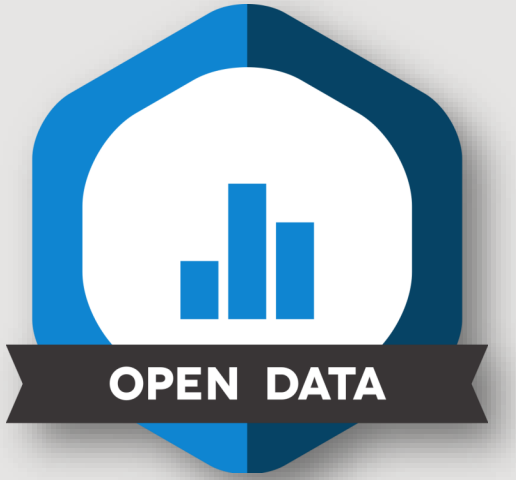
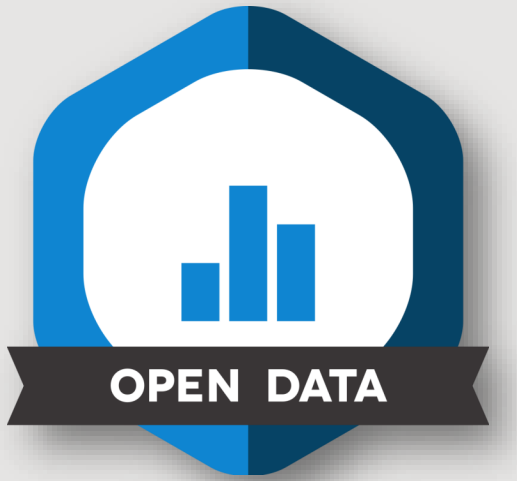
Kidwell, Mallory et al. “Badges to Acknowledge Open Practices: A Simple, Low Cost, Effective Method for Increasing Transparency.” OSF, 18 Aug. 2016. Web.

BIH QUEST
Center for Responsible Research



Open Data Definition

Open Data sind online kostenfrei zugängliche Daten, die verwendet, nachgenutzt und verbreitet werden können, mit allenfalls den Auflagen der Quellennennung und der Weitergabe unter gleichen Bedingungen (share-alike).



Open Data Definition

Für die Zuweisung der Open-Data-Incentivierung als zusätzlicher Indikator der LOM (Leistungsorientierte Mittelvergabe) Forschung an der Charité müssen die untenstehenden daten- und publikationsbezogenen Kriterien erfüllt sein. Allgemeine Kriterien für die LOM-Berechtigung der jeweiligen Person finden Sie auf der Intranetseite unter Interne Forschungsförderung. Die Kriterien für die Open-Data-Incentivierung sind wie folgt (Stand 2022): Forschungsdaten wurden von Forscherinnen und Forschern der Charité frei zugänglich gemacht ODER die Daten wurden zugangsbeschränkt geteilt und erfüllen die folgenden Voraussetzungen: die Daten sind in einem externen Repositorium (bzw. Archiv, Datenbank, Register) abgelegt
es ist ein standardisierter Zugangsweg benannt, d.h. die Zugangsvoraussetzungen, der Ablauf eines Antrags und die verantwortlichen Personen bzw. Stellen sind beschrieben der Grund für den beschränkten Zugang wird genannt oder ist unmittelbar aus dem Personenbezug ersichtlich der Zugang ist für alle akademisch Forschenden – mindestens des Europäischen Wirtschaftsraums – möglich Ko-Autor:innenschaft bei entstehenden Fachartikeln wird nicht zur Bedingung für die Bereitstellung der Daten gemacht die Bereitstellung erfolgt kostenlos oder gegen Aufwandsentschädigung Es kann sich um Rohdaten, Primärdaten oder Sekundärdaten (z.B. aus Analysen frei verfügbarer Datensätze, Meta-Analysen oder Health Technology Assessments) handeln; die Daten würden somit eine analytische Replikation (Nachvollziehung der Analyseschritte) zumindest eines Teils der Ergebnisse der Studie ermöglichen; die Nennung statistischer Zahlenwerte (Mittelwerte, Standardabweichungen, p-Werte etc.) reicht hierfür nicht aus. Daten wurden zu einer Artikelpublikation geteilt; somit fallen für sich stehende Datensätze ohne Artikelbezug nicht darunter. Die Daten sind auch unabhängig von der Publikation auffindbar; somit sind Supplementary Materials nur zulässig, wenn sie in einem Repositorium (Archiv) abgelegt und auch über dieses Repositorium auffindbar sind. • In der Publikation wurde explizit auf die Datensätze hingewiesen; ein Verweis auf z.B. Supplementary Materials ohne weitere Erläuterung reicht nicht aus, ebenso wie der Verweis auf eine Datenbank ohne Nennung von Datensatz, Accession code oder genauen Sucheinstellungen. Die Daten sind tatsächlich zugänglich und können zum Zeitpunkt der Überprüfung abgerufen werden (bei Daten unter Embargo muss dieses spätestens zum 31.7. ablaufen). Die Daten wurden in einem maschinenlesbaren Format geteilt; für Tabellen z.B. CSV-, Excel- oder Word-Dateien, nicht jedoch PDFs oder Bildformate. Im Folgenden fassen wir einige Aspekte zusammen, die nicht unter unsere Definition von Open Data fallen: Analyseskripte, Computerprogramme, Modelle und andere Methoden, Materialien oder Protokolle, auch wenn ihre Entwicklung Ziel des Forschungsprojekts war und/oder ihre Darstellung Hauptgegenstand der Publikation ist. Wenn für die Entwicklung oder Validierung Daten erhoben und geteilt wurden, können diese jedoch unter die Open-Data-Definition fallen, sofern sie für sich stehend nachvollziehbar und nachnutzbar sind. Daten im Artikeltext selbst, sofern es sich nicht um eingebettete Tabellen handelt, die zugleich auch als für sich stehende digitale Objekte abgerufen werden können. Bildliche, audiovisuelle und andere Daten, die primär der Illustration dienen. Daten zu Fallberichten (case reports), sofern sie nicht in Repositorien aus der jeweiligen Fachdisziplin abgelegt wurden. Bei Systematic Reviews und Meta-Analysen: Listen mit Quellenangaben oder andere allgemeine Informationen zu den untersuchten Studien wie Erhebungsmethode oder Teilnehmendenzahl (LOM-fähig sind dagegen neu aus der Originalliteratur zusammengesetzte Datensätze, die eine Nachvollziehbarkeit der Analyse sicherstellen, wie z.B. extrahierte Textstellen oder statistische Werte). Daten, die nur auf Nachfrage und/oder bei Erfüllung von Voraussetzungen verfügbar sind, es sei denn, es handelt sich um personenbezogene oder anderweitig sensible Daten, die über einen standardisierten Nachfrageweg verfügbar sind (darunter fällt z.B. nicht die Verfügbarkeit „upon request“). Daten aus Datensammlungen von Konsortien („Datenpools“), wenn unklar ist, ob die Autor:innen selbst einen Beitrag zum Pool geleistet haben. Daten, für die nur ein „private Link“ geteilt wird, so dass sie nicht im Repositorium gefunden werden können, sondern ausschließlich über die Publikation zugänglich sind.
Daten, die vor dem betrachteten LOM-Zeitraum geteilt wurden (hierdurch soll sichergestellt werden, dass nur eine begrenzte Zahl von Artikeln für das Teilen eines bestimmten Datensatzes belohnt werden kann).

