ETH BIBLIOTHEK

Formatvalidierung bei Forschungsdaten: Wann und wozu?

106. Deutscher Bibliothekartag
Frankfurt am Main, 1. Juni 2017

Dr. Matthias Töwe, ETH-Bibliothek, ETH Zürich

Überblick

- Formatidentifizierung und -validierung – ideal
- Workflows in der Praxis
  - Formatinformation: zu wenig, zu spät und dann zu viel?
  - Interesse bei Datenproduzentinnen?
- Mögliche Ansatzpunkte und Grenzen

Aufgabe Langzeitarchivierung

- **Fachstelle «Digitaler Datenerhalt»** der ETH-Bibliothek
- **ETH Data Archive** als Infrastruktur für die **Langzeitarchivierung des wissenschaftlichen und historischen Erbes** der ETH Zürich mit
  - **Forschungsdaten («Long tail»)**
  - **Archivunterlagen**
  - **Bibliotheksinhalten**

➢ ***Offensichtlich unterschiedliche Voraussetzungen***

Formatidentifizierung und -validierung

- **«Know your data!»**
- Möglichst vollständige **Charakterisierung von Dateien beim Ingest** in das Langzeitarchiv
- Gewonnene **technische Metadaten** als Basis für **Risikoabschätzung** im Hinblick auf langfristige Nutzbarkeit
- **Dokumentation von** inneren und äusseren **Abhängigkeiten** (z.B. von spezieller Software)
- Wichtige Information zur **Planung von Erhaltungsmassnahmen wie Formatmigration oder Emulation…**
- **…sowie insgesamt zur späteren Nutzung**

Beispiel: Masterfiles aus der Digitalisierung

- **Sehr grosse, einheitliche Serien**
- **Verarbeitung automatisierbar**
- **Gebräuchliche und gut dokumentierte Formate**
- **Werkzeuge zur Charakterisierung verfügbar –** und wenn nicht, so lohnt sich der Aufwand, weitere zu integrieren
- **Ganzer Workflow im Haus**, d.h. Probleme können gemeldet und behoben werden

➢ ***Formatidentifizierung und –validierung sind vollständig möglich***

Forschungsdaten – «Long tail»

- **Individuell aufbereitete Pakete mit heterogenem Inhalt:**
  «Be prepared for anything»
- **Technische Grenzen**:
  - Wichtige binäre Formate bisher nicht identifiziert (z.B. NetCDF4)
  - Textfiles bieten wenig Möglichkeiten für differenzierte Charakterisierung
  - Ohne Identifizierung auch keine Validierung
- **Interessenlage**:
  Langzeitarchivierung nicht im Fokus, vor allem dort nicht, wo Nachnutzung bisher die Ausnahme ist. Verantwortung wird bei späteren Usern gesehen.

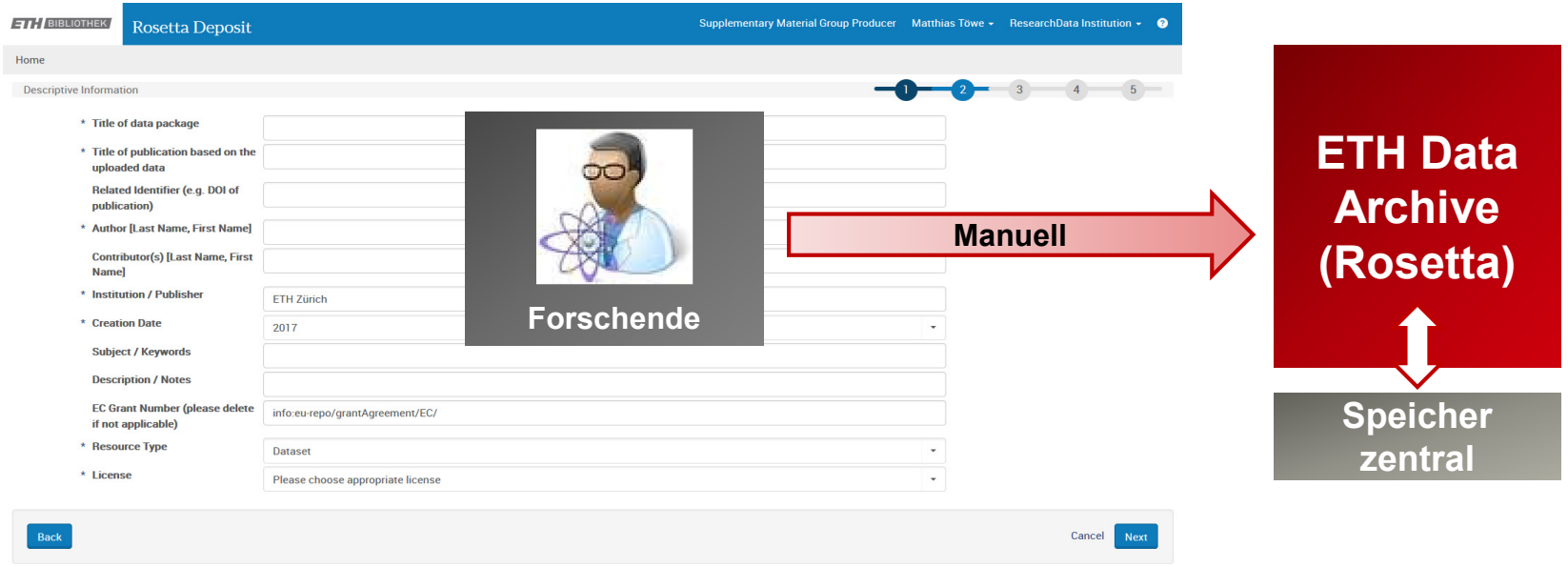
➢ ***Formatidentifizierung und –validierung sind stark limitiert***

