Kompetenznetzwerk
Langzeitarchivierung
- Qualifizierung -

Was sind Forschungsdaten und was bedeutet es, sie zu managen oder zu archivieren?

Jens Ludwig

nestor/DINI AG Forschungsdaten &
Staatsbibliothek zu Berlin - Stiftung
Preußischer Kulturbesitz

Inhaltsübersicht

- Was sind Forschungsdaten?
- Was ist Langzeitarchivierung?
- Forschungsdatenmanagement und Unterschiede zur Langzeitarchivierung
- Einige Überlegungen zur Informationsinfrastruktur

Inhaltsübersicht

- Was sind Forschungsdaten?
- Was ist Langzeitarchivierung?
- Forschungsdatenmanagement und Unterschiede zur Langzeitarchivierung
- Einige Überlegungen zur Informationsinfrastruktur

Beispiele für Forschungsdaten

Statistiken, Interviews, Simulationen, Messdaten aus Experimenten, Beobachtungsdaten aus Instrumenten, Text mit semantischen Annotationen, 3D-Scans, Video, Audio, Bilder, Tabellen, Dokumente, Binärdaten, Software, Textdateien, ...

Der erste Eindruck ist, dass fast alles als Forschungsdaten angesehen werden kann.

Abgrenzung von analogen Forschungsdaten

- Auch mit Bleistift und Papier notierte Messdaten sind eigentlich Forschungsdaten.
- Analoge Forschungsdaten werden im Folgenden ignoriert.
- Viele Fragestellungen des Forschungsdatenmanagements stellen sich nicht in gleicher Form und im gleichen Maße für analoge Forschungsdaten.
- Hauptgrund: Massenproduktion und -verarbeitung von Forschungsdaten ist digital effizienter, ermöglicht neue Verfahren und ist der neue Standard.

Was ist ein digitales Objekt? (nach Thibodeau 2002)

conceptual object:

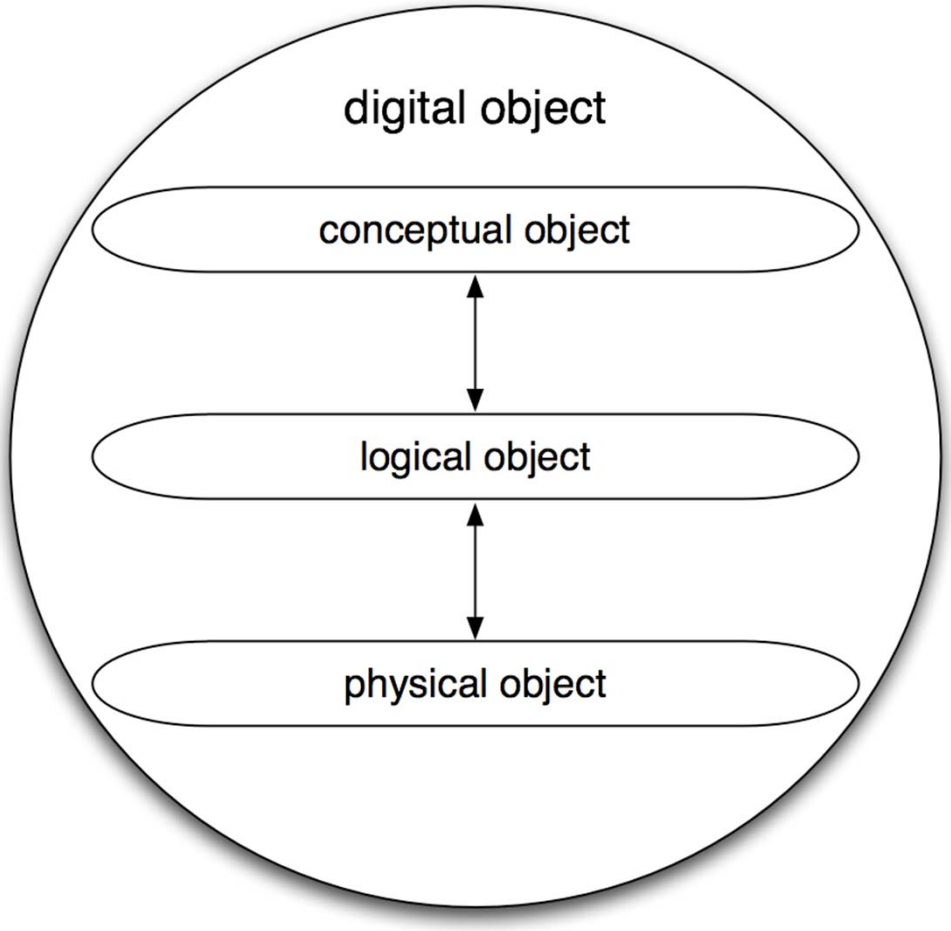
für Menschen bedeutsames Objekt

logical object:

von Software und Technik verstehbares Objekt

physical object:

materielle Zeichen und Trägermedium



Beispiel: Ebenen eines digitalen Buchs

conceptual object:

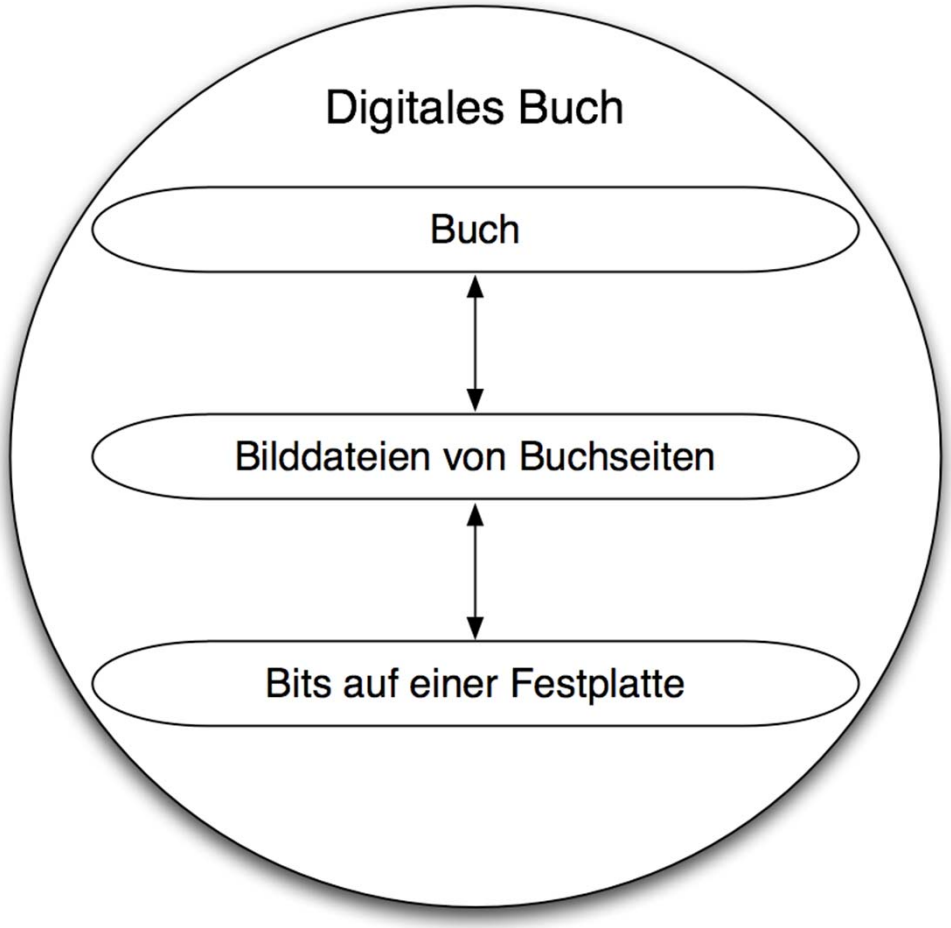
Buch mit z.B. navigierbaren Verzeichnissen, Abbildungen, ...