Open Data operationalisieren

Erfahrungen mit der Formulierung überprüfbarer Kriterien

Evgeny Bobrov

BIH in der Charité, QUEST Center

3. Sächsische FDM-Tagung, Leipzig

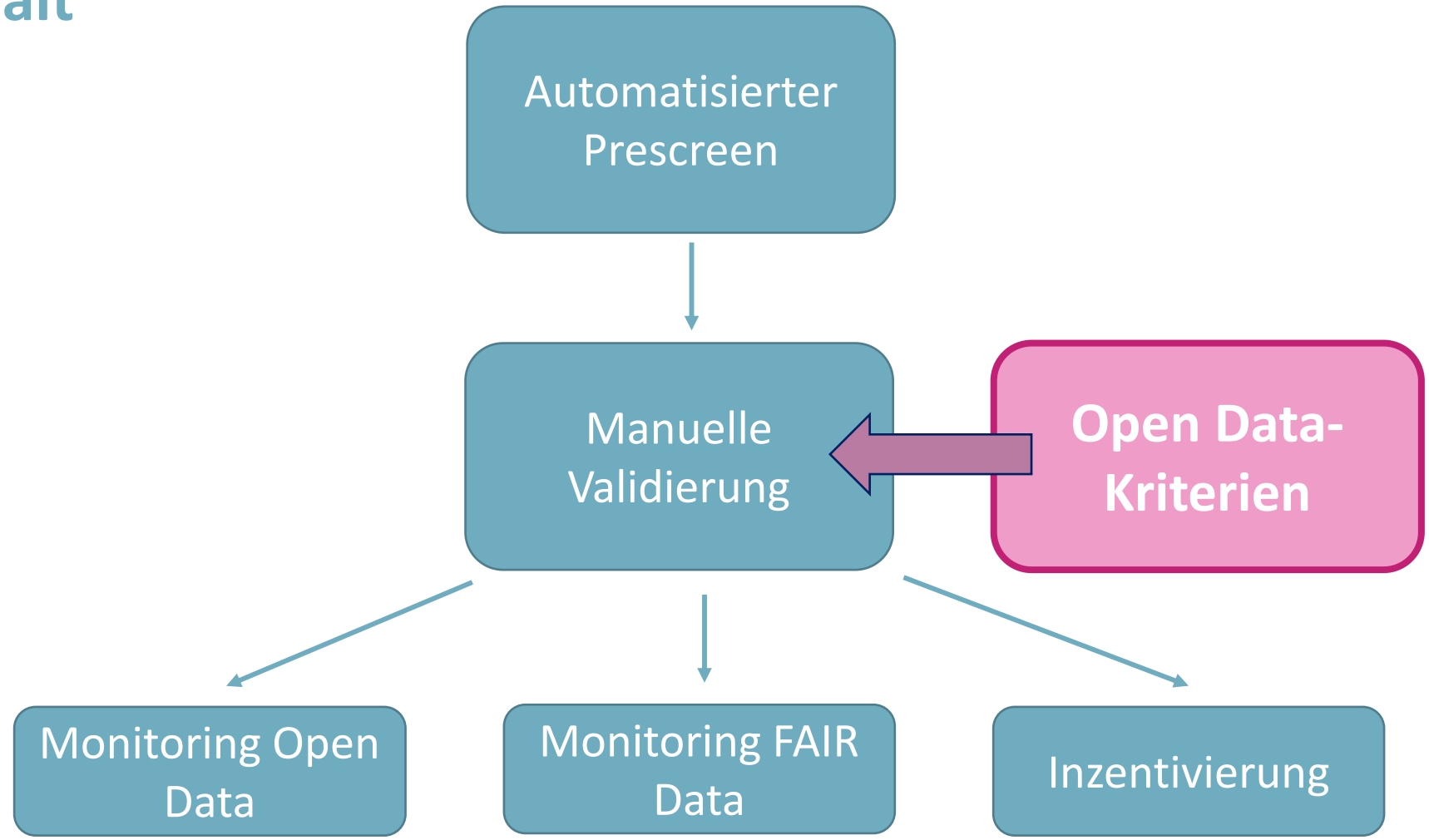
22.09.2022

BIH QUEST
Center for Responsible Research

BIH Berlin Institute of Health @Charité

Aus Forschung wird Gesundheit

Inhalt



BIH QUEST Center for Responsible Research

Automatisierter Prescreen

ODDPub – Open Data Detection in Publications

- Open Source-Tool in R, am QUEST entwickelt
- Detektion von Schlagworten in Artikeln → Detektion von Open Data (und Open Code)
- Open Data detektiert als Kombination von Begriffen zu Verfügbarkeit und Ablageort, sowie ggf. Accession Code
