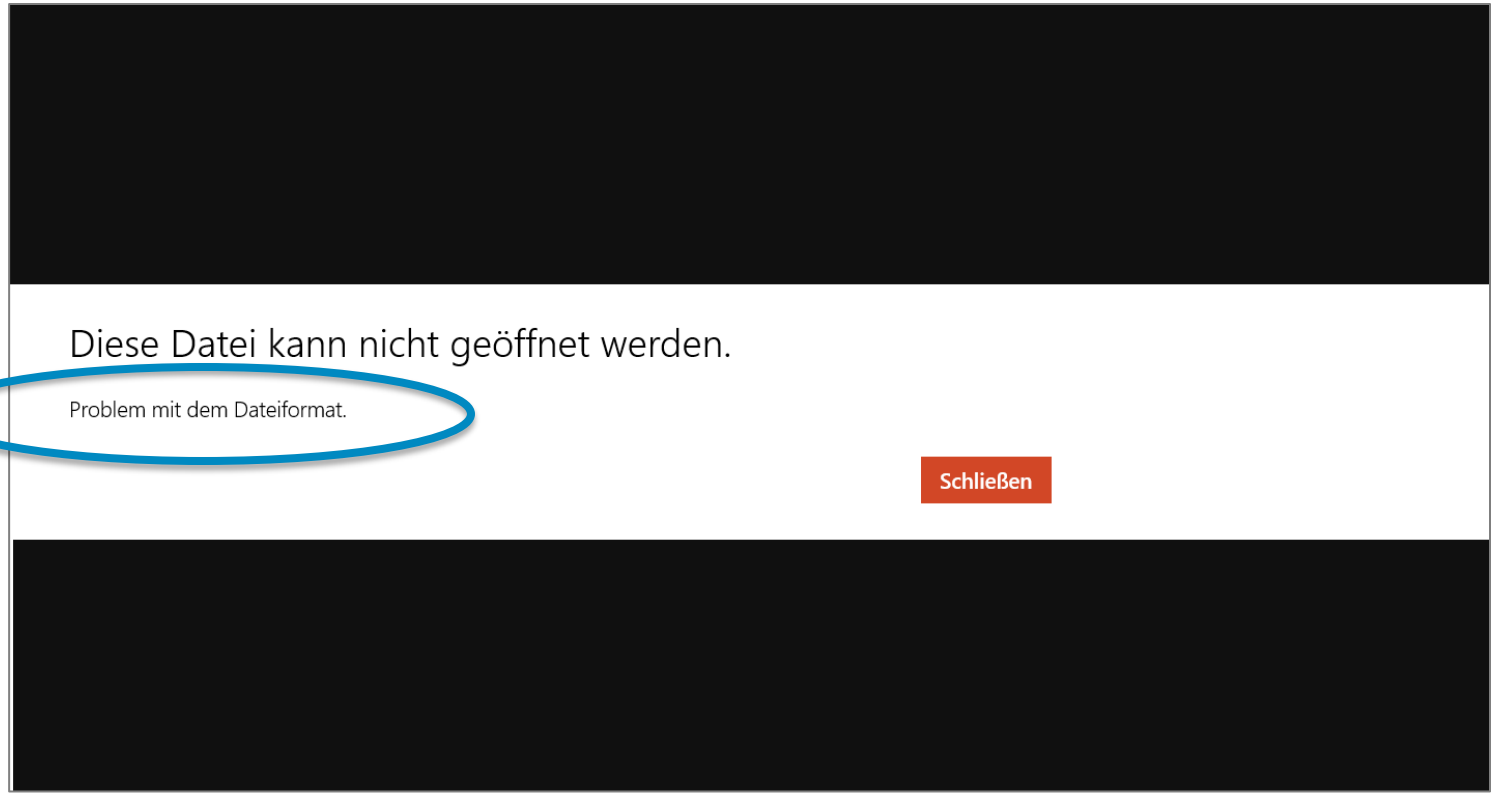
Freistaat Thüringen Landesarchiv

Validierst du noch oder archivierst du schon…

Der Bedarf eines Workflow-Managements in der Formatverifikation

Bedarf eines Workflow-Managements in der Formatverifikation

Freistaat Thüringen Landesarchiv



Unterschiedliche Qualitäten der Formatverifikation

Freistaat Thüringen Landesarchiv

Formaterkennung / Formatidentifizierung

= Ermittlung des Dateiformats anhand folgender Möglichkeiten:

- Dateinamenserweiterung (File Extension)
  → nur Formatfamilie
  → unzuverlässig, da leicht und beliebig veränderbar
- Magische Zahl im Quellcode
- Integrierte Metadaten

