

Stefan Hein

# Metadaten für die Langzeitarchivierung

Konzepte, Standards, Erzeugung und Anwendungsszenarien

18. April 2016



# Gliederung

## 1. Einführung

Motivation, Was sind Metadaten?

## 2. Metadatenkonzepte

OAIS-Datenmodell, Metadatenklassen

Theorie

## 3. Metadatenstandards

DC, METS, PREMIS und Co.

## 4. Generierung von Metadaten

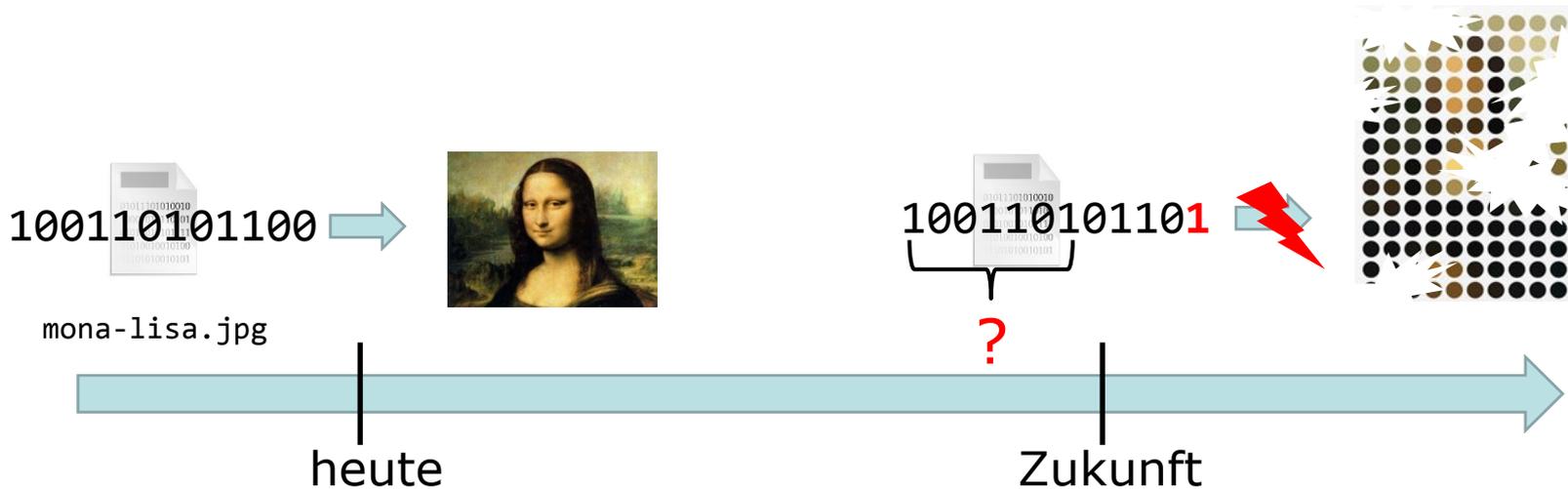
Metadatentools

Praxis

## 5. Anwendungsszenarien

und nun ?

