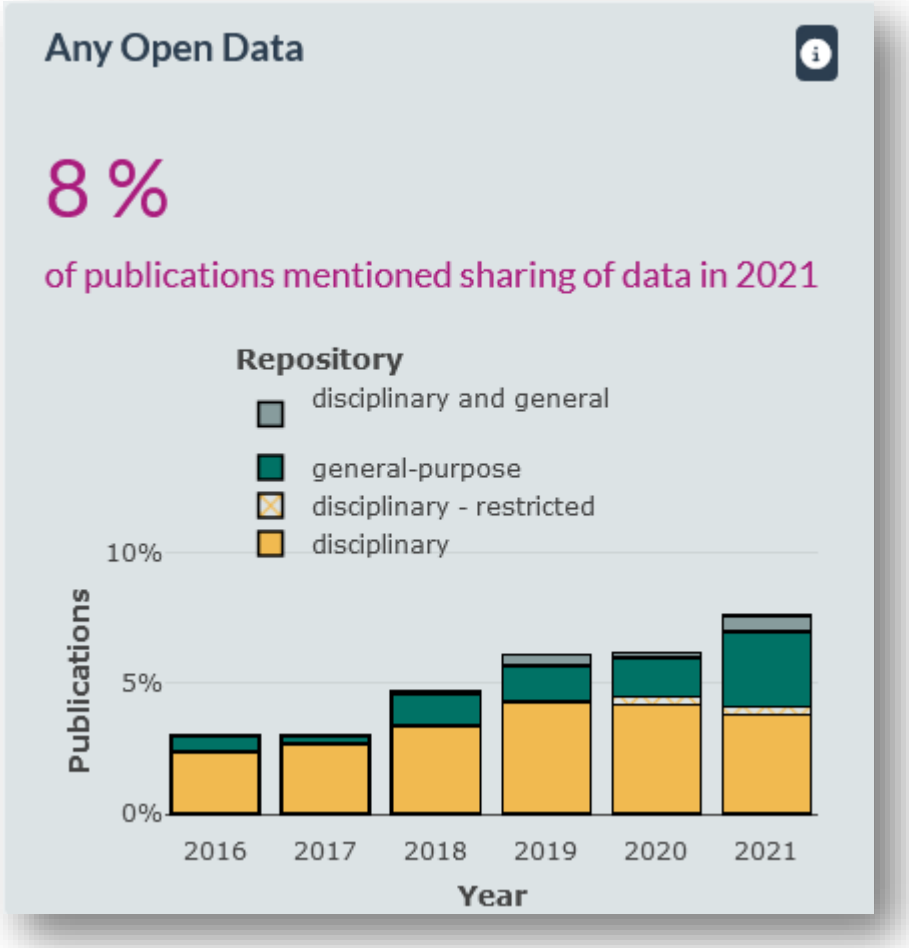
Aus Forschung wird Gesundheit



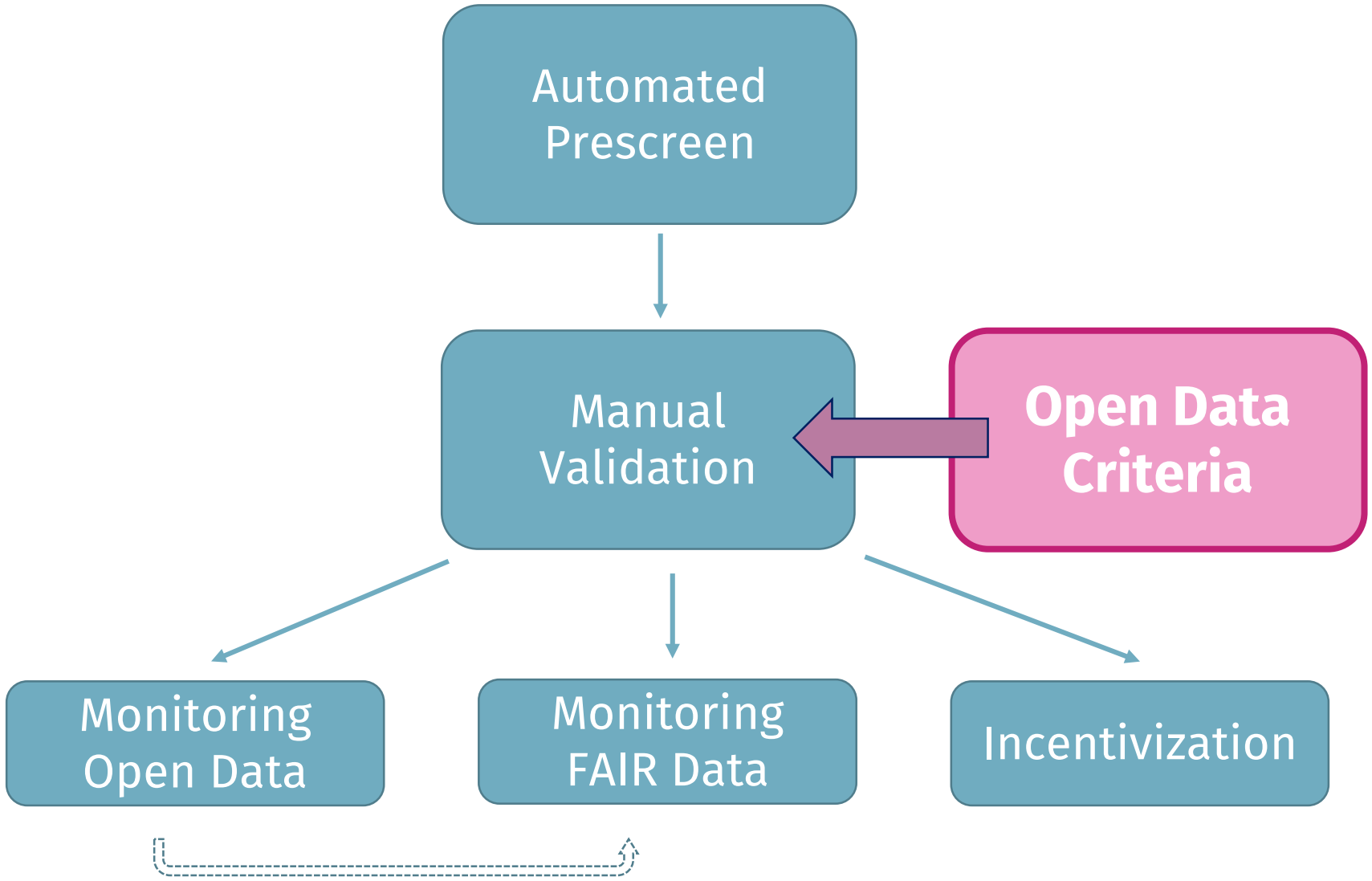
https://quest-dashboard.charite.de/#tabStart

BIH QUEST
Center for Responsible Research

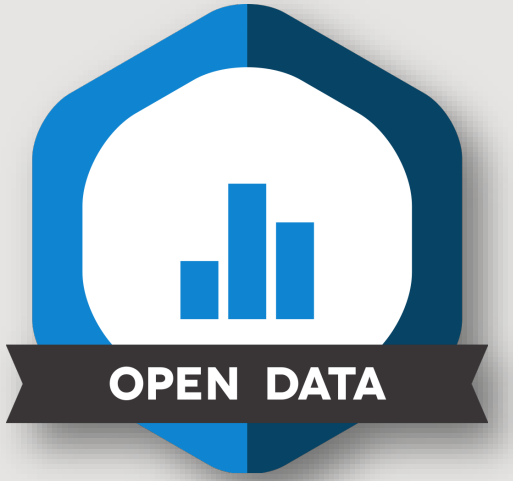
Charité Dashboard on Responsible Research



BIH QUEST
Center for Responsible Research

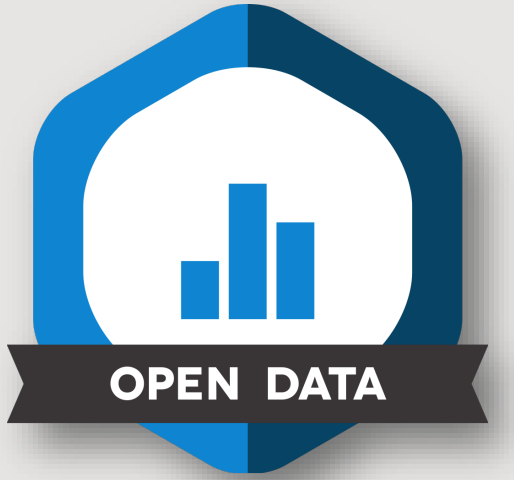


BIH QUEST
Center for Responsible Research



Open Data Definition

Open data is data that can be freely used, re-used and re-distributed by anyone - at most restricted by the obligation to name sources and “share-alike”.



