

# Anticipating digital research Der Codex Sinaiticus gibt seine Daten frei

**Schröer, Annika**

[schroer@ub.uni-leipzig.de](mailto:schroer@ub.uni-leipzig.de)

Universitätsbibliothek Leipzig, Deutschland

ORCID: 0000-0003-2293-4093

## Einleitung

Die älteste Überlieferung des vollständigen neuen Testaments hat eine bis heute bewegte Geschichte: Im 4. Jahrhundert in griechischer Sprache auf über 400 Pergamentblätter geschrieben, zu einem großformatigen Codex gebunden und über 8 Jahrhunderte hinweg ausgesprochen reich kommentiert; Mitte des 19. Jahrhunderts im Katharinenkloster auf dem Sinai durch Konstantin von Tischendorf entdeckt und aus dem Kloster entfernt, wird der Codex Sinaiticus heute in vier Teilen aufbewahrt: in der British Library in London (BL), der Universitätsbibliothek Leipzig (UBL), dem Katharinenkloster sowie der Russischen Nationalbibliothek in St. Petersburg (Parker, 2010). Die überaus wertvolle Quelle für Forscher:innen verschiedener geisteswissenschaftlicher Disziplinen ist digitalisiert und transkribiert seit Ende der 2000er Jahre wieder als Gesamtheit online verfügbar – allerdings primär zur Betrachtung über die Weboberfläche.

Seit 2022 wird das Webportal im Rahmen eines mehrjährigen, DFG-geförderten Projekts komplett überarbeitet. Dadurch soll nicht nur das User Interface auf heutige Standards von Usability, Responsivität und Barrierearmut gehoben werden, sondern vor allem die reichhaltige Datengrundlage für die maschinelle Nutzung in den Digital Humanities verfügbar werden.

## Ausgangslage und Ziel

Die exzellenten Daten sind das Ergebnis des von 2003 bis 2009 durchgeführten ersten Projekts, das die aktuelle Webseite entwickelt hat und an dem neben den vier Partnerinstitutionen zahlreiche weitere internationale Expert:innen beteiligt waren. Unter anderem wurden alle Seiten des Codex hochauflösend digitalisiert und der gesamte Text inklusive der Kommentare transkribiert. Die Transkription wurde als XML auf Basis des TEI-Standards kodiert und über Pixelkoordinaten wortgenau auf die Digitalisate bezogen. An einigen Stellen wurde aufgrund von Besonderheiten des Codex vom TEI-Standard abgewichen, weshalb das XML-

Dokument nicht durchweg kompatibel ist. Es steht als Gesamtdatei zum Download zur Verfügung, doch es existieren keine weiteren Schnittstellen.

Das aktuelle Projekt verfolgt datenbezogen das Ziel, alle Informationen – Images sowie Transkriptionsdaten – möglichst atomar, standardisiert und unter Verwendung quell-offener Software für maschinelle Nachnutzung aufzubereiten. Die Entwicklung findet, begleitet durch eine erneute Projektpartnerschaft mit der BL und dem Katharinenkloster auf dem Sinai, an der Universitätsbibliothek Leipzig statt.<sup>1</sup> Für fachliche Fragen steht dem Projekt ein wissenschaftlicher Beirat aus internationalen Expert:innen zur Seite, der die Anforderungen spezifischer Zielgruppen vertritt.

## Herangehensweise

Der Einsatz des International Image Interoperability Frameworks (IIIF) zur Einbindung der Digitalisate ist eines der grundlegenden Konzepte im Projekt. Es ermöglicht unter anderem die Verlinkung beliebiger Bildbereiche und die direkte Einbindung von Inhalten in andere Anwendungen. (IIIF Consortium, o. J.) Essenziell dabei ist die Entscheidung für ein gemeinsames Manifest für den Gesamtcodex. Dieses muss für alle Seiten des Codex eindeutige, mit URI identifizierte Canvases enthalten, die wiederum die Bilder verknüpfen. Da sowohl die BL als auch die UBL für ihre jeweiligen Teile des Codex bereits über IIIF-Repräsentationen verfügen, existieren auch bereits Canvas und URI für jede davon abgedeckte Seite. Anstatt, wie häufig praktiziert, projektspezifisch neue Daten zu erzeugen, werden die URIs der vorhandenen Canvases in das gemeinsame Manifest integriert. Dadurch beziehen sich Referenzen auf IIIF-Codexseiten aus Leipzig oder London auch auf den virtuellen Gesamtcodex und umgekehrt.

