# Forschungsdaten an der ETH Zürich: Das Projekt *Digitaler Datenerhalt* der ETH-Bibliothek zwischen Datenmanagement und Langzeitarchivierung

**Symposium: "Big Data, Smart Data und das semantische Web"**

**Bern, 21. November 2012**

**Dr. Matthias Töwe**



ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

# GLIEDERUNG

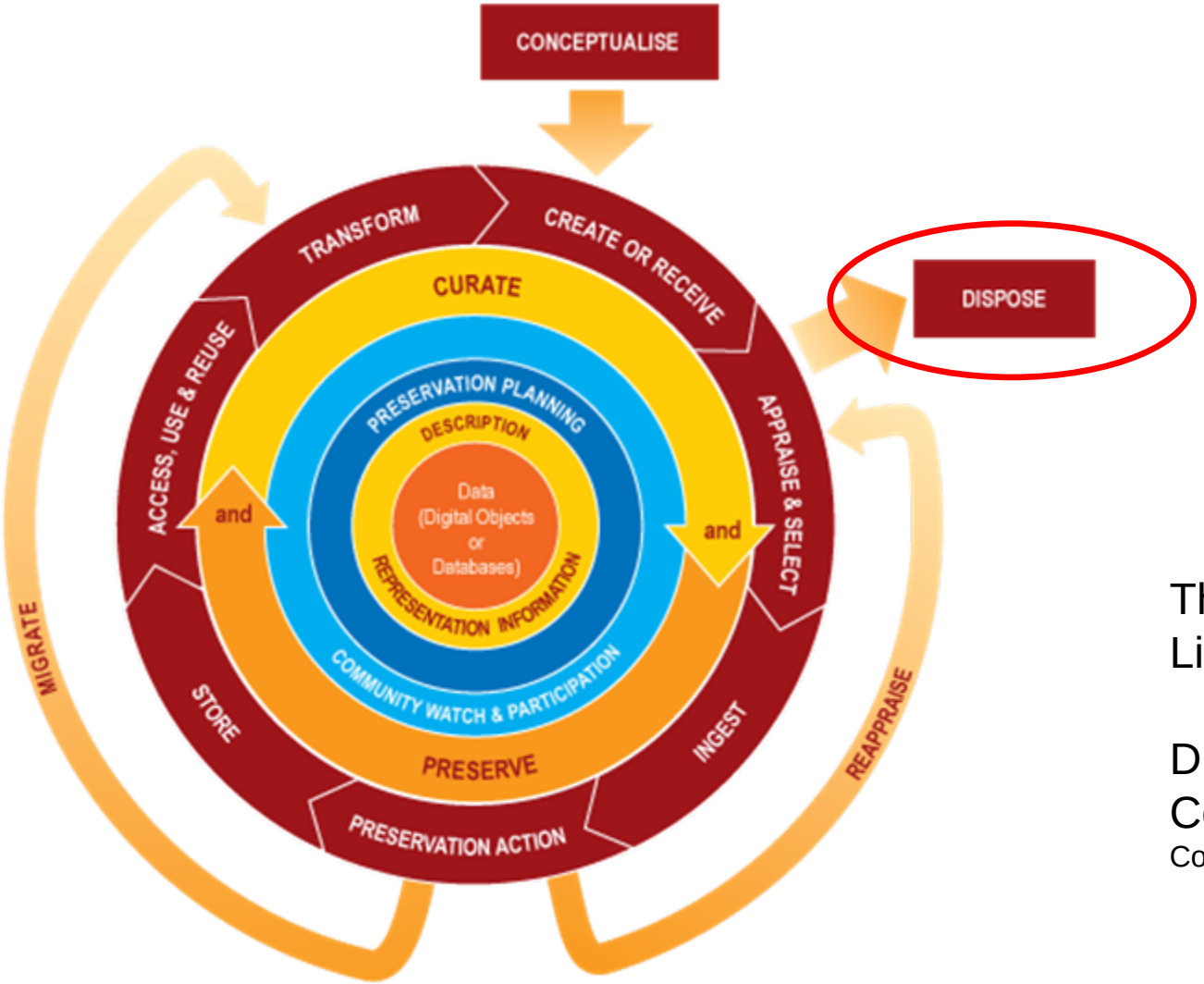
1. Hintergrund
2. Ziele
3. Anforderungen an die Erhaltung
4. Lokales Datenmanagement und zentrale Langzeitarchivierung
5. Aktueller Stand

# HINTERGRUND

## Herausforderungen

- **Forschungsprozess stützt sich umfassend auf digitale Daten**
- **Gute wissenschaftliche Praxis** verlangt Aufbewahrung von Daten in nutzbarer Form (z.B. Richtlinien ETH Zürich)
- **Förderorganisationen fordern Datenmanagementpläne**
- Z.T. nicht wiederbeschaffbare **Daten mit dauerhaftem Wert**
- **Veröffentlichte Daten oder referenziertes Zusatzmaterial** müssen zitierbar sein und verfügbar bleiben

