SÄCHSISCHES
STAATSARCHIV

Freistaat
SACHSEN

Webarchivierung per ISO-Image

Stephanie Kortyla

19. AK „Archivierung von Unterlagen aus digitalen Systemen“
am 10. und 11. März 2015 in Wien

Gliederung

- Ausgangslage
- Prozessschritte unserer Webarchivierung
- Auswahlkriterien für Formate
  – hier zweckbezogen: Wrapper/ Encapsulation
- ISO-Image
- Fazit
- Ausblick

Ausgangslage

- Sächsisches Archivgesetz (Stand 2014)
- Fachkonzept „Archivierung des Auftritts des Freistaats Sachsen im World Wide Web“ (2009)
- Präsentation zur „el_sta Webarchivierung“ auf AK AUdS 2013
  - ➔ *erste Testläufe: Crawler und PDF/A-Konvertierung*
  - ➔ *heutige Ergebnisse nicht zufriedenstellend*

Ausgangslage

- Bestehendes elektronisches Archiv (el_sta): Systemumgebung und Abläufe

➔ Webarchivierung

  ➔ Integration in System, keine extra „Archivlandschaft“ schaffen (Ausnahme: Crawler)

  ➔ so unkompliziert wie möglich

Prozessschritte einer Webarchivierung

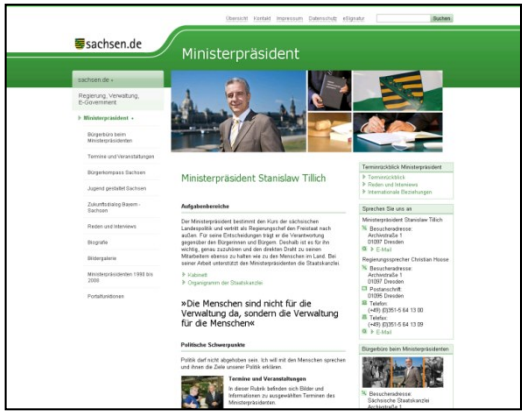
(Bewertung)

1. Erfassung
2. Organisation, Aufbereitung
3. Ingest, Storage
4. Access

➔ ausgerichtet auf signifikante Eigenschaften

➔ Probleme?

1. Erfassung



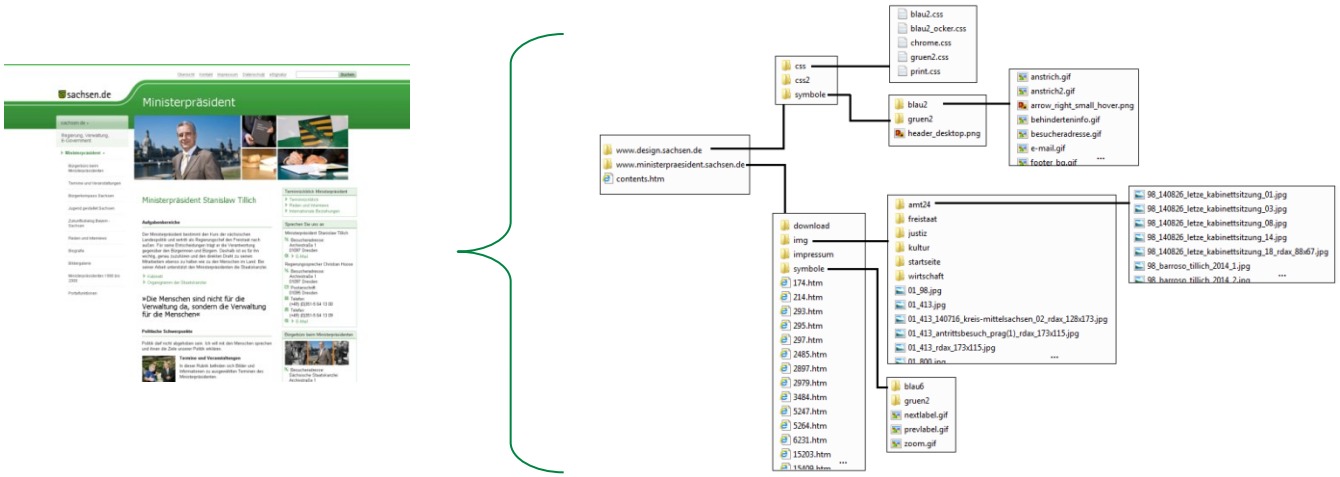
- Crawler (client-side)

