Protokoll zum 2. Treffen

am 4. November 2014, 11.00-16.00 Uhr in Köln

Teilnehmende

• Claire Roethlisberger (KOST)
• Christine Rost (Thür. Hauptstaatsarchiv Weimar)
• Stefan Hein (DNB)
• Martin Hoppenheit (LA NRW)
• Simone Thierfelder (Gast, ZB MED)
• Tim Hasler (Zuse Institut Berlin)
• Sina Westphal (Bundesarchiv)
• Joachim Rausch (Bundesarchiv)
• Marion Germies (AbbVie)
• Heinz Werner Kramski (DLA Marbach)
• Adnane Bousfiha (LABW)
• Yvonne Friese (ZBW, *Moderation*)
• Andre Müller (Gesis, *Protokoll*)

Agenda

TOP 1: Mögliche Aufgaben der AG Formaterkennung

TOP 1.1: Inhaltliches

Formaterkennung
• Diskussion über bestehende Lücken und Bedarf an Zuarbeit
• Grundsatzdiskussion an die Anforderungen der File Format Registries / Formaterkennung
• Mögliche Zuarbeit NSLA, Pronom und Co.

Aufbau einer Wissensdatenbank (Themen: Pronom, UDFT, Tools, Kompaktwissen zu wichtigen Archivformaten)
• nestor Knowledge Base: Welche Institution kennt sich mit welchen Formaten aus? Info ins Wiki inkl. Kontaktdaten. (Formatexperten-Datenbank?)
• Pronom neue Signaturen Pattern erstellen. YF hält dies weiterhin für wichtig, z. B. für Tei-Dateien und NITF-Dateien, die zurzeit nicht korrekt erkannt werden bzw. keine Pronom PUID haben. Zwar basieren diese auf fmt/101, sollten aber möglicherweise genauer kategorisiert werden.

Ausbau der Dateien-Sample, verwalten, testen und Fokus festlegen/ausbauen
• Speicherort für Dateien-Sample (deutlich als nestor-Arbeitsergebnis erkennbar)

• Strukturierung der Dateien in den Samples: Nach Jahrgang? Format? PUID? Typ?
• "Let's Solve the File Format Problem!" (s. TOP 2) hat schon sehr viele Beispiele.

PDF-Validierung (Tools, Tool-Samples, Tests und Veröffentlichung dieser Tests)
Validierung anderer Formate?
• Recherche zu Tiff-Dateien (Baseline-Tiff, Tags, Wissensdatenbank hierzu, Programme zum Prüfen der Tiffs) (geplant: Workshop zu Bildformaten wie Tiff und JPEG200 im Herbst 2015 in Köln von Goportis. Aus der AG gibt es hier auch Interesse bzgl. der Teilnahme.
• Epub-Dateien langzeitarchivieren
• Wer hat Erfahrungen mit Audio-Dateien und LZA von Audio-CDs? (Austausch mit nestor AG Media?)
Fazit: Entwurf des Mission Statement der AG

TOP 1.2: Organisatorisches

Treffen in größerer Runde (ganze AG, mind. 12 Leute) 2x im Jahr?
Unter-AGs für gezielte Aufgaben? Bei kleinerer Gruppenstärke auch Telkos möglich.
Falls Idee Anhängerinnen/Anhänger findet: Sammlung der Themen für die UAGs + Interessierte, YF legt entsprechende Seiten im Wiki an. Mögliche UAG-Themen:
• Dateien-Sample
• PDF-Validierung/Analyse (Zusammenstellung und Analyse von Tools)
• Bildformate (Überblick über Formate und Analyse-Tools -> Wissensdatenbank)

TOP 2: Vorstellung für die AG relevanter Projekte

• TREASURES (noch nicht angenommen) (DNB/Stefan Hein)
• PREFORMA (PREservation FORMAts for culture information/e-archives)
• Let's Solve the File Format Problem! ("[...] there is a lot of spread-out information about file formats in the world, and almost universal acknowledgement that there are too many to keep track of and too much information in too spread-out an area for it ever to be assembled. This Wiki is an attempt to bring together all that spread out information.")

