FAIRe Forschungsdaten Die ersten 2400 Briefe an Goethe als TEI-XML-Volltexte im Akademienvorhaben PROPYLÄEN: Goethes Biographica

Thomas, Christian
thomas@bbaw.de
Klassik Stiftung Weimar, Berlin-Brandenburgische Akademie der Wissenschaften
ORCID: 0000-0002-1761-0222

Prell, Martin
Martin.Prell@klassik-stiftung.de
Klassik Stiftung Weimar, Sächsische Akademie der Wissenschaften
ORCID: 0000-0003-3152-6542

Hofmann-Polster, Katharina
Katharina.Hofmann-Polster@klassik-stiftung.de
Sächsische Akademie der Wissenschaften

Häfner, Claudia
Claudia.Haefner@klassik-stiftung.de
Klassik Stiftung Weimar

[1]Mit der Veröffentlichung einer neuen Version der Editions- und Forschungsplattform *PROPYLÄEN: Goethes Biographica*[2] ging im September 2024 erstmals ein Korpus TEI-XML-kodierter Forschungsdaten aus diesem Akademienvorhaben online. Die neue Version der Plattform bietet somit neben mehreren Erweiterungen und Optimierungen erstmals einen ganzen Schwung frischer, FAIRer[3] und frei verfügbarer Forschungsdaten (Lizenz: CC BY 4.0), die im Zentrum der Poster-Präsentation stehen.

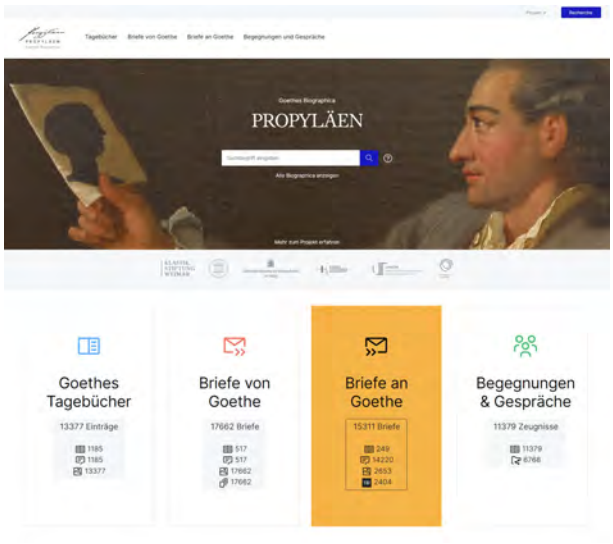

Abbildung 1: *PROPYLÄEN*-Startseite (Detail) mit Informationen zu den aktuell verfügbaren Datensätzen der vier Teilprojekte bzw. -editionen.

Das *PROPYLÄEN*-Vorhaben ist Teil des Akademienprogramms und mit einer Gesamtlaufzeit von 25 Jahren bis 2039 konzipiert; es vereint vier Editionsvorhaben auf einer gemeinsamen digitalen Plattform: Die „Begegnungen und Gespräche", die Briefe von und an Goethe sowie dessen Tagebücher (vgl. Koltes u. a. 2023). Auf der *PROPYLÄEN*-Plattform finden Nutzer:innen bereits jetzt mehrere tausend Datensätze (vgl. Abb.1) in einer Volltext-Repräsentation mitsamt strukturierten Metadaten, die aus den retrodigitalisierten Daten zuvor gedruckter Bände erarbeitet wurden (vgl. Neuber u. a. 2020).

Bis zum Projektende werden diese Datensätze – dem Tagungsmotto „Under Construction" entsprechend – sukzessive um weitere retrodigitalisierte Daten sowie genuin digital erstellte Datensätze ergänzt. Im Laufe seines Lebens empfing der „Hochwohlgeborene Geheimerath" mehr als 20000 Briefe von etwa 3500 Absender:innen aus aller Welt. Der erste publizierte Forschungsdatensatz aus dem „born digital"-Workflow umfasst 2406 „Briefe an Goethe" aus dem gleichnamigen Teilprojekt, die auch über die Programmierschnittstelle (API) des *PROPYLÄEN*-Vorhabens bezogen werden können.[4]

Überblick: Datensatz „Briefe an Goethe" zur Veröffentlichung Sept. 2024.

