# Die offene Edition. Vernetzung, Datenpublikation und Transparenz in der edition humboldt digital

**Dumont, Stefan**
dumont@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

**Kraft, Tobias**
kraft@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

**Seifert, Sabine**
sabine.seifert@uni-potsdam.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

**Thomas, Christian**
thomas@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

**Wierzoch, Jan**
jan.wierzoch@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Das an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) angesiedelte Akademienvorhaben *Alexander von Humboldt auf Reisen – Wissenschaft aus der Bewegung* erschließt und ediert die amerikanischen, russisch-sibirischen und europäischen Reisetagebücher des preußischen Naturforschers und Entdeckers Alexander von Humboldt.[1] Begleitet werden die Tagebücher von thematisch zugehörigen Briefen aus seinem weltumspannenden Korrespondentennetz sowie von Manuskripten aus seinem umfangreichen Nachlass, von denen viele bis dato nie veröffentlicht wurden. Ergänzt werden diese edierten Texte durch Forschungsbeiträge, eine Chronologie zu Humboldts Leben und umfangreiche Register. Die Publikationsstrategie ist "digital first", d. h. die veröffentlichten Dokumente erscheinen zuerst online ohne Verzögerung durch 'Moving Walls' oder sonstige verlagsseitige Einschränkungen, und die gesamte Ausgabe ist unter einer Creative-Commons-Lizenz[2] vollständig frei zugänglich. Mit der Veröffentlichung der ersten Bände der Print-Edition im Verlag Springer Nature/J. B. Metzler[3] wurde die Hybridstrategie des Projekts umgesetzt. Realisiert wird die digitale Edition mit ediarum[4], das u. a. auf X-Technologien, der freien XML-Datenbank eXistdb und der Software Oxygen XML Author aufsetzt.

Fünf Jahre nach dem Launch der *edition humboldt digital* ( *ehd*) haben wir acht Versionen dieser digitalen, textkritisch-dokumentarischen Edition[5] vorgelegt. Mit der aktuellen Version 8 der *ehd* (veröffentlicht im Mai 2022)[6] lösen wir nun das Versprechen ein, Humboldts komplexe handschriftliche und schwer zu entziffernde Texte auch *als Daten* bereitzustellen, indem wir (1) die kommentierten Texttranskriptionen von mehr als 500 Dokumenten (ca. 2.800 Seiten), (2) die umfassende Alexander von Humboldt-Chronologie mit ca. 1.600 datierten Ereignissen aus Humboldts fast 90-jährigem Leben und (3) ca. 18.000 Indexeinträge (z. B. Personen, Orte, Institutionen, bibliographische Einträge) auf GitHub (Ette et al. 2022) und (ab Winter 2022/23) auf Zenodo zur Verfügung stellen. Alle Datensätze liegen im TEI-XML-Format vor. Dem Single-Source-Prinzip folgend, basieren sowohl die digitale als auch die gedruckte Komponente (Buch, PDF und eBook-Derivate) vollständig auf denselben TEI-XML-kodierten Daten. Das TEI-XML-Subset der *ehd* wurde durch Übernahme etablierter TEI-Spezifikationen, v. a. des Basisformats für Manuskripte des Deutschen Textarchivs (DTABf-M[7]; Thomas/Haaf 2016-2019), entwickelt, um ein Höchstmaß an Standardisierung, Nachnutzbarkeit und Interoperabilität der Daten zu gewährleisten (Dumont/Haaf/Kraft/Czmiel/Thomas/Boenig 2016). Eine umfassende Dokumentation der Transkriptions- und Kodierungsrichtlinien steht zur Verfügung[8] und kann von anderen, ähnlich gelagerten Projekten nachgenutzt werden[9]. Gleichzeitig bringen sich die Projektmitarbeiter:innen aktiv in die Verbesserung von bestehenden Richtlinien ein[10]. Dadurch fließen auf zweierlei Wegen Erfahrungen aus der editorischen Praxis in die Community zurück.