# Einführung - Motivation

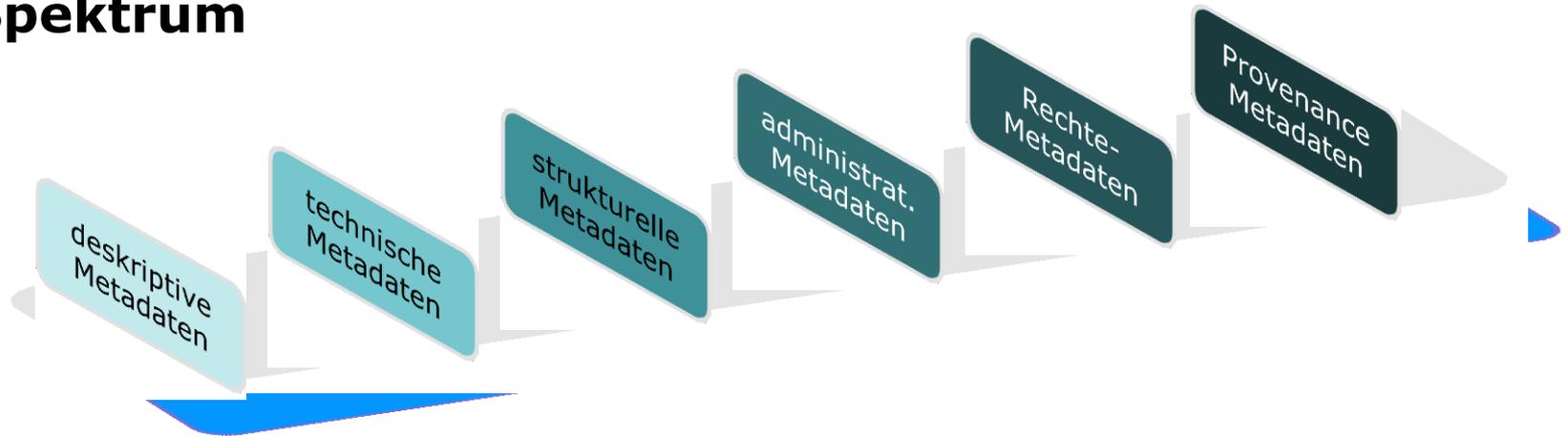


→ Welche Informationen brauchen wir, um den Strom aus Nullen und Einsen auch in Zukunft **unverfälscht** noch **interpretieren** zu können?

→ Wie können wir diese Informationen gewinnen?

## 2. Metadatenkonzept (3)

### Spektrum



- Metadatenklassen dienen unterschiedlichen Bedürfnissen
- Metadaten zur Langzeitarchivierung (Preservation Metadata) bilden einen Querschnitt durch alle Klassen

➔ **Metadaten** als Konzept, die im OAIS-Datenmodell definierten Informationen zu erfassen und zu speichern.

# Gliederung

## 1. Einführung

Motivation, Was sind Metadaten?

## 2. Metadatenkonzepte

OAIS-Datenmodell, Metadatenklassen

Theorie

## 3. Metadatenstandards

DC, METS, PREMIS und Co.

## 4. Generierung von Metadaten

Metadatentools

Praxis

## 5. Anwendungsszenarien

und nun ?

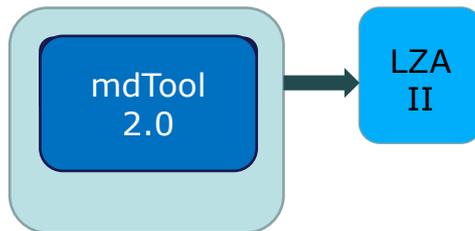
### 3. Metadatenstandards (1)

- Standards definieren und vereinheitlichen die einzelnen Elemente, den Wertebereich und den strukturellen Aufbau

