ODDPub – detection of Open Data in biomedical publications

Nico Riedel

Helmholtz Open Science Webinars

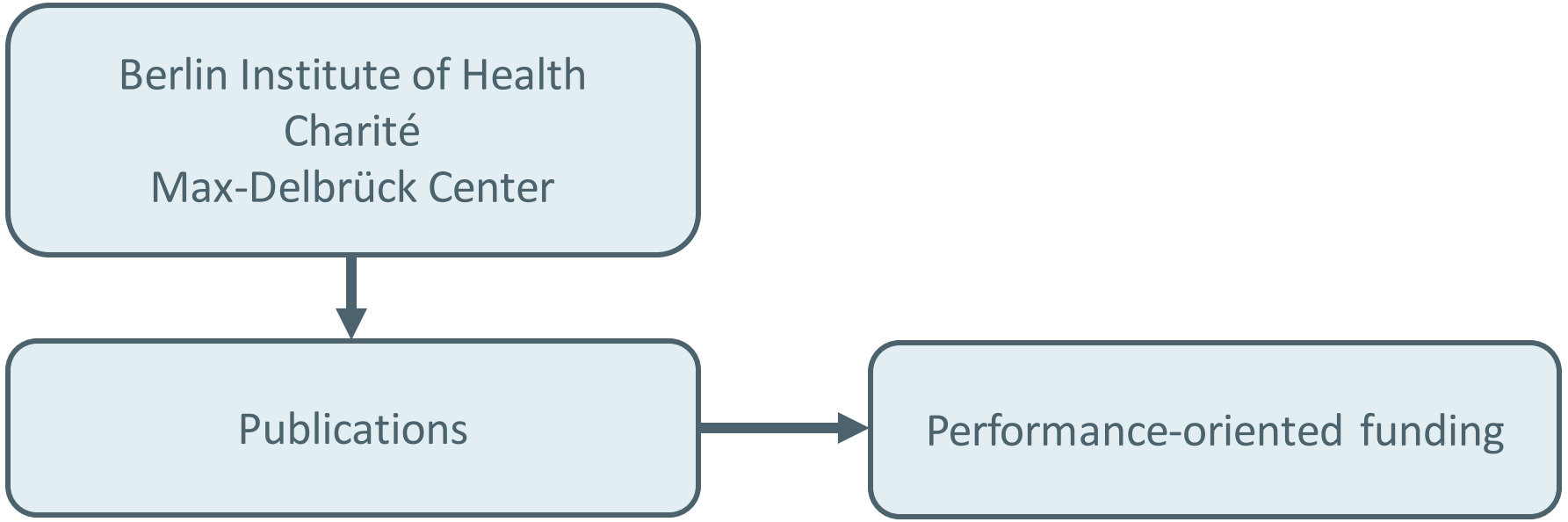
Webinar 54 – 23/27 April 2020

BIH QUEST
Transforming Biomedical Research