# LEBENSZYKLUS




The DCC Curation Lifecycle Model

Digital Curation Centre, UK

# TRENDS

**Veränderungen in der Wahrnehmung von Daten**

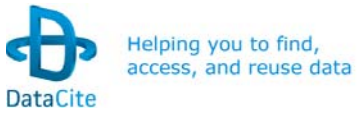
- **Starke Abhängigkeit vom jeweiligen Fachgebiet** mit seinen Methoden und Traditionen → *keine Verallgemeinerungen*
- **Bewusstsein des potentiellen Wertes von Daten über die erste Publikation hinaus**
  - Auswertung mit **anderen Methoden** und **durch andere Forschende**
  - Erleichtert durch **Online-Austausch, bessere Zitierbarkeit**
- **Bereitschaft zur öffentlichen Bereitstellung wächst - *langsam***

# «BIG» ODER «SMALL»?

**Innerhalb der Hochschule**

- **Daten** werden überwiegend **in Einzelprojekten und kleineren Kooperationsprojekten** produziert → *Heterogenität*
- **Konzepte für «Big Data» nicht immer übertragbar**
- **Interessen der Hochschule nicht immer im Einklang mit denen der *einzelnen* Forschenden**
- Beträchtliche **Risiken** sind **nicht technischer Natur…**
  - Verlust Kontextinformation, fehlende Dokumentation, unklare Versionen und Redundanzen, Lücken…
- **…und nur zum Teil mit technischen Mitteln**

# ZIELE

- **Forschende durch Dienstleistung entlasten**
  - **Datenmanagement ←→ Langzeitarchivierung**
  - Rechenschaft und **Nachprüfbarkeit erleichtern**
  - **Zitierbarkeit** von Daten gewährleisten
    → DOI-Registrierung bereits in Betrieb
    
    
  - Eigene **Nachnutzung,** Zugriff durch Kollegen bis zu echten *Open Data* **unterstützen**
  - **Durchgängige Services** aus Sicht der Forschenden (*one stop*)
  - ➢ *Erwartung:* ***Anforderungen*** *von Förderern und Hochschulen* ***ans Datenmanagement wachsen***