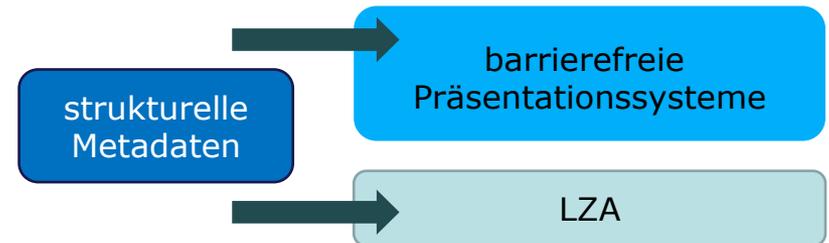
- Interoperabilität



- Austauschbarkeit

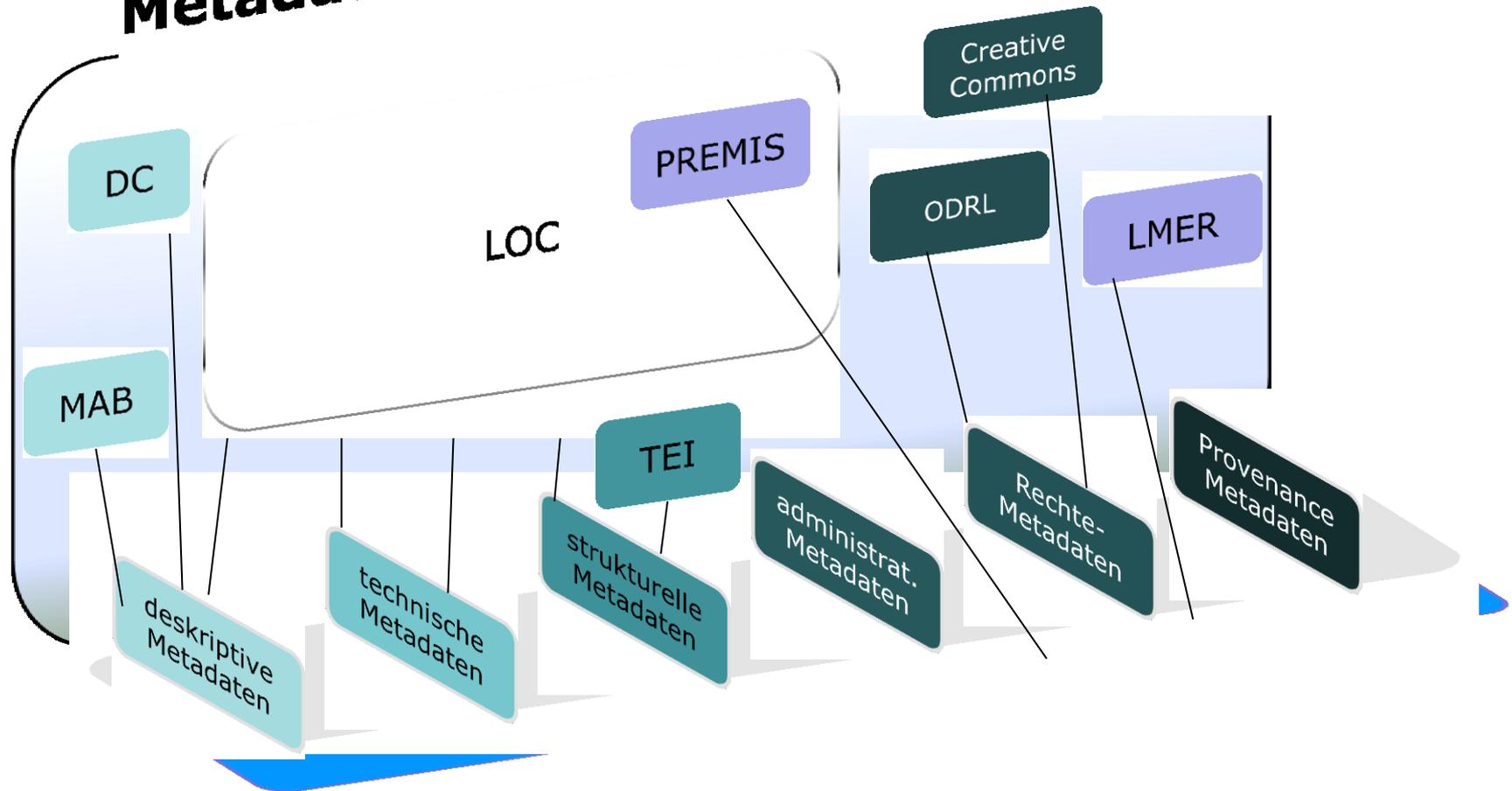


- Wiederverwendbarkeit



- Transparenz
- Integrationsfähigkeit

# Metadatenstandards



## 3. Metadatenstandards (2)

### *Dublin Core Metadata Element Set (DCMES)*

#### **Inhaltsbeschreibung**

*title*: Name der Ressource  
*subject*: Thema der Ressource  
*coverage*: Geltungsbereich  
*description*: Beschreibung der Ressource  
*language*: Sprache

#### **Personen und Rechte**

*creator*: Urheberin / Urheber  
*publisher*: Verlegerin / Verleger  
*contributer*: Mitwirkende /  
Mitwirkender  
*rights*: Rechte

