**027.7 Zeitschrift für Bibliothekskultur / Journal for Library Culture •
Ausgabe 8(2)**

Faithful Transcriptions

Nicole Eichenberger¹, Hedwig Suwelack¹, Annika Schröer²

¹Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, ²Universitätsbibliothek Leipzig

Published on: Oct 20, 2021

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

Von Mai bis Juli 2021 richteten die Staatsbibliothek zu Berlin und die Universitätsbibliothek Leipzig den Transkribathon „Faithful Transcriptions“ aus, ein digitales Crowd-Sourcing-Projekt zu theologischen Handschriften des Mittelalters. Über 100 Teilnehmende fertigten dabei TEI/XML-codierte Transkriptionen von 181 Handschriftenseiten an. Als Arbeitsplattform diente der IIF-basierte Arbeitsplatz des aktuell in Entwicklung befindlichen Handschriftenportals. Dieser Beitrag berichtet über Konzeption und Durchführung der Veranstaltung und stellt die Transkribathon-Datensets vor.

From May to July 2021, the Berlin State Library and the Leipzig University Library organized the Transcribathon 'Faithful Transcriptions', a digital crowd-sourcing project on medieval theological manuscripts. More than 100 participants produced TEI/XML-encoded transcriptions of 181 manuscript pages. The IIF-based workspace of the Handschriftenportal, which is currently under development, served as a working platform. This article reports on the conception and implementation of the event and presents the Transcribathon data sets.

1. Konzeption und Ziel des Transkribathons

Vor dem Hintergrund aktueller Infrastrukturprojekte wie dem von der Deutschen Forschungsgemeinschaft geförderten „Handschriftenportal“¹ und der Einrichtung von Library Labs kommt dem Feld der Aufbereitung, Kuratierung und Präsentation digitaler Daten für Bibliotheken eine stets wachsende Bedeutung zu. Die genannten Plattformen bilden überdies eine Schnittstelle zwischen Bibliothek, Forschung und interessierter Öffentlichkeit, indem sie kooperatives Arbeiten an digitalen historischen Beständen sowie die Präsentation von nutzergeneriertem Content ermöglichen. ([Rapp, 2021](#)) Dies zeigt sich etwa in Crowd-Sourcing-Projekten, die in den letzten Jahren gerade auch für Kulturerbeeinrichtungen an Relevanz gewonnen haben. (([Waidmann, 2020, 126](#)); [Owens, 2014](#))

Zugleich hat die digitale Transformation auch die geisteswissenschaftliche und insbesondere die mediävistische Forschungspraxis grundsätzlich verändert. ([Burrichter, Gebert, Mackert, & Viehhauser, 2021](#)) So werden die als Quellenmaterial verfügbaren digitalisierten historischen Bestände immer umfangreicher und die Methoden, sie mit computergestützten Verfahren zu erschließen und zu analysieren, immer differenzierter.

Dennoch gibt es noch deutliche Unterschiede in der Forschungstiefe und der praktischen Einsetzbarkeit der informatischen Methoden, etwa bezüglich der Medienarten und der historischen Epochen: Während die Optical Character Recognition (OCR)-Verfahren zur automatischen Texterkennung zum Beispiel für historische Drucke schon sehr gut funktionieren ([Engl, 2020](#)), ist die automatische Erkennung von – insbesondere vormodernen – Handschriften immer noch eine technische Herausforderung, nicht zuletzt auch deshalb, weil das Erstellen von Ground Truth- und Trainingsdaten in diesem Bereich deutlich aufwändiger ist und mehr fachliche Expertise erfordert als bei Druckschriften.

Dieses Forschungsdesiderat sowie die Möglichkeit, als Bibliothek in einem Projekt zu historischen Beständen nicht nur als Datengeberin zu agieren, sondern einen virtuellen Raum zu schaffen, in dem Studierende, Forschende und andere Interessierte gemeinsam Content erarbeiten und als Grundlage für die weitere Forschung zur Verfügung stellen, führten zu der Idee, eine Crowd-Sourcing-Veranstaltung zur Erstellung von Volltexten mittelalterlicher Handschriften zu organisieren. Die Veranstaltung sollte den Teilnehmenden zugleich die Möglichkeit bieten, sich in die Auszeichnungssprache XML und das in den Geisteswissenschaften und Digital Humanities als de facto-Standardformat etablierte Schema der Text Encoding Initiative (TEI)² einzuarbeiten sowie Erfahrungen im Lesen und Transkribieren mittelalterlicher Schriften zu sammeln. Dabei lag es inhaltlich nahe, das Feld der noch verhältnismäßig wenig erforschten und oft unedierten theologischen Handschriften als Materialgrundlage zu wählen.

Ein Rahmenprogramm mit Vorträgen rund um die Themen XML/TEI, digitale Edition und maschinelle Texterkennungsverfahren sollte die Transkriptionsarbeit begleiten und den Teilnehmenden weiterführende Einblicke in diese Forschungsfelder geben.

2. Organisatorischer Rahmen und technische Umsetzung

Den institutionellen und organisatorischen Rahmen der Veranstaltung bildeten zwei Bibliotheken, die Staatsbibliothek zu Berlin – Preußischer Kulturbesitz und die Universitätsbibliothek Leipzig. Als Infrastruktur und technische Umgebung diente die Alpha-Version des Handschriftenportals, der aktuell in einem DFG-geförderten Projekt von diesen beiden Einrichtungen sowie der Bayerischen Staatsbibliothek München und der Herzog August Bibliothek Wolfenbüttel entwickelten zentralen Plattform zu mittelalterlichen und neuzeitlichen Buchhandschriften aus deutschen Sammlungen.

