Forschungsdaten-Management: Formen, Akteure und Werkzeuge

Jens Ludwig

ludwig@sub.uni-goettingen.de
ludwig@mmg.mpg.de
19. Juni 2013

mpimmg

NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN | SUB

Schwierigkeiten in der Diskussion über Forschungsdaten

Formen: mangelnde Unterscheidung zwischen unterschiedlichen Aktivitäten

Akteure: mangelnde Unterscheidung zwischen Akteure und ihrer Interaktion

Werkzeuge: (zu) starker Fokus auf technische Werkzeuge; zu wenige übertragbare und organisatorische Instrumente

Formen: mangelnde Unterscheidung zwischen unterschiedlichen Aktivitäten

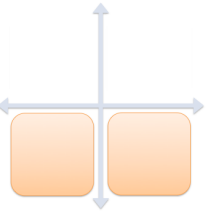
Akteure: mangelnde Unterscheidung zwischen Akteure und ihrer Interaktion

Werkzeuge: (zu) starker Fokus auf technische Werkzeuge; zu wenige übertragbare und organisatorische Instrumente

Was umfasst Forschungsdaten-Management?

Service-Gruppe: Dokumentation

Ziel: Nachvollziehbarkeit für Verantwortungszwecke



Zielgruppe: Institutionen (weniger Wissenschaftler)

Hohes Volumen, sehr seltene Datenanfragen, begrenzte Dauer

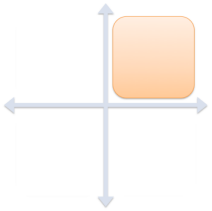
- einfacher und schneller Zugriff wird Kostenersparnis geopfert
- voraussichtlich Bitstream Preservation + basale Metadaten + Hard-/Software-Museum

Aufwand und Kosten für Dokumentation werden mit Kosten des Verantwortungsfalls abgewogen

Service-Gruppe: Nachnutzung

Ziele:

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



Zielgruppe: Fach-Communities

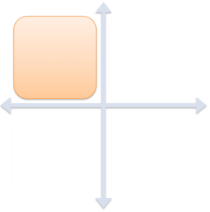
Daten werden selten benutzt, aber ohne klares Enddatum

Geringeres Datenvolumen, da Aufwand pro Datensatz hoch

Service-Gruppe: eResearch

Ziele:

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



Zielgruppe: wissenschaftliche Arbeitsgruppen, z.B. SFBs

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

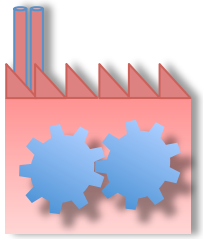
Verknüpfung mit kollaborativen Forschungsumgebungen und Werkzeugen

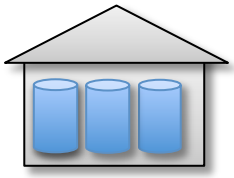
Formen: mangelnde Unterscheidung zwischen unterschiedlichen Aktivitäten

Akteure: mangelnde Unterscheidung zwischen Akteure und ihrer Interaktion

Werkzeuge: (zu) starker Fokus auf technische Werkzeuge; zu wenige übertragbare und organisatorische Instrumente