Freistaat Thüringen Landesarchiv

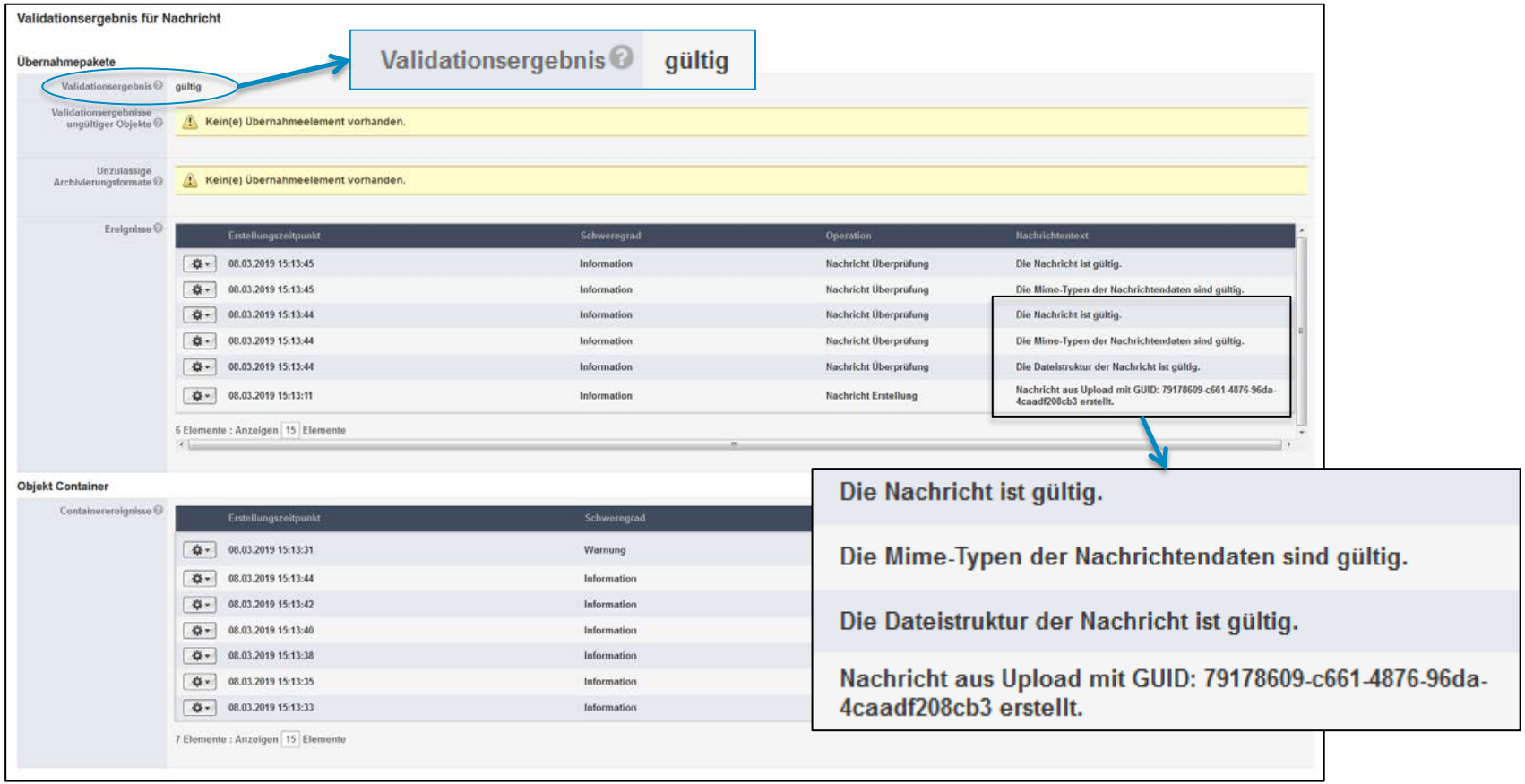
Formatvalidierung