Open Data Definition

Criterion: “*Research data have been made **freely accessible** by **researchers of the Charité***” Main question: What does free access mean in the given context? Definition: Free access is understood as making the dataset available in such a fashion that any human can access it online and use it for any purpose without revealing one’s own identity, as long as this purpose does not explicitly infringe on legitimate rights of authors (e.g., their right to demand citation) or study subjects (e.g. not to be re-identified). Also see demarcations under 1d. Limitation: Criteria for cases where restricted access is considered justified are excluded, and are discussed further below. Demarcation : **Access requirements. Our definition includes** cases, where an agreement has to be accepted, as long as this agreement does not restrict reusers in the type of study or analysis [note that this is not yet included in the published version of the criteria in Kip et al. (2022, June 16)]. Reasoning: for machine-readability, any kind of agreement would be a considerable hurdle, and such obstacles should be avoided. However, query of datasets by humans is so far the norm, and for humans it is not a large obstacle to accept certain terms, although it is not “open” in the full sense. **Our definition excludes** cases where a registration is necessary [this is also not yet included in Kip et al. (2022, June 16)]. Reasoning: the threshold is too high for researchers who would like to look at the data to assess its reuse value for them, especially given that most datasets do not have sufficient metadata to assess their utility purely on these grounds. **Data availability upon request. Excludes** data available upon request. Reasoning: such data are not freely available, but rather only upon the discretion of the authors. The literature shows consistently that data upon request are difficult and sometimes impossible to come by (Vines et al., 2014 Tedersoo et al, 2021; Gabelica et al., 2022). Minimally, it is difficult and one has to reveal information about oneself. **Licenses. Does not require** any specific license. Reasoning: Many repositories do not provide standardized licenses at all. Thus, assessing the openness of licensing and use terms is time-consuming and very difficult to standardize. We do not encourage non-commercial (NC) licenses, similarly to many others (see e.g. Margoni & Tsiavos, 2018), and we are aware that it comes at a cost to reuse to apply them (Hagedorn et al., 2011; The Open Data Institute, 2015; Matthews, 2022). However, they do sometimes occur in biomedical research, and given the availability of data for research purposes, we prefer to incentivize them at this point. In addition, researchers who attached a restrictive license should not be disadvantaged compared to those who did not attach a license at all. No derivatives (ND) licenses would be inappropriate for data reuse, but we have never come across such licensing (Charité Metrics Dashboard - Data Reusability, no date), and as this so extremely rare for datasets, we decided not to address this case. **Data authorship. Requires** clarity about authorship of data by authors of the article. Thus, the definition **excludes** “data from data collections of consortia (“data pools”), if it is unclear whether the authors themselves have contributed to the pool.” Reasoning: It is impossible otherwise to distinguish between data sharing and data reuse. However, if it is explicitly stated in the article that the authors contributed to the data pool, this is considered sufficient. **Does not require** data authorship as listed in repositories specifically by the Charité-affiliated authors of the respective article. Reasoning: Author lists of datasets are often unavailable or list only one person, and it cannot be determined, whether the person listed as data depositor is actually the (only) creator or collector of the data. “*The data can be **raw, primary, or secondary data** (e.g. from analyses of freely available datasets, meta-analyses, or health technology assessments); the **data would thus allow the analytical replication** (retracing of analysis steps) for at least a part of the study’s results; reporting of statistical values (means, standard deviations, p-values etc.) is not sufficient*” Please note, that in our experience this is the most difficult of the criteria to check, and one cannot fully avoid heuristics which depend on the specific subfield and are influenced by assessor experience. However, the application of below definition and demarcations substantially constrains the assessor degrees of freedom. Main question: When does data allow analytical replication? Definition: Analytical replication is here understood as the re-tracing of the analysis (quantitative or qualitative) from shared data to the results presented in an article. This re-tracing can be based on raw data as they were collected, or data which were in any way cleaned, normalized or otherwise processed. Demarcation: **“Source data” (tables with data points underlying figures). Includes** so-called “source data”, which list individual measurements underlying figures, and are quite common in biomedical journals. Reasoning: these are considered open data, as they provide additional information to the article and its figures, and could be pooled with other data, even though these data might already be highly derived. **Statistical values. Excludes** statistical values. Reasoning: statistical values do not constitute additional data compared to what is normally reported in articles, and do not allow computational reproduction, as they already constitute the result. Reuse is possible in a meta-analytic way, but is greatly reduced compared to individual observations. **Other outputs than data. Excludes** analysis scripts, computer programs, and other methods, materials, and protocols, even if their development was the goal of the research project and/or their presentation was the focus of the publication. If data has been collected and shared for development or validation, these can, however, be included. Reasoning: the assessment in question focuses on openly available data. We acknowledge the importance, even the essentiality of

