Universität Bielefeld

CITEC

# Kolloqium Wissensinfrastruktur

25.11.2016

# Langzeitarchivierung von Forschungsdaten - ein Pilotversuch

**Cord Wiljes**

**CITEC, Universität Bielefeld**

# Ziel

Entwicklung einer institutionellen Lösung, um der Nachweispflicht für Forschungsdaten nachzukommen.

# Gute Wissenschaftliche Praxis

Mitgliederversammlung der DFG:
"*Sicherung guter wissenschaftlicher Praxis*"
(17. Juli 1998, Ergänzung 3. Juli 2013):

**Empfehlung 7: Sicherung und Aufbewahrung von Primärdaten**

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

➔ http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf

# DFG Kommentar zu Empfehlung 7 (3. Juli 2013)

- Auf die Aufzeichnungen später zurückgreifen zu können, ist schon aus Gründen der Arbeitsökonomie in einer Gruppe ein zwingendes Gebot. Noch wichtiger wird dies, wenn veröffentlichte Resultate von anderen angezweifelt werden.
- (…) Daher hat jedes Forschungsinstitut, in dem lege artis gearbeitet wird, **klare Regeln** über die Aufzeichnungen, die zu führen sind, und über die Aufbewahrung sowie den Zugang zu den Originaldaten und Datenträgern. (…)

# DFG Kommentar zu Empfehlung 7 (3. Juli 2013)

- In renommierten Labors hat sich die Regel bewährt, dass der **komplette Datensatz**, der einer aus dem Labor hervorgegangenen Publikation zugrunde liegt, als Doppel **zusammen mit dem Publikationsmanuskript** und der dazu geführten Korrespondenz archiviert wird.
- Schon deshalb ist die Feststellung wichtig, dass das Abhandenkommen von Originaldaten aus einem Labor gegen Grundregeln wissenschaftlicher Sorgfalt verstößt und prima facie einen Verdacht unredlichen oder grob fahrlässigen Verhaltens rechtfertigt

# Universität Bielefeld

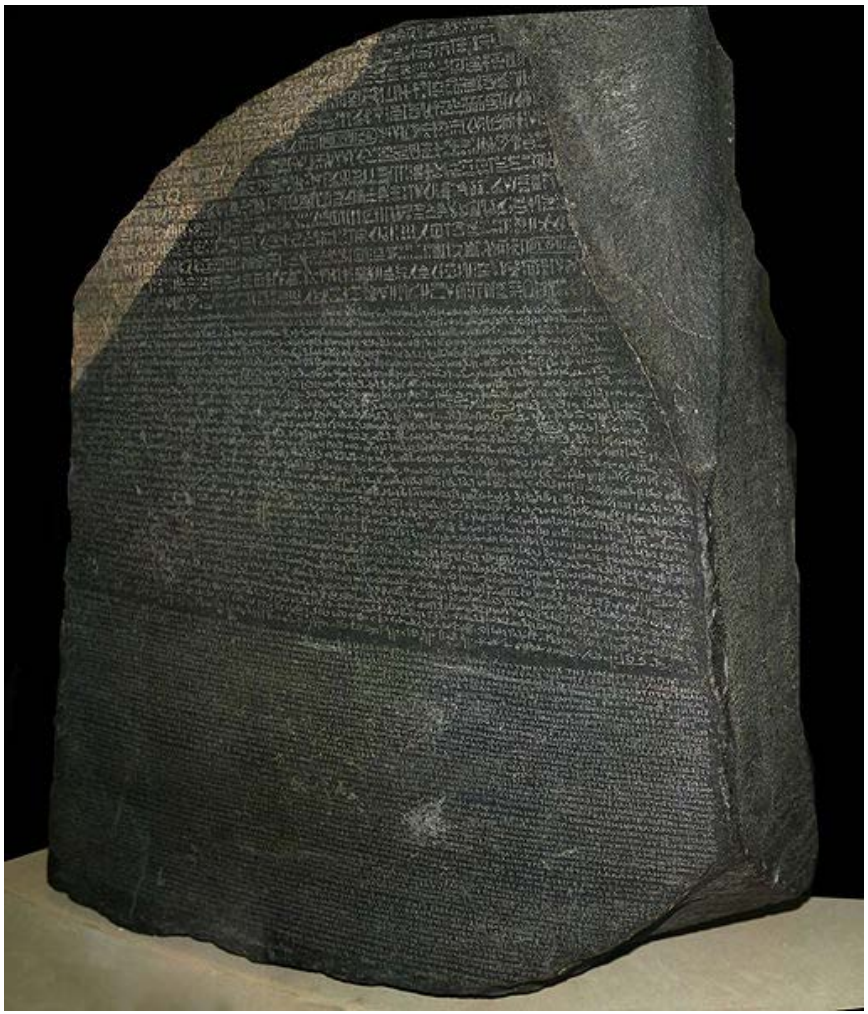
Senat der Universität Bielefeld: *"Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Universität Bielefeld"* (2. Feb. 2000):

