Ergebnisprotokoll „3_01_Automatic_Diploma_Digitalisation“

1. Ausgangssituation

Mit dem Gesetz zur Verbesserung des Onlinezugangs zu Verwaltungsleistungen (Onlinezugangsgesetz– OZG) sind die Hochschulen (unter anderen) verpflichtet, Verwaltungsleistungen auch digital anzubieten. Hier schließt das Projekt „Automatic Diploma Digitalisation“ an. Ziel ist es, die Digitalisierung von Zeugnissen deutlich effizienter zu gestalten. Dazu soll eine Applikation entwickelt werden, die es ermöglicht Bild- oder PDF-Dateien von Zeugnissen auszulesen und die wichtigsten Daten in ein XHochschule-konformes XML-Format zu übermitteln.

2. Zielsetzung

Vision

In Vollendung soll die Applikation dazu in der Lage sein, Scans oder Fotografien von Zeugnissen aus dem In- und Ausland auszulesen, wichtige Daten zu erkennen und diese in ein XML-Format zu überführen. Dazu sollen die folgenden Informationen ausgelesen werden können:

Attribut	Ausprägungen
Name Studierender	Titel, Vorname 1, 2, 3, Nachname
Gesamtnote	dezimal
Hochschulname	Codeliste / String?
Geburtsdatum	Festes Datumsformat
Studiengang	Text
Art des Abschlusses	Bachelor, Master, Diplom, Staatsexamen, Magister, o.A.
Datum der Ausstellung	Festes Datumsformat
Ort der Ausstellung	Bekannte Orte
Ggfs. Prädikat	Text

Die Vorteile liegen auf der Hand:

- Studierende müssen die Daten nicht eintippen, sondern können bequem ein Foto oder PDF hochladen
  → Minimierung von Tippfehlern
  → Komfort für Studierende
- Die Hochschulen bekommen sowohl eine Abbildung des Zeugnisses als auch direkt die benötigten Daten im passenden Format
  → Weniger Arbeitsschritte
  → Erleichterter Austausch zwischen Hochschulen

Das Minimal Viable Product (MVP)

Da die Entwicklung einer solchen Software sowohl kosten- als auch zeitintensiv ist, sollte sich für ein Minimal Viable Product (MVP) zunächst ausschließlich auf deutsche Hochschulen beschränkt werden. Dies hat unter anderem den Vorteil, dass die Zeugnisse fast ausschließlich in römischer Schrift ausgestellt werden.

Des Weiteren sollte sich zunächst auf einige „Musterhochschulen“ beschränkt werden. Denn eine Hürde in der Texterkennung liegt darin, in designtechnisch unterschiedlich aufgebauten Zeugnissen die richtigen Informationen zu finden.

3. Stakeholder

Damit die Applikation auch an die Bedürfnisse der Anwender*innen angepasst ist, müssen zunächst die zentralen Zielgruppen identifiziert werden:

- (Ehemalige) Studierende
  - Nationale Studierende (MVP)
  - Internationale Studierende (Vollkommenes Produkt)
- Hochschulen (MVP)
  - Prüfungsämter
  - Studierendensekretariate und -services
  - International Offices

Darüber hinaus gelten auch die Hochschul-IT, die für die Implementierung verantwortlich ist, das Projekt XHochschule sowie Bund und Länder als wichtige Stakeholder.

4. Scope

Die Applikation soll als Plugin mit serverseitiger Verarbeitung in die bestehenden Bewerbungsportale integriert werden können. Dabei soll die Applikation „nur“ die Informationen aus Bild- oder PDF-Dateien erkennen und konvertieren können. Eine Validierung der Daten findet nicht statt. Des Weiteren soll keine Datenbereinigung durch die Applikation vorgenommen werden, die beispielsweise automatisch aus „Uni“ das Wort „Universität“ macht.

