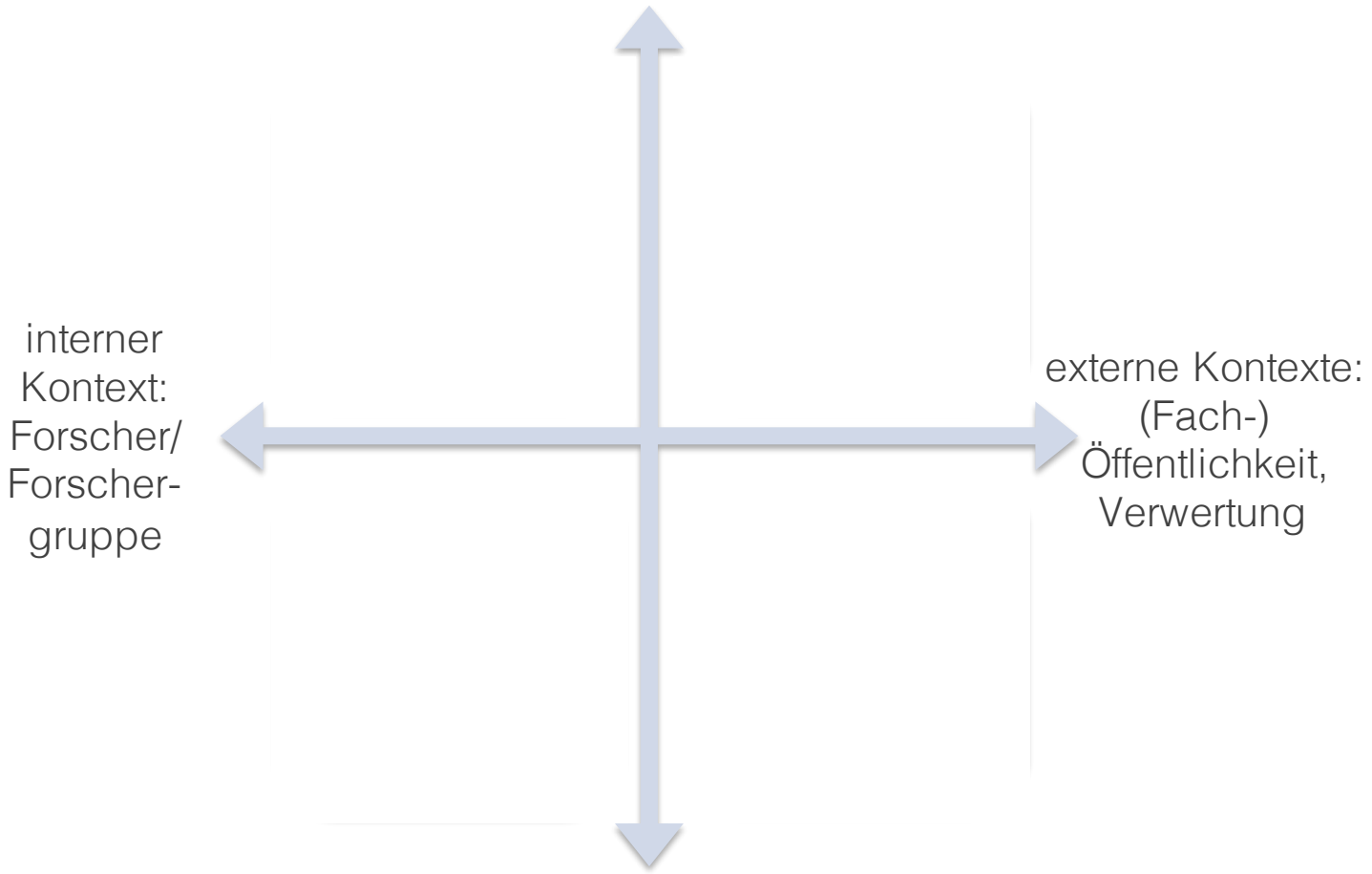
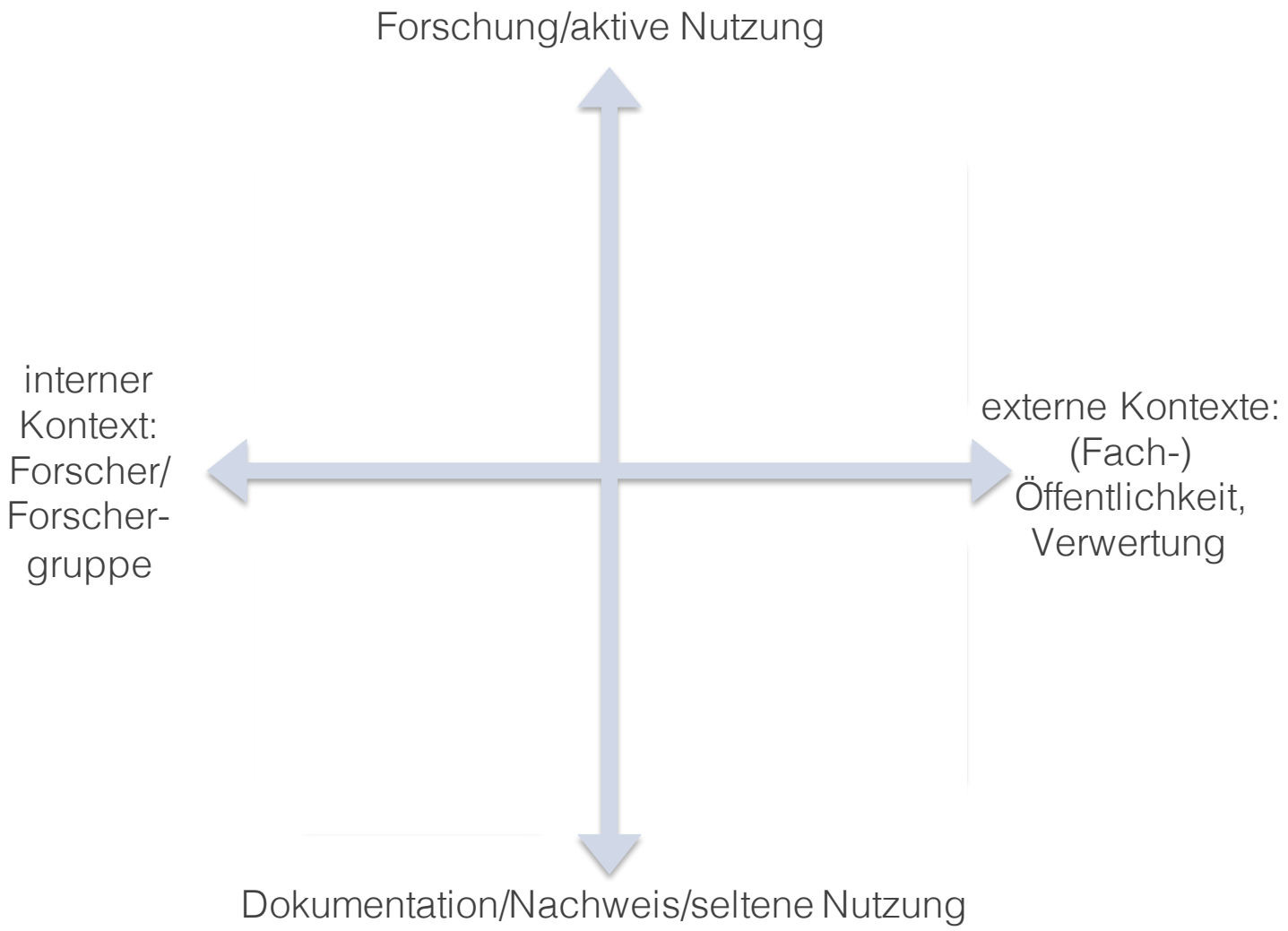
Forschungsdatenmanagement: Formen und offene Fragen