Die Organisation und Koordination der Veranstaltung lagen bei einem Team aus dem Referat Wissenschaftliche Dienste der Staatsbibliothek, das ebenfalls für die Auswahl der

acht zu transkribierenden Handschriften aus dem Bestand dieser Bibliothek verantwortlich war. Das Team des Handschriftenzentrums der Universitätsbibliothek Leipzig wählte vier Handschriften aus dem Leipziger Bestand aus und leistete überdies fachliche Unterstützung während der Veranstaltung. Die Abteilung Digitale Dienste der UB, die innerhalb des Handschriftenportal-Projekts für die Entwicklung von Präsentation und Suche und damit auch den virtuellen Arbeitsplatz zuständig ist, war der technische Kooperationspartner.³ Dank der Unterstützung, Offenheit und Experimentierfreude der Kolleg:innen aus dem Handschriftenportal-Projekt war es möglich, die technische Umgebung des aktuell noch in Entwicklung befindlichen Handschriftenportals für den Transkribathon zu nutzen.

Aus der Perspektive des Handschriftenportals kam dem Transkribathon eine doppelte Funktion zu: neben der Erzeugung der eigentlichen Transkriptionen sollten auch Impulse für die nächste Ausbaustufe des Portals gesammelt werden, in der die Unterstützung kooperativer Arbeit an nutzergenerierten Inhalten im Fokus stehen wird.

Das Handschriftenportal als zentrale Plattform für die Arbeit mit dezentral vorgehaltenen Handschriftendigitalisaten ist ein idealer Anwendungsfall für das International Image Interoperability Framework (IIIF)⁴, das daher ein Fundament der Portal-Infrastruktur darstellt. Über die Spezifikation mehrerer Schnittstellen ermöglicht es IIIF, Bilder aus verteilten Repositorien einheitlich zu verarbeiten und somit unter anderem in eine gemeinsame Umgebung einbinden zu können. Auch im Transkribathon konnten darüber auf verschiedenen institutionellen Servern vorgehaltene Handschriftendigitalisate gemeinsam bearbeitet werden.

Der im Transkribathon verwendete virtuelle Arbeitsplatz des Handschriftenportals ([Seige, 2020](#)) implementiert den IIIF-Viewer Mirador ⁵, eine Open Source Software mit internationaler Entwicklergemeinschaft, die sich mit ihrer modernen, nutzerfreundlichen Oberfläche auf React-Basis gut in andere Webanwendungen integrieren sowie über ein Plugin-System erweitern lässt. Zur Erfüllung der Anforderungen des Transkribathons wurde auf ein existentes Annotationsplugin⁶ aus der Community zurückgegriffen, das leicht bedienbare Funktionalitäten für das Anlegen pixelbasierter Annotationen nach dem Web Annotation Data Model⁷ bietet. Dabei können Bildbereiche verschiedener geometrischer Formen markiert und mit Text verknüpft werden. Zur Persistierung der Annotationen ist die Anbindung einer Datenbank notwendig, worauf für den Transkribathon zugunsten einer temporären lokalen Speicherung im Local Storage des Browsers verzichtet wurde, um schnell und flexibel eine Arbeitsumgebung schaffen zu können, ohne spätere technische Entscheidungen im Projekt vorwegnehmen zu müssen.

Die Verwendung von Mirador für den auf die Erstellung von TEI-Dateien abzielenden Transkribathon stellte dabei eine grundlegende konzeptuelle Herausforderung dar: auf der einen Seite die atomare pixel- und canvasbezogene Herangehensweise unter Nutzung des Web Annotations Data Models, in dem jede Annotation selbständig auf einen Bildausschnitt bezogen ist und Kontext zwischen Annotationen nur durch das gemeinsam zugrundeliegende Image gegeben ist – auf der anderen Seite das dokumentenbasierte Zielformat XML/TEI, das die Vorlage als Ganzes abbildet und aus diesem Gesamtkontext heraus auf einzelne Stellen des Images verweist. Diesen Gegensatz gewinnbringend zu gestalten wird eine Herausforderung auch für spätere Transkriptionsfunktionalitäten innerhalb des Handschriftenportals bleiben; die Umsetzung für den Transkribathon stellte eine erste experimentelle Annäherung an das Problem dar.

Der gewählte Weg nutzt die Annotationsoberfläche von Mirador primär als Tool zur Erstellung einer Transkription, deren finale Form eine einzelne XML/TEI-Datei sein soll. Die pro Zeile erstellten, atomaren Annotationen fungieren dabei als Bausteine, die über einen am Ende der Bearbeitung stehenden Export in die korrekte Reihenfolge gebracht und zu einem die gesamte Seite umfassenden XML/TEI-Snippet zusammengefügt werden.

Dazu wurde das Community-Plugin sehr spezifisch für den konkreten Anwendungsfall erweitert: Die ergänzte interaktive Exportfunktion fügt die Annotationen anhand bereits in den Annotationstext eingefügter und nach dem inhaltlichen Textverlauf nummerierter <line>-Elemente in eine XML-Struktur ein. Sie ergänzt neben Wort-Tags unter anderem Image-URLs zur korrekten Identifizierung der Bezüge auch außerhalb des Erstellungskontextes sowie interne XML-IDs nach Angabe eines seitenbezogenen Prefixes. Die Pixelkoordinaten der Annotationen werden beim Export von xywh zu dem von der TEI geforderten Format ulx, uly, lrx, lry konvertiert. Gewünschte zusätzliche Informationen zu Illuminationen, Hervorhebungen oder Abkürzungen werden wie die Zeilennummern bereits im Annotationstext über XML/TEI-Auszeichnungen erfasst und fließen unverändert in den Export ein.

