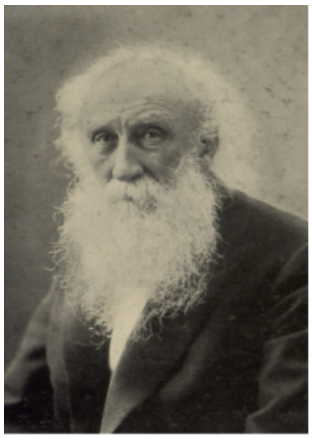
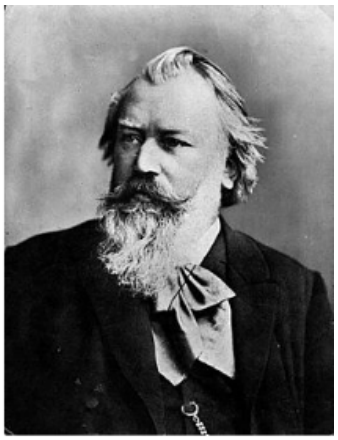
Musikwissenschaft Münster

Universität Münster

## Zertifikat „Digital Humanities“

# Eine explorative Studie zur digitalen Aufbereitung analoger Briefeditionen am Beispiel des Brahms-Grimm-Briefwechsels





Vorgelegt am 22. Februar 2024 von:

Maximilian Greshake
Matr.-Nr.: 440160
m_gres09@uni-muenster.de

# Inhaltsverzeichnis

# 1 Einführung

Briefe und ähnliche Korrespondenzen bieten wertvolle Informationen zur Erforschung von Personen des öffentlichen Lebens, seien es biografische, soziokulturelle, ästhetische oder politische Aspekte. Das Briefeditionswesen hat daher für verschiedene wissenschaftliche Disziplinen einen hohen Stellenwert. Entsprechend groß ist das Interesse an qualitativ hochwertigen, möglichst vollständig erfassten Briefwechseln, wobei hierbei zwischen dem Verzeichnis und der Edition zu differenzieren ist. Während ein Briefverzeichnis primär die Erfassung und Katalogisierung sämtlicher existierender Korrespondenzen anstrebt – also gewissermaßen ein besonderes Augenmerk auf die Metadaten wirft –, sieht die Edition hingegen eine vollständige Transkription des Korrespondenztextes vor, mitsamt wissenschaftlicher Einordnung – d. h. Ergänzung von kontextuellen Informationen und Beschäftigung mit dem Textinhalt. Briefeditionen stellen in den Literaturwissenschaften ein festes, primäres Forschungsfeld dar. In musikwissenschaftlichen Forschungskontexten erscheinen sie dagegen häufiger als parallele, supplementäre Elemente, etwa im Zuge von Werkgesamtausgaben[1], oder im Kontext der Erforschung von Komponistenbiografien.

In den letzten Jahrzehnten war das Editionswesen grundlegenden Veränderungen unterworfen. Durch die Digitialisierung und das Aufkommen von Auszeichnungssprachen – also maschinenlesbaren Sprachen für die Gliederung und Formatierung von Textdaten – eröffneten sich zahlreiche neue Möglichkeiten der Erfassung, Katalogisierung, Edition und Darstellung von lexikographischen Texten. Im Zuge dessen etablierte sich das XML-Format TEI[2] als Standard für digitale Briefeditionen. Für Musikalische Notentexte wurde derweil das XML-Format MEI[3] zum wissenschaftlichen Standard. Jedes TEI-Dokument enthält umfangreiche Metadaten zum jeweiligen Dokument, die eine schnelle Auffindbarkeit ermöglichen und weitere Informationen bereithalten. Durch den Kodierungsstandard TEI ist zugleich die einheitliche Verwertung der Texte gewährleistet. Nicht zuletzt stehen dadurch zahlreiche Analyse- und Auswertungsmethoden zur Verfügung. Dies beginnt bei der vielseitigen Durchsuchbarkeit durch Query-Languages (Pfadbeschreibungssprachen wie xPath) und endet bei maschinengestützten Inhaltsanalysen, etwa durch Large Language Modelle (Sprachmodelle wie ChatGPT oder Google Bard).

