Agenda des 5. Treffen

9. Mai 2016 in Kiel, 11:00 bis 16:00 Uhr

zugesagt: Christine, Konrad und Svenia Pohlkamp (Thüringisches Hauptstaatsarchiv), Heinz, Michelle, Stefan H. (DNB), Martin, Andre, Rainer, Yvonne

entschuldigt: Marion, Claire, Mario, Christine, Tim

Begrüßung

Alle Teilnehmerinnen und Teilnehmer stellen sich vor. Svenia und Rainer (Sinas Nachfolger) sind zum ersten Mal dabei. Das Protokoll vom letzten Mal kann im öffentlich sichtbaren Teil veröffentlicht werden. Andre richtet für die "internen" Protokolle eine Archivseite ein.

Aktueller Stand aller Working Groups

AP 1 Tool Registries

Tabelle

Die Tabelle soll für Anwender gut lesbar sein, daher benötigen wir eine weitere Spalte mit dem Namen "Benutzerschnittstelle" (Typ wurde als zu generisch verworfen). Hierbei ist gemeint, ob es sich um ein Tool mit GUI handelt, ein kommandobasiertes Tool oder eine Library - und wenn Library, dann ob es eine Java-Library ist oder eine Format Library etc. Natürlich können einige Dinge auch mehreres gleichzeitig sein, wie z. B. JHOVE ein Tool mit GUI ist, aber ebenfalls als Java-Library dienen kann. Der Benutzer sollte aber z. B. gleich erkennen können, ob und wie er das Tool mit seinem Hintergrund benutzen kann.

Dafür sollte die letzte Spalte (Tool-Übersicht) wegfallen. Tool-Suiten sollten besser in eine separate Liste.

Fragebogen für Erkennungsworkflow-Umfrage

Svenia stellt den Fragebogen vor. Sie hat sich auch nach einem Online-Werkzeug hierfür umgeschaut, hier sind die Möglichkeiten zahlreich und daher wurde noch keine endgültige Entscheidung getroffen. Da angedacht ist, den Fragebogen nicht nur einmal zu bewerben, sondern vielleicht sogar einmal im Jahr durchzuführen und Anpassungen vornehmen zu können, bietet sich hier ein Online-Tool an. Das ist flexibler und die Auswertung ist einfacher als z. B. bei einem PDF-Fragebogen.

Der Plan ist, möglichst wenig Freitext zu haben, um die Auswertung zu erleichtern.

Um nicht in den gleichen Gewässern zu fischen wie die OPF sollten wir den Bogen nur an deutschsprachige Institutionen verschicken und hier eher den Dialog mit der OPF suchen, um Ergebnisse zu vergleichen und ggf. auch die OPF dabei zu unterstützen, mehr Rücklauf auch im internationalen Raum zu erhalten. Eine persönliche Ansprache erzeugt doch sehr viel mehr Rücklauf als eine Rundmail mit der Bitte um das Ausfüllen eines Fragebogens. Ggf. ist das sogar mit Confluence möglich, die Frage an die nestor Geschäftsstelle wurde gestellt (Update:

Das wurde inzwischen verworfen und andere Möglichkeiten werden getestet, siehe Protokoll des 6. Treffens).

Aktionen wie der Fragebogen sollen unter dem mission statement der AG als "laufende Aktivitäten der AG" festgehalten werden.

Der Fragebogen soll als Entwurf kurz vor dem nächsten Treffen im September bereitgestellt werden.

AP 2 Testsuites (*André*)

Andre und Heinz hatten eine praktische Aufgabe mit einer konkreten Datei (ist im internen wiki zu finden). Beiden ist es auf zwei unterschiedlichen Wegen gelungen, den Inhalt zu rekonstruieren, obwohl die Datei auf den ersten Blick obsolet erschien. Andre plant einen Beitrag im wiki über die beiden Wege und die Datei kann dabei auch im internen wiki veröffentlicht bzw. extern zur Gesis verlinkt werden (Datei "hirdina" in der Tabelle).

Weg Heinz: das ursprüngliche System in der Virtuellen Maschine laufen lassen und die Datei als rtf gespeichert.

Weg Andre: Datei "nackt" angeschaut, eine magic number repariert (2 Bits waren gekippt) und als zip gespeichert, dann als xml geöffnet.