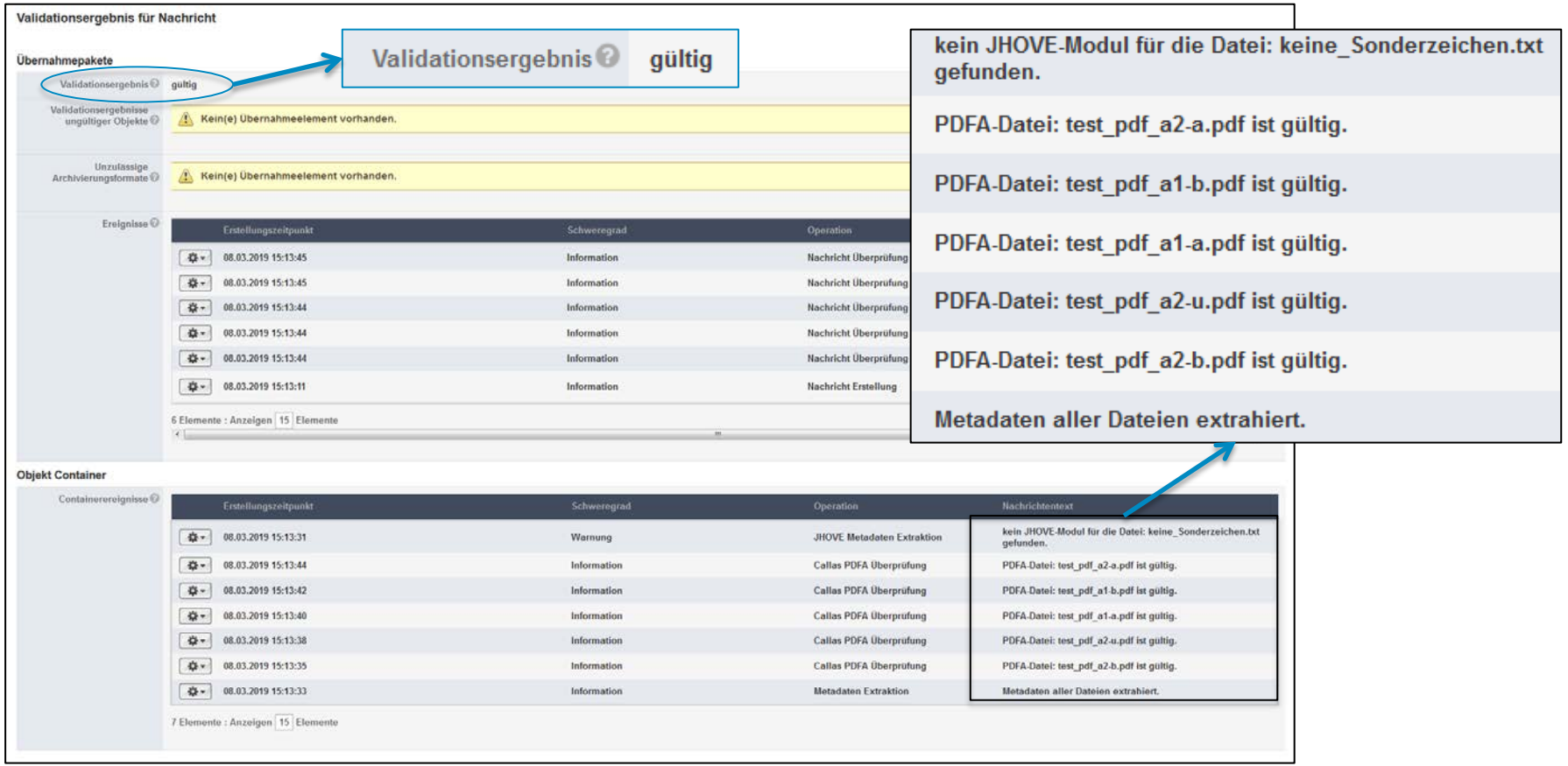
= Prüfung des Dateiformats gegen die Spezifikation des jeweiligen Formats

