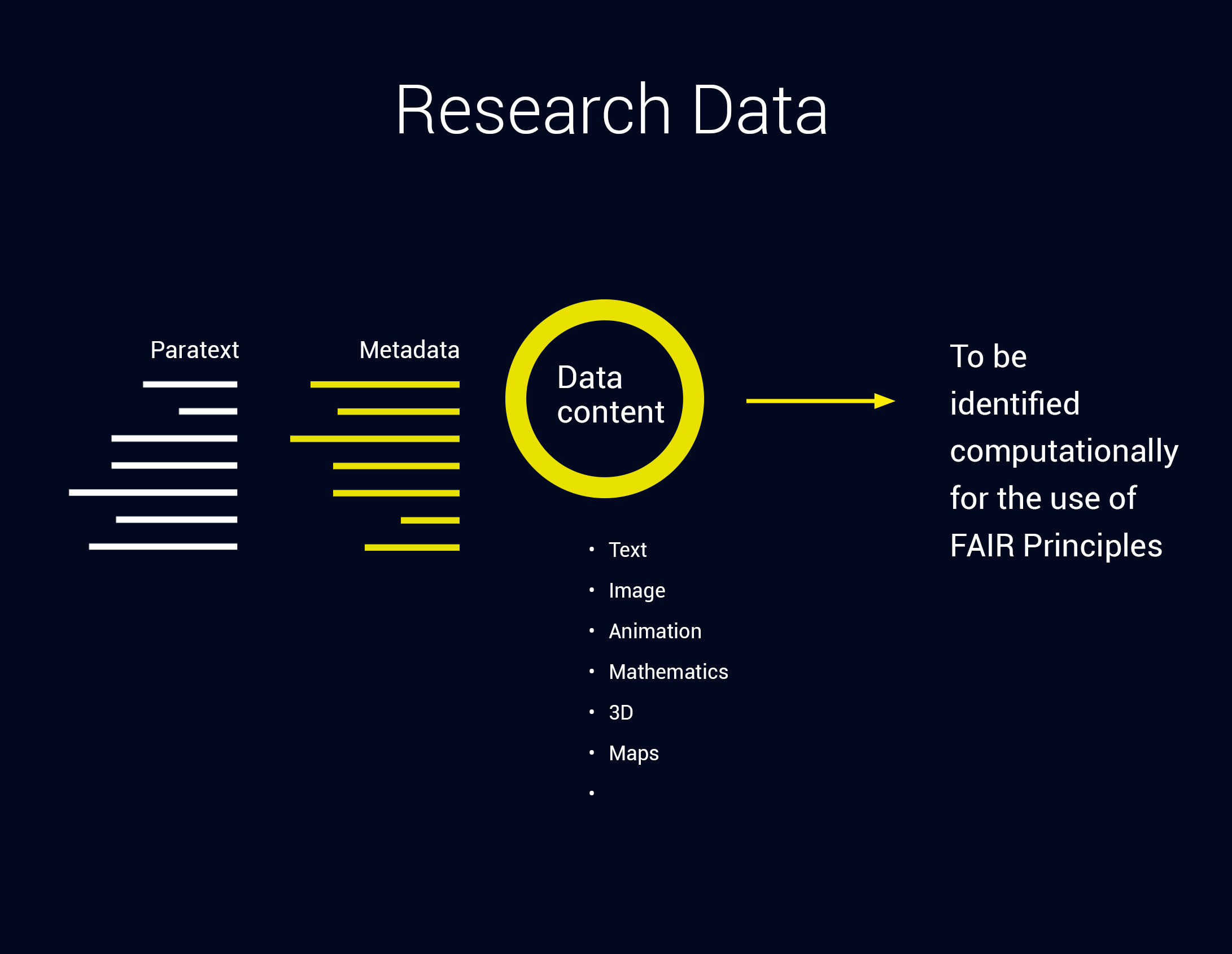
# Research Data



sd

## Epistemology of research data

Forschungsdaten wurden in der Wissenschaftenschaftsgeschichte seit den ersten Anfängen als eigenes wissenschaftliches Genre publiziert und genutzt. Sie sind damit keine Erfindung des digitalen Zeitalters. Datenkataloge wie Ptolemy’s Geography haben die Orte und Wege der antiken Welt geordnet, und mithilfe seiner Handy Tables bestimmten Gelehrte die Konstellation der Gestirne. Auch Kalender und damit verbundene religiöse Feste orientierten sich an Forschungsdaten, und Lexika normierten nicht zuletzt das Fachverständnis und die Kommunikation von Wissen. Umfangreiche Sammlungen von Forschungsdaten schufen den Erfolg der Wissenschaft im engen Zusammenspiel mit der Entwicklung großer Theorien – doch sie trugen als kollaborative Gemeinschaftswerke häufig keinen Autorennamen. Forschungsdaten werden in allen Wissenschaftskulturen mit größter Sorgfalt epistemisch geprüft. Es sind Wissensbestände, die regelmäßig untersucht werden, um Irrtümer möglichst schnell zu identifizieren und validierte Daten auszuzeichnen. Forschungsdaten sind demnach nicht einfach Informationen, die im digitalen Zeitalter im Internet verbreitet, mit anderen Daten verlinkt und offen zugänglich ohne Ansehen ihres Wahrheitsgehalts weiter genutzt werden. Es handelt sich auch nicht um “closed data”, die wie ein esoterischer Schatz nur auf Basis von Privilegien zugänglich sind.

Die Wissenschaft hat den Begriff des Wissens eingeführt, um diesen besonderen Status von Informationen über Gesellschaft und Welt zu kennzeichnen. Mit wissenschaftstheoretischer Allgemeinheit können wir Forschungsdaten als die Beschreibungen von Gegenstandsklassen jeder Art charakterisieren, die als Wissen qualifiziert werden oder deren Prüfung auf diesen Status aussteht. Forschungsdaten als Wissen über die Eigenschaften von Forschungsobjekten mit allen methodischen Bewertungen durch die wissenschaftliche Gemeinschaft werden dabei von beliebigen Informationen unterschieden: Forschungsdaten werden kollaborativ, offen und kritisch geprüft; mit ihnen wird ein Geltungsanspruch vertreten, der jederzeit revidiert werden kann. Die Organisation wissenschaftlicher Kommunikation und das Ethos im Umgang mit Wissen ist daraufhin optimiert, Irrtümer zu minimieren. Die etablierten methodologischen Mechanismen sollten auch für Forschungsdaten in allen Wissenschaften gelten. Für das NFDI4SD ist dieser besondere Status von Forschungsdaten im Konzert aller wissenschaftlicher Disziplinen – auch und besonders der sogenannten kleinen Fächer – Hauptaugenmerk für die Aufgaben einer zukünftigen Forschungsdateninfrastruktur.

