Staatsbibliothek zu Berlin
Preußischer Kulturbesitz

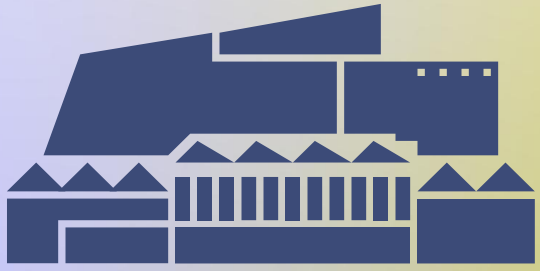
StabiV

Ein Open-Source-Werkzeug zur automatisierten, massenweisen Validierung und Qualitätskontrolle von Digitalisaten

Achim Hatzius (Digitalisierungszentrum SBB)

Thomas Low (knopflogik GmbH)

05.11.2026





- Stand heute
  - ca. 250.000 Digitalisate von Drucken, Handschriften, Musiknoten, Karten und anderen Medien
  - ca. 650.000 Ausgaben von über 200 digitalisierten Zeitungen
- TIFF 6.0 Baseline-konforme Dateien
- orientiert an den Praxisregeln Digitalisierung der DFG, Metamorfoze und ISO 19264-1:2021
- Inhouse-Digitalisierung, Dienstleister-Digitalisierung und Übernahme von Digitalisaten

2025

- insgesamt ca. 3,4 Millionen neu produzierte Images
- je ca. 1,7 Millionen Images Inhouse / durch Dienstleister
- Inhouse:
  - ca. 600.000 Images von Rollfilm (Zeitungen)
  - ca. 1,1 Millionen Images von monographischen und Einblattmaterialien
- ca. 100.000 übernommene Digitalisate unterschiedlicher Provenienz

2027

- perspektivisch steigende Volumina durch verstärkte Auslastung der Scanarbeitsplätze Inhouse (Randzeiten)
- zusätzlich bis zu 2,5 Millionen Images aus Dienstleister-Digitalisierung

Anzahl Mitarbeitende in der Bildkontrolle

- 2 Vollzeitkräfte
- 1 befristete Teilzeitkraft

→ **Unterstützung durch Automatisierung bei**

- Datenübernahme
- Datei-Validierung
- Prüfung Image-Eigenschaften
- Prüfung Image-Metadaten
- Prüfung der Konsistenz über Vorgänge hinweg

→ **StabiV**

- basiert auf JHOVE und Exiftool
- validiert und prüft:
  - Einzeldateien
  - Einzelverzeichnisse
  - Unterverzeichnisse (batch)
- erzeugt:
  - einen formatierten Excel-Report zur schnellen Erfassung der Prüfergebisse (im batch für alle Verzeichnisse)
  - einen CSV-Report (im batch für alle Verzeichnisse)
  - bei erfolgreicher Prüfung je fehlerfreiem Verzeichnis eine Prüfsummendatei
  - Exit-codes
- ist flexibel Konfigurierbar:
  - Prüfkriterien auf Datei und Verzeichnisebene in Konfigurationsdateien, die für den jeweiligen Prüfungszeck individuell erstellt werden können (yaml)
  - Ausgabeformate und Dateibenennung der Excel und CSV Reports
  - Excel-Report: Anzeige und Benennung der Prüfkriterien (Spalten) und Formatierung

StabiV – Implementierung im Datenübernahme-Workflow der SBB

Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

I.) PPN in der Datenbank suchen
wenn kein Images dann weitermachen
wenn Images → fertig

II) shalsum * sha1 testPh
Erfolgreich weitermachen
↳ Erfolgreich PPN ⇒ Fehler kopieren
Mail an DL, IIF

III) StabiV rufen
Erfolgreich → weitermachen
↳ Erfolgreich ⇒ Mail an DL, IIF
PPN ⇒ Fehler kopieren

Ia) PPN prüfen ob in Fehler vorhanden
wenn ja sha1 Datei ≠ → II
wenn sha1 Datei = → fertig
wenn nein → II