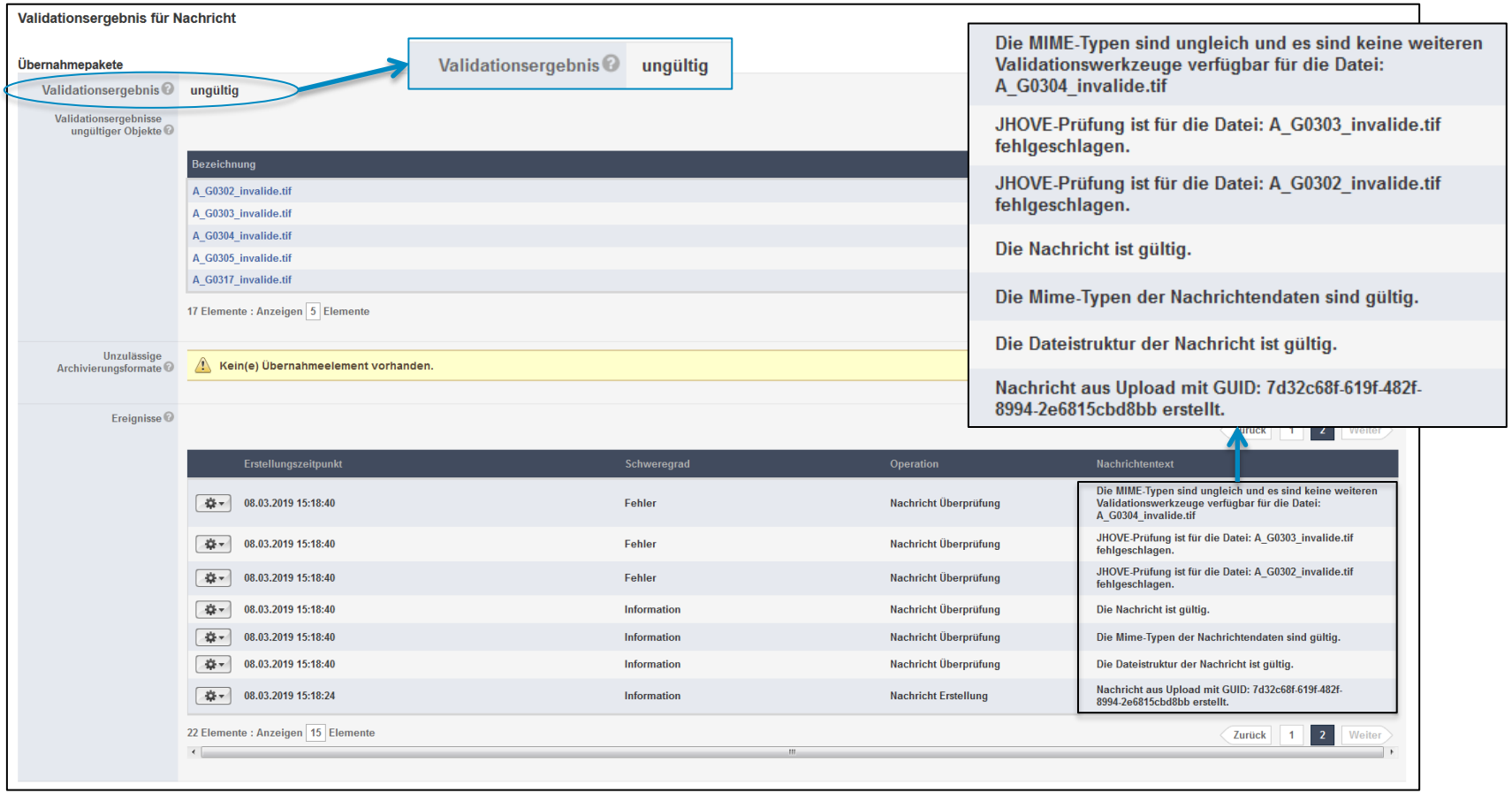
Eine Datei ist hinsichtlich des Formats valide, wenn sie **keine** Bestimmung der entsprechenden Formatspezifikation verletzt!

Freistaat Thüringen Landesarchiv

Aktuelle Marktsituation


Validationsergebnis für Nachricht
Übernahmepakete
Validationsergebnis gültig
Validationsergebnis gültig
Validationsergebnisse ungültiger Objekte
Kein(e) Übernahmeelement vorhanden.
Unzulässige Archivierungsformate
Kein(e) Übernahmeelement vorhanden.
Ereignisse
Erstellungszeitpunkt | Schweregrad | Operation | Nachrichtentext
08.03.2019 15:13:45 | Information | Nachricht Überprüfung | Die Nachricht ist gültig.
08.03.2019 15:13:45 | Information | Nachricht Überprüfung | Die Mime-Typen der Nachrichtendaten sind gültig.
08.03.2019 15:13:44 | Information | Nachricht Überprüfung | Die Nachricht ist gültig.
08.03.2019 15:13:44 | Information | Nachricht Überprüfung | Die Mime-Typen der Nachrichtendaten sind gültig.
08.03.2019 15:13:44 | Information | Nachricht Überprüfung | Die Dateistruktur der Nachricht ist gültig.
08.03.2019 15:13:11 | Information | Nachricht Erstellung | Nachricht aus Upload mit GUID: 79178609-c661-4876-96da-4caadf208cb3 erstellt.
6 Elemente : Anzeigen 15 Elemente
Objekt Container
Containerereignisse
Erstellungszeitpunkt | Schweregrad
08.03.2019 15:13:31 | Warnung
08.03.2019 15:13:44 | Information
08.03.2019 15:13:42 | Information
08.03.2019 15:13:40 | Information
08.03.2019 15:13:38 | Information
08.03.2019 15:13:35 | Information
08.03.2019 15:13:33 | Information
7 Elemente : Anzeigen 15 Elemente
Die Nachricht ist gültig.
Die Mime-Typen der Nachrichtendaten sind gültig.
Die Dateistruktur der Nachricht ist gültig.
Nachricht aus Upload mit GUID: 79178609-c661-4876-96da-4caadf208cb3 erstellt.