Das Fundament aller textbasierten Daten bildet die XML-Transkription. Zu Beginn des aktuellen Projekts wurde die bisherige Modellierung durch Partner unter anderem in Birmingham iterativ verbessert. Sie ist nun in allen Teilen TEI-P5-konform und zudem innerhalb des XML durch automatisiertes Tagging zu großen Teilen auf Wortebene mit griechischen Wörterbuchformen und Strongs' Numbers angereichert, um die Suchfunktionalitäten zu verbessern.

Der auf Apache Solr basierenden Suchindex hält Dokumente auf Versebene vor, zu denen neben den eigentlichen Inhalten und Metadaten weitere Informationen ergänzt werden, beispielsweise Varianten der enthaltenen Nomina Sacra sowie die im TEI angereicherten, grammatisch normalisierten Wörterbuchformen und Strongs' Numbers mit Übersetzungen. Dies ermöglicht Treffer für Suchterme mit nicht direkt im Text vorhandenen Schreibweisen und Flexionen, und sogar Suchen in Fremdsprachen.

Zur Darstellung und Verknüpfung der Transkriptionen mit Bilddaten und Suche werden die im TEI einzeln ausgezeichneten Wörter zu separaten, jeweils über eine URI referenzierten Annotationen nach dem Open Web Annotation Standard konvertiert. Dabei wird explizit keine komplexe Repräsentation der gesamten TEI-Semantik (Ciotto

& Tomasi, 2016) angestrebt, sondern ein zusätzlicher flacher Einstieg in die Daten auf Wortebene geschaffen. Jede Annotation enthält, in LOD-Vokabularen modelliert (Allemang, 2011), alle vorhandenen Angaben zu dem spezifischen Wort: die transkribierten Buchstaben, editorische Auszeichnungen, Dictionary Form, Strongs' Number und Übersetzungen sowie neben weiteren Metadaten die URI-codierten Koordinaten zu den Canvas-Bereichen des korrespondierenden Bildes.

## Ausblick

Alle beschriebenen Daten – IIIF-Images, TEI, Annotationen – werden über Schnittstellen zur freien Nachnutzung veröffentlicht. Parallel zu den im Bereich des User Interface stattfindenden Usability Tests sollen während des Beta-Stadiums der APIs Vertreter:innen der DH-Communities nach weiteren datenzentrierten Use Cases und Verbesserungsansätzen befragt werden.

Dadurch sollen über die Anforderungen des wissenschaftlichen Beirats hinaus die aktuellen Bedarfe der Digital Humanities erkannt und berücksichtigt werden – im Sinne des Titels „Anticipating digital research“.

## Fußnoten

1. Aufgrund der aktuellen politischen Lage besteht momentan kein Kontakt zur Russischen Nationalbibliothek in St. Petersburg.

## Bibliographie

**Allemang, Dean.** 2011. „Semantic Web for the working ontologist: Effective modeling in RDFS and OWL.“ Amsterdam: Morgan Kaufmann/Elsevier.

**Böttlich, Gottfried.** 2011. „Der Jahrhundertfund: Entdeckung und Geschichte des Codex Sinaiticus.“ Leipzig: Evangelische Verlagsanstalt.

**Ciotti, Fabio and Francesca Tomasi.** 2016. „Formal Ontologies, Linked Data, and TEI Semantics“ *Journal of the Text Encoding Initiative* 10.4000/jtei.1480.

**IIIF Consortium. o. J.** „How It Works: A plain-language guide to how the APIs work.“ <https://iiif.io/get-started/how-iiif-works> (zugegriffen: 24.07.2024).

**Parker, David C.** 2010. „Codex Sinaiticus: the story of the world's oldest bible.“ London: The British Library.