# Zwischen Datenmanagement und Langzeitarchivierung:

## Das Projekt «Data Curation» der ETH-Bibliothek

**Dr. Wolfram Neubauer/ETH Zürich, ETH-Bibliothek und Sammlungen/Berlin, 29.04.2013**

# ALLGEMEINE RANDBEDINGUNGEN AN DER ETH ZÜRICH

- mittelgrosse Universität mit etwa 17.000 Studierenden
- ETH Zürich ist eine «Research University»
- die Verantwortung für die Sicherung digitaler Daten liegt beim einzelnen Wissenschaftler/ der einzelnen Wissenschaftlerin
- d.h.: formal ist die Sicherung geklärt, real ist die Sicherung desparat, unklar, unübersichtlich

# SITUATION AUS SICHT DER ETH-BIBLIOTHEK (UM DAS JAHR 2002)

- die digitale Transformation Jahre erfordert eine verstärkte Fokussierung auf den Langzeitzugriff bzw. die Langzeitsicherung elektronischer Daten (Forschungsdaten/administrative Daten/Bibliotheksdaten)
- kein «Produkt von Stange» auf dem Markt
- keine Kapazitäten für eine Eigenentwicklung
- Diskussion: Wer ist für die Aufgaben an einer Universität zuständig?

# HINTERGRUND/RANDBEDINGUNGEN

## Herausforderungen

- Forschungsprozess stützt sich auf digitale Daten
- Gute wissenschaftliche Praxis verlangt Aufbewahrung von Daten in nutzbarer Form (z.B. Richtlinien ETH Zürich)
- Teilweise gesetzliche Vorgaben zur Datenarchivierung
- Förderorganisationen fordern Datenmanagementpläne
- Nachnutzung soll gefördert werden
- teilweise nicht wiederbeschaffbare Daten mit dauerhaftem Wert
- Veröffentlichte Daten oder referenziertes Zusatzmaterial müssen zitierbar sein und verfügbar bleiben

# ZIELE DES PROJEKTES «DATA CURATION» AN DER ETH ZÜRICH (1/2)

## Forschende durch Dienstleistung entlasten

- Erfüllung von Anforderungen der Förderorganisationen (Datenaufbewahrung und -bereitstellung, Nachprüfbarkeit)
- Zitierbarkeit von Daten gewährleisten
  - DOI-Registrierung
- Nachnutzung erleichtern:
  Nach Bedarf für eigene Zwecke, für Kollegen oder global

# ZIELE DES PROJEKTES «DATA CURATION» AN DER ETH ZÜRICH (2/2)

## Forschende durch Dienstleistung entlasten

- Notwendige Dokumentation und Beschreibung unterstützen
- Keine Konkurrenz zu fachspezifischen Datendiensten
- Ergänzende interne Lösungen für Fächer ohne solche Datenarchive sowie für Daten, die nicht in globale Gefässe sollen oder dürfen
- Beratung und Unterstützung bei der Nutzung externer oder interner Dienste, bei der Auswahl langzeittauglicher Formate und bei der Qualitätskontrolle

# WARUM INSTITUTIONELLE ANSÄTZE?

- Brauchen «uns» die Forschenden?
  - Forschende können viele Aufgaben selbst erledigen…,
  - …was aber zu Lasten ihres «Kerngeschäfts» geht

➢ *Infrastruktureinrichtungen sollten sie entlasten*

- Fachspezifische Repositorien und Dienstleister
  - …gibt es nur für bestimmte Fächer…
  - …machen Vorgaben zur Veröffentlichung…
  - …beschränken zum Teil die Art der abzuliefernden Daten

➢ *Ein erheblicher Teil der Bedürfnisse ist nicht abgedeckt*

# WARUM LOKALE ANSÄTZE?

- Abgrenzung zu internationalen fachspezifischen Angeboten
  - keine Konkurrenz zu bereits akzeptierten Lösungen
  - potentielle Nutzer informieren, wenn nicht bekannt

➢ *«Fachcommunity» entscheidet, was ihre Lösungen sind*

- Nationale Infrastruktur
  - fehlt mehr oder weniger komplett
  - Akzeptanz heute eher gegeben, als vor vielleicht fünf Jahren…
  - …aber lokale Ansprechpartner werden dadurch eher wichtiger

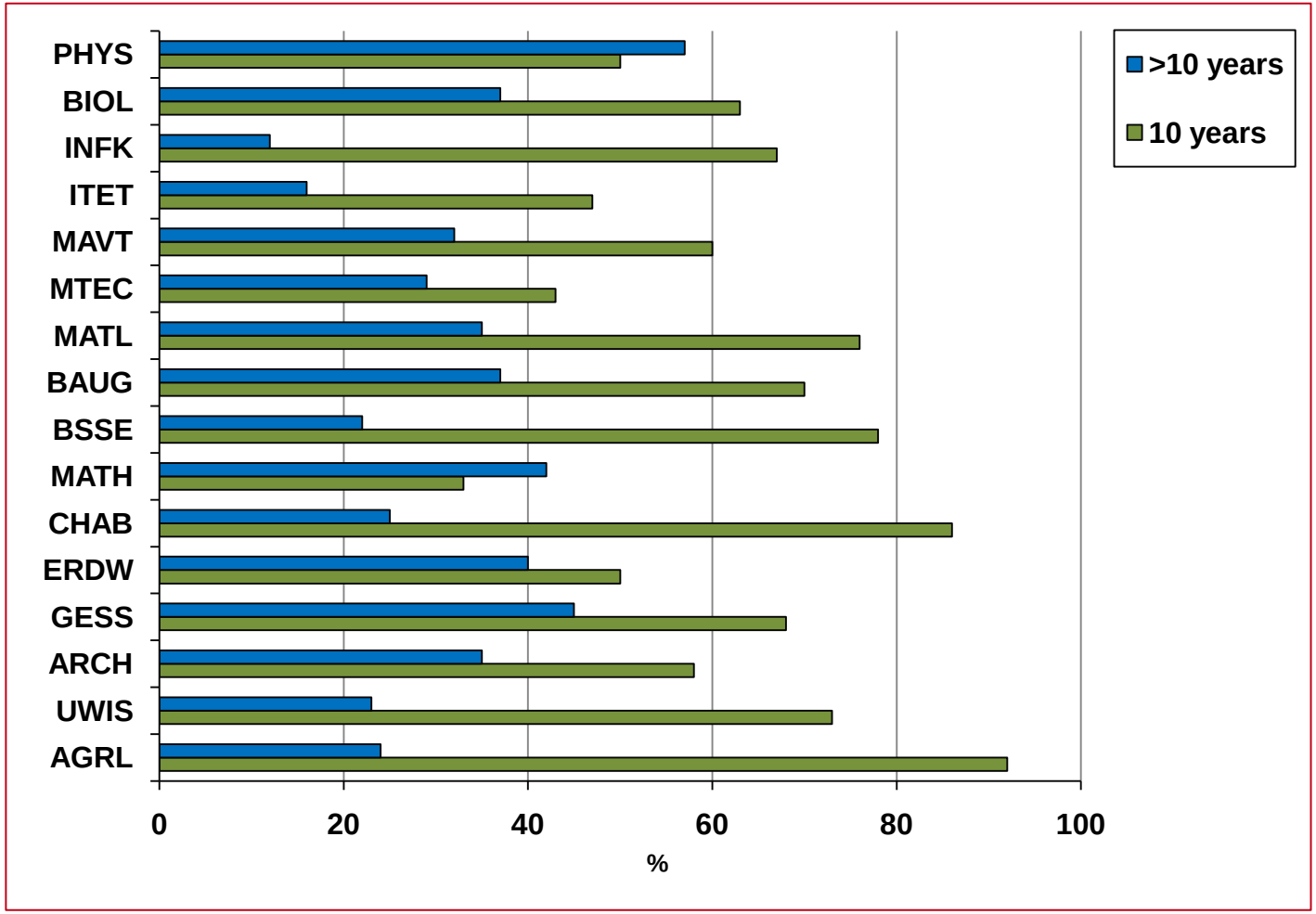
➢ *Implikationen auch für SUK-Programm «Wissenschaftliche Information»: Betrachtung nicht auf technische Aspekte reduzieren*