AP 3 Register und Kontaktbörse für obsoleszente Speichertechniken (»R.O.S.T.«) (*Heinz*)

Am Anfang der Formaterkennung steht ein erfolgreicher Lesevorgang - dies beginnt also mit der Hardware, dem Lesegerät.

Heinz stellt den von ihm entwickelten Fragebogen vor, der innerhalb des deutschsprachigen Raums verbreitet werden soll und Institutionen dahingehend befragt, welche Lesegeräte sie haben, welche Datenträger sie haben und ob sie ihre Lesegeräte ggf. auch für Externe zur Verfügung stellen würden. Ziel ist, das Register und die Kontaktbörse aufzubauen und auszubauen. Heinz stellt den Fragebogen noch ins wiki. Andre wird den Fragebogen in ein besser auswertbares Format umwandeln.

Das Keep-Projekt hatte international mal etwas Ähnliches vor, sie sind aber leider nicht über das Framework hinausgekommen.

Heinz stellt sich hier einen pragmatischen Ansatz vor, einfach einen ersten Wurf, um einen Überblick zu bekommen.

Das LA BW hatte z. B. mal ein Problem mit Dat-Kassetten.

AP 4 Format Registries (*Martin*)

Martin hat begonnen, eine ausführlichere deutschsprachige Anleitung zu erstellen. Es gibt auch eine kurze Einführung, die auch nicht so sehr technisch vorgebildete Leserinnen und Leser abholt. Die Hürde bei PRONOM etwas einzureichen ist nicht so hoch wie man nach der Anleitung von Jay oder gar Adrian annehmen mag - ein Sample aus mehreren Quellen ist z.

B. nicht nötig, wenn die Datei-Spezifikation bekannt ist oder man die Patterns aus dem Programm File auslesen kann. Ggf. ist es auch möglich, nur anhand der Extension das Dateiformat zu bestimmen, sofern man aufgrund eines zu kleinen Samples die magic number leider nicht mit Sicherheit bestimmen kann - dies muss man aber noch in der Praxis testen, ob die TNA sich darauf einließe.

AP 5 Formatvalidierung (*Stefan*)

Sina nimmt leider nicht mehr teil, Rainer stößt aber gern als Unterstützung dazu.

Stefan stellt zur Diskussion, ob das AP in andere aufgehen sollte oder man besser hier eine Art weiterführendes Paket aufsetzend auf AP1 (dann nur für die Validierungstools) erstellt und den Scope schärft. Die AG spricht sich für die Schärfung des Scope aus. Die ersten vier Arbeitspakete haben eher einen Listencharakter, lediglich AP5 und AP6 erfordern tiefergehende inhaltliche Arbeit. Eine Zuarbeit zu OPF ist auch von beiden Seiten erwünscht. Man könnte auch Blog-Einträge auf der Seite der OPF erstellen.

Wir planen eine Telefonkonferenz, nachdem wir alle ein wenig über das Profil des AP5 nachgedacht haben. Einige lose Ideen wurden bei der Sitzung laut. So z. B. die Idee zu erforschen (oder zu verstehen), wie Validierung eigentlich funktioniert. Die DNB hat versucht einen mpeg Validator zu bauen, dies war aber zu komplex. Für epub war dies möglich, auch wenn der DNB epub-Checker eher rudimentär lediglich das Paket auspackt, schaut ob alles drin ist was hinein gehört und das dazugehörige xml prüft. Andere epub-Checker machen mehr oder auch weniger. Es ist ein wenig eine Policy-Frage wie viel man checkt. Michelle führt auch die Frage der Policy ein wenig aus, dies ist zurzeit auch ein Thema für veraPDF. Es kann institutionelle Policies geben, welche Dateien man in welchem Zustand akzeptiert und hier gibt es diverse Abstufungsmöglichkeiten.

AP 6 Best Practices in der Identifizierung

Heinz stellt NSRL kurz vor. Obwohl beispielsweise bei dem Kittler-Nachlass die Software sehr oft customized war, konnten selber da 30% der Dateien als nicht unikal identifiziert werden, was eine große erste Orientierungshilfe darstellte.

Stefan und Michelle erwähnen das Projekt "Presenter", das ein von den teilnehmenden Institutionen (NLNZ, BL, DNB, KLB und die australische Nationalbibliothek) selber finanziertes Nachfolgeprojekt zum abgelehnten Horizon 2020 Projekt "Treasures" ist und sich wieder um Formatidentifizierung dreht. Stefan fragt Tobias (DNB) nach Einzelheiten und Michelle fragt im Juni Steve Knight (NLNZ).