Seit Abschluss der Betaphase im Mai 2017 wird die *edition humboldt digital* versioniert publiziert, d. h. die Daten werden nicht einfach aktualisiert und überschrieben, sondern durch jedes Update (mittlerweile einmal im Jahr) wird eine neue, zusätzliche Datenschicht hinzugefügt. Gleichzeitig werden alle vorangehenden Versionen weiterhin bereitgehalten und lassen sich über die Web-Oberfläche aufrufen – bis hin zu den Registereinträgen mit ihren dynamischen Verlinkungen. Die Datensätze werden in Zukunft immer als neue, zusätzliche Version publiziert. Ergänzt wird diese Versionierung seit 2022 durch die Einführung von "Editionsstufen", die die unterschiedlichen Bearbeitungszustände systematisch abbilden.[11] Zusammen mit der umfangreichen Dokumentation wird damit der gesamte Forschungs- und Editionsprozess offen gelegt und analysierbar. Einen ersten summarischen Überblick über den Fortschritt der *ehd* gibt die mit Version 8 neu hinzugekommene "Versionsgeschichte", die die acht publizierten Versionen auch quantitativ auswertet.[12]

Das nun zur Verfügung gestellte Datenset wird nicht direkt aus der eXistdb-Datenbank exportiert, sondern über die öffentlich zugängliche API abgerufen.[13] Das ermöglicht diverse Optimierungen am Datenbestand für die externe Nachnutzung. So werden z. B. alle projektinternen

IDs durch URIs aus Normdateien ersetzt – sofern eine solche in den einschlägigen Normdaten-Beständen vorhanden ist. Ist dies nicht der Fall, werden die projektinternen IDs als vollständige URIs ausgegeben und so immerhin technische Interoperabilität gewährleistet.

Die Registereinträge werden, wo immer verfügbar, mit URIs aus Normdateien versehen. Insbesondere das Personenregister weist dabei großes Potenzial für die unmittelbare Nachnutzung auf, da zahlreiche historische Personen in Humboldts Texten noch nicht in der wichtigsten Normdatei für die deutschsprachige Forschungsgemeinschaft, der GND[14] der Deutschen Nationalbibliothek, dokumentiert sind.[15] Diese ergänzenden Daten können dazu beitragen, die Normdateien der Community, wie GND, Wikidata etc. zu verbessern. Dabei stellen sich unterschiedlich große Hürden für die Zuarbeit: während Wikidata von Prinzip her ein offenes Communityprojekt ist, ist die GND institutionell angesiedelt und wird redaktionell betreut. Die Zuarbeit zur GND wurde im Projekt GND4C grundsätzlich für nicht-bibliothekarische Bereiche geöffnet, in naher Zukunft sollen die Ergänzungsmöglichkeiten seitens des Editionsvorhaben ausgelotet werden. Leider stehen solche Zuarbeiten immer unter dem Vorbehalt der Projektkapazitäten, da sie bei der *ehd* – ebenso wie in den meisten anderen Projekten – eigentlich nicht vorgesehen sind.

Die API selbst bietet nicht 'nur' den Volltext an, sondern ebenfalls diverse Metadaten, um eine umfassende Vernetzung der Edition zu gewährleisten. So bietet eine Schnittstelle die Metadaten zu den edierten Texten und Forschungsbeiträgen unter anderem im Dublin Core-Format via OAI-PMH an. Dadurch werden alle diese Texte in der Open Access-Suchmaschine BASE nachgewiesen. Daneben werden eine BEACON-Schnittstelle und eine CMIF-Schnittstelle für correspSearch (Dumont, Grabsch und Müller-Laackman 2021) angeboten, die mittlerweile zu den 'klassischen' Ausstattungen einer digitalen Edition zählen dürfen. Eine Schnittstelle ins Semantic Web gibt es derzeit noch nicht.[16] Grund dafür ist v.a., dass andere Schnittstellen und vor allem Funktionen der *ehd* bisher im Fokus der Entwicklungsarbeit standen. Das Vorhaben läuft bis 2032 und lässt es daher grundsätzlich zu, in Zukunft auch diesen Bereich anzugehen. Denkbar wäre es, z.B. die Einträge der Chronologie oder der Register noch stärker zu schematisieren und tiefer zu kodieren, um dieses Wissen als Linked Open Data bereitzustellen. Das würde aber nicht nur Entwicklungsaufwand bedeuten, sondern auch erhebliche redaktionelle Arbeit, für die entsprechende Ressourcen geschaffen werden müssten. Die edierten Texte an sich werden für solch eine zusätzliche Aufbereitung - im Sinne einer "assertive edition" (Vogeler 2019) - wohl leider nicht in Frage kommen, dafür wäre der Aufwand (gegenüber dem Projektplan) zu hoch.