- Charité: ca. 5500 Publikationen/Jahr → ca. 770 Hits
- → Müssen manuell gescreent werden

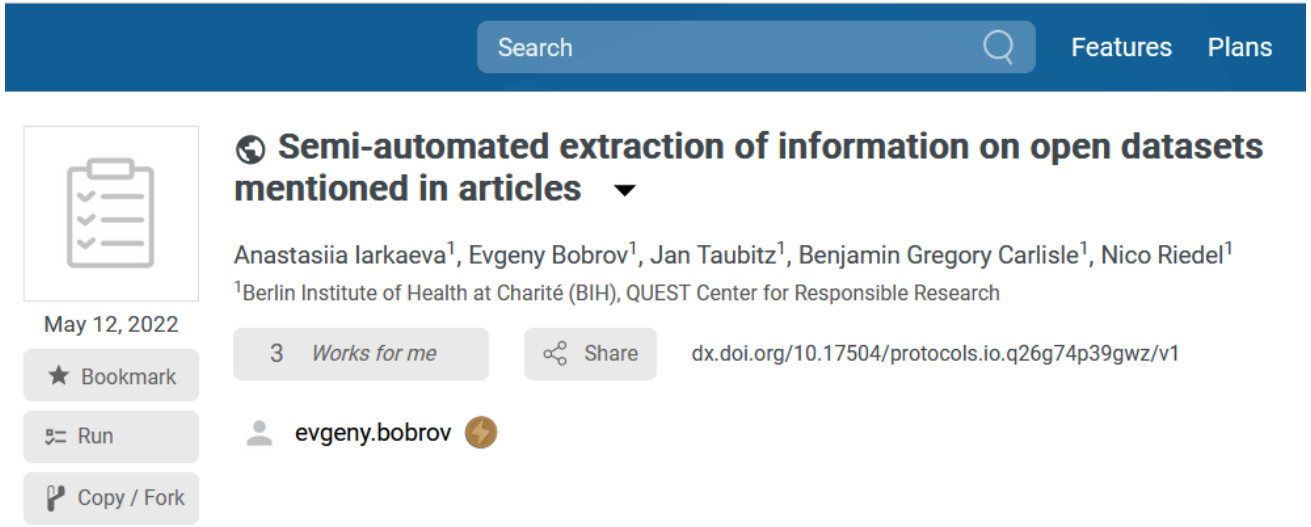
Nico Riedel

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

Manuelle Validierung

Screening von Open Data-Statements in Numbat

- Open Source-Tool in PHP, ursprünglich entwickelt für Systematic Reviews
- Adaption für Open Data-Screening
- → Detailliertes Protokoll des gesamten Workflows in protocols.io



Anastasiia Iarkaeva

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

Kidwell, Mallory et al. “Badges to Acknowledge Open Practices: A Simple, Low Cost, Effective Method for Increasing Transparency.” OSF, 18 Aug. 2016. Web.

