# Sitzungsprotokoll

Sitzung am: 6.5.2015 um: 13.00Uhr

Ort: Büro Manuel Burghardt

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

Thema: Erstes Stakeholdermeeting: Konkretisierung der Projektaufgabe

Fragen:

-Vorstellungen zur Aufgabe

-Welche Vorarbeit gibt es?

-Zielgruppe? Wer soll die Anwendung nutzen?

-Gibt es andere Experten? (Ansprechpartner?)

-Programmiersprache? Server??

-GitHub ok? - Will er dem Repository auch beitreten? Oder existiert bereits eines?

-Wie sollte die Usabilitystudie aussehen?

-Gibt es relevante Forschungsliteratur?

-Wie sollen die Liedblätter annotiert werden?

-Nach was/ wie sollen sie durchsucht werden können?

-Anforderungsanalyse; Prototyp?

Verlauf:

**Backgroundinformationen über das Projekt und die zu digitalisierende Liedblattsammlung:**

Der Musikethnologe Hoerburger hat Weltweit Lieder gesammelt und aufgeschrieben.

Liedblätter liegen auf Papier vor.

Gelagert in der Bibliothek. Kontaktpersonen zu dem Projekt in der Bibliothek: (Fr Steinmaus versieht mit Signaturen; Herr Schröder ist Leiter; Herr Schupfner ist für IT zuständig und als Ansprechpartner für die Metadaten)

Kommunikation mit der Bibliothek wird über Herrn Burghardt koordiniert.

Blätter wurden bereits kulturwissenschaftlich aufbereitet.

Metadaten (Metadaten: Datum, Herkunft, „Sangesort“, Was ist dabei? Text? Melodie? Beides? ; fortlaufende Nummer 🡪 sind vorhanden (mittels AUGIAS wurden sie digitalisiert)) zu den Liedblättern liegen auf Karteikarten vor

Dfg-Projekt hat bereits aufbereitet. 🡪 liegen digital vor

Liedblätter sind ein Teilaspekt eines größeren Gesamtprojekts zur Erfassung der Hoerburger-Dokumente.

Nicht alle Liedblätter sind von Hoerburger.

Es sind meist Volksmusiklieder = Einstimmige Lieder (monophon)

Insgesamt 250000 verschiedene Liedblätter.

Nur ein Bruchteil der Blätter liegen in der Regensburger Universitätsbibliothek.

**Ziel**: Digitalisieren der Liedblätter.

Die Blätter selbst können via Batchscanner schnell gescannt werden. (Einer oder zwei Projektgruppenmitglieder können beim Scannen dabei sein wird aber von Bibliothek durchgeführt)

**Konkrete Projektaufgabe:**

Verknüpfen von Scans und Metadaten. Zu diesem Zweck muss ein geeignetes Format für die Daten überlegen werden!

Ultimatives Ziel: Texte und Noten maschinenlesbar digitalisieren.

Texte via OCR (Optical Character Recognition) digitalisieren 🡪 Eine OCR-Software steht in der Bibliothek zur Verfügung… (Character Recognition wurde bei Notenerkennung noch nicht in Betracht gezogen)

Noten via ONR digitalisieren? 🡪 Herr Burghardt hat schon in vorherigen Tests versucht diese mittels diesem System zu testen 🡪 der Erfolg ist ausgeblieben 🡪 Genauigkeit 50-60 %

Fragestellung 1: Wie zuverlässig arbeitet die Software für verschiedene Textarten (Schreibmaschine/ Handschrift)? (Vermutung: ok für Schreibmaschine, jedoch problematisch bei handschriftlichen Passagen)

Fragestellung 2: Wie beeinflussen Eigenschaften der Bilder wie z.B. Kontrast und Helligkeit usw. die Qualität der Erkennung? Wie sollten diese eingestellt werden um eine möglichst gute Erkennung zu ermöglichen?

Fragestellung 3: Gibt es alternative OCR-Software, die besser arbeitet oder für unsere Zwecke besser geeignet ist?

**Vorgehen:**

Anhand eines exemplarischen Korpus OCR-Systeme evaluieren. (Kleiner Korpus. Verschiedene System und Einstellungen werden daran getestet. Ergebnisse manuell/ intellektuell überprüft)

Analog für die Erkennung der Noten: ONR(Optical Note(?) Recognition)/OMR(Optical Music Recognition) für die Notenerkennung 🡪 Existierende Tools an Testkorpus vergleichen.

Theorie: Erkennung nicht sehr präzise (>50%) 🡪 möglicher Lösungsweg: Crowdsourcing zur korrekten Annotation von Noten und Handschriften einsetzen.

Für die Datenerhebung via Crowdsourcing ein Tool erstellen.

**Unsere Ziele:**

1. Pilotkorpus: Evaluieren welche Tools wie gut sind.
2. Zusammenführen der Formate/Daten/
3. Konzeptionelle Arbeit: Datenbank
4. Frontend für die Annotation durch Experten (Webinterface)

Frontend: Hier sollen die Nutzer die Liedblätter sortieren und auch nach einem Suchen können oder nach Jahren filtern können.

**Vorbereitende Fragen:**

-Wie viele Blätter/ Tools sollen bei der Evaluation eingesetzt werden? 🡪 Testkorpus: ca. max 100 Blätter

-Welche OCRs/ONRs gibt es? Wie ist deren Verfügbarkeit? 🡪 Die Bibliothek verfügt über einige oder hat Budget um benötigte Software zu beschaffen.

- Handschrift Erkennung: Kontrast einstellen um alle Text bestmöglich eruieren zu können.

Für weitere Informationen über das Projekt s.h. auch Abstract von Herrn Burghardt 🡪 Hier auch Related work mit einbeziehen.

**Nächste Schritte:**

Schritt 1: Scanvorgang mit Dokumentenbenennung usw. konzipieren.

Schritt 2: Literaturrecherche zum Thema und Methoden 🡪 Literatur: Vgl Paper Quelle 1 und 2

Schritt 3: OCR und ONR Tools heraussuchen. (Auch speziell Handschriftenerkennung)

**Zwischenziel:**

Crowdsourcing im Browser mit auditiver Qualitätskontrolle 🡨 Empirisch untermauern wann diese benötigt (weil automatische Erkennung versagt)

## Debriefing

Nachbesprechung des ersten Stakeholdermeetings und Aufgabenverteilung:

**Anstehende Aufgaben:**  
-Anforderungsdokument erstellen

-Tools suchen

-Literatur suchen (Gruppenmitglieder aufteilen nach OCR/ONR/Handschriften)

🡪 Wie kann man bei den Vergleichen statistisch vorgehen?

🡪 Gibt es aktuelle Literatur zum Vergleich von OCR/ONR Software?

🡪 …

-Qualitätskriterien für Richtigkeit der Erkennung festlegen

-Tools testen (an kleinem vorab Testkorpus (ca. 30))

-Ideen für das sinnvolle Zusammenführen von Scans und Metadaten

-Evaluation der Tools mit Evaluationskorpus

**Festlegungen**:

Dropboxordner mit Literatur. Ordnerstruktur nach Themen

Literaturbennenung in der Dropbox: Autor, Jahr, Titel

Verwendung von BitBucket für das gemeinsame Arbeiten an Ergebnissen und Code (Da private Repositories hier möglich) 🡪 Dozenten sollen Zugang zu Repository bekommen.