Risiken und Nebenwirkungen von Datenmigrationen

Yvonne Friese

Digitale Langzeitarchivierung

BID Kongress 2013

Leipzig

12.03.2012

Inhalt

Langzeitarchivierung in Goportis

Verbund der drei Zentralen Fachbibliotheken in Deutschland

TIB Hannover

ZBW Kiel

ZBW Hamburg

ZB MED Köln

ZB MED Bonn

Langzeitarchivierung an der ZBW (2012 + 2013)
PDF-Dateien im Fokus

Open-Access-Dokumentenserver EconStor (50.000 PDF-Dateien)

- Stand 03/2013: 36.000 PDF-Dokumente im Rosetta-Langzeitarchiv
- Rund 20 % hiervon weisen Fehler auf:
    - 16 % nicht valide und nicht wohlgeformt
    - 4 % wohlgeformt, aber nicht valide

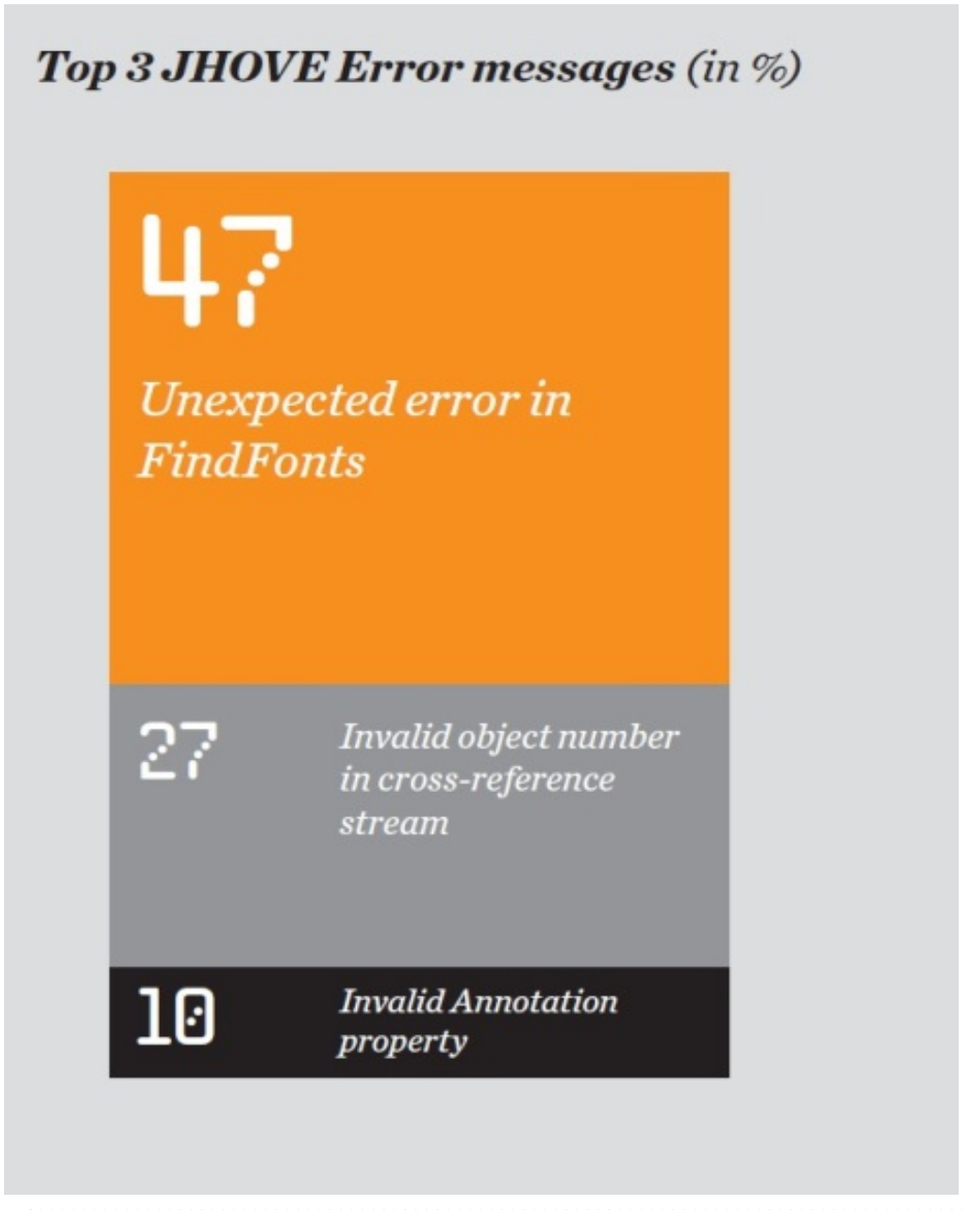
Was ist das? Ist es das wirklich?

