



Rosetta @ ETH Data Archive

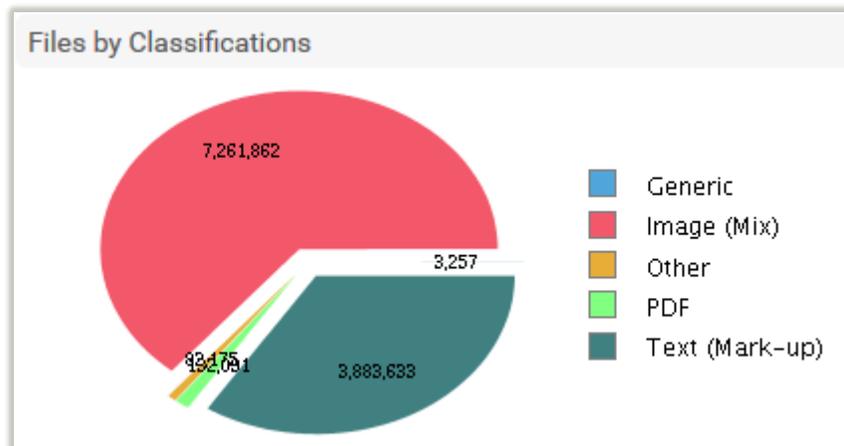
Memoriav Fachtagung 2019

Biel, 19. März 2019

Dr. Franziska Geisser, ETH-Bibliothek, Forschungsdatenmanagement und Datenerhalt

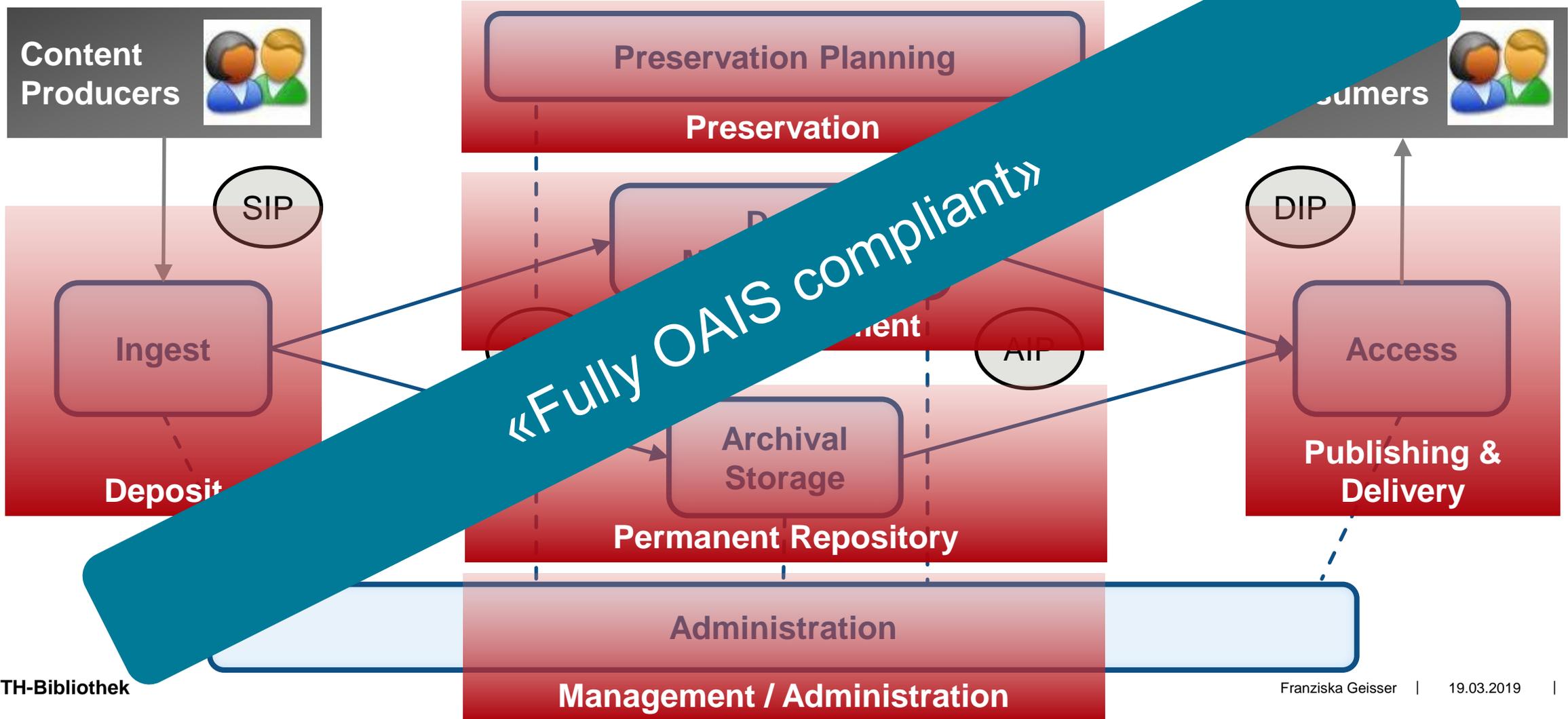
Rosetta an der ETH Zürich

- «ETH Data Archive» als Dienstleistung der ETH-Bibliothek für die ETH Zürich
- Produktiv seit 2012
- Stand März 2019:
 - Insgesamt 160 TB Daten
 - 11 Mio. Dateien
 - 562'000 archivierte Objekte (AIPs)