Konrad und Christine wünschen sich ein Tool, das erst die Identifizierung erledigt und je nach Ergebnis einen passenden Validator benutzt, sofern es einen gibt. Martin berichtet, dass es am LA NRW so einen Workflow gibt. Rosetta, Preservica und Archivematica arbeiten ebenfalls so, hier handelt es sich aber jeweils um Workflowleistungen des Archivierungssystems, das die Tools in Abhängigkeit voneinander einsetzt - in der Regel erst DROID und dann JHOVE mit entsprechend vorausgewähltem Modul. Sofern für jede Datei einfach jedes Modul angesprochen wird, kommt es zu starken Performancebelastungen. Die DNB mit einem Ingestvolumen von 1.500 - 10.000 Dateien täglich hat hier leider entsprechende Erfahrungen.

Claire hat die Frage gestellt, ob die Community beim Entwickeln eines Tools wie von ihr gewünscht unterstützen würde. Sie stellt sich ein Tool vor, das vom Groben ins Feine arbeitet,

z. B. erst einmal die Formatfamilie feststellt, beispielsweise PDF und dann erst untersucht, um welches PDF es sich handelt. Sofern die genaue PDF-Version dann nicht festgestellt werden kann, hat man wenigstens das Ergebnis, dass es sich überhaupt um ein PDF handelt. DROID, Siegfried und überhaupt jene auf PRONOM aufbauende Tools handeln zurzeit nach dem Alles-oder-Nichts-Prinzip.

Die AG-Mitglieder bezweifeln, dass Entwicklungsarbeit Aufgabe der AG sein kann, würden aber gern in einer Telko mit Claire besprechen, was sie sich eigentlich vorstellt. Martin, Andre, Michelle, Stefan und Yvonne möchten sich hierzu mit Claire zu einer verabreden.

Die Beschreibungen der Arbeitspakete können größtenteils so veröffentlicht werden. So z. B. AP 3 und AP 4.

nestor AG Formaterkennung und die OPF

Yvonne berichtet von der OPF Document Interest Group. Die OPF wünscht sich, dass wir veraPDF testen und am besten gleich in Github Bugs reporten. Die JHOVE Fehlermeldungen werden von der OPF noch in Github übertragen, dort können wir dann auch weiterhin zuarbeiten.

Öffentlichkeitsarbeit der nestor AG

Es gibt selten auf Zuruf aus der Community direkte Fragen an die AG, aber es kommt vor. So z. B. gab es eine Frage zum GEDCom Format in 2016 und in 2015 gab es die Frage zum Thema Persistenz von PUIDs. Insgesamt ist die AG und ihre Arbeit aber noch nicht übertrieben gut sichtbar. Daher wurden mehrere Vorschläge zur besseren Öffentlichkeitsarbeit gemacht:

- Kurzartikel für die nestor Seite. Stefan macht den Anfang mit Qualitätsmanagement-Tools. Yvonne würde mit Preservation Planning fortsetzen.
- Für den diesjährigen Praktikertag können wir nicht mehr bekommen als eine kurze Zusammenfassung unserer Arbeit sowie alle AGs sie haben. In 2017 wäre ein Vortrag denkbar. Sabrina wurde gefragt, ob vielleicht sogar ein zweistündiger Workshop denkbar ist (Update: Workshop kann sogar halbtägig gemacht werden)
- für die iPRES in Bern würde ein Paper eingereicht. Ergebnis gibt es bis Ende Mai (Update: Paper wurde leider nicht angenommen)
- Für den Bibliothekartag 2017 in Frankfurt würde Yvonne gern etwas einreichen
- Die Archiving müsste 2017 auch wieder in Europa sein
- AUdS 2017 in Basel? (s. http://www.staatsarchiv.sg.ch/home/auds.html)
- Digitale Bibliothek in Graz ca. im Februar 2017
- Österreichischer Bibliothekartag
- OPF Blog (Michelle)

Außerdem stellte Yvonne die Frage, ob unser Ziel zu schwammig sei, da für die nestor AG kein Endziel und somit auch keine Deadline angegeben ist. Wir möchten auch weiterhin kein Ende der AG haben, sondern weiterhin eine fortlaufende nestor AG bilden, setzen uns aber Meilensteine und Zwischenziele wie nun die Umfrage, den Ausbau des Registers, Artikel und Blogs.