BIH QUEST Center for Responsible Research

Open Data-Kriterien

Primäre Nutzung

Verteilung Leistungsorientierter Mittel (LOM) an Einrichtungen der Charité für Open Science-Praktiken → Inzentivierung & Bewusstseinsbildung

Weitere Nutzung: Dashboard für Monitoring und Kommunikation

Grundlegende Entscheidungen

- Einheit der Bewertung: Artikel vs. Datensatz
- Binäre vs. abgestufte Bewertung
- Nur Offenheit vs. auch Vollständigkeit

Kip M, Riedel N, König S, Bobrov E (2022). Including open data as an additional indicator for the performance-based allocation of funds (1.0). Zenodo. https://doi.org/10.5281/zenodo.6651941 ***[Article in preparation]***

Open Data-Kriterien: Grundsatzentscheidungen

Einheit der Bewertung: **Artikel, nicht Datensatz**

Hauptgrund: Implementierung innerhalb bestehender LOM, die ebenfalls auf Artikeln basiert

Weitere Gründe:

- Detektion für sich stehender Datensätze 2019 schwierig; inzwischen über DataCite evtl. besser möglich (z.B. Data Monitor eines großen Verlages)
- Einfachere Kommunikation: Artikel als Referenz des Wissenschaftsbetriebs
- Mindestmaß an indirekter Qualitätssicherung über Peer Review

Kip M, Riedel N, König S, Bobrov E (2022). Including open data as an additional indicator for the performance-based allocation of funds (1.0). Zenodo. https://doi.org/10.5281/zenodo.6651941 ***[Article in preparation]***

Open Data-Kriterien: Grundsatzentscheidungen

Binäre vs. abgestufte Bewertung: **Binär (d.h. ein Artikel hat entweder Open Data-Status, oder hat keinen Open Data-Status)**

Gründe:

- Einfachere Berechnung und Verteilung finanzieller Förderung
- Einfachere Kommunikation
- (Einfachere Visualisierung)

Kip M, Riedel N, König S, Bobrov E (2022). Including open data as an additional indicator for the performance-based allocation of funds (1.0). Zenodo. https://doi.org/10.5281/zenodo.6651941 ***[Article in preparation]***

Open Data-Kriterien: Grundsatzentscheidungen

Nur Offenheit oder auch Vollständigkeit: **Nur Offenheit**

Gründe:

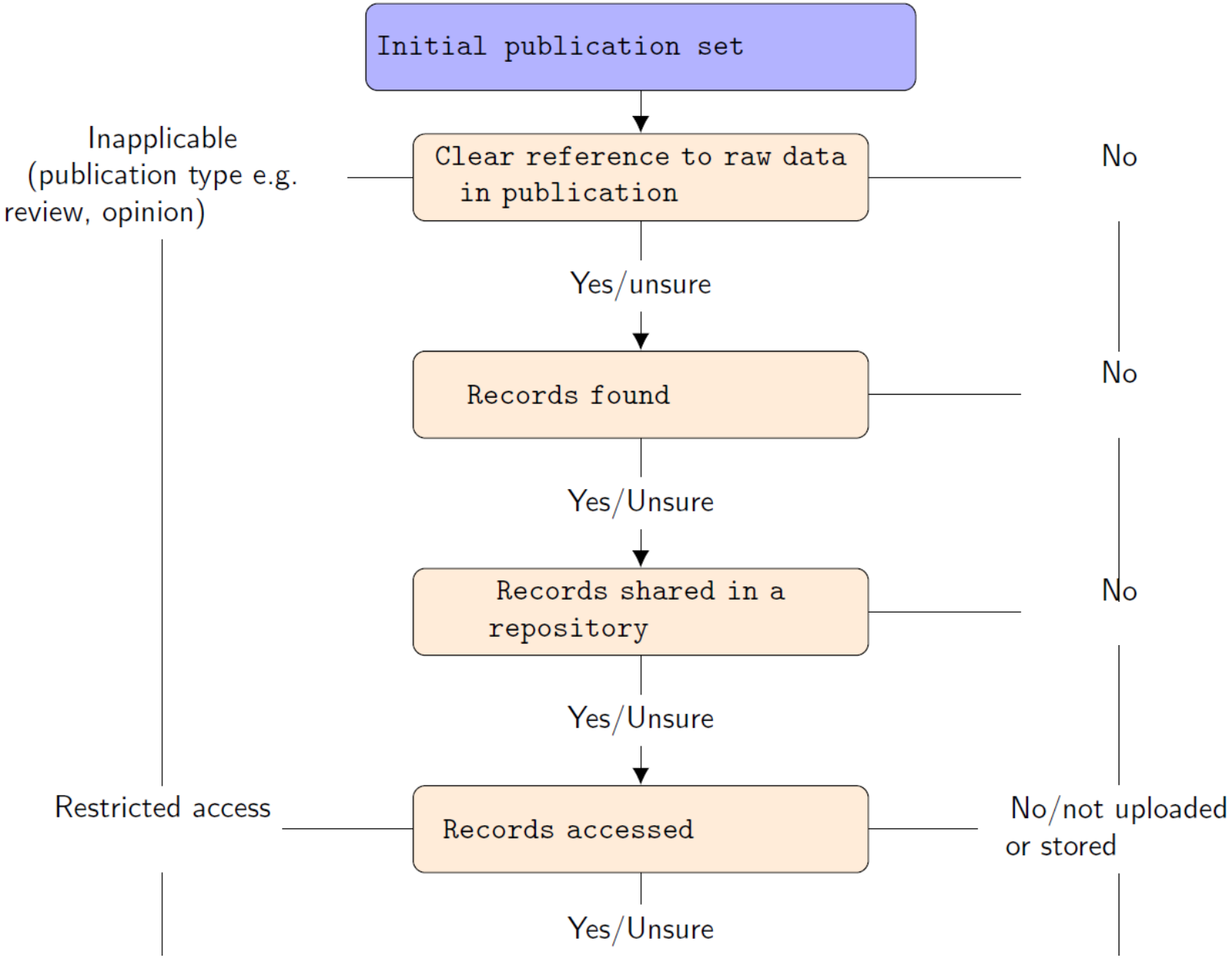
- Umsetzbarkeit: ohne disziplinäres Wissen und sehr großen Zeitaufwand ist Vollständigkeit nicht zu bewerten
- Zudem auch Fairness – nicht immer können alle Daten geteilt werden, und Inzentivierung sollte nicht zu hochschwellig sein

