ETH zürich

Rosetta @ ETH Data Archive

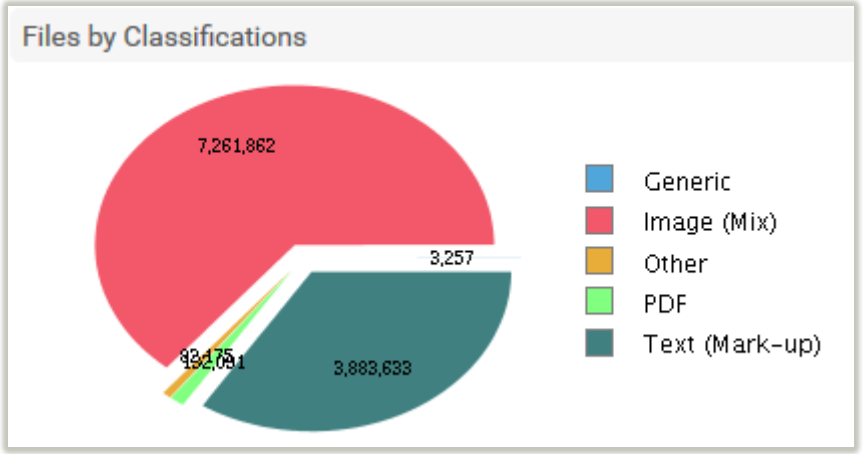
Memoriav Fachtagung 2019

Biel, 19. März 2019

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

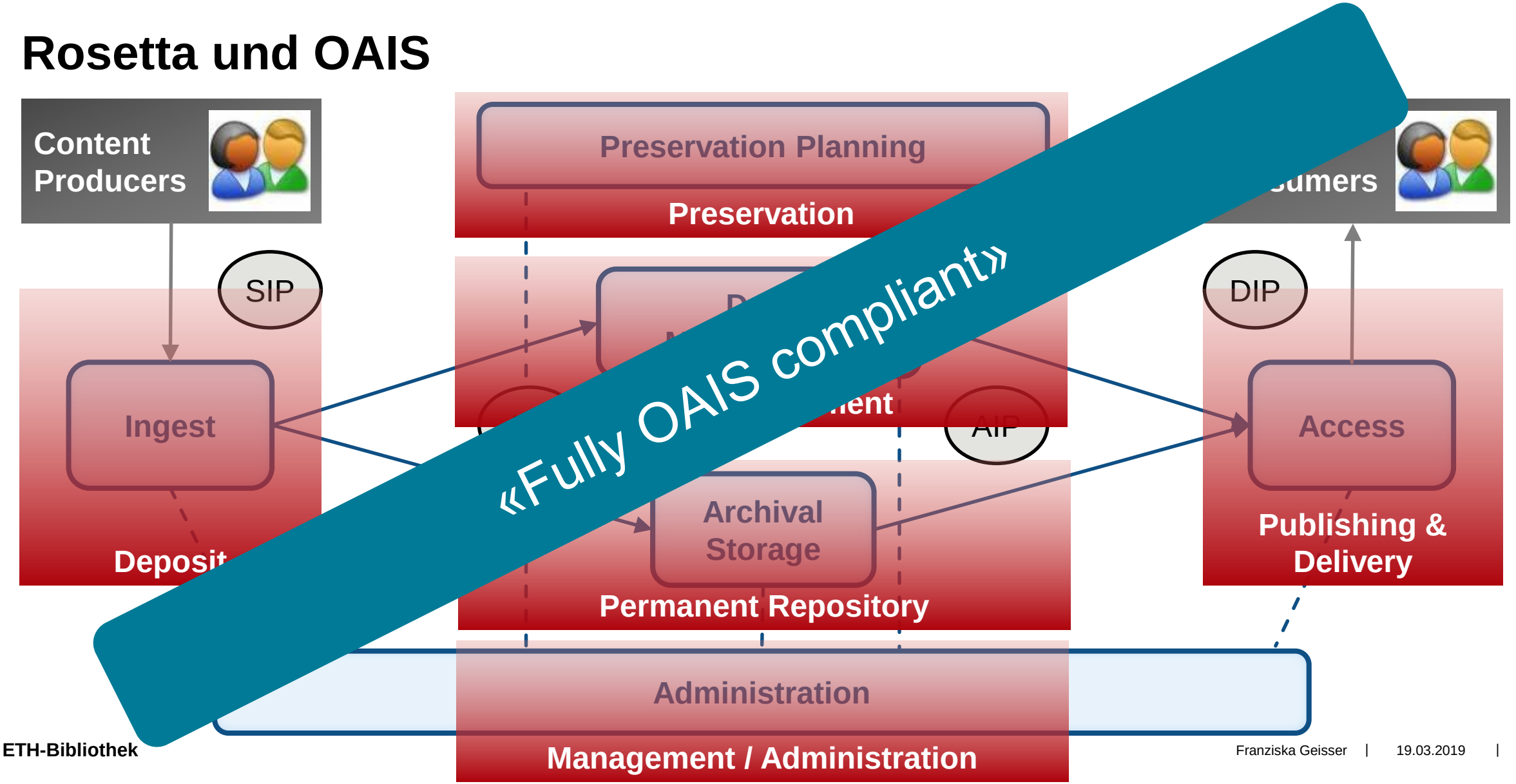
Rosetta an der ETH Zürich

- «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

AIP = IE

Archival Information Package — **Intellectual Entity**

Representation 1 (Preservation Master)

Files
Image1.tif
Image2.tif
Image3.tif
…

Representation 2 (Modified Master)

