ETH BIBLIOTHEK

Forschungsdaten zwischen «dunkler» Archivierung und Open Data

104. Deutscher Bibliothekartag

Nürnberg, 26. Juni 2015

Dr. Matthias Töwe, ETH-Bibliothek, ETH Zurich

Gliederung

- Ausgangslage
- Verhältnis von Offenheit und langfristiger Erhaltung
- Anwendungsfälle
- Schlussfolgerungen für Dienstleistungen

Ausgangslage

- **Forderungen**
  (Förderer, Politik, Hochschulen, Herausgeber, z.T. Forschende)
  - Möglichst **freier Zugang** zu Forschungsdaten
  - Möglichst **unbeschränkte Nutzbarkeit** von Forschungsdaten
  - Mittelfristige **Aufbewahrung für die** unmittelbare **Überprüfbarkeit**
  - **Langfristige Aufbewahrung von nicht wieder zu gewinnenden Daten**
  - **Erhaltung publizierter Daten** analog zu anderen Veröffentlichungen
- **Ausrichtung Dienstleistung ETH Data Archive**
  - **Langzeitarchivierung und Zugänglichkeit von statischen Daten**
    - ***Also alles ok?***

Offenheit *und* Langzeitarchivierung

- **Daten erhalten,** aber für Dritte **nicht zugänglich**
  - **nur für die Produzenten** von Nutzen
- Daten erhalten und **zugänglich,** aber ohne **Kontextdokumentation**
  - **Nutzung u.U. selbst für die ursprünglichen Produzenten** schwierig
- Daten werden **zugänglich gemacht, aber nicht erhalten**
  - **Nicht zitierbar, nicht verlässlich nachnutzbar**

- **Zugänglichkeit und dauerhafte Erhaltung sind nicht zu trennen**

Herausforderungen

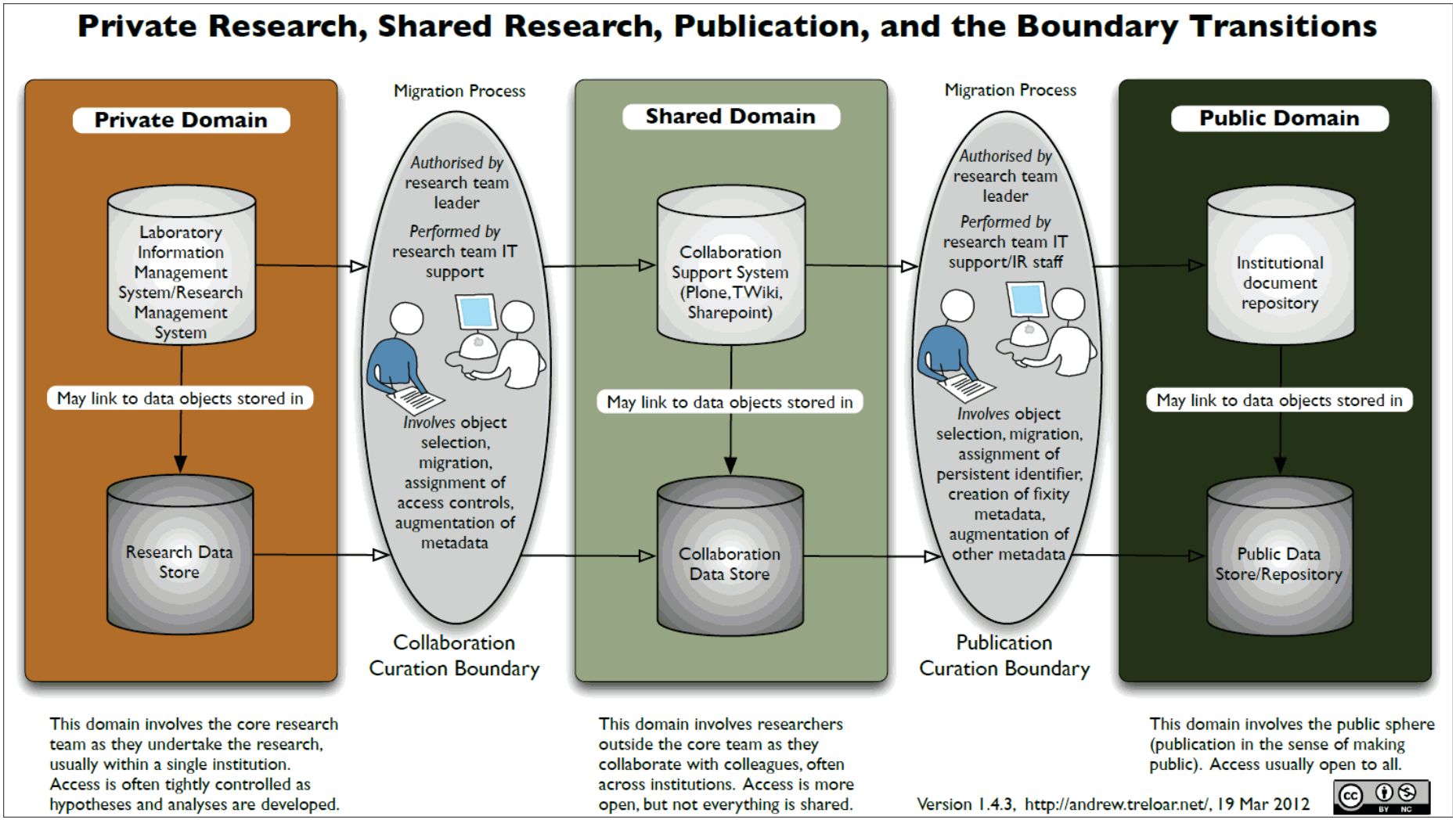
- **Etablierte Standards** für das Vorgehen **nur in wenigen Fächern**
- **«Vertrauenskultur» sehr unterschiedlich ausgeprägt**
- **Kundenerwartungen sowie technische und organisatorische Anforderungen für «Open Data» und Langzeitarchivierung stimmen nicht überein…**
- **…obwohl sich beide Aspekte nicht trennen lassen**
- Bei Dienstleistungen ist bisher **viel Einzelfallbehandlung nötig**
  - **Sowohl unseren Kunden als auch uns fehlt Erfahrung**