Jens Ludwig, Staatsbibliothek zu Berlin
jludwig1@gmail.com
Berlin, 24.06.2015

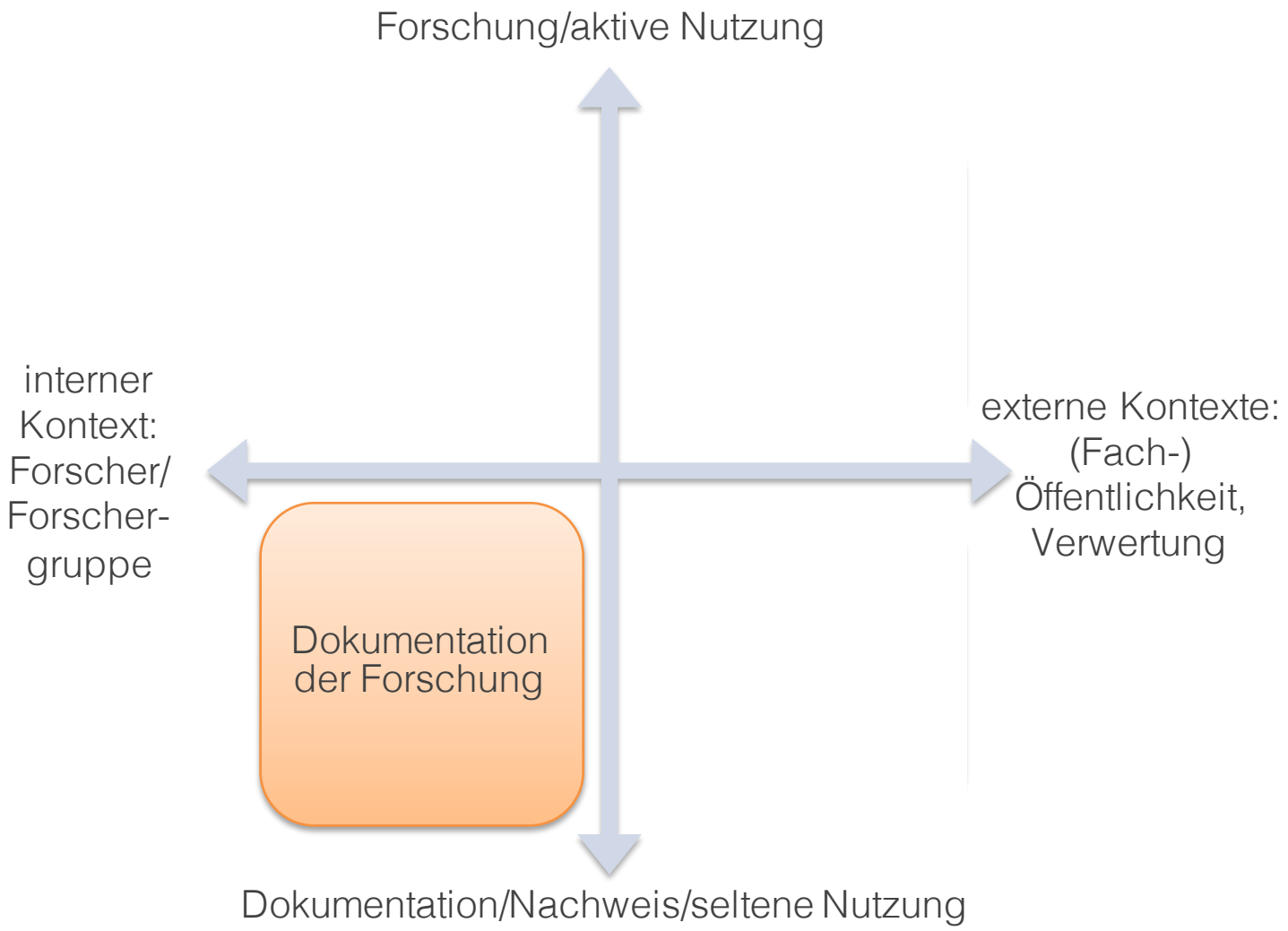
Formen des Forschungsdaten-Managements



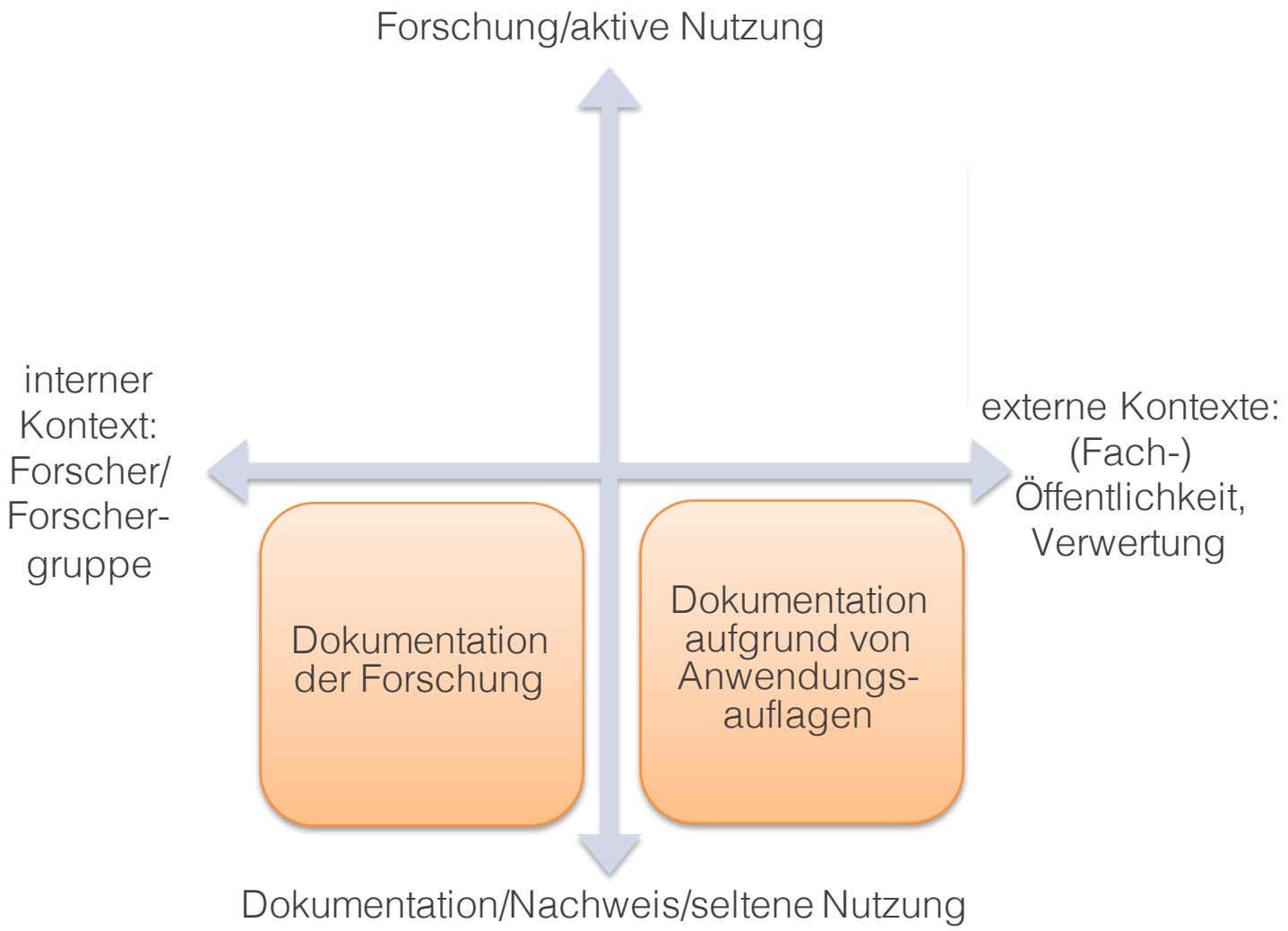
Formen des Forschungsdaten-Managements



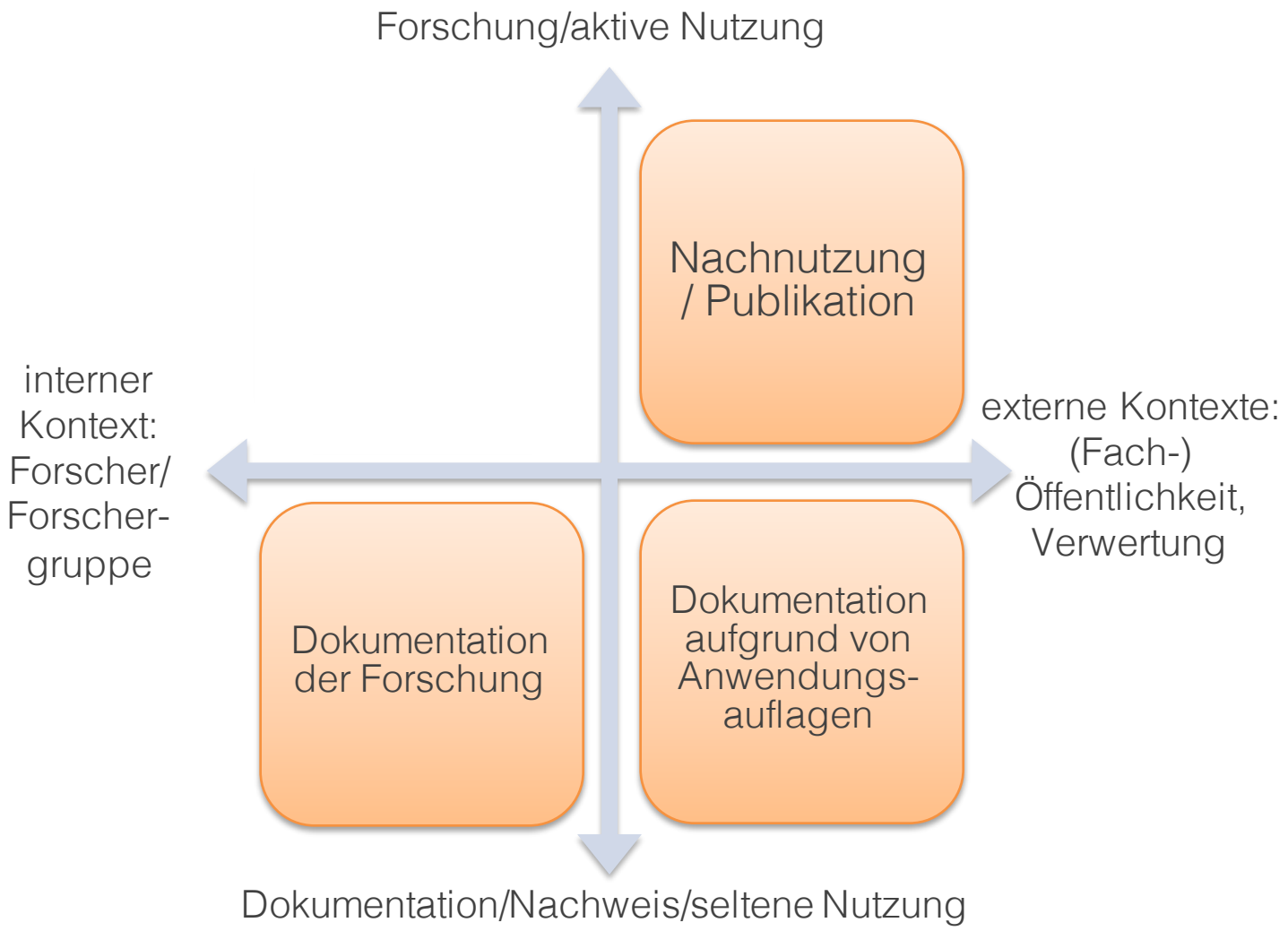
Formen des Forschungsdaten-Managements



Formen des Forschungsdaten-Managements

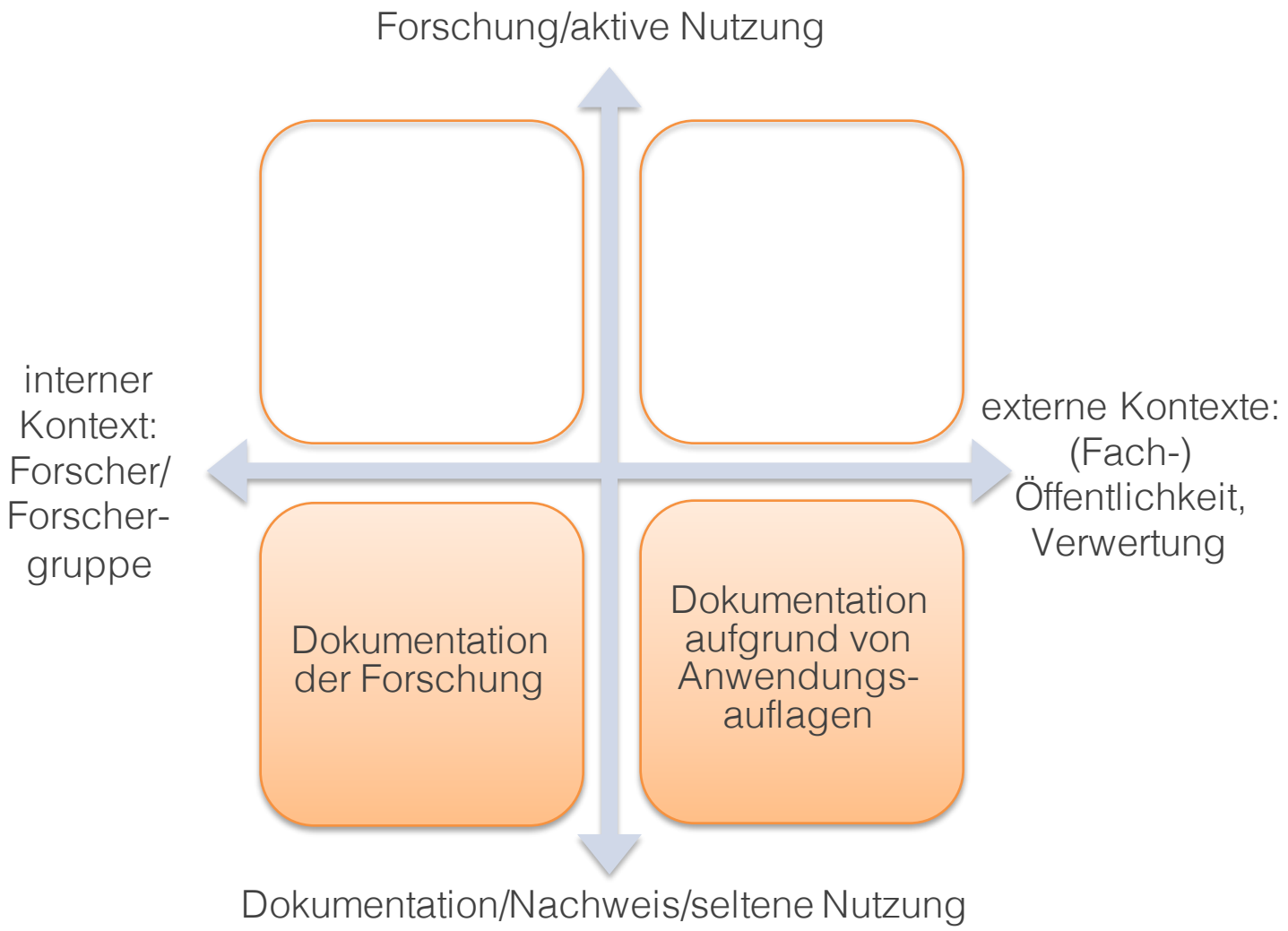


Formen des Forschungsdaten-Managements



Formen des Forschungsdaten-Managements

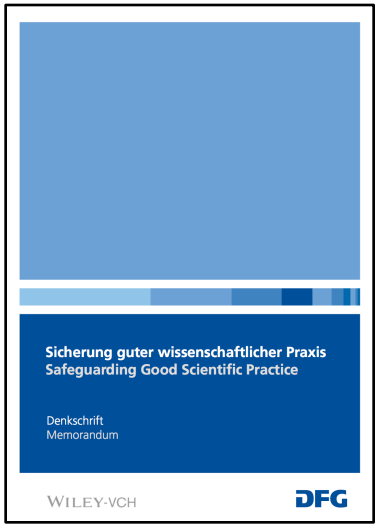
Service-Klasse: Dokumentation



DFG, Sicherung guter wissenschaftlicher Praxis, 2013

„Ein in der Öffentlichkeit im In- und Ausland breit diskutierter Fall **wissenschaftlichen Fehlverhaltens** hat das Präsidium der Deutschen Forschungsgemeinschaft (DFG) veranlasst, eine international zusammengesetzte Kommission unter Vorsitz des Präsidenten zu berufen […]" (Vorwort zur ersten Auflage, 1997)

„Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für zehn Jahre aufbewahrt werden."



