Informationsinfrastrukturen für Forschungsdaten

Ansätze und Strategien der DFG

Open-Access-Tage Wien, 26. - 27. September 2012
Dr. Stefan Winkler-Nees

Bitte beachten:

Zu Nutzungsrechten der verwendeten Bilder und Inhalte:

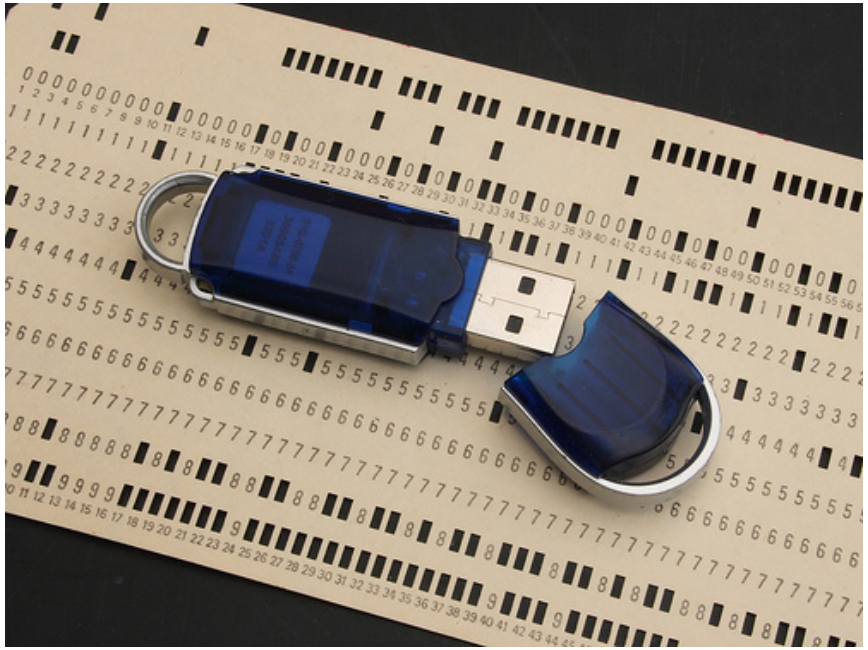
Zum Zeitpunkt der Präsentation dieser Folien unterlagen die verwendeten Bilder den jeweils angegebenen Nutzungsrechten (Creative Commons, Gemeinfreiheit, Bilderfundus der DFG). Vor einer evtl. Weiternutzung müssen diese Nutzungsrechte erneut überprüft werden.

Die darüber hinaus gehenden Inhalte unterliegen der Creative Commons Attribution 3.0 Germany (CC BY 3.0).

1. Um was geht es?
Forschungsdaten ...

Der Versuch einer Definition:

... Unter Forschungsdaten sind [...] digitale und elektronisch speicherbare Daten zu verstehen, die im Zuge eines wissenschaftlichen Vorhabens z.B. durch Quellenforschungen, Experimente, Messungen, Erhebungen oder Befragungen entstehen. ...



Ian-S; http://www.flickr.com/photos/ian-s/2152798588; cc: by-nc-nd

2. Um was geht es?
Forschungsdaten ...

- ... sind die Grundlage wissenschaftlicher Erkenntnis.
- ... werden heute nur unzureichend genutzt.
- ... sind nur eingeschränkt zugänglich.
- ... stehen nicht langfristig zur Verfügung.