logical object:

Dateien in verschiedenen Formaten (TIFF, TXT, ...)

physical object:

Bits auf einer Festplatte, die ein bestimmtes Dateisystem hat



Definitionsversuch 1

- Forschungsdaten sind einfach alle Daten, die im Forschungsprozess benutzt werden.

- Das wird dem Phänomen gerecht, dass ein und dieselben digitalen Daten je nach Kontext Forschungsdaten sein können oder nicht. Z.B. Urlaubsfoto eines Denkmals in der privaten Fotosammlung versus Foto als Teil eines Trainingsdatensatzes für Bilderkennungsverfahren.

- Es werden aber viele Dinge als Forschungsdaten erfasst, die nicht als Forschungsdaten bezeichnet werden sollten. Z.B. Literatur in Form von PDF-Dateien, die von Forscherinnen und Forscher während des Forschungsprozesses konsultiert wird.
  - Natürlich können digitale Texte auch Forschungsdaten sein, wenn sie z.B. Teil eines Textkorpus sind, der in der Computerlinguistik benutzt wird.
  - Forschungsdaten können auch in Texten enthalten sein, z.B. chemische Strukturformeln in Fachartikeln.

Zwei Arten von Forschungsdatenbezeichnungen und Definitionsversuch 2

- Beobachtung: Wenn man FachvertreterInnen befragt, was Forschungsdaten in ihrer Disziplin sind, erhält man grob zwei Gruppen von Antworten (Ludwig, in: Neuroth et al 2012, S. 299ff).
  1. Forschungsdaten werden durch ihren Medientyp (z.B. Video, Markup, Tabellen, 3D-Modelle)
  2. oder durch die erzeugende Forschungsmethode bezeichnet (z.B. Simulationsdaten, Beobachtungsdaten, Statistikdaten).

- Definitionsversuch 2: Forschungsdaten sind alle diejenigen Daten, die durch die Anwendung einer wissenschaftlichen Methode im Forschungsprozess entstehen.
  - Ähnlich: „Forschungsdaten sind Daten, die im Zuge wissenschaftlicher Vorhaben z.B. durch Digitalisierung, Quellenforschungen, Experimente, Messungen, Erhebungen oder Befragungen entstehen.“ (http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten.html)

- Aber nicht alle Forschungsdaten entstehen in Forschungsprozessen und durch wissenschaftliche Methoden. Z.B. Behördendaten, die später für sozialwissenschaftliche Forschung benutzt werden. Oder: „Help scientists recover Arctic and worldwide weather observations made by United States ships since the mid-19th century by transcribing ships' logs.“ (http://www.oldweather.org/)

Vielfalt der Forschungsdaten und Definitionsversuch 3

- Forschungsdaten sind so vielfältig wie die Forschung selbst.
  - Es kann (fast?) alles Objekt wissenschaftlicher Untersuchung sein und es kann Forschungsdaten über fast alles geben.
  - Die Forschungsmethoden sind nicht beliebig, aber die dabei entstehenden Forschungsdaten sind sehr vielfältig.

- Forschungsdaten sind alle Daten, die mit einer wissenschaftlichen Methode über ein Forschungsobjekt erzeugt oder verarbeitet werden.

Abgrenzungsversuch

- Die Klausel „über ein Forschungsobjekt“ soll von reinen digitalen Stellvertretern des Forschungsobjekts abgrenzen.

- Beispiel: Digitalisate einer mittelalterlichen Handschrift sind Daten, aber meist nicht „über ein Forschungsobjekt“, sondern eher eine Repräsentationsform des Forschungsobjekts.
  - Das hängt natürlich von der Forschung ab. Ein Literaturwissenschaftler erfährt durch das Digitalisat vermutlich nichts Neues „über sein Forschungsobjekt”, den Text, während das für eine Handschriftenforscherin, die das materielle Objekt untersucht, anders sein kann.

Wozu überhaupt eine Definition?

- Definitionen sind ein Instrument und damit nicht zweckfrei und neutral.

- Definitionen können zwar in der Kommunikation hilfreich sein. Aber auch ohne oder mit sehr unscharfen Definitionen kann Kommunikation erfolgreich sein und werden Fortschritte gemacht.

- Wozu dient uns die Definition? Wozu dient uns der Forschungsdaten-Begriff?

Position: Forschungsdatenmanagement als der Sinn des Forschungsdatenbegriffs

- Der Definitionsvorschlag 3 ist sinnvoll, weil er hilft die neuen und bisher in klassischen Informationseinrichtungen nicht behandelten Aufgaben zu verstehen.

- Für Digitalisate, digitale Texte und Textpublikationen gibt es sehr viele etablierte Verfahren und Methoden, um sie zu managen. Diese Aufgaben sind aber sehr verschieden von vielen anderen Aufgaben des Forschungsdatenmanagements.

- Die kürzeste Definition wäre deshalb, dass Forschungsdaten die Objekte des Forschungsdatenmanagements sind.