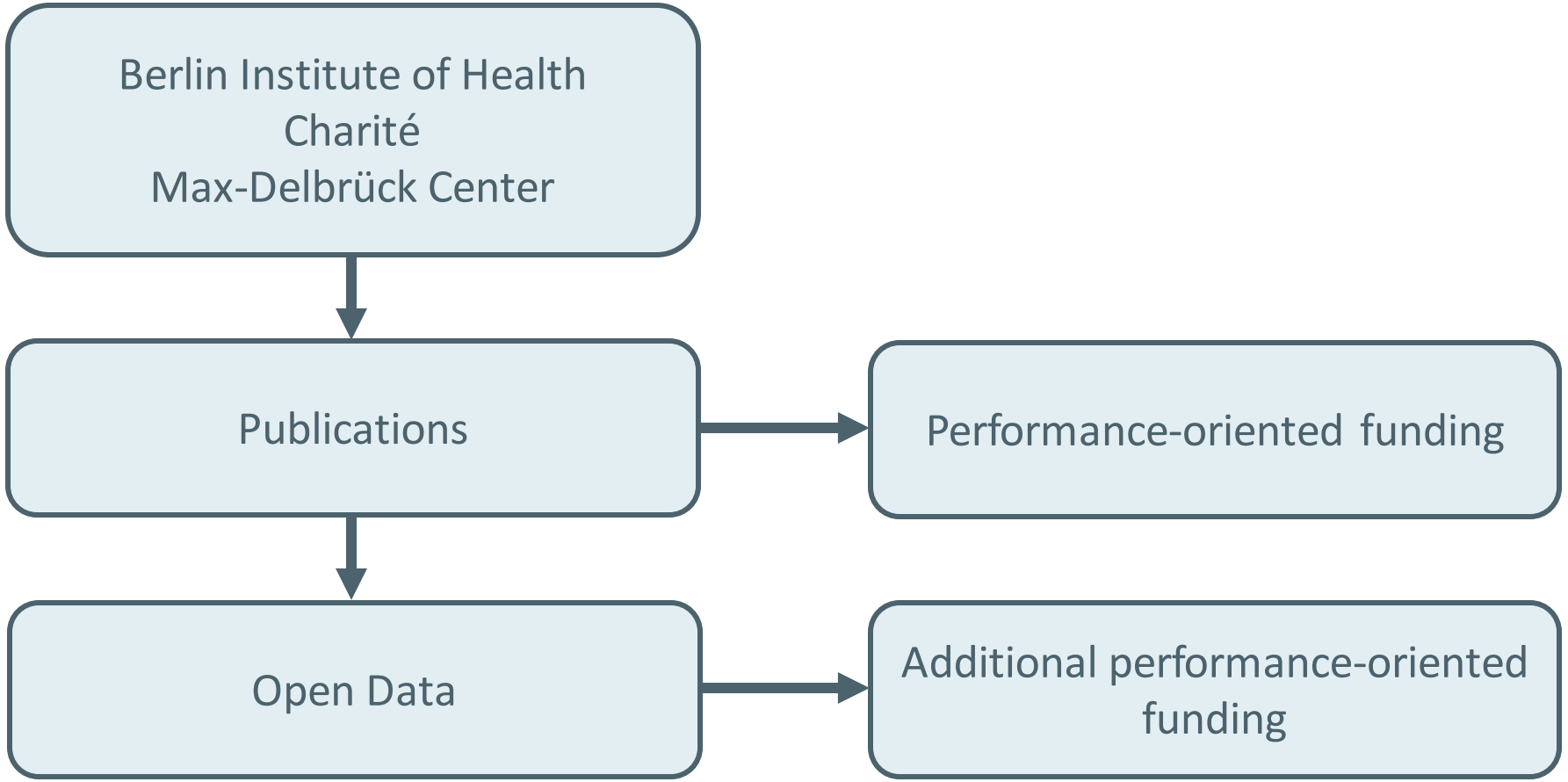
HELMHOLTZ
Open Science

BIH Berlin Institute of Health
Charité & MDC

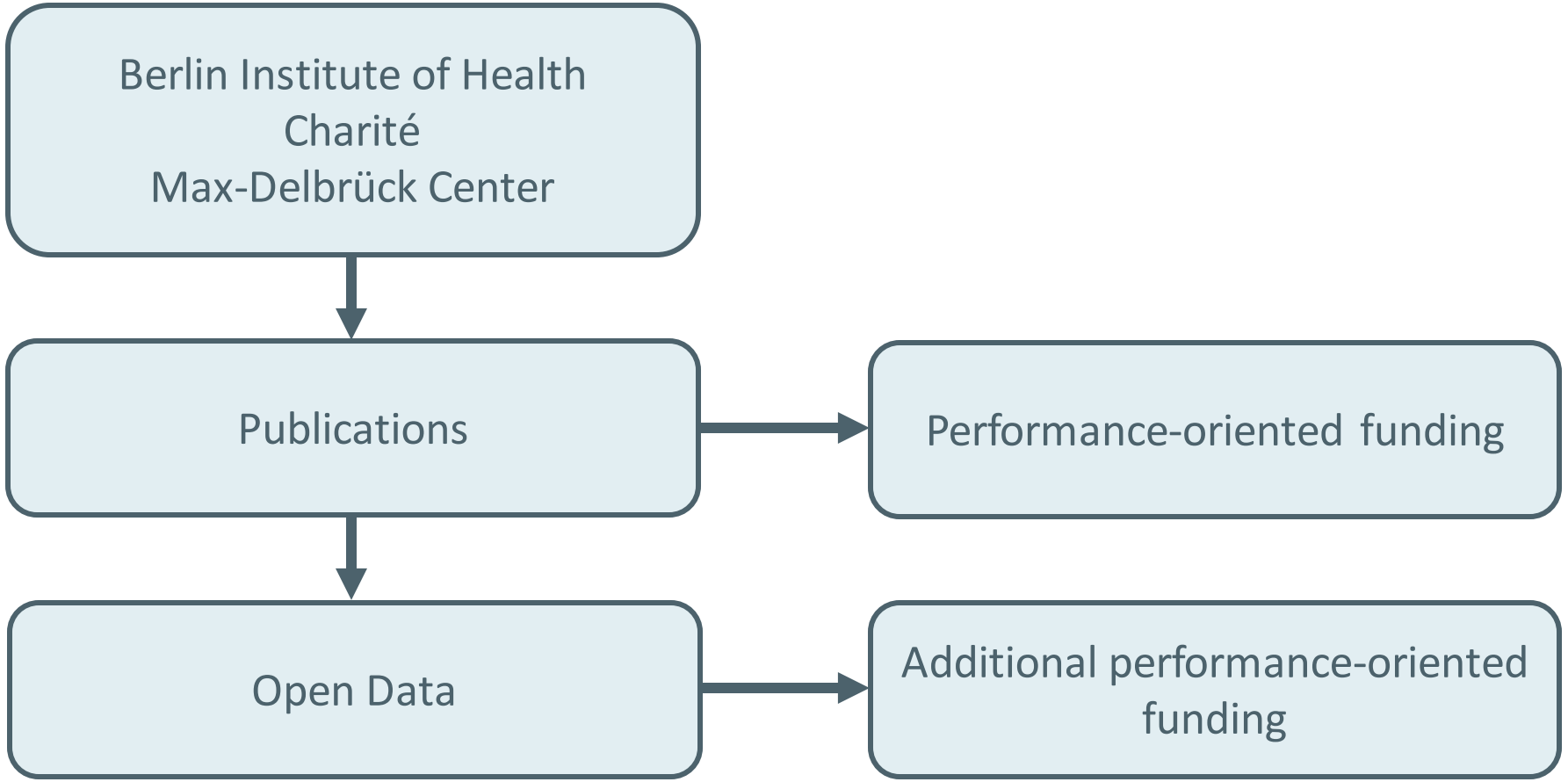
Aus Forschung wird Gesundheit



BIH QUEST
Transforming Biomedical Research



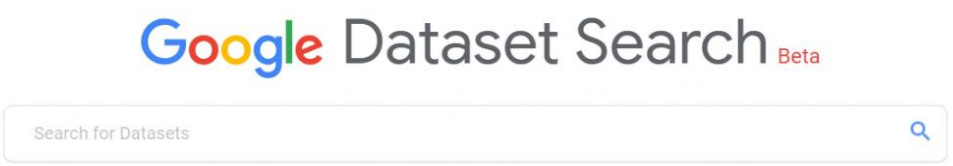
BIH QUEST
Transforming Biomedical Research



Problem: Which publications at our institution share their research data?

BIH QUEST
Transforming Biomedical Research

Where to look for Open Data?



