SLUB
Wir führen Wissen.

Die Eignung des Dateiformates PDF für die Langzeitarchivierung – Probleme und Lösungen

Dr. Henrike Berthold, Sabine Krug, Christian Beissert,
Michaela Voigt, Anke Wartenberg
Sächsische Landesbibliothek –
Staats- und Universitätsbibliothek Dresden (SLUB)

Juni 2014

SLUB
Wir führen Wissen.

Europäischer Fonds für regionale Entwicklung – EFRE Konzeption und Realisierung einer effizienten Langzeitarchivierung

Europäische Union

Europa fördert Sachsen.
EFRE
Europäischer Fonds für regionale Entwicklung

SLUB
Wir führen Wissen.

Einführung

Definition. Archivfähigkeit. Stand an der SLUB.

Digitale Langzeitarchivierung
Definition und Ziele

Definition eines digitalen Langzeitarchivs (nach DIN 31644)

„**Organisation** (bestehend aus Personen und technischen Systemen), die die Verantwortung für den Langzeiterhalt und die Langzeitverfügbarkeit von Information in digitaler Form sowie die Bereitstellung für eine bestimmte **Zielgruppe** übernommen hat.“

Ziele

- Sicherung der langfristigen Verfügbarkeit und Nutzung von digitalen Objekten (50 Jahre +)
- Berücksichtigung zukünftiger Nutzungsszenarien

=> Erhalt der Korrektheit (Bitstream Preservation) und

=> Erhalt der Interpretierbarkeit und Nutzbarkeit (Content Preservation)

© by SLUB Dresden

Digitale Langzeitarchivierung
Archivfähigkeit eines Datenformates

Archivfähigkeit = Kriterien für die Festlegung von akzeptierten Datenformaten

- Offene Spezifikation; kein proprietäres Format
- Große Verbreitung
- Geringe Komplexität
- Keine Schutzmechanismen wie Kopierschutz, Verschlüsslung, DRM
- Selbstdokumentation
- Robustheit
- Geringe Abhängigkeiten von spezieller Hardware, Software oder anderen Ressourcen

Quelle: Kapitel 7.3 des nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, Hrsg: Heike Neuroth, Achim Oßwald, Regine Scheffel, Stefan Strathmann, Karsten Huth, Version 2.3, 2010, http://nestor.sub.uni-goettingen.de/handbuch/

© by SLUB Dresden

Digitales Langzeitarchiv SLUB
Ziele und Stand

Aufbau des Digitalen Langzeitarchivs der SLUB erfolgt im Rahmen eines EFRE-geförderten Projektes (Mai 2012 bis Oktober 2014)

Ziele

- Sichern der Langzeitverfügbarkeit der Digitalen Sammlungen der SLUB (Digitalisierung mit Goobi, Elektronische Publikationen, Digitale Sammlung der Deutschen Fotothek, Digitales Audio/Video-Material der Mediathek)
- Vorbereitung einer Dienstleistung für andere sächsische Institutionen

Stand:

- Implementierung und Test von Workflows für die Digitalisierung mit Goobi und Elektronische Publikationen (Qucosa) erfolgreich
- Erweiterter Testbetrieb für Goobi-Workflow; Inbetriebnahme Q3/2014
- Anpassung und Inbetriebnahme des Qucosa-Workflows nach Umstellung des Repositories; Inbetriebnahme Q1/2015

© by SLUB Dresden

Digitales Langzeitarchiv SLUB
Grundsätze

Verwendung als Dark Archive, in dem die Masterdaten verwaltet und archiviert werden – die Präsentationsdaten bleiben in einem separaten Repository, können aber aus den Masterdaten erzeugt werden

Automatisierung des Ingest, d.h. der Übernahme ins Langzeitarchiv, und des Access, d.h. des Zugriffs auf die Daten aus dem Langzeitarchiv (bis auf Fehlerfälle)

Prüfsummen werden bereits im Produktionsprozess (bei der Digitalisierung) bzw. der Annahme (bei Elektronischen Publikationen) erzeugt und bei der Übernahme ins Langzeitarchiv geprüft

Unterstützung einer definierten Menge von LZA-fähigen Datenformaten

Übernahme ins Langzeitarchiv nur für erfolgreich geprüfte Dokumente

© by SLUB Dresden

SLUB
Wir führen Wissen.