IV) Dateinamen prüfen
↳ → kopieren nach Werkbank
↳ ⇒ Mail an DL, IIF
kopie nach Fehler

Skizze: Armin Möller, mit freundlicher Genehmigung des Autors

skriptbasierter Datenübernahme-Workflow am Beispiel Dienstleister-Digitalisierung

- Einsatz bei Datenlieferanten zur Vorprüfung sowie im automatisierten Datenübernahme-Workflow
- automatisierter Sync der in Vorgangsordnern auf SFTP bereitgestellten Images mit einem Speicherbereich der SBB (zeitgesteuert)
- hier Prüfung der Vorgangsordner und Images:
  - Prüfsummendatei vorhanden? (Annahmekriterium)
  - Prüfsummen korrekt?
  - StabiV auf noch nicht übernommene Vorgänge/Images inkl. der Erzeugung von XLSX und CSV Report
  - Vorgang fehlerfrei: Übernahme der Images in vorbereitete Kitodo Vorgänge
  - Vorgang fehlerhaft: Ablage in Nachweis-Verzeichnis
- Erzeugen einer E-Mail an Dienstleister und Bildkontrolle SBB
  - neu übernommene Vorgänge
  - fehlerhafte Vorgänge
  - bereits übernommene Vorgänge

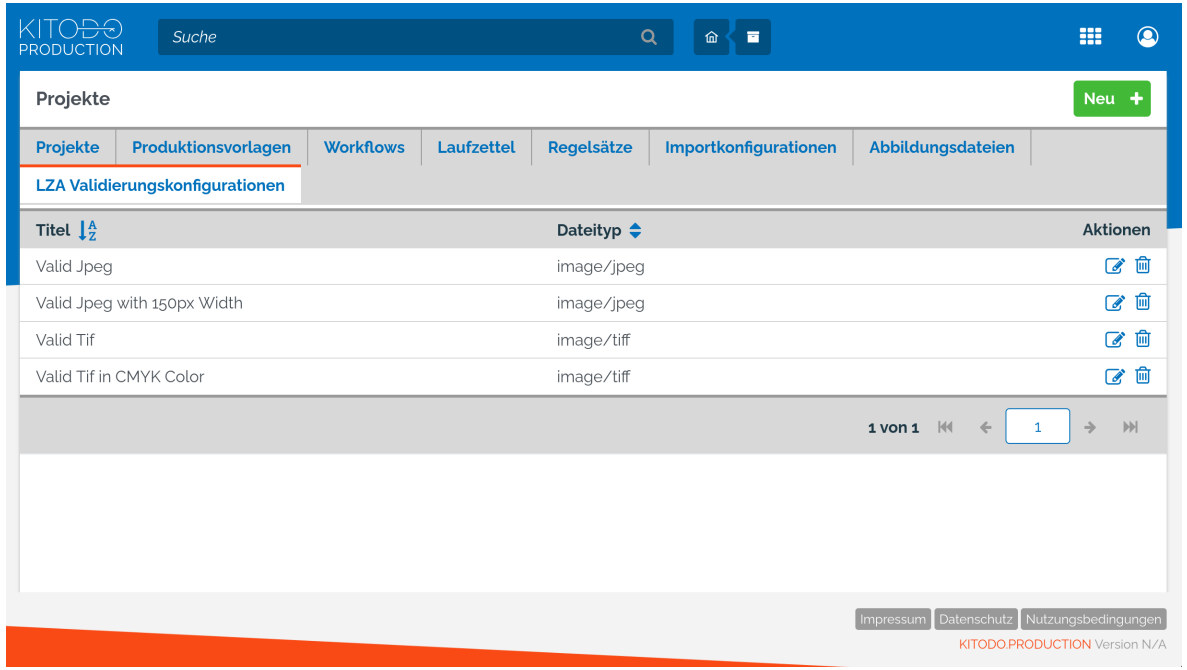
StabiV

Teil 1: Technische Details

Kitodo.Production: Bildvalidierung

Stiftung Preußischer Kulturbesitz

StabiV orientiert sich an Kitodo.Production ab 3.9 [1]