Idealisierter Forschungs- und Forschungsdatenzyklus





Daten und Artikel publizieren/ teilen

Daten

Daten übernehmen

(virtuelle/kollaborative) Forschungsumgebung

Forschungs-datenzentrum

Forschung durchführen

recherchieren und neue Ideen generieren

Daten bereitstellen

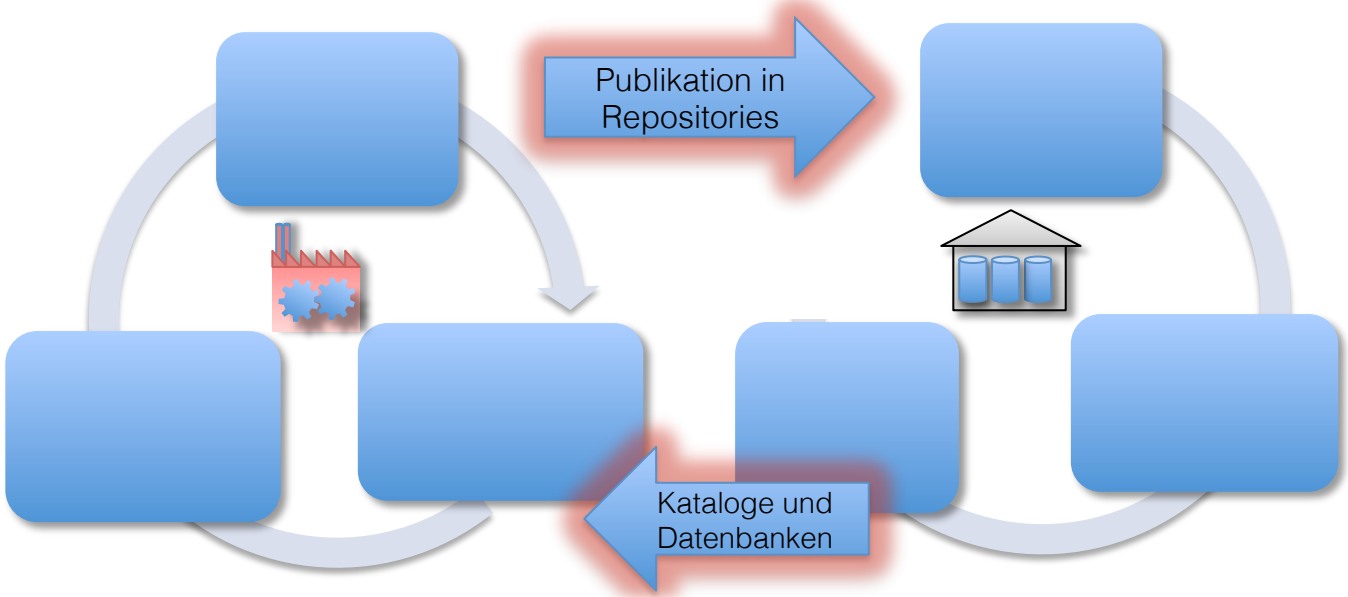
Daten aufbewahren und anreichern

Daten

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?



Anforderungen von Forschungsdaten-Nachnutzung

Klassischer Ansatz wird disziplinspezifischen Anforderungen für Nachnutzung nicht gerecht.

Hohe Disziplinverankerung und großes Disziplinwissen notwendig:

- aufwändige Aufbereitung und Dokumentation von Datensätzen
- Qualitätskontrolle von Forschungsdaten
- Pflege langfristiger Zeitreihen
- Interaktion mit Zielgruppe und Begleitung der Entwicklung der Disziplin (Anforderungen, Terminologie etc)
- ...

Aufgaben und Akteure

Dokumentation:

- Aufgabe: klar und halbwegs standardisierbar
- Anbieter: lokale Infrastruktur z.B. Rechenzentren, Bibliotheken

Nachnutzung/Publikation:

- Aufgabe: komplex, begrenzt standardisierbar
- Anbieter: spezialisierte, in den Disziplinen verankerte Zentren

eResearch:

- Aufgabe: komplex, sehr projektspezifisch
- Anbieter: lokale eResearch-Support-Teams
  - Mischrolle aus Forschung und Infrastruktur
  - können an lokaler Infrastruktur sein
  - müssen aber Vermittler und dürfen nicht Vertreter sein

Strategie für Nachnutzung: In eResearch investieren

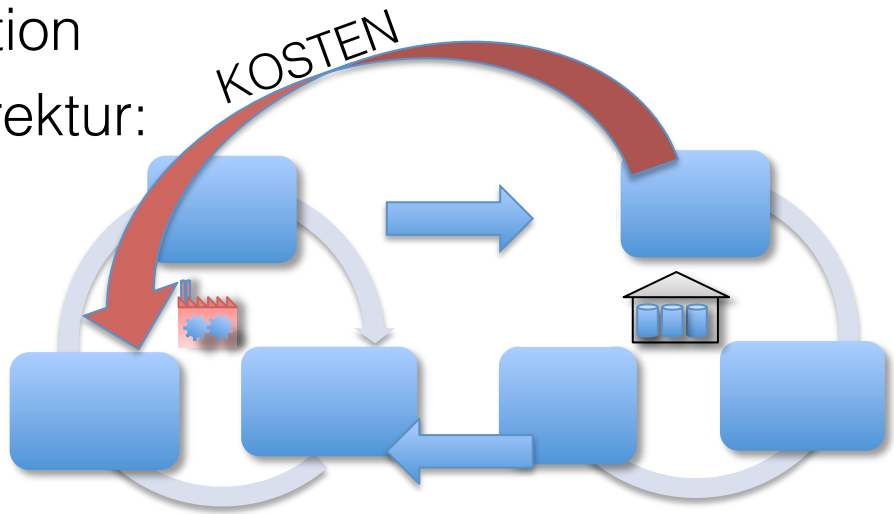
Kosten für Nachnutzung in den Forschungszyklus verschieben.

