ETH BIBLIOTHEK

Forschungsdaten und Open Access: Ordnung ist… ein Teil davon

Open Access Tage 2015

Zürich, 7. und 8. September 2015

Dr. Matthias Töwe, ETH-Bibliothek, ETH Zürich

Überblick

- Ziele
- Worum geht es bei der «Ordnung»?
- Herausforderungen
- Handlungsoptionen

ETH-Bibliothek

- Gegründet 1855
- Hauptbibliothek der ETH Zürich
- Nationales Zentrum für technische und naturwissenschaftliche Informationen
- Sammelgebiete
  - Architektur und Bauwissenschaften
  - Ingenieurwissenschaften
  - Naturwissenschaften und Mathematik
  - Systemorientierte Naturwissenschaften
  - Management- und Sozialwissenschaften

Fachstelle Digitaler Datenerhalt: Unterstützung für Forschende

- **Forschende durch Dienstleistung entlasten**
  - Beratung und praktische Unterstützung des **Datenmanagements** bis hin zu **Veröffentlichung und Langzeitarchivierung**
  - **Rechenschaft und Nachprüfbarkeit** erleichtern
  - **Zitierbarkeit** von Daten gewährleisten
    → **DOI-Registrierung via DataCite**

  - Eigene **Nachnutzung**, Zugriff durch Kollegen bis zu echten ***Open Data***
  - **Durchgängige Services von ETH-Bibliothek und Informatikdiensten**

➢ ***Erwartung: Anforderungen von Förderern und Hochschulen ans Datenmanagement und an die Publikation von Daten wachsen***

Was ist mit «Ordnung» gemeint?

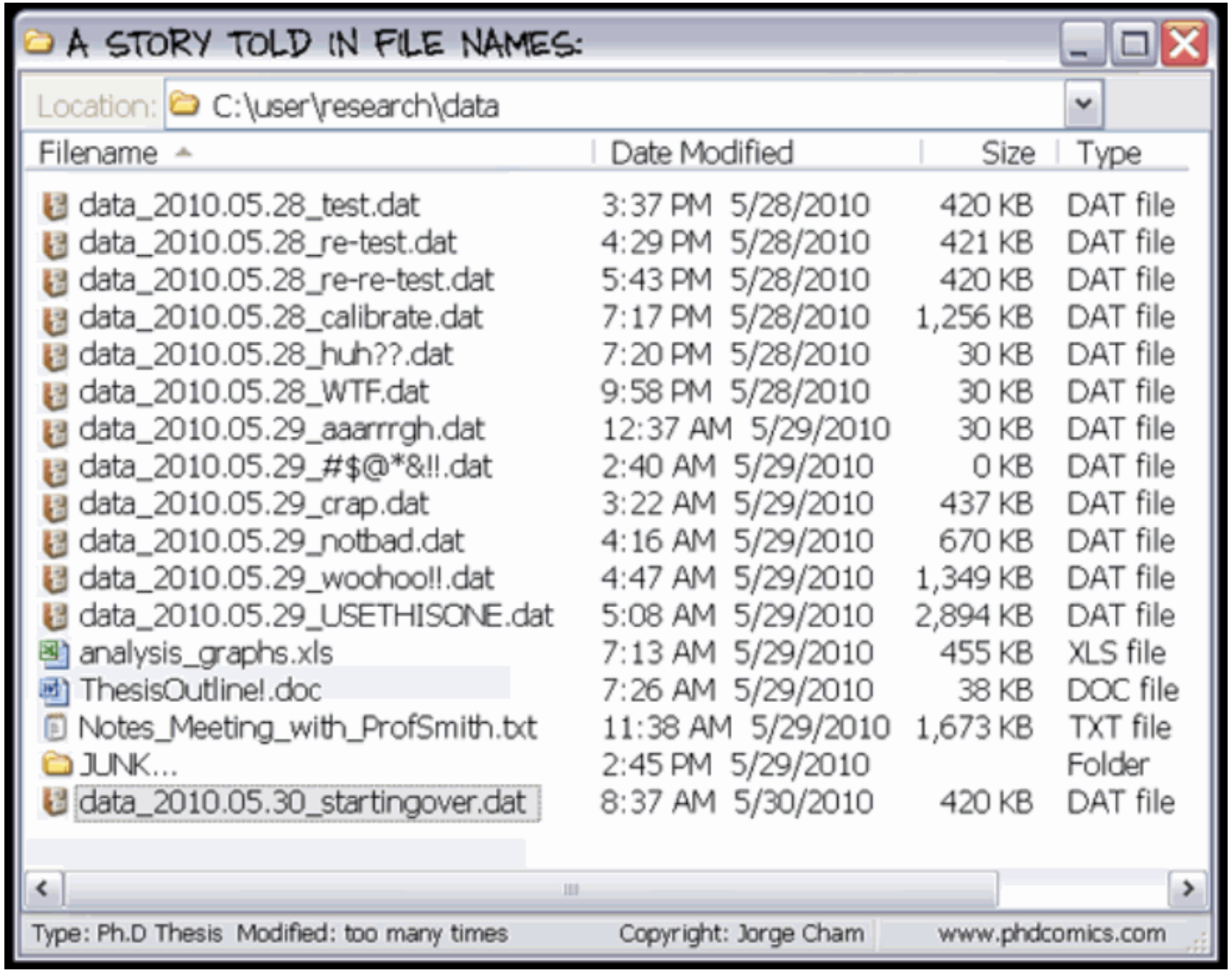
Formal:

- Eine bewusste **Strukturierung** wurde vorgenommen
- Daten sind **identifizierbar bezeichnet**
- Wo nötig, sind **Versionen unterscheidbar**

Inhaltlich:

- **Auswahl** aufgrund nachvollziehbarer Kriterien ist erfolgt
- Inhaltlicher **Kontext** ist für eine Nachnutzung **dokumentiert**
- Interne und externe **Abhängigkeiten sind dokumentiert**

Kommt Ihnen das bekannt vor?



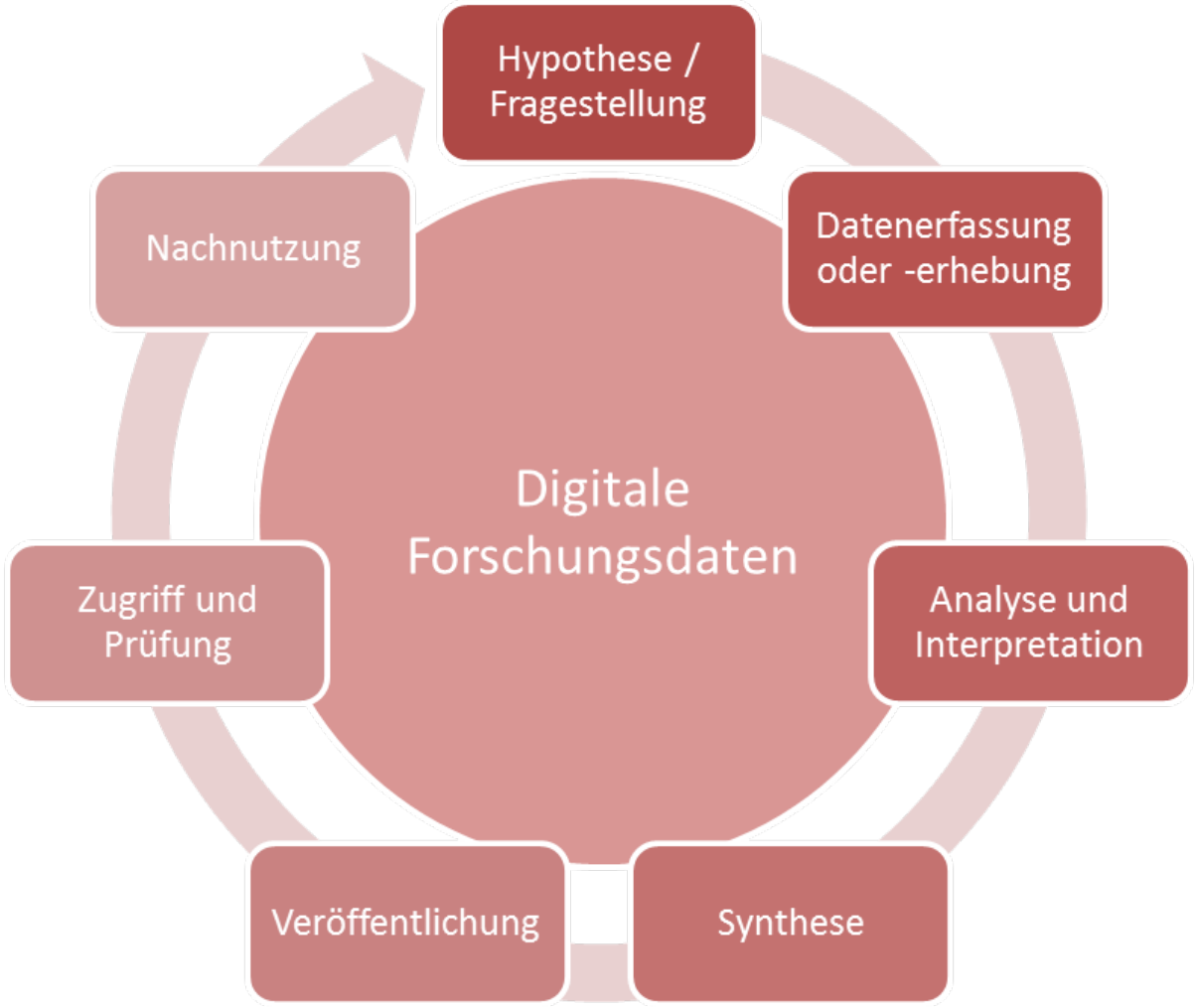
"A story told in file names":

Source:
"Piled Higher and Deeper" by Jorge Cham
www.phdcomics.com
http://www.phdcomics.com/comics/archive.php?comicid=1323

Der Idealfall?

- **Forschungsdaten sind so vorbereitet, dass sie mit wenig Aufwand «auf Knopfdruck» geteilt und publiziert werden können**
- Die Daten liegen **zu einem frühen Zeitpunkt offen und dokumentiert** vor
- Sie können mit angemessener Sachkenntnis und **mit vertretbarem Aufwand** verstanden und bei Bedarf **nachgenutzt** werden
- Die Daten ermöglichen sowohl **Sekundär- als auch Metaanalysen**

