Zurückweisung beschränkt archivfähiger Dateien? Vom Umgang mit technisch problematischen Daten.

-Ein Werkstattbericht-

Jens Peters, M.A., LVR-InfoKom

Jahrestagung AK AUdS, Prag, März, 2019

Agenda

- Einführung
- Problemstellung
- Qualitätskategorien
- Fazit
- Fragen & Antworten

DA-NRW Software Suite (DNS)

- Die **DA-NRW Software Suite (DNS)** ist ein Modul der Langzeitarchivierungslösungen des Landes Nordrhein-Westfalen (NRW) im Lösungsverbund DA NRW www.danrw.de
  - Mehrfach redundanter, dezentraler Archivspeicher
  - Zentrales Erhaltungsmanagement
  - Servicecharakter
  - Wird von LVR-InfoKom im Auftrag weiterentwickelt
- LVR-InfoKom ist das Systemhaus des Landschaftsverbandes Rheinland (LVR) und einer der Servicegeber und Entwicklungspartner.

Einführung

- Würde ein klassisches Archiv ein analoges, archivierungswürdiges, beschränkt archivfähiges Dokument zurückweisen?

- Nein, dieses würde konservatorisch behandelt und fachgerecht archiviert werden.

Kernfrage

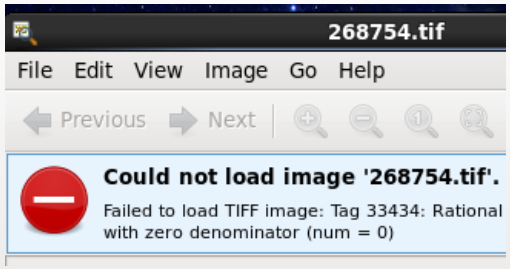
- Wie gehen Digitale Archivierungssysteme mit nur beschränkt archivfähigen Dateien um?

- Nebenaspekt: Was ist überhaupt ein „Fehler“ und wie wird dieser erkannt?

Was bedeutet „technisch problematisch“?

- Bspw. eine Datei eines **validen** Abgabepakets (SIP)
  - enthält **Lesefehler**
- Entdeckung beim Zugriff auf die gespeicherte Information:
  - lässt sich nicht (erfolgreich und korrekt) identifizieren oder validieren
  - Ein Codec wird nicht erfolgreich erkannt.
  - Die Datei entspricht nicht den Formatvorgaben des Services
  - (…)

- Insbesondere der Punkt „Lesefehler“ ist
  - Versionsabhängig -> ständige Weiterentwicklung
  - Werkzeugabhängig
  - Systemumgebungsabhängig
  - Ist ggfs. vom **Zeitpunkt** der Prüfung abhängig
  - Kurz: er ist in hohem Maße „interpretationsfreudig“



Problemstellung (Sicht OAIS, CCSDS 650.0-M-2, 2012)

» Wer ist aus Sicht des OAIS Referenzmodells zuständig für die Reparatur problematischer Dateien?

» Im OAIS Ingestprozesses, hier der *audit function,* wird ausgeführt:

"AIP/SIP reviews from Preservation Planning and may also **involve an outside committee** e.g., science and technical review (…) The Audit process may determine that some portions of the SIP are **not appropriate for inclusion** in the Archive and must be **resubmitted** or **excluded** (…) After the audit process is completed, any *liens* are reported to the Producer, **who will then resubmit the SIP"** (Vgl. S. 56, CCSDS 650.0-M-2)

» OAIS weist diese Aufgabe zur Lösung eindeutig an den Einlieferer zurück.

Problemstellung in der Praxis

Einliefernde Stelle

- Hohe Datenmenge zur Abgabe
- Neuerzeugung der Datei ggfs. gar nicht (mehr) möglich.
  - z.B. eine „bessere“ Version der Datei ist nicht greifbar.
- Werkzeuge (Tools) und Know-How nicht (mehr) vorhanden.

Digitales Archivierungssystem

- (unkategorisierte) *bitstream preservation* fehlerhafter Dateien vermeiden!
- Hoher Arbeitsaufwand in der nachgelagerten Analyse technisch problematischer Dateien.
- Ggfs. Verarbeitungsstau (im Ingest), da Fehlerpostkörbe nicht schnell genug abgearbeitet werden können.
- Wie kann die Bestandserhaltung optimal unterstützt werden?

Lösungsszenario bei der DA-NRW Software Suite (DNS)

- Einführung von Qualitätskategorien (z.Zt. fünf)
- Arbeitet im Ingest und bei späteren Formatmigrationen im Preservation Management und bei „Delta“ (=Ergänzung von Abgaben)
- Jede Kategorie ist von Qualitätsmerkmalen gekennzeichnet, als Ausdruck der Eignung als digitales Archivgut.
- Kenntlichmachung und Dokumentation und in PREMIS
- Transparenz der Zuordnung und deren Merkmale
- Festlegung der Kategorien in einem Fachgremium

Lösungsszenario: Vorgehen

» Anforderung: Eine qualitative Unterscheidung der AIP innerhalb der DNS war gewünscht.

Bsp. Qualitätskategorien bei DNS, Implementation

Kriterium	Kat. 0	Kat. 1	Kat. 2	Kat. 3	Kat. 4	Kat. 5
SIP ist virenfrei und konform zur DNS SIP-Spezifikation	Ablehnung	X	X	X	X	X
Fehlerfreie **Dateiidentifikation** (inkl. Codecs)			X	X	X	X
Alle Dateien unterstützter Formate im SIP sind nach Aussage eines Validators zu der jeweiligen Formatspezifikation konform. (**Validation**)			X		X	X
Alle Dateien unterstützter Formate können (falls notwendig) zu LZA-Format migriert werden. (LZA-Migrierbarkeit)				X	X	X
Alle Dateien im AIP sind im Sinne der DNS Spezifikation unterstützte Formate						X
Eignung zur dig. LZA in DNS	Keine, Neueinl.	Bitstream pres.	Teilw.	Voll.	Voll.	Voll.

Fazit

- „Problematische“ Dateien sind eher Normalfall als die Ausnahme
- Zurückweisung führt nicht unbedingt zur Lösung
- Alle Systembausteine & Verfahren verursachen s.g. „false positives“ und „false negatives“
- Unterstützung der Bestandsmanagements durch Kategorisierung von möglichen Fehlerquellen sinnvoll – und auch möglich.

Mehr unter:

www.danrw.de

www.Github.com/da-nrw/DNSCore

Fragen & Antworten

Vielen Dank!