Inhaltsübersicht

- Was sind Forschungsdaten?
- Was ist Langzeitarchivierung?
- Forschungsdatenmanagement und Unterschiede zur Langzeitarchivierung
- Einige Überlegungen zur Informationsinfrastruktur

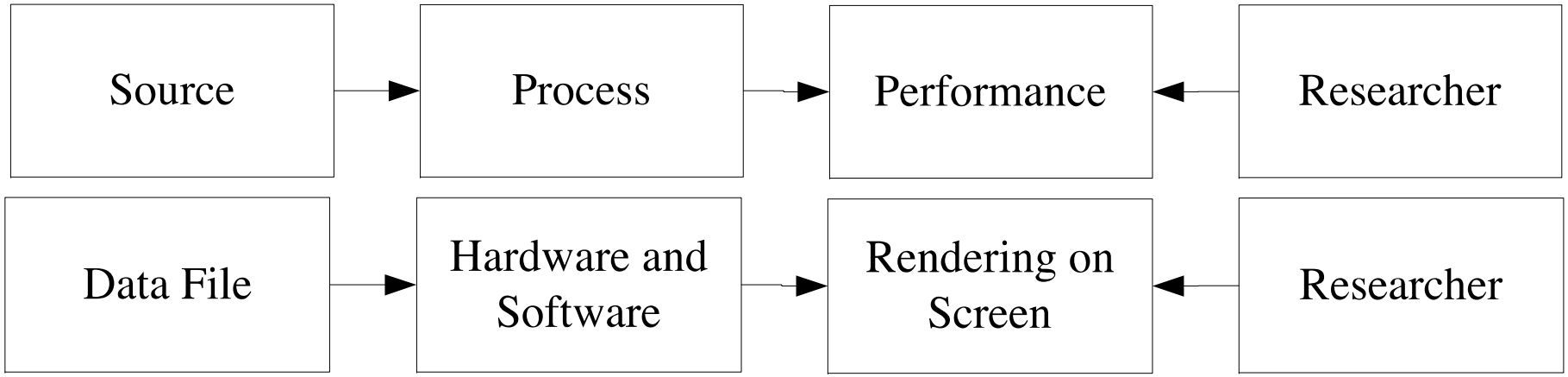
Langzeitarchivierung

„Langzeit“ ist die Umschreibung eines nicht näher fixierten Zeitraumes, währenddessen wesentliche, nicht vorhersehbare technologische und soziokulturelle Veränderungen eintreten [...]

„Archivieren“ bedeutet [...] mehr als nur die dauerhafte Speicherung digitaler Informationen auf einem Datenträger. Vielmehr schließt es die Erhaltung der dauerhaften Verfügbarkeit und damit eine Nachnutzung und Interpretierbarkeit der digitalen Ressourcen mit ein.

Aus: Hans Liegmann, Heike Neuroth, Einführung, in: nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung hg. v. H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, K. Huth urn:nbn:de:0008-2010030508

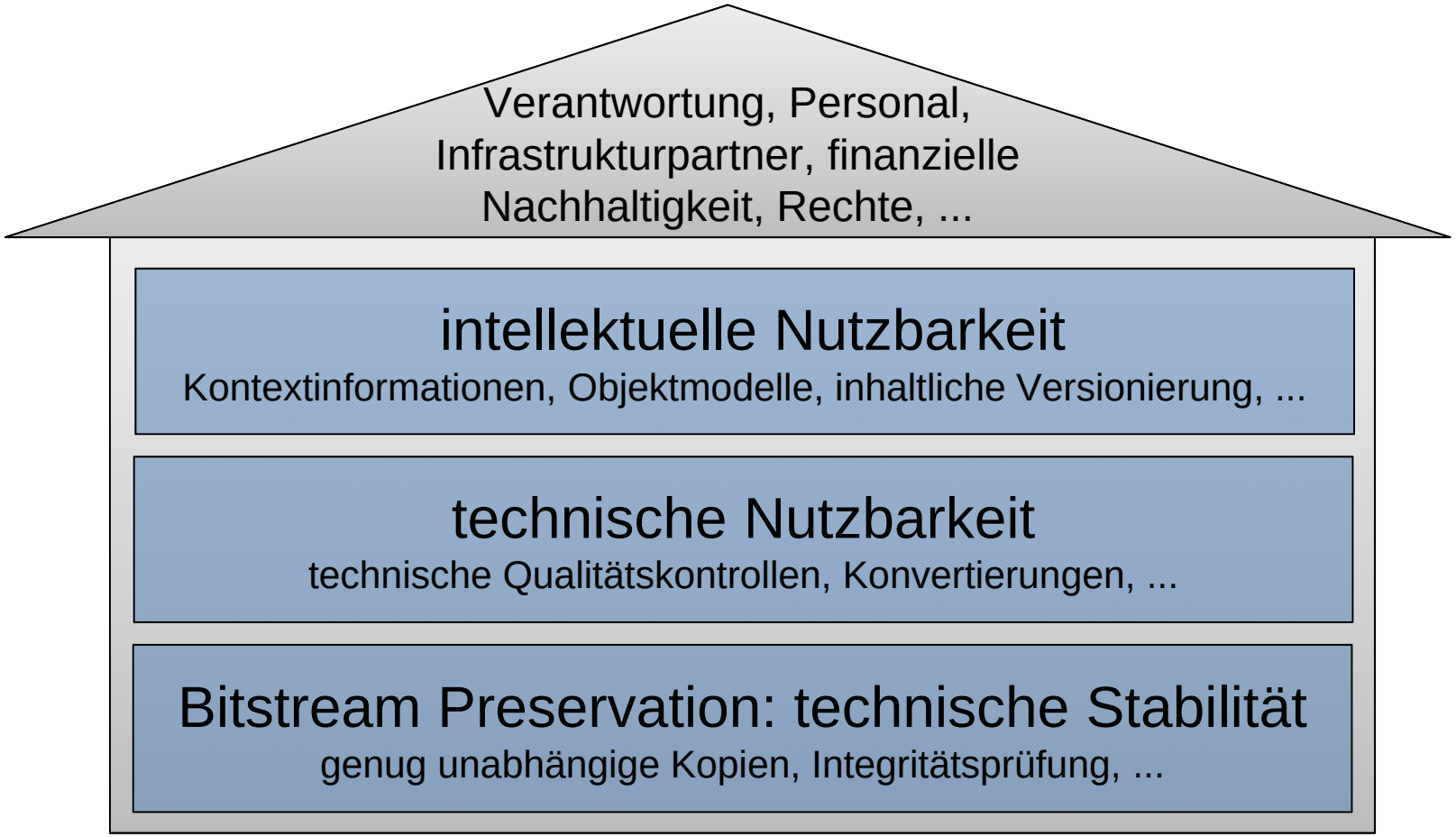
Digitale Objekte existeren nur in der Ausführung



Quelle: Heslop, Helen; Davis, Simon; Andrew Wilson: An Approach to the Preservation of Digital Records. National Archives of Australia, 2002

- Ein Grundproblem der LZA: Das gleiche digitale Objekt kann abhängig vom Kontext und der Umgebung zu ganz unterschiedlichen bzw. gar keinen Ergebnissen führen.
- Vom Drei-Ebenen-Modell digitaler Objekte von Thibodeau kann man unterschiedliche Aufgaben der LZA ableiten.