# KONKRETE BEDÜRFNISSE DER FORSCHENDEN: DIE UMFRAGE

- befragt wurden 450 Professorinnen/Professoren und Forschungsgruppen (Rücklauf ca. 80%) im Jahr 2011
- Fragen
  Datentypen/Gibt es Metadatenstandards?/Gibt es überhaupt Datenmaterial?/Gibt es «Data Policies»?/Welche Datenformate sind im Einsatz?/Notwendige Aufbewahrungszeiträume/Wie und in welcher Form werden die Daten archiviert?/usw.
- Interpretation der Antworten als Handlungsempfehlung

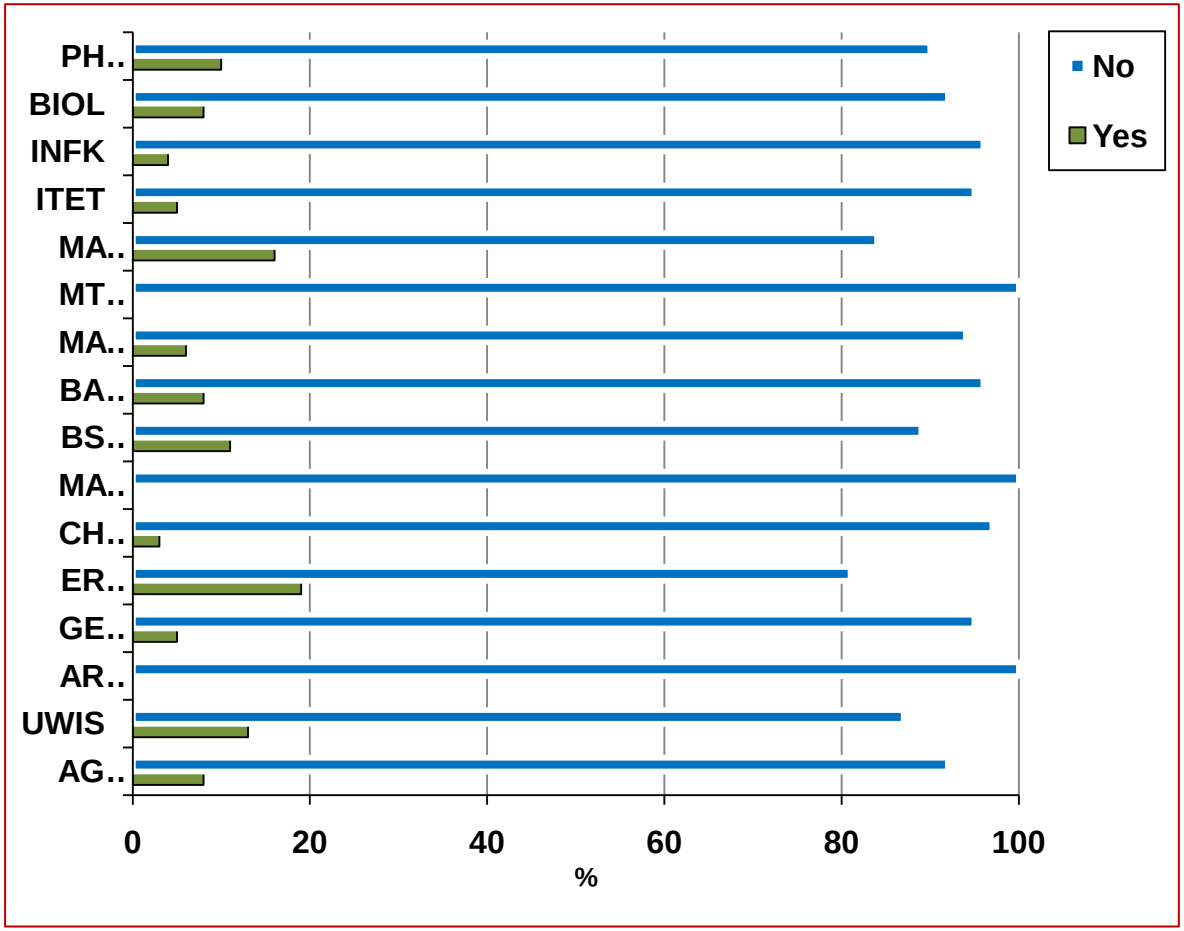
# UMFRAGEERGEBNISSE: BEISPIEL 1/2

**Question: Which period of time have you or your research group in mind for storing data?**



# UMFRAGEERGEBNISSE: BEISPIEL 2/2

**Question: 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?**



# RELEVANTE ASPEKTE/RANDBEDINGUNGEN FÜR DIE REALISIERUNG EINER APPLIKATION

Viele Wissenschaftlerinnen/Wissenschaftler…

- möchten die Kontrolle, wer auf ihre Daten zugreift (auch wenn sie Open Data im Prinzip unterstützen)
- stellen ihre Daten nicht generell Dritten zur Verfügung
- sind in ganz unterschiedlicher Intensität mit dem Phänomen vertraut
- sind an einer Unterstützung bei der Datensicherung und der Qualitätskontrolle interessiert (Checklisten, Hilfestellung bei der Metadatengenerierung etc.)