Kip M, Riedel N, König S, Bobrov E (2022). Including open data as an additional indicator for the performance-based allocation of funds (1.0). Zenodo. https://doi.org/10.5281/zenodo.6651941 ***[Article in preparation]***

Open Data-Kriterien: Übersicht

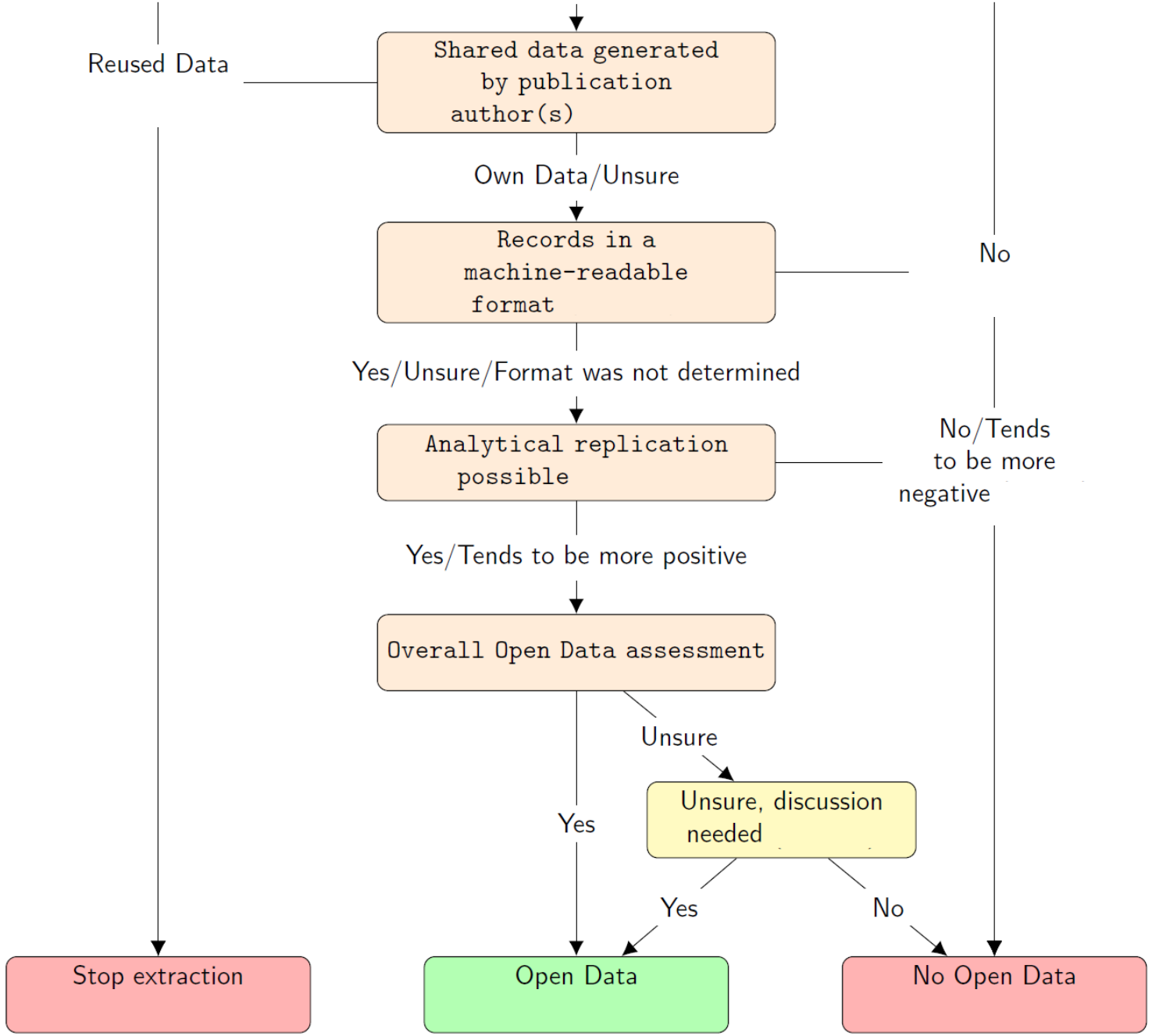
1. Auf Verfügbarkeit von Daten wird im Artikel klar hingewiesen
2. Daten sind auffindbar
3. Daten wurden über ein Repositorium geteilt
4. Daten wurden von Autor*innen des Artikels generiert oder gesammelt
5. Daten sind zugänglich
6. Daten sind in einem maschinenlesbaren Format
7. Es handelt such um Rohdaten (d.h. sie erlauben die analytische Replikation)
8. Oder, als Alternative zu (5) - (7): personenbezogene Daten, die mit Auflagen zugänglich sind