Die Vision der Wissenschaftsorganisationen

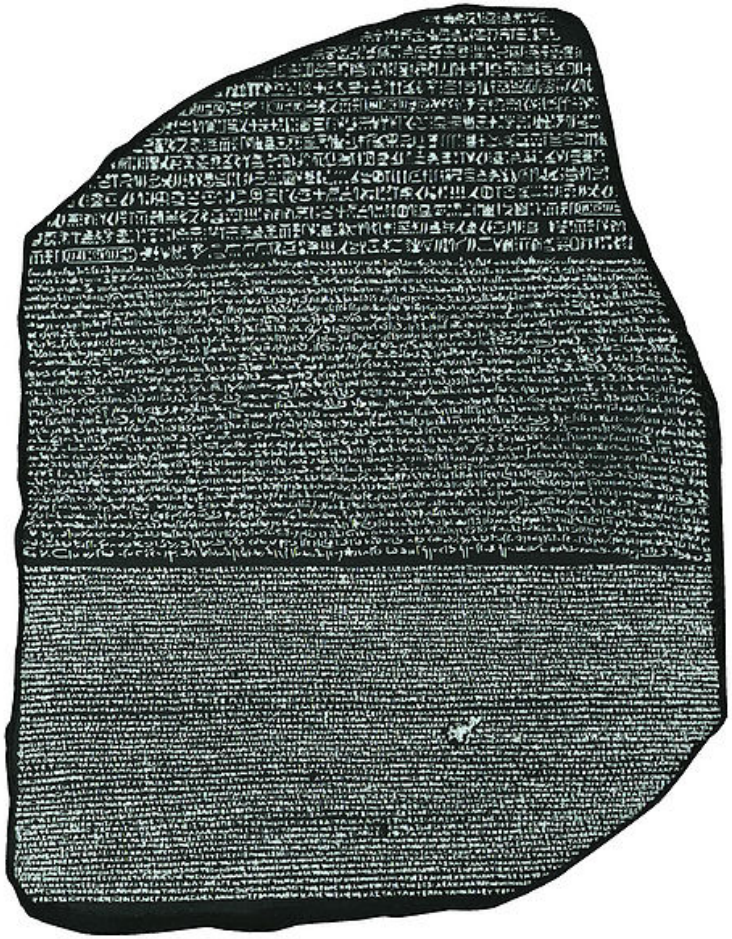
- Forschungsdaten sollen möglichst frei und überregional zugänglich und langfristig verfügbar sein.

Nehmen wir an, ...

... Sicherung und Archivierung, Teilen und Nachnutzen von Forschungsdaten wäre möglich und fände statt!

- Wir hätte mehr Daten aus einer Disziplin und könnten unterschiedlichste Daten vergleichen.
- Doppel-Untersuchungen würden vermieden, eine Qualitätskontrolle wäre gewährleistet.
- Andere Datensätze könnten eigene Ergebnisse unterstützen.
- Einzigartige, nicht reproduzierbare Ergebnisse wären dokumentiert.
- Neue Interpretationen werden möglich.
- Wissenschaftliches Arbeiten wird erweitert: Experiment, Theorie, Modellierung und nun „data-driven science“ - „The Fourth Paradigm“
- Ein offener Zugang zu Daten erfüllte die Anforderungen „Guter Wissenschaftlichen Praxis“.

Die Lebenszeit von Information



Der **Rosettastein** ist eine halbrunde, steinerne Stele mit einem in drei Schriften (Altgriechisch, Demotisch, Hieroglyphen) eingemeißelten Priesterdekret als Ehrung des ägyptischen Königs Ptolemaios V. sowie seiner Frau und deren Ahnen. Der Rosettastein trug maßgeblich zur Übersetzung der ägyptischen Hieroglyphen bei. Er befindet sich heute im British Museum in London. Er stammt aus dem Jahr 196 v. Chr.