.....

Automated Prescreen

ODDPub – Open Data Detection in Publications

- Open Source Tool in R, developed at QUEST
- Detection of keywords in articles → Open Data (and Open Code)
- Open Data detected as combination of terms regarding availability and deposit location, as well as partly identifiers
- Charité: ca. 5500 publications/year → ca. 770 Hits
- → Have to be screened manually

Riedel, N., Kip, M. and Bobrov, E., 2020. ODDPub – a Text-Mining Algorithm to Detect Data Sharing in Biomedical Publications. *Data Science Journal*, 19(1), p.42. DOI: http://doi.org/10.5334/dsj-2020-042

Manual validation

Screening of open data statements in Numbat

- Open source tool in PHP, originally developed for systematic reviews
- Adaptation for open data screening
- → Detailed protocol of the complete workflow in protocols.io

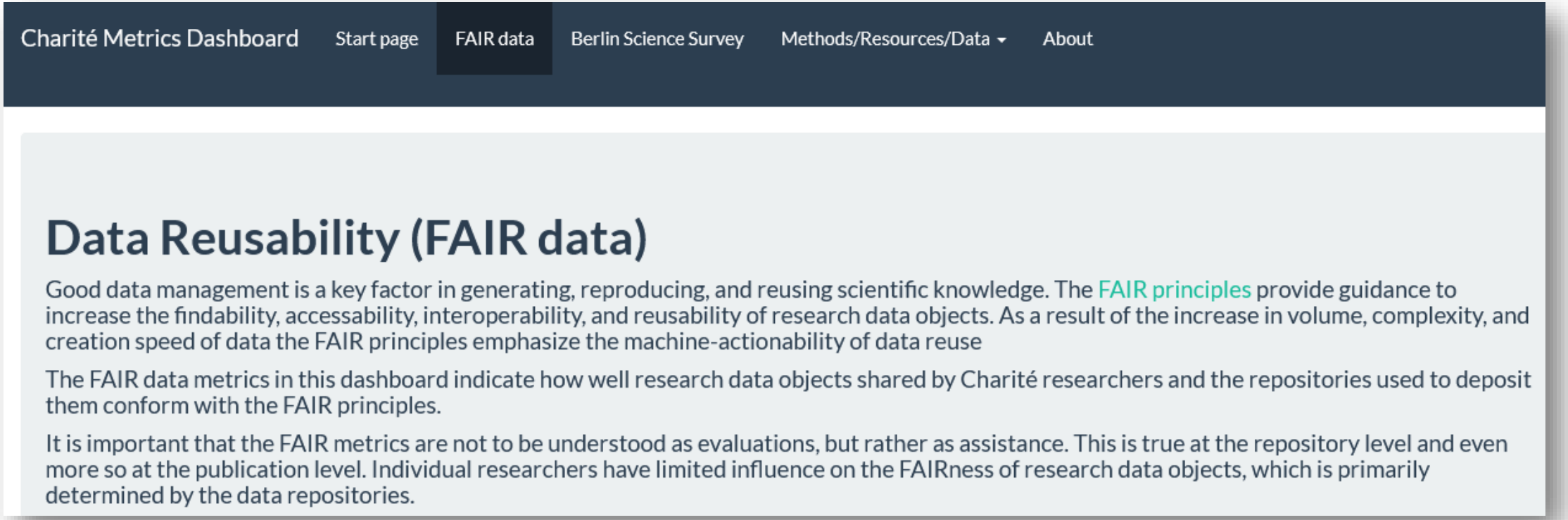
Iarkaeva A, Bobrov E, Taubitz J, Carlisle BC, Riedel N 2022. Semi-automated extraction of information on open datasets mentioned in articles. **protocols.io**https://dx.doi.org/10.17504/protocols.io.q26g74p39gwz/v1