Open Data-Kriterien: Flowchart (1)



BIH QUEST Center for Responsible Research

Open Data-Kriterien: Flowchart (2)



BIH QUEST Center for Responsible Research

Die Abgrenzung operationalisieren #1 – Auffindbarkeit

Supplementary Materials müssen in Repositorien abgelegt worden sein und auch für sich stehend auffindbar sein

Auffindbarkeit unabhängig vom Artikel == eines von beiden muss gegeben sein:

(1) Bei Suche im Repositorium nach den ersten 5 Worten des Artikeltitels plus Namen von Erst- und Letztautor*in erscheint der Datensatz unter den ersten 10 Hits

(2) Bei Verwendung des Accession Codes oder Identifiers plus ggf. Name des Repositoriums ist der Datensatz bei einer Suche mit einer Suchmaschine unter den ersten 5 Hits; wird bestätigt durch z.B. Autor*innen, Titel oder Jahr

Es ist offensichtlich, dass die genauen Suchsettings arbiträr sind

Konsequenz: Supplements in Figshare sind nach dieser Definition Open Data, auch wenn schlecht auffindbar und meist wenig FAIR

Die Abgrenzung operationalisieren #2 – Autorenschaft

Daten wurden von Autor*innen des Artikels generiert oder gesammelt

Erfordert: Klarheit über die Autorenschaft von Daten durch Autor*innen des Artikels

Schließt aus: Daten aus Sammlungen von Konsortien ("data pools"), wenn unklar, ob Autor*unnen selbst zum Pool beigetragen haben. Wenn im Artikel gesagt wird, dass Autor*innen zum Pool beigetragen haben, reicht dies jedoch aus.

Erfordert nicht: Autor*innen des Datensatz müssen Charité-Affiliation haben

Konsequenz: Open Data-Kategorisierung auch dann, wenn Größe des Beitrag zum Datenpool sehr klein oder unbekannt ist und/oder durch Autor*innen anderer Einrichtungen erfolgte

Die Abgrenzung operationalisieren #3 – Andere Outputs

Es handelt sich um Roh-, Primär- oder Sekundärdaten, die analytische Replikation ermöglichen

Schließt aus: Andere Outputs als Daten, z.B. Analyseskripte, Software und andere Methoden oder Protokolle, selbst wenn ihre Entwicklung Fokus des Projekts oder Artikels ist. Daten für Entwicklung oder Validierung werden nicht ausgeschlossen.

Konsequenz: Keine Open Data-Kategorisierung von Artikeln, die andere Openness-Praktiken umsetzen, aber keine eigenen Daten geteilt haben (teilweise Data Reuse). Open Data-Kategoriserung vergleichbarer Artikel, die z.B. in Github einen Validierungsdatensatz aufgenommen haben (ggf. nachgenutzt mit offener Lizenz).

Die Abgrenzung operationalisieren #4 etc.