Datenformat PDF

Historie. Formatfamilie PDF/A. Bewertung.

Datenformat PDF
Was ist es für ein Format?

Sprache zur Beschreibung von Seiten: Inhalt, Struktur und Metadaten

- Erstellung des Bildes direkt vor dem Darstellen am Bildschirm oder dem Drucken
- Ermöglicht Anpassungen zur Darstellungszeit z.B. Skalierung, unterschiedliche Auflösungen

Plattformneutrales Austauschformat

Einige Vorteile

- Selektiver Zugriff auf Seiten
- Fonts können eingebettet werden, damit wird die korrekte Darstellung gesichert
- Volltextsuche, Copy&Paste (auch bei Bildern ist ein OCR-Text integrierbar)

© by SLUB Dresden

Datenformat PDF
Historie und Produktfamilie

- 1993: Erste Version
- 2000: Version 1.3
- 2001: PDF/X-1a (PDF for Exchange; ISO 15930-1), spezifiziert eine Teilmenge von PDF 1.3
- 2001: Version 1.4
- 2005: PDF/A-1 (PDF for Archive; ISO 19005-1); Teilmenge von PDF 1.4
- 2008: Version 1.7, Erster ISO-Standard ISO 32000-1
- 2010: PDF/A-2; Teilmenge von PDF 1.7; ISO 19005-2
- 2012: PDF/A-3; Teilmenge von PDF 1.7; ISO 19005-3
- Weitere: PDF/E (Engineering), PDF/VT (Variable and transactional printing); PDF/UA-Standard (Universal Accessibility)

© by SLUB Dresden

Datenformat PDF
Produktfamilie PDF/A

PDF-A/1 (Conformance level a – komplett und b – basic)

- Beinhaltet alle für die Darstellung relevanten Daten zu Schriftarten und Farben
- Keine Verschlüsslung
- Keine Referenzierung von externen Daten
- XMP-Metadaten sind zwingend erforderlich

PDF-A/2 (Conformance level a – komplett, u – unicode, b – basic)

- Transparenz erlaubt
- Erweiterte Möglichkeiten der Komprimierung bspw. JPEG2000

PDF-A/3

- Einbettung von zusätzlichen Daten im beliebigen Datenformaten z.B. XML-Daten

© by SLUB Dresden

Datenformat PDF
Exkurs Metadaten

Document Information Dictionary (ab PDF 1.0, erweitert in PDF 1.1)

- Optional; für das gesamte Dokument
- Elemente: Author, CreationDate, Creator (Software, mit der das Dokument erstellt wurde), Producer (Software, mit der das PDF erstellt wurde), Title, Subject, Keywords, ModDate

Metadaten als Datenstrom-Objekt

- Zugeordnet zum Dokument oder Objekten
- Da Datenstromobjekt, meist auch komprimiert; Dokumentmetadaten sollten aber unkomprimiert vorliegen
- Inhalt sind XML-Daten gemäß einem XMP-Schema (Extensible Metadata Platform; ISO 16684-1)
- Mehrere Schemata definiert + Vorgehen zum Definieren eines eigenen Schemas

© by SLUB Dresden

Datenformat PDF

Unsere Bewertung der Archivfähigkeit

Kriterien	PDF 1.7	PDF/A-1	PDF/A-2	PDF/A-3
Offene Spezifikation	++	++	++	++
Große Verbreitung	++	o	o	o
Geringe Komplexität	--	+	+	o
Keine Schutzmechanismen	-	++	++	++
Selbstdokumentation	+ (XMP möglich)	++ (XMP zwingend)	++ (XMP zwingend)	++ (XMP zwingend)
Robustheit	- (Kompr.)	- (Kompr.)	- (Kompr.)	- (Kompr.)
Keine Abhängigkeiten	-- (Referenzen, zusätzl. Dateien)	++	++	- (Zusätzl. Dateien)