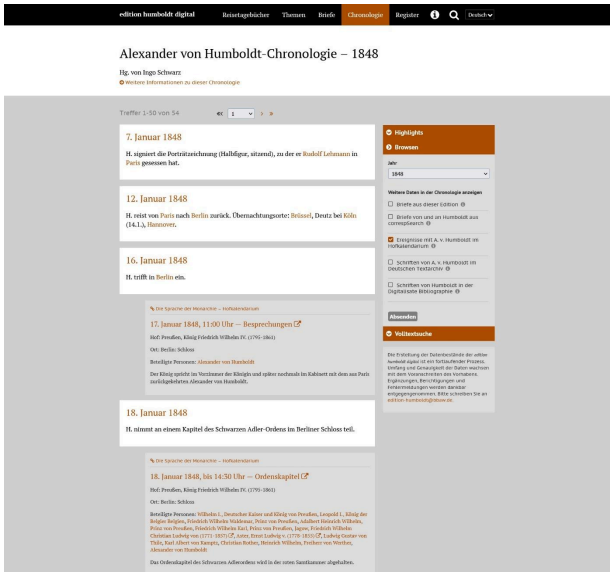


Abb. 1: Screenshot der Alexander von Humboldt-Chronologie in der *edition humboldt digital* mit eingeblendeten Daten aus dem Hofkalendarium 1848.

Die *edition humboldt digital* stellt nicht nur ihre eigenen Daten zur Verfügung, sondern nutzt auch andere Daten nach und verwendet externe Webservices zur Anreicherung oder zur Vergrößerung des eigenen Funktionsumfangs. So werden die TEI-XML-Dateien zur Lemmatisierung an den Webservice DTA::CAB[17] (Jurish 2011) geschickt, dort analysiert und angereichert und wieder zurück in die Datenbank gespeichert. Dadurch wird eine lemmabasierte Suche ermöglicht, die die unterschiedlichen Schreibweisen und Flektionsformen bei der Suche abfangen kann.[18]

Abgesehen von der Suche ist vor allem die Chronologie zu Alexander von Humboldts Leben ein zentraler Vernetzungs- und Integrationspunkt externer Ressourcen. Diese Chronologie wurde bereits in den 1960er Jahren an der damaligen Deutschen Akademie der Wissenschaften zu Berlin begonnen, in den 2000ern überarbeitet als HTML-Version im Web veröffentlicht und 2015/16 schließlich in TEI-XML überführt und in die *edition humboldt digital* integriert. Dort wird sie fortlaufend gepflegt, erweitert, mit externen Quellen sowie mit den edierten Texten und Registereinträgen verlinkt. Darüber hinaus integriert sie automatisiert verschiedene externe Angebote und Dienste in die Edition, wie z. B. die Metadaten der publizierten Korrespondenz Alexander von Humboldts aus correspSearch[19], Schriften Humboldts aus dem Deutschen Textarchiv[20] und Einträge aus dem Hofkalendarium der preußischen Monarchie[21] mit Bezug zu Humboldts Leben (siehe Abb. 1). Dadurch öffnet die Chronologie die *edition humboldt digital* nach außen hin zu zahlreichen extern vorliegenden Materialien und Informationen.