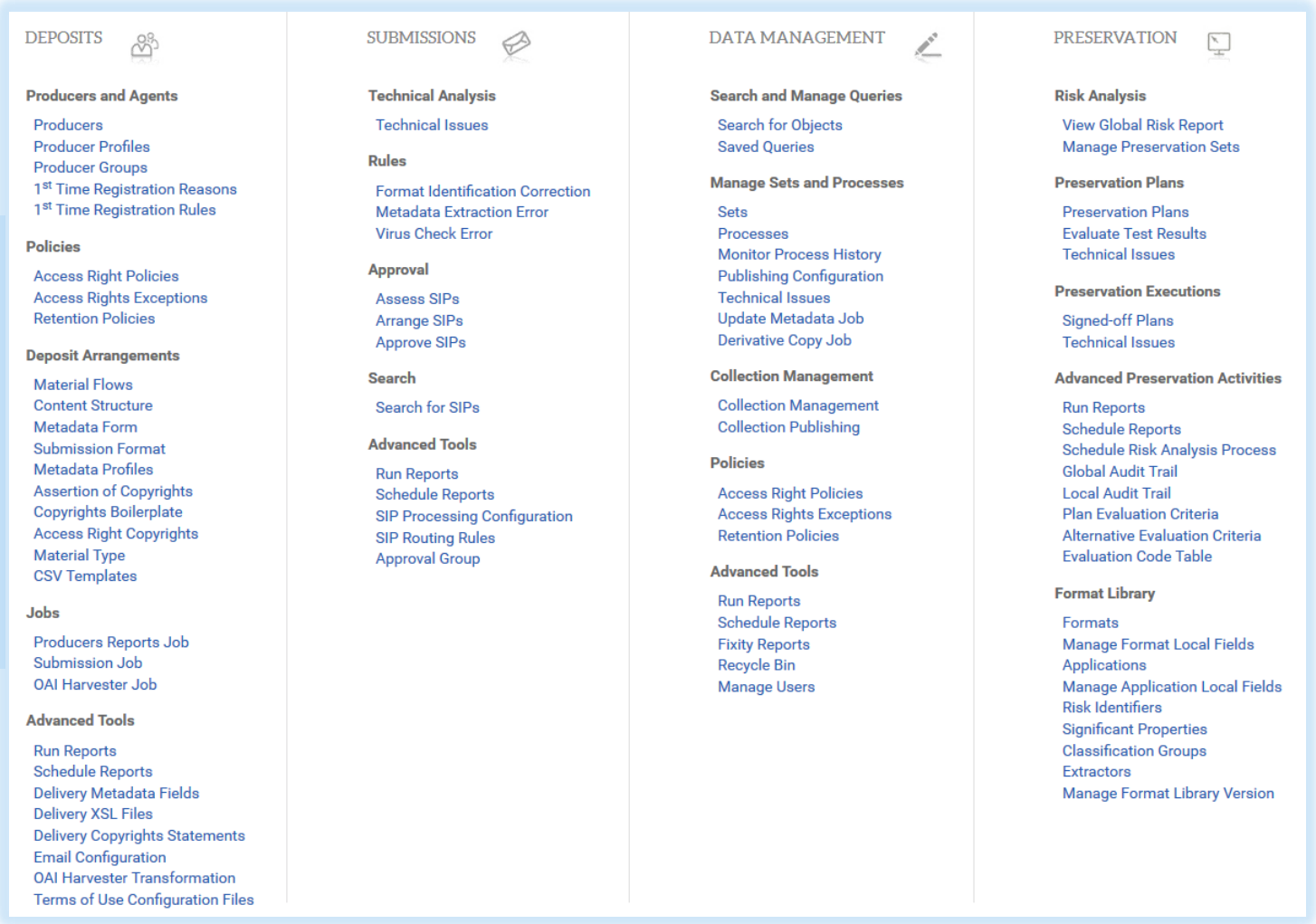
Files
Book.pdf

«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>
```

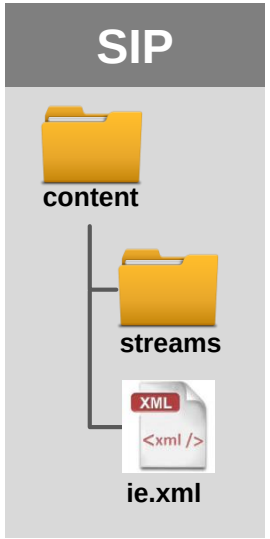
Rosetta Module

- **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

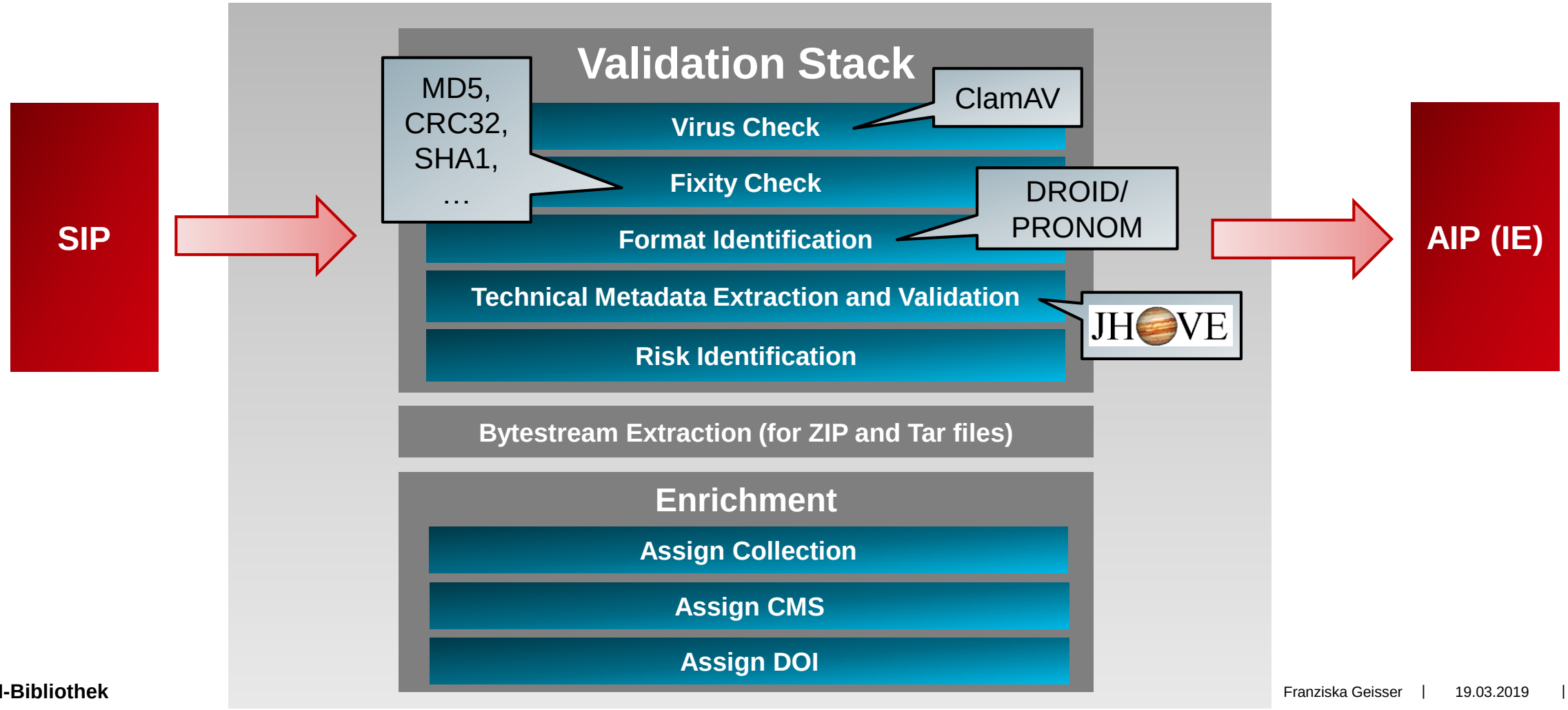