Our solution

Develop own text-mining tool
to screen publications for Open Data

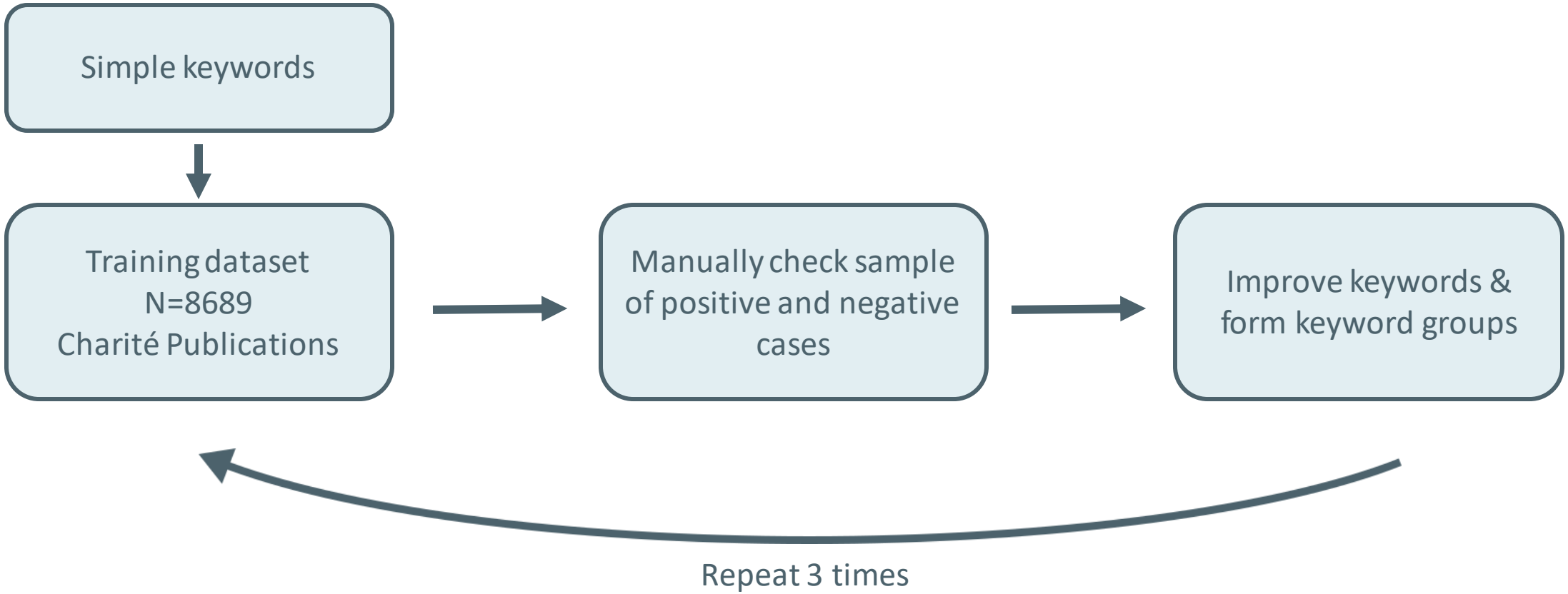
Open Data definition used

Different levels of stringency possible, we sat barrier rather low

Data have to be

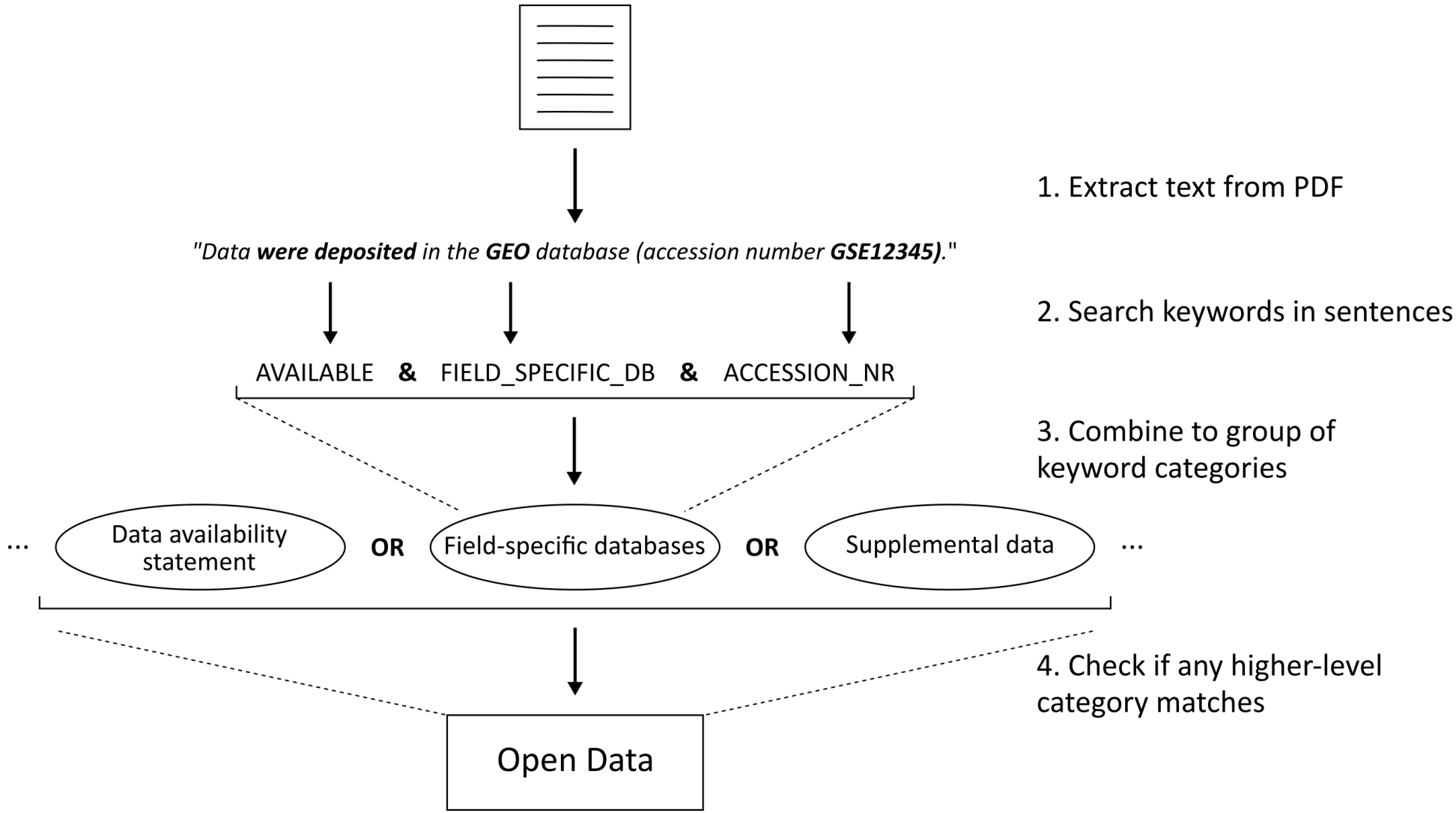
- findable through DOI, URL, database ID/accession code or a reference to a specific file in the supplement
- freely accessible to everyone (no registration or sharing upon request)
- not aggregated and allow reproducibility of at least part of the study
- 'machine-readable' (no PDF or image, but Excel or Word table ok)
- Created specifically for the publication