➔ Look + Feel der Homepage, wie sie der Nutzer sieht
(Snapshot/ statisches„finales HTML“ mit weiteren Objekttypen → Filter)

➔ Kein Erhalt serverseitiger Funktionen (z.B. Suchfunktionen),
ggf. späteres Aufsetzen von Indexierprogramm & Suchmaschine

2. Organisation und Aufbereitung

- Spiegelung von (Teil-)Domänen zur lokalen Ablage
- Erhalt der ursprünglichen Webseitenstruktur,
  → Verzeichnisse und Objekte, Referenzierungen/ Verlinkungen
  (absolute → relative Links)



2. Organisation und Aufbereitung

1. Navigation
2. Inhalt
3. Herausgeber

- Übersicht
- Kontakt
- Impressum
- Datenschutz
- eSignatur

Suchen

sachsen.de

Ministerpräsident

sachsen.de

Ministerpräsident

Navigation

- sachsen.de

Regierung, Verwaltung, E-Government

- Ministerpräsident -
  - Bürgerbüro beim Ministerpräsidenten
  - Termine und Veranstaltungen
  - Bürgerkommission Sachsen
  - Dialog gestaltet Sachsen
  - Eidesfreiheitsdienst Sachsen
  - Reden und Interviews
  - Biografie
  - Bildergalerie
  - Ministerpräsidenten 1990 bis 2008
  - Portalfunktionen

Inhalt

Ministerpräsident Stanislaw Tillich

Aufgabenbereiche

Der Ministerpräsident bestimmt den Kurs der sächsischen Landespolitik und vertritt als Regierungschef den Freistaat nach außen. Für seine Ente unterstützt den Ministerpräsidenten die Staatskanzlei.

- Kabinett
- Organisation der Staatskanzlei

»Die Menschen sind nicht für die Verwaltung da, sondern die Verwaltung für die Menschen«

Politische Schwerpunkte

Politik darf nicht abgehoben sein. Ich will mit den Menschen sprechen und ihnen die Ziele unserer Politik erklären.

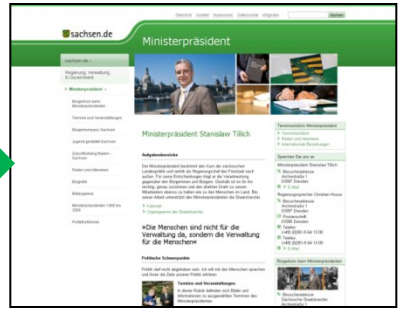
Termine und Veranstaltungen

In dieser Rubrik befinden sich Bilder und Informationen zu ausgewählten Terminen des Ministerpräsidenten.

- Termine und Veranstaltungen

Crawlergebnis

2. Organisation und Aufbereitung

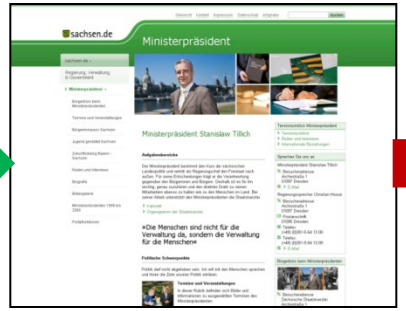


Crawlergebnis

aufbereiteter Crawl

2. Organisation und Aufbereitung

Crawlergebnis



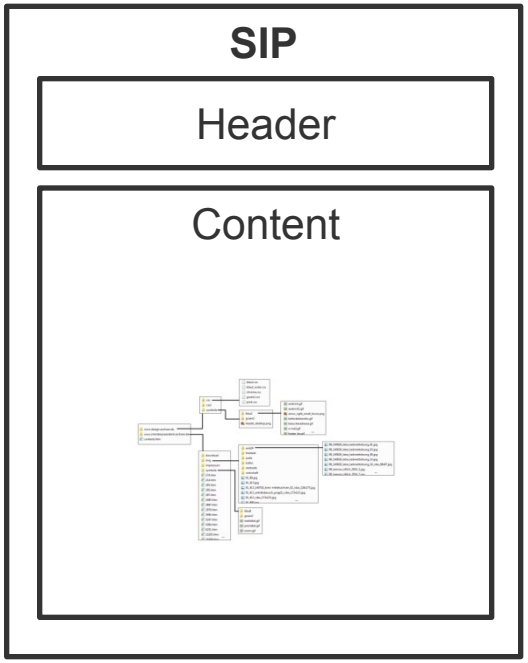
aufbereiteter Crawl

nach PDF/A konvertierter Crawl

3. Ingest und Storage

- Große Objektanzahl → Systemperformance
  Bsp. „kleiner“ Crawl: 1100 Dateien in 23 Verzeichnissen (~75MB)