#### **Technische Metadaten**

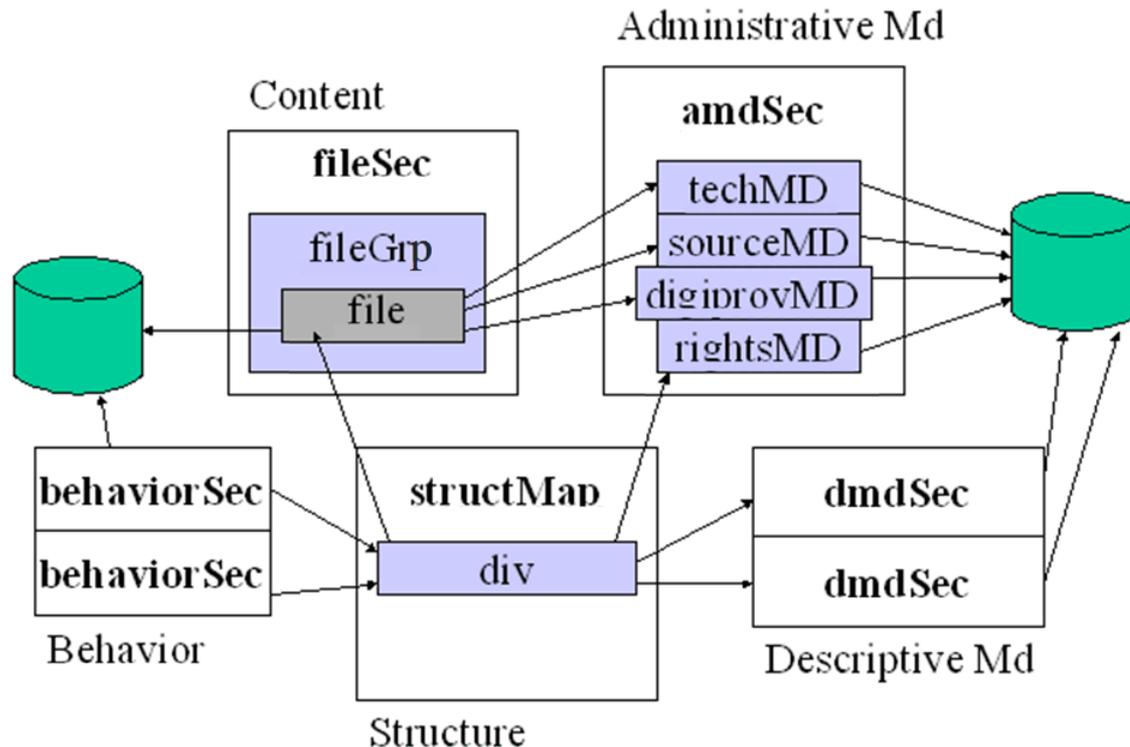
*type*: Art und Gattung der Ressource  
*identifier*: Identifikator  
*format*: Dateiformat, Datenträger  
oder Umfang der Ressource  
*date*: Zeitangabe

#### **Vernetzung**

*source*: Quelle  
*relation*: Beziehung

# 3. Metadatenstandards (3)

## *Metadata Encoding & Transmission Standard (METS)*

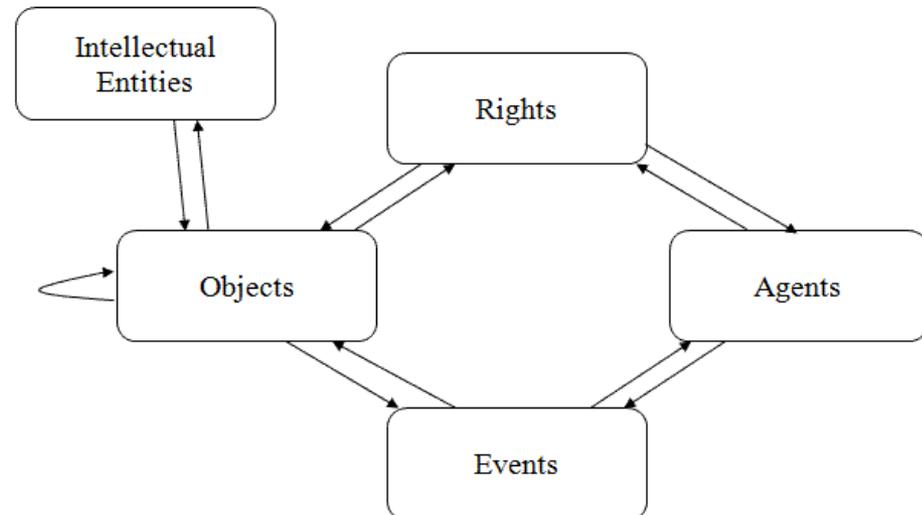


METS-Diagramm (Quelle: modifiziert nach [Beaubien 2002: 17])