BIH QUEST Center for Responsible Research

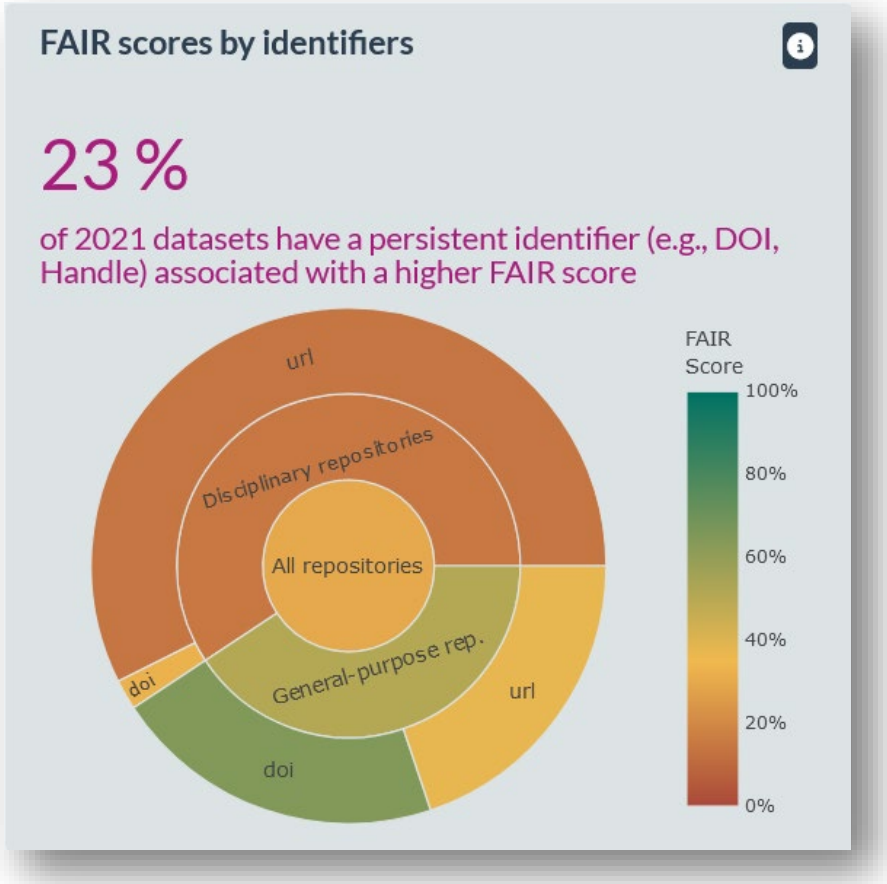
BUA Open Science Dashboards

- Project funded by Berlin University Alliance (BUA), Objective 3
- **Systematic Monitoring of open science practices** → Application of (semi-)automated tools + presentation of open science indicators in dashboards
- with Open Access Office Berlin, 10/2021-9/2023