- möchten ihre Daten vor dem Ingest umstrukturieren, seklektieren und dokumentieren können
- benötigen die Daten häufig nur für begrenzte Zeit (z.B. 10 bis 12 Jahre)
- betrachten Archivierungsfragen häufig im Kontext zu Eigenpublikationen und möchten Daten/Materialien dauerhaft referenzieren
- haben eine differenzierte Meinung zu einer universitätsweiten Data Policy
- möchten keinen zusätzlichen Arbeitsaufwand ohne Mehrwert

# WAS VERSTEHEN DIE VERSCHIEDNEN STAKEHOLDER UNTER «ARCHIVIERUNG»?

## Es besteht eine erhebliche Begriffsunschärfe

- Forschende?
- Informatik(-dienste)
  - ein gemeinsames Verständnis wurde durch frühere Projekte gefördert
  - Weiterhin viel Gewicht auf Speichermanagement: Langzeitarchivierung als ein Kriterium zur Steuerung eines hierarchischen Speichermanagements (Entlastung von Online-Speicher)
- «Gedächtnisinstitutionen»
  - Langzeitarchivierung adressiert Aufbewahrung und Nutzbarerhaltung über Lebenszyklen der technischen Komponenten hinweg (Formate, Software, Betriebssysteme, Hardware)

➢ *Konsequenz:*

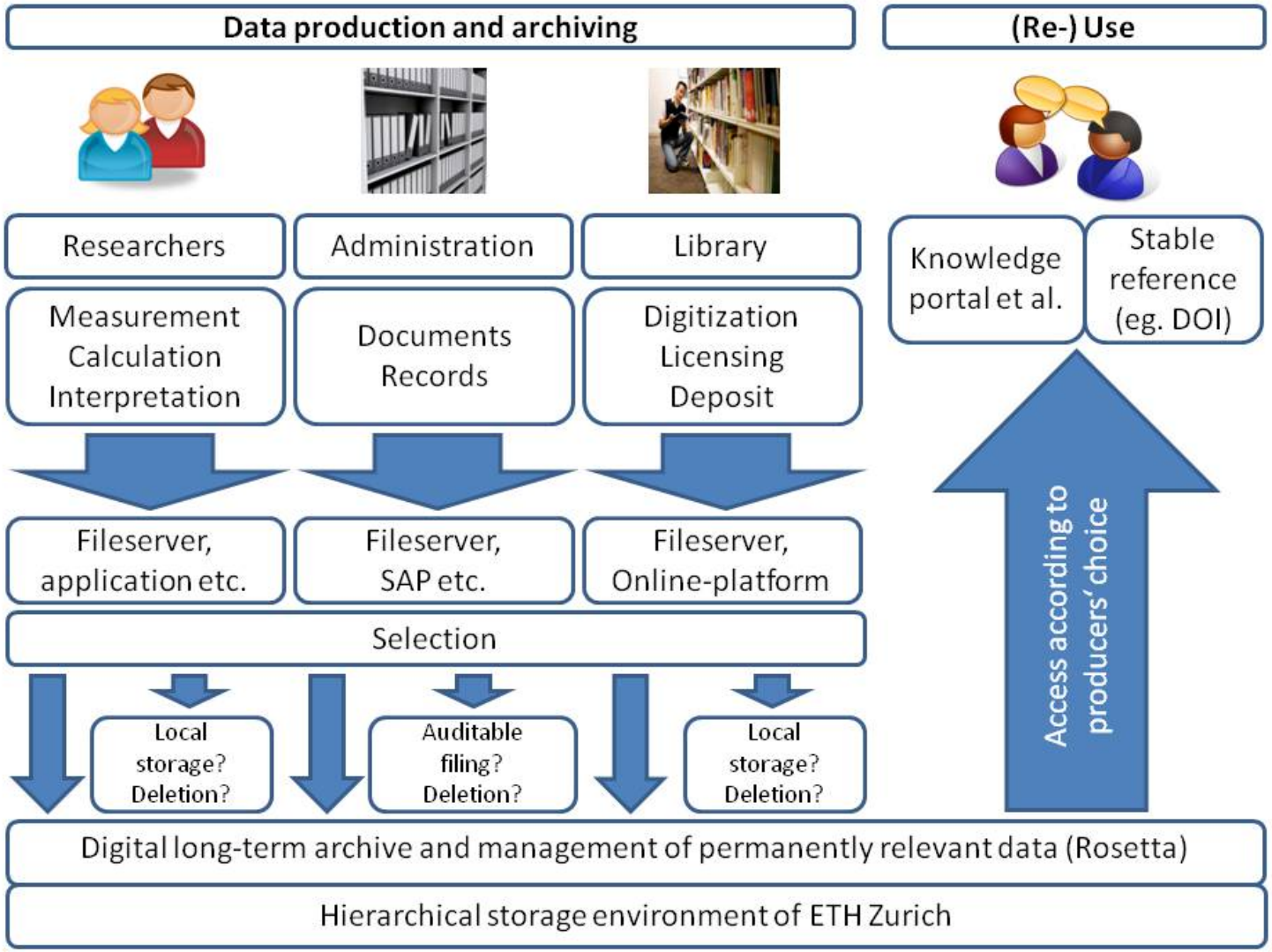
*«Nicht eigenes Verständnis übertragen»*