Exportierte Seitentranskriptionen, die von unterschiedlichen Teilnehmenden erstellt und eingereicht werden konnten, wurden schließlich manuell zu einer einzigen, validen XML/TEI-Datei für die jeweilige transkribierte Handschrift zusammengefügt, die alle Verweise als global gültige Verlinkungen enthält und dadurch unabhängig für sich als Ergebnis der Transkriptionsarbeit stehen, weitergegeben und nachgenutzt werden kann.

Die IIF-Umgebung des Handschriftenportals auch über diese Funktion als reines Erstellungswerkzeug hinaus für die Arbeit mit Transkriptionen und zu deren Anzeige nutzen zu können, ist Ziel der geplanten Erweiterungen. Die Erfahrungen, Ideen und

Anforderungen der Transkribathon-Teilnehmenden, die während der Veranstaltung auch in Feedbackrunden aufgenommen wurden, werden in die Konzeption des weiteren Ausbaus der Annotationsfunktionalitäten einfließen.

3. Durchführung der Veranstaltung

Schon bei der Planung und Terminierung des Transkribathons wurde die zentrale Zielgruppe, der akademische Nachwuchs, berücksichtigt. Um eine Teilnahme von Studierenden aus Seminargruppen zu ermöglichen, wurde die zweimonatige Veranstaltung mit Transkriptionsphase und Rahmenprogramm in die Semesterzeit gelegt. Mit einem Beginn Anfang Mai blieb den Seminargruppen außerdem nach Beginn des Semesters genügend Zeit, sich zusammenzufinden und auf den Beginn des Transkribathons vorzubereiten. Mit einem Vorlauf von mehreren Monaten wurden Dozent:innen mit entsprechenden thematischen Schwerpunkten im Leipziger und Berliner Raum direkt angeschrieben und über die Möglichkeit, mit einer Seminargruppe am Transkribathon teilzunehmen, informiert. Eine Integration in den Seminarbetrieb wurde schließlich mit Gruppen von Germanistikstudierenden (Prof. Dr. Katharina Philipowski, Dr. Inci Bozkaya, Universität Potsdam) und von Kunstgeschichte-Studierenden (Prof. Dr. Kathrin Müller, Humboldt-Universität zu Berlin) verwirklicht. Für die Seminargruppen wurden in Absprache mit den Dozentinnen im Vorfeld Textabschnitte reserviert. Neben diesen gezielten Anfragen für Lehrkooperationen erfolgte die Werbung zur Teilnehmenden-Gewinnung über Ankündigungen in Mailinglisten (h-soz-kult, H-Germanistik und DHd) sowie auf den Social Media-Kanälen der Staatsbibliothek, der UB Leipzig und des Handschriftenportals. Der grundsätzlich offene Call for Participation erfolgte so bereits über Medien, die auf die Zielgruppen zugeschnitten waren.

Die Anmeldung für individuelle Teilnehmende war mit der Auswahl einer spezifischen Textstelle aus einer Handschrift verknüpft. Diese Textstellen wurden vom Organisationsteam vorbereitet und auf der Veranstaltungswebseite zur Auswahl angeboten. So hatten die Teilnehmenden die Möglichkeit, sich Handschriften und Texte je nach ihrer Interessenslage auszusuchen, und aus organisatorischer Sicht konnte sichergestellt werden, dass Textstellen nicht doppelt transkribiert wurden. Bei den zur Auswahl stehenden Textstellen handelte es sich jeweils um circa 5 Seiten einer Handschrift. Für die Teilnahme am Transkribathon wurden Grundkenntnisse im Lesen von mittelalterlichen Handschriften sowie in Mittelhochdeutsch bzw. Latein vorausgesetzt. Kenntnisse in XML-Codierung waren hingegen nicht erforderlich. Insgesamt gab es 65 Anmeldungen von individuellen Teilnehmenden; dazu kamen 72

Personen, die im Rahmen von Lehrveranstaltungen teilnahmen. Von den 137 angemeldeten Teilnehmenden haben 107 Personen im Rahmen der Veranstaltung auch Transkriptionen beigesteuert. Einige Personen haben sich während oder nach der Veranstaltung abgemeldet; in vielen Fällen wurde nicht der gesamte Textabschnitt transkribiert, sondern nur ein bis zwei Seiten.

Die Transkriptionsarbeit wurde von einem Rahmenprogramm begleitet, das zwei Bereiche umfasste: zum einen Anleitung und Erfahrungsaustausch zur praktischen Transkriptionsarbeit in Form virtueller Tutorien, zum anderen begleitende Fachvorträge zu Mediävistik und Digital Humanities.

Zum Auftakt des Transkribathons fand ein virtuelles Tutorium statt, bei dem die Arbeitsumgebung im Handschriftenportal und die Transkriptionsregeln vorgestellt wurden. Am Ende des Tutoriums gab es für die Teilnehmenden die Möglichkeit, in sogenannten „Breakout-Rooms“ in Kleingruppen zusammenzukommen. Dazu wurden die Räume nach Handschriften, aus denen die Teilnehmenden ihre Transkriptionsabschnitte ausgewählt hatten, benannt. Da die Vernetzung der Teilnehmenden untereinander ein wichtiger Benefit für die Crowdworker:innen ist ([Waidmann, 2020, 126](#)), wurde so eine einfache Möglichkeit für die Teilnehmenden geschaffen, sich auch in der komplett virtuell stattfindenden Veranstaltung untereinander zu vernetzen. Dieses Angebot wurde von zahlreichen Teilnehmenden genutzt. Die Gruppen blieben über unterschiedliche selbstgewählte Kanäle dann auch während der Transkriptionsphase in Kontakt. Im Verlauf der Veranstaltung boten zwei weitere Tutorien die Möglichkeit, auf Fragen einzugehen, die während der Transkriptionsarbeit aufkamen, und konkrete Leseprobleme in den Handschriften gemeinsam zu besprechen. Um die stetige Vernetzung der Teilnehmenden weiter zu unterstützen, begannen die Tutorien stets mit einer 15-minütigen Session in den virtuellen Handschriftengruppenräumen.