Lebenszyklus von Forschungsdaten



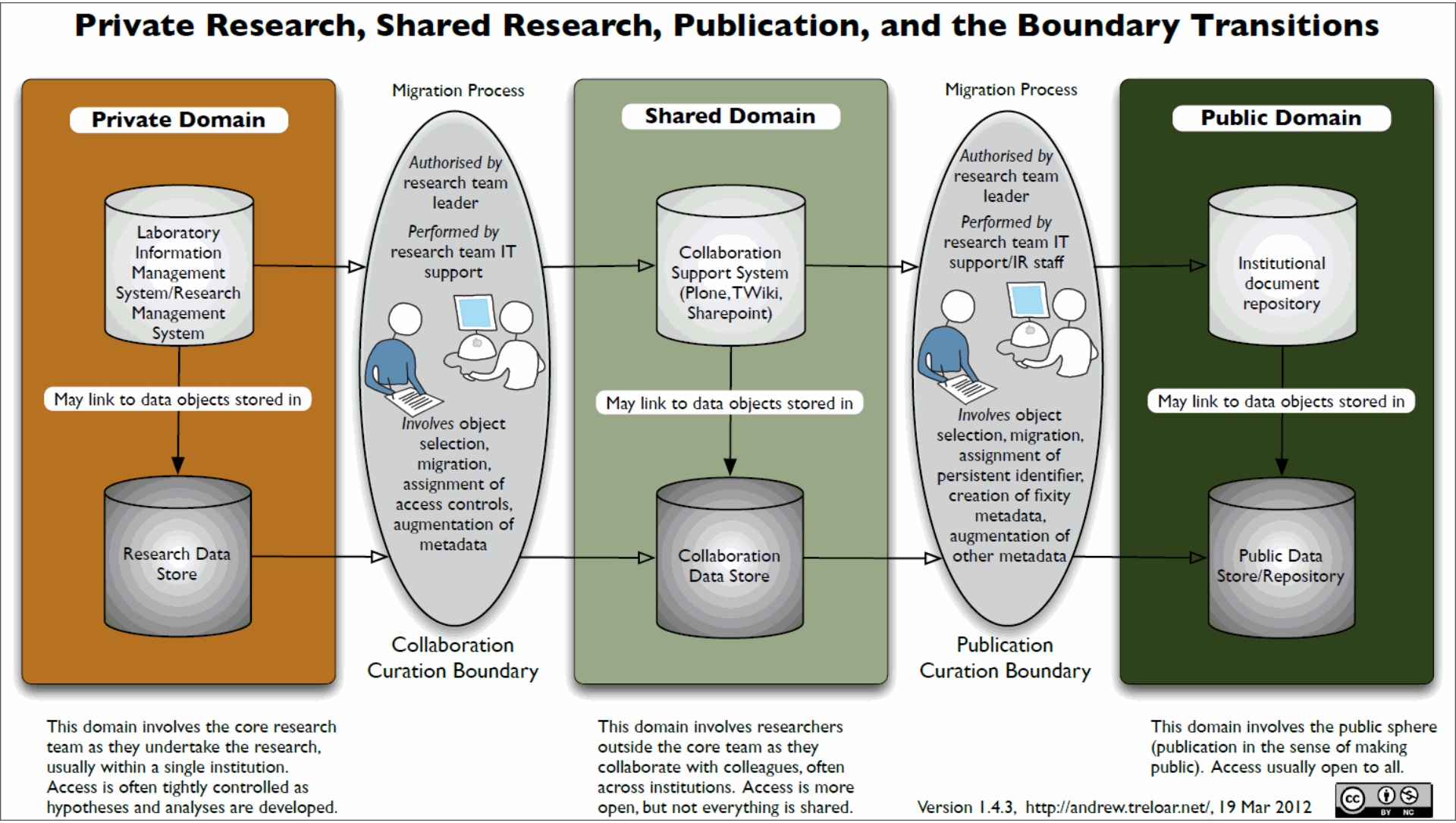
Herausforderungen für Open Data

- **Voraussetzungen je nach Fach extrem un**terschiedlich
- **Augenmerk** richtet sich u.U. zu sehr **auf technische Herausforderungen,** auch im Hinblick auf Langzeitarchivierung
- **Hochwertige Daten können für die Nachnutzung verloren sein**, wenn Kontextinformation fehlt
- **Forschende wissen, dass es** für eine Nachnutzung **nicht reicht, ungeordnete Daten online zu stellen -** und lassen es im Zweifelsfall
- **Qualität von Daten schwerer zu beurteilen** als die von Publikationen

Forschungsdaten sind keine Publikationen

- **Forschungsdaten** werden bisher überwiegend **nicht zum Zwecke der Veröffentlichung** erzeugt
- Die spätere **Veröffentlichung** ist daher meist **mit Zusatzaufwand** verbunden
- **Nachträgliche Verbesserungen** von Ordnung und Erschliessung sind **aufwändig** – und praktisch nur den Produzenten möglich
- Aufwand wird nur betrieben bei entsprechender **Motivation oder Anreiz** – Anerkennung ist noch beschränkt auf bestimmte Fächer
- **Es fehlen eingespielte Prozesse**

Privat oder öffentlich?