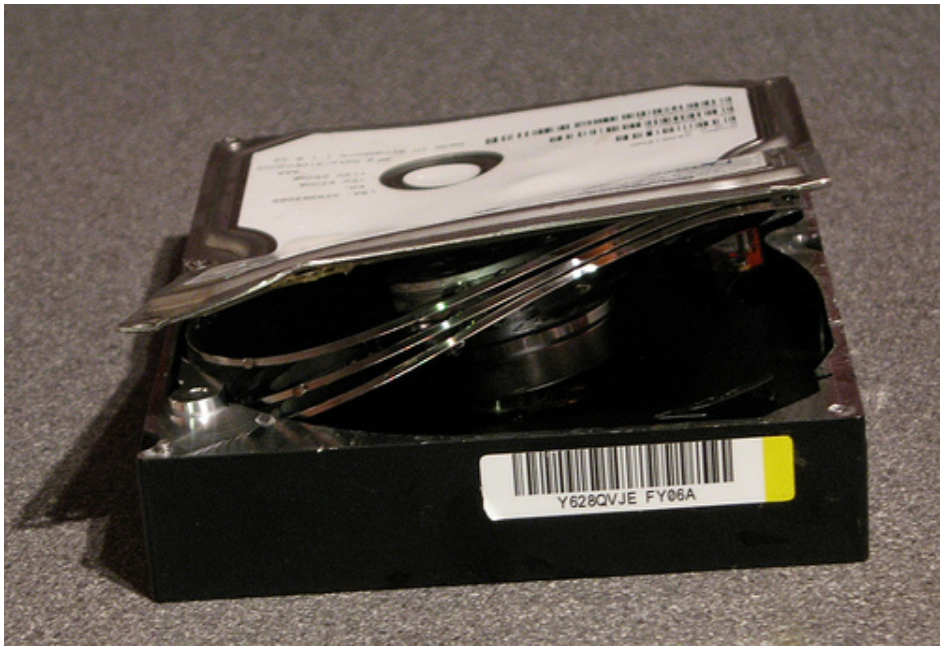
Alter: 2.208 Jahre

http://upload.wikimedia.org/wikipedia/commons/c/ca/Rosetta_Stone_BW.jpeg (gemeinfrei)

Die Lebenszeit digitalen Wissens

"Digital information lasts forever - or five years, whichever comes first."

Jeff Rothenberg, RAND Corp., 1997



Stinging Eyes; http://www.flickr.com/photos/martinlatter/299981441 cc: by-sa

Aber, ... das sind doch MEINE Daten!!!

Und es gibt gute Gründe Daten nicht zu teilen, sprich „wegzugeben“:

- Darin stecken xxx Euro und yyy Jahre an Arbeit.
- Sie könnten so leicht missverstanden werden.
- Sie könnten fehlerhaft sein und jemand Fremdes könnte das feststellen.
- Irgendjemand könnte etwas Interessanteres darin finden.
- Irgendjemand könnte sie vor mir veröffentlichen.
- Es braucht viel zu viel Zeit, sie in einen neuen Kontext zu setzen und sie neu zu formatieren.
- Wie und wo soll ich meine Daten überhaupt einstellen?
- Ich habe ja überhaupt keine Kontrolle mehr über meine Daten.

Viele offene Fragen

- **Wissenschaftliche Freiheit**
  (GG § 5). Auch die Freiheit, Daten nicht freizugeben?

- **Wem gehören die Daten?**
  Förderorganisation, Institution/Einrichtung, (Haupt)-Antragsteller, Wissenschaftlicher Mitarbeiter, Verlag? Alles ist möglich.

- ***Wie* werden Daten zur Verfügbar gestellt?**
  Was für Rechte werden für die Nachnutzung gewährleistet? Welche Lizenzen verwendet? Die „Offenheit“ von Daten ist nicht nur eine technische Frage.

- **Qualitätssicherung**
  Alle Daten? Ist das überhaupt möglich? Wünschenswert? Wer entscheidet, welche Daten wo und wie lange gespeichert werden?

Die Deutsche Forschungsgemeinschaft

- **1998: DFG Denkschrift: “Sicherung guter Wissenschaftlicher Praxis”**
- **2003: Berliner Erklärung über den offenen Zugang zu wissenschaftlichem Wissen**
- **2006: DFG Positionspapier**
- **seit 2007: regelmäßige Workshops und Expertenrundgespräche**
- **2008: Nationale Schwerpunktinitiative „Digitale Information“ der Allianz der Deutschen Wissenschaftsorganisationen**
- **2010: GWK Kommission Zukunft der Informationsinfrastruktur („KII“)**
- **2011: Internationale Vernetzung (Knowledge Exchange, EC, G8 & O5)**
- **2012: DFG Positionspapier**
- **2013: Förderprogramm ?**

DFG Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsdaten

Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme

Unterausschuss für Informationsmanagement

Januar 2009

- Definitionen zu Forschungsdaten, Organisationskonzept, Metadaten und Standards
- Rechtewahrung der Wissenschaftlerinnen und Wissenschaftler
- Überregionale Bereitstellung
- Qualitätssicherung

www.dfg.de/lis/ unter „Veröffentlichungen / Informationsmanagement“

Aktuelle Aktivitäten der DFG

Ausschreibung: Informationsinfrastrukturen für Forschungsdaten

Die Deutsche Forschungsgemeinschaft (DFG) unterstützt mit dieser Ausschreibung im Förderbereich Wissenschaftliche Literaturversorgungs- und Informationssysteme (LIS) Vorhaben zur Entwicklung und Optimierung von Informationsinfrastruktur, die auf einen effizienten und nachhaltigen Umgang mit Forschungsdaten abzielen.