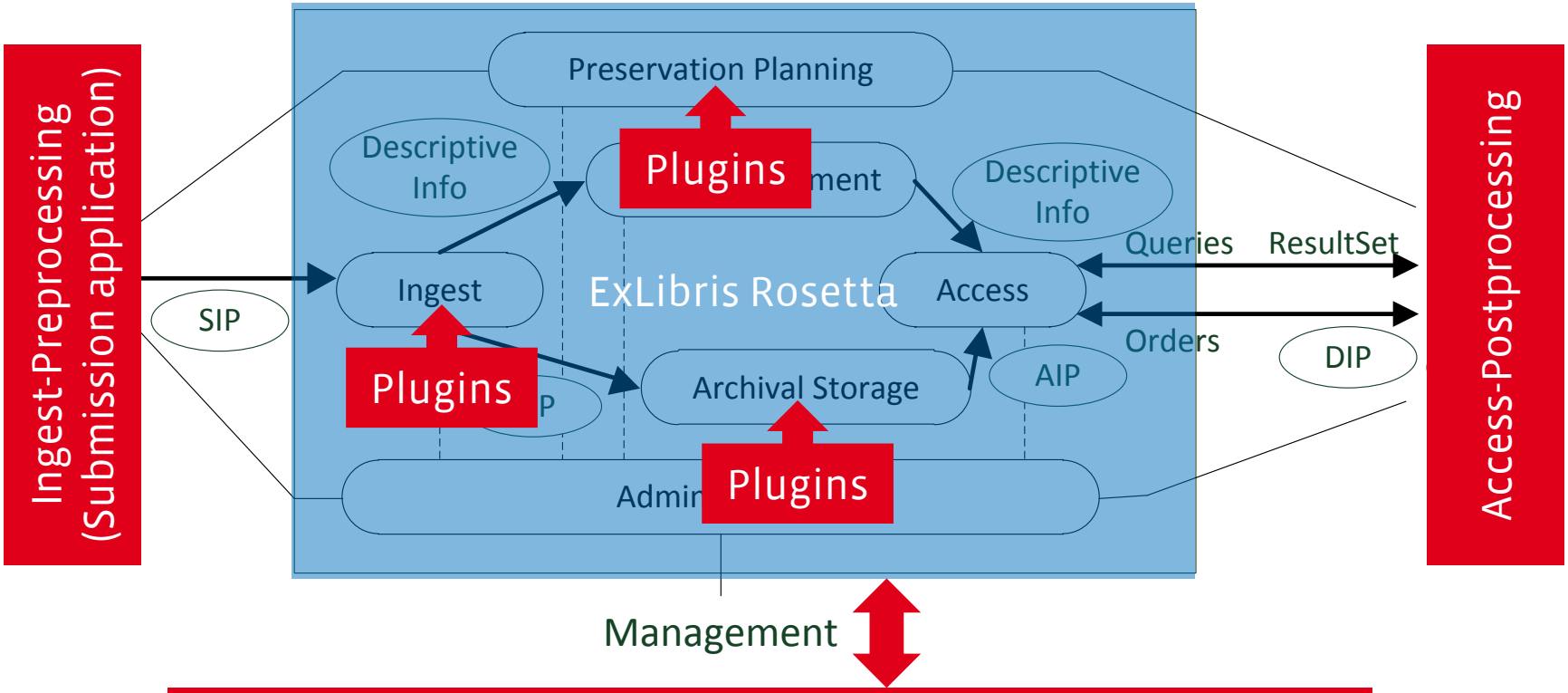
© by SLUB Dresden

SLUB
Wir führen Wissen.

Digitale Langzeitarchivierung von pdf-Dokumenten

Architektur SLUB Langzeitarchiv. Validierung.

Digitale Langzeitarchivierung
Architektur SLUB Langzeitarchiv



IT-Basisinfrastruktur bestehend aus Servern, SAN-Speicher, NAS-Speicher, Bandsystemen

Kooperation mit dem Rechenzentrum der TU Dresden zum Produktivsystem

© by SLUB Dresden

Digitale Langzeitarchivierung
Qucosa Ingest-Preprocessing

Daten

- pdf-Dokument und ergänzende Daten in anderen Formaten
- Metadaten derzeit in spezifischem Format

Ingest Pre-Processing = Automatische Vorbereitung der Datenpakete durch Submission application

- Prüfung der Vollständigkeit der Daten
- Transformation der Metadaten
- Anstoßen der automatischen Übernahme durch die Langzeitarchivsoftware