# BEDÜRFNISSE VON FORSCHENDEN*

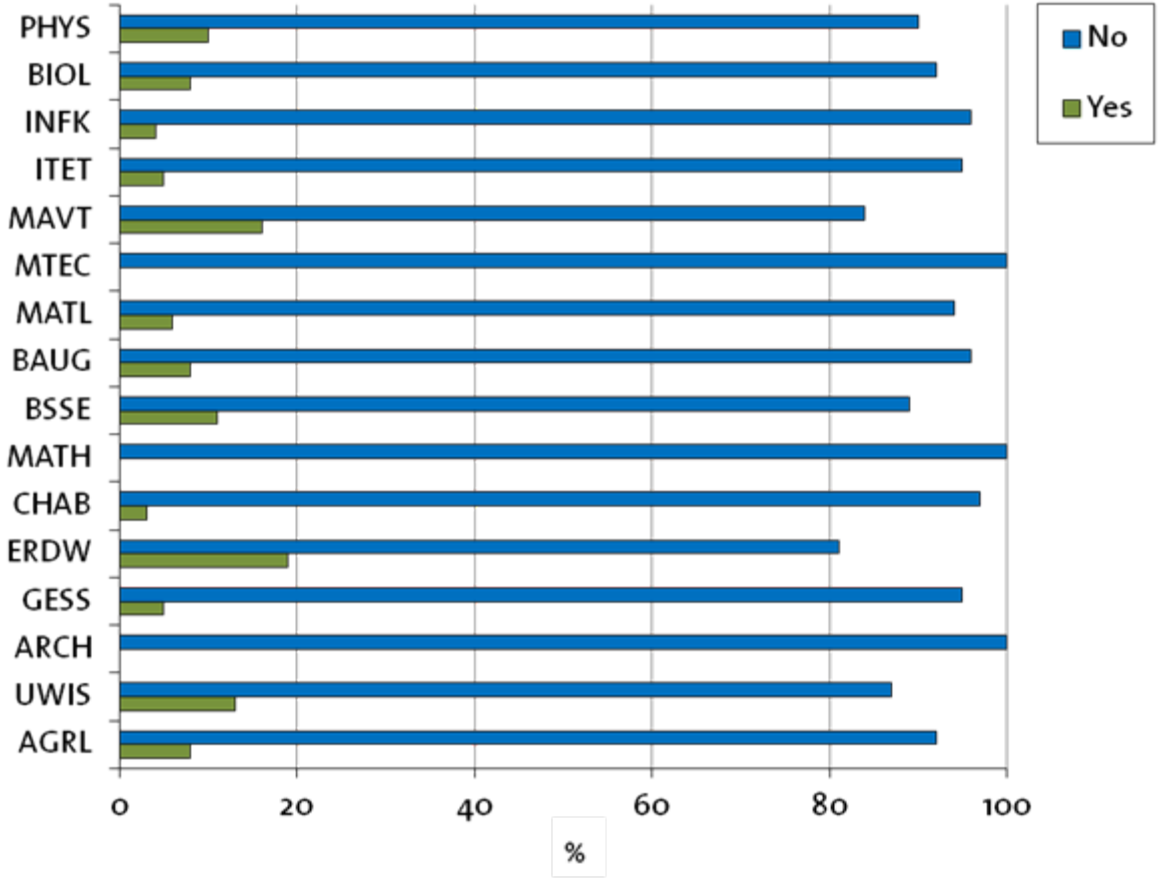
**Sehr viele Forschende…**

- müssen Daten für **begrenzte Zeit** aufbewahren (z.B. 10 bis 12 Jahre)
- möchten **publizierte Daten dauerhaft referenzieren und archivieren**
- möchten Daten vor dem Ingest **umstrukturieren, auswählen**, **dokumentieren**
- möchten **Metadaten editieren und Daten hinzufügen**
- möchten **Kontrolle** darüber behalten, wer auf ihre Daten zugreift…
- …und **stellen ihre Daten nicht generell für Dritte zur Verfügung**
- sind **interessiert an Unterstützung bei Erhaltung und Qualitätskontrolle**
- wollen **keinen zusätzlichen Arbeitsaufwand** ohne Mehrwert

**(Gemäss Umfrageantworten von 80% der Forschungsgruppen und Diskussion mit Pilotpartnern)*

# UMFRAGE: POLICIES DER FORSCHUNGSGRUPPEN

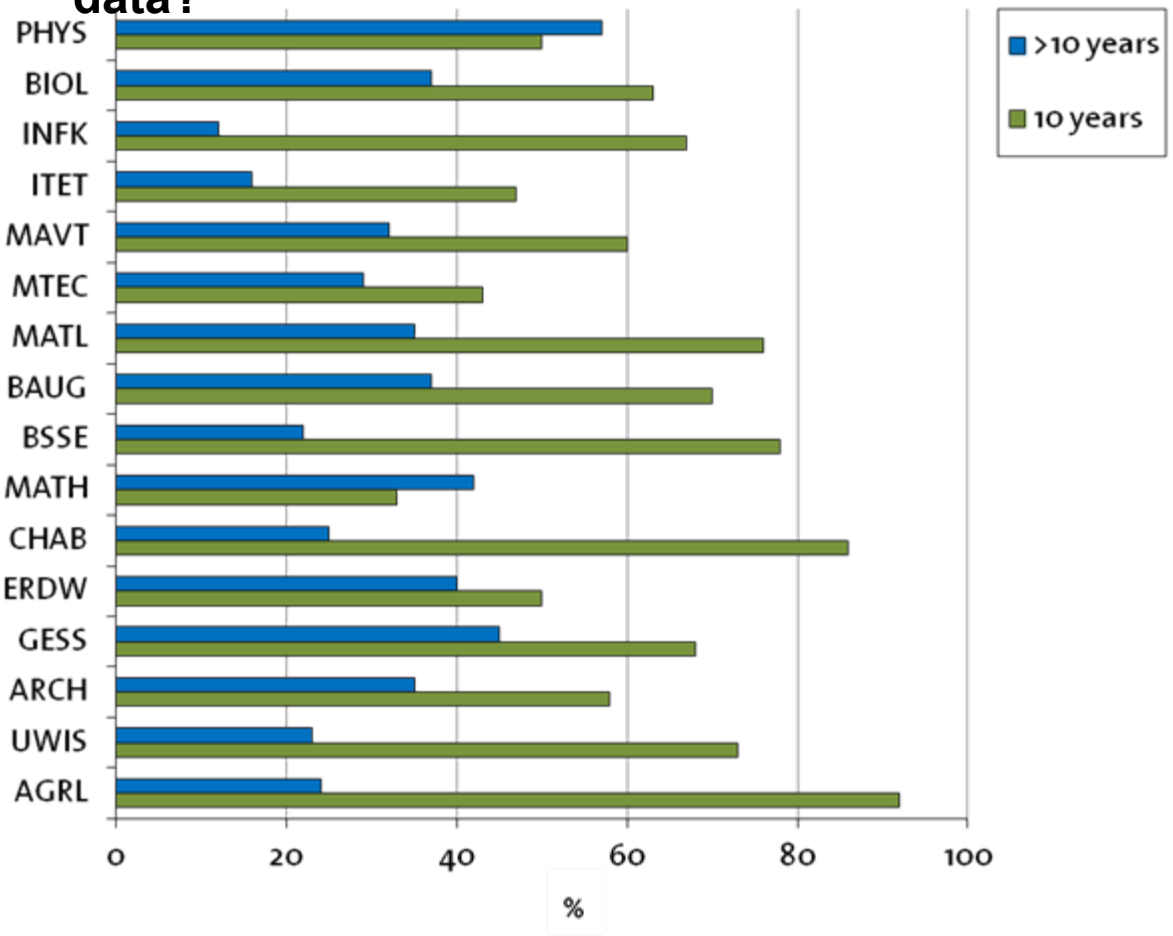
1. ***Part*** **- Does your research group have a written 'Data management policy' on handling its own or external research data, or similar documents describing how to handle research data?**