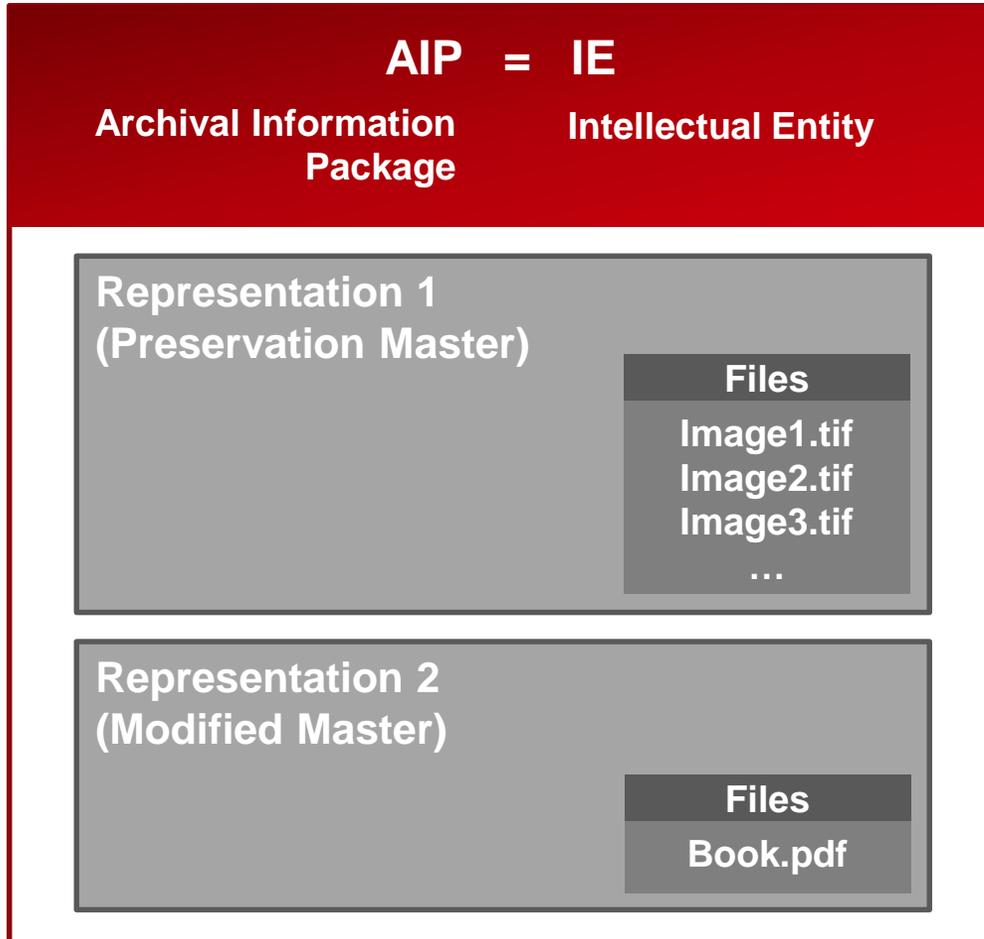


	Format	Anzahl Dateien
1	TIFF	7'084'815
2	XML	3'883'112
3	PDF	130'328
4	WARC	826
5	Log File	744
6	Matroska FFV1 (.mkv)	544
7	ZIP	525
8	Unknown	492
9	Plain Text (.txt et al.)	484
10	MPEG-2 (.mpg, .mpeg)	448

Rosetta und OAIS



Rosetta und PREMIS



«DNX»-Metadaten im METS.xml:

```

- <mets:digiprovMD ID="FL998806-amd-digiprov">
  - <mets:mdWrap MDTYPE="OTHER" OTHERMDTYPE="dnx">
    - <mets:xmlData>
      - <dnx>
        - <section id="event">
          - <record>
            <key id="eventDateTime">2016-04-08 16:14:27</key>
            <key id="eventType">VALIDATION</key>
            <key id="eventIdentifierType">DPS</key>
            <key id="eventIdentifierValue">27</key>
            <key id="eventOutcome1">SUCCESS</key>
          - <key id="eventOutcomeDetail1">
              IE_PID=IE997956;COPY_ID=null;ALGORITHM_NAME=MD5;DEPOSIT
              16:14:27;STATUS=SUCCESS;REP_PID=REP997957;TASK_ID=1;PROCE
            </key>
            <key id="eventDescription">Fixity check performed on file</key>
            <key id="linkingAgentIdentifierType1">SOFTWARE</key>
            <key id="linkingAgentIdentifierValue1">REG_SA_JAVA5_FIXITY</key>
          </record>
        </section>
      </dnx>
    </mets:xmlData>
  </mets:mdWrap>
</mets:digiprovMD>

```

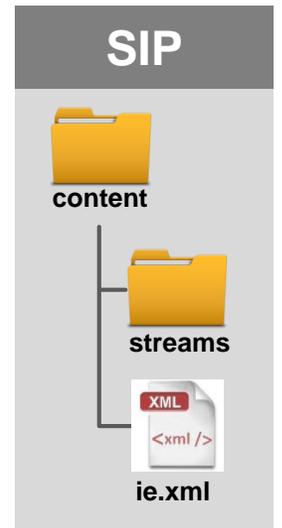
Rosetta Module

- **Deposit:** Manuelle Uploads durch registrierte User
- **Management (Ebene «Institution»)**
 - Konfiguration von Upload-Workflows
 - Analyse von Formatidentifikations- und Validierungsproblemen
 - Datenmanagement
 - User-Management
 - Preservation Planning & Execution
- **Administration:** Übergreifende Einstellungen und Konfigurationen (Ebene «Consortium»)
 - z.B. Speicherverwaltung