Die Vorträge im Rahmenprogramm hatten eine andere Funktion: Hier sollten den Teilnehmenden wichtige Themen, die im Zusammenhang mit dem Transkribathon standen, von Fachexpert:innen präsentiert werden. Die offizielle Eröffnungsveranstaltung des Transkribathons mit dem Keynote-Vortrag „*alte kuntschaft und ouch niuwe*. Digitale Perspektiven auf die Mediävistik“ von Prof. Dr. Andrea Rapp und Dr. Luise Borek (Technische Universität Darmstadt) fand wenige Tage nach dem ersten Tutorium statt. Es folgten Vorträge zur Text Encoding Initiative (Torsten Schaßan, Herzog August Bibliothek Wolfenbüttel), zu digitalen Editionen (Dr. Jakub Šimek, Universitätsbibliothek Heidelberg) und zu OCR-Verfahren und -Instrumenten für historische Dokumente (Dr.

Christian Reul, Universität Würzburg). Zum Abschluss des Transkribathons fanden eine virtuelle Ergebnispräsentation und Podiumsdiskussion statt, bei der die Arbeitsabläufe im Transkribathon rekapituliert und die Ergebnisse der gemeinsamen Arbeit vorgestellt wurden. Mehrere Teilnehmende steuerten Berichte über ihre Erfahrungen mit der Transkriptionsumgebung und der kollaborativen Arbeitsweise bei. In der anschließenden Podiumsdiskussion zur Zukunft mediävistisch-digitaler Projekte in Bibliotheken und Forschungseinrichtungen kam Prof. Dr. Racha Kirakosian (Universität Freiburg) als Vertreterin der Mediävistik zu Wort, Leander Seige (Universitätsbibliothek Leipzig) brachte die Perspektive der Digitalen Dienste in Bibliotheken ein und Dr. Christian Reul (Universität Würzburg) berichtete von ersten Tests mit den während des Transkribathons erstellten Datensets mit der OCR4all-Software. Moderiert wurde die Podiumsdiskussion von Dr. Christoph Mackert (Handschriftenzentrum der Universitätsbibliothek Leipzig). So bot das Rahmenprogramm einen konkreten Nutzen für die Teilnehmenden, indem es ihnen die Möglichkeit gab, ihre eigenen Kenntnisse zu erweitern und verschiedene Forschungsperspektiven kennenzulernen. ([Waidmann, 2020, 126](#)).

Die Transkriptionsarbeit wurde von den Teilnehmenden asynchron in Einzelarbeit geleistet. Das offizielle Ende der Transkriptionsphase war auf den 30. Mai terminiert, so dass die Teilnehmenden ungefähr einen Monat Zeit für das Transkribieren der fünf ausgewählten Seiten hatten. Dabei konnten die Teilnehmenden bei Fragen jederzeit das Organisationsteam per E-Mail erreichen. Da konkrete Leseprobleme in den Tutorien besprochen wurden, handelte es sich bei den aufkommenden Fragen hauptsächlich um technische Probleme, beispielsweise mit dem Export⁸ der TEI-Dokumente aus dem Handschriftenportal.

Zur Vorbereitung und zur Unterstützung der Transkriptionsarbeit standen den Teilnehmenden umfangreiche Anleitungen und Materialien zur Verfügung. Das für den Transkribathon zusammengestellte TEI-Handbuch versammelte neben sämtlichen genutzten TEI-Elementen und -Attributen eine Zusammenstellung der Transkriptionsregeln, ein Beispiel-TEI-Dokument und eine Übersicht zu möglichen Sonderzeichen. Eine ausführliche Anleitung zur Arbeit im Handschriftenportal stand sowohl als PDF-Dokument als auch als Webseite mit Screencasts zur Verfügung. Hier wurden gezielt Anwendungsfälle wie beispielsweise das Eingeben einer Annotation für eine Zeile mit Schmuckelement oder eine Annotation für eine Zeile mit Initialen vorgestellt. Im Laufe des Transkribathons wurden Handbuch und Anleitungen aktualisiert und ergänzt, wenn sich aus dem Kreis der Teilnehmenden Fragen oder Sonderfälle ergeben hatten, die vorher noch nicht berücksichtigt worden waren. So wurde beispielsweise das ursprünglich vor allem für Worttrennungszeichen vorgesehene

Element <metamark> um die Möglichkeit erweitert, Zeilenfüllsel zu kennzeichnen (function="line-filler").