Survey and diagrams:
S. Scheid

# UMFRAGE: AUFBEWAHRUNGSDAUER

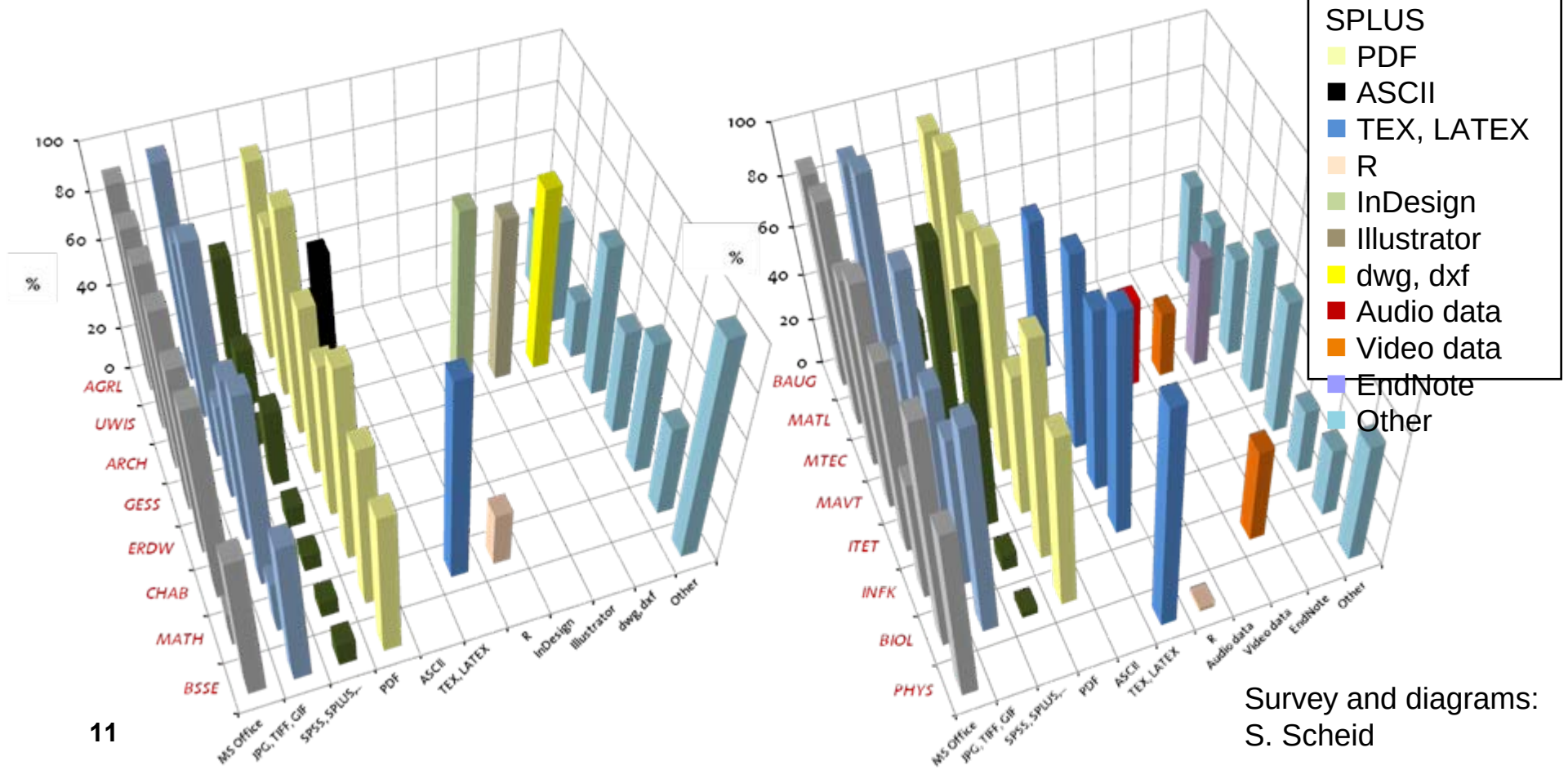
2. *Part* - **How long a period do you or your research group have in mind for storing data?**