Beispiele:

- DKRZ, MPI Psycholingu., GESIS etc bieten disziplinspezifsche Forschungswerkzeuge an, die ihnen später die Arbeit erleichtern
- INF-Projekte („embedded data managers“) in SFBs

Nutzen:

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



Notwendige Interaktion der Akteure

Zentrale Forschungsdatenzentren können nicht überall/ dezentral eResearch unterstützen, sind aber darauf angewiesen.

Basis-Infrastruktur für Dokumentation ist Aufgabe der jeweiligen Institution.

Lokale eResearch-Teams können weder

- Fachkompetenz der Forschungsdatenzentren noch
- Angebote der Forschungsdatenzentren und Infrastruktur

ersetzen.

Formen: mangelnde Unterscheidung zwischen unterschiedlichen Aktivitäten

Akteure: mangelnde Unterscheidung zwischen Akteure und ihrer Interaktion

Werkzeuge: (zu) starker Fokus auf technische Werkzeuge; zu wenige übertragbare und organisatorische Instrumente

„Data Management Plans“

Es existieren im englisch-sprachigen Raum sehr viele Checklisten und Leitfäden, Schwerpunkt: „data management plans“.

Hintergrund: Förderer haben stärkere Auflagen und verlangen „data management plans“ als Teil des Antrags.

Persönliche Einschätzung: Diese Pläne sind auch „nur“ Teil des Antrags und wenig instruktiv für die eigentliche Durchführung.

WissGrid-Publikationen

Primäres Ziel war deutschsprachige Quelle zu schaffen.

Generische Checkliste und Leitfaden. Muss angepasst werden.

Richtet sich an Wissenschaftler und Infrastruktur. Muss angepasst werden.

Aktuelle Arbeiten

Angepasste, „verzehrfertige“ Checkliste für gesamtes Datenmanagement:

- stärkere Interviewstruktur mit Einstiegsfragen
- Folgeaufgaben und Vertiefungsfragen
- Implementation als Web-Portal

Metadaten

- nehmen Informationen aus Checkliste auf
- für Nachnutzung, Dokumentaton und eResearch
- sowohl für Wissenschaftler als auch Verwaltung, Datenmanager oder IT

Generische Arbeitsabläufe

- Definition von auslösenden Ereignissen
- Definition der Aktivitäten
- involvierte Objekte und Hilfsmittel

Kern- und Einstiegsfragen

1. What kind of data is it and what is the content?
2. What kind of data is it from a technical viewpoint?
3. Where will the data be stored and who is responsible for it?
4. Is the data sensitive or access limited for legal, ethical or other reasons? Are there other obligations?
5. What documentation and context information is necessary to understand and use the data?
6. Is there an interest in support for using and managing the data during the research phase?

Ausschnitt: What kind of data is it and what is the content?

Aufgaben:

1. *Check whether domain expertise is available for this content and if it fits with existing collections*
2. *Determine whether the data is valuable for reuse.*
    1. Vertiefungsfrage: *Could the data be relevant for other people's research? Which people, groups or institutions might be interested in using the data?*
    2. Vertiefungsfrage: *Can the data be reproduced efficiently (in contrast to e.g. observations of non-recurring events)?*

3. and 4. Encourage search for prior data and *determine which institution is the best host or curator for the data*

1. Vertiefungsfrage: *Can already existing data sets be reused?*
2. Vertiefungsfrage: *Do other institutions have similiar data collections?*

...

Resultierende Metadaten z.B. Datentyp, inhaltliche Stichwörter, vorläufiger Service Level

Vielen Dank!