Nach Ablauf der Transkriptionsphase schickten die Teilnehmenden die über das Handschriftenportal erstellten TEI-Dokumente per E-Mail an das Organisationsteam, das die Qualitätskontrolle übernahm. Hierbei wurden die eingegangenen Transkripte auf XML-Validität geprüft, Fehler im Tagging eliminiert und Lesefehler korrigiert. Zudem wurde bei Worttrennungen eine Verknüpfung der Wortteile über die Attribute @xml:id sowie @prev und @next eingefügt. Außerdem wurde sichergestellt, dass trotz individueller Transkriptionseigenheiten, die die Crowdworker:innen einbrachten, eine größtmögliche Einheitlichkeit in der Umsetzung der TEI-Transkription gewahrt blieb. Da die Arbeitsumgebung des Handschriftenportals momentan noch keine Möglichkeit für den Import und die Anzeige von Annotationen bietet, wurden die Koordinaten der Annotationen mithilfe des LAREX-Tools⁹ überprüft und korrigiert, einem Layoutanalyse-Werkzeug aus dem OCR-Kontext ([Reul, Springmann, & Puppe, 2017](#)). Die fertigen XML-Dokumente wurden in einem nächsten Schritt zu einer Gesamtdatei für die jeweilige Handschrift zusammengefügt, und über eine XSLT-Transformation wurden zudem HTML-Dateien erstellt, die eine menschenlesbare Bildschirmausgabe der Transkriptionen mit einfacher grafischer Aufbereitung (zum Beispiel der Schmuckelemente) darstellen.

Die Datensets mit den Transkriptionen wurden nach Abschluss der Qualitätskontrolle zunächst im Lab der Staatsbibliothek¹⁰ und in einem zweiten Schritt auf dem Repositorium Zenodo unter der Lizenz CC0 veröffentlicht.¹¹

4. Beschreibung der entstandenen Datensets und des Transkriptionsschemas

Im Rahmen des Transkribathons wurden von den Teilnehmenden insgesamt 181 Seiten aus zwölf mittelalterlichen Handschriften transkribiert. Die Gesamtzahl der transkribierten Textzeilen beträgt 8.952. Im Folgenden werden die Transkribathon-Datensets hinsichtlich ihres inhaltlichen Zuschnitts sowie ihrer formalen Gestalt (Transkriptionsschema) beschrieben (siehe auch die Übersichtstabellen im Anhang und im SBB-Lab¹²).

Bei der Auswahl der Materialgrundlage für den Transkribathon lag der Fokus auf theologischen Texten, die in den letzten Jahren vermehrt ins Blickfeld der mediävistischen Forschung rücken, jedoch bisher größtenteils noch unediert sind. Um dem breiten Spektrum theologischer Literatur gerecht zu werden, wurden unterschiedliche Handschriften-/Texttypen ausgewählt: illustrierte Historienbibeln,

Gebet- und Stundenbücher, katechetische Textsammlungen sowie Predigten. Dabei wurden sowohl deutschsprachige als auch lateinische Handschriften berücksichtigt. Die Entstehungszeit der Handschriften reicht von circa 1320 bis zum Ende des 15. Jahrhunderts, die verwendeten Schriften umfassen mit Textura, Textualis, Gotischer Kursive und Bastarda sowohl Buchschriften auf hohem kalligraphischem Niveau als auch weniger elaborierte Kursivschriften.

Aus jeder Handschrift wurden vom Organisationsteam Abschnitte für die Transkription vorgeschlagen und von den Teilnehmenden bei der Anmeldung zum Transkribathon ausgewählt. Durch das offene Auswahlverfahren und den unterschiedlich erfolgreichen Rücklauf ergaben sich deutliche Abweichungen bei der Gesamtzahl der pro Handschrift tatsächlich transkribierten Seiten; sie liegt zwischen 5 und 36.

Als Grundlage für die Codierung der Transkriptionen in der Auszeichnungssprache XML wurde ein auf den Transkribathon zugeschnittenes Transkriptionsschema entworfen, das mit den P5 Guidelines der Text Encoding Initiative konform ist. Das Schema ist in einem TEI-Handbuch dokumentiert und liegt zudem als RNG-Schema vor.¹³ Nachfolgend werden die konzeptionellen Grundsätze, die Struktur und die wichtigsten Elemente des Transkriptionsschemas beschrieben.

Da die potentielle Nachnutzung der Transkriptionen als Ground Truth- beziehungsweise Trainingsdaten für automatische Texterkennungsverfahren ein wichtiges Ziel des Transkribathons war, wurde bei der Gestaltung des Transkriptionsschemas in erster Linie eine größtmögliche Nähe der Transkription zum Layout der Handschrift angestrebt. Daher wurde als Grundlage des Schemas das TEI-Modul Representation of Primary Sources¹⁴ gewählt. Es wurden immer ganze Seiten – unabhängig von inhaltlichen Abschnitten – transkribiert, und die Transkription wurde zeilengenau mit den Pixelkoordinaten des Digitalisats verknüpft. Um eine Nachnutzung außerhalb der Geisteswissenschaften zu erleichtern, wurde außerdem bei der Konzeption des TEI-Schemas auch, soweit möglich, auf strukturelle Ähnlichkeit mit Standardformaten aus dem Bereich der OCR-Verfahren wie Alto und PageXML geachtet.

Das Transkriptionsschema enthält die für TEI-Dokumente üblichen Bestandteile <teiHeader> für die Metadaten sowie <sourceDoc> für die Transkription. Jede Handschriftenseite wurde innerhalb des <sourceDoc> als <surface>-Element codiert, das mittels des Attributs @facs auf das Digitalisat der Seite und mittels @source auf die Canvas-Id aus dem IIIF-Manifest der Handschrift verweist.