**§ 5: Sicherung und Aufbewahrung von Primärdaten**

Primärdaten als Grundlagen für Veröffentlichungen sind auf haltbaren und gesicherten Trägern in der Institution, in der sie entstanden sind, für zehn Jahre aufzubewahren, soweit nicht gesetzliche Bestimmungen entgegenstehen. Wann immer möglich, sollen Präparate, mit denen Primärdaten erzielt wurden, für denselben Zeitraum aufbewahrt werden.

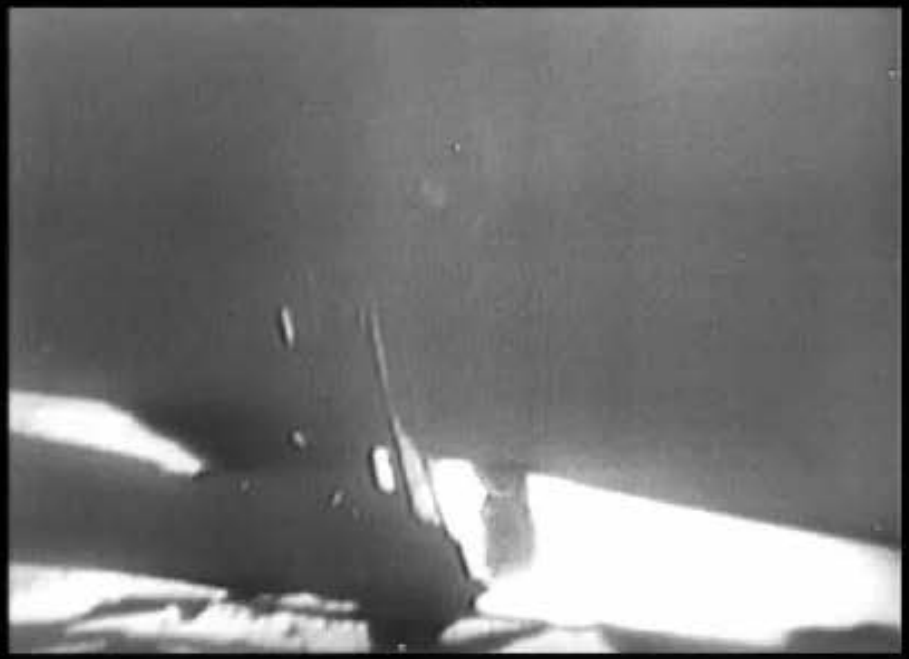
➔ http://www.uni-bielefeld.de/gute_wiss_praxis/

# Langzeitarchivierung



Stein von Rosette

# Apollo 11 EVA Television Comparison 1



NASA Archive

2009 Restoration

Neil Armstrong descends the Apollo 11 Lunar Module ladder

# Filmaufnahmen der Mondlandung

- Apollo 11 sendete einen live stream, konvertiert aus dem slow-scan television (SSTV)
- Die Originalaufnahmen waren auf ca. 45 telemetry data tapes gespeichert
- Im Jahr 2000 suchte die Nasa diese Originalaufnahmen …