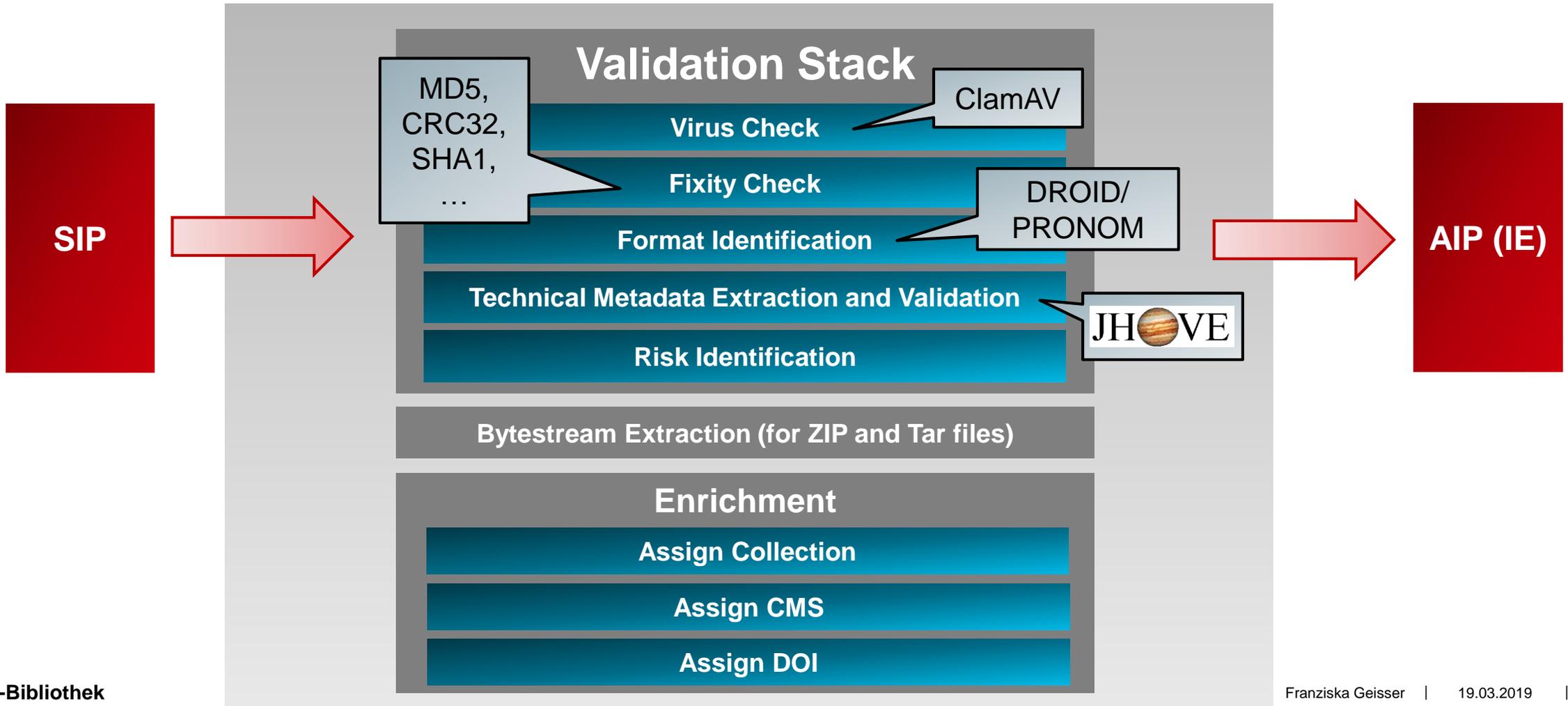
DEPOSITS 	SUBMISSIONS 	DATA MANAGEMENT 	PRESERVATION 
Producers and Agents Producers Producer Profiles Producer Groups 1 st Time Registration Reasons 1 st Time Registration Rules	Technical Analysis Technical Issues	Search and Manage Queries Search for Objects Saved Queries	Risk Analysis View Global Risk Report Manage Preservation Sets
Policies Access Right Policies Access Rights Exceptions Retention Policies	Rules Format Identification Correction Metadata Extraction Error Virus Check Error	Manage Sets and Processes Sets Processes Monitor Process History Publishing Configuration Technical Issues Update Metadata Job Derivative Copy Job	Preservation Plans Preservation Plans Evaluate Test Results Technical Issues
Deposit Arrangements Material Flows Content Structure Metadata Form Submission Format Metadata Profiles Assertion of Copyrights Copyrights Boilerplate Access Right Copyrights Material Type CSV Templates	Approval Assess SIPs Arrange SIPs Approve SIPs	Collection Management Collection Management Collection Publishing	Preservation Executions Signed-off Plans Technical Issues
Jobs Producers Reports Job Submission Job OAI Harvester Job	Search Search for SIPs	Policies Access Right Policies Access Rights Exceptions Retention Policies	Advanced Preservation Activities Run Reports Schedule Reports Schedule Risk Analysis Process Global Audit Trail Local Audit Trail Plan Evaluation Criteria Alternative Evaluation Criteria Evaluation Code Table
Advanced Tools Run Reports Schedule Reports Delivery Metadata Fields Delivery XSL Files Delivery Copyrights Statements Email Configuration OAI Harvester Transformation Terms of Use Configuration Files	Advanced Tools Run Reports Schedule Reports SIP Processing Configuration SIP Routing Rules Approval Group	Advanced Tools Run Reports Schedule Reports Fixity Reports Recycle Bin Manage Users	Format Library Formats Manage Format Local Fields Applications Manage Application Local Fields Risk Identifiers Significant Properties Classification Groups Extractors Manage Format Library Version