TOP 3: Berichte von Konferenzen und Treffen

Bericht von der iPRES (bzgl. Formatfragen), z. B.:
• Modelling file formats and technical environments using the NSLA Digital Preservation Technical Registry (DPTR)
• Achieving Canonical PDF Validation, Duff Johnson
• Panel: Getting to Digital Preservation Tools that "just work"
• Converting WordStar to HTML4
• Sustainability Assessments at the British Library: Formats, Frameworks & Findings
• A Model for Format Endangerment Analysis using Fuzzy Logic
• Occam's Razor and File Format Endangerment Factors. Heather Ryan
• Preservation of ebooks: from digitized to born-digital
PDF Technical Conference

TOP 4: Sonstiges

nestor AG digitale Bestandserhaltung (C. Keitel) plant ggf. eine Schnittstelle zu uns. Sie kommen bei Bedarf auf uns zu.
TOP 5: nächstes Treffen

Protokoll

Problemstellung
Viele Archive haben Dateien, deren Dateityp bzw. -untertyp (Version) nicht oder nicht hinreichend genau automatisch erkannt werden können.

Desiderata und Lösungsansätze
Pronom/Droid alleine reicht nicht aus, da es bei der Erkennung grundsätzlich auf identifizierende Marker im Bitstream angewiesen ist. Das ist nur bei einigen Dateiformaten gegeben.
Für eine leistungsfähigere Formaterkennung soll eine "Formaterkennungskaskade" mit mehreren Erkennungs-Helfern wie etwa den Unix-Fileutils arbeiten, die darüber hinaus auch andere Eigenschaften einer Datei in die Analyse einbeziehen. Es herrscht Konsens, dass Pronom/Droid hier Bestandteil sein sollte. Die Erstellung einer neuen Registry wird als zu ambitioniert und nicht sinnvoll angesehen.
Zusammenstellungen von vorhandenen Tools und ihrer Möglichkeiten und Grenzen finden sich etwa bei COPTR (http://coptr.digipres.org/Main_Page). COPTR wird offen auf einer Wiki-Plattform entwickelt; hier können eigene Beiträge eingestellt werden. Eine Zuarbeit wird unterstützt; eine weitere "Wissens-Insel" ist weniger pfiffig. Im Rahmen des Formaterkennungs-Wikis sollen die vorhandenen Ressourcen verlinkt werden. Claire Roethlisberger und Christine Rost bauen eine Seite zu COPTR auf und die anderen AG-Teilnehmerinnen und Teilnehmer können hier zuarbeiten (Arbeitspaket 1). Ein großes Problem für die Formaterkennung sind eingebettete Dateien. Hierfür ist keine generelle Lösung in Sicht bzw. bekannt.

Arbeit mit problematischen Dateien
Andre Müller koordiniert die Sammlung und Dokumentation von Dateien bzw. Dateiformaten, die sich einer Erkennung entziehen oder nicht mehr ohne weiteres gelesen werden können (etwa das .odb-Format, ein MS-Office-Metaformat (Arbeitsmappen)) (Arbeitspaket 2).
Weiterhin sollen fehlerhafte PDF-Dateien zusammengetragen werden. Hier wurde auch die Grenze zwischen Erkennung und Validierung diskutiert. Ein PDF-Validierer wird aber als deutlich zu aufwändig gesehen.
Die Frage, ob die AG sich mit der Validierung überhaupt bzw. für einzelne Dateitypen befassen soll, soll in Arbeitspaket 5 (s.u.) diskutiert werden.

"Hardware-Formate/Abspielgeräte und Speichermedien"
Wir wollen bestimmen, welche Medien von welchen Organisationen noch gelesen werden können und wo es hilfsbereite Experten gibt. Dazu soll eine Registry bzw. Kontaktbörse für Abspielgeräte und Speichermedien entstehen. Heinz Werner Kramski koordiniert die Arbeiten. (Arbeitspaket 3)

"Dateiformate-Grundgesamtheit"
Es sollen die in den Archiven und Bibliotheken vorhandenen Dateiformate gesammelt werden, um festzustellen, mit welchen Dateitypen wir uns befassen müssen bzw. können. Im Umkehrschluss kann so positiv festgestellt werden, mit welchen Formaten wir uns nicht auseinandersetzen müssen. Die AG-Teilnehmer sind aufgerufen, die vorhandenen Dateitypen zu listen und anzugeben, ob und wieweit Expertise zum Umgang mit den Dateien vorhanden ist (Arbeitspaket 4: Formatsammlung - wer hat was).
Es ist zweifelhaft, ob dies je den Status eines Kanons erlangen kann; hilfreich ist die Listung auf jeden Fall. Hiermit wird sich eine ganze Gruppe um Yvonne Friese befassen.
In der Diskussion kommt die AG überein, dass die Arbeit sich auf bekannte und vorhandene Fälle beschränken soll. Eine Suche nach obskuren Grenzfällen wird nicht als sinnvoll gesehen.