Survey and diagrams:
S. Scheid

# UMFRAGE: DATEIFORMATE / SOFTWARE

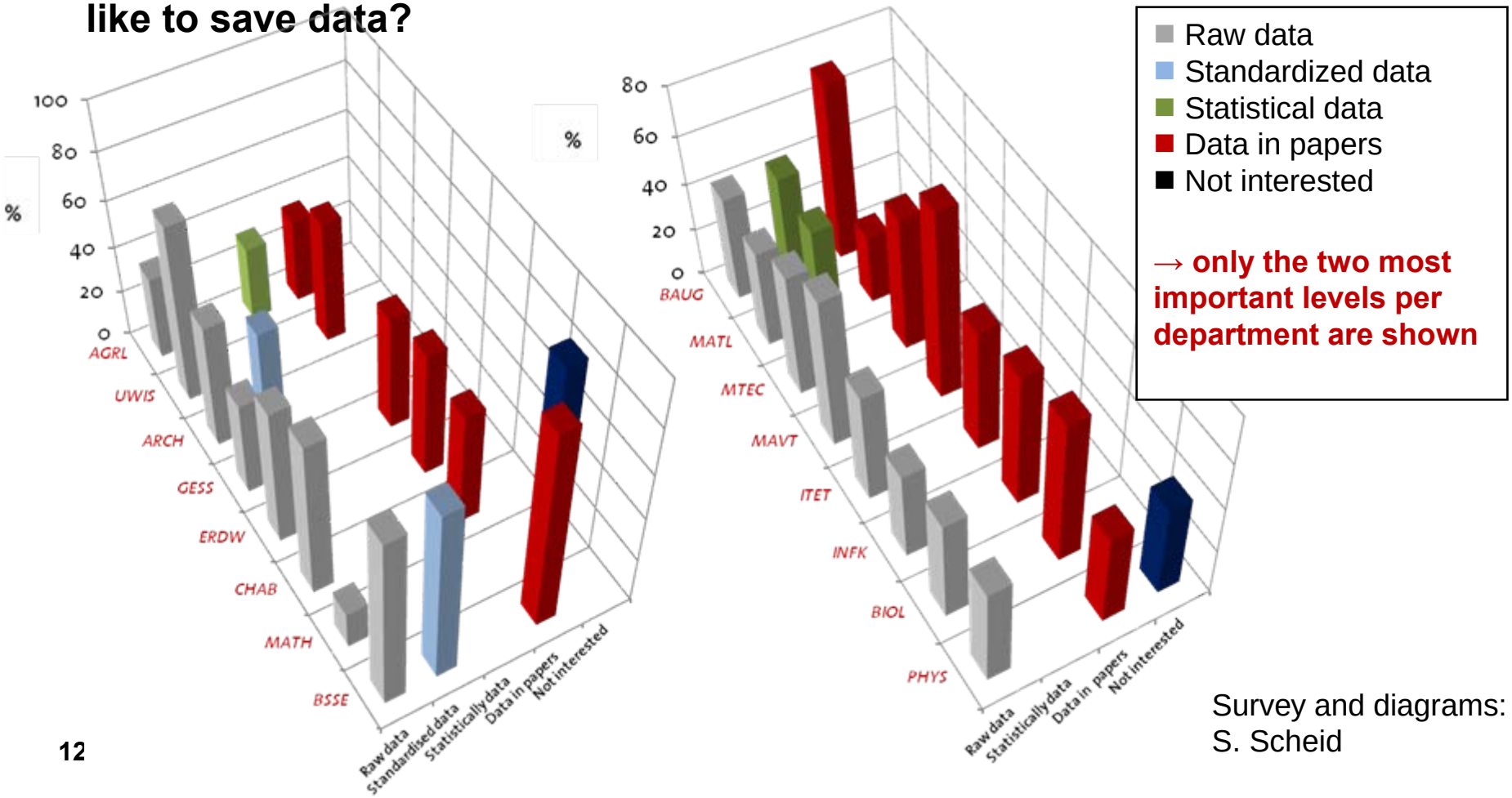
**2. *Part* - Which formats do you and your research group work wit**




Survey and diagrams: S. Scheid

# UMFRAGE: STATUS DER DATEN

**2. *Part* - If the ETH-Bibliothek were to provide you with a database for storing your research data, at which level would you or your research group like to save data?**