(Grafik von Andrew Treloar)

Aus der Frühzeit (?) des Open Access zu Publikationen: Mehr als ein Image-Problem

- Bewusst irreführende oder unbewusst **falsche Gleichsetzung**:
- «***Open Access*** *heisst:*
  Nur elektronisch = Ungeprüft = Unwissenschaftlich = Schlecht»
- Angeblich fehlende **Qualität als Killer-Argument**
- **Akzeptanz darum zäher** als nötig

➢ ***Reale Probleme bei der Aufbereitung von Forschungsdaten*** *können das Konzept von Open Data schädigen und* ***sollten vermieden werden***

Was tun?
Datenmanagementplanung

- Data Sharing und Publishing müssen **Teil der Datenmanagementplanung** zu Projektbeginn sein
  - **Vorgaben** der Förderer, der Community und der eigenen Institution **prüfen**
  - **Was kann und darf später öffentlich gemacht werden?**
  - Welche **Wege zur Publikation** sollen genutzt werden?
  - Gibt es **Folgekosten**?
  - Welches sind **Voraussetzungen**, **um eine Nachnutzung zu ermöglichen**?
  - Sind **Einschränkungen für die Nutzung** nötig?

Was tun?
Datenerzeugung und Ablage

- **Vom Beginn der Datenerzeugung an** muss die Veröffentlichung der Daten mitgedacht und vorbereitet werden
  - **Einhaltung allfälliger Vorgaben** von Dritten (Förderorganisation, Hochschule, Community, Gesetzgeber)
  - Verwendung nachvollziehbarer, dokumentierter **Strukturen**
  - Anwendung transparenter, sprechender oder dokumentierter **Benennungen**
  - **Versionierung**, die eine einfache Auswahl der «richtigen» Daten für die Veröffentlichung erlaubt
  - **Vermeidung von unklaren Redundanzen**

Was tun?
Gruppen-«Policy»

- **Selbstkritische Frage:**
  - **Wie müssen Daten aussehen**, damit wir sie mit **fachlicher Überzeugung und mit Vertrauen** nachnutzen?
  - **Trifft das auf unsere Daten zu? Was fehlt?**
- **In der Gruppe**
  - **Vereinbaren** und schriftlich festhalten, **welche Regeln gelten**
  - Neue **Mitglieder einweisen**
  - «**Geordnet archivieren oder löschen**» - keine Friedhöfe entstehen lassen
  - Umsetzung **überprüfen**

Eine altruistische Übung?

- **Eigenen Nutzen klarmachen**
  - **Die eigenen Daten sind transparent im Griff** – ob öffentlich oder nicht
  - **Daten können innerhalb der eigenen Gruppe besser gehandhabt und weitergegeben werden**
  - **Freigabe** kann **ohne weiteren Aufwand** auch kurzfristig erfolgen
  - **Sichtbarkeit der eigenen Arbeit** durch Open Access zu transparent organisierten und qualitativ relevanten Daten steigt
- **Vorgaben der Förderer und Institutionen können ohne grossen Zusatzaufwand umgesetzt werden**

Grenzen der Planung – wer soll die Zielgruppe sein?

- **«Designated Communities»**
  - **Datenproduzenten treffen explizit oder implizit Annahmen** darüber, wer voraussichtlich mit ihren Daten arbeiten wird und arbeiten können sollte
  - Die **Perspektive des eigenen Fachs** dürfte bestimmend sein
  - Eine **vollständige Unabhängigkeit der Daten** von Grundlagen und Gepflogenheiten des eigenen Fachs **ist nicht realistisch**
  - Eine **spezifische Aufbereitung von Daten** für eine Community mit anderen Methoden und Erwartungen wird allenfalls innerhalb einer vereinbarten Zusammenarbeit erfolgen können, aber nicht «auf Vorrat» ohne Anlass