Die Layout-Regionen der Handschriftenseite (Seite, Textzeilen, Buchschmuck, Paratexte) wurden als rechteckige Annotationen (Bounding Boxes) durch die Teilnehmenden im Handschriftenportal-Arbeitsplatz angelegt. TEI-Grundelement der Layout-Regionen ist die `<zone type="page">` für die ganze Handschriftenseite. Weitere Layoutelemente wie Illustrationen, Paratexte, Seitenzahlen u.ä. wurden ebenfalls als `<zone>`-Elemente annotiert und jeweils mit dem entsprechenden Attribut `@type="ornament | figure | fw"` versehen. Textzeilen wurden als `<line>`-Elemente annotiert. Beim Export der Annotationen aus dem Handschriftenportal wurden automatisiert die Bildkoordinaten der Bounding Boxes als Attribute (`@ulx, uly, lrx, lry`) der jeweiligen TEI-Elemente exportiert sowie eine eindeutige `xml:id` für jede Layoutregion erzeugt, um eine Referenzierung innerhalb des TEI-Dokuments zu ermöglichen.

Innerhalb der Layout-Elemente wurde dann entweder per `<note>`-Element eine Beschreibung eingefügt (zum Beispiel bei Illustrationen) und/oder der handschriftliche Text transkribiert. Für die Transkription galt die Grundregel der Buchstabentreue; dies wurde von den Teilnehmenden in unterschiedlichem Präzisionsgrad umgesetzt. Bei der Qualitätskontrolle wurden i/j-, u/v- und s/f-Schreibung korrigiert, granularere Unterscheidungen (wie zum Beispiel runder/gerader Schaft bei r und d oder punktförmige/schräggestellte/buchstabenförmige Umlautzeichen über Vokalen) jedoch in der von den Teilnehmenden gewählten Form belassen. Für die Codierung von Abkürzungen wurde das Element `<choice>` verwendet, das mit `<abbr>` und `<expan>` die parallele Transkription von Abkürzung und Auflösung erlaubt. Innerhalb der Zeile wurde jedes Wort mit einem `<w>`-Element umgeben; bei getrennten Wörtern wurden die Wortteile mittels Attributen aufeinander bezogen. Interpunktionszeichen wurden mit `<pc>` codiert, Worttrennungszeichen, Zeilenfüllsel und Verweiszeichen mit `<metamark>`.

Hervorhebungen im Text wie Initialen, Rubrizierung, Unterstreichungen und Ähnliche wurden mit dem Element `<hi>` und unterschiedlichen Attributen codiert. Wenn ein Buchschmuckelement wie zum Beispiel eine historisierte Initiale zugleich als eigene `<zone>` angelegt worden war, konnte darauf mittels des Attributs `@corresp` innerhalb des `<hi>`-Elements verwiesen werden.

Optional konnten die Teilnehmenden im Text enthaltene Eigennamen (Personen, Orte, Körperschaften, Werke) mit Normdaten (GND-IDs) versehen. Dies wurde jedoch nur von einem Bruchteil der Teilnehmenden umgesetzt; insgesamt wurden 60 Verlinkungen vorgenommen. Bei der Qualitätskontrolle wurden zwar die vorhandenen Verlinkungen überprüft, aber aus Zeitgründen keine zusätzlichen Normdatenverknüpfungen angelegt.

5. Nachnutzungsszenarien und Fazit

Das Crowd-Sourcing-Projekt „Faithful Transcriptions“ hatte zum Ziel, Volltext-Datensets von handschriftlichem Material zu erzeugen und für die Nachnutzung in unterschiedlichen Forschungskontexten frei verfügbar zu machen. Dabei sind Nachnutzungsszenarien in der geisteswissenschaftlich ausgerichteten Mediävistik (zum Beispiel für Editionsprojekte oder Textanalysen) ebenso denkbar und wünschenswert wie eine Weiterverwertung der Daten in der Informatik beziehungsweise den Digital Humanities (zum Beispiel als Ground Truth-Daten für Handwritten Text Recognition und Layout-Analyse sowie als Grundlage für computerlinguistische Analysen).

Um dieses Ziel zu erreichen, wurde das in den Geisteswissenschaften als de facto-Standard akzeptierte XML-Format TEI verwendet. Damit die Daten für bild- beziehungsweise layoutorientierte Ansätze nachnutzbar sind, wurde eine große Nähe der Transkription zum visuellen Erscheinungsbild der Handschrift angestrebt; für eher textorientierte beziehungsweise linguistische Nachnutzungsszenarien wurden auch sprachliche Spezifika berücksichtigt, indem zum Beispiel getrennte Wortteile aufeinander bezogen und Abkürzungen stets auch aufgelöst wurden. So sollte auch die Erstellung eines Lesetextes, zum Beispiel für den Gebrauch in der universitären Lehre, ermöglicht werden. Dieses Bemühen um eine flexible Einsetzbarkeit der Daten in unterschiedlichen Kontexten führte zwangsläufig zu pragmatischen Kompromissen bei der Erstellung des Schemas und damit auch zum Verzicht auf viele Ausdifferenzierungsmöglichkeiten, die das Format TEI eigentlich bietet. Die geringe Komplexität des hier vorgestellten Transkriptionsschemas erleichtert jedoch auch die Konvertierung der Dokumente in andere Formate und kann überdies je nach Kontext und Forschungsinteressen problemlos spezifiziert und angereichert werden.

