SACHSEN-ANHALT

Landesarchiv

Prüfung signifikanter Eigenschaften mittels Open-Source-Anwendungen

Björn Steffenhagen

Landesarchiv Sachsen-Anhalt, Abt. 1 Zentrale Dienste

26. Tagung des Arbeitskreises „Archivierung von Unterlagen aus digitalen Systemen“ (AK AUdS)

21.03.-22.03.2023 in Mannheim

Hintergrund

SACHSEN-ANHALT

Landesarchiv

- Konversion ist die am häufigsten angewandte Bestandserhaltungsmethodik
- Entwicklung von Massenkonvertern und anderen Tools zur dig. Bestandserhaltung: Archivematica, Preservica, startext SORI, Bestandserhaltungsmodul des DIMAG-Verbunds uvm.
- hauseigene Auseinandersetzungen mit sign. Eigenschaften

Problematisierung/Erkenntnisinteresse

SACHSEN-ANHALT

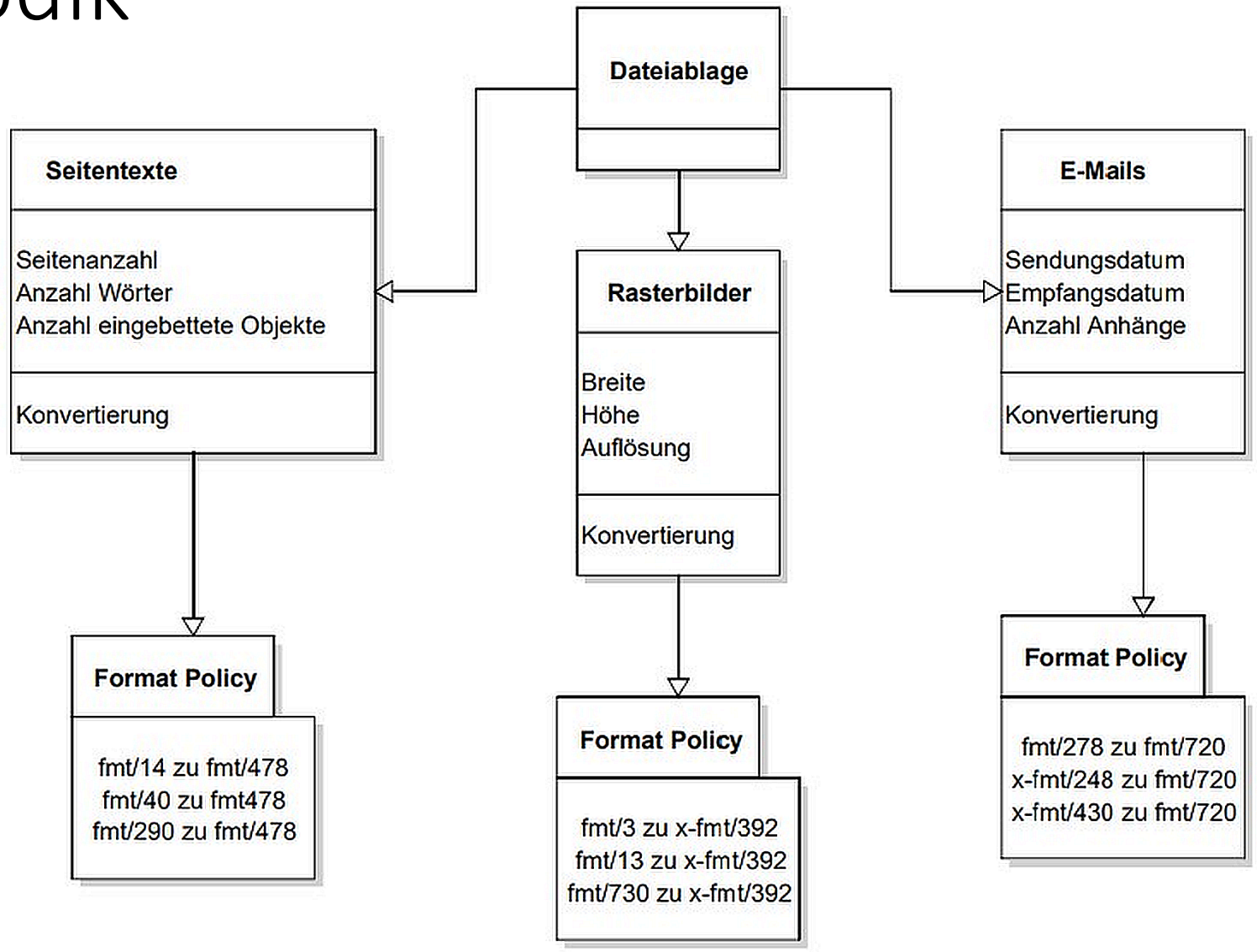
Landesarchiv

1. Lassen sich signifikante Eigenschaften (zumindest stichprobenartig) aus den technischen Metadaten von Dateien erheben?
2. Lassen sich signifikante Eigenschaften durch Open-Source-Lösungen erheben?
3. Geben Wrapper-Lösungen die gleichen Ergebnisse aus, wie Stand-Alone-Varianten?
4. Lassen sich signifikante Eigenschaften nach einer Konversion in ein langzeitstabiles Dateiformat mit den gleichen Tools bzw. der gleichen Methodik erheben und vergleichen?
5. Ist das Vorgehen für kleinere und mittlere Archivverwaltungen umsetzbar?
6. Für größere Archivverwaltungen: Lässt sich die Methodik automatisieren?

Methodik

SACHSEN-ANHALT

Landesarchiv



Quelle: Steffenhagen, Björn: Die Modellierung des Kontexts. Ein objektorientierter Ansatz zur automatisierten Verarbeitung digitaler Archivalien mittels signifikanter Eigenschaften, in: ABI Technik 43 (2023) 1, S. 46-54, hier S. 53. URL: https://doi.org/10.1515/abitech-2023-0006.

Methodik

SACHSEN-ANHALT

Landesarchiv

