# Datencharakteristika

Art, Struktur, Größe und Verteilung der Daten.

Accuracy and precision

Validity

Consistency

Reliability

Velocity

Volume

Variety

Relevance

Accessibility

Timeliness

Completeness

Veracity

## WBS

Natürlich, hier ist eine mögliche Work Breakdown Structure (WBS) für das NLP-Projekt zur Konvertierung eines Literaturverzeichnisses in BibTeX:

1. Projektinitiierung

- Projektdefinition und Zielsetzung festlegen

- Ressourcenplanung und Teamzusammensetzung

2. Anforderungsanalyse

- Literaturverzeichnisformate analysieren (z.B. APA, MLA) -> Es gibt verschiedene BibTex Styles! Alle haben zwar gleiche BibTex, aber das Einlesen vor allem fur OCR anders

- BibTeX-Spezifikationen verstehen

- Benutzeranforderungen sammeln

3. Datensammlung und -vorverarbeitung

- Sammlung von Beispieldaten (Literaturverzeichnisse)

- Datenbereinigung und -formatierung

- Extrahierung von Metadaten (Titel, Autor, Jahr usw.)

4. NLP-Modellentwicklung

- Auswahl eines NLP-Modells (z.B. Named Entity Recognition)

- Training des Modells mit den Beispieldaten

- Evaluation der Modellleistung

5. Implementierung der Konvertierungsfunktion

- Entwicklung des Algorithmus zur Konvertierung von Metadaten in BibTeX-Format

- Implementierung einer Benutzerschnittstelle (optional)

6. Testing und Qualitätssicherung

- Funktionstests des Konvertierungsalgorithmus

- Benutzertests durchführen

- Fehlerbehebung und Optimierung

7. Dokumentation und Abschluss

- Erstellen von Projektdokumentationen (Benutzerhandbuch, Entwicklerdokumentation)

- Abschlussbericht erstellen

- Projektabschlussbesprechung und Übergabe an den Kunden

Das ist eine grundlegende WBS und kann je nach den spezifischen Anforderungen und dem Umfang des Projekts angepasst werden.

**Example 3: Software Product Development:**

1. Project Management
   1. Requirements Gathering
   2. Project Planning
   3. Resource Planning
2. Product Design
   1. UX Design
   2. UI Design
   3. Prototyping
3. Development
   1. Front-end Development
   2. Back-end Development
   3. API Development
   4. Database Design
4. Quality Assurance
   1. Testing
   2. Bug Fixing
   3. User Acceptance Testing
5. Deployment and Maintenance
   1. Deployment Planning
   2. Release Management
   3. Maintenance and Support

# Meeting 06.05.24, Projektplan

## Grobe Lösungsskizze

Was kann unser Endprodukt? Welche Technologien setzen wir ein? Wie kommen wir dahin?

* Online-Suche zur Vervollständigung von Abkürzungen? Aufbau einer Datenbank mit CiteDev? Datenstruktur aufbauen?
* OCR-Schnittstelle? -> Task: OCR-Tool suchen
* Datenspeicher: Git oder Cloud (DVC)
* Welche Bibliotheken anbinden,
* Anaconda
* Grobe wie Beschreibung wie: Tokenisierung des Strings bzw. Repräsentation der Ein- und Ausgabedaten, Pipeline -> auch als Task
* Aufbau / Definition einer einheitlichen Umgebung
* Visualisierung der Ergebnisse?
* Zielbild? Wie soll es für den Anwender aussehen?

## Daten

* Verzeichnis mit verschiedenen BibTex-Styles durch Latex generieren und dann als Trainingsdaten -> Bias?
* Daten Repräsentativ?
* Datenqualität wie messen?

## WBS

### Datensammlung und -vorverarbeitung

* Datenquellen bestimmen
* Sammlung von Beispieldaten (Literaturverzeichnisse)
  + Stichprobenumfang?
  + Welche Styles?
  + Ablage der Beispieldaten wo? Wie erfolgt der Zugriff?
* Datenbereinigung und -formatierung
* Extrahierung von Metadaten (Titel, Autor, Jahr usw.)

Evaluation

Metrik definieren (Ähnlichkeitsmaß) für Testdaten