## 3. Metadatenstandards (7)

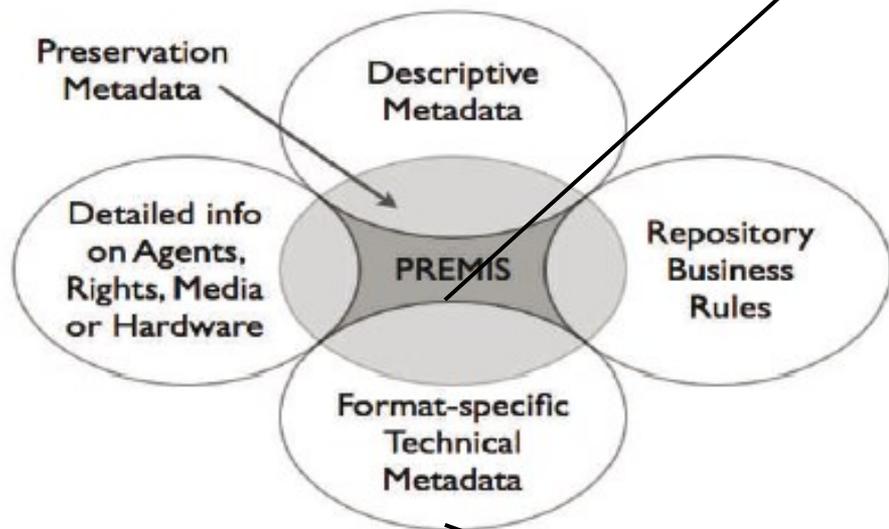
### *PREservation Metadata: Implementation Strategies (PREMIS)*

- international renommiertester Standard für LZA-Metadaten
- Schwerpunkt auf Unterstützung von Erhaltungsstrategien wie Migration und Emulation
- Datenmodell:



# 3. Metadatenstandards (8)

## Data Dictionary



PREMIS als Teilmenge verschiedener Metadattentypen (Quelle: [Caplan 2009: 5])

Semantic unit	1.5.3 size		
Semantic components	None		
Definition	The size in bytes of the file or bitstream stored in the repository.		
Rationale	Size is useful for ensuring the correct number of bytes from storage have been retrieved and that an application has enough room to move or process files. It might also be used when billing for storage.		
Data constraint	Integer		
Object category	Representation	File	Bitstream
Applicability	Not applicable	Applicable	Applicable
Examples		2038937	
Repeatability		Not repeatable	Not repeatable
Obligation		Optional	Optional
Creation / Maintenance notes	Automatically obtained by the repository.		
Usage notes	Defining this semantic unit as size in bytes makes it unnecessary to record a unit of measurement. However, for the purpose of data exchange the unit of measurement should be stated or understood by both partners.		

Die semantische Einheit "size" (Quelle: [Caplan 2009: 13])

# Gliederung

## 1. Einführung

Motivation, Was sind Metadaten?

## 2. Metadatenkonzepte

OAIS-Datenmodell, Metadatenklassen

Theorie

## 3. Metadatenstandards

DC, METS, PREMIS und Co.

## 4. Generierung von Metadaten

Metadatentools

Praxis

## 5. Anwendungsszenarien

und nun ?

## 4. Generierung von Metadaten

*Was lässt sich automatisiert generieren bzw. erfassen?*

- Dateiformat (Identifizierung)
- format[un]spezifische **technische** Metadaten
- **strukturelle** Metadaten
- Dokumentenbeschränkungen
- Dateiformat-Validität
- administrative Metadaten
- Provenance Metadaten

## 4. Generierung von Metadaten (2)

### *Wie generieren und ablegen?*

- Integration von **Metadatentools** in den Verarbeitungsworkflow (Ingest, Preservation Action)
- Metadaten sollten maschinenlesbar (XML) kodiert werden und wenn möglich einem anerkannten Metadatenstandard entsprechen
- ggf. Normalisierung des Outputs
- Speicherung im AIP und im Data-Management des LZA