Validationsergebnis für Nachricht
Übernahmepakete
Validationsergebnis gültig
Validationsergebnisse ungültiger Objekte
Kein(e) Übernahmeelement vorhanden.
Unzulässige Archivierungsformate
Kein(e) Übernahmeelement vorhanden.
Ereignisse
Erstellungszeitpunkt Schweregrad Operation
08.03.2019 15:13:45 Information Nachricht Überprüfung
08.03.2019 15:13:45 Information Nachricht Überprüfung
08.03.2019 15:13:44 Information Nachricht Überprüfung
08.03.2019 15:13:44 Information Nachricht Überprüfung
08.03.2019 15:13:44 Information Nachricht Überprüfung
08.03.2019 15:13:11 Information Nachricht Erstellung
6 Elemente : Anzeigen 15 Elemente
Objekt Container
Containerereignisse
Erstellungszeitpunkt Schweregrad Operation Nachrichtentext
08.03.2019 15:13:31 Warnung JHOVE Metadaten Extraktion kein JHOVE-Modul für die Datei: keine_Sonderzeichen.txt gefunden.
08.03.2019 15:13:44 Information Callas PDFA Überprüfung PDFA-Datei: test_pdf_a2-a.pdf ist gültig.
08.03.2019 15:13:42 Information Callas PDFA Überprüfung PDFA-Datei: test_pdf_a1-b.pdf ist gültig.
08.03.2019 15:13:40 Information Callas PDFA Überprüfung PDFA-Datei: test_pdf_a1-a.pdf ist gültig.
08.03.2019 15:13:38 Information Callas PDFA Überprüfung PDFA-Datei: test_pdf_a2-u.pdf ist gültig.
08.03.2019 15:13:35 Information Callas PDFA Überprüfung PDFA-Datei: test_pdf_a2-b.pdf ist gültig.
08.03.2019 15:13:33 Information Metadaten Extraktion Metadaten aller Dateien extrahiert.
7 Elemente : Anzeigen 15 Elemente
kein JHOVE-Modul für die Datei: keine_Sonderzeichen.txt gefunden.
PDFA-Datei: test_pdf_a2-a.pdf ist gültig.
PDFA-Datei: test_pdf_a1-b.pdf ist gültig.
PDFA-Datei: test_pdf_a1-a.pdf ist gültig.
PDFA-Datei: test_pdf_a2-u.pdf ist gültig.
PDFA-Datei: test_pdf_a2-b.pdf ist gültig.
Metadaten aller Dateien extrahiert.


Validationsergebnis für Nachricht
Übernahmepakete
Validationsergebnis ungültig
Validationsergebnisse ungültiger Objekte
Bezeichnung
A_G0302_invalide.tif
A_G0303_invalide.tif
A_G0304_invalide.tif
A_G0305_invalide.tif
A_G0317_invalide.tif
17 Elemente : Anzeigen 5 Elemente
Unzulässige Archivierungsformate
Kein(e) Übernahmeelement vorhanden.
Ereignisse
Erstellungszeitpunkt
Schweregrad
Operation
Nachrichtentext
08.03.2019 15:18:40 Fehler Nachricht Überprüfung Die MIME-Typen sind ungleich und es sind keine weiteren Validationswerkzeuge verfügbar für die Datei: A_G0304_invalide.tif
08.03.2019 15:18:40 Fehler Nachricht Überprüfung JHOVE-Prüfung ist für die Datei: A_G0303_invalide.tif fehlgeschlagen.
08.03.2019 15:18:40 Fehler Nachricht Überprüfung JHOVE-Prüfung ist für die Datei: A_G0302_invalide.tif fehlgeschlagen.
08.03.2019 15:18:40 Information Nachricht Überprüfung Die Nachricht ist gültig.
08.03.2019 15:18:40 Information Nachricht Überprüfung Die Mime-Typen der Nachrichtendaten sind gültig.
08.03.2019 15:18:40 Information Nachricht Überprüfung Die Dateistruktur der Nachricht ist gültig.
08.03.2019 15:18:24 Information Nachricht Erstellung Nachricht aus Upload mit GUID: 7d32c68f-619f-482f-8994-2e6815cbd8bb erstellt.
22 Elemente : Anzeigen 15 Elemente
Zurück 1 2 Weiter
Die MIME-Typen sind ungleich und es sind keine weiteren Validationswerkzeuge verfügbar für die Datei: A_G0304_invalide.tif
JHOVE-Prüfung ist für die Datei: A_G0303_invalide.tif fehlgeschlagen.
JHOVE-Prüfung ist für die Datei: A_G0302_invalide.tif fehlgeschlagen.
Die Nachricht ist gültig.
Die Mime-Typen der Nachrichtendaten sind gültig.
Die Dateistruktur der Nachricht ist gültig.
Nachricht aus Upload mit GUID: 7d32c68f-619f-482f-8994-2e6815cbd8bb erstellt.