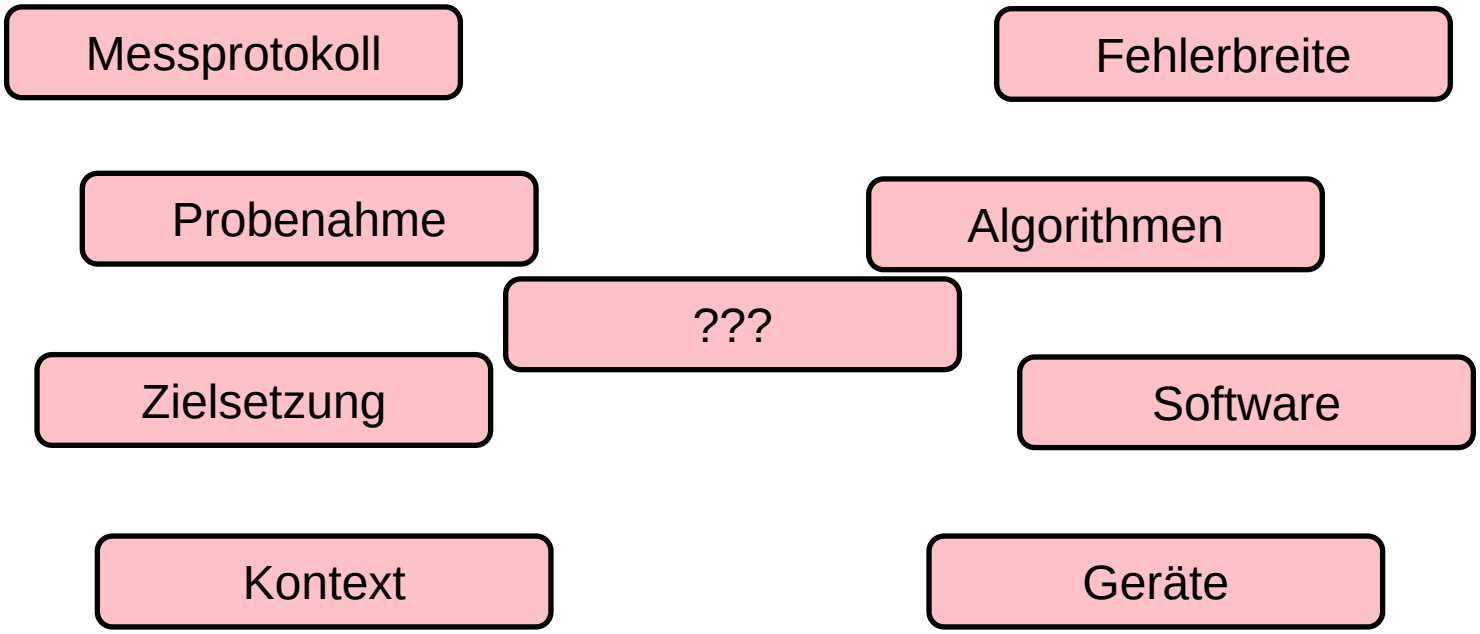



Survey and diagrams:
S. Scheid

# INHALTLICHER KONTEXT

**Umfassende Dokumentation notwendig:**

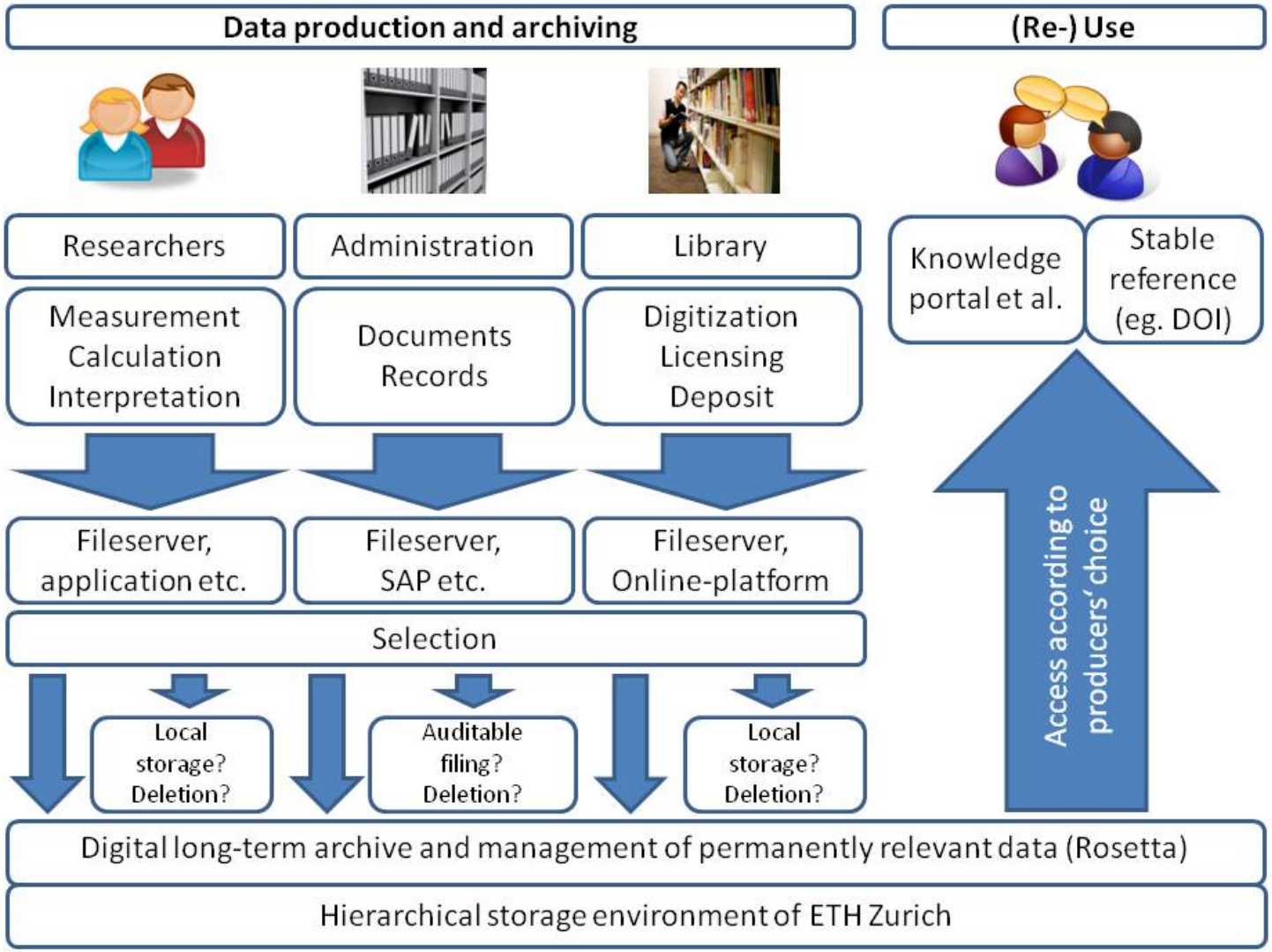
- **Beispiel: apparative Messung einer physischen Probe**
  **→ *idealerweise* durch Publikationen abgedeckt, *real* nur bedingt**