Aspekte der Langzeitarchivierung

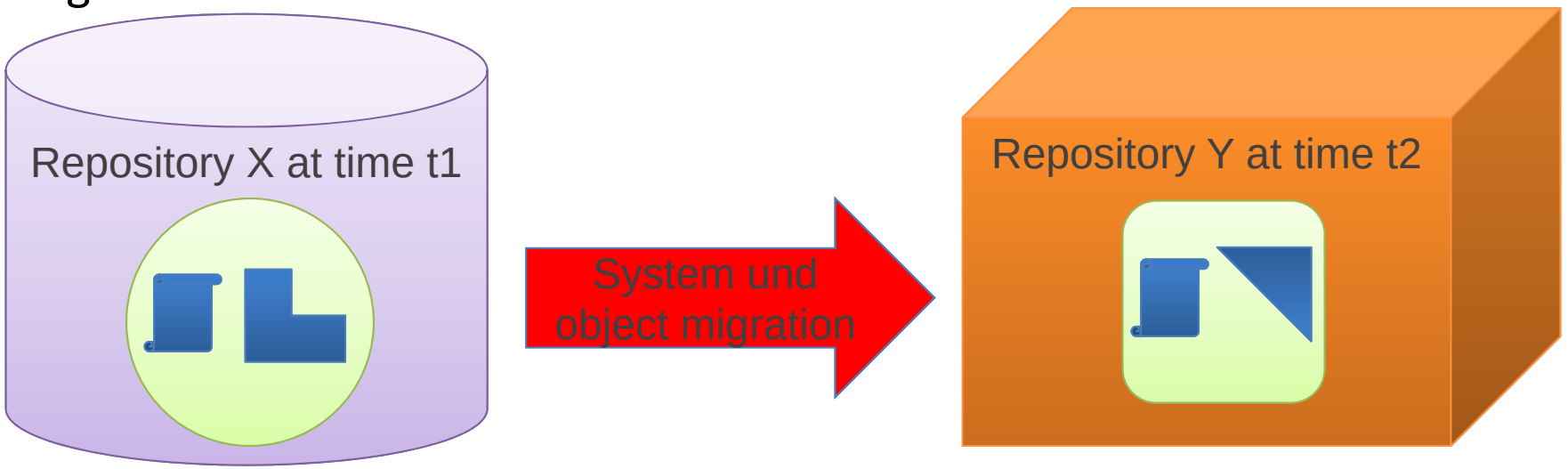


Was ist Langzeitarchivierung nicht?

„Preservation is not a Place“ (Stephen Abrams et al., doi:10.2218/ijdc.v4i1.72)

Auch Archivsysteme veralten und nicht immer ist ein separates Archivsystem zusätzlich zum existierenden Repository sinnvoll.

Deshalb: Kein technisches System macht Langzeitarchivierung, sondern Organisationen.

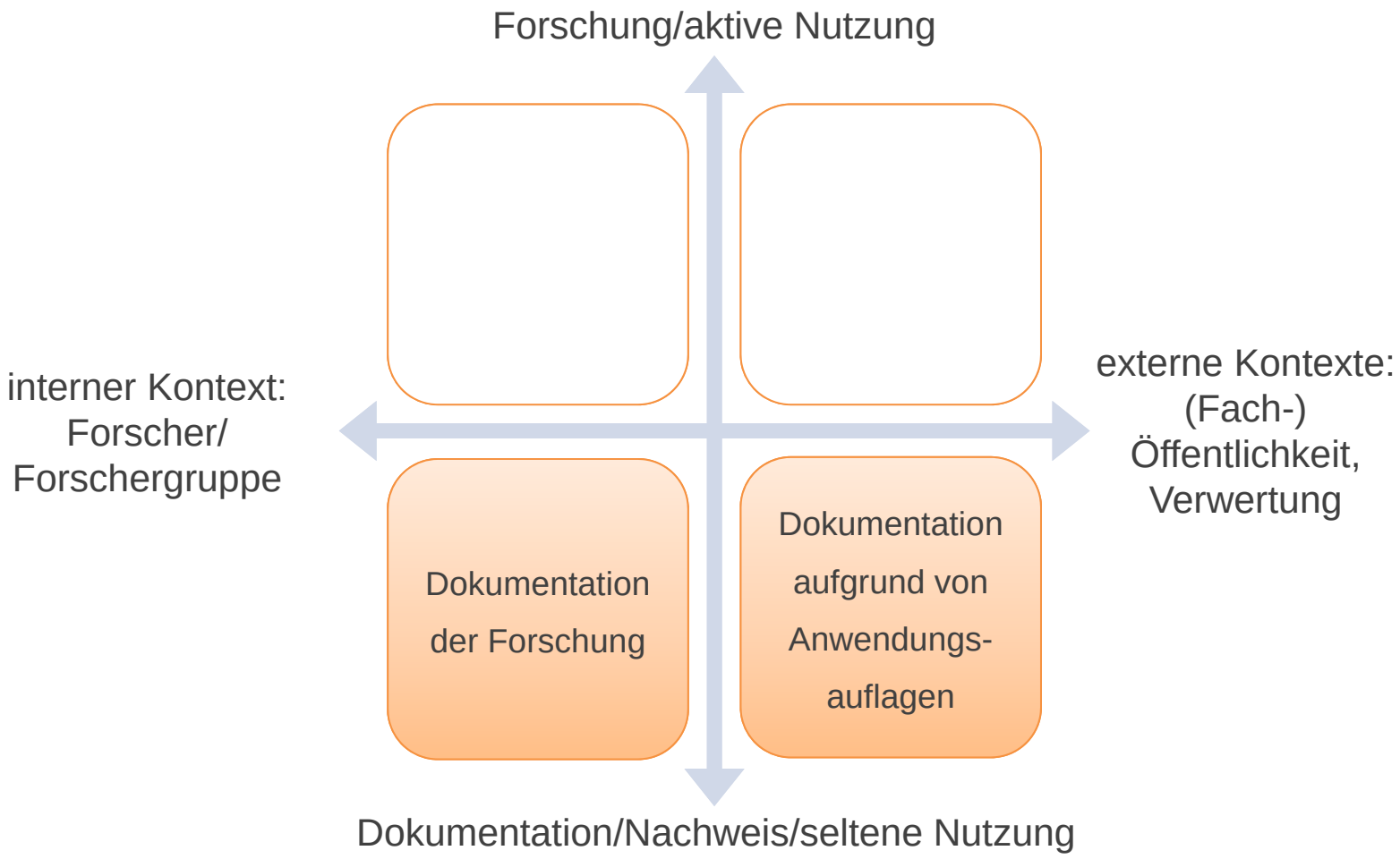


Inhaltsübersicht

- Was sind Forschungsdaten?
- Was ist Langzeitarchivierung?
- Forschungsdatenmanagement und Unterschiede zur Langzeitarchivierung
- Einige Überlegungen zur Informationsinfrastruktur

Wieso Forschungsdatenmanagement?