Wege für Deposit und Upload

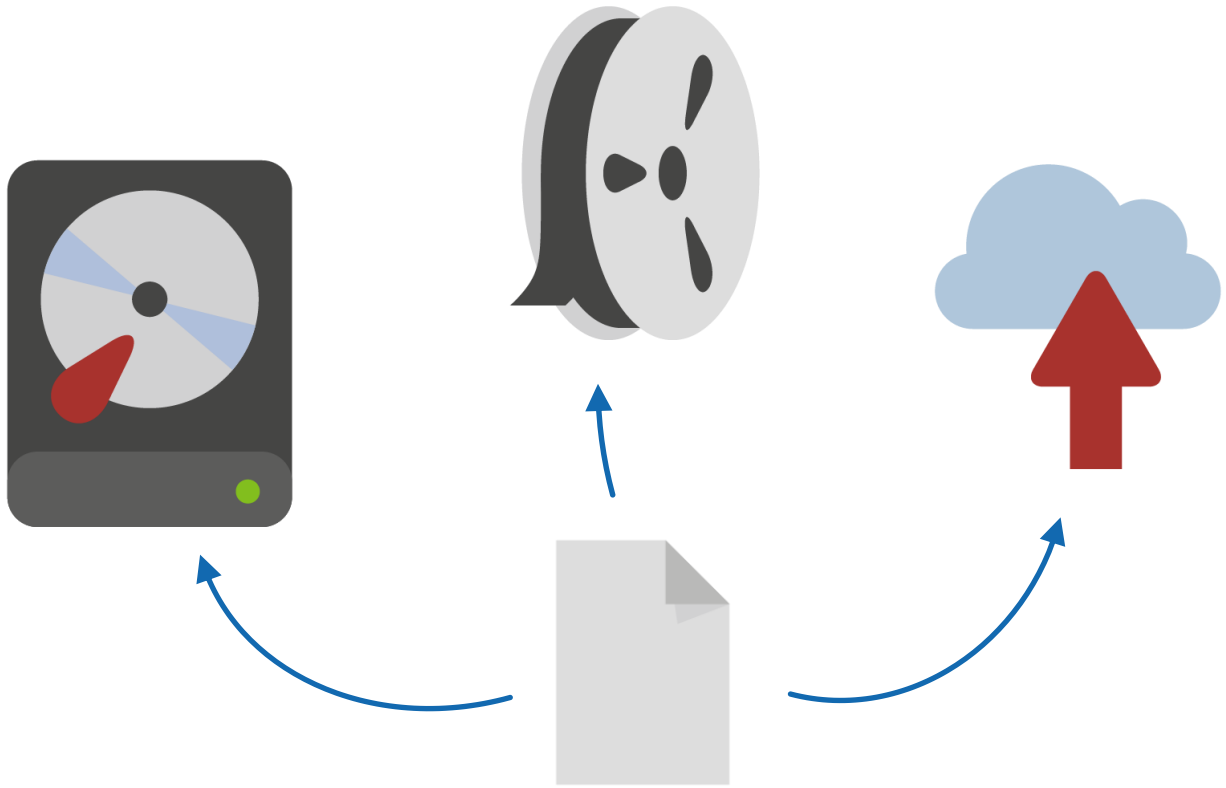
- **Manuell**
  Webdialog zum Hochladen und zur Metadatenerfassung
- **Halbautomatisch**
  Batch-Upload von Dateien mit vorhandenen Metadaten (CSV, Bagit)