- Hierarchische Struktur bei Ingest
  nicht verarbeitbar

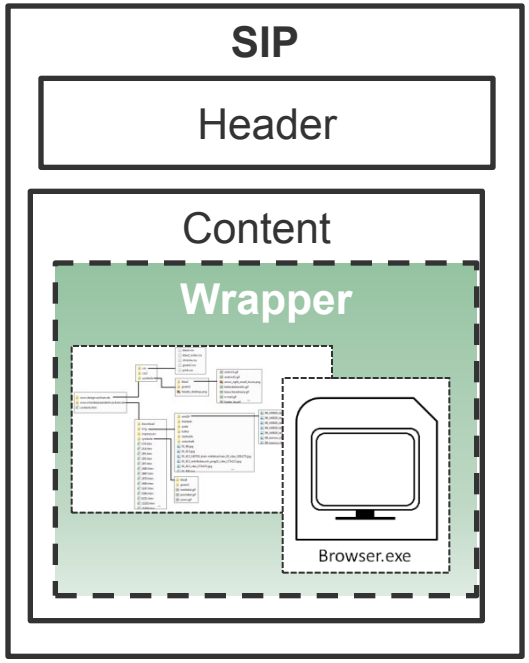


➔ Flache Struktur (z.B. PDF/A) fällt weg,
stattdessen Wrapper für Content?

3. Ingest und Storage

- Große Objektanzahl → Systemperformance
  Bsp. „kleiner“ Crawl: 1100 Dateien in 23 Verzeichnissen (~75MB)

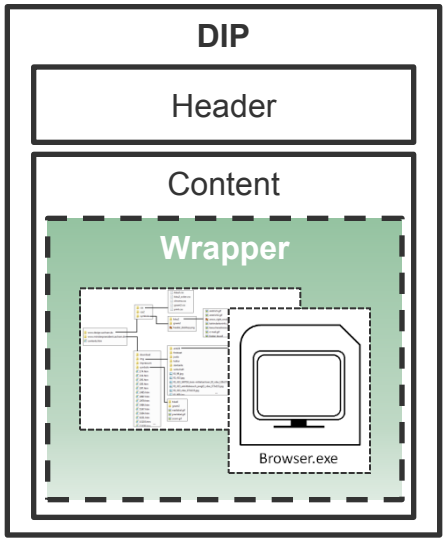
- Hierarchische Struktur bei Ingest
  nicht verarbeitbar



→ Flache Struktur (z.B. PDF/A) fällt weg,
stattdessen Wrapper für Content?

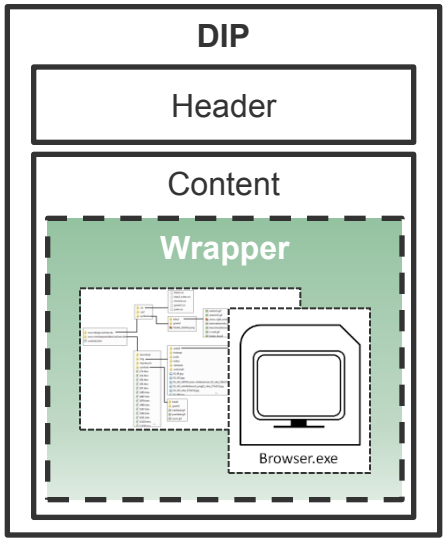
4. Access

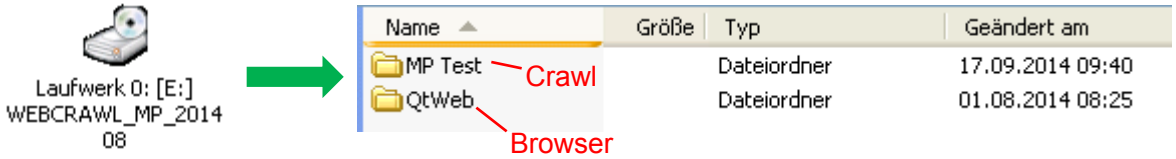
- Vor-Ort-Nutzung, kein Onlinearchiv
- Bereitstellung eines zeitgemäßen Browsers
  → gewohnte Nutzungsumgebung