1. Informationstypen (jeweils 50 Dateien):
    - Texte (.doc, .docx, .pdf)
    - Rastergrafiken (.jpg, .gif, .png, .tiff)
    - E-Mails (.msg, .eml)
2. Metadaten überprüfen
3. Konversion in langzeitstabile Dateiformate
    - Texte (PDF/A-2b)
    - Rastergrafiken (TIFF V6)
    - E-Mails (MBOX)
4. Metadaten überprüfen

Methodik – genutzte Tools

SACHSEN-ANHALT

Landesarchiv

- DROID
- IngestList
- FITS
- Apache Tika
- PDF-Analyzer
- JHOVE
- MediaInfo

- diff-pdf
- KOST-Simy

- PDF 24 Konverter
- XNView MP
- Thunderbird + AddOn ImportExportTools NG

Methodik – Kategorisierung sign. Eigenschaften

SACHSEN-ANHALT

Landesarchiv

Nach InSPECT:

- Erscheinung (Appearance)
- Verhalten (Behavior)
- Kontext (Context)
- Inhalte (Content)
- Struktur (Structure)

Signifikante Eigenschaften:

Orientierung an van Veenendaal et. al.: Significant Significant Properties auf der iPres 2018 in Boston.

Quellen: van Veenendaal, Remco et. al.: Significant Significant Properties, in: iPRES 2018. URL: https://openpreservation.org/wp-content/uploads/2018/10/Significant-Significant-Properties.pdf. Siehe weiterführend auch das Poster der Präsentation: https://osf.io/mc2ea ; Grace, Stephen; Knight, Gareth; Montague, Lynne: InSPECT Final Report, London 2009. URL: https://www.webarchive.org.uk/wayback/archive/20130423072330/http://www.significantproperties.org.uk/methodology.html

Ergebnisse: Gesamtübersicht

SACHSEN-ANHALT

Landesarchiv

	Tika	ExifTool	FITS	JHOVE	MediaInfo	IngestList	KOST-Simy	PDF-Analyzer	PDF-Diff	Image Magick
Texte										
Rasterbilder										
E-Mails										

Legende:
Grün = produktiv einsetzbar
Gelb = eingeschränkt prod. einsetzbar
Rot = nicht produktiv einsetzbar
Blau = Verwendung für Informationstyp nicht möglich