Workflows für Forschungsdaten im «Long tail»

- **Typisch**
  - Daten werden am **Ende eines Projekts** abgeliefert, oft unter **Zeitdruck**
  - Doktorierende liefern u.U. **nur einmal** etwas ab…
  - …daher auf beiden Seiten **keine** Entwicklung einer **Routine** möglich
- **Motivation für weiteren Aufwand ist gering, Anreize gibt es nicht**
- **Verwendung von offenen und dokumentierten Formaten? Hängt ab von**
  - **Gepflogenheiten** im Fach
  - **Bewusster Aufbereitung für die Publikation** und Nachnutzung
  - **Verfügbarkeit und Praxistauglichkeit** entsprechender Formate

Workflow 1: Web-Upload ins ETH Data Archive

- **Bisher:**
  Direkte Interaktion von Kunden mit dem ETH Data Archive und seinem Personal



Workflow 1: Web-Upload ins ETH Data Archive

- **Konkrete Rückmeldung zu Formaten** wäre möglich…
- **…ist aber** quantitativ im Detail **nicht leistbar…**
- **…und Forschende haben nur geringen Spielraum,** zu diesem Zeitpunkt etwas zu ändern
  - **Inhaltlich, weil es bereits Abhängigkeiten in den Daten gibt**
  - **Zeitlich, weil meistens ein Druck besteht für Veröffentlichung und Zitierung**
- **Workflows des Langzeitarchivs sind** korrekterweise eher **starr**
- Es werden **Container** gebildet und **enthaltene Files so weit wie möglich identifizier**t. Es erfolgt **keine** weitere **Nachbearbeitung**.

Workflow 2: Repository als Frontend

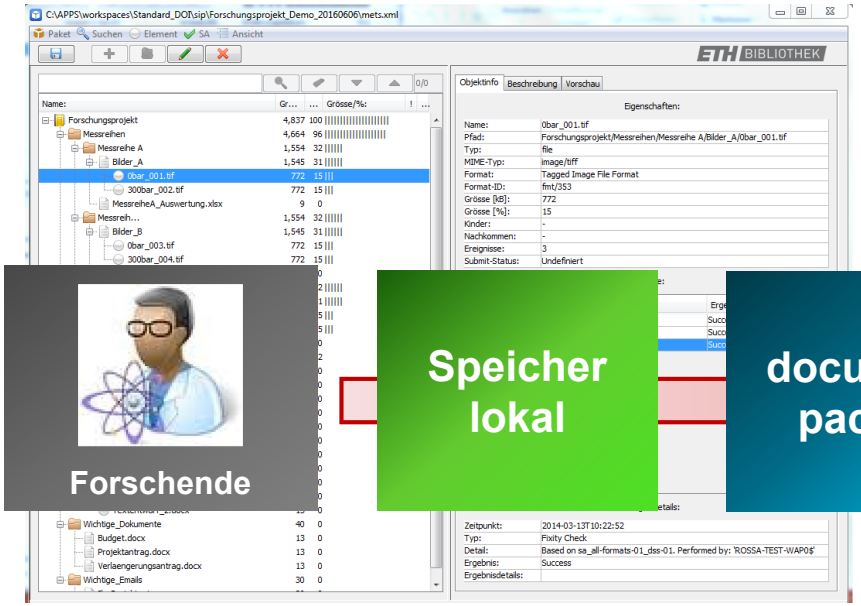
- **Neu ab Mitte 2017:**
  Web-Upload von Forschungsdaten **in institutionelles Repositorium «Research Collection»**
- **Anschliessend Ingest** aus dem Repositorium **ins Langzeitarchiv**

Workflow 2: Repository als Frontend

- **Neu ab Mitte 2017:**
  Web-Upload von Forschungsdaten **in institutionelles Repositorium**
- **Anschliessend Ingest** aus dem Repositorium **ins Langzeitarchiv**
- **Vorteil:**
  Komfort und Flexibilität für Kunden und hohe Automatisierbarkeit der Datenübergabe
- **Nachteil:**
  Rückmeldung zu Formaten und ihres Status hinsichtlich der Langzeitarchivierung nur auf Basis der **Dateiendung.**
  (Ausnahme: Sicht- und Funktionsprüfung von Dissertationen in PDF)