Wege für Deposit und Upload

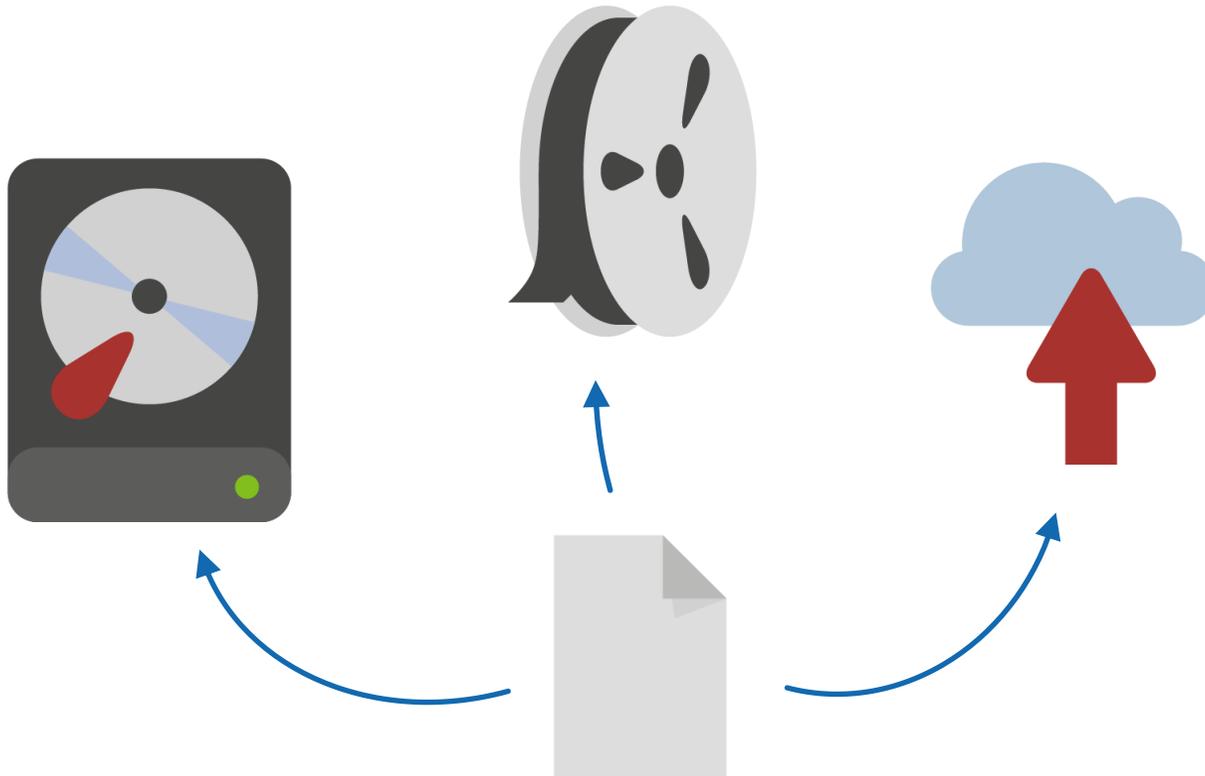
- **Manuell**
Webdialog zum Hochladen und zur Metadatenerfassung
- **Halbautomatisch**
Batch-Upload von Dateien mit vorhandenen Metadaten (CSV, Bagit)
- **Automatisch nach manueller Vorbereitung** im docuteam packer
(lokaler Viewer und Editor für File-Struktur und Metadaten)
- **Automatisch via Submission Application**
Dateien aus verschiedenen Quellsystemen werden mit vorhandenen Metadaten im METS-XML-Format zu einem Rosetta-SIP gebündelt