FAIR Data Dashboard



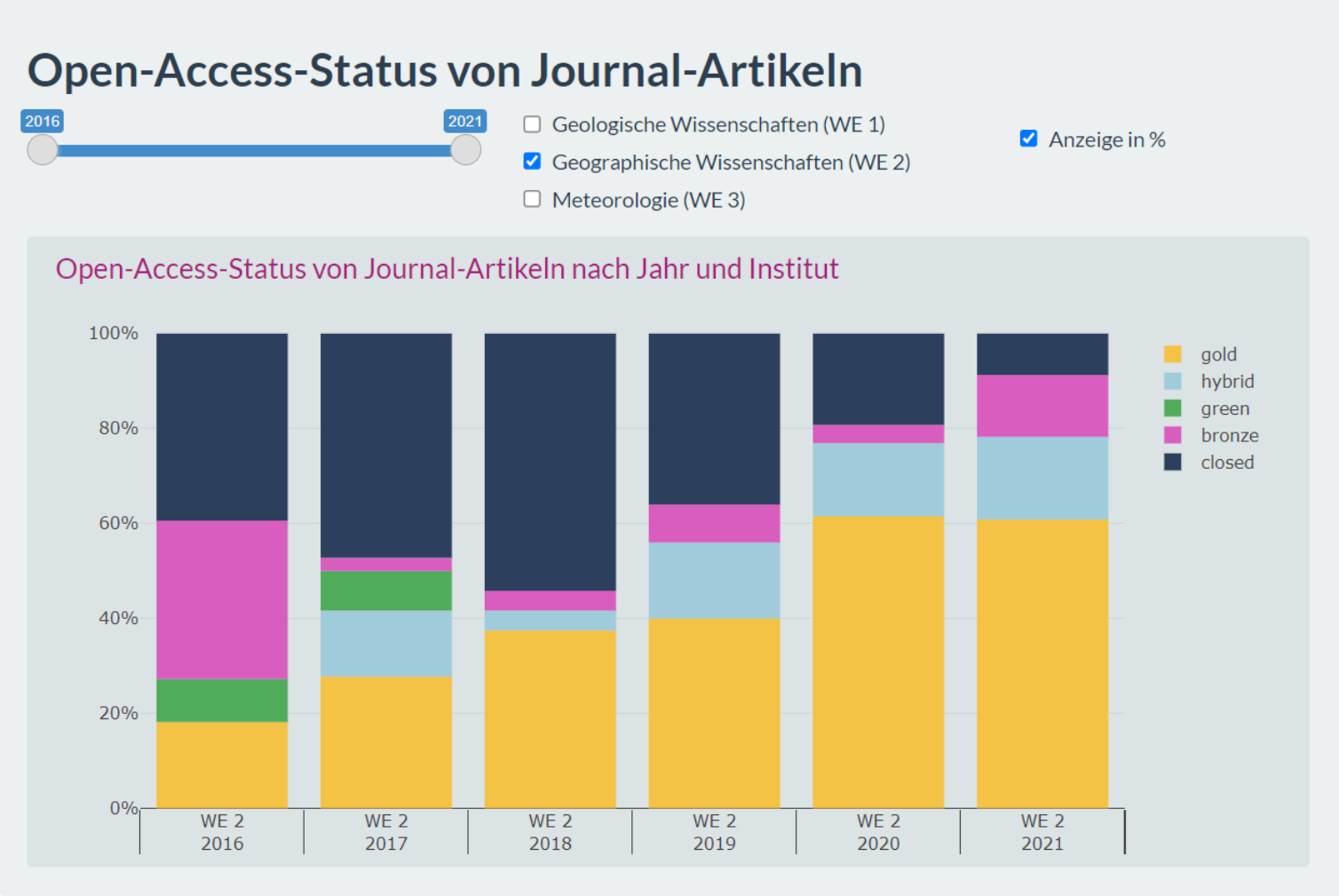
FAIR Data Dashboard



https://quest-dashboard.charite.de/#tabFAIR

BIH QUEST
Center for Responsible Research

Open Science Dashboard in other fields

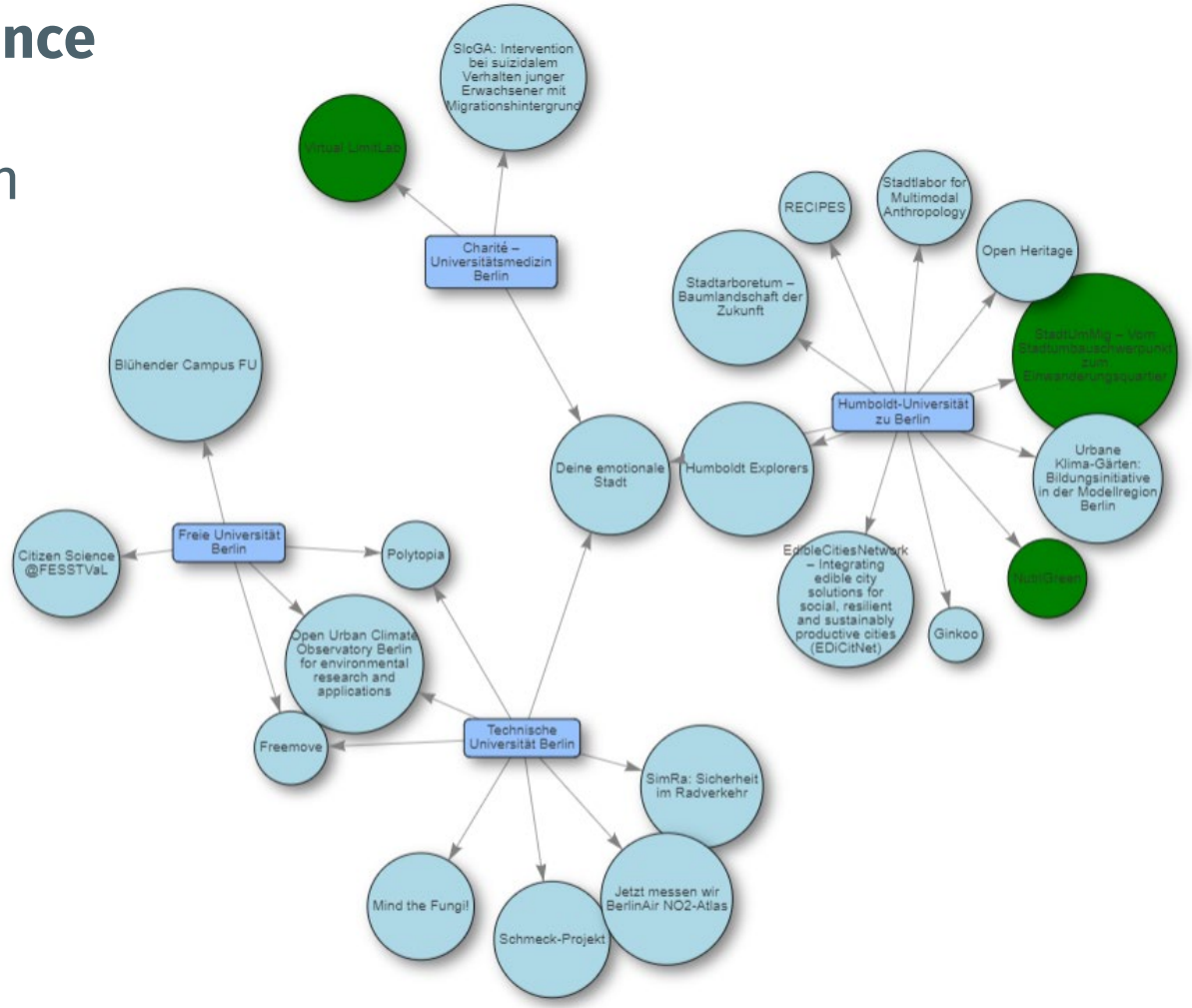


BIH QUEST
Center for Responsible Research

From Dashboards to „Magnifiers“

Alternative ways of making open practices visible

BUA Citizen Science Network Graph
→ Aimed at both citizens and scientists



Highlighting of **best practice examples**

BIH QUEST
Center for Responsible Research

Thanks to collaborators and team members!

Anastasiia Iarkaeva

Project Data Scientist

Alumnus:
Jan Taubitz

Project Data Scientist

Vladislav Nachev

BIH QUEST Data Scientist

Alumnus:
Nico Riedel

BIH QUEST Data Scientist

- *Project Team Open Science Dashboards @Open Access Office Berlin*
- *Miriam Kip (Open Data criteria)*
- *Benjamin Carlisle (Numbat)*
- *+ many others for validation of open data extraction*

Thank you!

quest.bihealth.org/

quest-dashboard.charite.de

BIH QUEST
Center for Responsible Research

BIH Berlin Institute of Health @Charité

Aus Forschung wird Gesundheit