  … **fand sie aber nicht!**
- Die Bänder wurden von der NASA Anfang der 1980er gelöscht und wiederverwertet – weil zu dieser Zeit Mangel an Magnetbändern herrschte
- Die überlebenden Sequenzen sind qualitativ geringer wertige Aufnahmen von Fernsehstationen

➔ https://en.wikipedia.org/wiki/Apollo_11_missing_tapes

# Elektronische Langzeitarchivierung

- Unter Langzeitarchivierung (LZA) versteht man die Erfassung, die langfristige Aufbewahrung und die Erhaltung der dauerhaften Verfügbarkeit von Informationen.
- „Langzeit“ = Entwicklung von Strategien, die den beständigen, vom Informationsmarkt verursachten Wandel bewältigen können
- Garantien sind nicht möglich!

# Herausforderungen

- physischer Verfall des Trägermediums
  => “*bitstream preservation*”
- Trägermedium nicht mehr lesbar

# Lebenszeit von Speichermedien

| | |
|---|---|
| floppy disk | 10–30 Jahre |
| HD | 2–10 Jahre (Durchschnitt: 5 Jahre)<br>10-30 Jahre (ohne Nutzung) |
| USB-stick | 10-30 Jahre |
| CD±R | up to 25 Jahre (oft nur 5-10 Jahre) |
| DVD±R | dto. |
| Blu-Ray R | 5-10 Jahre |
| Magnetic Tapes | > 30 Jahre |
| M-Disc | 1.000 Jahre (Herstellerangaben) |