Auch in den Registern werden externe Daten nachgenutzt. Zum einen werden die BEACON-Schnittstellen anderer ausgewählter digitaler Publikationen, wie die Kosmos-Vorlesungen Alexander von Humboldts im Deutschen Textarchiv, abgerufen und automatisiert verknüpft.[22] Darüber hinaus werden ganze Datensätze der

GND (also nicht nur die bloßen URIs) nachgenutzt. Mit ihrer Hilfe werden die Registereinträge der *ehd* automatisiert *untereinander* verlinkt – nämlich anhand der in der GND notierten familiären und freundschaftlichen Beziehungen. Außerdem können so Porträts von Wikimedia eingebunden werden. Ein Register, das in besonders großem Maße auf externe Dienste zurückgreift, ist das Pflanzenregister. Hier liegt in der Edition die Besonderheit vor, dass die Pflanzen nicht mit eigenen Registereinträgen versehen werden, sondern dieses Register ausschließlich automatisch über die taxonomischen Namen generiert wird. Dazu werden alle Pflanzennamen in den Transkriptionen von den Editor:innen auf ihren regulären Namen ergänzt oder korrigiert (natürlich nachverfolgbar). Anhand der taxonomischen Namen werden dann verschiedene Webservices abgefragt und automatisiert verknüpft. Damit öffnet die Edition v. a. die Tagebücher Humboldts für verschiedene Disziplinen wie die Biodiversitätsforschung.

Mit der intensiven Nachnutzung externer Webservices und Daten erhöht sich der Nutzen einer digitalen Edition signifikant. Gleichzeitig stellt diese Nachnutzung neue Probleme und Herausforderungen an die Entwicklung und den Betrieb digitaler Editionen, da externe Dienste sich grundsätzlich ändern können (aktualisierte Schnittstellen, Änderungen im Format etc.). Das - und Fragen der Performance - führt dazu, dass diese externen Daten auch in der *edition humboldt digital* vorgehalten werden müssen. Fraglich ist dann aber weiter, ob und in welchem Rahmen diese externen Daten auch in der Datenpublikation mitveröffentlicht werden können und müssen. Darüber hinaus kann man nicht garantieren, dass ein externer Dienst auch in Zukunft verfügbar sein wird. Das ist insbesondere ein Problem vor dem Hintergrund der relativ langen Laufzeit des Projekts: Werden die externen Services zur Anreicherung von Texten und Daten, die erst in den nächsten Jahren hinzukommen, noch vorhanden sein? Der Vortrag möchte am Beispiel der *edition humboldt digital* diese und weitere Herausforderungen und Chancen einer 'offenen Edition' vorstellen und die damit zusammenhängenden, skizzierten Themenfelder Datenpublikation, Bereitstellung und Nachnutzung von APIs und externen Daten sowie Transparenz im Editions- und Forschungsprozess diskutieren.