LZA vs. Zugänglichkeit vs. Openness

	Interne Ablage «Schliess-fach»	Kurz-fristiger Austausch / «Sharing»	Dauerhaft, aber gesperrt	Dauerhaft zugäng-lich	Dauerhaft nutzbar als «Open Data»
LZA	-	-	+	+	+
Daten-publikation	-	(+)	-	+	+
Open Data	-	-	-	-	+

- **Kunden erwarten *Lösungen*** – am liebsten ***EINE* Lösung**

Domänenmodell



Anwendungsfall: Interne Ablage – ‘Schliessfach’

	Interne Ablage «Schliess-fach»
LZA	-
Daten-publikation	-
Open Data	-

- Charakteristik
  - Aufbewahrungsfrist ca. **10 Jahre**
  - **Archivierung «für den Fall der Fälle»**
  - Zugriff für **ausgewählte Personen** einer Forschungsgruppe
  - **Seltene Nutzung**
  - **Metadaten unsichtbar** für Dritte

- Kann ETH Data Archive u.U. trotzdem ein passender Ort sein?
  - Daten sind statisch, werden nicht verändert und selten genutzt
  - Vereinbarter **Verzicht auf Aufwand für Formatanalyse und** alle weiteren **Erhaltungsmassnahmen**
  - Nachteil für Kunden: **Keine Selbstverwaltung**

Anwendungsfall: Interne Ablage - Beispiel

- **Ausgangslage**
  - Sammlung von **Dissertationen mit dazugehörigen Daten auf CD-ROM**
  - **Erfahrung:** In den letzten zehn Jahren gab es eine Nachfrage
  - **Metadaten auf der Ebene der ganzen Doktorarbeit**
  - **Tausende bis zehntausende von Dateien** pro CD-ROM
  - Zum grossen Teil **spezifische und selbst entwickelte Formate**
  - **Web-Zugriff** für die Verantwortlichen **gewünscht**
  - Nur der **Professor kann** mit einiger Sicherheit in den Daten **navigieren**
  - **Nachträgliche Anreicherung mit Metainformation oder Selektion der Daten nicht realistisch**

Anwendungsfall: Interne Ablage - Ergebnis

- Alternative Lösung direkt auf kostengünstigem Speicher:
  Keine Metadaten, kein Webzugriff, Objekte zu klein → Nicht gewünscht
- Umsetzung im **Langzeitarchiv**:
  - **Formatidentifizierung und -validierung mehrheitlich nicht möglich**
  - **Beschränkung auf Erhaltung «wie geliefert»**
  - **Filezahlen und -strukturen nur handhabbar in Containern (ZIP/TAR)**
  - **Definierte Retention Period**
  - **Zugriff für Professor und letzten Mitarbeiter**
  - **Wenn vorhanden: Prüfsummenvergleich möglich**