Freistaat Thüringen Landesarchiv

We have a dream…

Workflowentwurf

Möglichkeiten und Grenzen der Umsetzung

Freistaat Thüringen Landesarchiv

DROID

<meta name="**mime-type**" content="**application/pdf**"/>
<meta name="**name" content**="**Acrobat PDF/A - Portable Document Format**"/>
<meta name="**puid**" content="**fmt/95**"/>
<meta name="**version**" content="**1b**"/>

Apache TIKA

<meta name="**pdf:PDFVersion**" content="**1.4**"/>
<meta name="**pdfa:PDFVersion**" content="**A-1b**"/>
<meta name="**dc:format**" content="**application/pdf**; version=**1.4**"/>
<meta name="**dc:format**" content="**application/pdf**; version="**A-1b**""/>
<meta name="**Content-Type**" content="**application/pdf**"/>
<meta name="**X-Parsed-By**" content="**org.apache.tika.parser.pdf.PDFParse**r"/>
<meta name="**pdfaid:conformance**" content="**B**"/>
<meta name="**pdfaid:part**" content="**1**"/>

Freistaat Thüringen Landesarchiv

JHOVE

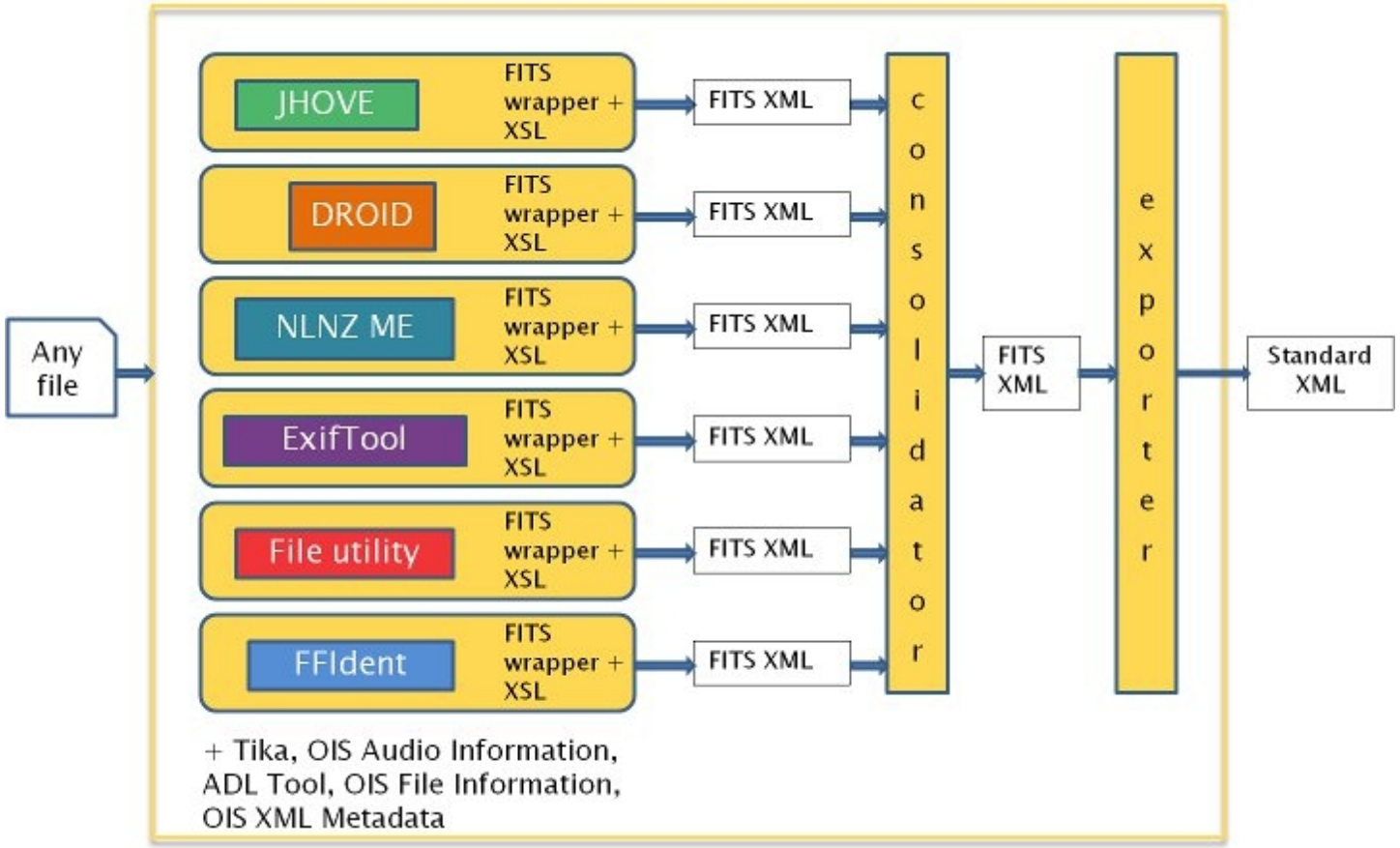
<reportingModule release="1.9" date="2017-07-20">**PDF-hul**</reportingModule>

