Protokollentwurf des 1. Treffen

12.05.2014 11:30 - 16:00 Uhr an der DNB in Frankfurt a.M.

Teilnehmer

Adnane Bousfiha (LA BW), Marion Germies (abbVie), Georg Büchler (KOST), Yvonne Friese (ZBW) [Moderation], Heinz Werner Kramski (DLA), Michelle Lindlar (TIB) [Protokoll], Andre Müller (gesis), Tobias Steinke (DNB), Steffen Schilke (hzg), Christoph Schmitt (LAV NRW), Armin Straube (nestor), entschuldigt: Mario Röhrle (Staatliche Akademie der Bildenden Künste, Stüttgart), Tim Hasler (KZI)

TOP 1: Vorstellung Teilnehmende und Vorstellung der AG-Arbeit

Die Vorstellungsrunde zeigte, dass in den vertretenden Institutionen unterschiedle Spektren an Formatvielfalt abgedeckt sind - rangierend von wenigen, überschaubauren (KOST, gesis) bis hin zu einer großen Vielfalt inklusive älteren (DLA) und hoch spezialisierten (AbbVie) Formaten.

Zielvorstellungen der Teilnehmer beinhalten neben dem Erfahrungsaustausch und Erfahrungsgewinn zum State-of-the-Art einige spezifische Fragestellungen. So wurden Tools und Software für den Umgang mit AV-Formaten (LA BW), Verfahren im Umgang mit Formaten aus veralteten Dateisystemem wie 7-bit (gesis), und ein robuster und frei verfügbarer PDF/A Validator (hzg) als Vorstellungen geäussert.

In der anschließenden Diskussion wurden 4 Themen als mögliche Kernthemen genannt:

- Formaterkennung
- Formatvalidierung
- Emulation / Migration
- QA

TOP 2: Begrifflichkeiten

Erste Diskussionen zeigten die Notwendigkeit einer Begriffsdefinition innerhalb der AG. Hierbei wurde folgender Konsens erreicht:

1.) Erkennung / Identifizierung

- Es wurde festgestellt, dass die Begrifflichkeiten im Deutschen generell synonym verwendet werden, eine Granularitätsunterscheidung aber möglich sei.
- "Erkennung" beschreibt eine allgemeinere Zuordnung eines Objekts zu einem Dateiformat, bspw. basierend auf äusserlichen Merkmalen wie der Dateiendung
- "Identifizierung" beschreibt eine genauere Zuordnung, bspw. basierend auf dem Abgleich gegen eine Formatbeschreibung / Formatbibliothek.

2.) Charakterisierung

- Meint die Generierung technischer Metadaten, dies kann auch Informationen aus den Prozessen Erkennung/Identifizierung und Validierung beinhalten.

3.) Validierung

- Meint die Prüfung gegen eine Formatspezifikation
- Die Möglichkeit der Validierung via der Lesbarkeit des Formats durch Software des Herstellers zu prüfen (z.B. Adobe Reader für pdf) wurde kontrovers diskutiert. In diesem Rahmen ist insbesondere die Toleranz, auch von Herstellersoftware, gegenüber Fehlern in der Objektstruktur zu beachten.

Es wurde festgehalten, dass die Definitionen der Begrifflichkeiten im Wiki im ersten Schritt für die AG selbst, zu späterer Zeit ggf. auch für Externe erfasst und zugänglich gemacht wird.

TOP 3: Öffentlich zugängliche Testsets

Die Möglichkeit, ein öffentlich zugängliches Testset (z.B. äquivalent zu PDF Cabinet of Horrors oder Isator des PDF Competence Centers) zu erstellen. Das Testset sollte nicht nur kaputte Dateien beinhalten, sondern auch valide, um als Benchmark fungieren zu können.

Das genaue Umfangsspektrum muss noch festgelegt werden. Es wurde angemerkt, das bei alten Objekten z.B. aus den 80ern auch unterschiedliche Kodierungen betrachtet werden sollten.

TOP 4: File Format Registries

Es wurde ein kurzer historischer Abriß zum Thema File Fromat Registries gegeben (GDFR, UDFR, PRONOM), sowie zwei aktuelle Bestrebungen berichtet:

1. Eine PRONOM linked data Weiterentwicklung durch Tessella im Rahmen von Preservica, allerdings ohne absehbare Bestrebungen, diese Weiterentwicklung über den Kundenkreis hinaus zugänglich zu machen
2. Das Modell der National and State Libraries of Australasia (NSLA) mit dem Aufruf an diverse europäische Nationalbibliotheken, sich hier im Rahmen eines Horizon 2020 Antrags zu beteiligen.

Die DNB beabsichtigt auf jeden Fall, sich an letzterem zu beteiligen.

Es wurde diskutieret, wo die AG ein mögliches Arbeitsfeld im Zusammenhang mit FIle Format Registries sieht, z.B. ob nur die Nutzung oder auch die Zuarbeit angedeacht ist. Vorschläge zu diesem Thema beinhalten eine deutschsprache Anleitung zur Signaturerstellung.

TOP 5: Weiteres Vorgehen

Basierend auf den Diskussionen zu den unterschiedlichen Themen wurde in der AG folgendes beschlossen:

Die AG arbeitet zunächst ausschließlich im Wiki. Hierzu wird ein AG-interner Bereich eingerichtet, welcher nur für AG-Mitglieder zugänglich ist.

Die Sichtbarkeit einzelner Wiki-Inhalte nach aussen wird innerhalb der AG beschlossen.

Als konkrete Aufgaben und Zuständigkeiten für erste Wiki-Inhalte wurden folgende festgelegt:

- Einführungen der beschlossenen Definition zu Begrifflichkeiten --> NN
- Verlinkung auf COPTR --> NN (Vorschlag fri: alle jeweils aus dem selbst erstellten Text heraus verlinken)
- kurze Beschreibung zu UDFR --> Steinke
- kurze Beschreibung zu PRONOM --> Friese
- kritische Ideen zum Thema PRONOM ergänzen --> Büchler
- kurze Beschreibung NIST (National Software Registry Library) --> Kramski
- Use Cases zur Formaterkennung / Formatvalidierung --> alle

--> beinhaltet Nutzen; Grund für Identifizierung wird dargestellt, weitere Arbeitsfelder werden identifiziert

- Sammlung von Fehlermeldungen in jhove --> Friese macht hier den Anfang für PDF, alle sind zum Beitrag aufgerufen

Als Zeitplan für die Wiki-Arbeit wurde festgelegt, dass die Struktur sowie erste Einträge bis Ende der Sommerferien (Ende August) zur Verfügung stehen sollten.

Des Weiteren wurde beschlossen, das Thema "Testsets" weiter zu behandeln. Hier sind alle AG Mitglieder aufgerufen, geeignete Objekte zu sammeln.

TOP6: Nächstes Treffen

Das nächste Treffen findet bei der GESIS in Köln statt. Eine Doodleumfrage für mögliche Termine ab dem 22. Oktober wird erfolgen. Als Wochentage kommen Dienstag, Mittwoch und Donnerstag in Betracht.

Ziel: beim nächsten Treffen soll das mission statement festgelegt werden.