Service-Klasse: Dokumentation



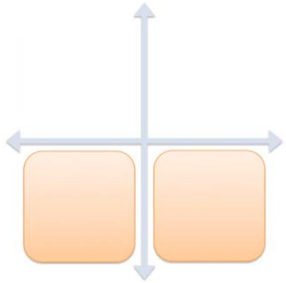
DFG, Sicherung guter wissenschaftlicher Praxis, 2013

„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)

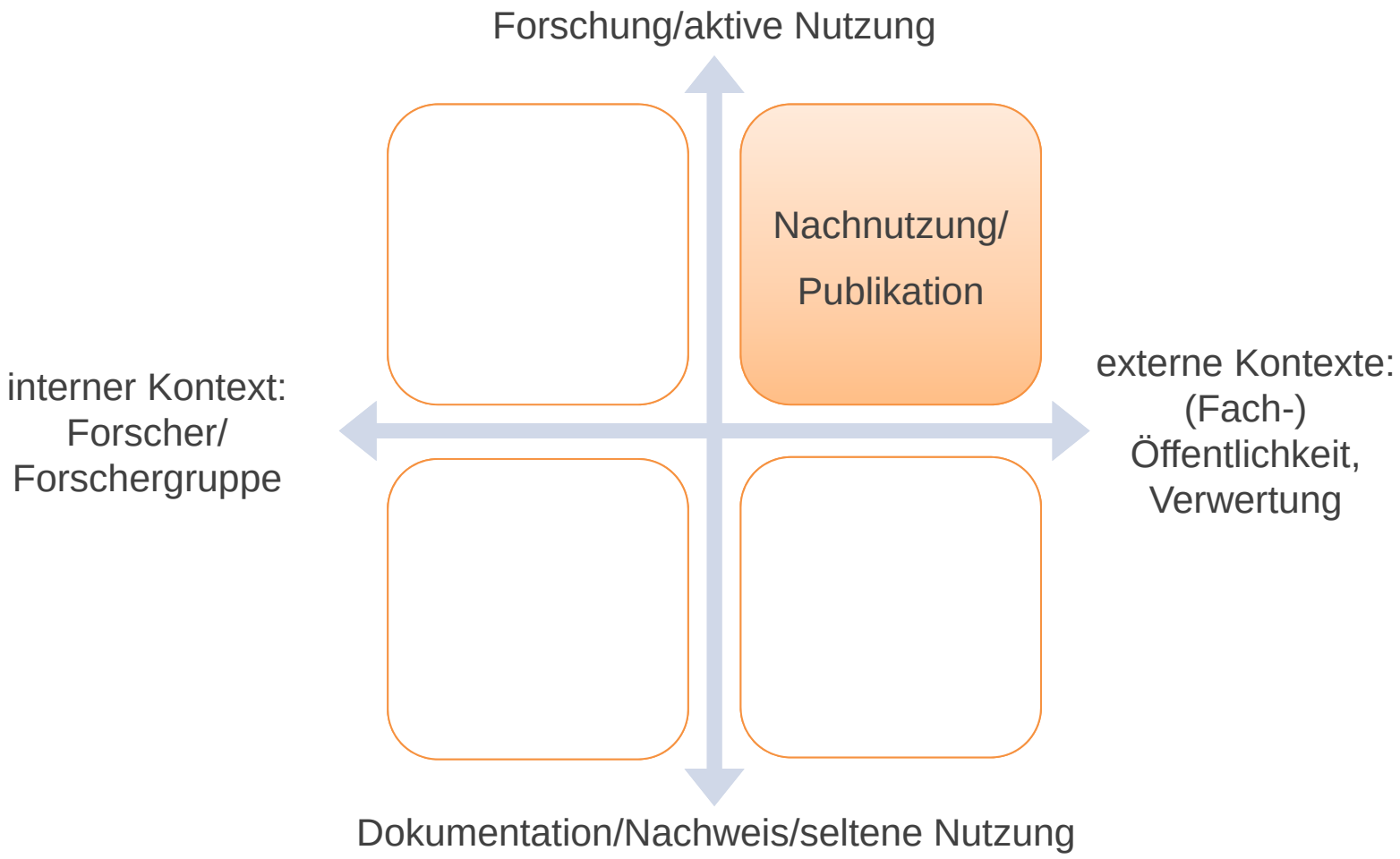
Anbieter: lokale Infrastruktur z.B. Rechenzentren, Bibliotheken

Hohes Volumen, sehr seltene Datenanfragen, begrenzte Dauer

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

Faktisch werden Aufwand und Kosten für Dokumentation mit Kosten des Verantwortungsfalls abgewogen.

Service-Klasse: Nachnutzung



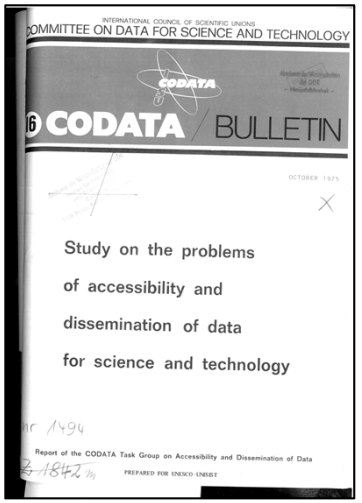
Geschichte der organisierten Nachnutzung digitaler Forschungsdaten

Geschichte der organisierten Nachnutzung digitaler Forschungsdaten

„One of the reasons for the ever-increasing efficiency of science 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.“

Nachnutzbarkeit ist kein neues Aufgabengebiet:

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

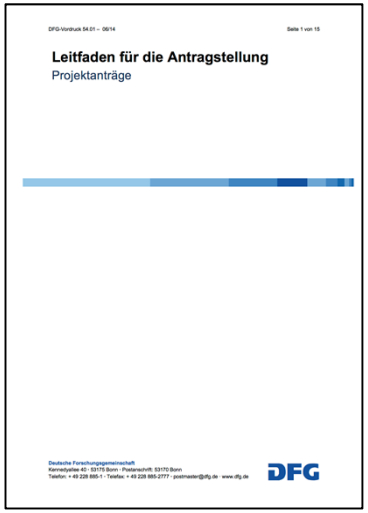


DFG-Anträge seit 2010

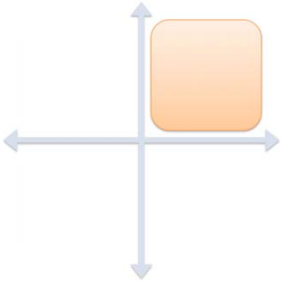
„2.4 Umgang mit den im Projekt erzielten Forschungsdaten

[...] 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.” DFG, Leitfaden für die Antragstellung, Version 06/2014



Service-Klasse: Nachnutzung



Ziele:

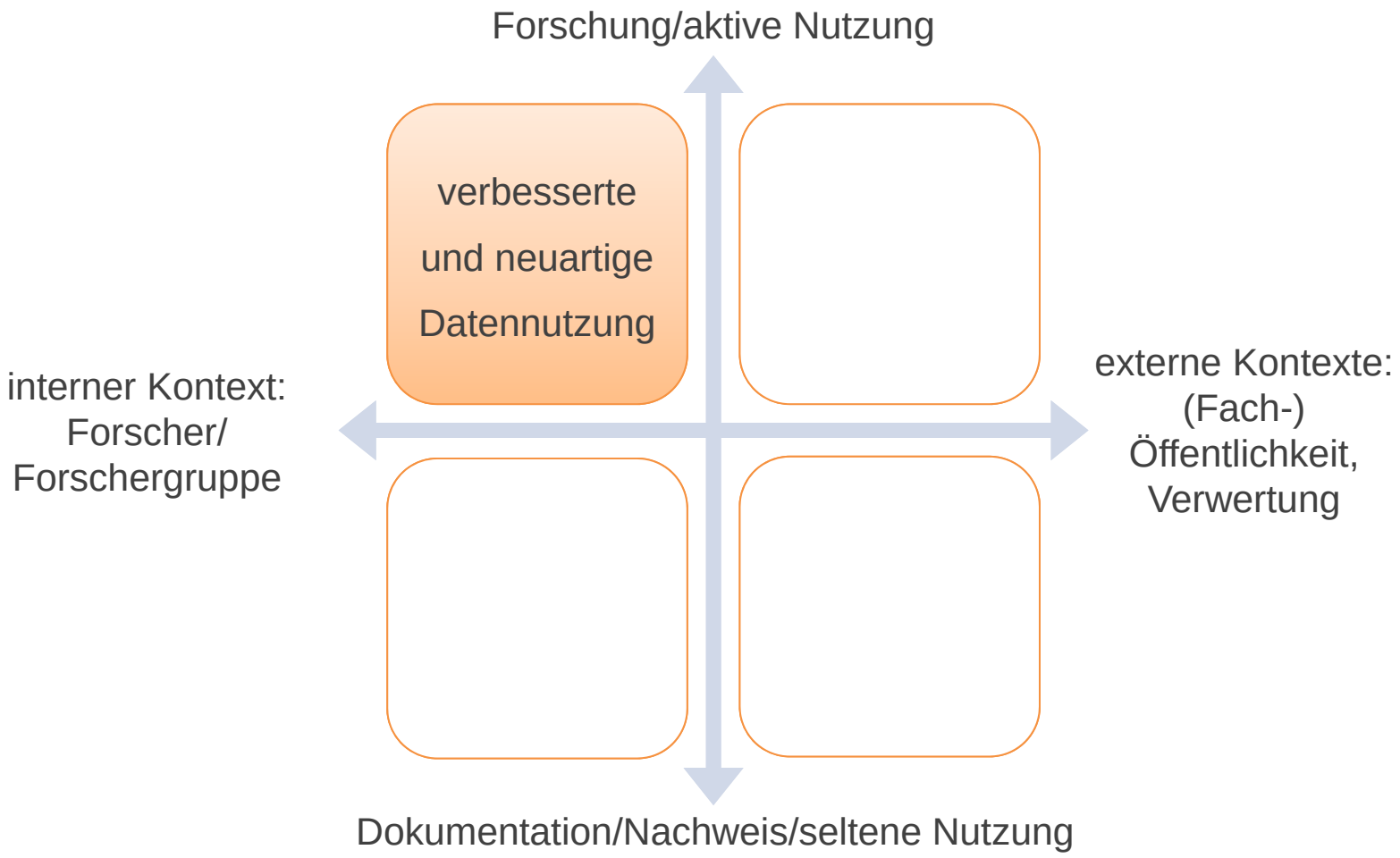
- zitierfähige Datenpublikation
- erneute wissenschaftliche Nutzung von Daten oder Datenproduktion als Dienstleistung
- Förderersicht: erhöhte Effizienz
- Bewahrung nicht reproduzierbarer Daten

Zielgruppe: Fach-Communities

Anbieter: spezialisierte, in den Disziplinen verankerte Zentren

Daten werden selten benutzt, aber ohne klares Enddatum. Geringeres Datenvolumen, da Aufwand pro Datensatz hoch

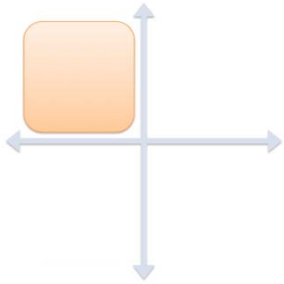
Service-Klasse: verbesserte Nutzung



Service-Klasse: verbesserte Nutzung

Ziele:

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



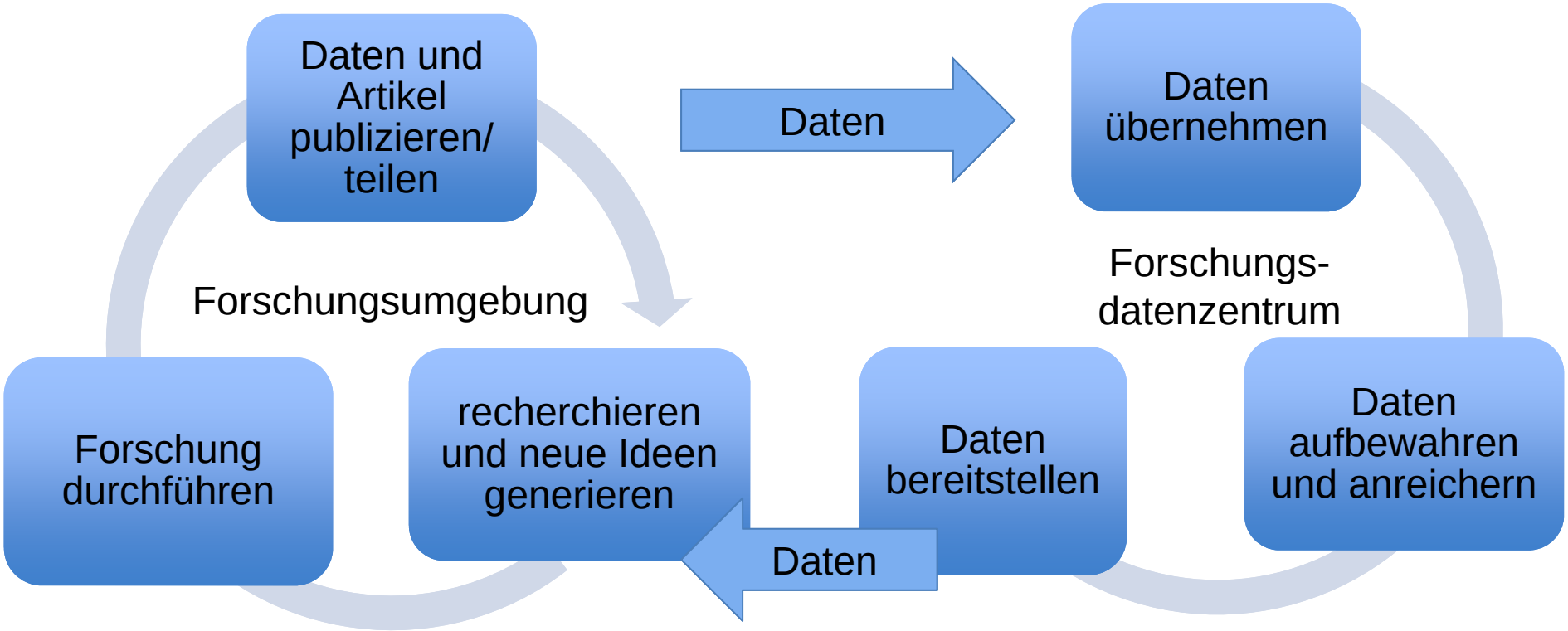
Zielgruppe: wissenschaftliche Arbeitsgruppen, z.B. SFBs