Texte	Tika	ExifTool	FITS	PDF-Analyzer	JHOVE	MediaInfo	IngestList
Erscheinung							
Name Schriftarten							
Anzahl Schriftarten				Auswahl			
Anzahl Annotationen			Nur Angabe, ob ja oder nein				
Anzahl Ebenen							

Texte	Tika	ExifTool	FITS	PDF-Analyzer	JHOVE	MediaInfo	IngestList
Struktur							
Anzahl Seiten							
Anzahl Wörter	Bei .doc-Dateien	Bei .doc-Dateien					
Anzahl Tabellen		Bei .doc.-Dateien					
Anzahl eingebettete Grafiken							
Anzahl eingebettete Objekte							
Anzahl Zeilen		Bei .doc-Dateien					
Anzahl Seitenumbrüche		Bei .doc-Dateien					
Seitenformat							

Texte	Tika	ExifTool	FITS	PDF-Analyzer	JHOVE	MediaInfo	IngestList
Inhalte							
Dateiname							
Titel							
Autor							
zuletzt geändert von		Bei .doc-Dateien, mit Historie					
erstellt am	mit Historie	mit Historie					
zuletzt geändert am							ohne Historie
Firma		Bei .doc-Dateien					
erzeugende Anwendung							
Beschreibung							

Bilddateien	Tika	ExifTool	FITS	JHOVE	MediaInfo	IngestList
Erscheinung						
Auflösung						
Bildhöhe						
Bildbreite						
Farbwiedergabe (samples per pixel)						
Farbwiedergabe (bits per sample)	nur ein Kanal genannt	nur ein Kanal genannt			nur ein Kanal genannt	
Inhalte						
Dateiname						
erstellt am						
geändert am						ohne Historie

E-Mail	Tika	ExifTool	FITS	JHOVE	MediaInfo	IngestList
Erscheinung						
Anzahl Anhänge	nur Angabe, ob Anhänge vorhanden sind					
Struktur						
Anzahl Wörter		nur .eml				
Anzahl Zeilen		nur .eml	nur .eml			
Inhalte						
Dateiname						
zuletzt geändert von						
erstellt am						
geändert am						ohne Historie
erzeugende Anwendung						

E-Mail	Tika	ExifTool	FITS	JHOVE	MediaInfo	IngestList
Kontext						
Angabe Sender, Empfänger, Ersteller						
Angabe local-part						
Angabe domain-part						
Angabe domain-Adresse						
Betreff						
Datum Versand Empfang						
trace-field						
Angabe Einstufungen						

Ergebnisse nach Konversion

SACHSEN-ANHALT

Landesarchiv

- Texte zu PDF/A
  - Inhaltliche Angaben gingen komplett verloren
  - Vergleich zur Ausgangsdatei mit diff-pdf nicht erfolgreich
- Rasterbilder zu TIFF
  - Homogenes Ergebnis, kaum Informationsverluste
  - Bildvergleich mit KOST-Simy rel. gut, mit 46 pos. Ergebnisse bei optischem Vergleich
- E-Mail zu MBOX
  - Anzahl Zeilen und Wörter vergleichbar
- E-Mail zu PDF/A (stichprobenartig)
  - wie Ergebnis zu PDF/A

Beantwortung der Ausgangsfragen

SACHSEN-ANHALT

Landesarchiv

Lassen sich signifikante Eigenschaften (zumindest stichprobenartig) aus den technischen Metadaten von Dateien erheben?

Ja, zu den genannten sign. Eigenschaften lassen sich teilweise technische Eigenschaften zuordnen.

Lassen sich signifikante Eigenschaften durch Open-Source-Lösungen erheben?

Ja, jedoch in unterschiedlicher Qualität. Eine genaue Beurteilung nach erforderlicher Eigenschaft und Dateiformat ist nötig.

Geben Wrapper-Lösungen die gleichen Ergebnisse aus, wie Stand-Alone-Varianten?

Ja, in den meisten Fällen geben Wrapper-Lösungen das gleiche Ergebnis aus, wie ihre Stand-Alone-Varianten. Weitere Anpassungen sind durch technische Eingriffe möglich.

Beantwortung der Ausgangsfragen

SACHSEN-ANHALT

Landesarchiv

Lassen sich signifikante Eigenschaften nach einer Konversion in ein langzeitstabiles Dateiformat mit den gleichen Tools bzw. der gleichen Methodik erheben und vergleichen?

