Wie Urkunden im Langzeitvorhaben Formulae – Litterae – Chartae digitalisiert werden

Mit dem Ziel, Forschung zum formelhaften Schreiben im frühen Mittelalter zu erleichtern, stellt das Langzeitvorhaben Formulae – Litterae – Chartae neben der neuen Edition der frühmittelalterlichen Formelsammlungen auch retrodigitalisierte Sammlungen bereits publizierter Editionen frühmittelalterlicher Urkundenfonds zur Verfügung. Bisher hat das Projekt 14.392 Urkunden mit mehr als 2,6 Millionen Wörtern online gestellt, wovon 10.858 Urkunden mit fast 1,8 Millionen Wörtern frei verfügbar sind. Die übrigen Urkunden sind zwar für alle Benutzer durchsuchbar, können aber aus Copyrightgründen nur von den Projektmitgliedern eingesehen werden. Andere Benutzer erhalten einen bibliographischen Verweis auf die einschlägige Edition, in welcher sie die betreffende Urkunde finden können. Dieser Beitrag fasst den Digitalisierungsprozess des Projektes für Urkunden zusammen, um den Benutzern der Werkstatt einen Einblick in das aufwendige Vorgehen zu geben, das nötig ist, um zuverlässige Urkundentexte online zugänglich zu machen.

Der Digitalisierungsprozess des Langzeitvorhabens zeichnet sich durch eine regelmäßige Abfolge von immer gleichen Schritten aus, die alle aufeinander aufbauen. Liegt eine Edition vor, wird sie mithilfe eines OCR-Programms, wie z.B. ABBYY FineReader, eingescannt. Diese qualitativ hochwertigen Scans werden mit Optical Character Recognition (OCR) bearbeitet. Aus dem nun durchsuchbaren Scan werden im nächsten Schritt die Texte herauskopiert und in ein MS Word-Dokument eingefügt.

Dabei treten häufig die folgenden Probleme auf, die es zu lösen gilt:

1. Trotz sehr guter Scanqualität werden einige Buchstaben oder auch Buchstabenabfolgen oft falsch erkannt, so beispielsweise in → m, e → c, c → e, ti → h, h → b, m → rn, i/t → l etc. Diese fehlerhaften Umwandlungen müssen händisch beseitigt werden.
2. Manchmal wurden bei der OCR-Bearbeitung ganze Zeilen nicht gelesen, sodass diese auch nicht Teil des durchsuchbaren Dokumentes sind. Die fehlenden Worte und Zeilen müssen ebenfalls händisch in den Text eingefügt werden. Ebenso müssen nicht benötigte Textteile (z.B. Verweise auf dem Apparat) entfernt werden.
3. Texte, die in heute ungebräuchlichen Schrifttypen, wie beispielsweise in Fraktur, gesetzt sind, werden von den OCR-Programmen nur schlecht oder gar nicht erkannt. In diesem Fall erfolgt die gesamte Erfassung händisch – durch Abtippen und Abgleichen jedes einzelnen Buchstabens.

Nach Aufbereitung des gesamten Dokuments in MS Word, werden die einzelnen Urkunden in ein im Formulae-Projekt entwickeltes Urkundentemplate eingefügt. Dabei handelt es sich um zwei Tabellen in MS-Word, von denen die erste Informationen zur zugrundeliegenden Edition und die zweite die eigentlichen Texte samt weiterer Informationen (sogenannte Metadaten), wie beispielsweise den Seitenangaben, enthält. Sind alle Urkunden eingefügt, werden diese von

einer weiteren Person gegengelesen, kontrolliert und letzte Erkennungs- oder Übertragungsfehler bereinigt. Im Anschluss folgt die Transformation zu XML.

Dieser mehrstufige Transformationsprozess produziert die notwendige Ordner- und Dateistruktur: Der erste Schritt ist die Umwandlung des Word-Dokuments mithilfe von OxGarage (https://oxgarage.tei-c.org/), ein online Werkzeug von der Text Encoding Initiative (TEI), das unter anderem MS-Word (DOCX) Dokumente ins TEI-XML automatisch umwandelt. Dieser erste Schritt produziert ein XML-Dokument mit zwei Tabellen, die den beiden Tabellen aus der DOCX-Datei entsprechen. Die notwendigen Metadaten- und Textdateien werden aus diesem XML-Dokument mittels 4 XSLT-Skripten erzeugt, die unter anderem einen eindeutigen Identifikator für jeden Text produzieren und eindeutige Datierungsangaben zum ISO-Standard (z.B. 12.03.850 zum 0850-03-12) transformieren. Zudem erkennt die XSLT-Transformation, ob zwei Reihen in der Tabelle voneinander unabhängige Urkunden oder zwei verschiedene Versionen (z.B. aus verschiedenen Handschriften) bzw. zwei verschiedene Teile (z.B. ein Privileg mit späterer Bestätigung) derselben Urkunde sind. Diese Transformation produziert neben den Textdateien für jede Urkunde auch Metadatendateien, die Angaben zur Datierung, Herkunft, Bibliographie usw. enthalten. Eine weitere Metadatendatei beinhaltet den Titel des Corpus sowie Links zu den einzelnen Urkunden.

An diesem Punkt wären die Urkunden schon bereit, in unsere Werkstatt eingepflegt zu werden. Aber vor der Veröffentlichung gibt es noch zwei Nachprüfungsschritte. Zunächst müssen die ISO-Datumsangaben geprüft und manchmal händisch eingetragen werden. Wenn z.B. in der Urkundenedition nur der *terminus ante quem*, wie z.B. “vor 820”, angegeben wird, kann keine automatische Transformation erfolgen, weil unsere Suchmaschine auch den *terminus post quem* braucht. Daher muss eine Entscheidung über diesen aufgrund der Information in der Urkunde und in der Edition getroffen werden. Die Angabe “vor 820” könnte beispielsweise als `notBefore=“0800-01-01” notAfter=“0819-12-31”` in die XML-Datei eingetragen werden. Obwohl die Benutzer der Werkstatt diese Datumsangaben nie sehen, sind sie für die Datumssuche sehr wichtig.

Der letzte Schritt vor der Veröffentlichung ist die Prüfung der Texte online in unserer Entwicklungsumgebung. Hier können die Mitglieder des Projekts die Texte so sehen, wie sie später in der Werkstatt präsentiert werden, um Formatierungs- und Rechtschreibfehler einfacher zu erkennen. Nach diesem letzten Prüfungsschritt werden die Texte in die Werkstatt eingepflegt, wo sie dann für die Benutzer lesbar und durchsuchbar sind. Übrigens unterliegt jeder öffentlich verfügbare Urkundentext einer sehr offenen Creative Commons Namensnennungslizenz (https://creativecommons.org/licenses/by/3.0/de/), die es den Benutzern erlaubt, den Text weiter zu benutzen oder neu zu publizieren, solange sie das Langzeitvorhaben als Quelle des Textes nennen. Wir wünsche Ihnen viel Spaß beim Lesen, Forschen und Weiterbenutzen und stehen für weitere Fragen zur Verfügung.