# DIE FRAGE NACH DEM ARCHIVIERUNGSSYSTEM

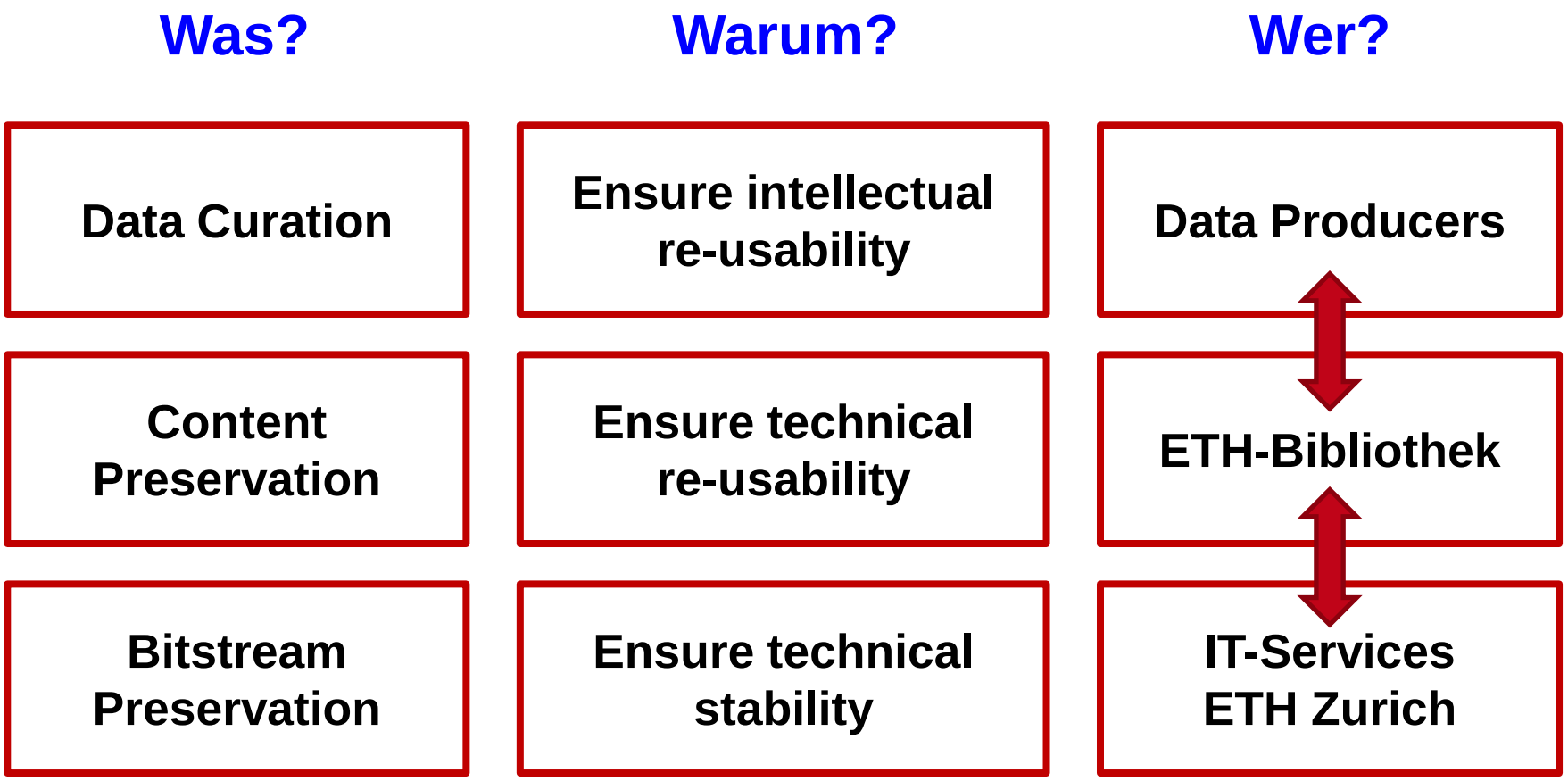
## Warum Rosetta?

- Systemkonzept ist konform mit dem OAIS Reference Model
- ist in Kooperation mit der Bibliotheks- bzw. Archiv-Community entwickelt worden
- unterstützt wesentliche Funktionen bei der Datensicherung
- erfüllt eine Reihe von technischen Voraussetzungen
- bietet die Gelegenheit für eine Entwicklungspartnerschaft
- die Anzahl der Objekte und das Datenvolumen sind skalierbar
- der Betrieb ist auf virtuellen Servern möglich (ETH-Strategie)
- Erfahrung mit verwandten Produkten ist in der Bibliothek vorhanden
- ETH-Strategie setzt grundsätzlich auf herstellerunterstützte Applikationen

# VISION: ROSETTA ALS GEMEINSAME BASIS



# FUNKTIONALE EBENEN




Adaptiert nach Jens Ludwig: Wissgrid

# 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“ | |



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

# RANDBEDINGUNGEN

- Bei Text-Dokumenten ist nachträglich eine gewisse Verbesserung von Metadaten möglich
- Bei Forschungsdaten können technische Metadaten ergänzt werden, aber keine inhaltliche Dokumentation:
  „Garbage in, garbage out“

- *Ein strukturierter Ablauf mit dem Ziel der Langzeitarchivierung muss früh ansetzen: Datenmanagement*
- *Qualität hängt von den Produzentinnen und Produzenten ab, Bibliothek kann diese nur eingeschränkt überprüfen*
- *Langzeitarchivierung hat prinzipielle Grenzen und hängt stark von der heutigen Vorbereitung ab*