4. Access

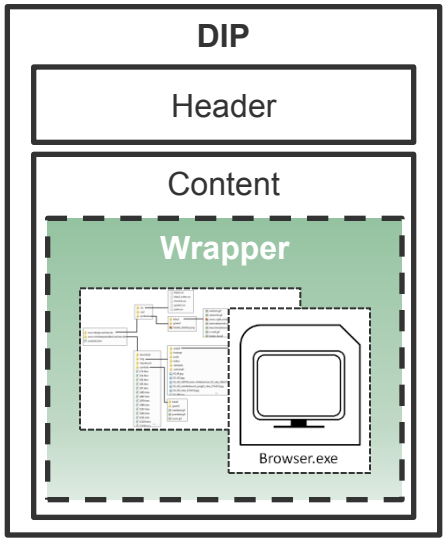
- Vor-Ort-Nutzung, kein Onlinearchiv
- Bereitstellung eines zeitgemäßen Browsers
  → gewohnte Nutzungsumgebung
- Zugriff über **virtuelles Laufwerk** („mounten“), Entpackersoftware oder
  mittels optischem Datenträger (z.B. CD-Rom)

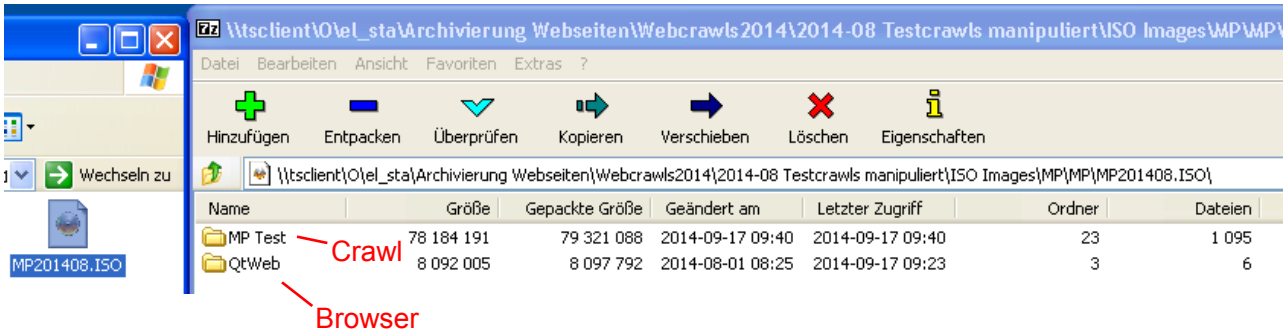




4. Access

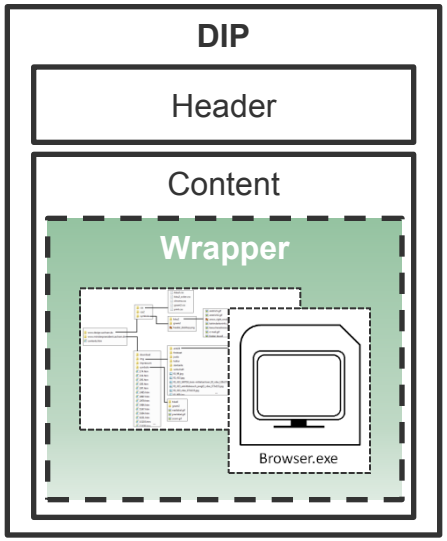
- Vor-Ort-Nutzung, kein Onlinearchiv
- Bereitstellung eines zeitgemäßen Browsers
  → gewohnte Nutzungsumgebung
- Zugriff über virtuelles Laufwerk („mounten“),
  Entpackersoftware oder
  mittels optischem Datenträger (z.B. CD-Rom)