Nutzung von Open Source Tools während der Überführung der Dateien in das Langzeitarchiv:

- zur **Identifikation des Dateiformats** (DROID, Format Library Pronom / The National Archives London)

- zur **Validierung des Dateiformats:** Entspricht die Datei den Formatspezifikationen? (JHOVE, entwickelt von der Harvard University Library)

Prüfung nicht bestanden: Abgelehnt



- 20% der PDF-Dateien werden von JHOVE abgelehnt

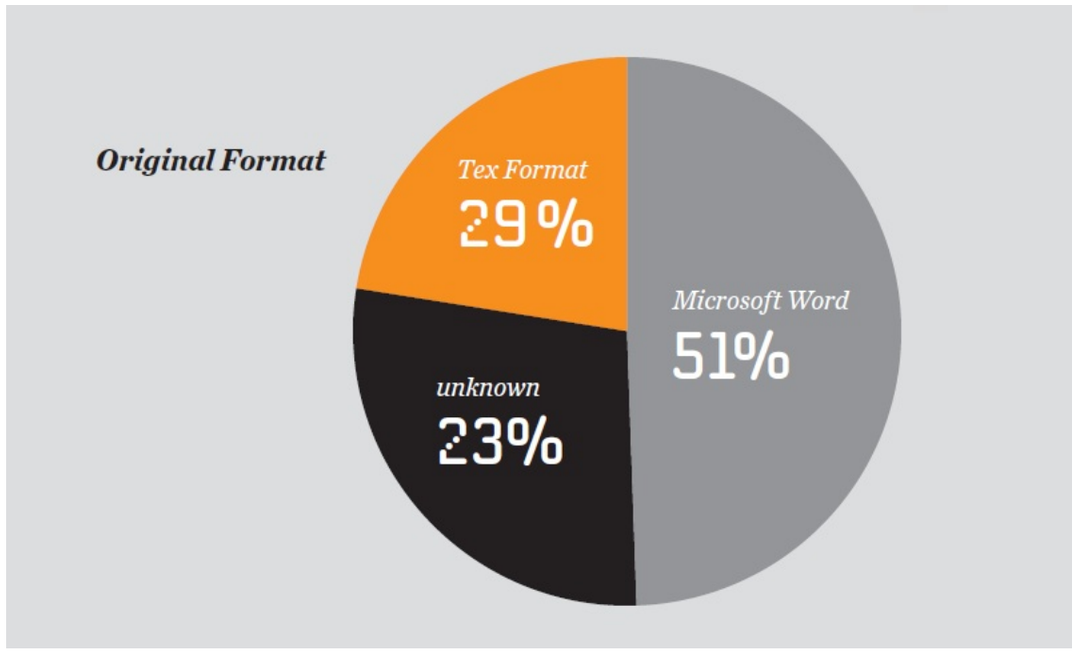
JHOVE ist nicht das perfekte Werkzeug für die PDF-Validitäts-prüfung:

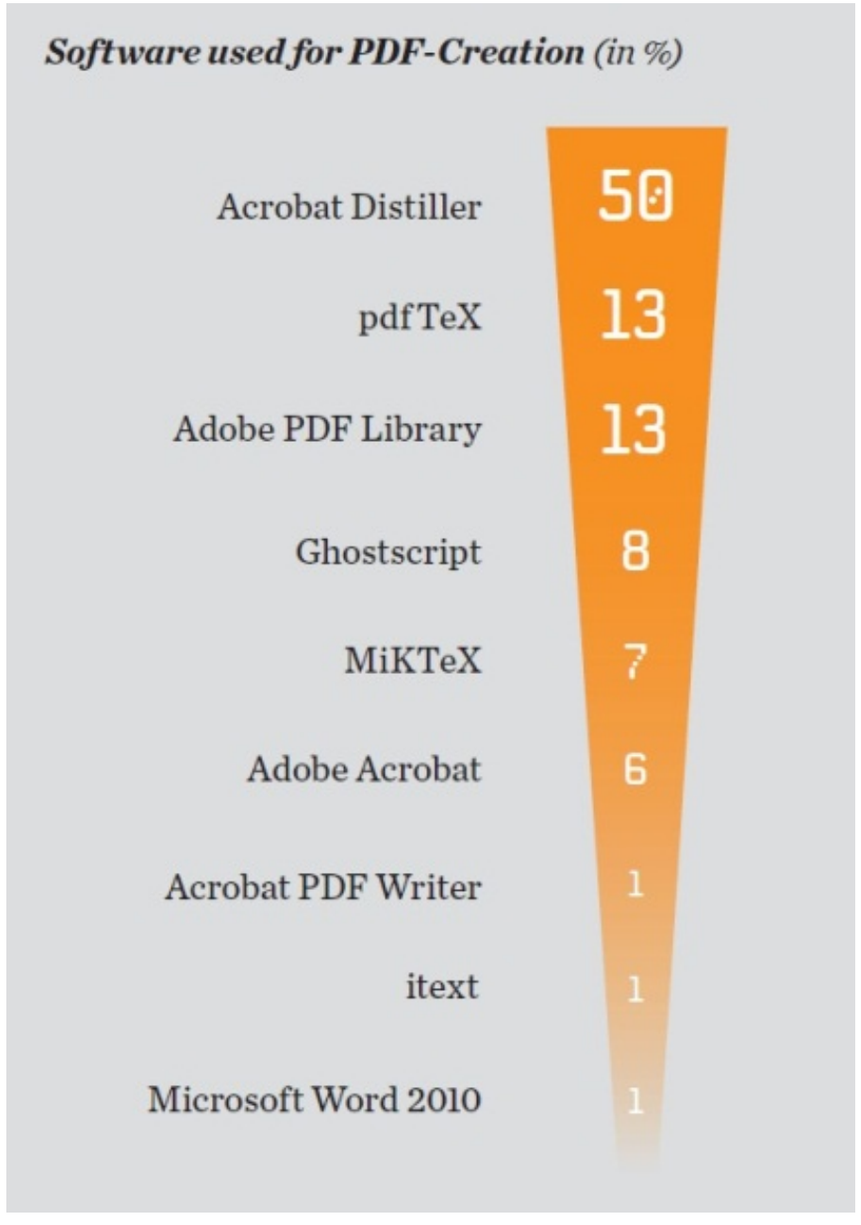
- Erkennt nicht jedes Problem (false positive)
- Ablehnung zu Unrecht (false negativ)
- Bedeutung der Fehler-meldungen oft nicht bekannt

Was war es vorher?