# UNTERSCHIEDE ZWISCHEN DEN DATENTYPEN?

| Was? | Forschungsdaten | Bibliotheksobjekte |
|---|---|---|
| Data Curation | Comprehensive documentation by producers required | Full control of metadata and context |
| Content Preservation | More and less common formats | Mainly standard formats |
| | Same preservation procedures apply | |
| Bitstream Preservation | „Any object is just bits“ | |

# VISION: ROSETTA ALS GEMEINSAME BASIS



# DATENMANAGEMENT UND OAIS?

- **Mit Pilotpartnern erarbeitete Anforderungen** betreffen kaum Funktionen innerhalb des OAIS-Rahmens
- Dagegen hohe **Flexibilität erforderlich im Pre-Ingest** oder davor

➢ **Auswirkungen auf die Rolle des LZA-Systems Rosetta:**

- **Wenig sinnvoll, Komplexität für die Langzeitarchivierung durch neue Funktionen weiter zu erhöhen**
- Sofern vorhanden, **Daten aus vorgelagerten Anwendungen übernehmen**
- Bei Bedarf **Flexibilität im lokalen Datenmanagement** erreichen, nicht in zentraler Anwendung

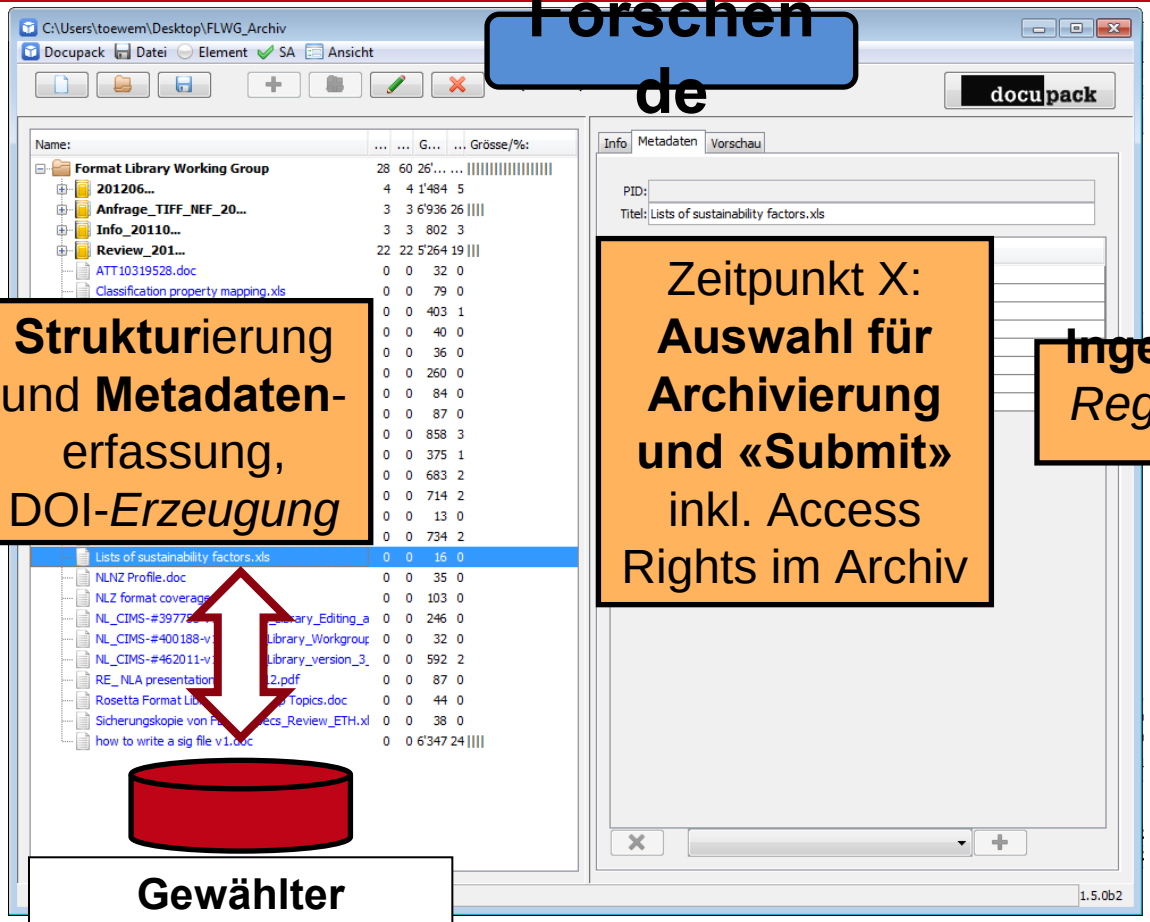
# LOKALES DATENMANAGEMENT?