Ergänzung des Antragsmusters

„ ... Wenn aus Projektmitteln systematisch (Mess-)Daten erhoben werden, die für die Nachnutzung geeignet sind, legen Sie bitte dar, welche Maßnahmen ergriffen wurden bzw. während der Laufzeit des Projektes getroffen werden, um die Daten nachhaltig zu sichern und ggf. für eine erneute Nutzung bereit zu stellen. Bitte berücksichtigen Sie dabei auch – sofern vorhanden – die in Ihrer Fachdisziplin existierenden Standards und die Angebote bestehender Datenrepositorien.“

Forschungsdaten: Ansätze und Strategien

- Begleitung eines Gestaltungsprozesses mit dem Ziel, digitale Ressourcen besser zu nutzen, Infrastrukturen aufzubauen und Werkzeuge zu entwickeln und Daten bereit zu stellen:

  Sicherung und Archivierung und Nachnutzung

- Enge Einbindung der Wissenschaftler als Datenproduzenten und Nutzer der Datenrepositorien gemeinsam mit Experten aus dem Informationsmanagement in die Gestaltung dieses Prozesses.
- Fachspezifische Anforderungen und Bedürfnisse formulieren und in den Prozess mit einbringen.
- Entwicklung von Publikationsformen für wissenschaftlichen Daten (inkl. Peer-Review).
- Auf existierenden Ansätzen aufbauen und von Erfahrungen profitieren.
- Pilotprojekte und Explorationsprojekte initiieren.

Die DFG fördert die systematische Zusammenarbeit zwischen Wissenschaftlern und Informationsspezialisten.

Informationsspezialisten? Z.B. Bibliotheken ...

Professionelles Informationsmanagement …

Der alte Lesesaal der British Library, London
http://en.wikipedia.org/wiki/File:British_Museum_Reading_Room_Panorama_Feb_2006.jpg cc: by-sa

PURDUE UNIVERSITY **Libraries** *Access. Knowledge. Success.*

INTRODUCTION

Librarians at Purdue University are beginning to identify the scientific datasets that are being generated by our faculty and researchers as information assets to be collected, preserved, and made accessible as a function of the library's collection development. These librarians are subject-area specialists, and many have advanced degrees in their respective disciplines in addition to a degree in library science. They have all been trained in collection management; however, much of this training was related to traditional formats such as monographs and serials and not datasets. In our experience, one of the most effective tactics for eliciting datasets for the collection is a simple librarian-researcher interview. In this poster, we share a set of ten questions that a librarian can use as a starting point for such a "data interview". It is not a comprehensive strategy but instead a practical tool to draw out information that needs to be considered in order to evaluate the suitability of a dataset for the collection and the requirements for the infrastructure and services that will be needed for data curation.

#1 What is the story of the data?

Begin the interview with an open-ended question that allows the researcher to talk freely about his or her research, scientific workflow, and community of practice. This lends some insight into the value of the dataset and how it may fit into your collection and be used, and it provides the ***context*** for understanding how and why the dataset was created and how it was processed and analyzed.

#2 What form and format are the data in?

What computing environments (e.g., software) are required to use the data? If the data are in proprietary structures, you may consider reformatting them into agnostic formats or ones that can be more easily ***re-versioned***. Is there any existing ***metadata***, either external to the data or description that could be extracted from it? Ideally the data could be described to be discoverable by researchers from another discipline.

#3 What is the expected lifespan of the dataset?

In many cases, there are distinctions in the utility of a dataset as it begins in a raw state and then is analyzed and processed into new forms and versions as a result of different steps in the research workflow. Different entities may have custody of the data and use it for different purposes at different times, affecting its ***provenance***. Funding agencies may require that data be archived for a prescribed period of time or you may forecast its future value and the amount of time it should be retained. The data may be described and archived for effective ***preservation*** to ensure its accessibility and integrity over time.

#4 How could the data be used, reused, and repurposed?