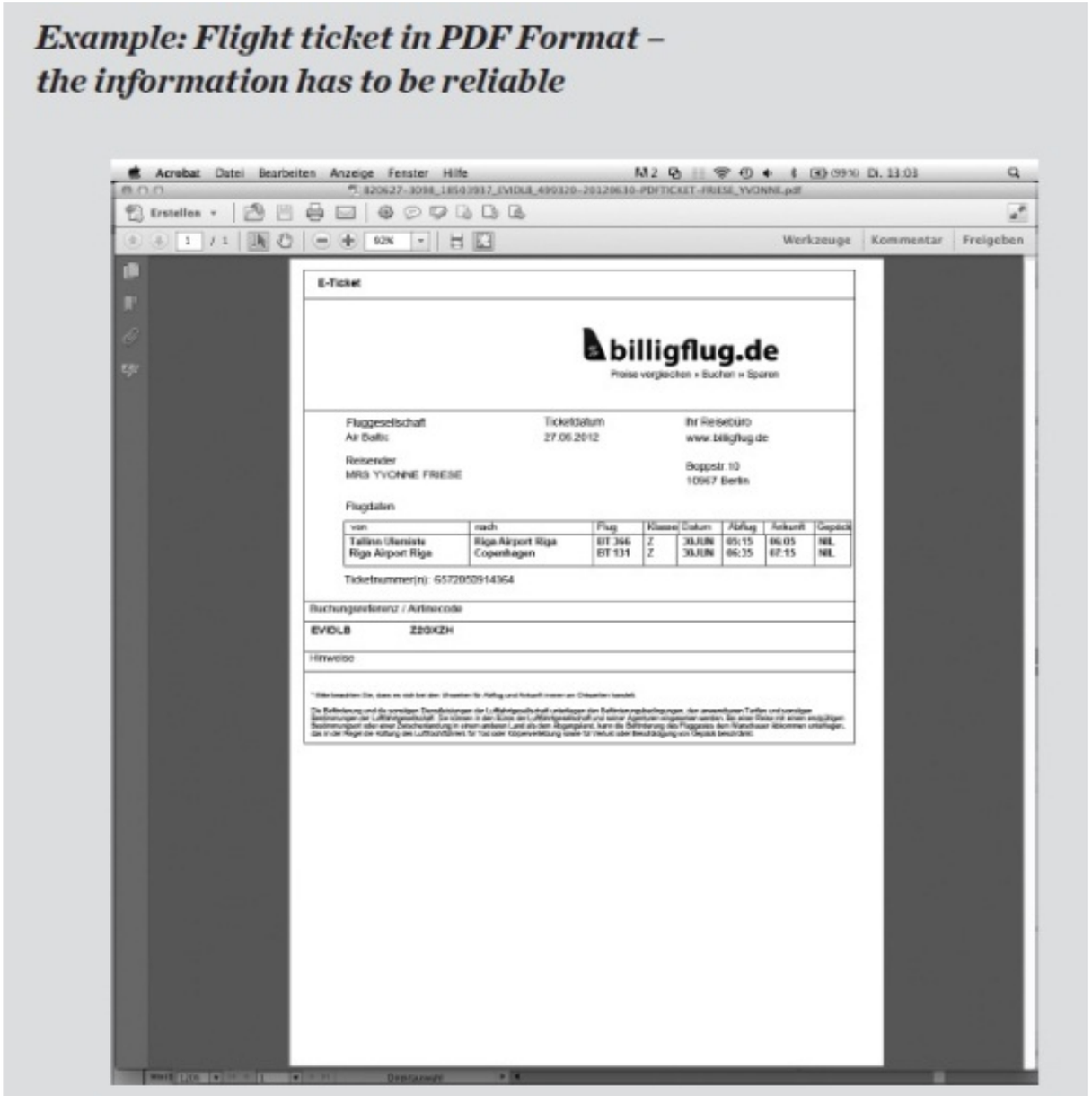
Heterogenität der Dokumente: Die zur PDF-Erstellung genutzte Software ist sehr unterschiedlich.

Das Ursprungsformat ist bei jedem vierten Objekt unbekannt.





Anspruch:
Valide und wohlgeformte Dateien im Langzeitarchiv



– Migration betroffener Dateien in valide und wohlgeformte PDF-Dateien (via Itext-Plugin)

– Bewahrung Original-PDF + migriertes PDF

– Dokumentation des Migrations-Vorgangs in den Metadaten

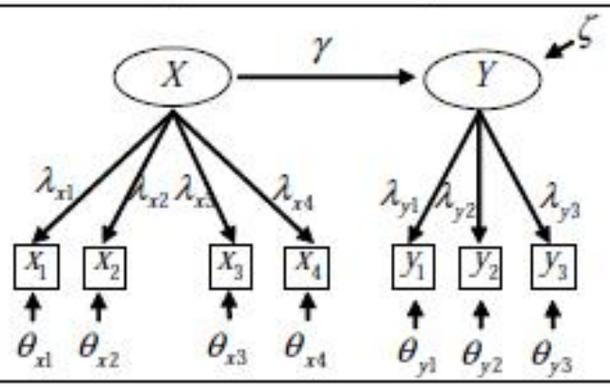
– Nutzung Risiko-Management-Modul (u. a. für spätere Fehlerbehebung)

Eineiiger Zwilling?