[1] Anwenderdokumentation "Bild validieren"

Projekte

Neu +

Titel	Dateityp	Aktionen
Valid Jpeg	image/jpeg	
Valid Jpeg with 150px Width	image/jpeg	
Valid Tif	image/tiff	
Valid Tif in CMYK Color	image/tiff	

LZA Validierungskonfiguration bearbeiten (Valid Tif in CMYK Color)

Abbrechen | Speichern

Details | Alle Prüfkriterien

Titel *

Valid Tif in CMYK Color

Dateityp

Tagged Image File Format (image/tiff, *.tif)

Einstellungen zum Workflow

Abschließen einer Aufgabe des Typs "Bilder validieren" bei Validierungsfehlern verhindern

Hochladen einer neuen Datei abbrechen, falls ein Validierungsfehler auftritt

Grundlegende Prüfkriterien

Schweregrad, falls eine Datei nicht wohlgeformt ist

Fehler

Schweregrad, falls eine Datei nicht valide ist

Fehler

Vergleichsmuster für Dateinamen inkl. Dateiendung

Regulärer Ausdruck für Dateinamen

Schweregrad, falls ein Dateiname nicht mit dem Muster übereinstimmt

-- Keine Prüfung --

LZA Validierungskonfiguration bearbeiten (Valid Tif in CMYK Color)

Abbrechen | Speichern

Details | Alle Prüfkriterien

Eigenschaft	Vergleichsoperation	Vergleichswerte	Schweregrad	Aktionen
valid	ist exakt	true	Fehler	
wellFormed	ist exakt	true	Fehler	
colorSpace	ist exakt	CMYK	Warnung	

Neues Prüfkriterium hinzufügen

Aufgabendetails

Aufgabendetails

Titel

Validate

Aktionen

Korrekturmeldun

Bilder validieren

Die Bearbeitung

Die Bearbeitung

tatus

Bearbeitung

Ergebnis der Validierung

Es wurden 49 valide Datei(en) und 7 Datei(en) mit Problemen gefunden:

Verzeichnis: images/(processtitle)_tif

81299/images/Towaensu_1920706593_tif/test.jpg

Prüfkriterien, die nicht zutreffen:

- Warnung: Eigenschaft "wellformed" ist exakt "true" - Wert "false" verletzt das Prüfkriterium
- Warnung: Eigenschaft "valid" ist exakt "true" - Wert "false" verletzt das Prüfkriterium
- Warnung: Eigenschaft "XSamplingFrequency_long" konnte nicht gefunden werden
- Warnung: Eigenschaft "XSamplingFrequency_double" konnte nicht gefunden werden
- Warnung: Eigenschaft "XSamplingFrequency" konnte nicht gefunden werden
- Fehler: Eigenschaft "filename" passt zu ".*.tif" - Wert "test.jpg" verletzt das Prüfkriterium

Zusätzliche Informationen:

Schließen

Kommandozeilen-Tool

- Keine Graphische Benutzeroberfläche
- Verfügbar für Linux und Windows [1]
- Nutzt JHove [2] und ExifTool [3]

```
stabiv.bat ^
  --imageRule="$.jhove.wellformed/equal/true" ^
  --include "**.tif" ^
  /path/to/directory
```

[1] https://github.com/StabiBerlin/StabiV
[2] https://jhove.openpreservation.org/
[3] https://exiftool.org/

Validieren

- Einzelne oder mehrere Bilder validieren
- Einzelne oder mehrere Verzeichnisse valideren
- Mit Unterscheidung zwischen "Warnung" und "Fehler"

Zusätzlich

- Eine Checksummen-Datei pro Ordner erzeugen
- Die Ergebnisse zur weiteren Verarbeitung als JSON bereitstellen
- Einen CSV- oder Excel-Report erzeugen

Ergebnis: Excel-Bericht (vereinfacht)