Ein mustergültiges Beispiel für eine digitale Briefedition im musikwissenschaftlichen Kontext stellt die Carl Maria von Weber Gesamtausgabe dar.[4] Sie umfasst unter anderem Verzeichnis und Edition der Korrespondenzen Webers sowie Drittbriefe und einen lexikalischen Apparat zur Auszeichnung von Personen, Werken, Orten und weiteren Aspekten.

[1] Siehe beispielsweise die Haydn Briefausgabe, https://www.haydn-institut.de/index.php/gesamtausgabe/briefausgabe, letzter Aufruf am 22. Februar 2024.
[2] Siehe https://tei-c.org/.
[3] Siehe https://music-encoding.org/.
[4] Siehe https://weber-gesamtausgabe.de/de/A002068.html#correspondence.

Dadurch findet eine weitreichende Verlinkung zwischen den Dokumenten statt, die sogleich ein vernetztes Erforschen der Dokumente begünstigt.

Der Begriff des „Digitalen“ erfährt mitunter vielschichtige Deutungen. Es fällt daher oft schwer, genuine digitale Editionen als solche zu dinstinguieren. Das Scannen, also die optomechanische Erfassung eines materiellen, manifesten Dokuments, ist terminologisch zwar die „Digitalisierung“ eines Dokuments. Der Scan einer analogen Edition stellt jedoch zunächst keine „digitale“ Edition dar; es gilt also: digitalisiert $\neq$ digital. Da „echte“ digitale Editionen ihren Editionsgegenstand von vornherein anders modellieren, unterscheiden sich auch die zur Verfügung stehenden Nutzungsformen. Aber auch ein Text im PDF-Format macht noch keine digitale Edition. Ebenso wenig stellt er einen „kodierten“, also vollständig maschinenlesbarer Text dar. Dies ist erst ansatzweise mit der Übertragung in eine Auszeichnungssprache wie HTML, XML oder TEI-XML erreicht. Erst dann kann der Computer eine Anrede, ein Datum oder einen Ort als solchen „verstehen“; und erst dann gelten die Informationen für geisteswissenschaftliche Kontexte als operationalisiert. Die Kodierung von Metadaten wiederum ermöglicht die Suche nach Korrespondenzen im Internet, beispielsweise qua CorrespSearch[5]. Die Berücksichtigung von (Kodierungs-)Standards ist demnach bei digitalen Editionen von noch größerer Bedeutung als es in der herkömmlichen Editionspraxis der Fall war. Zuletzt unterscheiden sich digitale Editionen durch neuartige Recherche- und Visualisierungsmöglichkeiten, die mittels anpassbarer Benutzeroberfläche ein individuelleren Zugriff erlauben. So bieten etwa dynamische Editionen im Vergleich zu herkömmlichen, analogen Editionen eine Vielfalt an Visualisierungsoptionen, die mit einem gedruckten Buch nie realisiert werden könnten. Zusammengefasst kann also erst von einer tatsächlichen „digitalen Edition“ gesprochen werden, wenn das Editionsprojekt ein Datenmodell bzw. ein zugeschnittenes Kodierungsschema vorweisen kann und mit weiteren digitalen Methoden bzw. Werkzeugen arbeitet.