Service-Klasse: Dokumentation

Ziel: Nachvollziehbarkeit für Verantwortungszwecke

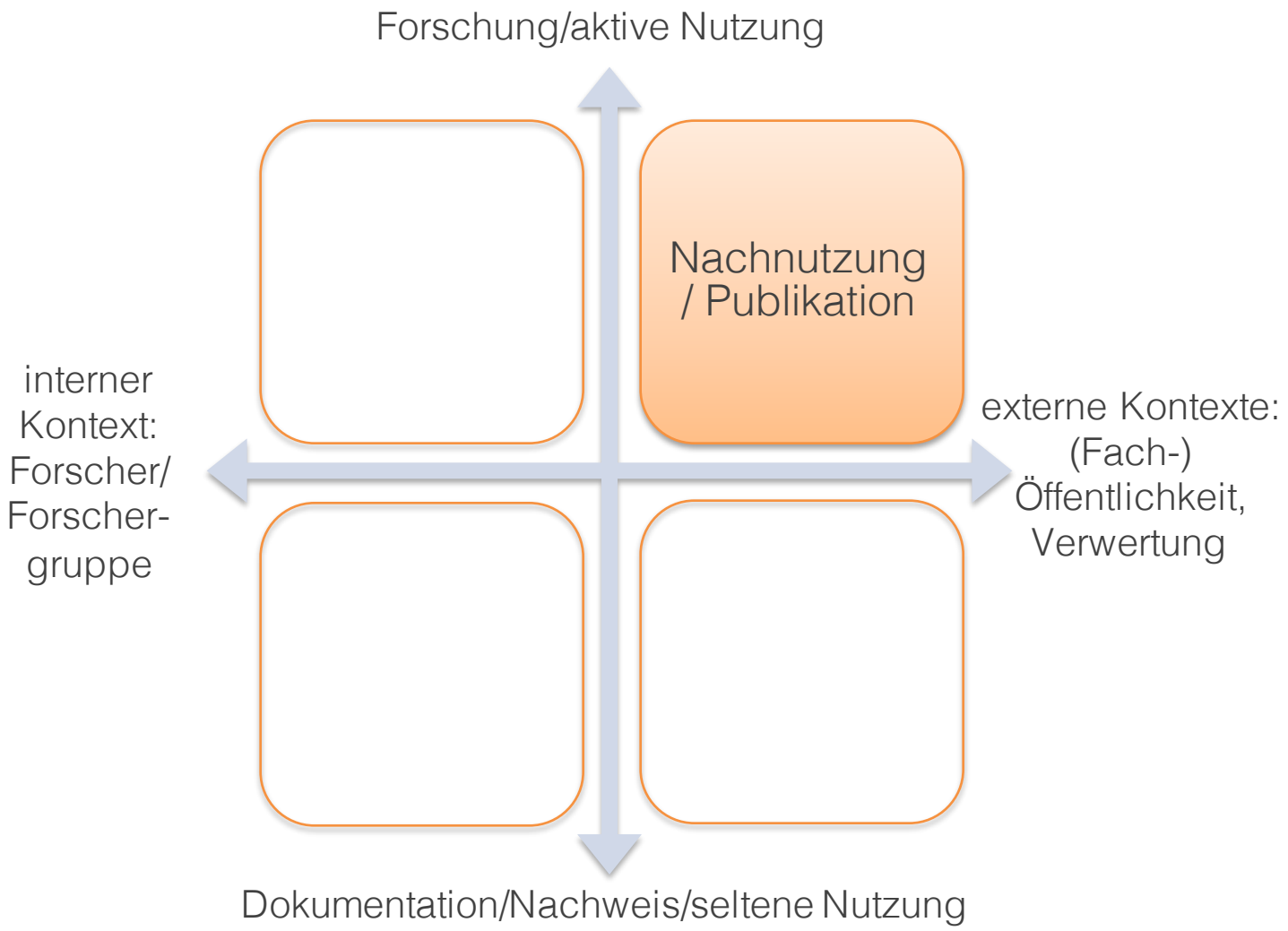
Zielgruppe: Institutionen (weniger Wissenschaftler)

Hohes Volumen, sehr seltene Datenanfragen, begrenzte Dauer

- einfacher und schneller Zugriff wird Kostenersparnis geopfert
- voraussichtlich Bitstream Preservation + basale Metadaten + Hard-/Software-Museum

Aufwand und Kosten für Dokumentation werden mit Kosten des Verantwortungsfalls abgewogen

Service-Klasse: Nachnutzung



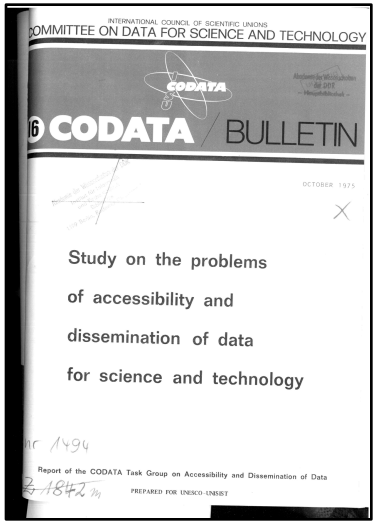
Die Relevanz von Nachnutzung

„One of the reasons for the ever-increasing efficiency of sciene and technology is that, whenever a new venture is undertaken, the results of past efforts are systematically retrieved and utilized. Such useful and important results are to be found, in concentrated form, in compilations of reliable data.“

Die Relevanz von Nachnutzung

„One of the reasons for the ever-increasing efficiency of sciene and technology is that, whenever a new venture is undertaken, the results of past efforts are systematically retrieved and utilized. Such useful and important results are to be found, in concentrated form, in compilations of reliable data.“

CODATA, „Study on the problems of accessibility and dissemination of data for science and technology“, **1975**



DFG-Anträge seit 2010

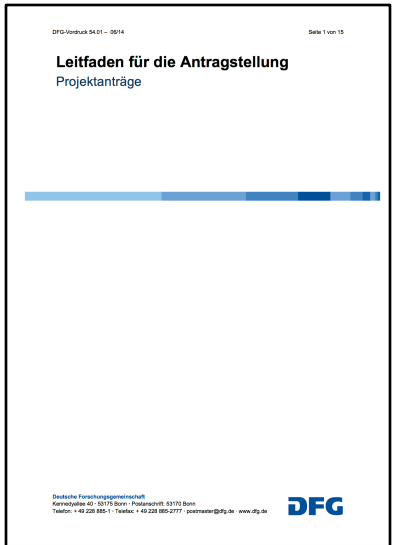
„2.4 Umgang mit den im Projekt erzielten Forschungsdaten

[...] Die DFG ist daher bestrebt, durch ihre Förderung auch zur Sicherung, Aufbewahrung und Nachnutzbarkeit von Forschungsdaten beizutragen.

Wenn aus Projektmitteln systematisch Forschungsdaten oder Informationen gewonnen werden, die für die Nachnutzung durch andere Wissenschaftlerinnen und Wissenschaftler geeignet sind, legen Sie bitte dar, ob und auf welche Weise diese für andere zur Verfügung gestellt werden. Bitte berücksichtigen Sie dabei auch - sofern vorhanden - die in Ihrer Fachdisziplin existierenden Standards und die Angebote existierender Datenrepositorien oder Archive.

[...]

Die für die Nachnutzung der Forschungsdaten anfallenden projektspezifischen Kosten können Sie im Rahmen des Projekts beantragen. Stellen Sie in diesem Fall bitte auch dar, in welcher Form eine Unterstützung beim Daten- und Informationsmanagement durch die am Projekt beteiligten Institutionen geleistet wird.” Leitfaden für die Antragstellung, Version 06/2014


DFG-Vordruck 54.01 – 06/14 Seite 1 von 15

Leitfaden für die Antragstellung

Projektanträge

Deutsche Forschungsgemeinschaft
Kennedyallee 40 · 53175 Bonn · Postanschrift: 53170 Bonn
Telefon: + 49 228 885-1 · Telefax: + 49 228 885-2777 · postmaster@dfg.de · www.dfg.de

DFG

DFG-Anträge seit 2010

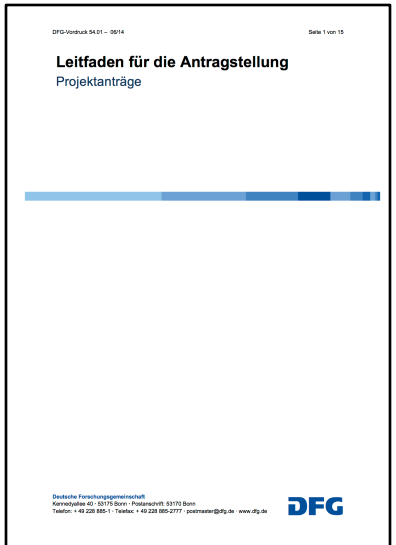
„2.4 Umgang mit den im Projekt erzielten Forschungsdaten

[...] Die DFG ist daher bestrebt, durch ihre Förderung auch zur Sicherung, Aufbewahrung und **Nachnutzbarkeit von Forschungsdaten** beizutragen.

Wenn aus Projektmitteln systematisch Forschungsdaten oder Informationen gewonnen werden, die für die Nachnutzung durch andere Wissenschaftlerinnen und Wissenschaftler geeignet sind, legen Sie bitte dar, ob und auf welche Weise diese für andere zur Verfügung gestellt werden. **Bitte berücksichtigen Sie dabei auch - sofern vorhanden - die in Ihrer Fachdisziplin existierenden Standards und die Angebote existierender Datenrepositorien oder Archive.**

[...]