© by SLUB Dresden

Digitale Langzeitarchivierung
Qucosa Ingest

Automatische Übernahme (muß für jeden Workflow konfiguriert und über Plugins/Programme angepasst werden)

- Prüfen der Vollständigkeit und Integrität der Dateien (Prüfsummen)
- Virusprüfung
- Identifikation des Datenformates (PRONOM-ID)
- Validierung = Prüfen der Korrektheit der Daten (Jhove)
- Extraktion von technischen Metadaten
- Erstellen eines Archivpaketes
- Speicherung im Archivbereich

Manuelle Nachbearbeitung von Fehlern

© by SLUB Dresden

Digitale Langzeitarchivierung
Qucosa Ingest

Automatische Übernahme (muß für jeden Workflow konfiguriert und über Plugins/Programme angepasst werden)

- Prüfen der Vollständigkeit und Integrität der Daten
- Virusprüfung
- Identifikation des Datenformates (PRONOM-ID)
- Validierung = Prüfen der Korrektheit der Daten (Jhove)
- Extraktion von technischen Metadaten
- Erstellen eines Archivpaketes
- Speicherung im Archivbereich

Manuelle Nachbearbeitung von Fehlern

Fehlerrate bei der Jhove-Validierung von den pdf-Dokumenten sehr hoch

© by SLUB Dresden

Digitale Langzeitarchivierung
Fehleranalyse

Komplexität des PDF-Formates sehr hoch -> fehlerhafte Dokumente

- PDF-Struktur nicht korrekt

Validierungstools sind teilweise nur für PDF/A entwickelt => Freiheitsgrade von PDF führen zu Validierungsfehlern, die aber zugleich auf mangelnde LZA-Fähigkeit hinweisen

- Fonts nicht eingebettet
- Fehlende Zuordnung eines Zeichenkodes zu einem Schriftzeichen
- Glyphe (graphische Darstellung) für Schriftzeichen fehlt
- Einsatz von Verschlüsselung
- Einsatz von Formularfeldern

Validierungsergebnisse unterscheiden sich bei verschiedenen Validierungstools

© by SLUB Dresden

SLUB
Wir führen Wissen.

Lösungen

Organisation. Werkzeuge zur Korrektur.

Organisatorische Lösungen
Festlegen von technischen Standards und Umsetzung

Festlegung von Standards auf technischer Ebene

- Formatversionen: PDF/A-1 (ISO 19005-1:2005), PDF/A-2 (ISO 19005-2:2011)
- PRONOM-Formate: fmt/354, fmt/95
- Randbedingungen: (keine)

Erstellung von Hinweisen für die korrekte Erstellung von Dokumenten gemäß dieser Anforderungen

Erarbeitung und Durchführung von Schulungen

© by SLUB Dresden

Organisatorische Lösungen
Service-Level-Konzept

Rückmeldung über die Qualität des abgelieferten Dokumentes an den Abliefernden und die Konsequenzen für die Langzeitarchivierung bspw.

- Garantierte Langzeitarchivierung nur für valides Dokument in akzeptiertem Format
- Archivierung (bis 10 Jahre) für anderes Dokument

© by SLUB Dresden

Technische Lösungen
Automatische Korrektur/Umwandlung

Untersuchung von Korrekturmechanismen: bis zu 90% der Dateien können korrigiert werden, eine Sichtprüfung ist trotzdem erforderlich

Auswahl von inkorrekten Dokumenten (1024 Dokumente) und die Erfolgsrate:

Erfolg der Validierung	Validator		
	3-Heights	Callas	JHOVE
des Originals	0	0	61 %
nach Korrektur mit 3-Heights	98 %	93 %	91 %
nach Korrektur mit Callas	78 %	95 %	66 %
nach Korrektur mit PDFtk	0	0	96 %