➔ http://de.wikipedia.org/wiki/Langzeitarchivierung

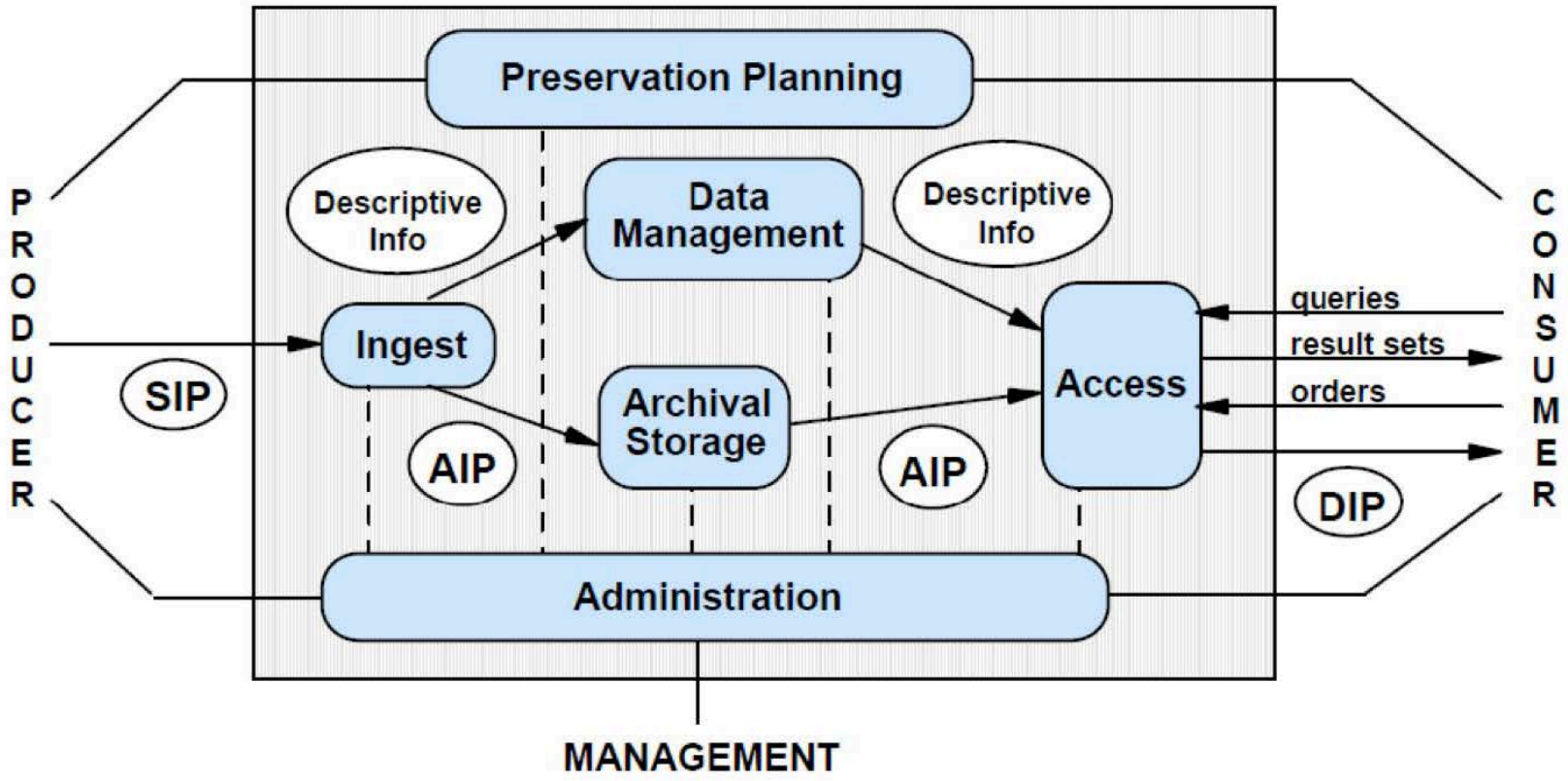
# Weitere Probleme

- veraltete Formate, Software bzw. Betriebssysteme
- proprietäre Formate
- unzureichend dokumentierte Formate
- urheberrechtliche Beschränkungen (z.B. DRM)
- datenschutzrechtliche Vorgaben
- Daten sind nicht auffindbar
- Daten sind nicht mehr verständlich

# OAIS Referenzmodell

- **O**pen **A**rchival **I**nformation **S**ystem
- Initiiert von NASA und ESA
- erster Entwurf 1999
- im August 2012 als ISO „14721:2012” übernommen
- wichtigster Standard für elektronische Archivierung
- Das Referenzmodell beschreibt ein Archiv als Organisation, in dem Menschen und Systeme zusammenwirken, um einer definierten Nutzerschaft Archivgut verfügbar zu machen.
- Die Implementierung eines OAIS-konformen Archivs ist dabei jedoch nicht festgelegt.

# OAIS




➔ OAIS Version 2012: CCSDS Magenta Book: Reference Model for an Open Archival Information System (OAIS)

# nestor

- deutsches Kompetenznetzwerk zur digitalen Langzeitarchivierung
- Kooperationsverbund: Bundesarchiv, GESIS, Deutsche Nationalbibliothek, Landesarchiv Nordrhein-Westfalen, Humboldt Uni, Uni Göttingen , …
- Publikationen: Handbuch, Ratgeber, …
- Fortbildungen
- Arbeitsgruppen: AG Forschungsdaten, AG Zertifizierung, …

# nestor



➔ http://www.langzeitarchivierung.de

# Bisherige Projekte un der UniBI

- Pilot LOCKSS (“Lots of Copies Keep Stuff Safe”) der LibTec
- SFB 673: Archivierung der Daten des 2014 abgeschlossenen SFB

# Anforderungen an eine Archivierung

- Nachweispflichten erfüllen
- geringer Aufwand für die Forschenden
- Anforderungen des Datenschutzes erfüllen