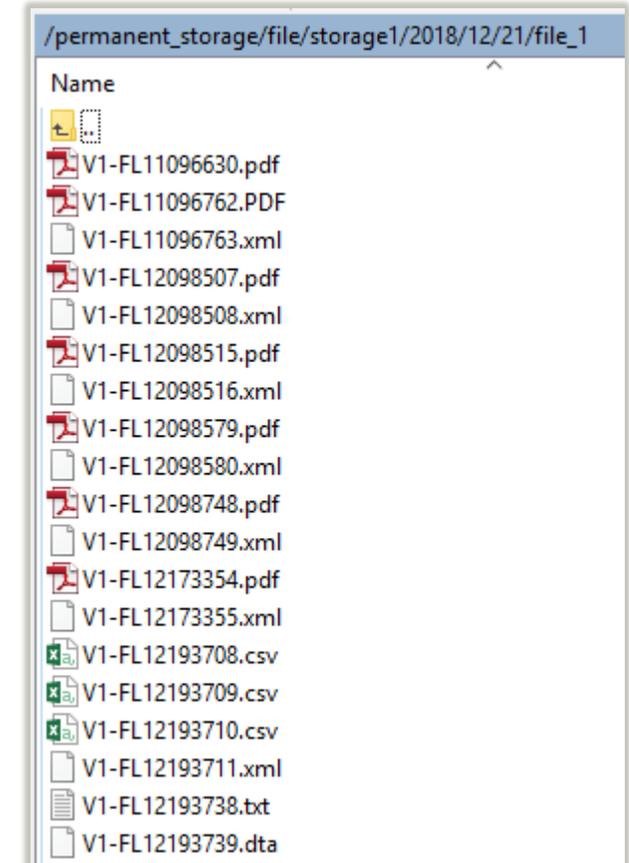
SIP Processing



Permanent Repository



Ablage auf dem Rosetta-Filesystem (default):



Preservation

Preservation Planning mit Rosetta:

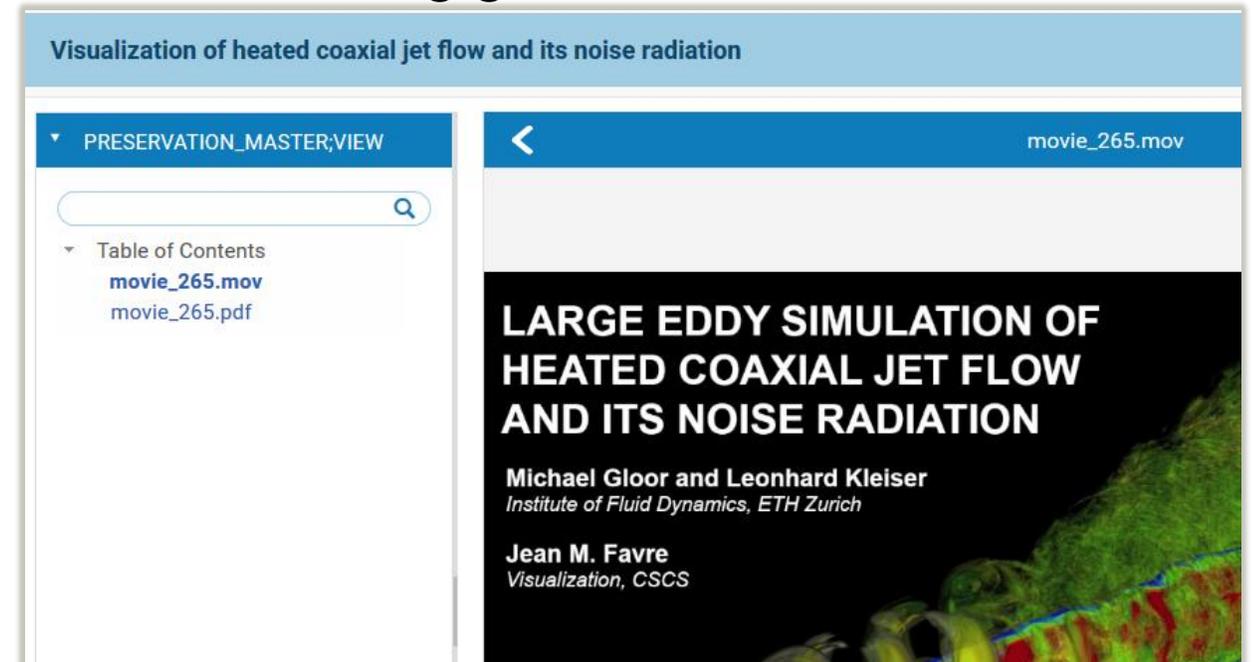
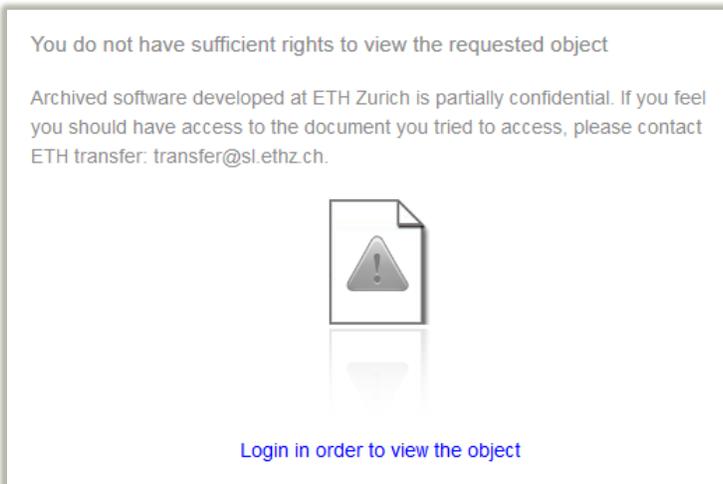
- Risikoanalyse als Ausgangspunkt, basierend auf Formaten und Eigenschaften
- Definieren und Testen von Preservation Plans (Migrationsszenarien)
- Formatmigration «intern» (mittels Plugin) oder «extern» (Export – Reimport)
- Preservation Action erzeugt neue Version

Preservation Planning als intellektuelle Aufgabe:

- Entsprechendes Wissen muss im Team aufgebaut werden
- Bewertung von Formate und ihren Risiken ist eine Daueraufgabe
- Migrationen müssen vom Personal geplant, ausgeführt und geprüft werden

Discovery und Delivery – Zugriff auf Inhalte in Rosetta

- Langzeitarchivierung ist oft «Dark Archive»
- OAI-PMH Publishing von DublinCore-Metadaten, z.B. nach Primo
- Ansicht von Inhalten im Rosetta Viewer – oft abhängig vom Browser
- Rosetta unterstützt differenzierte Access Rights Policies



Rosetta - Bilanz

Pro

- Support
- Vieles ist out-of-the-box vorhanden (z.B. Plugins für Formatidentifikation und Validierung)
- Erweiterbar durch offene Schnittstellen / APIs
- Skalierbarkeit
- Relativ robust
- Aktive internationale Community

Contra

- Kosten
- Geschlossenes System
- z.T. mangelhafte Dokumentation
- Komplexität
- Ressourcenintensiv

Lessons learned

- Nicht zu viel auf einmal wollen
- Genügend IT-Ressourcen einplanen
- Von den Erfahrungen anderer Anwender lernen
- Getrennte Speicherpfade pro Datenproduzent definieren
- Pre-Ingest: Gute Vorbereitung ist die halbe Archivierung!

Infos zu Rosetta

<https://knowledge.exlibrisgroup.com/Rosetta>

Fragen?

Dr. Franziska Geisser
Forschungsdatenmanagement und Datenerhalt
ETH-Bibliothek
Rämistrasse 101
8092 Zürich
044 632 35 96

geisser@library.ethz.ch

<http://www.library.ethz.ch/Forschungsdatenmanagement-und-Datenerhalt>