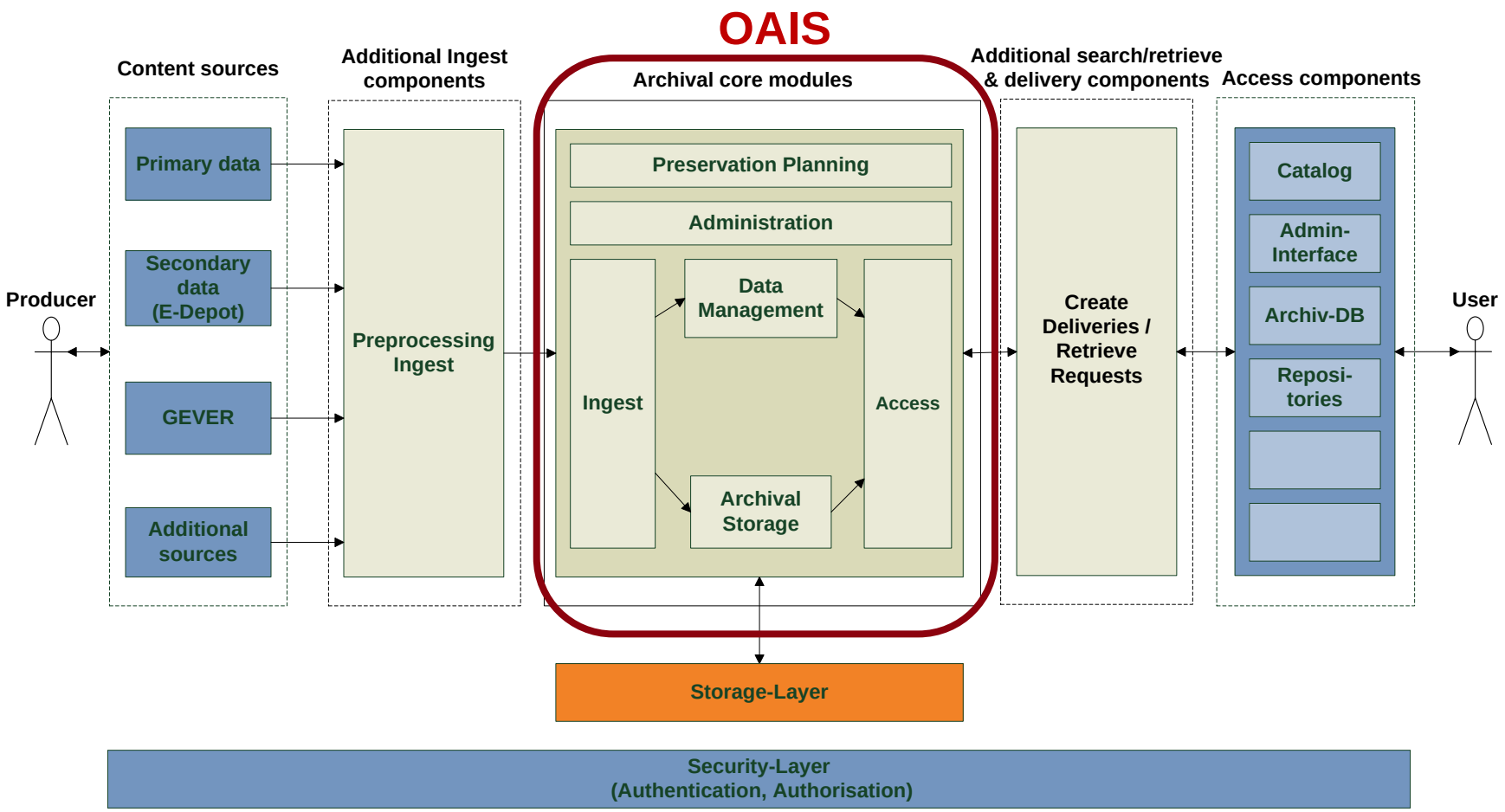
# FORSCHUNGSDATEN UND OAIS?

- Anforderungen betreffen weniger die Funktionen innerhalb des OAIS (ISO-Referenzmodell Open Archival Information System) – aber OAIS-Rahmen ist zu eng:
- 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

# VISION DER ETH-ANWENDUNG




Final report on second phase „Pilot Langzeitarchivierung“, S. 23f; Aliesch, P. et al., 2007: Projekt „Pilot Langzeitarchivierung“. Internal.

# EINE LÖSUNG FÜR ALLE «ARCHIVIERUNGSFÄLLE»?

- für «Big Data» (= sehr grosse Mengen, eher unstrukturierter Daten, innerhalb oder ausserhalb von Applikationen
  - **Bedarf prüfen, Lösungen noch unklar**
- für strukturierte Daten aus bestehenden Anwendungen
  - z.B. LIMS (Laboratory Information Management Systems), Digitale Laborjournale, Datenplattformen wie openBIS (D-BSSE), B-Fabric (FGCZ) o. ä.
  - **grosses Potential für Automatisierung über flexible Schnittstellen**
  - **Pilotprojekt Ende 2013**
- für eher kleinteilig strukturierte Daten
  - **manuelle Vorbereitung in Docupack**
  - **produktiv voraussichtlich ab Herbst 2013**

# MÖGLICHKEITEN DER «DATENABLAGE»

- manuell

  Webdialog zum Hochladen und zur Metadatenerfassung

- halbautomatisch

  Batch-Upload von Dateien mit vorhandenen Metadaten

- automatisch

  - jeweils angepasste Submission Application packt strukturierte Dateien mit vorhandenen Metadaten im XML-Format
  - Submission Application kann auch direkt als Schnittstelle zu bestehenden Quell-Applikationen angelegt werden

# GESAMTPROZESS «SMALL DATA»

**Auswahl und Kontextdokumentation**

**Erhaltung der Nutzbarkeit**

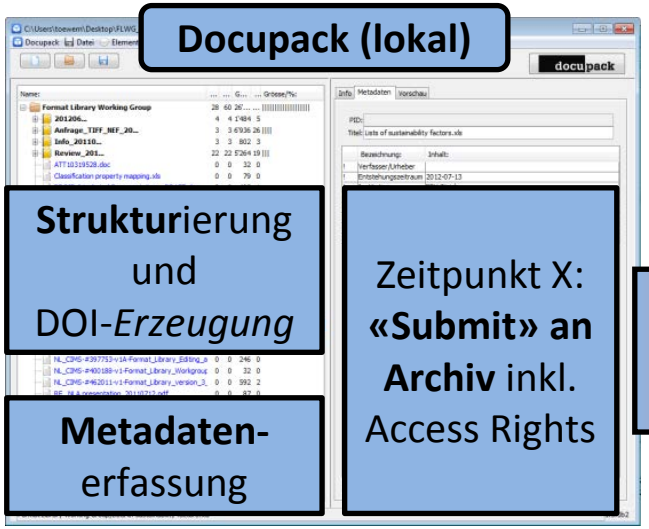
**Bitstromerhaltung**

**Forschende**

**Bibliothek**

**Informatikdienste**

**Docupack (lokal)**



**Strukturierung** und DOI-*Erzeugung*

**Metadaten-**erfassung

Zeitpunkt X: **«Submit» an Archiv** inkl. Access Rights

**«Lokaler» Speicher**

**Über-gabe**

**Documill (zentral)**

**Ingest**

**Digitales Archiv (Rosetta)**

**Server**

**Speicher**

**Netz-werk**

**Zugriff** via DOI



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

# LOKALES DATENMANAGEMENT?

- SIP Package Handler («Docupack» (Java)) + Documill
- Viewer und Editor für Strukturierung und Metadaten-erfassung
- Daten und Metadaten bleiben beliebig lange auf dem lokalen Speicher jeder Forschungsgruppe
- Erzeugt bei Anstoss der Archivierung ein SIP (Submission Information Package) für den Ingest in Rosetta
- Ziel aus Sicht der Langzeitarchivierung:
  Erzeugung von Struktur und von Metadaten für die automatische Verarbeitung (METS)