!!! note “research data” **RD1**: With Research data scholars and scientists claim fallible knowledge. Research data can be justified and serve as justifying grounds. Research data are open to epistemic qualification. For this they must be computationally identified in all times independent of their symbolic form. All epistemic qualification are transparent to the community.

Mit dem Begriff “epistemic qualification” können alle zusätzlichen Attribute zusammengefasst werden, die für eine Auszeichnung und Behandlung von Forschungsdaten als Wissen relevant sind. Forschungsdaten lassen sich nach den Komponenten des Gehalts der Daten, ihrer Metadaten und der epistemischen Qualifikation unterscheiden. Der Gehalt der Forschungsdaten meint die von den Daten beschriebenen Eigenschaften der Forschungsgegenstände. Forschungsdaten umfassen immer beides: ihren Gehalt und die zusätzlichen Attribute. Eine Infrastruktur von Forschungsdaten muss entsprechend diese Integrität berücksichtigen und sorgsam beibehalten.

Hier kann eine Analogie zur *Paratext*-Debatte [@zotero-40395] der letzten Jahre hergestellt werden: Basierend auf dem von Gérard Genette (1930–2018) eingeführten philologischen Begriff werden Paratexte (unterteilt in *Peri-* und *Epitext*) als jene einem (ursprünglich literarischen) Basistext beigefügten Elemente bezeichnet, welche Rezeption und Vertrieb eines Werkes entscheidend mitprägen und steuern, darunter Informationen zum Autor, Verlag, Titelei, Vorwort, Dank etc.[@genette1997] Dieses Konzept, das größere Aufmerksamkeit auf Produktionsprozesse und Autorisierungsinstanzen lenkt, wurde in der weiteren Forschung sukzessive um Textgattungen und Formate neuer Medien erweitert. Im Kontext der Digitalisierung wissenschaftlicher Daten heißt das: Gerade die Prüfung, Autorisierung und nachhaltige Präsentation von Forschungsdaten wird entscheidenden Einfluss auf die Wahrung und Weiterentwicklung nationaler wie internationaler Standards und Qualitätsansprüche haben.