Algorithm development



BIH QUEST
Transforming Biomedical Research

Final structure



BIH QUEST Transforming Biomedical Research

Open Data categories

Open Data	Open Code
Field-specific repositories	Source code availability
General-purpose repositories	Supplementary source code
Supplemental Datasets	
Data availability statement	
Data journals	

Performance of ODDPub

ODDPub validation sample:

- random PubMed sample from 2018
- 792 downloaded journal article full texts manually screened

		ODDPub	
	Open Data	Yes	No
Human rater	Yes	67	24
	No	23	678

- Sensitivity: 0.73; Specificity: 0.97; F_1-score: 0.73

Open Data Detection in Publications (ODDPub)

ODDPub is a text mining algorithm that parses a set of publications and detects which publications disseminated Open Data or Open Code together with the publication. It is tailored towards biomedical literature.

Authors

Nico Riedel (nico.riedel@bihealth.de), Miriam Kip, Evgeny Bobrov - QUEST Center for Transforming Biomedical Research, Berlin Institute of Health

Installation

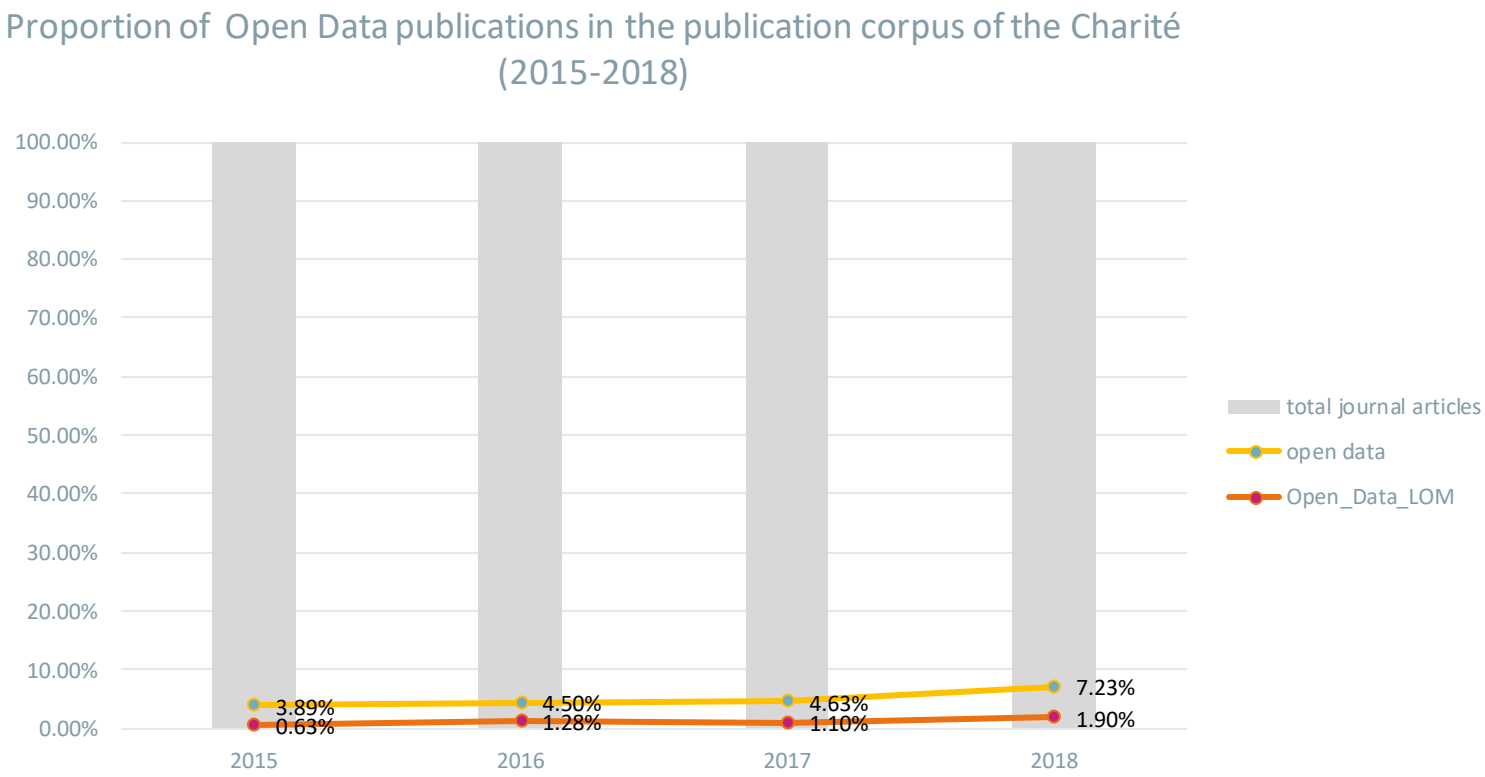
The latest version of the algorithm is structured as an R package and can easily be installed with the following command:

```
# install.packages("devtools") # if devtools currently not installed
devtools::install_github("quest-bih/oddpub")
```

How much Open Data did we find?

Validation sample (PubMed 2018):

- Open Data: 11.5% (n=91), Open Code: 1.4% (n=11) of publications



Kip et al., in preparation

BIH QUEST Transforming Biomedical Research

How are data shared?

For the validation sample:

Category	Number of occurrences
Supplemental Data	42
Field-specific repository	40
General-purpose repository (including GitHub)	14
Institutional repository	0
Personal/project-specific website	1
Data journal	0

Data sharing is most common for specific fields (genetics) and journals (e.g. PLoS)

Performance-oriented funding at Charité & MDC

- Performance-oriented funding at Charité traditionally rewarded third party funding and journal impact factor
- In 2018 we were able to distribute 200,000€ for Open Data Publications (Charité: 120,000€, MDC: 80,000€)
- We rewarded 156 publications from 2015-17 (96 Charité, 60 MDC)
- We could repeat this in 2019 (290,000€ in total)
- In the optimal case Open Data becomes part of the regular performance-oriented funding at Charité

The future of Open Data -
How would data sharing ideally look like?

Standardized reporting

- ID for each dataset (DOI or repository-specific ID)
- No more data sharing through the supplement
- Data availability statement in each publication
- Sufficient metadata describing the data-sets

→ Data-set search engine

FAIR & quality of shared data

Frequent data sharing is not all – shared data need to be reusable!

- FAIR principles are our best guidance here
- How well documented are the data?
- Actual machine-readable data
- Field-specific standards and repositories are a plus

Thank you

Nico Riedel
Data Scientist

QUEST Center -
Berlin Institute of Health
(BIH)

nico.riedel@bihealth.org

www.bihealth.org/quest

BIH QUEST
Transforming Biomedical Research

BIH Berlin Institute of Health
Charité & MDC

Aus Forschung wird Gesundheit