Anbieter: lokale Forschungsdaten-Support-Teams

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

Verknüpfung mit kollaborativen Forschungsumgebungen und Werkzeugen

Forschungsdatenmanagement betrifft den gesamten (idealisierten) Forschungs- und Forschungsdatenzyklus



Unterschiede zwischen Langzeitarchivierung und Forschungsdatenmanagement

- Langzeitarchivierung hat eine abstraktere Fragestellung (langfristigere Nutzbarkeit beliebiger digitaler Objekte) und sieht von dem Management des weiteren Anwendungsgebietes ab.

- Forschungsdatenmanagement beschränkt sich auf Forschungszwecke, aber kümmert sich viel stärker um spezielle Anforderungen und Prozesse der Zielgruppe.

- Tendenz: Forschungsdaten werden als instrumenteller Wert für die Forschung betrachtet. In der Langzeitarchivierung ist eher das digitale Objekt der Maßstab und wird als historische Überlieferung behandelt.

Beispiel: Integrität und Authentizität

- Im Digitalen müssen wir das Identitätskriterium festlegen. Z.B. Identität...
  - des physical objects: dasselbe Trägermedium
  - des logical objects: dieselbe Bitfolge
  - des conceptual objects: dieselben wesentlichen Eigenschaften, unabhängig vom Dateiformat

- Bei Forschungsdaten sind dieselben wesentlichen Eigenschaften eher die Aussagen über das Forschungsobjekt für einen Forschungszweck. Oft ist es daher legitim und sinnvoll Forschungsdaten zu verändern, um sie zu korrigieren, zu optimieren oder zu erweitern.

Beispiel: Dateiformate

- In der Langzeitarchivierung sind Dateiformate ein großes Aufgabenfeld:
  - Was sind die besten Dateiformate für die Langzeitarchivierung?
  - Wie bewerkstelligt man die Konvertierung von Dateiformaten?
  - ...

- Bei Forschungsdaten gibt es diese Fragen auch und z.T. sehr spezielle Dateiformat. Aber oft sind Dateiformate auch selbst entwickelte und etablierte Community-Standards und Werkzeuge werden von der Community selbst entwickelt. Dann stellen sich viele typische Fragen der Langzeitarchivierung nicht.

Beispiel: disziplinspezifische Anforderungen des Forschungsdatenmanagements

- aufwändige Metadaten und Aufbereitung von Datensätzen
- inhaltliche und technische Qualitätskontrolle von Forschungsdaten
- Pflege langfristiger Zeitreihen
- Bereitstellung der notwendigen Werkzeuge zur Nachnutzung
- Unterstützung des Forschungsprozesses
- Interaktion mit Zielgruppe und Begleitung der Disziplinentwicklung (Anforderungen, Terminologie etc.)

Inhaltsübersicht

- Was sind Forschungsdaten?
- Was ist Langzeitarchivierung?
- Forschungsdatenmanagement und Unterschiede zur Langzeitarchivierung
- Einige Überlegungen zur Informationsinfrastruktur

Wieso kann nicht jede Infrastruktureinrichtung einfach das Forschungsdatenmanagement übernehmen?

- Die Aufgabe ist zu groß.
- Die Aufgabe wird zu pauschal betrachtet.
- Nicht jede Teilaufgabe ist richtig für jede Infrastruktureinrichtung.

Daten- und Kostentrends

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

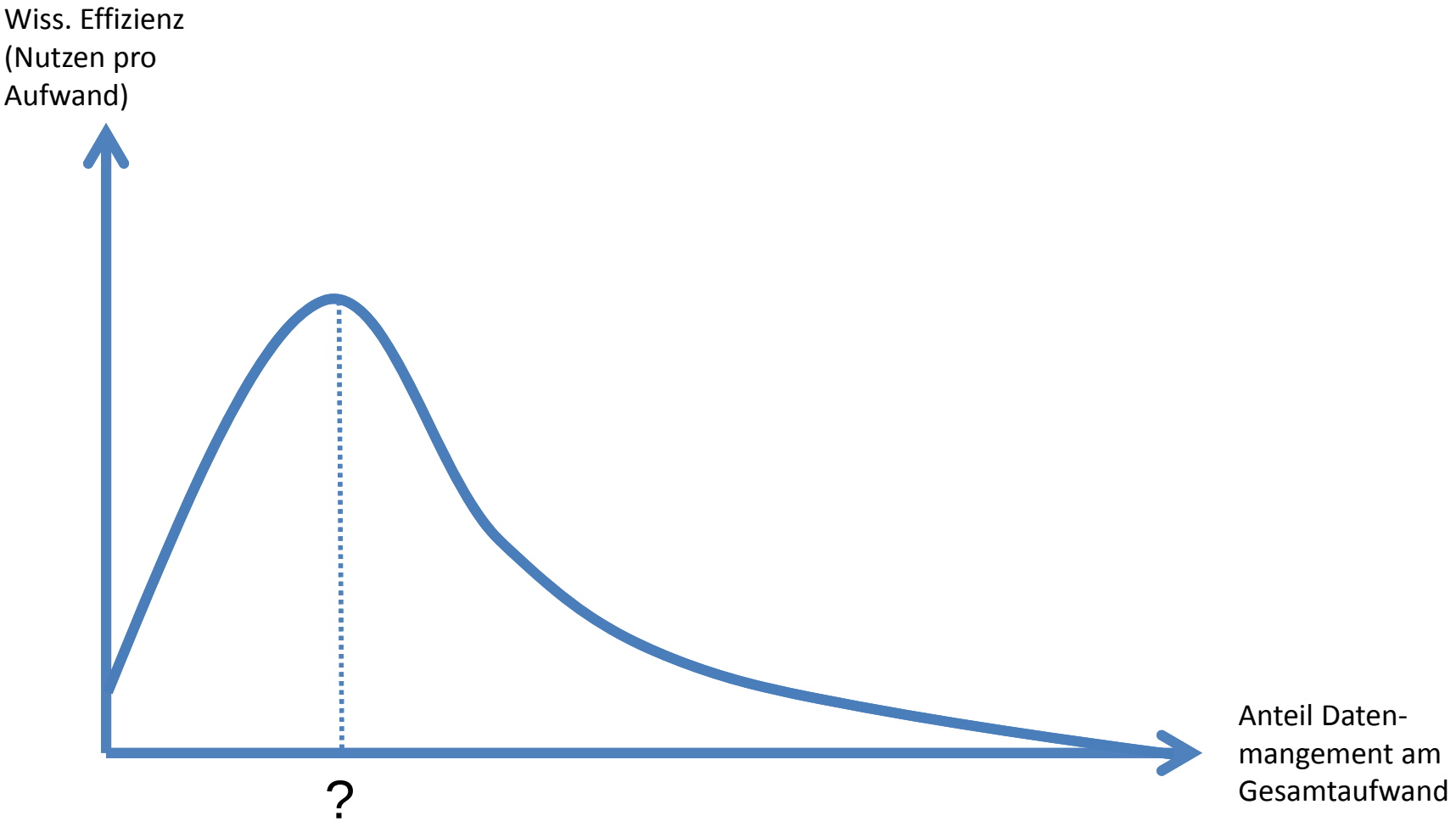
- Speicherplatzbedarf steigt jedes Jahr um 60%
- Die Speicherdichte wird nicht mehr als 20% pro Jahr steigen für die nächsten fünf Jahre. (Speicherdichte und Kosten sind eng verwandt.)
- IT-Budgets steigen ... quasi nicht.