Zugangsvoraus-setzungen

Daten im Artikel selbst

Daten mit privatem Link

Maschinenlesbar-keit des Formats

Lizenz des Datensatzes

„Available upon request“

Systematische Reviews & Meta-Analysen

Illustrative Daten

„Source Data“

Daten ohne Artikelbezug

Daten unter Embargo

Klarer Hinweis auf Daten

Daten in Fallbeispielen

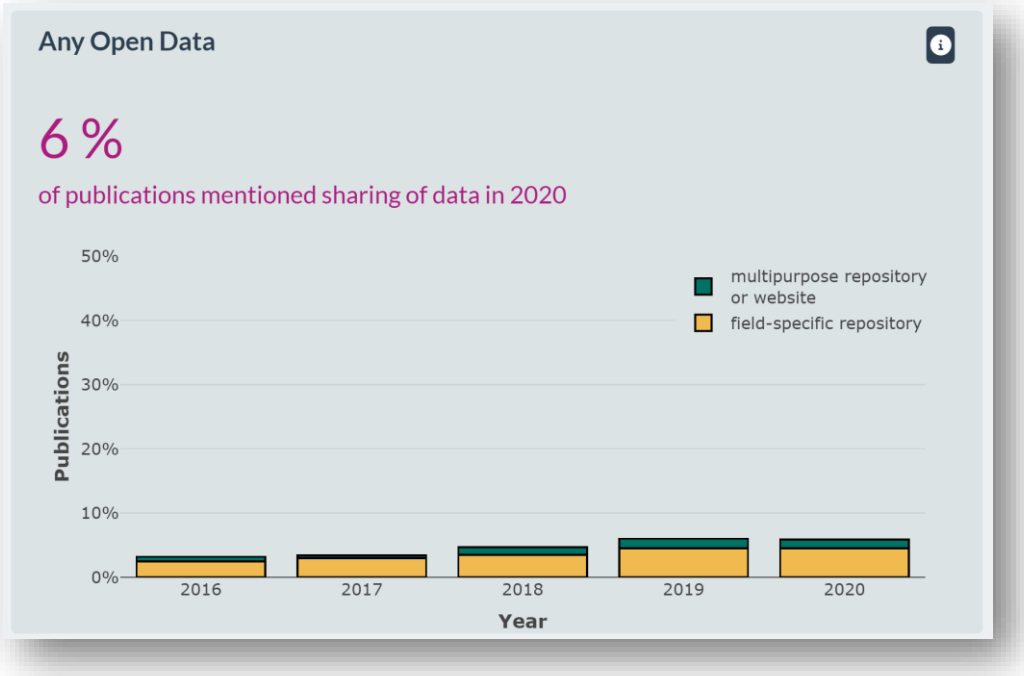
Statistische Werte

BIH QUEST Center for Responsible Research

Die Abgrenzung weiter operationalisieren

Restricted Access

Monitoring Open Data



Vladislav Nachev

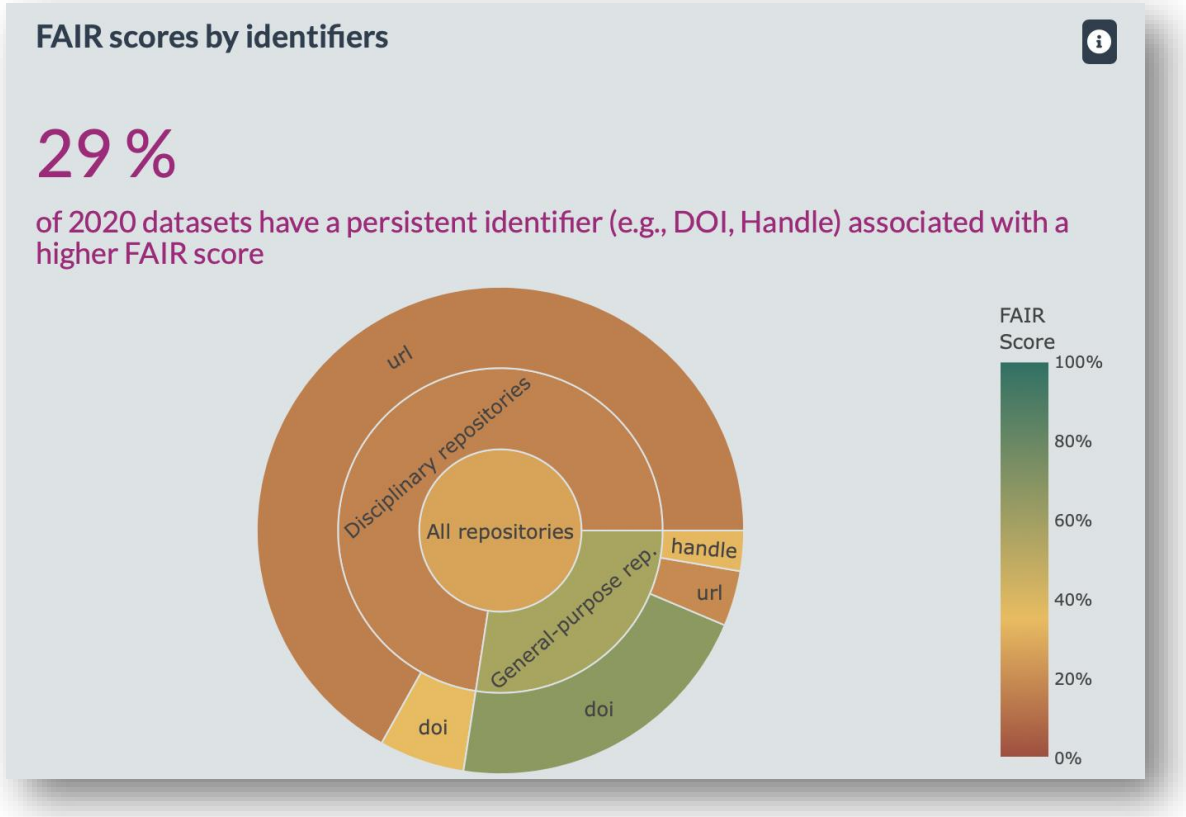
Für Methoden und weitere Metriken wie Open Code siehe