Stiftung Preußischer Kulturbesitz

Verzeichnis	Datei	Valide	Meldungen	#	FileType	# IFDs	ImageWidth	ImageHeight	...
Vorgang1		ja							
Vorgang1	001.tif	ja			TIFF	1	512	512	
Vorgang1	002.tif	ja			TIFF	1	512	512	
Vorgang1	003.tif	Ja			TIFF	1	512	512	
Vorgang2		nein	Lücke	3					
Vorgang2	001.tif	ja			TIFF	1	256	256	
Vorgang2	002.tif	nein	Zu viele IFDs	2	TIFF	2	256	512	
Vorgang2	004.tif	ja			TIFF	1	256	256	

Farbliches Hervorheben von Problemen

- Warnungen (gelb), Fehlern (rot)

Anzeige aller Fehlermeldungen und Warnungen

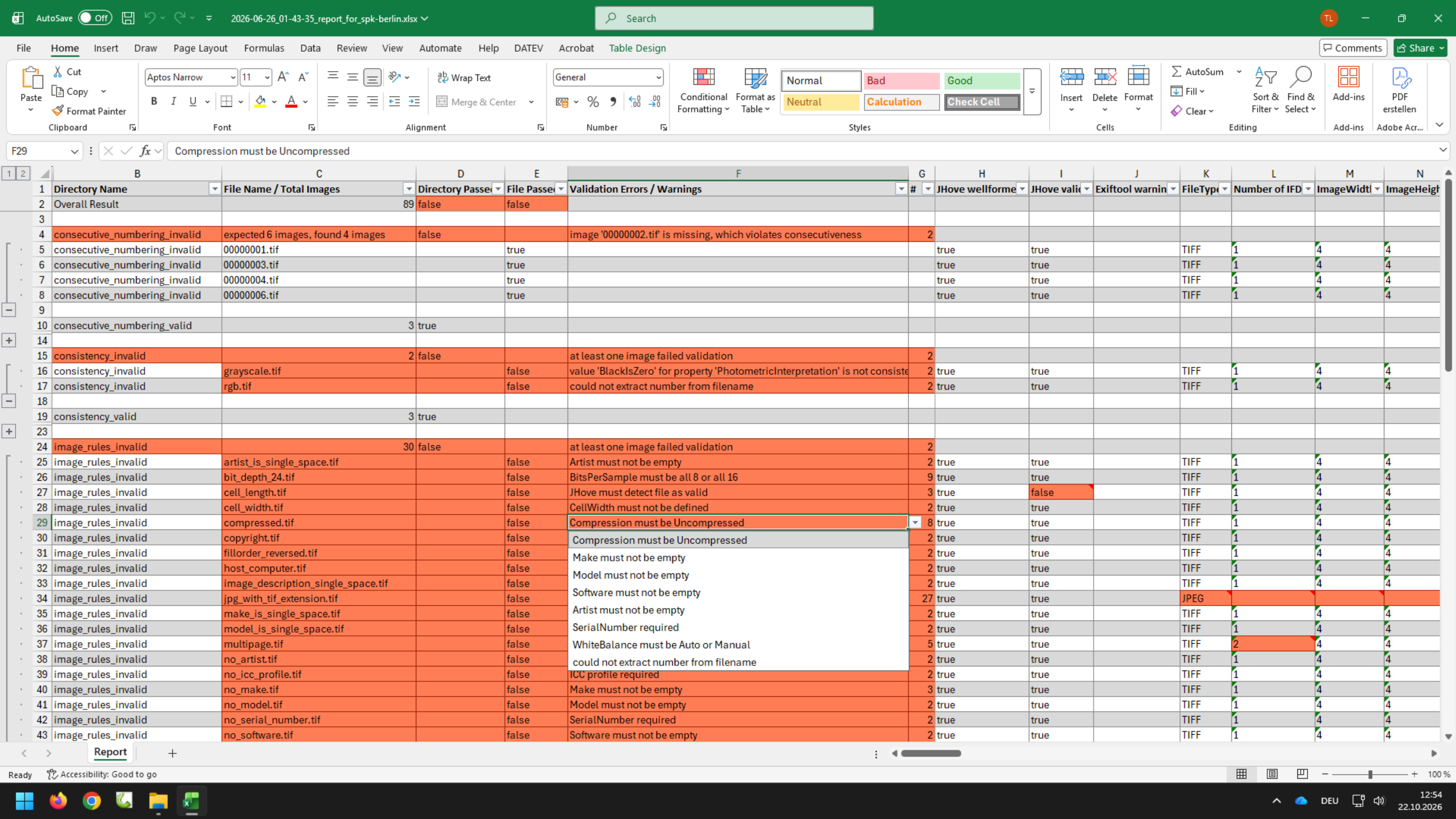
- als Drop-Down-Menu

Anzeige der Fehler für ein Metadatum

- als Tooltip-Kommentare

Interaktives Ein/Ausklappen der Verzeichnisse

- Verzeichnissen ohne Fehler sind standardmäßig zusammengeklappt



	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Directory Name	File Name / Total Images	Directory Passe	File Passe	Validation Errors / Warnings	#	JHove wellforme	JHove vali	Exiftool warnin	FileTyp	Number of IFD	ImageWidt	ImageHeig
2	Overall Result	89	false	false									
3													
4	consecutive_numbering_invalid	expected 6 images, found 4 images	false		image '00000002.tif' is missing, which violates consecutiveness	2							
5	consecutive_numbering_invalid	00000001.tif		true			true	true		TIFF	1	4	4
6	consecutive_numbering_invalid	00000003.tif		true			true	true		TIFF	1	4	4
7	consecutive_numbering_invalid	00000004.tif		true			true	true		TIFF	1	4	4
8	consecutive_numbering_invalid	00000006.tif		true			true	true		TIFF	1	4	4
9													
10	consecutive_numbering_valid	3	true										
14													
15	consistency_invalid	2	false		at least one image failed validation	2							
16	consistency_invalid	grayscale.tif		false	value 'BlackIsZero' for property 'PhotometricInterpretation' is not consiste	2	true	true		TIFF	1	4	4
17	consistency_invalid	rgb.tif		false	could not extract number from filename	2	true	true		TIFF	1	4	4