# Fußnoten

1. Projektbeschreibung auf den Seiten der BBAW: http://www.bbaw.de/forschung/avh-r/uebersicht (Zugriff für alle im Abstract angegebenen Links: 03. August 2022); *edition humboldt digital* : https://edition-humboldt.de/ . Vgl. dazu Kraft/Dumont 2020; zu dem auch im vorliegenden Abstract zentralen Aspekt der Vernetzung siehe die Zusammenfassung und Illustration dieses Ansatzes in Kraft/Dumont 2017.
2. CC-BY-SA 4.0 ( https://creativecommons.org/licenses/by-sa/4.0/ ) für die TEI-XML-Daten; CC-0 für die Register-Einträge und Metadaten zu den Dokumenten.
3. Buchreihe *edition humboldt print* , https://www.springer.com/series/16345 . 2020 erschien der erste Band über die *Geographie der Pflanzen* , herausgegeben von Ulrich Päßler. 2022 wurde Band 1 der Amerikanischen Reisetagebücher von Carmen Götz 2022 herausgegeben, gefolgt von Band 1 der Russisch-Sibirischen Reisetagebücher, herausgegeben von Tobias Kraft und Florian Schnee, in 2023.
4. https://www.ediarum.org/ ; zur Erfassungssoftware siehe Dumont et. al 2021.
5. Siehe für eine Orientierung zum Editionsmodell der *ehd* Sahle 2016 sowie insbesondere zum Konzept der 'documentary edition' beispielsweise Pierazzo 2011.
6. Vgl. den Überblick zur Version 8 sowie den vorhergehenden Versionen der *ehd* unter https://edition-humboldt.de/H0020382 ; API: https://edition-humboldt.de/about/index.xql?id=api ; TEI-XML (der jeweils aktuellen Version der *ehd* ) https://edition-humboldt.de/api/v1.1/tei-xml.xql .
7. https://www.deutschestextarchiv.de/doku/basisformat/ .
8. Editionsrichtlinien der *ehd* , v. 1.1.2 (9.5.2022), https://edition-humboldt.de/richtlinien/index.html .
9. So orientiert sich beispielsweise das Akademienvorhaben *Propyläen: Goethes Biographica* ( https://goethe-biographica.de/ ) im Zuge seiner Entwicklung eines TEI-XML-Datenmodells für diese Hybrid-Edition v. a. für die Briefe von und an Goethe sowie dessen Tagebücher an den Richtlinien der *ehd* .
10. Z. B. das TEIC/TEI Issue #2028 "@calendar should allow multiple values", https://github.com/TEIC/TEI/issues/2028 , das aufgrund der Diskussion auf der TEI-Mailingliste (August 2020) entstand und zur Implementierung in die TEI P5 Guidelines v. 4.3.0 (2021-08-31) führte.
11. Mit Version 8 wurde dieses Feature erstmals umgesetzt, zunächst nur bei den Tagebüchern. Zu den Editionsstufen siehe https://edition-humboldt.de/richtlinien/ediarum.AVHR/editionsstufen.html .
12. https://edition-humboldt.de/H0020382 .
13. Von Axelle Lecroq wurde dafür die bereits seit Version 1 vorhandene API optimiert und ein entsprechendes Skript zum Abruf der Daten entwickelt.
14. https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html .
15. Von derzeit 9022 vorliegenden Personenregistereinträgen sind 5581 mit einer GND-URI versehen, das entspricht rund 61% aller Personen. 663 Datensätze verfügen nur über eine VIAF-URI. Damit verfügen rund 40% über gar keine Norm-ID – i. d. R., weil kein Normdatensatz vorhanden ist.
16. Ein sehr kleiner Anfang konnte dennoch schon gemacht werden: In Wikidata wurden seitens der Freiwilligen dort bei den entsprechenden Personen die Perma-IDs der *ehd* eingetragen, siehe z.B. https://www.wikidata.org/wiki/Q132197 . Das war möglich, weil die ehd konsequent GND-IDs zu den Personen einträgt, falls vorhanden.
17. https://www.deutschestextarchiv.de/cab/ .
18. Die Suchfunktionalität auf der *ehd* -Seite wurde im Zuge der Version 8 (2022) grundlegend überarbeitet; siehe zur Einführung "ehd – explained. Kapitel 4: Die Suche" von Tobias Kraft, aufgenommen beim Humboldt-Tag am 16. September 2022 in der BBAW, verfügbar unter https://youtu.be/l1D0zGd7osA .
19. https://correspsearch.net/de/suche.html?s=http://d-nb.info/gnd/118554700 .
20. https://edition-humboldt.de/chronologie/index.xql?jahr=1827&dta=on .

21. Vgl. Einleitung Hofkalendarium, https://actaborussica.bbaw.de/v5/P0006298 , in Akademienvorhaben 2021.
22. Siehe z. B. den Registereintrag zu August Böckh: https://edition-humboldt.de/H0003413 .