- **Automatisch nach manueller Vorbereitung** im docuteam packer (lokaler Viewer und Editor für File-Struktur und Metadaten)
- **Automatisch via Submission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



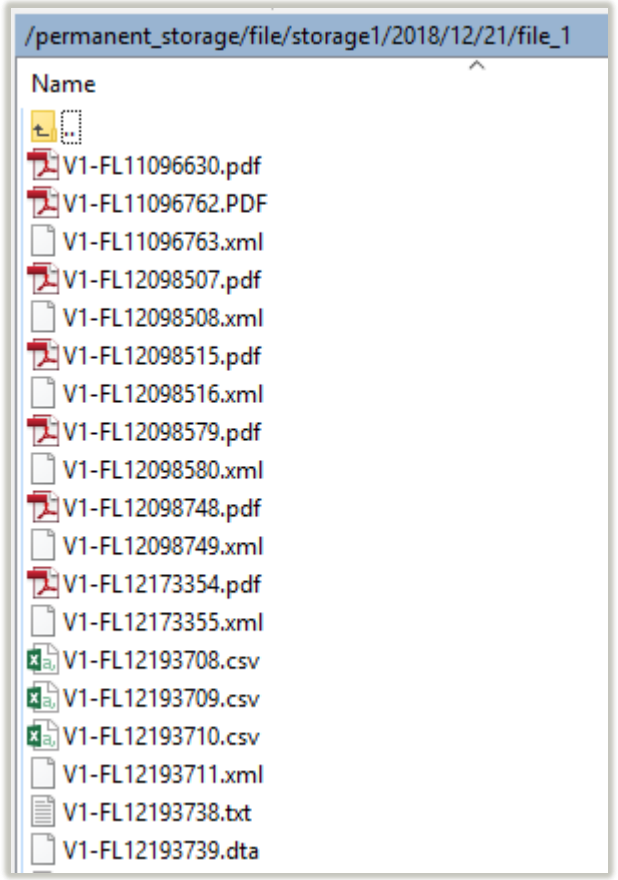
SIP Processing



Permanent Repository