<format>**PDF**</format>
<version>**1.4**</version>
<status>**Well-Formed and valid**</status>
<**mimeType**>**application/pdf**</mimeType>
<profile>**ISO PDF/A-1, Level B**</profile>
<profile>**ISO PDF/A-1, Level A**</profile>

Callas PDFPilot

<pdfa>
<entry key="**xmp_pdfaid_part**">**1**</entry>
<entry key="**xmp_pdfaid_conformance**">**B**</entry>
</pdfa>
<display_name>**PDF document is compliant with PDF/A-1b (2005)**</display_name>
<display_comment>**Checks whether the PDF file is compliant with PDF/A-1b (2005)**
</display_comment>

Beispiel Tool-Set FITS



FITS

```
 <identification>
  <identity format="PDF/A" mimetype="application/pdf" toolname="FITS"
toolversion="1.4.0">
    <tool toolname="Droid" toolversion="6.4" />
    <tool toolname="Jhove" toolversion="1.20.1" />
    <tool toolname="Exiftool" toolversion="11.14" />
    <tool toolname="Tika" toolversion="1.19.1" />
    <version toolname="Droid" toolversion="6.4">1b</version>
    <externalIdentifier toolname="Droid" toolversion="6.4"
type="puid">fmt/354</externalIdentifier>
  </identity>
 </identification>

<filestatus>
  <well-formed toolname="Jhove" toolversion="1.20.1"
status="SINGLE_RESULT">true</well-formed>
  <valid toolname="Jhove" toolversion="1.20.1" status="SINGLE_RESULT">true</valid>
 </filestatus>
```

Freistaat Thüringen Landesarchiv

Was wurde im Beispiel gezeigt?

- Sofern mehrere Tools in einer Software integriert angesprochen werden, laufen diese parallel und werden nicht miteinander verknüpft eingesetzt.
- Im besten Fall werden Toolergebnisse zu einem Gesamtergebnis konsolidiert.

Wie kann dies zum Workflow ausgebaut werden?

- Die Tools werden in Abhängigkeit zueinander ausgeführt.
- Die Ergebnisse der Tools werden miteinander verknüpft.

Was ist dafür erforderlich?

- Durch das Mapping der Ergebnisstrings soll ein vergleichbares Vokabular erzeugt werden.

Mapping der Ergebnisstrings

Freistaat Thüringen Landesarchiv

DROID

<meta name="**mime-type**" content="**application/pdf**"/>
<meta name="**name" content="Acrobat PDF/A - Portable Document Format**"/>
<meta name="**puid**" content="**fmt/95**"/>
<meta name="**version**" content="**1b**"/>

Apache TIKA

<meta name="**pdf:PDFVersion**" content="**1.4**"/>
<meta name="**pdfa:PDFVersion**" content="**A-1b**"/>
<meta name="**dc:format**" content="**application/pdf**; version=**1.4**"/>
<meta name="**dc:format**" content="**application/pdf**; version="**A-1b**""/>
<meta name="**Content-Type**" content="**application/pdf**"/>
<meta name="**X-Parsed-By**" content="**org.apache.tika.parser.pdf.PDFParser**"/>
<meta name="**pdfaid:conformance**" content="**B**"/>
<meta name="**pdfaid:part**" content="**1**"/>

Mapping

MIME-Type = *application/pdf*

PDF Version = *A-1b*

Freistaat Thüringen Landesarchiv

JHOVE

<reportingModule release="1.9" date="2017-07-20">**PDF-hul**</reportingModule>

<format>**PDF**</format>
<version>**1.4**</version>
<status>**Well-Formed and valid**</status>
<**mimeType**>**application/pdf**</mimeType>
<profile>**ISO PDF/A-1, Level B**</profile>
<profile>**ISO PDF/A-1, Level A**</profile>