Ein Versuch der Einordnung und Charakterisierung von Dateiformaten soll am Beispiel von einem bis wenigen Dateiformaten durchgespielt werden. Wesentliche Leitfragen sind hierbei:
• komplexes (z.B. PDF, TIFF) oder einfaches (Text, PNG) Format?
• Bytesequenz mit Dateikennung direkt parsbar? (PDF ja, TIFF muss erst interpretiert werden) Dies zeigt insbesondere an, ob zu den Formattypen eine Erkennung via PRONOM der richtige Ansatz ist.
• Vorhandene Ressourcen/Tools zum Umgang mit den Formaten bestimmen.

Mission Statement

Das Mission Statement wurde grob entworfen und wird beim 3. Treffen verfeinert.

"Errata-Howto"

Soll eine "Fehlerdatenbank" für verbreitete Tools (Droid, Jove) aufgesetzt werden, um harmlose und relevante Fehlermeldungen erkennen zu können? Es wird das Problem gesehen, dass die Frage der Relevanz von Fehlern davon abhängt, welche Dokumenteneigenschaften für den jeweiligen Archivzweck als relevant angesehen werden (z.B. ist die Erhaltung von Layout/Fonts relevant?). Auch wird angemerkt, dass dies über die Erkennung hinausgeht und zum viel größeren Problem der Validierung gehört.
Eine Lagebestimmung soll in einer Gruppe "Validierungsfragen" vorgenommen werden.

Hinweise auf hilfreiche Ressourcen

• TREASURES über Probleme mit Pronom/Droid, Überlegungen zur Nachhaltigkeit, eines Community-Ansatzes, eines Business-Model
• UDFR (in Planung, mit 18 EU- und 2 nicht-EU-Partnern), Projekt zur Validierung von A/V-Formaten
• Das Archive Team bietet mit "Let's solve the file format problem" unter fileformats.archiveteam.org eine Informationssammlung zu Dateitypen.
• Archivematica ist eine Open Source-Alternative zu Rosetta. Die Software wird von einer Firma entwickelt. Neue Features können "gekauft" werden, d.h. man kann die Entwickler bezahlen, neue Features zu entwickeln, die dann zur freien Nutzung veröffentlicht werden. Die Webseite ist auch eine gute Quelle zu Referenzen zu Formatinformationen.
• bitcurator.net bietet eine Live-Linux-Distribution an, die viele Formaterkennungstools mitbringt. Der Schwerpunkt liegt aber eher auf forensischen Tools.
• Aus einem PDF-Hackathon im September ist ein Tool zur PDF-Analyse bzgl. PDF/A-Fehlermeldungen (Validator basiert auf Apache Tika) und PDF-Erstellungssoftware hervorgegangen, der Quellcode liegt auf github.
• Eine nestor-AG soll sich mit der Validierung von PDF/A befassen wollen. Das ist womöglich für andere nestor-Mitglieder zugänglich?

Info-Punkte

• Es gibt Planungen für einen Workshop zu Bildformaten (Tiff, JPEG und JPEG2000) im Herbst 2015 in Köln an ZB MED.
• Vom 8.-10. Juni 2015 wird im Maternushaus Köln eine Tech Conference zu PDFs veranstaltet.

Arbeitspakete und Verantwortliche

Überblick über die vorhandenen Tools (Arbeitspaket 1)

• Claire Roethlisberger (Koordination)
• Christine Rost (Koordination)
• Yvonne Friese
• Heinz Werner Kramski

• Joachim Rausch

Archivematica und Co (Arbeitspaket 2)
• Andre Müller (Koordination)
• Tim Hasler
• Heinz Werner Kramski
• Joachim Rausch

Hardware-Formate/Abspielgeräte und Speichermedien (Arbeitspaket 3)
• Heinz Werner Kramski (Koordination)
• Marion Germies
• Tim Hasler
• Stefan Hein
• Andre Müller
• Joachim Rausch

"Kontaktbörse Formate" (Arbeitspaket 4)
• Yvonne Friese (Koordination)
• alle
• Scope wird am 25.03. klarer festgelegt. Grob geht es darum, dass jede/r erfassen kann, mit welchen Formaten Erfahrungen bestehen und welche man (in welchen Mengen) vorliegen hat. So können sich Personen mit ähnlichen Formatproblemen innerhalb der AG besser zusammenfinden ggf. für speziellere Unter-AGs.

Arbeitspaket 4.1.: Zuarbeit zu Format Registries
Christine, Martin, Yvonne und ggf. TIB (da diese hier bereits Erfahrungen haben)

Validierungsfragen (Arbeitspaket 5)
• Yvonne Friese
• Tim Hasler
• Christine Rost
• Sina Westphal

Nächstes Treffen
Das nächste Treffen wird am 25.03.2015 in Marbach stattfinden.