Zugleich handelte es sich bei dem Transkribathon um ein Pilotprojekt, das die Rolle von Bibliotheken und ihren Infrastrukturangeboten als Raum für kollaboratives wissenschaftliches Arbeiten und gemeinsames Lernen auf verschiedenen Ebenen ausloten wollte. So wurden durch die Vorträge des Rahmenprogramms und die Tutorien einerseits Angebote zum Aufbau von Wissen und Kompetenzen für die Teilnehmenden gemacht, andererseits profitierten aber auch die Bibliotheken und das Handschriftenportal vom Input der Teilnehmenden, indem diese – neben der Anfertigung der Transkriptionen – zum Beispiel auch Feedback zur Weiterentwicklung des Arbeitsplatzes gaben. Nicht zuletzt wurde auch die Vernetzung der Teilnehmenden untereinander gefördert, sodass hier auch gegenseitige Unterstützung in den peer groups stattfinden konnte. Das digitale Format der Veranstaltung brachte – insbesondere für die

Vernetzungsaspekte – durchaus auch Herausforderungen mit sich, ermöglichte aber zugleich die Teilnahme und Zusammenarbeit zahlreicher, räumlich teilweise weit voneinander entfernter Personen.

Diese neuen, offenen und kollaborativen Formen des Arbeitens und des Wissensaustauschs werden in Zukunft gerade auch in den geisteswissenschaftlichen Fächern zunehmend an Bedeutung gewinnen. Der Transkribathon hat beispielhaft aufgezeigt, welche Rolle Bibliotheken in diesem Transformationsprozess spielen können, wenn sie ihr Potential als Vernetzungs- und Kreativraum für gemeinsames Arbeiten nutzen und den Prozess zusammen mit der Forschungscommunity aktiv gestalten.

Anhang: Datensets

Übersicht – alle Transkriptionen

Materialien:	Transkriptionen aus 12 mittelalterlichen Handschriften in deutscher, niederländischer und lateinischer Sprache
Schriften:	Textura, Textualis, Gotische Kursive, Bastarda
Anzahl Seiten:	181
Anzahl Zeilen:	8.952
Formate:	XML-codierte Transkriptionen nach TEI-Schema
Lizenzen:	CC0
Links:	Datenset auf Zenodo (zitierbar, XML-Dateien) DOI:10.5281/zenodo.5582483 Datenset im SBB-Lab (XML-Dateien und HTML-Ansicht)

Einzelne Handschriften

Handschrift:	<i>Berlin, SBB-PK, Ms. germ. fol. 516</i>
Inhalt:	Niederrheinische Historienbibel

Entstehungszeit:	um 1457-1460
Schrift:	Bastarda
Sprache:	Mittelhochdeutsch (westmitteldeutsch)
Anzahl Seiten:	36 (von 716)
Anzahl Zeilen:	2.252
Spezifika:	Illustrationen (kolorierte Federzeichnungen)

Handschrift:	<i>Berlin, SBB-PK, Ms. germ. fol. 1108</i>
Inhalt:	Historienbibel IIIb
Entstehungszeit:	1472
Schrift:	Bastarda
Sprache:	Mittelhochdeutsch (bairisch)
Anzahl Seiten:	21 (von 533)
Anzahl Zeilen:	1.544
Spezifika:	Illustrationen (Deckfarbenminiaturen)

Handschrift:	<i>Berlin, SBB-PK, Ms. germ. fol. 1413</i>
Inhalt:	Historienbibel
Entstehungszeit:	15. Jahrhundert (Mitte)
Schrift:	Bastarda
Sprache:	Mittelhochdeutsch (bairisch)
Anzahl Seiten:	27 (von 489)
Anzahl Zeilen:	1.368

Spezifika:	Illustrationen (teilweise kolorierte Federzeichnungen)
------------	--

Handschrift:	<i>Berlin, SBB-PK, Hdschr. 25</i>
Inhalt:	Stundenbuch
Entstehungszeit:	2. Hälfte 15. Jahrhundert
Schrift:	Textualis
Sprache:	Latein
Anzahl Seiten:	17 (von 310)
Anzahl Zeilen:	234
Spezifika:	Illustrationen (historisierte Initialen)

Handschrift:	<i>Berlin, SBB-PK, Ms. germ. fol. 76</i>
Inhalt:	Stundenbuch
Entstehungszeit:	1474
Schrift:	Bastarda
Sprache:	Niederdeutsch
Anzahl Seiten:	6 (von 251)
Anzahl Zeilen:	454

Handschrift:	<i>Berlin, SBB-PK, Ms. germ. oct. 6</i>
Inhalt:	Stundenbuch des Geert Grote
Entstehungszeit:	1480
Schrift:	Textura

Sprache:	Niederländisch
Anzahl Seiten:	5 (von 429)
Anzahl Zeilen:	76
Spezifika:	Illustrationen (historisierte Initialen, Zierseiten)

Handschrift:	<i>Berlin, SBB-PK, Ms. germ. oct. 511</i>
Inhalt:	Gebetbuch
Entstehungszeit:	15. Jahrhundert
Schrift:	Bastarda
Sprache:	Latein, Mittelhochdeutsch (bairisch)
Anzahl Seiten:	6 (von 315)
Anzahl Zeilen:	54
Spezifika:	Illustrationen (Deckfarbenminiaturen und Schmuckinitialen)

Handschrift:	<i>Berlin, SBB-PK, Ms. germ. fol. 19</i>
Inhalt:	Geistliche Sammelhandschrift
Entstehungszeit:	1448
Schrift:	Bastarda
Sprache:	Mittelhochdeutsch (südalemannisch)
Anzahl Seiten:	12 (von 533)
Anzahl Zeilen:	548
Spezifika:	Illustrationen (kolorierte Federzeichnungen)

Handschrift:	<i>Leipzig, UB, Ms 687</i>
Inhalt:	Predigtsammlungen
Entstehungszeit:	15. Jahrhundert (1. Viertel)
Schrift:	Gotische Kursive
Sprache:	Mittelhochdeutsch (ostmitteldeutsch)
Anzahl Seiten:	5 (von 418)
Anzahl Zeilen:	382