David Rosenthal, Talk "Costs: Why Do We Care?", 2014.
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.

Neil Beagrie, Brian Lavoie and Matthew Woollard, Keeping Research Data Safe 2, 2010.
http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf

In der Forschungsdaten-Diskussion implizit, aber offene Frage:
Nicht nur Kosten, sondern Investition und Nutzen.
Aber wieviel wäre optimal oder sinnvoll?



Eine Konsequenz: Auswahl und Bewertung von Daten

- Nicht alles kann aufbewahrt werden, weder technisch noch ökonomisch.
  - Bedauerlich, aber stellt nicht das Gesamtvorhaben in Frage.

- 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.

So wie Kooperation notwendig ist bei der Gewinnung von Forschungsdaten, …

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

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

... so ist Kooperation notwendig beim Forschungsdatenmanagement.

Datenvolumen

Technik

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

Disziplinanforderungen

- disziplinspezifische Expertise zu divers
- Disziplinverankerung notwendig

Service-Struktur

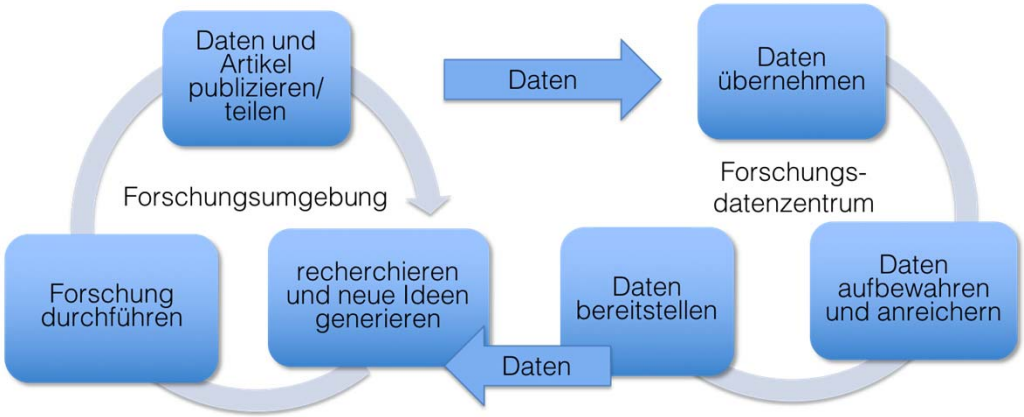
- Zielgruppe und Beratungs-/Betreuungsaufwand zu groß
- auch (kooperierende) Akteure mit sehr ähnlichem Angebot können sinnvoll sein (z.B. Ausfallsicherheit)

Häufiger Ansatz generischer Infrastruktureinrichtungen

Bibliotheken und Rechenzentren fokussieren meist auf klassisches Publikations- und Archivparadigma und ignorieren den Forschungsprozess (links):

- Wie können wir Forschungsdaten mit unseren bestehenden Speichermöglichkeiten übernehmen?
- Wie können wir Forschungsdaten mit unseren bestehenden Recherche-Instrumenten bereitstellen?

Das wird oft den disziplinspezifischen Anforderungen und dem Forschungsprozess nicht gerecht.



Optionen für generische Infrastruktureinrichtungen im Bereich Forschungsdaten

- eingeschränkte Basis-Infrastruktur (Dokumentation, dark archive)

- Spezialisierung zu überregionalem disziplin-spezifischem Forschungsdatenzentrum (um Disziplinanforderungen für Nachnutzung zu erfüllen)

ads ARCHAEOLOGY DATA SERVICE

- Publikations- und Archivparadigma verlassen und in Forschungsprozess gehen. Vermittler und lokaler Basis-Support zwischen Wissenschaftler und Forschungsdatenzentrum.

Fotos: Hochspannung, Thomas Kohler, cc-by https://www.flickr.com/photos/mecklenburg/4968331341/, Archaeology Data Service, SUB Göttingen

Vielen Dank!

Weiterführende Literatur

- Kenneth Thibodeau. Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years. Council on Library and Information Resources, 2002. – URL: http://www.clir.org/pubs/reports/pub107/thibodeau.html

- Heike Neuroth, Stefan Strathmann, Achim Oßwald, Regine Scheffel, Jens Klump, Jens Ludwig (Hrsg.). Langzeitarchivierung von Forschungsdaten. Boizenburg 2012, urn:nbn:de:0008-2012031401

- Jens Klump, Jens Ludwig. Forschungsdatenmanagement. S. 257 – 276, in: Heike Neuroth, Norbert Lossau, Andrea Rapp (Hg.). Evolution der Informationsinfrastruktur – Kooperation zwischen Bibliothek und Wissenschaft. Vwh-Verlag, Glückstadt 2013, dx.doi.org/10.3249/webdoc-39006