Callas PDFPilot

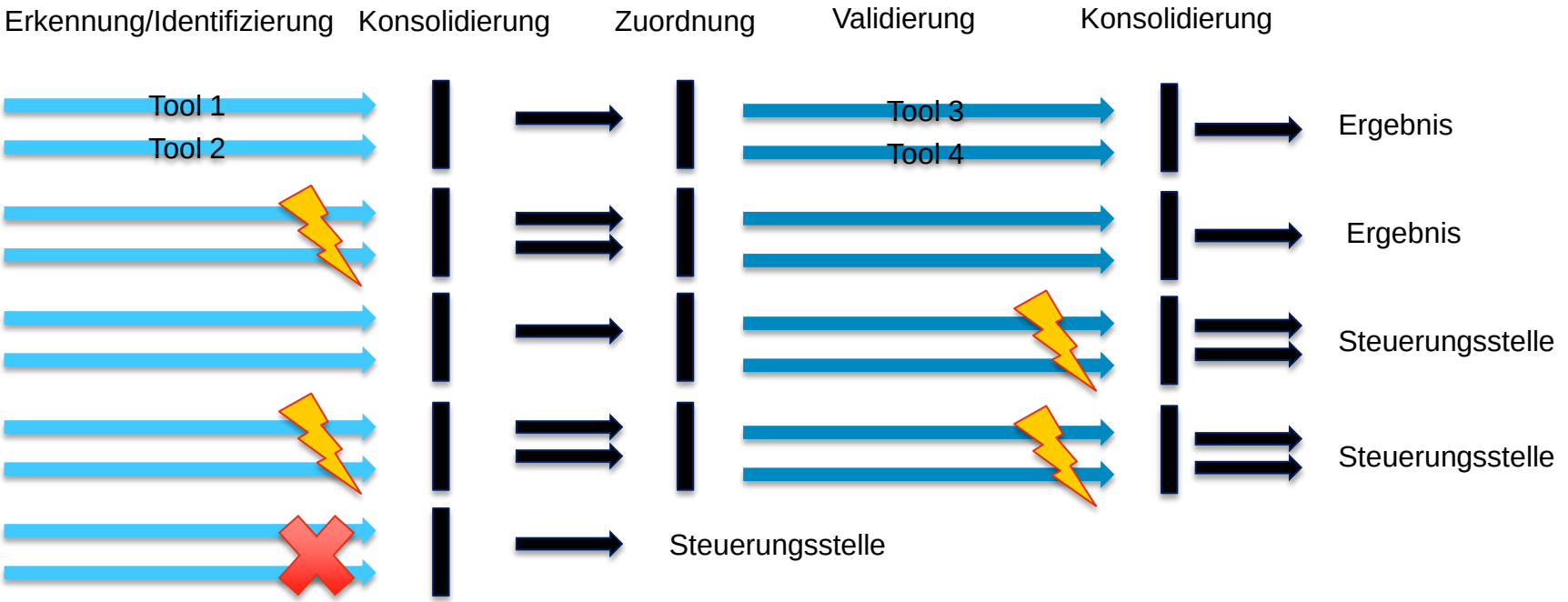
<pdfa>
<entry key="**xmp_pdfaid_part**">**1**</entry>
<entry key="**xmp_pdfaid_conformance**">**B**</entry>
</pdfa>
<display_name>**PDF document is compliant with PDF/A-1b (2005)**</display_name>
<display_comment>**Checks whether the PDF file is compliant with PDF/A-1b (2005)**
</display_comment>

Mapping

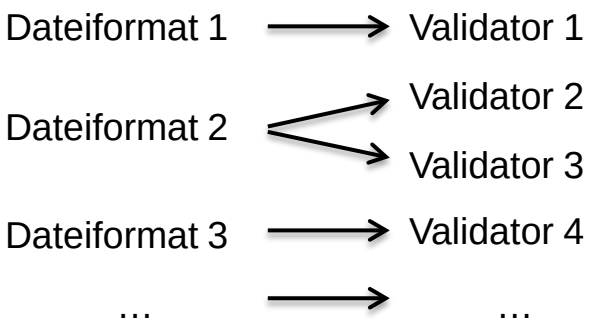
MIME-Type = *application/pdf*

PDF Version = *A-1b*

Status = *valid*


Erkennung/Identifizierung
Konsolidierung
Zuordnung
Validierung
Konsolidierung
Tool 1
Tool 2
Tool 3
Tool 4
Ergebnis
Ergebnis
Steuerungsstelle
Steuerungsstelle
Steuerungsstelle

- Mapping der Ergebnisstrings zur Herstellung einer Vergleichbarkeit der einzelnen Tools
- Zuordnung von Dateiformaten zu Validatoren



Freistaat Thüringen Landesarchiv

Landesarchiv Thüringen
Projekt Digitales Magazin
Marstallstraße 2
99423 Weimar
www.thueringen.de/landesarchiv

Christine Träger
Tel.: +49 (0)3643 870 135
christine.traeger@la.thueringen.de

Daniel Wittmann
Tel.: +49 (0)3643 870 163
daniel.wittmann@la.thueringen.de