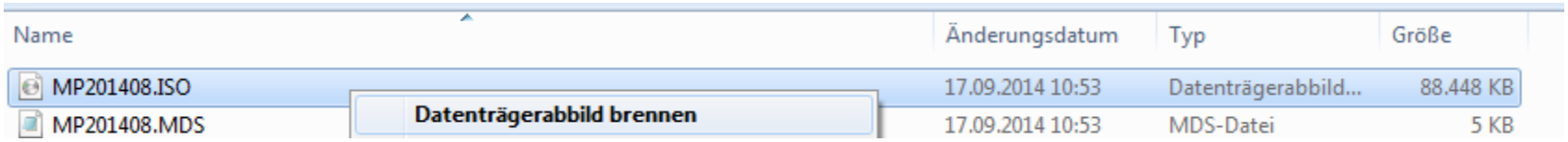




4. Access

- Vor-Ort-Nutzung, kein Onlinearchiv
- Bereitstellung eines zeitgemäßen Browsers
  → gewohnte Nutzungsumgebung
- Zugriff über virtuelles Laufwerk („mounten“), Entpackersoftware oder mittels **optischem Datenträger** (z.B. CD-Rom)





Auswahlkriterien für Formate

- Hier zweckbezogen: **Encapsulation**/ „einfacher Wrapper“

„…in the context of preserving digital materials is a technique of grouping together a digital object and anything else necessary to provide access to that object.“

Quelle: NLA, PADI 2011, http://pandora.nla.gov.au/pan/10691/20110824-1153/www.nla.gov.au/padi/topics/20.html

Auswahlkriterien für Formate

- Hier zweckbezogen: **Encapsulation**/ „einfacher Wrapper“

  „…in the context of preserving digital materials is a technique of grouping together a digital object and anything else necessary to provide access to that object.“

  Quelle: NLA, PADI 2011, http://pandora.nla.gov.au/pan/10691/20110824-1153/www.nla.gov.au/padi/topics/20.html

➔ Strukturerhalt, evtl. Hinzufügen weiterer Komponenten (Browser)

➔ Einfacher Access

Auswahlkriterien für Formate

- Hier zweckbezogen: **Encapsulation**/ „einfacher Wrapper“

  „…in the context of preserving digital materials is a technique of grouping together a digital object and anything else necessary to provide access to that object.“

  Quelle: NLA, PADI 2011, http://pandora.nla.gov.au/pan/10691/20110824-1153/www.nla.gov.au/padi/topics/20.html

➔ Strukturerhalt, evtl. Hinzufügen weiterer Komponenten (Browser)

➔ Einfacher Access

➔ Analyse verschiedener Containerformate

  - **z.B. 7z, AXF, MHT, SIRF, WARC**

Auswahlkriterien für Formate

- Hier zweckbezogen: **Encapsulation**/ „einfacher Wrapper“

  „…in the context of preserving digital materials is a technique of grouping together a digital object and anything else necessary to provide access to that object.“

  Quelle: NLA, PADI 2011, http://pandora.nla.gov.au/pan/10691/20110824-1153/www.nla.gov.au/padi/topics/20.html

➔ Strukturerhalt, evtl. Hinzufügen weiterer Komponenten (Browser)

➔ Einfacher Access

➔ Analyse verschiedener Containerformate

  - **z.B. 7z, AXF, MHT, SIRF, WARC**

➔ Bisher als „Spezial“-Container behandelt, jedoch für unsere Zwecke praktikabelste Lösung: **ISO-Image** *(vgl. auch KOST)*

ISO-Image

- ursprünglich = identisches Abbild einer CD oder DVD
- universelles Dateisystem: ISO 9660, plattformunabhängig
- Erweiterungen/ Dateisysteme: Joliet, Rockridge, etc.

ISO-Image

- ursprünglich = identisches Abbild einer CD oder DVD
- universelles Dateisystem: ISO 9660, plattformunabhängig
- Erweiterungen/ Dateisysteme: Joliet, Rockridge, etc.

➔ eine Datei

➔ Erhalt der Datei-/ Verzeichnisstruktur

  ➔ eignet sich auch für Verzeichnisse ➔ Webcrawls

ISO-Image

- Kriterien zur Formatbewertung u.a. nach
  - Kim/ Ross (2012): *Digital Forensics Formats: Seeking a Digital Preservation Storage Format for Web Archiving*
  - Zierau (2012): *Package Formats for Preserved Digital Material*
  - Koordinierungsstelle für die dauerhafte Archivierung elektronischer Unterlagen (KOST)
  - National Digital Information Infrastructure and Preservation Program (NDIIPP), Library of Congress
  - The National Archives, UK
  - National Library of Australia (NLA)
  - InterPARES 2 Project