Charité Dashboard on Responsible Research

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

BIH QUEST Center for Responsible Research

Monitoring FAIR Data

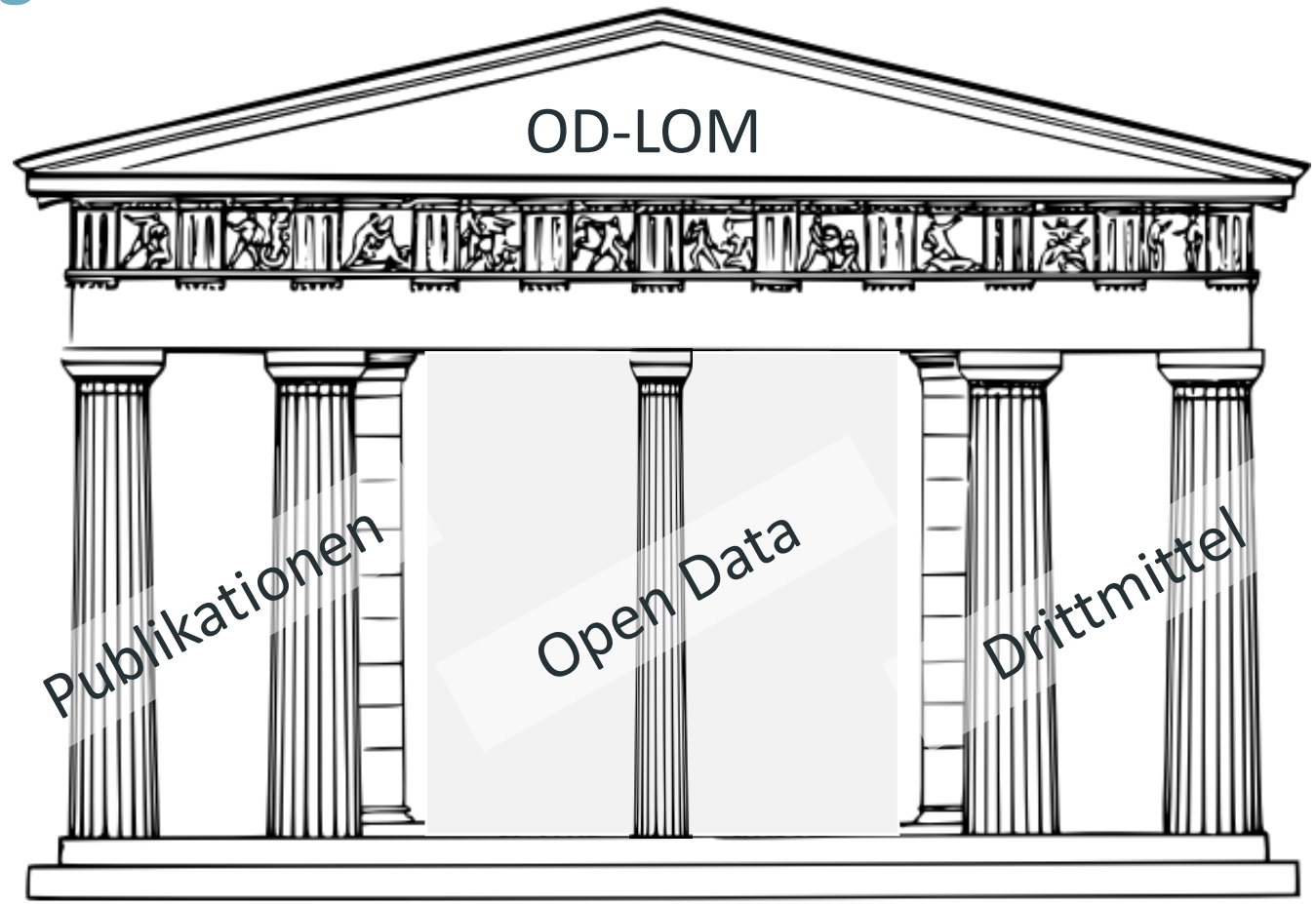


Für weitere Analysen siehe **Dashboard on Data Reusability (FAIR Data)**

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

BIH QUEST Center for Responsible Research

Inzentivierung



Miriam Kip

OD-LOM = Leistungsorientierte Mittelvergabe für Open Data

2022: 300.000€ als Inzentivierung an der Charité

BIH QUEST Center for Responsible Research

Zusammenfassung

Open Data zu operationalisieren ist…

…konzeptuell schwierig

…praktisch aufwändig

…abhängig vom use case

…work in progress

Vielen Dank!

quest.bihealth.org/

quest-dashboard.charite.de

BIH QUEST
Center for Responsible Research

BIH Berlin Institute of Health @Charité

Aus Forschung wird Gesundheit