# Sitzungsprotokoll

Sitzung am: 12.05.2015 um: 9.00Uhr

Ort: Büro Manuel Burghart

Teilnehmer: Manuel Burghart, Miriam Nickl, Katia Buchhop, Thomas Schmidt, Florian Fuchs

Thema: Zweites Stakeholdermeeting: Detailinformationen zu den Datenbeständen

Fragen:

-Wie sind die Daten (Metadaten) in der Universitätsbibliothek strukturiert?

-Welche Daten liegen vor?

-Wie können sie verarbeitet werden?

Verlauf:

-Link für das Regensburger Volksmusik-Portal:

//Bestehender Projektkontxt

-Literatur: Vgl Bibliogrphie-Dokument (OCR/ ONR (für Ausblick Music Encoding)

-Neue Infos von Schupfner:

- Archiv: Kartons mit Liedblättern und Schrank mit Karteikarten 🡪 Nummern auf Karte und Blatt

🡪 Sangesort, Titel, etc.

* Karteikarten Metadaten in Augias digitalisiert/ eingepflegt 🡪 Relationale Datenbank für den Import der Metadaten oder auch Exceltabellen.
* Excel Dokument mit Metadaten von den Kulturwissenschaftlern
* Für uns wichtig: A-Sammlung (Nr. 0 bis 200 000) 🡪 insg aber nur 28 000 Blätter
* Blätter haben meist schon Signaturen (Regensburger Verbunds Nummer) 🡪 Frau Steinmaus nach der genauen Menge der Blätter fragen. (oder auch in der Tabelle zu ermitteln)
* Wurden alle Karteikarten erfasst?
* 🡪 Wir sollten aber eher mit Pilotkorpus arbeiten

Tabelle:

* Findbuch (analog od digital): Ausdruck für Informationretrival (Enthält Metadaten zur Sammlung)
* Signatur: Stehen auch auf dem Blatt. Kennzeichnen Sammlung und Nummer darin.
* ADV/SIDM (erhalten am): Wann wurde das Blatt in die Sammlung aufgenommen? (Für uns nicht so wichtig)
* V.-Num: vermutlich nur Interne, fortlaufende Nummer
* Fehlt weil: Angabe der nicht vorhandenen Blätter
* Herkunft: Hier trivial… (Oder gibt es in dieser Sammlung auch nichtdeutschprachige Lieder?)
* Titel: Titel des Liedes
* Liedgattung: Hat in dieser Tabelle keinen Eintra… ?
* Umfasst: Meist Text und Melodie (Untersuchen ob es bei allen so ist/ Abweichungen?)
* Icipit: Erster Satz des Lieds
* Umfang: Anzahl der Seiten
* Sangesort: Wo wurde das Lied gehört/ transkribiert (Untersuchen/ Vergleichen mit Normdatenbanken: Genormte Schreibweisen für Namen (Personen/ Orte)) 🡪 Schreibweisen auf einen Nenner bringen. Damit bei der Suche alles gefunden werden kann (GND-Normdatenbank)
* Archivort: Wo kommt das Dokument her?
* **Nächster Arbeitsschritt: Tabelle durchzählen und Statistiken über die Daten**
* **Welche Informationen sind wichtig/ interessant? Was sollte später abfragbar/ anzeigbar sein???**
* 19 Tabellen: Organisiert in 10 000 Einheiten
* Testkorpus eigenständig einscannen. (Einzugsscanner: Wird später auch für alle Dokumente verwendet)
  + Einstellungen für OCR optimieren.
* Fürs Projekt auch über den Scanner schreiben. Funktionsweise, Einstellungen

Volkslieder werden in unterschiedliche Gattungen unterteilt. A-Sammlung größtenteils Volks- und Kinderlieder

Sammlung wurde im 3. Reich mussbraucht um Propaganda zu betreiben…

Nicht alle/Nur wenige Blätter sind von Hoerburger. Wurden nur in der Sammlung zusammengefasst…

-Scope-Erweiterung:

-4 Schwerpunkte (unter uns aufteilen)

- OCR/ONR (Welche gibt es? Wie gut sind sie?)

- Datenstruktur (Exceltabellen untersuchen)

-Plattform (Frameworks für die Annotation und Bearbeitung suchen?)

* Anforderungen:
  + InfoRetrivalSys: Metadaten, Volltexte (digitalisiert) 🡪 Suche/ Filtern nach Ort/ Zeit) (🡪 Must Have)
  + Nice to Have: Suche nach Melodie?
  + Digitla im Browser abspielen
  + Effizienz des OCR sollte genau untersucht und dokumentiert werden 🡪 Was muss dann durch Crowdsourcing korrigiert werden?
  + Workflow für das Bearbeiten optimieren. (Wie kann man große Mengen an Blättern gut digitalisieren/ einpflegen?

**Workflow**: Einscannen🡪 OCR 🡪 Metadaten verknüpfen 🡪 Ergebnisse manuell Prüfen 🡪 Tool für Crowdsourcing um Fehler des OCR auszubessern

* Vorsicht bei automatisch generierten Music-XML Tags! (Korrektheit überprüfen)
* Visualisierung der Metadaten: Dashboard für zusammengefasste Metadaten. (Überblick) 🡪 Explorative Suchergebnisse (Häufigste Wörter visualisieren. Geolocation der Sangesorte anzeigen etc. etc. etc.)

Schritt 1: Erste Datenanalyse: Exceltabelle auswerten. Welche Daten sind in welchem Umfang vorhanden/interessant (🡪 Experten fragen, was interessiert (Zielgruppeninterview) )? 🡪 Mit welchen Daten sollte gearbeitet werden? Was sollte visualisiert/ angezeigt/ erfasst werden?

Findbuch als State oft he Art. Was ist gut. Was kann verbessert werden?

Wichtige Frage: Wie bringt man Scan und Metadaten zusammen? Für den Workflow eine praktikable Benennungsmethode finden. (Kann man die Nummer automatisch erkennen? (Kann man Erste Textzeile (digitalisiert in den Metadaten) und Erste Textzeile im Scan (OCR) abgleichen.

Schritt 2: Bald Termin mit Herrn Knüttel(?)

Debrifing