18													
19	consistency_valid	3	true										
23													
24	image_rules_invalid	30	false		at least one image failed validation	2							
25	image_rules_invalid	artist_is_single_space.tif		false	Artist must not be empty	2	true	true		TIFF	1	4	4
26	image_rules_invalid	bit_depth_24.tif		false	BitsPerSample must be all 8 or all 16	9	true	true		TIFF	1	4	4
27	image_rules_invalid	cell_length.tif		false	JHove must detect file as valid	3	true	false		TIFF	1	4	4
28	image_rules_invalid	cell_width.tif		false	CellWidth must not be defined	2	true	true		TIFF	1	4	4
29	image_rules_invalid	compressed.tif		false	Compression must be Uncompressed	8	true	true		TIFF	1	4	4
30	image_rules_invalid	copyright.tif		false		2	true	true		TIFF	1	4	4
31	image_rules_invalid	fillorder_reversed.tif		false		2	true	true		TIFF	1	4	4
32	image_rules_invalid	host_computer.tif		false		2	true	true		TIFF	1	4	4
33	image_rules_invalid	image_description_single_space.tif		false		2	true	true		TIFF	1	4	4
34	image_rules_invalid	jpg_with_tif_extension.tif		false		27	true	true		JPEG			
35	image_rules_invalid	make_is_single_space.tif		false		2	true	true		TIFF	1	4	4
36	image_rules_invalid	model_is_single_space.tif		false		2	true	true		TIFF	1	4	4
37	image_rules_invalid	multipage.tif		false		5	true	true		TIFF	2	4	4
38	image_rules_invalid	no_artist.tif		false		2	true	true		TIFF	1	4	4
39	image_rules_invalid	no_icc_profile.tif		false	ICC profile required	2	true	true		TIFF	1	4	4
40	image_rules_invalid	no_make.tif		false	Make must not be empty	3	true	true		TIFF	1	4	4
41	image_rules_invalid	no_model.tif		false	Model must not be empty	2	true	true		TIFF	1	4	4
42	image_rules_invalid	no_serial_number.tif		false	SerialNumber required	2	true	true		TIFF	1	4	4
43	image_rules_invalid	no_software.tif		false	Software must not be empty	2	true	true		TIFF	1	4	4