Anwendungsfall: Interne Ablage - Ergebnis

- Alternative Lösung direkt auf kostengünstigem Speicher:
  Keine Metadaten, kein Webzugriff, Objekte zu klein → Nicht gewünscht
- Umsetzung im **Langzeitarchiv** (!):
  - **Formatidentifizierung und -validierung mehrheitlich nicht möglich**
  - **Beschränkung auf Erhaltung «wie geliefert»**
  - **Filezahlen nur handhabbar in Containern (ZIP)**
  - **Definierte Retention Period**
  - **Zugriff für Professor und letzten Mitarbeiter**
  - **Wenn vorhanden: Prüfsummenvergleich möglich**
- **Als ad hoc-Lösung 'OK' – aber können wir damit zufrieden sein?**

Anwendungsfall: Interne Ablage - Perspektive

- **Langfristig**
  - **Auf bewussteres Datenmanagement in den Gruppen hinarbeiten!**
  - **Frühzeitige Planung in der Gruppe**
    (unabhängig davon, ob ein Plan verlangt wird)
  - Eigene **Kriterien definieren und anwenden**
    - **Was soll und kann publiziert und archiviert werden?**
    - **Welche technische und inhaltliche Dokumentation ist dafür nötig?**
    - **Was soll zusätzlich befristet aufbewahrt werden?**
  - Zusätzliches **«Digitales Zwischenarchiv»** einrichten?

Fragen an die Datenproduzenten

- **Wie viel Vertrauen haben Forschende (und wir) in ihre Peers?**
  - **Wie viel und was wird frei zugänglich gemacht,** was nur engen Partnern?
  - **Als Datenproduzenten** (heute):
    Umfang und Nutzen der Kontextdokumentation; vorausschauende Entscheidung für oder gegen Veröffentlichung
  - **Als Datennutzer** (in der Zukunft):
    Bereitschaft, Aufwand zur Nachnutzung «alter» Daten ungewisser Qualität zu treiben; Verfügbarkeit der Kompetenzen, es auch zu tun
- **Was können Förderer oder Hochschulen «per Dekret» und ohne Nebenwirkungen für die Forschenden beeinflussen?**

Anwendungsfall: Kurzfristiger Austausch / «Sharing»

	Kurz-fristiger Austausch / «Sharing»
LZA	-
Daten-publikation	(+)
Open Data	-

- Charakteristik
  - Aufbewahrungsfrist < **10 Jahre**
  - Zugriff für **ausgewählte Personen** ausserhalb der Institution
  - **Metadaten unsichtbar** für Dritte
  - **Flexibilität** gefragt

- Langzeitarchiv ist in der Regel keine gute Lösung
  - **Daten werden** allenfalls (gemeinsam) **ergänzt, bearbeitet, aktualisiert**
  - **Kunden erwarten Möglichkeit zur Selbstverwaltung des Zugriffs**

Anwendungsfall: Kurzfristiger Austausch / «Sharing»

- **Ausgangslage**
  - **«Lebende» Daten** aus einem laufenden Projekt zur weiteren Ergänzung und Bearbeitung
  - **Nur ausgewählte Partner aus verschiedenen Institutionen** dürfen Zugriff haben
  - **Nur im Projektkontext interpretierbare Metadaten**
  - **Implizite Information in Ablagestruktur**
  - Nutzung in dieser Form **für maximal 2 Jahre**
  - **Web-Zugriff gewünscht**

Anwendungsfall: Kurzfristiger Austausch / «Sharing»

- **Ergebnis**
  - **Langzeitarchiv ist kein geeigneter Ablageort**
    - Gemäss seiner **Zielsetzung**
    - **Aus technischer Sicht**
  - **Alternative:**
    Erweiterung der für das gruppeninterne Datenmanagement genutzten **fachspezifischen Plattform (openBIS)** durch die Informatikdienste
  - Falls noch kein spezifisches Tool genutzt wird:
    - **Microsoft Sharepoint für einfache Dateien?**
    - **ETH polybox** (Own Cloud) **als niederschwelliger Weg zum Austausch**
  - ➢ ***Weit weg von Langzeitarchivierung!***