Digitale Briefeditionen besitzen gegenüber ihren analogen Pendants in vielerlei Hinsicht einen Vorteil. Dementsprechend sinnvoll erscheint es, bereits bestehende analoge Ausgaben auch im digitalen Format vorliegen zu haben, um ihre Nutzungsfrequenz sowie ihre Qualität zu potenzieren. Die Dokumente von Grund auf neu, digital zu edieren ist jedoch mit einem hohen zeitlichen Aufwand und einem großen finanziellen Investment verbunden. Als Ausweichmöglichkeit bietet sich daher an, die bereits bestehenden Texte zu scannen, automatisch per OCR in TEI zu kodieren und so weit wie möglich für den Gebrauch herzurichten, also „digital aufzubereiten“. Ziel wäre die Erstellung eines erweiterbaren Dokuments, welches auch nachträglich noch durch Archivinformationen, weitere Digitalisate und Korrekturen ergänzt werden kann.

Die Intention der vorliegenden Studie ist es, den Prozess einer solchen digitalen Aufbereitung am Beispiel einer bestehenden Briefausgabe zu erproben und dabei operable Methoden zu explorieren. Als Gegenstand wurde der Briefwechsel zwischen den Komponisten Johannes Brahms (1833-1897) und Julius Otto Grimm (1827-1903) gewählt, welcher von Richard Barth (1850-1923) ediert und 1908 (2. Auflage 1912) im Zuge des

[5] Ein Verzeichnis für Briefverzeichnisse, siehe https://correspsearch.net/de/start.html.

vierten Bandes des Johannes Brahms Briefwechsels herausgegeben wurde.[6] Die Studie versteht sich nicht als genuine digitale Briefedition und erhebt keine Ansprüche, wissenschaftliche Korrekturen oder Ergänzungen vorzunehmen, da sie den 1912 veröffentlichten Text zunächst unverändert übernimmt. Vielmehr sollen Möglichkeiten dargelegt werden, wie bestehende analoge Briefeditionen in digitalisierter Form für den Forschungsbetrieb oder für außeruniversitäre Interessenten zur Verfügung gestellt werden können. Die vereinfachte Durchsuchbarkeit sowie die Kompatibilität mit etablierten Kodierungsschematas bestehender digitaler Briefeditionen stehen dabei im Vordergrund. Als wesentliche Orientierungshilfe werden dafür die Editionsrichtlinien der Weber Gesamtausgabe hinzugezogen.[7] Einige zu erwartene Abschnitte einer digitalen Briefedition werden jedoch unweigerlich fehlen; vor allem der Metadaten-Apparat wird stark verkürzt sein.

## 1.1 Versuchsaufbau

Die Studie erfolgt anhand von drei Schritten. Sämtliche Dateien, die im Laufe der Arbeitsprozesse ausgegeben werden, sind über GitHub einsehbar und können dort heruntergeladen werden. Gleiches gilt für Skripte und sonstige Transformationsdateien.[8] Eine Bebilderung der einzelnen Schritte dokumentiert den Versuchsablauf.

Zunächst wird die Briefwechsel-Ausgabe in ihrer Buchform gescannt und im PDF-Format gespeichert. Diese Datei wird dann in Einzelseiten/-dateien aufgeteilt und auf der Plattform Transkribus hochgeladen, wo sie per OCR (Optical Character Recognition) transkribiert wird.[9] Der übertragene Text kann schließlich im TEI-Format ausgegeben werden. Im zweiten Schritt wird die TEI-Datei im Oxygen XML-Editor geöffnet und bereinigt. Neben der Verwendung der Suchen/Ersetzen-Funktion wird auch ein XSLT-Dokument (XSL-Transformation-Sheet) angelegt, um komplexere Eingriffe vorzunehmen. Ziel dieser Datenbereinigung ist es, überflüssige Codezeilen zu entfernen und den gewünschten Kerntext der Briefe zu filtern. Sodann wird die Aufbereitung des Textes vorgenommen. Die Briefwechselausgabe beinhaltet eine ausführliche Einleitung sowie insgesamt 128 Briefe. Diese werden vollständig nach einer formalen Vorlage (Template) kodiert, welches ausführlich besprochen wird. Zur abschließenden Bereinigung wird nochmals ein XSLT-Dokument genutzt, um nachträgliche Hinzufügungen oder Löschungen vorzunehmen.