!!! note “research data” **RD2**: Research-data are a composit of data-content, metadata and epistemic attributes. It unifies all those information about data-sources like texts including their paratexts. Their composition will adhere epistemic principles enhancing the requirements of obtaining, accessing and maintaining scientific knowledge.

Der Physiker und Philosoph Hans Reichenbach unterschied ab den 1920er-Jahren für die wissenschaftliche Beurteilung von Aussagen den Entdeckungszusammenhang vom Rechtfertigungszusammenhang. Der Entdeckungszusammenhang berücksichtigt Gründe, die zu einer neuen Einsicht und Beurteilung wissenschaftlicher Sachverhalte führen. Davon zu trennen sind die Gründe für die Rechtfertigung von Aussagen – unabhängig davon, ob es sich um theoretische Hypothesen oder den Gehalt von Forschungsdaten handelt. Aus der Perspektive der Forscher, die auf der Suche nach wissenschaftlichen Antworten sind und dazu Forschungsdaten erstellen und weiternutzen, steht der Entdeckungszusammenhang im Vordergrund. Dabei wird übersehen, dass auch ein Entdeckungsprozess immer wieder auf auf bereits gerechtfertigtes Wissen stützt. Forschungsdaten werden besonders sorgfältig validiert, wenn sie bereits während des Forschungsprozesses versioniert für die Publikation bearbeitet werden.

!!! note “research data” **RD3**: Der Wert von Forschungsdaten muss immer aus der Perspektive des weiteren Nutzens durch andere beurteilt werden, und nicht aus der Perspektive der ersten Person im Entdeckungsprozess.

Die Rechtfertigung von Wissen – und damit um Geltung von Forschungsdaten – ist immer eine kollaborative Aufgabe der gesamten Forschungsgemeinschaft. Die Forschungsdateninfrastruktur muss deshalb für Mittel sorgen, mit denen die Provenienz der Forschungsdaten und epistemische Abhängigkeiten von Dritten nachvollzogen werden können. Gerade dies ist derzeit selten gewährleistet.

!!! note “research data” **RD4**: Die Rechtfertigung der Daten durch andere Forschungsdaten soll möglichst transparent und nachvollziehbar sein. Ein “dependency graph” kann die Abhängigkeit der Forschungsdaten von anderen Forschungsdaten aufzeigen und damit die epistemischen Abhängigkeiten untereinander verdeutlichen. Es soll erkennbar sein, von welchen wissenschaftlichen Annahmen die Geltung von Forschungsdaten abhängt.

Innerhalb einer innovativen Wissenschaftsorganisation folgt daraus für neu gewonnene Forschungsdaten, dass diese möglichst schnell und vollständig hinsichtlich ihrer epistemischen Qualifikation anderen Forscherinnen und Forschern zugänglich gemacht weren müssen. Dieser Akt der Übergabe von Forschungsdaten aus den Händen der im Entdeckungszusammenhang arbeitenden Forschenden in die wissenschaftliche Öffentlichkeit verdient einen eigenen, besonderen Begriff: die *wissenschaftliche Publikation*.

!!! note “research data” **RD5**: Publication. Die Publikation von Forschungsdaten ist der Akt der Übergabe von erstellten Befunden von Daten im Entdeckungszusammenhang an die wissenschaftliche Gemeinschaft. Erst dadurch werden Ergebnisse zu rechtfertigbarem Wissen. Den Status von **research data** erhalten empirische Befunde erst durch die wissenschaftliche Publikation. Datenpublikation führen zu zitierfähigen Datenbeständen, die von anderen Forschenden wie die derzeitige Forschungsliteratur in einer Bibliographie gelistet wird.

Das letztgenannte Prinzip führt in der Konsequenz dazu, die Forschungsdatenpublikation möglichst rechtzeitig im Managementplan zu berücksichtigen. Eine sequentielle Abfolge von Prozessen, die eine Publikation von Forschungsdaten erst nach Abschluß eines Projektes nach allen anderen abgeschlossenen Tätigkeit vorsieht, ist somit wissenschaftstheoretisch unvernünftig und schadet einer innovativen Wissenschaft. Die Datenpublikation und die innovative Datenzitation werden dazu führen, dass Forschungsprojekte eine große wissenchaftliche Sichtbarkeit erlangen und die eigenen Datenpublikation den Impact des Projektes und der daran Beteiligten wesentlich steigern.