# Ideen

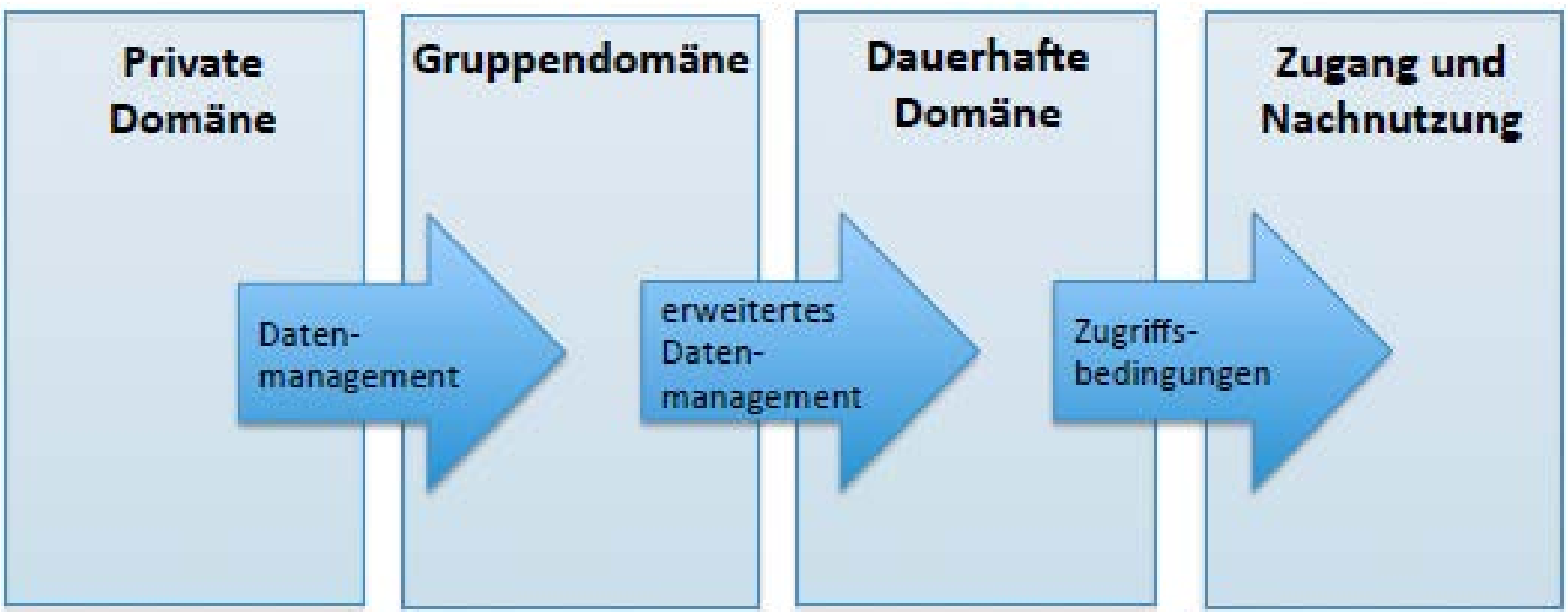
- Die Daten werden publikationsbezogen archiviert.
- Die Daten werden komplett archiviert, d.h. es findet keine Auswahl statt. Minimale Aufarbeiten/Strukturierung + Basis-Metadaten
- Die Daten werden vom HRZ einmal im Monat auf Band gesichert.
- Zugriff auf die archivierten Daten ist nur über schriftlichen Antrag beim Dekan der Technischen Fakultät möglich

# Verfahren

1. Direkt im Anschluss an eine erfolgreiche Publikation (d.h. gleichzeitig mit dem Entrag in PUB) übergibt der/die Forscher/in die Daten an den/die Datenmanager/in
2. Der/die Datenmanager/in stellt die Daten in ein Netzlaufwerk ein.
3. Das HRZ sichert die Daten von dort monatlich auf Band.