© by SLUB Dresden

Technische Lösungen
Automatische Umwandlung/Korrektur

Automatische Korrektur und erfolgreiche Validierung ersetzt keine „Sichtkontrolle“

of each of the grains
direction. Working or
on single crystals [9],

|e>

σ_{tBua}

- Niedrige
- Geringer

$$v = \frac{d_0 \cdot \pi \cdot n \cdot \cos\alpha}{6000}$$

Korrektur

of each of the grains
direction. Working or
on single crystals [9],

|e>

(tBua

→ Niedrige

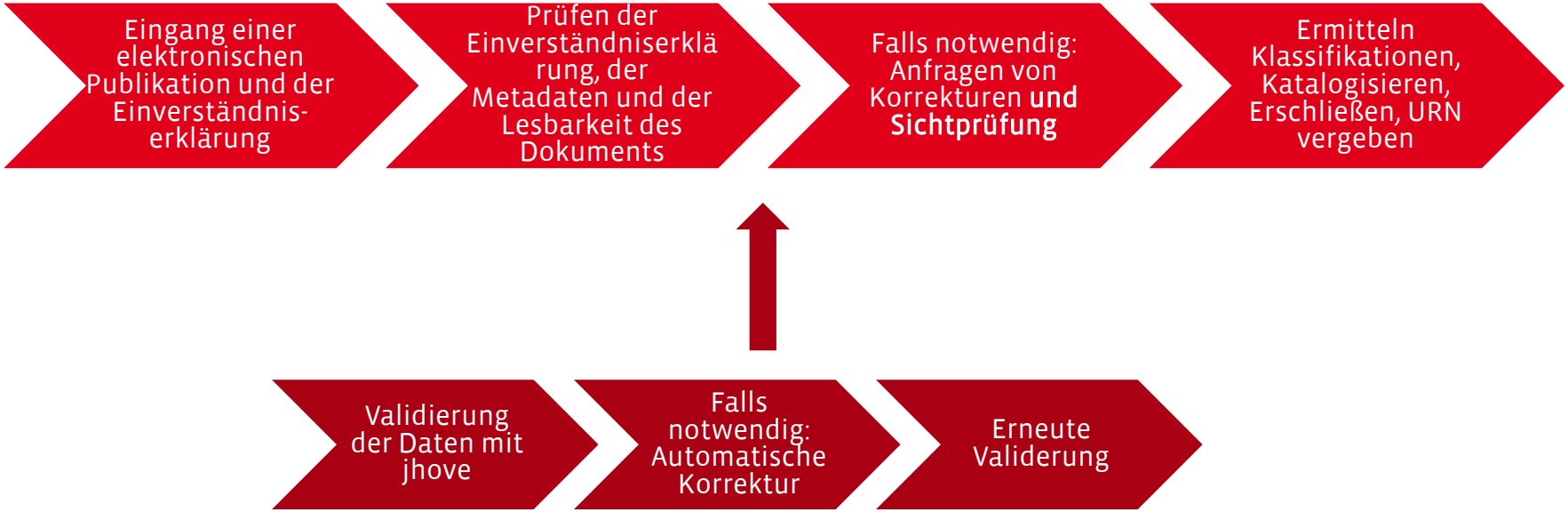
→ Geringer

$$v = \frac{d_0 \cdot \pi \cdot n \cdot \cos\Delta}{6000}$$

Organisatorische Lösungen
Veränderung des Annahmeworkflows

Eine mögliche Lösung

- Prüfung des Ergebnisses der Korrektur durch den Urheber



© by SLUB Dresden

SLUB
Wir führen Wissen.

Zusammenfassung und Ausblick

Zusammenfassung
Einsatz des Datenformates PDF für die LZA

PDF/A-1 und PDF/A-2 sind aus unserer Sicht LZA-fähige Datenformate

PDF selbst ist zu komplex, erlaubt Zugriffsbeschränkung und Komprimierung

Elektronische Ablieferung sollte im Hinblick auf die digitale Langzeitarchivierung in PDF/A-1 und PDF/A-2 erfolgen

Einflussnahme auf die Produktion ist empfehlenswert

Einführung einer Validierung und Korrektur von pdf-Dokumenten schon bei der Ablieferung/Annahme ist sinnvoll aber aufwendig

© by SLUB Dresden

SLUB
Wir führen Wissen.

Die Eignung des Dateiformates pdf für die Langzeitarchivierung – Probleme und Lösungen

Dr. Henrike Berthold, Sabine Krug, Christian Beissert,
Michaela Voigt, Anke Wartenberg
Sächsische Landesbibliothek –
Staats- und Universitätsbibliothek Dresden (SLUB)

Juni 2014