- **SIP Package Handler** («Docupack» (Java))
- **Viewer und Editor** für Strukturierung und Metadatenerfassung
- **Daten und Metadaten** beliebig lange **lokal** auf eigenem Speicher jeder Forschungsgruppe
- **Erzeugt** bei Anstoss der Archivierung **SIP (Submission Information Package)** für den Ingest in Rosetta
- **Herkunft aus dem archivischen Bereich** kein Zufall: Anforderungen sehr ähnlich


ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

# «UNDER CONSTRUCTION»

**Forschende**




**Strukturierung** und **Metadaten**-erfassung, DOI-*Erzeugung*

Zeitpunkt X: **Auswahl für Archivierung und «Submit»** inkl. Access Rights im Archiv

**Ingest**, DOI-*Registrierung*

**Bibliothek**

**Rosetta**

**Lieferung** via DOI, wenn Rechte ok

**Gewählter Speicher** (lokaler PC, Share); Berechtigungen über Dateisystem

# ANWENDUNGEN IN DER PRAXIS

- **Einheitliche Templatestruktur** z.B. für Doktorierende einer Gruppe
- Nutzung für **Daten in Verbindung mit einer Manuskripteinreichung**
- **Kontinuierliches Befüllen** während eines Projekts
- **Import** einer kompletten vorhandenen **Ordnerstruktur mit Dateien**
- **Gemeinsame Nutzung** innerhalb der Gruppe
- **Ablieferung zur Archivierung nach Bedarf**
- **Vorteil für Nachnutzung und Langzeitarchivierung:**
  **METS-Output** enthält Strukturinformation und weitere Metadaten und kann **automatisch weiter verarbeitet werden**

# ERWARTUNGEN AN DIE DIENSTLEISTUNG

- **Gewichtung von Services gegenüber Softwarebereitstellung**
- Bedarfsorientierte, **individuelle Konfiguration vs. Handhabbarkeit und Automatisierung** der Gesamtlösung
- Nicht **einfach ein weiteres Tool** zur Verfügung stellen
- Forschende erwarten **echte Unterstützung und Services**, die ihre Arbeit erleichtern…
- … und eine möglichst **nahtlose Integration** in ihre Arbeitsabläufe
- … sowie Lösungen nicht nur für Forschungsdaten, sondern auch für ihre **administrative Unterlagen**

# STRATEGISCHE FRAGEN

- Wie sieht das **Profil der Dienstleistung** «Datenerhalt» aus?
- **Bibliothek verändert ihre Rolle** noch stärker von der Informationsversorgerin hin zur Dienstanbieterin
  → bewusste Gestaltung nötig
- **Faktoren für die Akzeptanz** bei Forschenden?
- Positionierung der **Bibliothek in** einem stark **IT-lastigen Arbeitsfeld?**
- **Geeignete Partner** innerhalb und ausserhalb der Hochschule?
- **Langfristige Kosten?**
- **Kosten** der aufwändigen Dienstleistung müssen **gedeckt** werden **–** gleichzeitig soll das Angebot **keine Kostenbarriere** aufweisen

# GLOBALE HERAUSFORDERUNGEN

- **Formatanalyse, -konvertierung und Ergebnisvergleich**
  - Verfahren noch **nicht ausgereift**
  - **Know-how wenig verbreitet**
- **Daten und ihr Management in dauernder Entwicklung**
- **Langzeitarchivierung als Prozess** ohne einmalige und dauerhafte «Lösung», sondern als **Gegenstand der Forschung**
- **Wie geeignetes Personal aufbauen?**
  Forschungsnah, bibliothekarisch und archivarisch, IT-affin…
- **Kultur und Methoden der qualifizierten Nachnutzung fördern?**


ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

# BISHERIGE UND AKTUELLE SCHRITTE

- ✓ **DOI-Registrierung** durch ETH Zürich als Mitglied von DataCite
- ✓ **Umfrage bei** allen Forschungsgruppen (Profs.) der ETH Zürich
- ✓ Identifizierung von **Pilotpartnern; Ermittlung ihrer Anforderungen**
- ✓ Prüfung und Aktualisierung **des Inventars von Daten der Bibliothek**
- ✓ **Submission application** für die ETH E-Collection
- ✓ **«Gap analysis»** und Entscheidung über Entwicklungen

- o **Entwicklung und Testen** der Erweiterungen für Rosetta
- o **Entwicklung und Testen** des lokalen Datenmanagements

**Sofern erfolgreich:**

- o **Ausdehnung der Abdeckung** auf weitere Gruppen
- o Übergang vom Projekt zur **produktiven Dienstleistung**

# 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