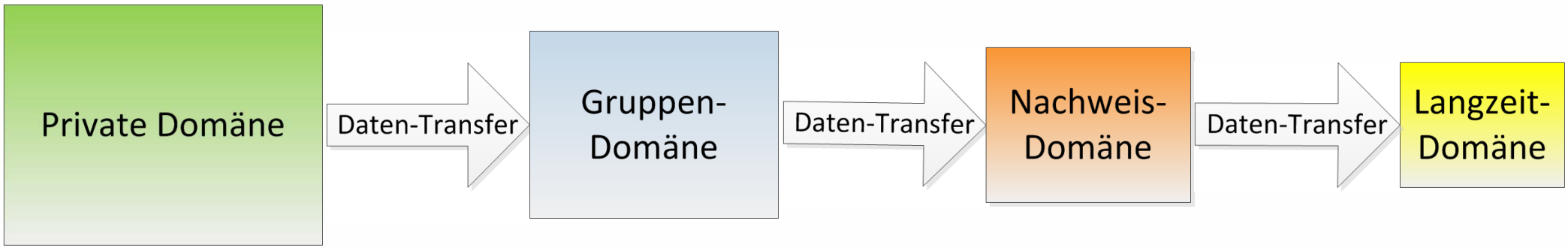
# WissGrid

## Das Curation Continuum




➔ WissGrid: Leitfaden zum Forschungsdaten-Management

# Die Datenveredelung




# Metadaten: Dublin Core

- Sammlung einfacher und standardisierter Konventionen zur Beschreibung von Dokumenten und anderen Objekten im Internet
- weit verbreitet

➔ http://www.langzeitarchivierung.de

# Metadaten: Datacite Schema

- umfangreiches Metadaten-Vokabular für Forschungsdaten
- gut dokumentiert
- wird ständig fortentwickelt
- kann direkt für Datenpublikationen (Beantragung der DOI) verwendet werden
- XML via XSLT in RDF/Linked Data (Semantic Web) umwandelbar

➔ http://schema.datacite.org

# Wo wird gespeichert

**Die Daten werden gespeichert:**

1. durch die Forschenden **an Speicherorten ihrer Wahl**
2. durch die Forschenden **auf gemeinsam genutztem Gruppenspeicher** (z.B. Sciebo)
3. durch den Datenmanager auf geeigneten Speichermedien (mobile hard disc, blu-ray/M-disc, ...) and deposited in the **zentralen, verschlossenen, feuersicheren Schrank** am CITEC.
4. Vom Datenmanager in ein gesondertes Netzlaufwerk eingestellt, um von dort durch das HRZ **auf Band gesichert** zu werden.

# Ordner-Struktur (Vorschlag)

Top-Folder nach PUB-ID benannt (z.B. “`/2681219/`”) enthält:

- `readme.txt`: Info-Datei
- `metadata.xml`: Metadaten im Datacite-Schema
- `/publication/`: die Publikation als pdf
  - `/src/`: Rhodatei bzw. Quelltext + Bilder der Publikation
- `/data/`: all associated research data and software, organized in individual sub-folders as necessary
  - `/raw/`: Rohdaten
  - `/bin/`: compiled software
  - `/src/`: Source Code der Software
  - `/results/`: Ergebnisse der Datenanalyse
  - `/doc/`: Dokumentation

➔ vgl.: Data Carpentry: Good Enough Practices for Scientific Computing

# Datenschutzrechtliche Fragen

- personenbezogene Forschungsdaten dürfen nur für den Zweck genutzt werden, für den sie erhoben wurden.
- nach Ablauf des Forschungsvorhabens sind sie zu vernichten … oder zu anonymisieren.
- In den Einverständniserklärungen muss die Archivierung explizit gennant sein.

# Aktueller Stand

**Bisherige Schritte:**

- Entwurf eines Policy-Dokuments / Anleitung
- fünf Gruppen des CITEC als Testpartner gewonnen (Biologie, Sportwissenschaft, Informatik)
- Anbinden eines Netzlaufwerkes zur Anlieferung der Daten an das HRZ
- Liste der jüngsten Eintrage in PUB

**Nächste Schritte:**

- Info-Veranstaltungen in den AGs
- Nachfassen: Daten aktiv einsammeln

# Bisherige Erkenntnisse

- Einsammeln der Daten von den Forschenden: Schwierig!
- Rechtliche Fragen zu Datenschutz + Archivierung: Juristisch komplex.

# Vielen Dank!

# Kontakt

Cord Wiljes
Tel.: 0521-106-12036
cwiljes@cit-ec.uni-bielefeld.de

➔ https://cit-ec.de/en/content/open-science

➔ http://data.uni-bielefeld.de