# BEISPIELE FÜR ANWENDUNGEN

## Forschungsgruppen

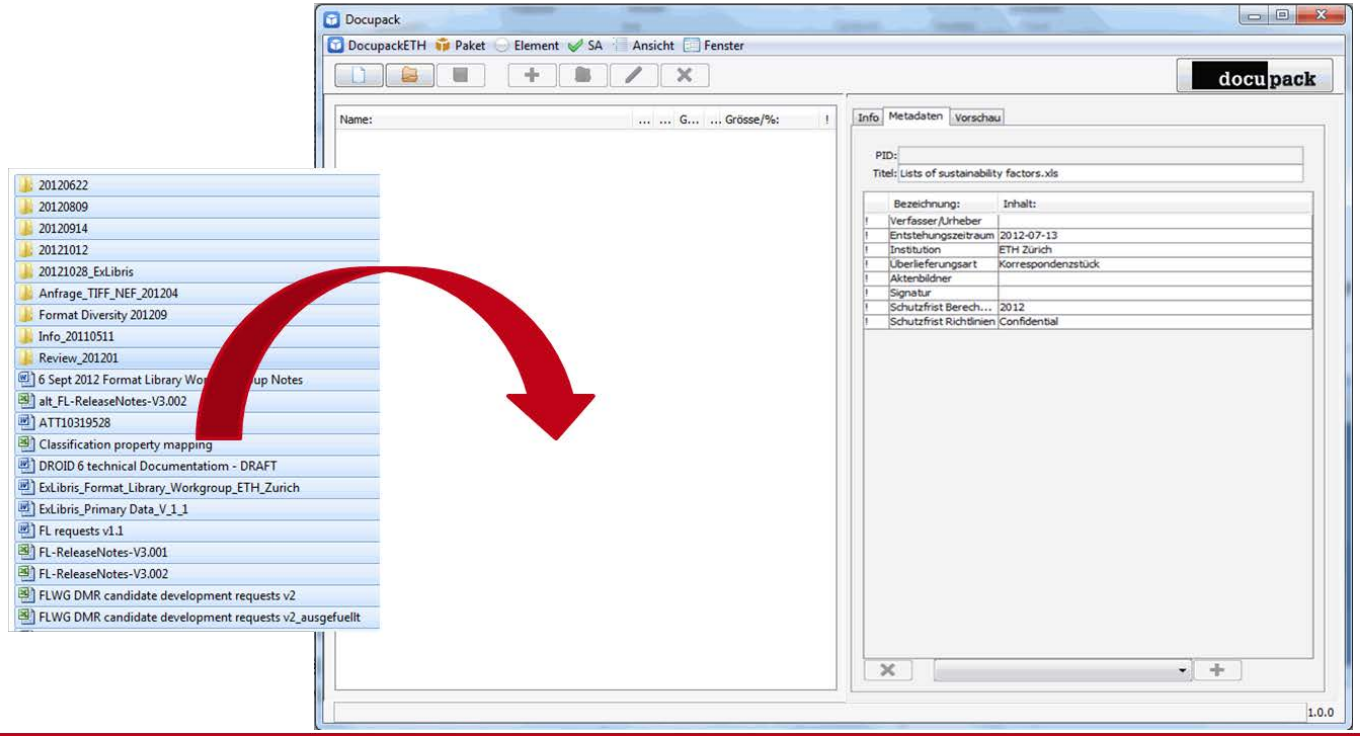
- Daten zu einer Manuskripteinreichung werden gesammelt, archiviert und via DOI zugänglich gemacht
- Forschungsgruppe verfügt über eine strukturierte Ablage ohne Metadaten, die ins Langzeitarchiv überführt werden soll

## Archiv der ETH Zürich

- abliefernde Stelle liefert strukturierte Daten an das ETH Archiv…
- …dessen Personal die Bewertung und Erschliessung vornimmt