# Metadatentool: FITS (1)

## Beispiel : JPEG-Datei

```

<?xml version="1.0" encoding="UTF-8" ?>
<fits xmlns="http://hul.harvard.edu/ois/xml/ns/fits/fits_output" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://hul.harvard.edu/ois/xml/ns/fits/fits_output
http://hul.harvard.edu/ois/xml/xsd/fits/fits_output.xsd" version="0.5.0" timestamp="24.03.11 11:32">
  <identification>
    <identity format="JPEG File Interchange Format" mimetype="image/jpeg">
      <tool toolname="Jhove" toolversion="1.5" />
      <tool toolname="file utility" toolversion="5.03" />
      <tool toolname="Exiftool" toolversion="7.74" />
      <tool toolname="Droid" toolversion="3.0" />
      <version toolname="Jhove" toolversion="1.5">1.01</version>
      <externalIdentifier toolname="Droid" toolversion="3.0" type="puid">fmt/43</externalIdentifier>
    </identity>
  </identification>
  <fileinfo>
    <size toolname="Jhove" toolversion="1.5">9071</size>
    <lastmodified toolname="Exiftool" toolversion="7.74" status="SINGLE_RESULT">2011:02:11 14:54:50+01:00</lastmodified>
    <filename toolname="OIS File Information" toolversion="0.1" status="SINGLE_RESULT">bauarbeiter.jpg</filename>
    <md5checksum toolname="OIS File Information" toolversion="0.1" status="SINGLE_RESULT">e63f2ff66703d55ee29aaad329cf4d9b</md5checksum>
    <fslastmodified toolname="OIS File Information" toolversion="0.1" status="SINGLE_RESULT">1297432490221</fslastmodified>
  </fileinfo>
  <filestatus>
    <well-formed toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">true</well-formed>
    <valid toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">true</valid>
  </filestatus>
  <metadata>
    <image>
      <byteOrder toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">big endian</byteOrder>
      <compressionScheme toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">JPEG (old-style)</compressionScheme>
      <imageWidth toolname="Jhove" toolversion="1.5">100</imageWidth>
      <imageHeight toolname="Jhove" toolversion="1.5">100</imageHeight>
      <colorSpace toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">YCbCr</colorSpace>
      <YCbCrSubSampling toolname="Exiftool" toolversion="7.74" status="SINGLE_RESULT">2 2</YCbCrSubSampling>
      <samplingFrequencyUnit toolname="Jhove" toolversion="1.5">in.</samplingFrequencyUnit>
      <xSamplingFrequency toolname="Jhove" toolversion="1.5">96</xSamplingFrequency>
      <ySamplingFrequency toolname="Jhove" toolversion="1.5">96</ySamplingFrequency>
      <bitsPerSample toolname="Jhove" toolversion="1.5">8 8 8</bitsPerSample>
      <samplesPerPixel toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">3</samplesPerPixel>
      <ingestLevel toolname="Didig0" toolversion="0.7" status="SINGLE_RESULT">4</ingestLevel>
    </image>
  </metadata>
</fits>

```

# Metadatentool: FITS (2)

Beispiel : JPEG-Datei

```
<identification>
  <identity format="JPEG File Interchange Format" mimetype="image/jpeg">
    <tool toolname="Jhove" toolversion="1.5" />
    <tool toolname="file utility" toolversion="5.03" />
    <tool toolname="Exiftool" toolversion="7.74" />
  </identity>
</identification>

<fileinfo>
  <size toolname="Jhove" toolversion="1.5">9071</size>
  <lastmodified toolname="Exiftool" toolversion="7.74" status="SINGLE_RESULT">2011:02:11 14:54:50+01:00</lastmodified>
</fileinfo>

<metadata>
  <image>
    <byteOrder toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">big endian</byteOrder>
    <compressionScheme toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">JPEG (old-style)</compressionScheme>
    <imageWidth toolname="Jhove" toolversion="1.5">100</imageWidth>
    <imageHeight toolname="Jhove" toolversion="1.5">100</imageHeight>
    <colorSpace toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">YCbCr</colorSpace>
    <YCbCrSubSampling toolname="Exiftool" toolversion="7.74" status="SINGLE_RESULT">2 2</YCbCrSubSampling>
    <samplingFrequencyUnit toolname="Jhove" toolversion="1.5">in.</samplingFrequencyUnit>
    <xSamplingFrequency toolname="Jhove" toolversion="1.5">96</xSamplingFrequency>
    <ySamplingFrequency toolname="Jhove" toolversion="1.5">96</ySamplingFrequency>
    <bitsPerSample toolname="Jhove" toolversion="1.5">8 8 8</bitsPerSample>
    <samplesPerPixel toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">3</samplesPerPixel>
    <ingestLevel toolname="Didig0" toolversion="0.7" status="SINGLE_RESULT">4</ingestLevel>
  </image>
</metadata>
```