ISO-Image

- package-/ storage-bezogen:
  - Externe Abhängigkeiten/ Interoperabilität:
    a) Software zum Öffnen oder virtuelles Laufwerk („mounten“)
    b) auf optischen Datenträger brennen (➔ Größenbeschränkung!)
  - Beschränkung hins. Dateiformat und -größe:
    content-neutral
    je nach Tool „custom“-Größe einstellbar, demnach auch > 4,7GB (DVD-ROM)
    Beschränkung hins. Anzahl der Verzeichnisse und Verzeichnistiefe
  - Datenvollständigkeit: exaktes Abbild

ISO-Image

- package-/ storage-bezogen:
  - Externe Abhängigkeiten/ Interoperabilität:
    a) Software zum Öffnen oder virtuelles Laufwerk („mounten“)
    b) auf optischen Datenträger brennen (➔ Größenbeschränkung!)
  - Beschränkung hins. Dateiformat und -größe:
    content-neutral
    je nach Tool „custom“-Größe einstellbar, demnach auch > 4,7GB (DVD-ROM)
    Beschränkung hins. Anzahl der Verzeichnisse und Verzeichnistiefe
  - Datenvollständigkeit: exaktes Abbild

ISO-Image

- preservation-bezogen:
  - Standardisierung: ECMA-119 (1987/1998) bzw. ISO Standard 9660 (1988)
  - Offenlegung (disclosure): viele einzelne, aber keine allumfassende/ allgemeingültige Spezifikation für verschiedene Varianten (Erweiterungen)
  - Transparenz: Öffnen mit Texteditor ➔ lesbarer Inhalt (sofern Objekte lesbar sind, z.B. Textdateien, Metadaten in Bildern), Log-Datei (Dateisysteme), Formatvalidierung erfolgreich, Checksummenerstellung
  - Verbreitung (adoption): frei verfügbare Tools zur Erstellung, zum Öffnen, zum „Mounten“ (Einhängen in virtuelles Laufwerk) bzw. kann jede gängige Brennsoftware ISO-Images verarbeiten
  - Selbstdokumentation: Metadatenaufnahme je nach Erstelltool (vordefiniert, kaum Spielraum) ➔ über beschreibende Metadaten des AIP auf Zugriffsmöglichkeit („Gebrauchsanweisung“) des ISO-Images hinweisen

ISO-Image

- preservation-bezogen:
  - Standardisierung: ECMA-119 (1987/1998) bzw. ISO Standard 9660 (1988)
  - Offenlegung (disclosure): viele einzelne, aber keine allumfassende/ allgemeingültige Spezifikation für verschiedene Varianten (Erweiterungen)
  - Transparenz: Öffnen mit Texteditor ➔ lesbarer Inhalt (sofern Objekte lesbar sind, z.B. Textdateien, Metadaten in Bildern), Log-Datei (Dateisysteme), Formatvalidierung erfolgreich, Checksummenerstellung
  - Verbreitung (adoption): frei verfügbare Tools zur Erstellung, zum Öffnen, zum „Mounten“ (Einhängen in virtuelles Laufwerk) bzw. kann jede gängige Brennsoftware ISO-Images verarbeiten
  - Selbstdokumentation: Metadatenaufnahme je nach Erstelltool (vordefiniert, kaum Spielraum) ➔ über beschreibende Metadaten des AIP auf Zugriffsmöglichkeit („Gebrauchsanweisung“) des ISO-Images hinweisen

Fazit

- Crawlstruktur durch Wrapper „eingefroren“ und Verarbeitbarkeit/ Ingest ermöglicht
- Access auf verschiedenen Wegen möglich
- Preservation Planning:
  - ➔ Ingestierte Formate lassen Weiterverarbeitung/ Migration zu, z.B. PDF/A

Fazit

- Crawlstruktur durch Wrapper „eingefroren“ und Verarbeitbarkeit/ Ingest ermöglicht
- Access auf verschiedenen Wegen möglich
- Preservation Planning:
  - ➔ Ingestierte Formate lassen Weiterverarbeitung/ Migration zu, z.B. PDF/A

➔ **Verfahren**

- ➔ **mit einfachen Mitteln, kostenfrei möglich, schnell umsetzbar**

Ausblick

- Bereitstellung eines Betriebssystems (Virtuelle Maschinen, Emulation)
- Aufgesetzte Suchmaschine

Vielen Dank!

stephanie.kortyla@sta.smi.sachsen.de