Anwendungsfall: Dauerhaft zu erhaltende, gesperrte Daten

	Dauerhaft, aber gesperrt
LZA	+
Daten-publikation	-
Open Data	-

- Charakteristik
  - Bei Forschungsdaten denkbar für **nicht anonymisierbare Daten**
  - **Typisch für Verwaltungs- und Nachlass-archive** (gesetzliche Sperrfristen)
  - **Metadaten zunächst unsichtbar** für Dritte
  - **Gesetzliche Vorgaben** zur Aufbewahrung

➢ **Volle Funktionalität des Langzeitarchivs** (OAIS) ist gefordert

➢ **Besondere Herausforderung: Management des Zugriffs auf Metadaten**

Anwendungsfall: Dauerhaft erhaltene, frei zugängliche Daten

	Dauerhaft zugänglich
LZA	+
Daten-publikation	+
Open Data	-

- Charakteristik
  - **Dauerhafte Zugänglichkeit und Nutzbarkeit**
  - **Metadaten werden frei verbreitet**
  - **Inhalte frei zugänglich…**
  - **…aber je nach System oder Lizenz nicht unmittelbar als Open Data verfügbar**

- **Volle Funktionalität des Langzeitarchivs** (OAIS) ist gefordert
  - **Umfassende Kontextinformation entscheidet über Nutzbarkeit**
  - **Allenfalls Verbindung zu anderen Systemen als Voraussetzung für Open Data-Nutzung?**

Anwendungsfall: Dauerhaft zu erhaltende, frei zugängliche Daten

- **Ausgangslage**
  - **Metadaten und Objekte sind frei zugänglich**
  - **Ursprünglicher Hauptanwendungsfall des ETH Data Archive!**
  - **Frei zugänglich – und doch nicht «open» gemäss Definition:**
    - *"Open data and content can be **freely used, modified, and shared** by **anyone** for **any purpose**"*
      (Kurzfassung der Open Definition, http://opendefinition.org/, Zugriff 22.05.2015)
    - **Eingeschränkt durch Lizenzbedingungen, z.B. Creative Commons NonCommercial**
  - **Praktische Beschränkung der Nutzbarkeit** durch Ausrichtung auf Einzeldateien: **Keine unmittelbar Nachnutzung einer Sammlung im Langzeitarchiv,** sondern erst nach Export der Dateien

Anwendungsfall: Dauerhaft erhalten und als Open Data nutzbar

	Dauerhaft nutzbar als «Open Data»
LZA	+
Daten-publikation	+
Open Data	+

- Charakteristik
  - **Breite Nachnutzungsmöglichkeit wird unterstützt** (technisch und durch entsprechende Lizenz)
  - Damit **implizite Notwendigkeit für dauerhafte Zugänglichkeit und Nutzbarkeit** als Datenquelle
  - **Metadaten werden frei verbreitet**

- **Funktionalität eines reinen Langzeitarchivs** reicht u.U. nicht
  - **Umfassende Kontextinformation entscheidet über Nutzbarkeit**
  - **LZA erhält zwar Beziehungen zwischen Objekten…**
  - **…für intensive Nutzung meist Export nötig**

Folgerungen für bedarfsgerechte Dienstleistungen

Voraussetzungen und Massnahmen:

- Ausreichendes **Verständnis der Anforderungen *im Einzelfall***
  - **Beratungsangebot zum Datenmanagement früh im Lebenszyklus**
  - **Frühzeitige Einblicke** sind hilfreich, **Kontakte vor Ort** stärken
- Eigene **Service-Policy** schärfen, einschliesslich ihrer **Grenzen**
  - Was können wir **mit vertretbarem Aufwand** für Kunden und uns anbieten?
- **Rechtslage** weiter **klären, z.B.:**
  Welche Lizenzen dürfen ETH-Angehörige vergeben?
  - ETH-Gesetz behält der ETH die meisten Immaterialgüterrechte vor

Folgerungen für bedarfsgerechte Dienstleistungen

Mittel- und langfristige Massnahmen:

- **Integration mit Anwendungen und Plattformen** für komplementäre Aufgaben **> *One-Stop-Shop***
  - **Elektronische Laborbücher (ELN) / Labor-Informations- und Management-Systeme (LIMS)**
  - **Elektronisches Publizieren**
  - **Open Data** Anwendungen
  - **Metadatenmanagement**
- **Redundante Ablage in mehreren Systemen** statt wachsender Komplexität durch verschiedenste Funktionen in der gleichen Anwendung?
- **Zertifizierung pro Workflow** statt für das Langzeitarchiv insgesamt?

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