Grenzen der Planung – Aufwand und Ertrag

- **Echte oder gefühlte Überforderung – und reale Grenzen**
  - **Selbsterfüllende Prophezeiungen:**
    - «Weil zu wenig Daten gut organisiert vorliegen, werden wenig Daten nachgenutzt»
    - «Weil so wenig Daten nachgenutzt werden, lohnt sich der Aufwand nicht»
  - **Je höher und abstrakter die Anforderungen, desto geringer die Bereitschaft**, Zeit zu investieren – **ausser man profitiert selber wahrnehmbar** von insgesamt besserer Qualität
  - **Gegenreaktionen möglich**: *«Wenn meine Daten so nicht gut genug sind, dann behalte ich sie für mich.»*
  - **Den meisten Beteiligten fehlt die Erfahrung**

Grenzen der Planung – das Know-how

- **Wir wissen zu wenig**
  - **Lernprozess in den Communities nötig**: Was funktioniert wirklich über die nächsten Jahre, was fehlt, was ist verzichtbar?
  - Es ist nicht klar, **wie viel Nachnutzung** in welchen Wissenschaftsbereichen wirklich stattfinden wird.
    - Sind die Erwartungen realistisch?
    - In welchen Fächern bzw. für welche Daten ist die Nachnutzung attraktiv?
  - Wo liegt eine für das Wissenschaftssystem und das einzelne Fach **sinnvolle Balance zwischen Aufwand und möglicher Nachnutzung**?
  - ➢ ***Forschende und Dienstleister können und müssen gemeinsam lernen***

Das alles ist – je nach Community - nicht neu

- **Fallbeispiel Poster Nummer 2** (Ana Sesartic)
- **Umsichtiges Datenmanagement** zur Sicherung der Langzeitarchivierung und Wiederverwendung seit **1988**
- **Wissenschaftlich motiviert** und als **integraler Bestandteil sauberen Arbeitens** verstanden
- Durch **Fach begünstigt**
- **Individueller Anspruch**: Konzept und Überprüfung als **«Chefsache»**
- **Probleme resultieren v.a. aus Zeitmangel:** Datenmanagement in **Konkurrenz zu Forschen, Publizieren, Betreuen, Lehren…**

Unterstützung für Forschende in den Hochschulen

- **Frühzeitige Beratung zur Datenmanagementplanung**
  (an der ETH Zürich z.B. durch die Fachstelle Digitaler Datenerhalt)
- **Bereitstellung von Tools**
  - **Datenmanagementplattformen / Forschungsumgebungen**
    (an der ETH Zürich z.B. openBIS der Scientific IT Services)
  - **Elektronische Laborjournale**
  - **Kollaborations- oder Dokumentenmanagementtools**
  - **Andere Tools zur Strukturierung und Metadatierung**
    (z.B. docuteam packer)

➢ ***Problem: Kaum ein Tool lässt sich generalisieren***

Angebote der Fachstelle Digitaler Datenerhalt

- **Beratung Datenmanagement** (im Aufbau)
- **ETH Data Archive** (Ex Libris Rosetta)
  - **Langzeitarchivierung oder befristete Aufbewahrung** für min. 10 Jahre
  - **Access Rights: Open Access / ETH-intern / individuell / zeitverzögert**
  - **Erhaltungsmassnahmen** (Formatmigration)
  - **Massenprozesse und Einzelaktionen**
  - **DOI-Registrierung** für freigegebene Inhalte via DataCite
  - **Metadaten im Wissensportal der ETH-Bibliothek**
- **docuteam packer**
  - **Viewer und Editor für** lokal erstellte **Dateistrukturen mit Metadaten**
  - **Vorbereitung** der Daten **für die Übergabe an das ETH Data Archive**

Vielen Dank!

Fragen?

Dr. Matthias Töwe
Leitung Digitaler Datenerhalt
ETH-Bibliothek
Rämistrasse 101
8092 Zürich
044 632 60 32

matthias.toewe@library.ethz.ch
http://www.library.ethz.ch/Digitaler-Datenerhalt