Ja, ein Vergleich ist möglich. Die Konversionsergebnisse sind massiv abhängig vom Konverter. Entsprechend umfangreiche Studien sind ein Desiderat.

Ist das Vorgehen für kleinere und mittlere Archivverwaltungen umsetzbar?

Ja, das Vorgehen ist auch für kleinere Einrichtungen umsetzbar, die nicht über umfangreiche IT-Kenntnisse verfügen, jedoch eine Herausforderung. Eine zentrale Plattform zum Nachweis fehlt bisher.

Für größere Archivverwaltungen: Lässt sich die Methodik automatisieren?

Prinzipiell ja. Die technische Machbarkeit (Lizenzsituation, Implementierbarkeit, Wartbarkeit) ist im Einzelfall zu klären.

Schlussfolgerungen

SACHSEN-ANHALT

Landesarchiv

- unter Voraussetzungen möglich
- unterschiedliche Qualität der Metadatenextraktion

->Prüfungskaskade?

- technische und organisatorische Herausforderung bei einer automatisierten Implementierung

->ggf. Einsatz mehrerer Tools notwendig?

->Alternative Eigenentwicklung?

Schlussfolgerungen

SACHSEN-ANHALT

Landesarchiv

- Zu beachten ist:
  - Die so erhobenen sign. Eigenschaften sind lediglich Indikatoren
  - Bei Änderung der Performance ändern sich auch die erhebbaren Eigenschaften
  - Der technische und der fachliche Kontext beeinflussen die geltenden sign. Eigenschaften maßgeblich.

-> Der vorgestellte Ansatz ist als Lösung hilfreich, jedoch nicht ausreichend.

->Eine stärkere Einbettung der sign. Eigenschaften in die Performance erscheint notwendig.

Kontakt

SACHSEN-ANHALT

Landesarchiv

Björn Steffenhagen, M. A.

https://orcid.org/0000-0002-5849-7123

Abteilung 1 - Zentrale Dienste

Landesarchiv Sachsen-Anhalt

Brückstraße 2

39114 Magdeburg

E-Mail: bjoern.steffenhagen@la.sachsen-anhalt.de

Quellen

SACHSEN-ANHALT

Landesarchiv

Grace, Stephen; Knight, Gareth; Montague, Lynne: InSPECT Final Report, London 2009. URL: https://www.webarchive.org.uk/wayback/archive/20130423072330/http://www.significantproperties.org.uk/methodology.html

Steffenhagen, Björn: Die Modellierung des Kontexts. Ein objektorientierter Ansatz zur automatisierten Verarbeitung digitaler Archivalien mittels signifikanter Eigenschaften, in: ABI-Technik 43 (2023) 1, S. 46-54. URL: https://doi.org/10.1515/abitech-2023-0006.

van Veenendaal, Remco et. al.: Significant Significant Properties, in: iPRES 2018. URL: https://openpreservation.org/wp-content/uploads/2018/10/Significant-Significant-Properties.pdf.

IT-Anwendungen

SACHSEN-ANHALT

Landesarchiv

Apache Tika 2.6.0: https://tika.apache.org/

ImageMagick 7.1.1: https://imagemagick.org/script/download.php

FITS 1.5.1: https://projects.iq.harvard.edu/fits/downloads

JHOVE 1.26: https://jhove.openpreservation.org/

IngestList 6.6.14: https://dimag-wiki.la-bw.de/xwiki/bin/view/%C3%96ffentliche%20Software%20und%20Informationen/IngestList

MediaInfo 22.12: https://mediaarea.net/de/MediaInfo

KOST-Simy 0.0.7: https://kost-ceco.ch/cms/kost_simy_de.html

DROID 6.5.2: https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/

PDF-Analyzer 5.0: https://www.is-soft.de/index.html

diff-pdf 0.5: https://vslavik.github.io/diff-pdf/

PDF 24 Konverter 11.10.2: https://tools.pdf24.org/de/

XNView MP 1.4.3: https://www.xnview.com/de/xnviewmp/#downloads

Thunderbird + AddOn ImportExport Tools NG 102.9: https://www.thunderbird.net/de/