5. Prozess

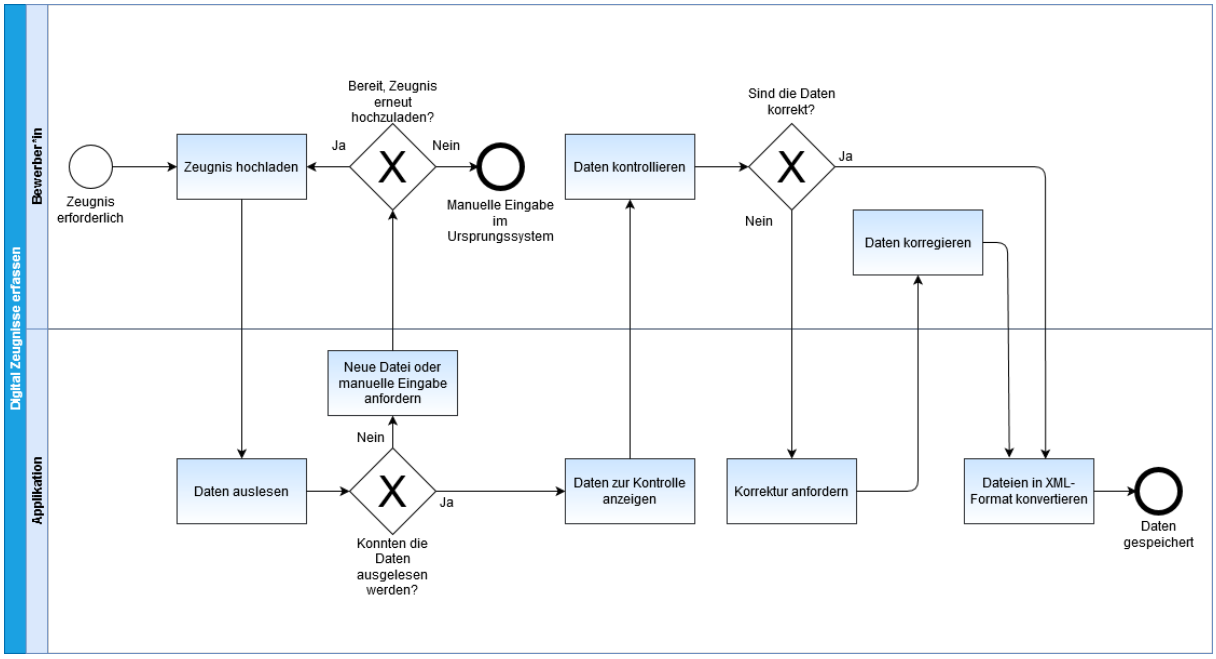

ABBILDUNG 1: DARSTELLUNG DES PROZESSES „DIGITAL ZEUGNISSE ERFASSEN“

6. Prototyp und erste Erkenntnisse

Im Rahmen des Hackathons „#Semesterhacks 2.0“ des Hochschulforum Digitalisierung am 12. Und 13. November 2020 wurden durch das Team bereits zwei Prototypen entwickelt:

1. Eine Demo für die grafische Oberfläche, wie die Bewerber*innen die Applikation später nutzen könnten.
2. Ein erstes Programm, das in der Lage ist, Informationen aus Bildern von Musterzeugnissen auszulesen.

Zur besseren Verständlichkeit wurde ein Video mit dem Titel „Automatische Zeugniserkennung – Prototyp-Demo“[1] erstellt. In diesem wird zunächst die Nutzung des Programms an einem sogenannten „Clickdummy“ demonstriert. Anschließend erfolgt eine Demonstration der Texterkennung und Einordnung anhand eines funktionierenden Prototyps.

Fortschritt

Der Prototyp wurde unter Verwendung der open-Source Software Tesseract[2] geschrieben. Dieses Texterkennungsprogramm wurde von Google entwickelt, steht aber unter einer Apache 2.0 Lizenz zur freien Nutzung zur Verfügung[3]. Mit der im Video aufgezeigten Version ist es möglich, Daten aus zwei Beispiel-Zeugnissen zu extrahieren und diese als „Dictionary“-Objekt zurückzugeben. Eine Weiterverarbeitung der Daten in den XHochschule XML-Standard[4] sollte in einem späteren Prozessschritt erfolgen.

Erkenntnisse

Der Prototyp zeigt, dass das Programm technisch implementierbar ist und mit angemessener Vorbereitung die gewünschten Ergebnisse bringt. Die verwendete Technik erlaubt potentiell große Zeiteinsparungen bei Studierenden und bei der Hochschulverwaltung. Das erarbeitete Skript wird allerdings nur bei Zeugnissen funktionieren, die nach einem gewissen Schema aufgebaut sind. Um das Programm auf andere Zeugnisse anwenden zu können, müsste die Lesefunktion um weitere Suchmuster und idealerweise semantische Plausibilitätsprüfung der Daten erweitert werden.

7. Ausblick

In ersten Versuchen hat sich gezeigt, dass eine Umsetzung der Idee durchaus Potenzial hat. Eine Entwicklung eines Minium Viable Products ist zu empfehlen. Darüber hinaus sollten die folgenden Fragen und Weiterentwicklungsmöglichkeiten bedacht werden:

- Wie kann die Richtigkeit der Daten geprüft werden? Wie können die Daten automatisch validiert werden?
- Die Plausibilität der Daten muss geprüft werden, indem man für jedes Feld einen akzeptablen Wertebereich zuordnet. Dabei könnte auch eine Rechtschreibeprüfung implementiert werden.
- Die Einbindung des Programms in XHochschule, PIM oder ähnliche Projekte, um auf Dauer Datenbanken für (ehemalige) Studierende aufzubauen
- Korrekturen, die von Studierenden durch manuelle Eingabe eingebracht werden, könnten als Basis zum automatischen Lernen der Schemata von neuen Zeugnissen benutzt werden – auf

[1] https://youtu.be/Xv-DTuTHUXI
[2] https://github.com/tesseract-ocr/tesseract
[3] https://github.com/tesseract-ocr/tesseract/blob/master/LICENSE
[4] http://xhochschule.de/def/xhochschule/0.0.05/xsd/

diese Weise könnte von der Notwendigkeit abgesehen werden, jedes Zeugnis manuell zu validieren.

8. Quellen

Bilder im Video

https://www.freepik.com/photos/business >Business photo created by benzoix

Genutzte Programme:

https://colab.research.google.com/

https://www.diagrams.net/

https://www.figma.com/file/Yjc7pI3zwdocGKyfF75u2R/Digital-Diploma-Reader?node-id=3%3A131

Icons im Video

https://icons-for-free.com/complete+done+green+success+valid+icon-1320183462969251652/

https://www.flaticon.com/free-icon/close_845648" title="Kiranshastry">Kiranshastry from www.flaticon.com

Musik im Video

https://www.free-stock-music.com/mixaund-success.html

Unser Team

Robin Dietrich

Consultant bei]init[im Projekt XHochschule

robin.dietrich@init.de

Daniel Schmedes

Consultant bei]init[im Projekt XHochschule

danielsamuel.schmedes@init.de

Felix Bitterer

Mitarbeiter der FH Bielefeld im Projekt „Digital Mobil @ FH Bielefeld“

felix.bitterer@uni-bielefeld.de