# ANWENDUNGSBEISPIELE 1/3

1. Import einer vorhandenen Datensammlung via Drag and drop und anschliessende Erschliessung



# ANWENDUNGSBEISPIELE 2/3

2. Import von Einzelobjekten in einheitliches Struktur-template für Mitglieder einer Forschungsgruppe

FLWGDataModelQuestionnaire_20111201

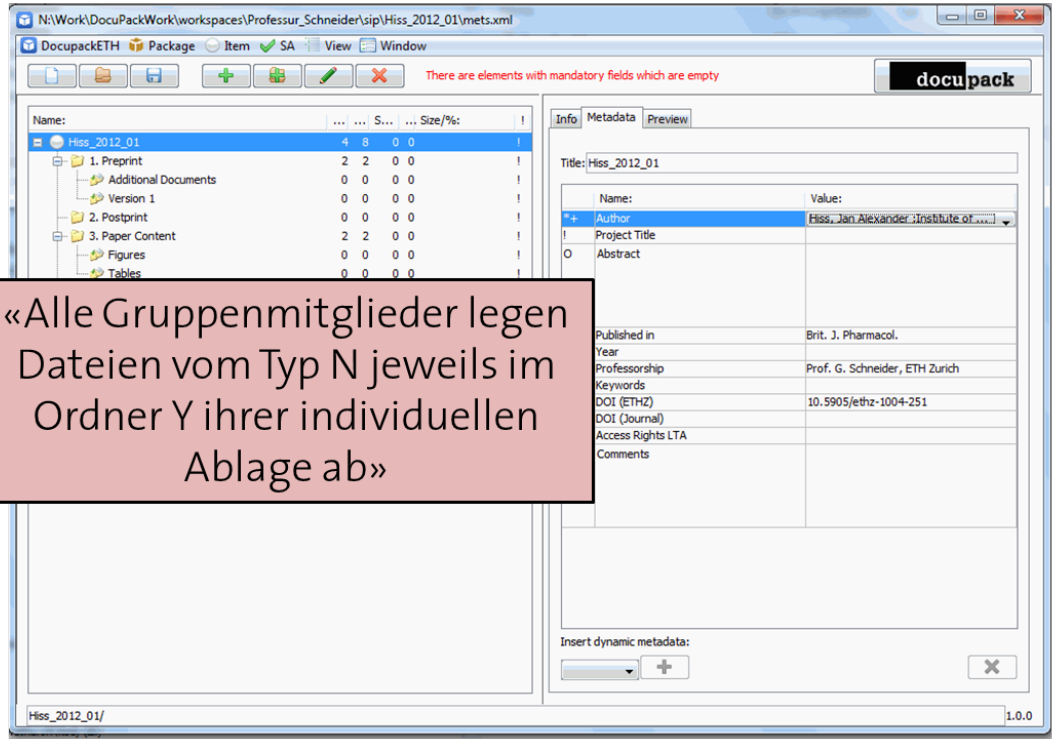
In Ordner X

Lists of sustainability factors

In Ordner Y

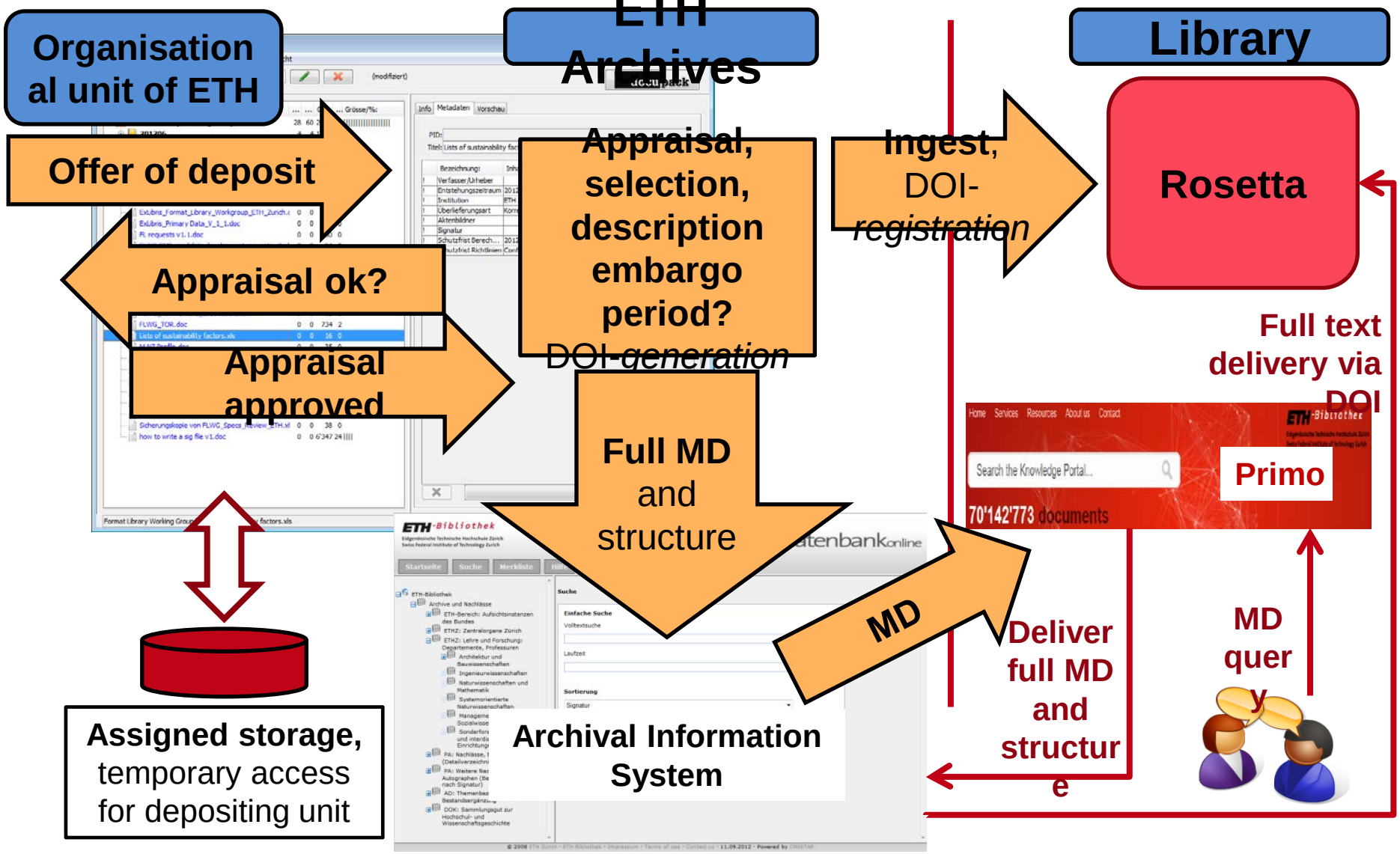
MBSpaperv1_1

In Ordner Z




«Alle Gruppenmitglieder legen Dateien vom Typ N jeweils im Ordner Y ihrer individuellen Ablage ab»

# ANWENDUNGSBEISPIELE 3/3



# WARUM DIESE KOMPLEXITÄT? 1/2

- «Datamanagement light»:
  Daten liegen strukturiert und erschlossen lokal für die Gruppe vor
- Gruppe behält volle Kontrolle, aber wichtige Vorarbeiten sind geleistet, um dauerhafte Erhaltung zu erleichtern
- Metadaten gemäss Vorgaben jeder Gruppe konfigurierbar

# WARUM DIESE KOMPLEXITÄT? 2/2

- Struktur- und Metadaten als METS-XML können automatisch an Rosetta übergeben werden
  (Documill als Submission Application)
- DOIs sind erzeugt, d.h. zur Zitierung verfügbar
  (Registrierung in Rosetta)
- Selektion der Daten für…
  - …kurzfristige lokale Aufbewahrung
  - …befristete zentrale Speicherung
    (Übergabe einer Retention Period an Rosetta)
  - …dauerhafte zentrale Erhaltung ohne oder mit öffentlicher Nutzung