# Gliederung

## 1. Einführung

Motivation, Was sind Metadaten?

## 2. Metadatenkonzepte

OAIS-Datenmodell, Metadatenklassen

Theorie

## 3. Metadatenstandards

DC, METS, PREMIS und Co.

## 4. Generierung von Metadaten

Metadatentools

Praxis

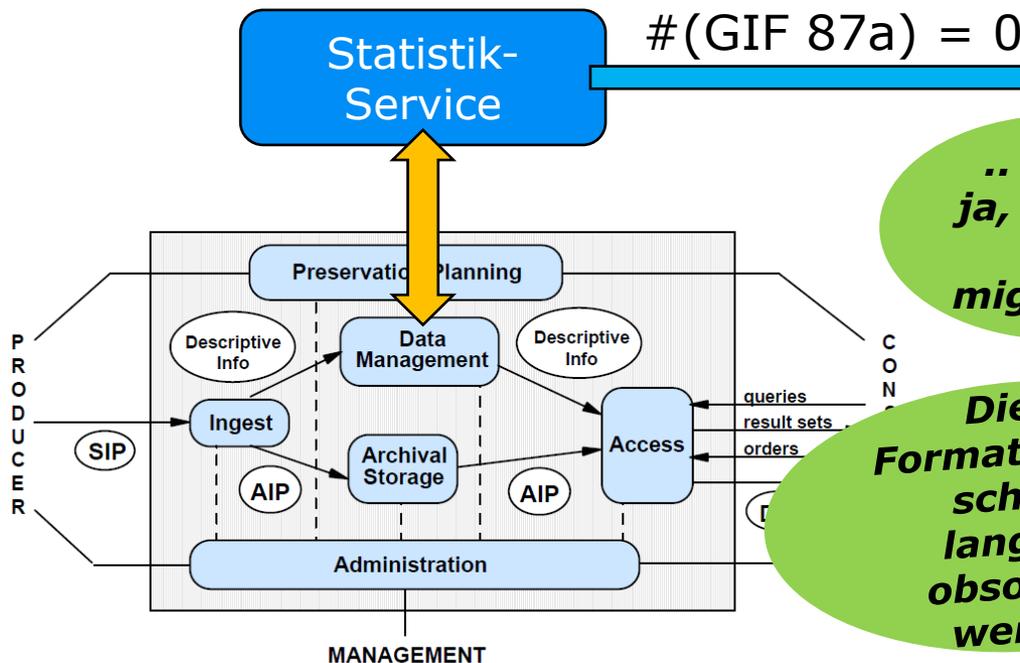
## 5. Anwendungsszenarien

und nun ?