Ablage auf dem Rosetta-Filesystem (default):



Preservation

Preservation Planning mit Rosetta:

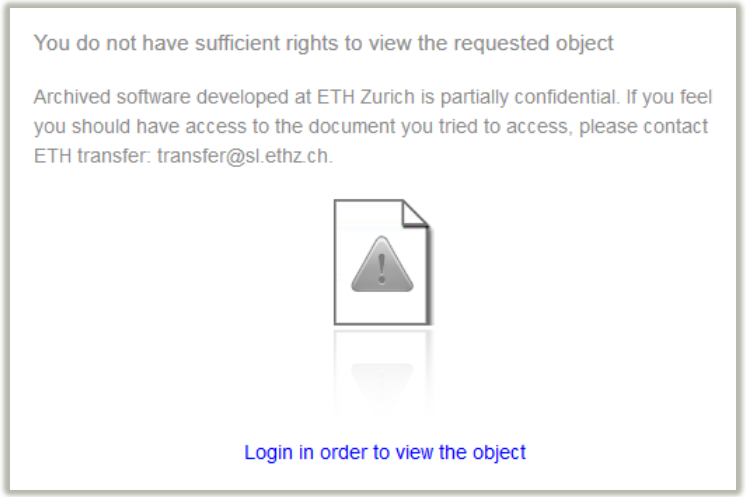
- Risikoanalyse als Ausgangspunkt, basierend auf Formaten und Eigenschaften
- Definieren und Testen von Preservation Plans (Migrationsszenarien)