Handschrift:	<i>Leipzig, UB, Ms 688</i>
Inhalt:	Predigthandschrift (Sermones de tempore)
Entstehungszeit:	14. Jahrhundert
Schrift:	Textualis
Sprache:	Mittelhochdeutsch (ostmitteldeutsch)
Anzahl Seiten:	26 (von 290)
Anzahl Zeilen:	1.382

Handschrift:	<i>Leipzig, UB, Ms 758</i>
Inhalt:	Theologische Sammelhandschrift
Entstehungszeit:	um 1320-1340
Schrift:	Textualis
Sprache:	Latein, Mittelhochdeutsch (böhmisch/mittelbairisch)
Anzahl Seiten:	15 (von 486)

Anzahl Zeilen:	547
Handschrift:	<i>Leipzig, UB, Ms 762</i>
Inhalt:	Predigten, Summa confessionis
Entstehungszeit:	Ende 14./Anfang 15. Jahrhundert
Schrift:	Gotische Kursive
Sprache:	Latein, Mittelhochdeutsch (ostmitteldeutsch)
Anzahl Seiten:	5 (von 518)
Anzahl Zeilen:	111

Footnotes

1. <https://handschriftenportal.de/> (abgerufen am 13.9.2021) [↗](#)
2. <https://tei-c.org/> (abgerufen am 13.9.2021) [↗](#)
3. Ein erster Einsatz des Mirador als kollaborative Arbeitsumgebung wurde bereits bei einem Sommerkurs an der UB Leipzig 2017 getestet; dabei handelte es sich allerdings noch um eine Vorgängerversion von Mirador: Eichenberger, N., Johnson, C.H., Seige, L. (2018). A week of manuscript studies with the digital workspace „mirador@ubleipzig “. *Informationspraxis*, 3(2). <https://doi.org/10.11588/ip.2017.2.43293> [↗](#)
4. <https://iiif.io/> (abgerufen am 05.10.2021) [↗](#)
5. <https://projectmirador.org/> (abgerufen am 05.10.2021) [↗](#)
6. <https://github.com/ProjectMirador/mirador-annotations> (abgerufen am 05.10.2021) [↗](#)
7. <https://www.w3.org/TR/annotation-model/> (abgerufen am 05.10.2021) [↗](#)
8. Beim TEI-Export wurde keine XML-Validierung durchgeführt, es mussten aber bestimmte Voraussetzungen erfüllt sein, beispielsweise musste eine Annotation ‚zone ‘ mit Typ ‚page ‘ und xml:id angelegt sein, ohne die der Download nicht möglich war.

Hier war es in einigen Fällen beispielsweise durch Tippfehler zu Problemen gekommen. [↵](#)

9. Das Tool ist verfügbar auf GitHub: <https://github.com/OCR4all/LAREX> (abgerufen am 19.10.2021). Für die Unterstützung bei der Verwendung des Tools für die Korrektur der Transkribathon-Datensets möchten wir uns bei Christian Reul (Universität Würzburg) bedanken. [↵](#)

10. <https://lab.sbb.berlin/datensets-transkribathon/> (abgerufen am 05.10.2021) [↵](#)

11. <https://doi.org/10.5281/zenodo.5582483> [↵](#)

12. <https://lab.sbb.berlin/datensets-transkribathon/> (abgerufen am 05.10.2021) [↵](#)

13. Beide Dateien sind im Faithful Transcriptions-Datenset verfügbar:
<https://www.doi.org/10.5281/zenodo.5582483> [↵](#)

14. <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html> (abgerufen am 13.9.2021) [↵](#)

Citations

1. Rapp, A. (2021). Digital Humanities und Bibliotheken: Traditionen und Transformationen. *027.7 Zeitschrift Für Bibliothekskultur / Journal for Library Culture*, 8(1). <https://doi.org/10.21428/1bfadeb6.486c17e5> [↵](#)
2. Waidmann, S. (2020). Benutzer/-innen als Kurator/-innen: Crowdsourcing in Bibliotheken. In K. U. Werner (Ed.), *Bibliotheken als Orte kuratorischer Praxis* (pp. 125–138). Berlin, Boston: De Gruyter Saur. <https://doi.org/10.1515/9783110673722-010> [↵](#)
3. Owens, T. (2014). Making Crowdsourcing Compatible with the Missions and Values of Cultural Heritage Organisations. In M. Ridge (Ed.), *Crowdsourcing our Cultural Heritage* (pp. 269–280). London: Routledge. [↵](#)
4. Burrichter, B., Gebert, B., Mackert, C., & Viehhauser, G. (2021). Digitale Mediävistik. *Das Mittelalter. Perspektiven Mediävistischer Forschung*, 26, 101–117. <https://doi.org/10.17885/heiup.mial.2021.1.24312> [↵](#)
5. Engl, E. (2020). OCR-D kompakt: Ergebnisse und Stand der Forschung in der Förderinitiative. *Bibliothek Forschung Und Praxis*, 44(2), 218–230. <https://doi.org/doi:10.1515/bfp-2020-0024> [↵](#)

6. Seige, L. (2020). Das physische Objekt im virtuellen Raum. Zur Entwicklung des IIIF-Viewers Mirador 3 für das Handschriftenportal. Retrieved from <https://handschriftenportal.de/mirador3/> ↵
7. Reul, C., Springmann, U., & Puppe, F. (2017). LAREX: A Semi-Automatic Open-Source Tool for Layout Analysis and Region Extraction on Early Printed Books. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage* (pp. 137–142). Göttingen, Germany: Association for Computing Machinery. <https://doi.org/10.1145/3078081.3078097> ↵