# 5. Anwendungsszenario 1

## Preservation Planning

Metadaten als wertvolle Datenquelle



**Wir müssen überprüfen, ob damit einhergehend auch die Unterstützung durch Viewer und Migrationstools rückläufig ist.**

**Wieviele GIF-Dateien der Version 87a wurden eigentlich in den letzten 3 Jahren archiviert?**

**.. wenn ja, sollten wir migrieren.**

**Diese Formatversion scheint langsam obsolet zu werden!**



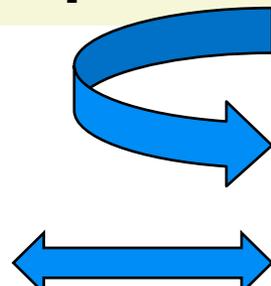
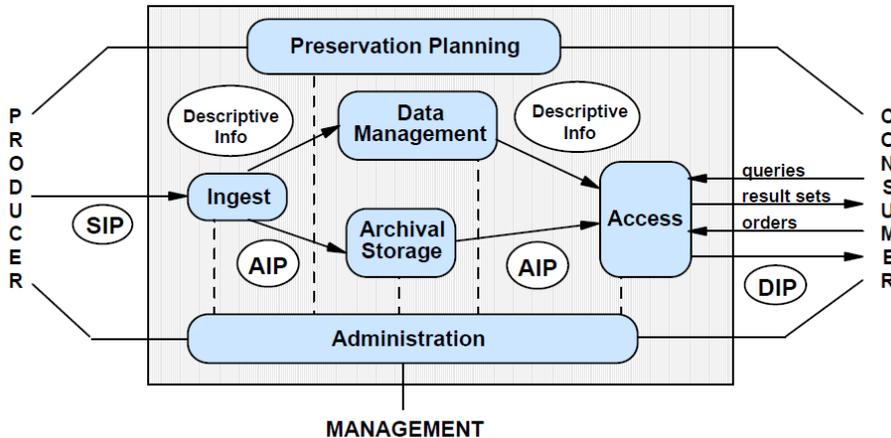
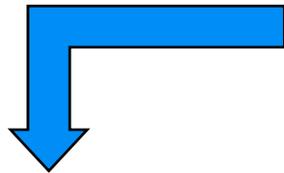
Technology Watch

# 5. Anwendungsszenario 2 - Migration

```
<imageWidth toolname="Jhove" toolversion="1.5">200 </imageWidth>  
<imageHeight toolname="Jhove" toolversion="1.5">40 </imageHeight>
```

*Überprüfung  
der  
signifikanten  
Eigenschaften*

**Qualitätssicherung**



**Migration  
Workflow**

```
SELECT *  
WHERE format="GIF"  
AND version="87a"
```

## 5. Anwendungsszenario 3

*Die Generierung von technischen Metadaten als ein Baustein des Risiko Managements*

beim Ingest – also noch **vor** der Archivierung:

- problembehaftete Objekte **frühzeitig** erkennen

```
<filestatus>  
<well-formed toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">false</well-formed>  
<valid toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">false</valid>  
<message toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">No PDF trailer offset=24608685</message>  
</filestatus>
```

- **frühzeitiges** Erkennen von Sicherheitsbarrieren

```
<isRightsManaged toolname="Exiftool" toolversion="7.74" status="SINGLE_RESULT">yes</isRightsManaged>  
<isProtected toolname="Exiftool" toolversion="7.74" status="SINGLE_RESULT">yes</isProtected>
```

# Quellen

**Beaubien, R. (2002):** METS: An Introduction. Part II. METS Mechanisms with XML. U. C. Berkeley Library Systems Office. <http://www.loc.gov/standards/mets/presentations/METSIntro2.ppt> [15.04.2016]

**Caplan, P. (2009):** Understanding PREMIS. Library of Congress Network Development and MARC Standards Office. <http://www.loc.gov/standards/premis/understanding-premis.pdf> [15.04.2016]

**FITS (2011):** File Information Tool Set (FITS). <http://projects.iq.harvard.edu/fits/home> [15.04.2016]

**Frodl, C./Fischer, T./Baker, T. und Rühle, S. (2007):** Deutsche Übersetzung des Dublin-Core-Metadaten-Elemente-Sets. Version 1.1. Kompetenznetzwerk Interoperable Metadaten. <http://nbn-resolving.de/urn:nbn:de:101:1-200911103125> [15.04.2016]

**LOC (Library of Congress) (2009):** METS: An Overview & Tutorial. Library of Congress, <http://www.loc.gov/standards/mets/METSOverview.v2.html> [15.04.2016]

**CCSDS (2012).** *Reference Model for an Open Archival Information System (OAIS)*. Recommended Practice (No. CCSDS 650.0-M-2).

## zum Nachlesen:

<http://www.ratswd.de/publikationen/langzeitarchivierung-von-forschungsdaten>



# Vielen Dank für Ihre Aufmerksamkeit

Haben Sie noch Fragen ???

Stefan Hein  
Deutsche Nationalbibliothek  
Informationsinfrastruktur  
Adickesallee 1  
D-60322 Frankfurt am Main  
Tel.: +49-69-1525-1722  
<mailto:s.hein@dnb.de>  
<http://www.dnb.de>