- Formatmigration «intern» (mittels Plugin) oder «extern» (Export – Reimport)
- Preservation Action erzeugt neue Version

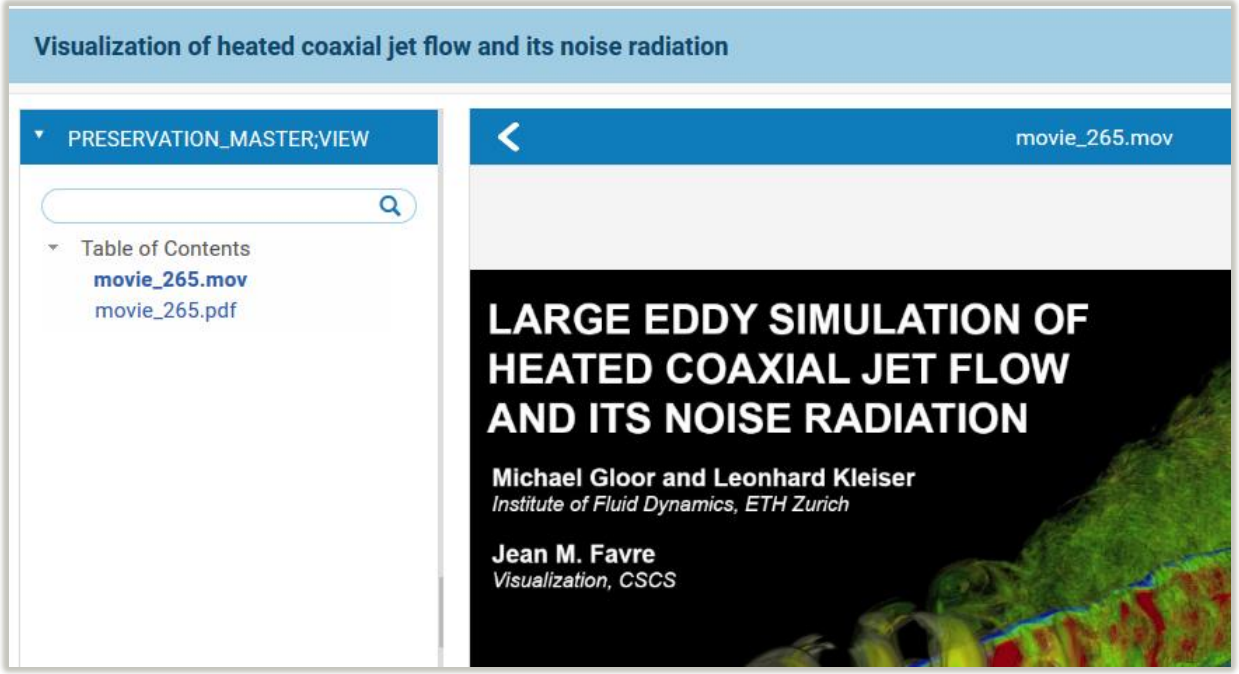
Preservation Planning als intellektuelle Aufgabe:

- 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