Die für die Nachnutzung der Forschungsdaten anfallenden projektspezifischen **Kosten können Sie im Rahmen des Projekts beantragen**. Stellen Sie in diesem Fall bitte auch dar, in welcher Form eine Unterstützung beim Daten- und Informationsmanagement durch die am Projekt beteiligten Institutionen geleistet wird." Leitfaden für die Antragstellung, Version 06/2014


DFG-Vordruck 54.01 – 06/14 Seite 1 von 15

Leitfaden für die Antragstellung

Projektanträge

Deutsche Forschungsgemeinschaft
Kennedyallee 40 · 53175 Bonn · Postanschrift: 53170 Bonn
Telefon: + 49 228 885-1 · Telefax: + 49 228 885-2777 · postmaster@dfg.de · www.dfg.de

DFG

Service-Klasse: Nachnutzung

Ziele:

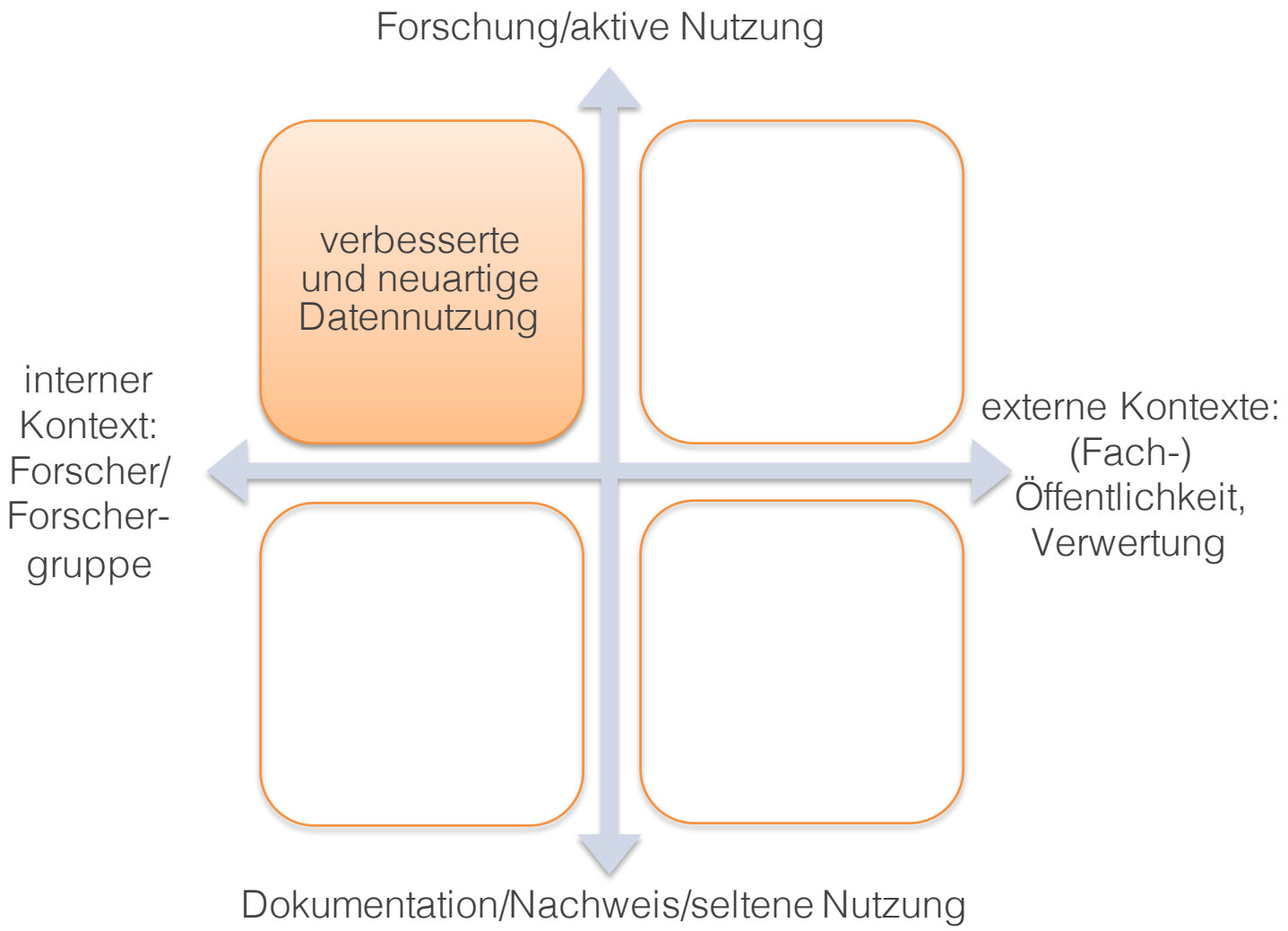
- zitierfähige Datenpublikation
- erneute wissenschaftliche Nutzung von Daten
- Förderersicht: erhöhte Effizienz
- Bewahrung nicht reproduzierbarer Daten

Zielgruppe: Fach-Communities

Daten werden selten benutzt, aber ohne klares Enddatum

Geringeres Datenvolumen, da Aufwand pro Datensatz hoch

Service-Klasse: eResearch, eScience, ...



Wired 2007

(fünf Jahre bevor „Big Data“ in Mode kam)

WIRED MAGAZINE: ISSUE 16.07, http://www.wired.com/wired/issue/16-07

Microsoft 2009

The

FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

Service-Klasse: eResearch, eScience, ...

Ziele:

- Erleichterung und Absicherung der Datennutzung
- Ermöglichung neuer Methoden/Analysen

Zielgruppe: wissenschaftliche Arbeitsgruppen, z.B. SFBs

Daten werden ständig benutzt und verändern sich, Aufbewahrungszeit durch Projektlaufzeit definiert

Verknüpfung mit kollaborativen Forschungsumgebungen und Werkzeugen

Was sind die eigentlich wichtigen offenen Fragen?

Ein Antwortversuch

Empfehlung des CODATA-Berichts 1975

1. mehr Förderung und Anerkennung durch Regierungen und Wissenschaftsorganisationen
2. Dreistufige Organisationsstruktur
    1. data evaluation centres
    2. (local) data dissemination centres
    3. a single data referral centre
3. bessere Verzeichnisse
4. Verbesserung von Standardisierung und Kompatibilität der Daten und Prozesse
5. besondere Maßnahmen zur Verbesserung der Datenverfügbarkeit in Entwicklungsländern

Was hat sich geändert?

Dissemination centres scheinen heute nicht mehr notwendig. Das Internet ermöglicht uns eine andere Organisationsstruktur.

„[T]he principal data compilations were extremely limited in number in and before the first half of this century [...] In the third quarter of this century, however, the amount of data in science and technology grew rapidly [...]“ Diese Datenmengen sind heute trivial, aber das Phänomen bleibt.

Was hat sich nicht geändert: Mehr Förderbedarf.

Das sind Hinweise auf das eigentlich Wichtige:

- Wir können weder technisch noch ökonomisch alle Daten bewahren.
- Um das bestmögliche Optimum zu bewahren, brauchen wir eine rationale Planung und Kooperation.

DOI:10.1145/1839676.1839692

Article development led by acmqueue
queue.acm.org

As storage systems grow larger and larger, protecting their data for long-term storage is becoming ever more challenging.

BY DAVID S.H. ROSENTHAL

Keeping Bits Safe: How Hard Can It Be?

„The requirement of a 50% chance that a petabyte will survive for a century translates into a bit half-life of 8×10^{17} years. The current estimate of the age of the universe is 1.4×10^{10} years [...]“

„We can place an upper bound on the bit half-life in [CERNs] storage systems [...] the result is 2×10^{8} [...]“

David Rosenthal, Keeping Bits Safe: How Hard Can It Be? Communications of the ACM, 2010, Vol. 53, No. 11

Daten- und Kostentrends

Verschiedene Quellen, mit Vorsicht zu genießen, aber Trend scheint klar:

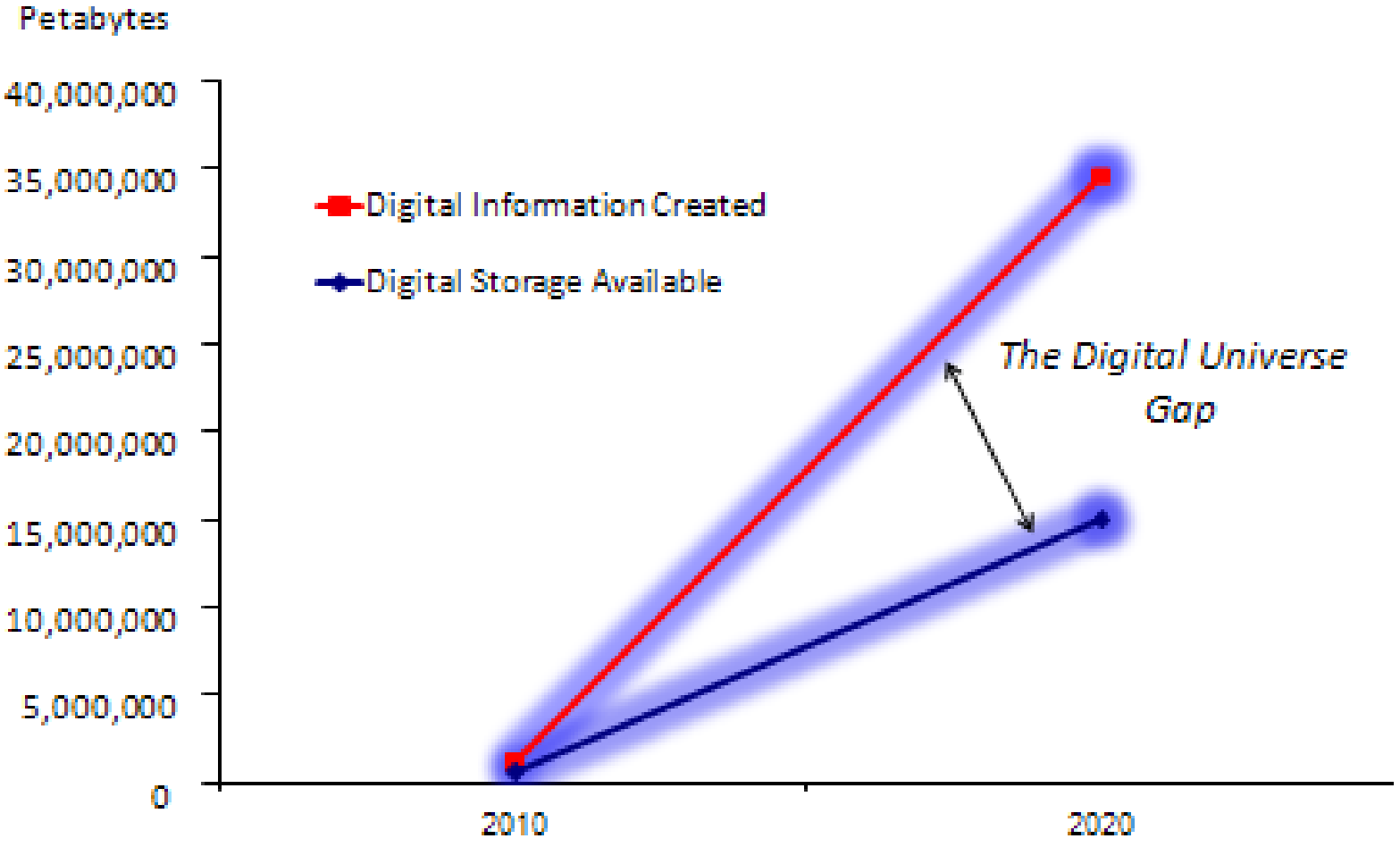
- Speicherplatzbedarf steigt jedes Jahr um 60%
- Die Speicherdichte wird nicht mehr als 20% pro Jahr steigen für die nächsten fünf Jahre. (Speicherdichte und Kosten sind eng verwandt.)
- IT-Budgets steigen ... quasi nicht.

http://blog.dshr.org/2014/11/talk-costs-why-do-we-care.html

Der Speicherplatz ist aber nur überlicherweise nur ein Bruchteil der Gesamtkosten des Datenmanagement.

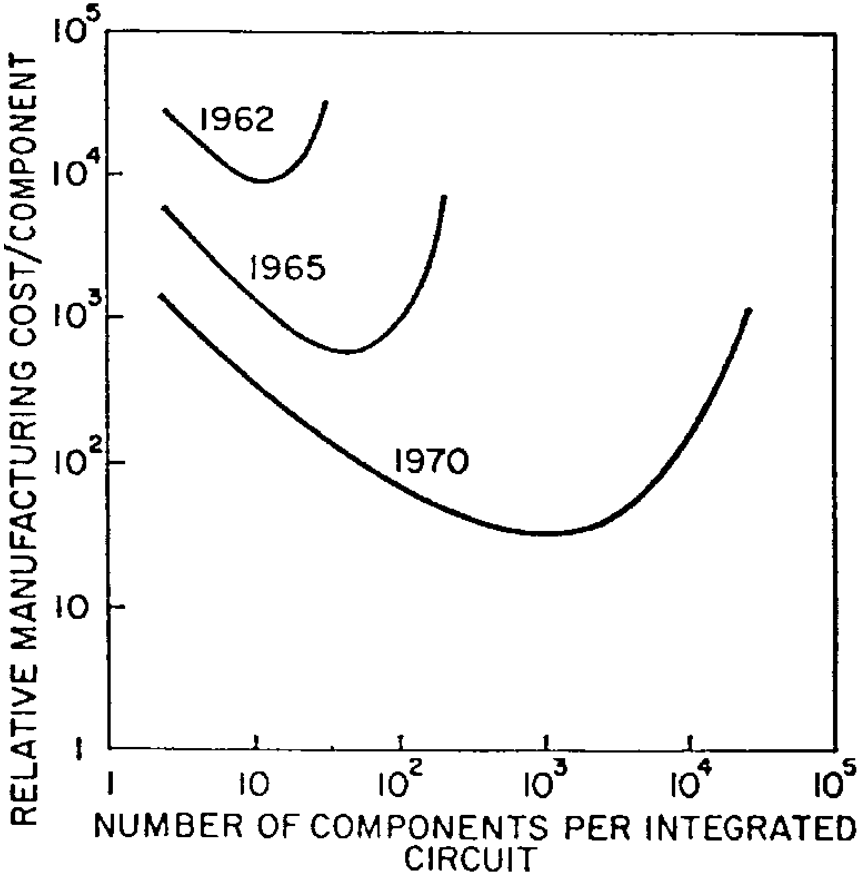
http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf

Information Creation > Storage Available



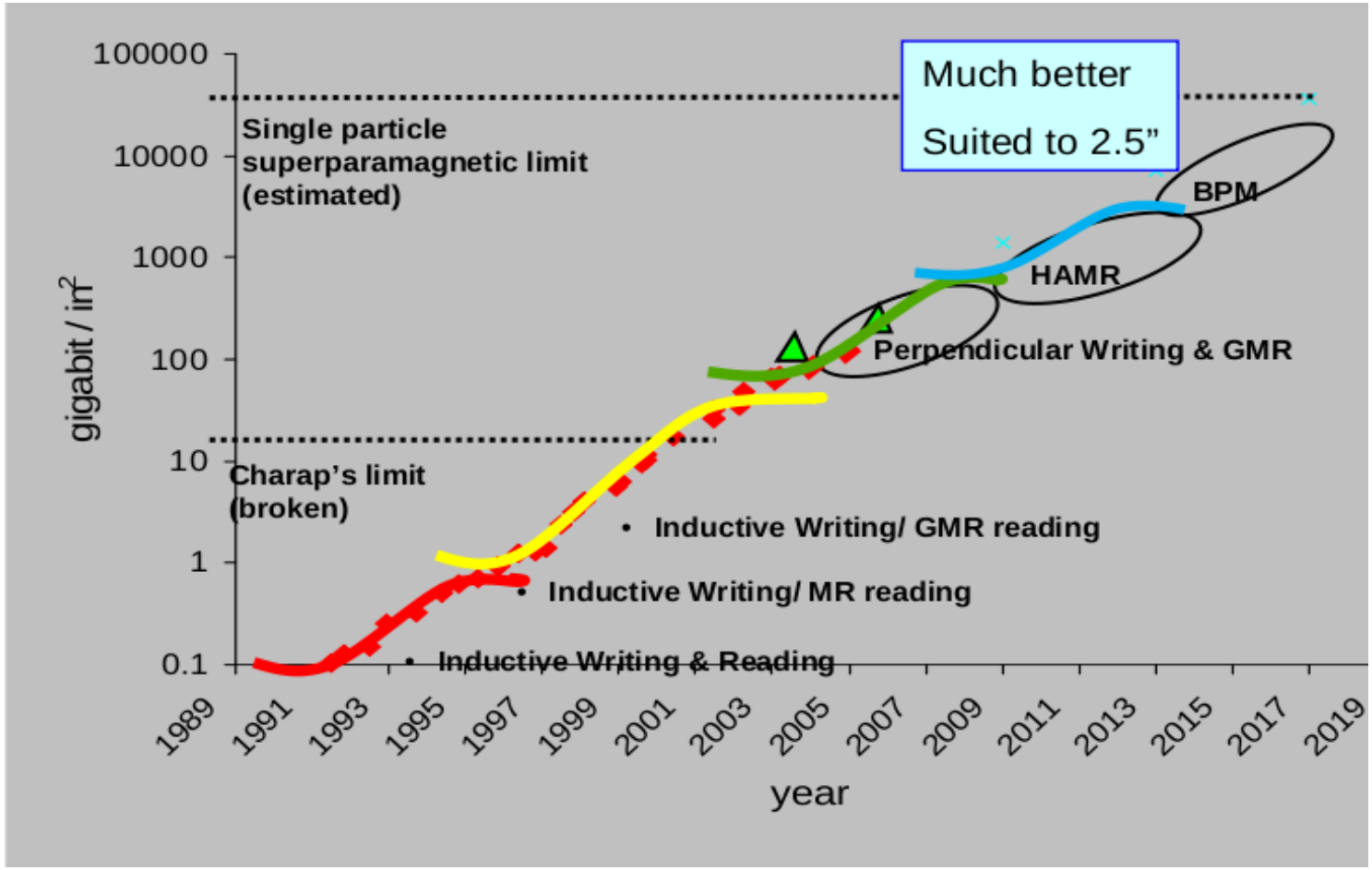
Source: IDC Digital Universe Study, sponsored by EMC, May 2010