Dokumente	2406
Zeitraum	September 1786–Ende 1797
Absender:innen	463
Faksimiles	7257
Zeichen	ca. 4,8 Mio.
Tokens[5]	ca. 871000
Types	ca. 76400

Die in TEI-XML annotierten Daten[6] stellen ein Novum und einen signifikanten Meilenstein dar. Erstens wurden diese zum größeren Teil nie zuvor veröffentlicht; zweitens war deren Veröffentlichung *als Volltexte* auch im Teilprojekt selbst zunächst nicht vorgesehen. Denn die Druck-

fassung der Ausgabe präsentiert lediglich *Regesten* der Brieftexte – d.h. kondensierte Inhaltszusammenfassungen –, ergänzt um Angaben zum Überlieferungsort der Vorlage, dem Entstehungsort, Absender- und Empfänger:innen[7] des Briefes sowie umfangreiche Registerpositionen. Die Publikation der Volltexte (samt Digitalisaten) ist somit ein Alleinstellungsmerkmal der Digitalen Plattform des als Hybrid-Edition konzipierten *PROPYLÄEN*-Vorhabens.

Abbildung 2: Screenshot (Detail): TEI-XML-Kodierung eines Beispielbriefs, Ansicht im *Oxygen XML Editor*. Oben: Regest (tei:abstract) mit Auflistung der erwähnten Personen (tei:listPerson) und Korrespondenz-Metadaten (tei:correspDesc); unten: Brief-Volltext mit Brief-spezifischen Annotationen (z.B. tei:opener, tei:salute, tei:dateline u.a.).

Die Gesamtmenge der Briefe wäre ohne automatisierte Texterkennung mittels *Transkribus*[8] bzw. *OCR4all*[9] nicht zu bewältigen. Zur Dokumentation und Vernetzung werden konsequent Normdaten verwendet, die in der Forschungsdatenbank*so:fie* der KSW[10] aggregiert werden. Ebenso werden Services wie *correspSearch* genutzt, wo bereits die Metadaten zu mehr als 15000 Briefen an Goethe aus dem *PROPYLÄEN*-Vorhaben bereitgestellt wurden, wodurch die Daten weithin sichtbar sind und zugleich im Kontext weiterer Briefeditionen durchsucht und analysiert werden können (vgl. das Beispiel in Abb. 3).[11]

Abbildung 3: Screenshot (Detail): *correspSearch: csVis (BETA)*: Netzwerk des *PROPYLÄEN*-Datensatzes mit Hervorhebung der Briefe von Christian Gottlob von Voigt und Goethe im Zentrum, URL .

Das Poster wird einen Überblick über das erste Teilkorpus aus dem *PROPYLÄEN*-Projekt bieten und einen Ausblick auf die in den kommenden Jahren zu erwartenden, noch sehr viel umfangreicheren und vielfältigeren Forschungsdaten aus allen vier Teilprojekten bzw. -editionen des Vorhabens geben. Dabei werden die technologischen, methodologischen und editorischen Prinzipien vorgestellt und erste Auswertungen der Daten präsentiert.