!!! note “research data” **RD6** Citing research data. The reuse of research data for other data or in textual publication should cite the origin. Citing research data should be listed in bibliographies or lists of references like any other scholarly source. Its publication format resembles the reference styles like any other bibliographical source. Citations can be managed and administered in citation databases or tools.

Während der Laufzeit von Forschungsprojekten entstehen viele Forschungsdaten zu anderen Daten. Die neuen Daten annotieren, verlinken oder werden in anderer Hinsicht für den weiteren Verlauf der Forschungen benötigt. Der Umfang dieser Forschungsdaten ist riesig, und dennoch werden sie derzeit nur in geringem Umfang selbst publiziert. Diese Datenbestände vergößern den Nutzen der mit ihnen verbundenen Daten erheblich.

!!! note “research data” **RD7** Enhancing research data. Modern research projects create during the research process rich additional research data.

## Institutional consequences: FAIR use

Wir verstehen die FAIR Prinzipien als Anforderungen an eine Infrastruktur, die dem Status von Forschungsdaten als Wissensbestand für eine kollaborative, offene Wissenschaft dienen muss.

Um diese Voraussetzungen zu schaffen, verpflichteten sich die internationalen Forschungsverbände und Institutionen auf die Umsetzung der FAIR principles. Zunehmend knüpfen Förderinstitutionen die Mittelvergabe an erfüllte FAIR Bedingungen. Diese können aber heute von Forschern allein nicht garantiert werden: Insbesondere agile Forschungsvorhaben in den kleinen Fächern verfügen nicht über die Ressourcen, um die besonderen Anforderungen der Bereitstellung und Nutzung von Forschungsdaten zu realisieren. Dazu braucht es zukünftig eine wissenschaftsgetragene Infrastruktur.

Von den vielen Implementationen der FAIR principles folgen wir den Vorschlägen der GO FAIR initiative.[[1]](#footnote-22) Diese formuliert ihre Vision wie folgt: “Fostering the coherent development of the global Internet of FAIR Data & Services (IFDS), with the main focus on early developments in the European Open Science Cloud (EOSC).” Citing its website:[[2]](#footnote-24)

!!! note “research data” **RD8**: The FAIR principles (below) regulate the implementation of technical usage of research data. They will guide the implementation of NFDI4SD research data infrastructure.

## Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

## Accessible

Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

## Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

## Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

Viele der genannten Prinzipien ermöglichen einen großen Interpretationsspielraum. Entsprechend lassen sich aus den Regeln keine technischen Spezifikationen ableiten. Ausdrücke für die Freigabe und den Zugriff auf Daten wie “released” und “accessible” sind hinsichtlich der Verfahren unterspezifiziert. NFDI4SD’s Partner ZENODO vom CERN entwickelte seine eigenen [“best effort principles”](https://about.zenodo.org/principles/).[[3]](#footnote-31) Ein massgeblicher Anteil der Datenpublikation von kleineren und mittelgroßen Forschungsdatensammlungen wird nach unserer Erwartung auf der Zenodo-Platform erfolgen. Umfangreiche Kollektionen werden in eigenen “data stores” gespeichert werden. Sie sollen im Rahmen einer ersten Approximation die Implementationsvorschläge von Zenodo umsetzen.

\bibliography

1. [go fair](https://www.go-fair.org/go-fair-initiative/) [↑](#footnote-ref-22)
2. [Go fair principles](https://www.go-fair.org/fair-principles/) [↑](#footnote-ref-24)
3. Die Zielsetzung einer Open-Access-Infrastruktur ist prominent im [plan-s](https://www.coalition-s.org/addendum-to-the-coalition-s-guidance-on-the-implementation-of-plan-s/principles-and-implementation/) formuliert und von vielen europäischen Wissenschaftsorganisationen auf die Agenda gesetzt worden. Die Ergebnisse einer Selbstevaluation (self assessment) hinsichtlich dieser Zielsetzungen werden zusammen mit den ermittelten Metriken der Forschungsdatennutzung regelmäßig publiziert. [↑](#footnote-ref-31)