Workflow 3: Lokale Erschliessung

- **Workflow mit docuteam packer:**
  - **Lokale Erzeugung von Submission Information Packages (SIP)**
  - **Übergabe ans ETH Data Archive**



Speicher lokal

docuteam packer

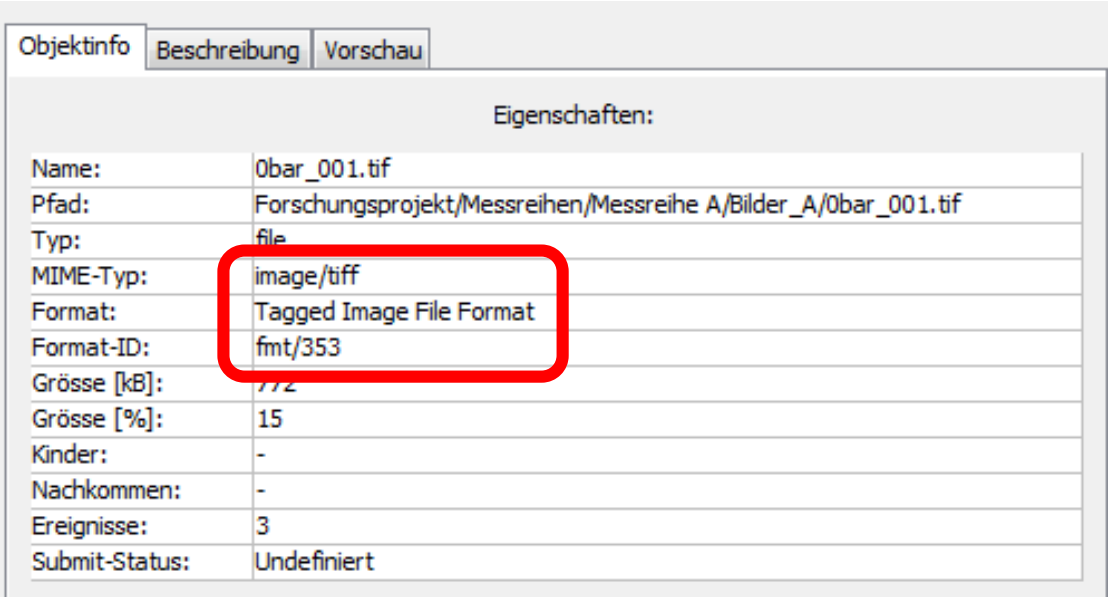
docuteam feeder

ETH Data Archive (Rosetta)

Speicher zentral

Workflow 3: Lokale Formatinformation

- **Workflow mit docuteam packer:**
  Formatinformation ist bereits lokal vorhanden ☺
- **Wird aber praktisch nicht genutzt ☹**



Interessieren die Formate niemanden?

- **Doch!** Aber zum Zeitpunkt der Archivierung ist es **zu spät**.
- **Informationen zur Archivtauglichkeit werden gewünscht und genutzt** und konkrete Fragen vorgelegt…
- **…wenn Forschende den Anspruch haben** oder aufgefordert sind**, die Langzeitverfügbarkeit sicherzustellen**
- **…wenn die Veröffentlichung von Daten früh geplant wird**
- **…in der Fachcommunity de facto oder formale Standards gelten**

➢ ***Typische Fragen zur Diskussion in einem Datenmanagementplan!***

Mögliche Ansatzpunkte

- **Beratung zu Datenmanagementplänen nutzen,** um Möglichkeiten und Grenzen mit Forschenden auszuloten…
- …und **Hintergründe** besser **zu verstehen**
- **Rückmeldung geben, auch wenn es im konkreten Fall zu spät ist:** Input für Diskussion der Forschungsgruppen und ihrer Fachcommunity
- **Ziel:**
  Mit der Zeit früher im Lebenszyklus Verbesserungen erreichen

➢ ***Langfristig denken und mit den Kundinnen lernen!***

Grenzen des Optimismus

- Ein **Zwang zur Verwendung** bestimmter Formate würde **die Forschungsmethodik beschränken**
- **Zusatzaufwand für Forschende** muss soweit wie möglich vermieden werden
- Massnahmen mit dem **Hauptargument der Langzeitarchivierung** wird es nur in eng begrenzten Gebieten geben

Oder doch lieber die Finger davon lassen?

- **Forschende:**
  «Spätere Nutzerinnen müssen Daten selbst aufbereiten».
  - Teilweise **Umkehrung unserer Idee von Langzeitarchivierung?**

- **Anspruch Formatcharakterisierung aufgeben? Nein!**
  - Wo sie gelingt, **ist diese Information wichtig für spätere Nutzer,** erst recht, wenn das Langzeitarchiv ein Format nicht selbst unterstützen kann.
  - Es geht weiterhin um die **Formatcharakterisierung für die spätere Nutzung**, allenfalls sind die **Adressaten aber andere** als wir!

Vielen Dank! - Gibt es Fragen?

http://www.library.ethz.ch/Digitaler-Datenerhalt

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

matthias.toewe@library.ethz.ch