Fußnoten

1. Contributor Roles: Christian Thomas (Conceptualization, Writing – original draft), Martin Prell, Claudia Häfner, Katharina Hofmann-Polster (Writing – review & editing).
2. https://goethe-biographica.de/.
3. (Vgl. Wilkinson, Dumontier und Aalbersberg 2016). Die Bereitstellung der Volltexte und Metadaten orientiert sich unter anderem an der vom Konsortium NFDI4Culture bereitgestellten Checkliste https://nfdi4culture.de/de/services/fair-check.html; alle Punkte werden aktuell bereits erfüllt bzw. in naher Zukunft im Rahmen der Gesamtstrategie der Klassik Stiftung Weimar (KSW) erfüllt werden.
4. API: https://goethe-biographica.de/projekt/how-to-use-api.html.
5. Ermittlung der Token-Zahlen: Zentrum für digitale Lexikographie (ZDL) / *Digitales Wörterbuch der deutschen Sprache* (DWDS) im Zuge einer probeweisen Indexierung der Daten; Types sind unique Tokens. Die Integration sämtlicher Forschungsdaten aus dem *PROPYLÄEN*-Vorhaben in die historischen Korpora des DWDS und des NFDI-Konsortiums *Text+* ist vorgesehen, was deren unmittelbare Nachnutzung sowie eine größere Nutz- und Sichtbarkeit gewährleistet. Diese Vernetzungsaktivitäten stehen im Kontext der Nationalen Forschungsdateninfrastruktur (NFDI), hier insbesondere dem Konsortium Text +, https://text-plus.org/.
6. Annotation gemäß Richtlinien der *Text Encoding Initiative* (TEI), orientiert am DTA-Basisformat für Manuskripte (vgl. Haaf und Thomas 2017), weitreichend ergänzt um Best Practices aus der *Carl-Maria-von-Weber-Gesamtausgabe*, der *edition humboldt digital*, dem

Handbuch *Encoding Correspondence* (Dumont, Haaf und Seifert 2020) u.a. Einige Erweiterungen der TEI-Spezifikationen aus dem *PROPYLÄEN*-Vorhaben flossen bereits in das jüngste Release 4.8.0 der *TEI Guidelines* ein.
7. Zumeist ist Goethe alleiniger Empfänger, einzelne Briefe gingen jedoch an Dritte, die Goethe die Inhalte der Briefe bzw. Teile derselben übermitteln sollten (vgl. Abb.3: Goethe ist im Zentrum, einige Briefwechsel, so wie der hervorgehobene, stehen außerhalb dieses Knotens).
8. READ-COOP: *Transkribus*, https://www.transkribus.org/.
9. *OCR4all: Optical Character Recognition (and more) for everyone*, https://www.ocr4all.org/.
10. *Forschungsdatenbank so:fie*, https://ores.klassik-stiftung.de/ords/f?p=900:1.
11. *correspSearch: Briefeditionen durchsuchen und vernetzen,* https://correspsearch.net/; zugrundeliegender Datensatz auf *Zenodo* (Project „PROPYLÄEN: Goethes Biographica“ 2023).

Bibliographie

Dumont, Stefan, Susanne Haaf und Sabine Seifert, Hrsg. 2020. Encoding Correspondence. A Manual for Encoding Letters and Postcards in TEI-XML and DTABf. doi:https://encoding-correspondence.bbaw.de/, .

Haaf, Susanne und Christian Thomas. 2017. Enabling the Encoding of Manuscripts within the DTABf: Extension and Modularization of the Format. *jTEI: Journal of the Text Encoding Initiative*. doi:https://doi.org/10.4000/jtei.1650, .

Koltes, Manfred, Ariane Ludwig, Yvonne Pietsch, Martin Prell und Bastian Röther. 2023. PROPYLÄEN. Ein Jahrhundertprojekt geht online. *Digitale Bibliothek Thüringen (DBT)*. doi:https://doi.org/10.22032/dbt.55585, .

Neuber, Frederike, Thorsten Schaßen, Dominik Kasper, Martina Gödel und Thomas Stäcker. 2020. Altbausanierung mit Niveau – die Digitalisierung gedruckter Editionen. In: *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. 7. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum“ (DHd 2020), Paderborn*, hg. von Patrick Helling und Christof Schöch. doi:https://zenodo.org/doi/10.5281/zenodo.4621821, .

Project „PROPYLÄEN: Goethes Biographica“, Hrsg. 2023. CMIF of Letters to Johann Wolfgang Goethe (Version 2) [Data set]. Zenodo. doi:https://doi.org/10.5281/zenodo.8063593, .

Wilkinson, Mark D., Michel Dumontier und Ijsbrand Jan Aalbersberg. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci data 3, 160018 (2016)*. doi:https://doi.org/10.1038/sdata.2016.18, .