# WIE EROLGT DER ZUGRIFF?

- gegenwärtig sind die Metadaten von Forschungsdaten nicht publiziert (ausser für die DOI-Registrierung)
  - *wenn Daten öffentlich sind, erfolgt Zugriff via DOI*
- Inhalte des ETH Archivs publiziert in Primo aus dem Archiv-informationssystem (CMI Star)
  - Primo verlinkt auf CMI Star für volle Metadaten und Struktur
  - CMI Star verlinkt via DOI auf Objekt in Rosetta
- Bibliotheksinhalte: Rosetta enthält das «dark archive»
  - Primo verlinkt in öffentliche Online-Anwendungen

# BISHERIGE UND AKTUELLE SCHRITTE 1/2

- ✓ DOI-Registrierung durch ETH Zürich als Mitglied von DataCite
- ✓ Umfrage bei allen Forschungsgruppen der ETH Zürich
- ✓ Identifizierung von Pilotpartnern, Ermittlung deren Anforderungen
- ✓ Prüfung und Aktualisierung des Inventars von Daten der Bibliothek
- ✓ Submission Application für die ETH E-Collection
- ✓ «Gap analysis» und Entscheidung über zukünftige Entwicklungen
- ✓ Entwicklung und Testen der Erweiterungen für Rosetta

# BISHERIGE UND AKTUELLE SCHRITTE 2/2

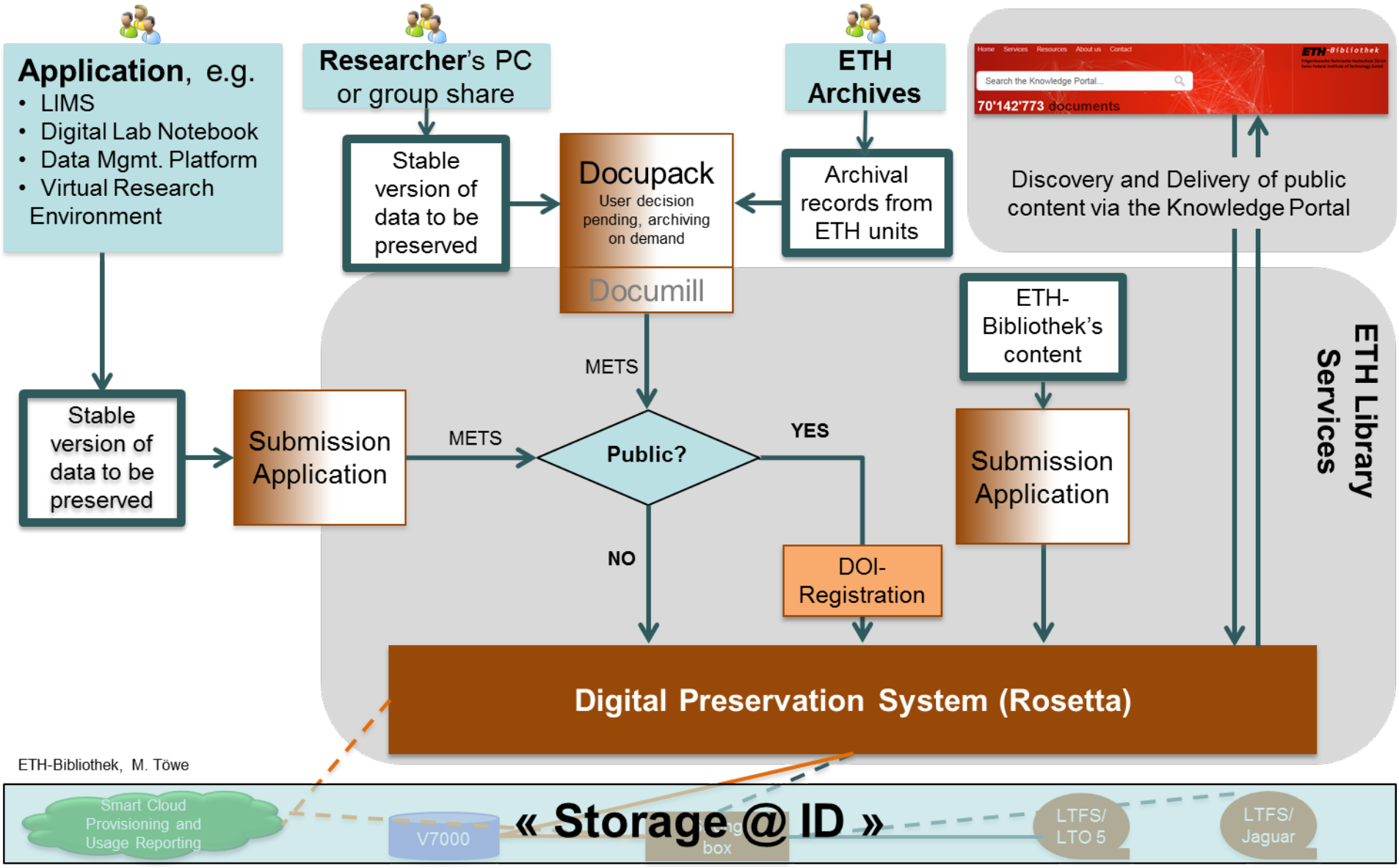
- Entwicklung und Testen des lokalen Datenmanagements

  → laufend

sofern erfolgreich:

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

# Services der ETH-Bibliothek



# STRATEGISCHE FRAGEN 1/2

- Wie soll das Profil des Services «Datenerhalt» aussehen?
  - Nur Erhaltung oder zukünftig verstärkt auch Datenmanagement?
  - Potential für eine E-Collection *plus* mit «*supplementary materials*»*?*
  - Positionierung mit Blick auf virtuelle Forschungsumgebungen?
  - Beratung als zentrale Aufgabe bereits bei Antragstellung für Forschungsprojekte?
  - Gewichtung zwischen Service und Softwarebereitstellung

# STRATEGISCHE FRAGEN 2/2

- Wie geeignetes Personal in ausreichendem Umfang aufbauen?
  (forschungsnah/bibliothekarisch und archivarisch/IT-affin…)
- Kosten der voraussichtlich aufwändigen Dienstleistung müssen gedeckt werden - gleichzeitig soll das Angebot keine Kostenbarriere darstellen
- Kultur der Zusammenarbeit unterschiedlichster Stakeholder
  (Forschende/Informatik Support der Departemente/Informatikdienste/ Bibliothek)


END
ROAD WORK
THANK YOU