Stiftung
Preußischer Kulturbesitz

Dateipfade oder Dateipfadmuster

- für Bilder und Verzeichnisse, z.B. `ordner/*.{tif,jpg}`

Prüfregeln

- für Bilder bzw. extrahierte Metadaten
- für Verzeichnisse

Sonstige Einstellungen

- Zielverzeichnis für Excel-Bericht
- Spalten für Excel-Bericht

Einzelne Eigenschaften

- `ImageWidth` : "512"
- `ImageLength` : "512"
- `SamplesPerPixel` : "3"
- `ColorSpace` : "RGB"

Untereigenschaften innerhalb der Hierarchie (von JHove und ExifTool)

- `ICC_Profile`
  `ProfileDescription` : "GIMP built-in sRGB"
  `ProfileCopyright` : "Public Domain"

Auswahl in der Hierarchie (mit JSONPath [1])

Stiftung Preußischer Kulturbesitz

Bildbreite des ersten Bildes

- `$.exiftool.TIFF.IFD[0].ImageWidth`

Eine Seriennummer in der Hierarchie (egal wo genau)

- `$..SerialNumber`

Explizit die EXIF-Seriennummer

- `$..ExifIFD.SerialNumber`

[1] https://en.wikipedia.org/wiki/JSONPath

Vergleiche (z.B. =, !=, <, <=, >, >=)

- `ImageWidth` ist größer als 0
- `XResolution` enthält den gleichen Wert wie `YResolution`

Operatoren (z.B. ist leer, existiert, Anzahl, ist eine Zahl, passt auf regulären Ausdruck)

- `Artist` ist nicht leer
- Anzahl der TIF `IFD`s ist genau 1

Logische Verknüpfungen (z.B. Und / Oder / Nicht)

- `PhotometricInterpretation` ist "BlackIsZero" oder "RGB"
- Entweder existiert `SubfileType` nicht, oder es ist "Full-resolution image"

Konsistenz

- `PhotometricInterpretation` ist für alle Bilder gleich

Maximale Abweichung

- `ImageWidth` weicht maximal um 10 Pixel vom Durchschnitt ab
- Ausnahme: abgesehen von Bild 1

Lückenlose Durchnummerierung

- Fortlaufende Bilddateien `000001.tif`, `000002.tif`, usw.

```
…
# Bildbreite
- message: ImageWidth must be positive integer
  severity: error
  condition:
    and:
      - property: $.TIFF.IFD[0].ImageWidth
        assert: isInteger
      - property: $.TIFF.IFD[0].ImageWidth
        assert: greaterThan
        value: 0
…
```

Verbesserungen

- Extraktion der Metadaten mit ExifTool
  - ICC-Profil
  - EXIF-Metadaten
  - XMP-Metadaten
- Hierarchische Auswahl der Metadaten
  - z.B. Unterstützung für mehrseitige TIF-Dateien
- Logische Verknüpfungen der Prüfregeln
- Verzeichnis-basierte Prüfregeln
  - Konsistenz, Maximale Abweichung
  - fortlaufende Nummerierung

Es gibt keine größeren Hürden

- Java / Kitodo Coding Standards
- JHove-Verarbeitung auf Basis von Kitodo.Production

Was fehlt

- Erweitern des Datenmodells für verbesserte Prüfregeln
- Erweitern der Benutzeroberfläche zur Eingabe der Prüfregeln

Showtime!

Danke

Stiftung
Preußischer Kulturbesitz

Ausprobieren

- GitHub: https://github.com/StabiBerlin/StabiV
- Dokumentation:
  https://github.com/StabiBerlin/StabiV/blob/main/documentation/userguide/

Fragen

- Sachlich: Achim Hatzius
  Achim.Hatzius@sbb.spk-berlin.de
- Technisch: Thomas Low
  thomas.low@knopflogik.de