Migrationen können Datei visuell verändern

(links Original-PDF, rechts migriertes PDF/A)

Figura 1. Modelo de investigación



siendo:

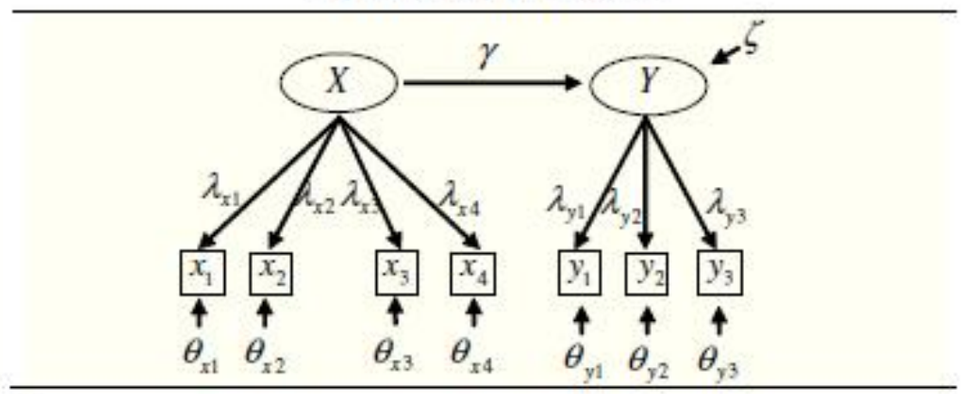
λ_i coeficiente de regresión entre el constructo y cada indicador
θ_i varianza de error de cada indicador
γ coeficiente de regresión entre las variables latentes
ζ varianza de error de la variable dependiente

indicadores de las dos variables, el investigador suele proceder al estudio de la validez discriminante a través de los siguientes métodos.

3.1. Comparación entre las correlaciones de los indicadores

Según las recomendaciones de Campbell y Fiske (1959), como las variables X e Y son indicadores de constructos distintos. existe validez discriminante si todas las correlaciones

Figura 1. Modelo de investigación



siendo:

λ_i coeficiente de regresión entre el constructo y cada indicador
θ_i varianza de error de cada indicador
γ coeficiente de regresión entre las variables latentes
ζ varianza de error de la variable dependiente

indicadores de las dos variables, el investigador suele proceder al est
discriminante a través de los siguientes métodos.

3.1. Comparación entre las correlaciones de los indicadores

Según las recomendaciones de Campbell y Fiske (1959), como las va
indicadores de constructos distintos, existe validez discriminante si t

Risiken und Nebenwirkung der gewählten Lösung

Pro-Argumente

+ Skalierbarkeit durch Automatisierung
+ Gewählte Tools basiert auf Open Source Software
+ Absicherung durch Mitarchivierung der Original-Datei

Contra-Argumente

- Gefahr der visuellen Veränderung durch Migration
- Sichtprüfung zeitaufwändig
- Rechtliche Schwierigkeiten
- Migrieren von Original-Datei in PDF nicht möglich, da Ursprungsdateien nicht verfügbar

Zusammenfassung und Ausblick

- Integration des Itext Plugins in das Preservation Planning Modul von Rosetta

- Automatische Sichtprüfung für PDF-Dokumente heute noch nicht in benötigter Form verfügbar

- Vorerst das Original bereitstellen und die migrierte Version mit archivieren

Fragen?

Yvonne Friese
Deutsche Zentralbibliothek
für Wirtschafts-
wissenschaften
(ZBW, Kiel)

Projektmanagerin
Digitale Langzeitarchivierung
y.friese@zbw.eu