This is a primary ***selection*** criterion that also impacts how the data are ***accessed*** and what ***policies*** may be needed to govern its use. As data are archived and shared, new and unintended uses for the data may increase its value. For example, a research dataset may be repurposed as a learning object.

#5 How large is the dataset, and what is its rate of growth?

It is important to quantify the size of the data for storage and network provisioning if you intend to ***ingest*** it into your repository. What is its physical (bits) and logical (records) ***scale***? Is the dataset static or dynamic? Ask for a sample of the data to examine.

#6 Who are the potential audiences for the data?

Information regarding potential users of the data and the users' needs is paramount. Along with potential uses for the data, this is another primary ***selection*** criterion. In some cases, the data may need to be embargoed or restricted to a limited group of users who are granted ***permission*** to access it.

#7 Who owns the data?

Establishing and maintaining the ***intellectual property*** represented by the data should be discussed at the earliest opportunity, and any conflicts should be resolved up-front. Many organizations have a submission policy that asks the contributor to verify that they own the data and have the right to submit it.

#8 Does the dataset include any sensitive information?

All data should be reviewed for information that violates ***confidentiality***, such as identification information on human subjects. Data curation activities should be informed by institutional review board requirements.

#9 What publications or discoveries have resulted from the data?

The researchers may have a bias regarding the importance of their data. The purpose of this question is to establish an objective metric for determining the value of the data for the collection. Different metrics may be more appropriate in determining the ***selection*** criteria for different kinds of data and data collections.

#10 How should the data be made accessible?

There is value in making data accessible using a conventional web-based user interface, but machine-to-machine interfaces should also be evaluated. These ***methods of access*** will be informed by the answers to the previous questions, and this question can be asked in an open-ended manner to fill in any gaps remaining at the conclusion of the interview.

SUMMARY

Although building robust collections of datasets present several complexities and challenges to resolve, the process of looking at scientific datasets as information assets and exploring what is needed to develop and manage data collections is similar to the traditional collection development practices that have been successfully employed by librarians for decades. We offer these ten "data interview" questions as a springboard for librarians to explore data curation in greater depth and specialization.

Michael Witt (mwitt@purdue.edu)
Assistant Professor of Library Science

Jake Carlson (jrcarlso@purdue.edu)
Data Research Scientist

Purdue University Libraries
Distributed Data Curation Center
http://d2c2.lib.purdue.edu

D2C2 Distributed Data Curation Center

"Conducting a Data Interview"

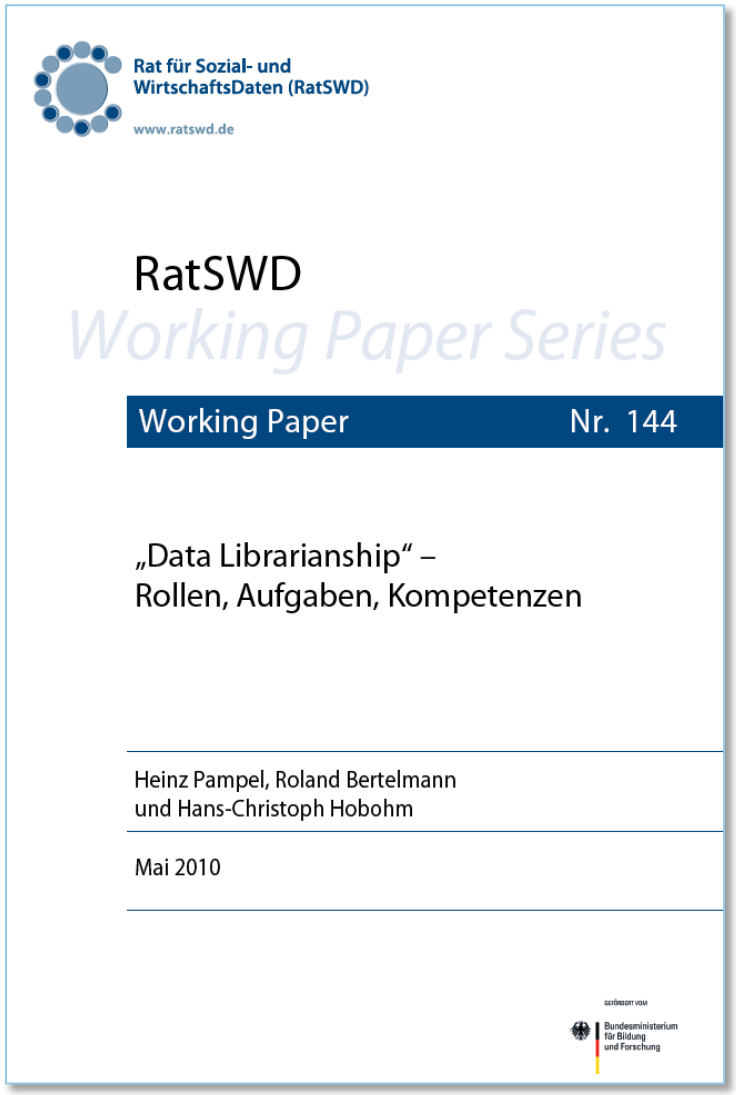
Michael Witt & Jake Carlson, Purdue University Libraries, West Lafayette, Indiana, USA