# Bibliographie

**Akademienvorhaben Anpassungsstrategien der späten mitteleuropäischen Monarchie am preußischen Beispiel (1786-1918) (Hg.).** 2021 . *Die Sprache der Monarchie (Version 5)*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. URL: https://actaborussica.bbaw.de/

**Dumont, Stefan und Susanne Haaf, Tobias Kraft, Alexander Czmiel, Christian Thomas, Matthias Boenig.** 2016. "Applying Standard Formats and Tools: 'Alexander von Humboldt auf Reisen' as an Example for the Collective Subsequent Use of DTABf and ediarum". Vortrag, *TEI Conference and Members' Meeting*, Vienna. Abstract (PDF): https://www.tei-c.org/Vault/MembersMeetings/2016/sites/default/files/TEIconf2016_BookOfAbstracts.pdf , 69–70 (zugegriffen: 03. August 2022).

**Dumont, Stefan und Nadine Arndt, Sascha Grabsch, Lou Klappenbach.** 2021. *ediarum.BASE.edit (Version 2.0.0)* [Computer software]. https://doi.org/10.5281/zenodo.5897100 (zugegriffen: 03. August 2022).

**Stefan Dumont, Sascha Grabsch und Jonas Müller-Laackman (Hg.).** 2021. *correspSearch – Briefeditionen vernetzen (2.0.0)* [Webservice]. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. https://correspSearch.net (zugegriffen: 03. August 2022).

**Ette, Ottmar (Hg.).** 2022. *edition humboldt digital* (Version 8). Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. https://edition-humboldt.de/ (zugegriffen: 03. August 2022).

**Ette, Ottmar und Stefan Dumont, Annika Geiser, Carmen Götz, Tobias Kraft, Ulrike Leitner, Ulrich Päßler, Florian Schnee, Christian Thomas (Hg.).** 2022. *TEI-XML-Datenset der Tagebücher, Briefe, Dokumente, Forschungsbeiträge, Chronologieeinträge und Register der edition humboldt digital (Version 8)*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. URL: https://github.com/telota/edition-humboldt-digital

**Jurish, Bryan.** 2011. *Finite-State Canonicalization Techniques for Historical German*. Potsdam: Universität Potsdam. urn:nbn:de:kobv:517-opus-55789 (zugegriffen: 03. August 2022).

**Kraft, Tobias und Stefan Dumont.** 2017. *Edition humboldt digital vernetzt*, Poster. Zenodo. http://doi.org/10.5281/zenodo.1035134 (zugegriffen: 03. August 2022).

**Kraft, Tobias und Stefan Dumont.** 2020. "The Humboldt Code. On creating a hybrid digital scholarly edition of a 19 th century globetrotter." In *Wiener Digitale Revue* 1. https://doi.org/10.25365/wdr-01-03-02 (zugegriffen: 03. August 2022).

**Pierazzo, Elena.** 2011. "A Rationale of Digital Documentary editions". In *Literary and Linguistic Computing*, 26,4, 463-477. https://doi.org/10.1093/llc/fqr033 (zugegriffen: 03. August 2022).

**Sahle, Patrick.** 2016. "What is a Scholarly Digital Edition?" In *Digital Scholarly Editing: Theories and Practices*, hg. von Matthew James Driscoll und Elena Pierazzo, 19-39. Cambridge, UK: Open Book Publishers. https://books.openedition.org/obp/3397 (zugegriffen: 03. August 2022).

**Thomas, Christian und Susanne Haaf.** 2016-2019. "Enabling the Encoding of Manuscripts within the DTABf: Extension and Modularization of the Format". In *Journal of the Text Encoding Initiative* [Online], Issue 10 | December 2016 - July 2019. https://doi.org/10.4000/jtei.1650 (zugegriffen: 03. August 2022).

**Vogeler, Georg.** 2019. „The 'Assertive Edition'. On the Consequences of Digital Methods in Scholarly Editing for Historians". *International Journal of Digital Humanities* 1 (2): 309–22. https://doi.org/10.1007/s42803-019-00025-5 .