Moore's bzw. Kryder's Law kein Ausweg: 1. Nicht technisch-physikalisches, sondern technisch-ökonomisches Gesetz



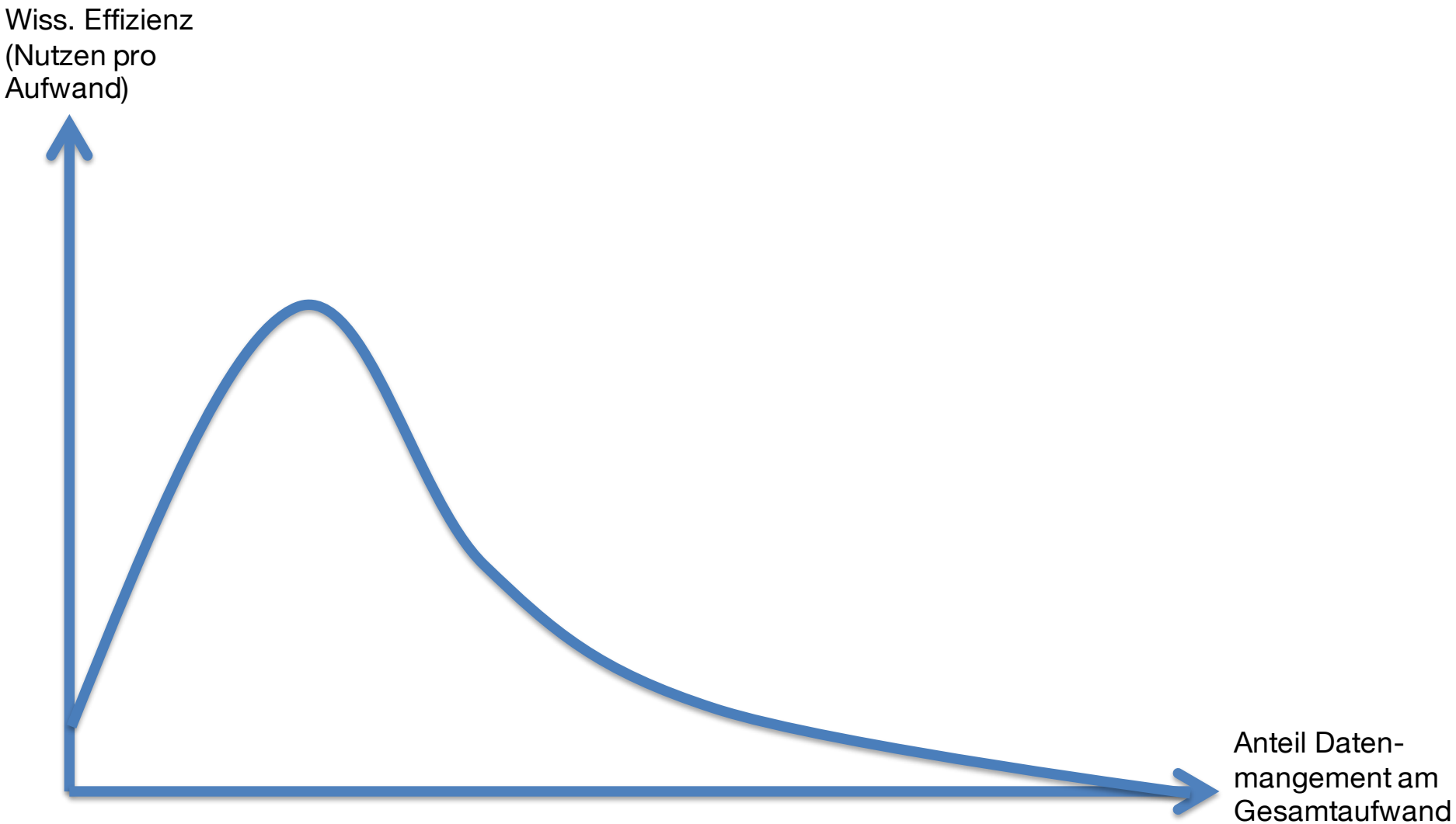
Gordon E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics,* pp. 114–117, April 19, 1965

Moore's bzw. Kryder's Law kein Ausweg: 2. Das technisch-ökonomische Gesetz beruht auf vorübergehend steigerbaren Technologien



Dave Anderson, in: David Rosenthal, Talk at Seagate, http://blog.dshr.org/2014/05/talk-at-seagate.html

In der Forschungsdaten-Diskussion implizit: Nicht nur Kosten, sondern Investiton und Nutzen



Eine Konsequenz: Auswahl und Bewertung von Daten

- Nicht alles kann aufbewahrt werden, sowohl technisch als auch ökonomisch.
  - Bedauerlich, aber stellt nicht das Gesamtvorhaben in Frage. (Bibliotheken werden seit sicherlich 3000 Jahren zerstört.)

- Es müssen auch nicht alle Daten aufbewahrt werden.
  - Manche Daten aber schon.

- Scheindebatte: Wer entscheidet?
  - Meist von Besitz- und Authoritätsansprüchen überschattet.
  - Vorüberlegungen zu einem rationalen Auswahlprozess klären Vieles.

Beispielfragen im Auswahl- und Bewertungsprozess
(sinnvolle Vorabklärung z.T. in Datenmanagementplänen)

1. Determine the obligations: legal and contractual obligations
    1. Is it person related data?
    2. How confidential is the data and for how long?
    3. How long is the rentention period?

2. Determine the value: intellectual evaluation
    1. Is the data valuable for reuse?
    2. Where is the best place for the data?

3. Determine the feasibility: technology and knowledge
    1. What technical tools and background knowledge do users/researches need to use the data?
    2. Can it be foreseen that these or better tools can be offered for the long-term?
    3. Will the data be sufficiently documented?

Eine andere Konsequenz: Kooperation

Wissenschaft ist erstmal kompetetiv. Kooperation z.B. aufgrund von:

- Aufwand der Datenerhebung/Instrumente
- Größe des Untersuchungs-gegenstand
- Interdisziplinärität der Fragestellung, Spezialisierung der Fachexpertise

Kooperationsgründe im Forschungsdatenmanagement

Technik

- technischer Aufwand zu groß (aber economy of scale?)
- technische Expertise zu divers

Disziplinanforderungen

- disziplinspezifische Expertise zu divers
- Disziplinverankerung notwendig

Service-Struktur

- Zielgruppe und Beratungs-/Betreuungsaufwand zu groß
- auch (kooperierende) Akteure mit sehr ähnlichem Angebot können sinnvoll sein (z.B. Ausfallsicherheit)

Weiterhin unklar: Welche der möglichen Kooperationsmodelle sind erfolgversprechend?

Arten der Aufgabenteilung:

- räumlich
- disziplinär
- funktional
- qualitativ/service level
- ...

Organisationsformen:

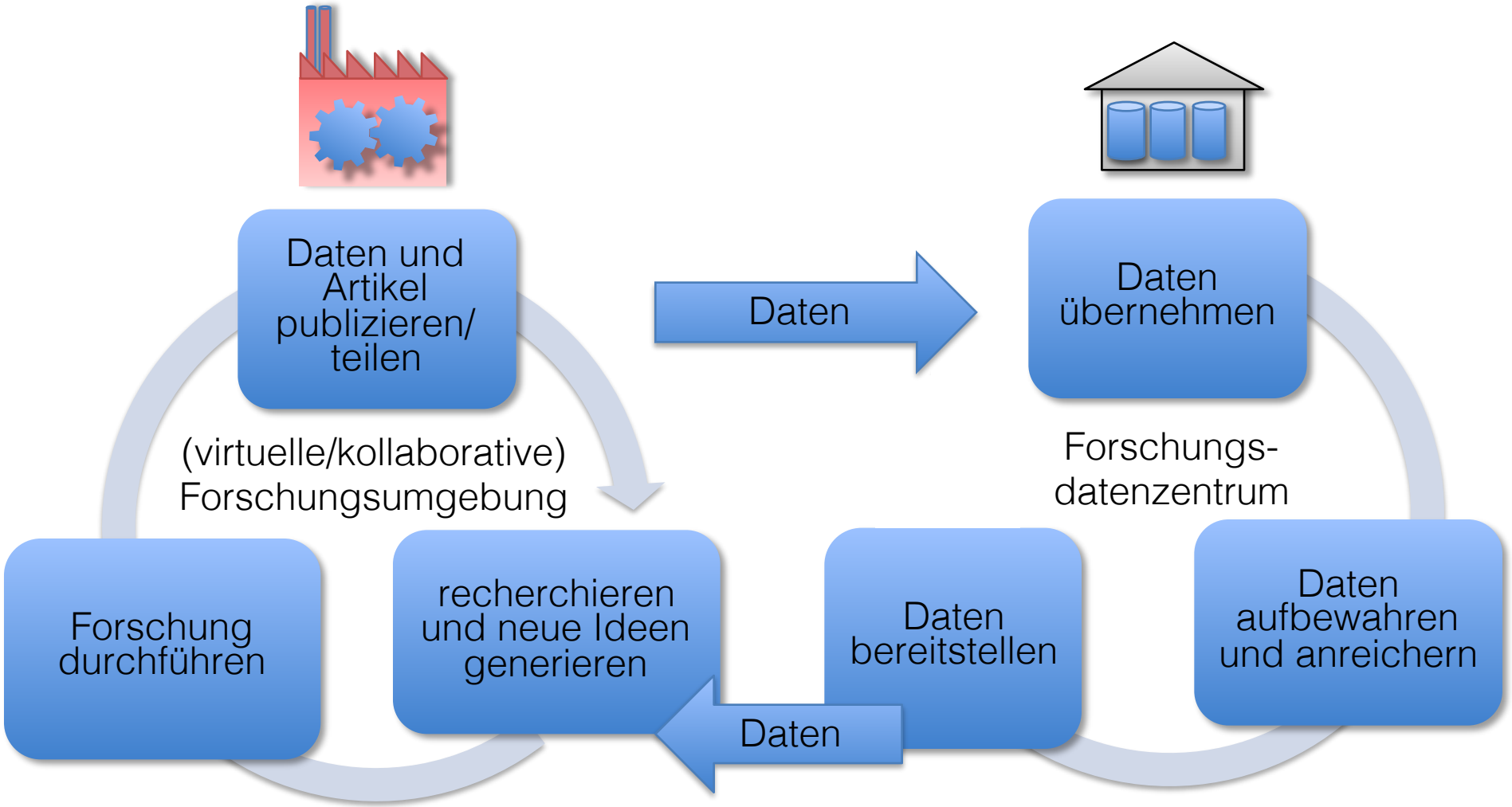
- Föderationen
- übergreifende Zentren (z.B. disziplinär)
- Institutionen für sich selbst
- Outsourcing
- Netzwerke
- ...

und Kombinationen davon.

Beispiele:

- GFBio: German Federation for the Curation of Biological Data
- WDC/WDS: World Data Centers/System
- RDA: Know-How Austausch

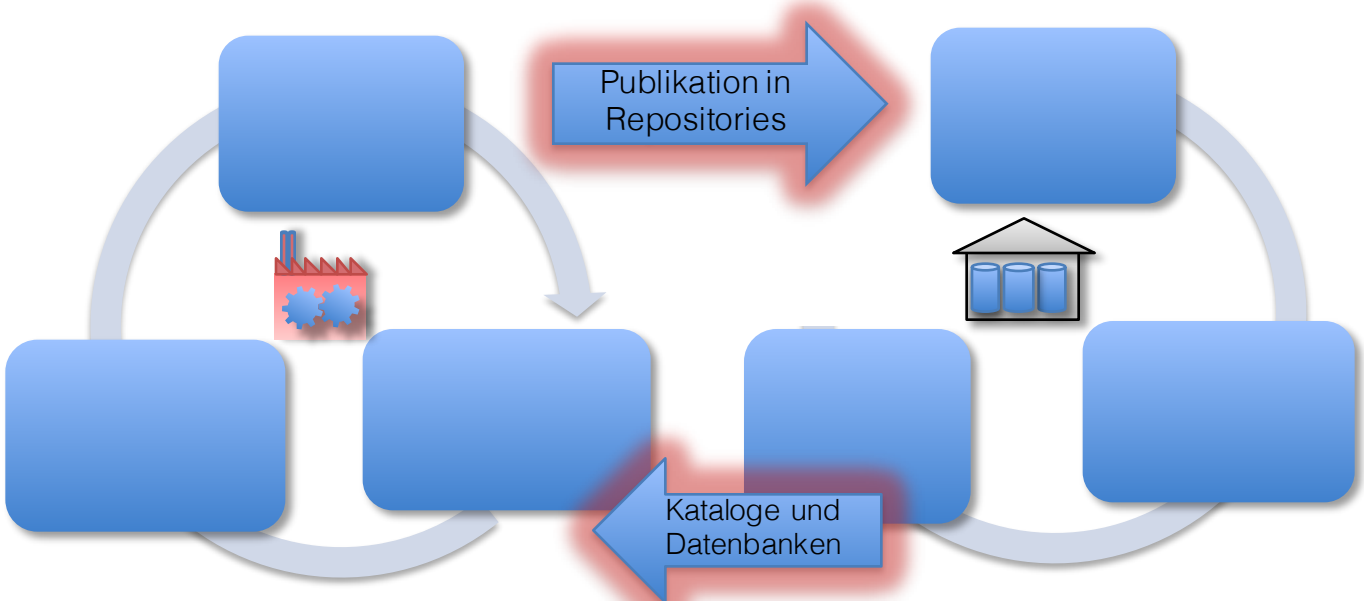
Idealisierter Forschungs- und Forschungsdatenzyklus



Häufiger Ansatz generischer Infrastruktureinrichtungen

Bibliotheken und Rechenzentren fokussieren meist auf klassisches Publikations- und Archivparadigma:

- Wie können wir unsere Speichermöglichkeiten für Forschungsdaten anbieten?
- Wie können wir Forschungsdaten in die bestehenden Recherche-Instrumenten einfügen?



Eignung unterschiedlicher Akteure für unterschiedliche Aufgaben

Zuschnitt der Aufgabenteilung unklar:

- fachspezifische Aufgaben → spezialisierte Einzellösungen/Zentren?
- generische Aufgaben → lokale disziplinübergreifende Infrastruktur? Oder zentrale Infrastruktur, da effizienter?
- Was zählt als Disziplin? Ab wann sind fachspezifische Angebote notwendig?

Dokumentation: für „lokale“ Verantwortung auch lokale Infrastruktur?

Nachnutzung/Publikation: für Fachöffentlichkeit in den Disziplinen verankerte Zentren notwendig?

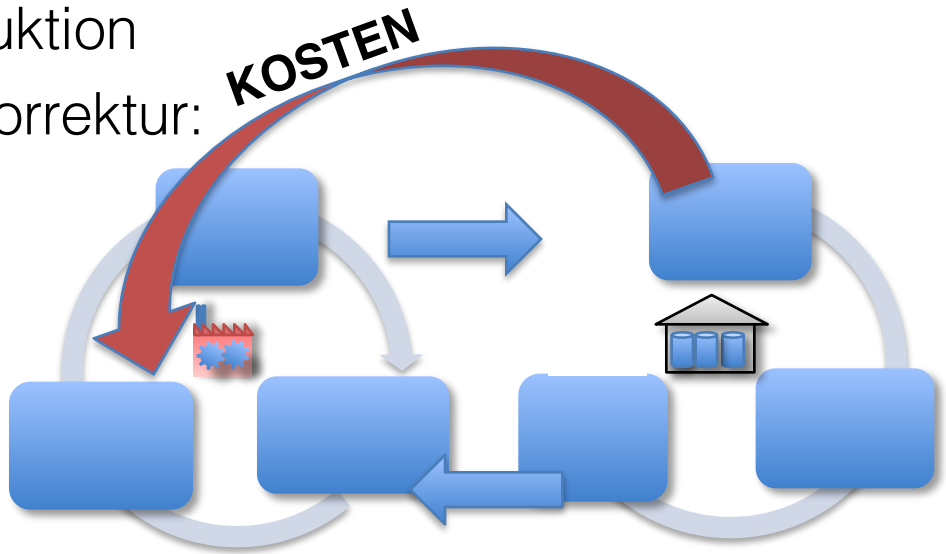
in Projekten: lokale/projekt-interne Teams (zuviele Projekte für ein Fachzentrum und lokale Infrastruktur nicht forschungsnah genug)?

Eine ökonomische Kooperationsstrategie:
Kosten für Nachnutzung in Projektphase verlagern

- Einige Zentren investieren in Werkzeuge, die Wissenschaftlern während des Projekts und später ihnen die Arbeit erleichtern.
- Lokale Teams (z.B. „embedded data managers“ wie INF-Projekte in SFBs) übernehmen Support.

Nutzen:

1. mögliche Gesamtkostenreduktion
2. hoffentlich Wahrnehmungskorrektur: Datenkosten sind Forschungskosten
   - 2.1 höhere Akzeptanz
   - 2.2 Skalierung der Finanzierung mit Projekten



Vielen Dank!