Skills, role and career structure of data scientists and curators

Eine Studie von KEY PERSPECTIVES für JISC

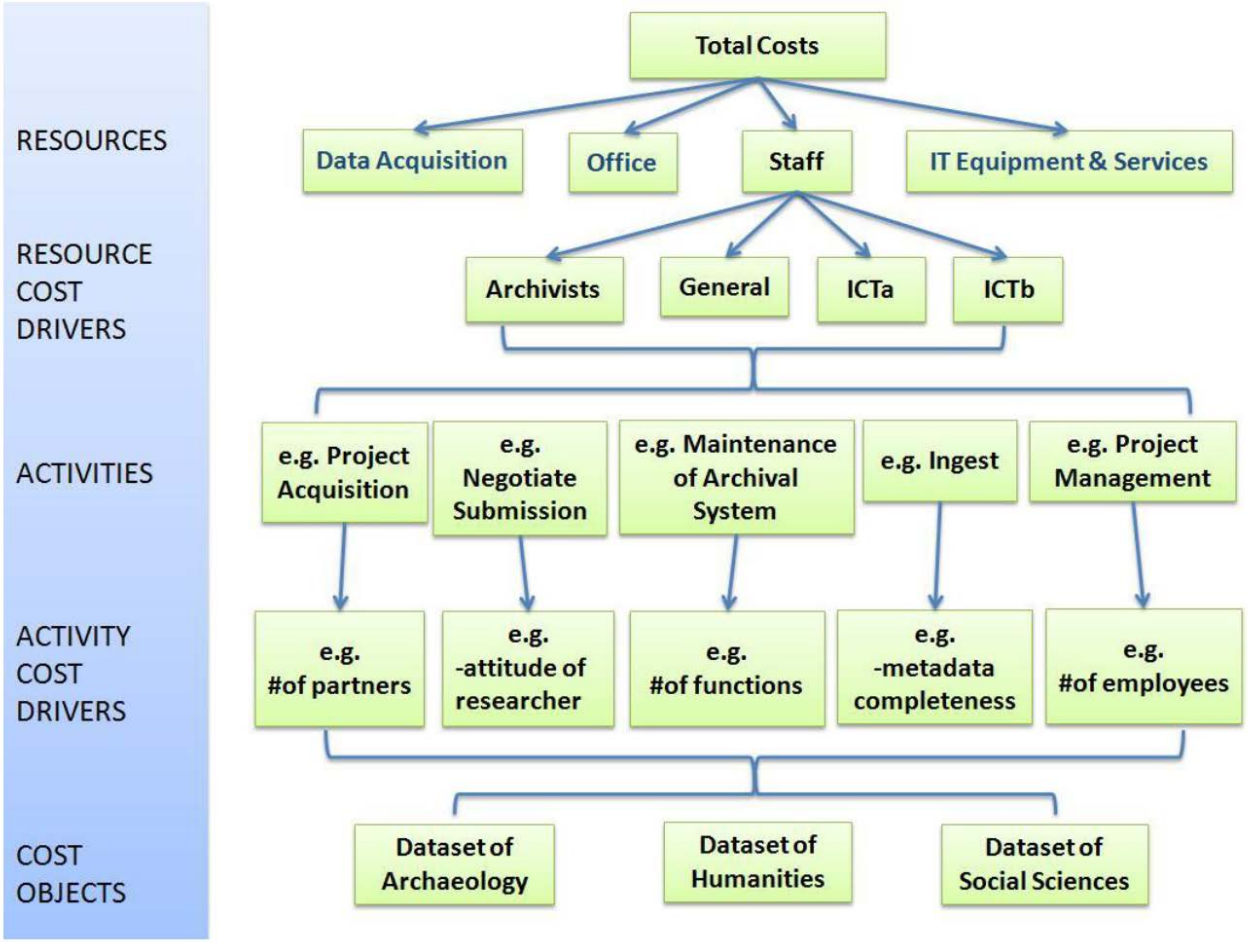
- **Data creator**
  Researchers with domain expertise who produce data. These people may have a high level of expertise in handling, manipulating and using data.
- **Data scientist**
  People who work where the research is carried out – or, in the case of data centre personnel, in close collaboration with the creators of the data – and may be involved in creative enquiry and analysis, enabling others to work with digital data, and developments in data base technology.
- **Data manager**
  Computer scientists, information technologists or information scientists and who take responsibility for computing facilities, storage, continuing access and preservation of data.
- **Data librarian**
  People originating from the library community, trained and specializing in the curation, preservation and archiving of data.

„Data Librarianship“ – Rollen, Aufgaben, Kompetenzen



... Die vielfältigen und häufig disziplin-spezifischen Herausforderungen beim Umgang mit wissenschaftlichen Daten fordern eine engere Kooperation zwischen Wissenschaft und infrastrukturellen Service-einrichtungen. Bibliotheken bietet sich die Chance, die Entwicklung organisatorischer und technischer Lösungen des Forschungs-datenmanagements aktiv zu gestalten und eine tragende Rolle in diesem Feld zu übernehmen. ...

Die Frage nach den Kosten ...



ABC - Activity Based Costing Model

- Improving tactical and strategic decision-making
- Understand the use of scarce organisational resources in various business activities

DANS

Data Archiving and Networked Services (KNAW, NWO; NL)

DFG

Eine erste Annäherung

Aus dem Gesamtkonzept der "Kommission Zukunft der Informationsinfrastruktur", April 2011: http://www.gwk-bonn.de/index.php?id=205

*„ ... Aus dem Bericht ist festzuhalten, dass der dauerhafte Betrieb von Forschungsdatenzentren als Teil der Forschungskosten etabliert werden muss und grob geschätzt einen dauerhaft zu finanzierenden Anteil von 5 % bis 10 % für den Bereich der „Datenpflege" an den Gesamtkosten für Forschung vorzusehen ist. Um international kompetitiv zu bleiben bedeutet dies, dass auch in Deutschland mittelfristig etwa **5 % bis 10 % der Forschungskosten** zusätzlich für nachhaltige „Datenbereitstellung" aufgebracht werden müssen. ... „*

⇒ ca. 400 – 800 Mio. €/a in Deutschland

Astrometrie-Satellit Gaia (ESA)
Kosten für die Mission einschließlich Start, Bodenkontrolle und Nutzlast: ca. 577 Millionen €.
Kosten für die wissenschaftliche Datenreduktion: ca. 120 Millionen €.

Was ist zu tun? Eine Zusammenfassung

- Zugang
  - Formen und Bedingungen des Zugangs regeln (“Open Access?”)
- Unterschiede in den Disziplinen
  - Art der Daten, Menge und Typus, Lebenszyklen und Nutzungscharakteristik definieren
- Wissenschaftliche Anerkennung
  - Die Bereitstellung von Daten für die Nachnutzung als selbstverständlichen Bestandteil wissenschaftlichen Arbeitens und Anerkennung etablieren (Wissenschaftskultur)
- Gebrauch von Standards, Entwicklung und Implementierung von Infrastruktur
  - Zusammenarbeit zwischen Wissenschaftlern und Informationsmanagementexperten
  - Sicherung der Interoperabilität in internationalen und interdisziplinären Netzwerken
- Lehre und Qualifikation
  - Forschende und Informationsdienstleister

Vielen Dank für die Aufmerksamkeit!

Stefan.Winkler-Nees@dfg.de