Limitationen und aufkommende Probleme werden bereits während des Versuchsablaufs erwähnt. In den Schlussbemerkungen werden sie aber noch genauer besprochen und eingeordnet. Zum Abschluss thematisiert die Studie kurz mögliche Wege der Veröffentlichung und Vernetzung der aufbereiteten Briefedition. Als vorläufiges Ergebnis wird – erneut per XSLT – eine einfache PDF-Version erstellt.

[6] *Johannes Brahms im Briefwechsel mit J. O. Grimm* (= Brahms-Briefwechsel IV, hrsg. von Richard Barth), Berlin, 2. Aufl. 1912.

[7] Siehe Editionsrichtlinien zur Ausgabe der Briefe, Tagebücher und Dokumente Webers, https://weber-gesamtausgabe.de/de/Projekt/Editionsrichtlinien_Text.html, letzter Aufruf am 22. Februar 2024.

[8] Siehe https://github.com/maxgreshake/bgbw_digital.

[9] Siehe https://readcoop.eu/de/transkribus/.

# 2 Digitale Aufbereitung des Brahms-Grimm-Briefwechsels

Richard Barth war zu Beginn seiner Musikerlaufbahn in Münster als Konzertmeister tätig, wo er eine annäherungsweise familiäre Beziehung zu dem damaligen Musikdirektor Julius Otto Grimm aufbaute. Infolgedessen kam er auch in engen Kontakt mit Johannes Brahms, einem langjährigen Freund Grimms. Zu Beginn des 20. Jahrhunderts widmete Barth sich der Sammlung und Edierung der Briefe, die zwischen Grimm und Brahms verschickt wurden. Einige der edierten Briefe gelten heute als verschollen – die Briefausgabe stellt in diesen Fällen die einzige Überlieferung dar. Sein Stand als Zeitgenosse der beiden Schreiber kommt durchaus in seinen Anmerkungen innerhalb der Briefausgabe zum Ausdruck. Neben einem ausführlichen Vorwort zu beiden Schreibern macht Barth in vielen Briefen zusätzliche Anmerkungen zu Personen, Werken und weiteren Hintergründen der Briefinhalte. Diese Fußnoten werden als Text übernommen und erscheinen in der digital aufbereiteten Version als `<note>` Element (siehe Anmerkungen zu den Editionsrichtlinien in Kapitel 2.3).

Der als PDF gescannte Briefwechsel beinhaltet drei Bilder, ein Titelblatt, eine Einleitung, 128 Briefe, ein Namensregister sowie eine Liste der Werke Julius Otto Grimms. Mit der Aufteilung des Gesamttextes in einzelne Seiten sind die Bilder, das Namensregister, die Werkliste (Supplement) sowie vereinzelte Leerseiten aussortiert worden. Die digital aufbereitete Version besteht demnach nur aus Titelblatt, Einleitung und den Briefen. Die Übernahme des analogen Namensregisters ist nichtig, da das finale TEI Dokument von sich aus ausführlichere Suchfunktionen zur Verfügung stellt. Somit werden insgesamt 171 Seiten bzw. Bilddateien verarbeitet.

## 2.1 Texterfassung per OCR

Die automatisierte Text- bzw. Schrifterkennung bedeutet die computergestützte Identifizierung von Texten respektive Textzeichen innerhalb von Bilddateien. Neben dem ursprünglichen OCR-Verfahren (Optical Character Recognition) können neuronale Netzwerke mittlerweile auch ganze Zeilen oder Textblöcke statt einzelner Zeichen verarbeiten. Handschriftliche Texte, die allgemein komplizierter und fehleranfälliger als gedruckte Schrift sind, werden durch sogenannte HTR-Modelle (Handwritten Text Recognition) verarbeitet. In der Regel liegen analoge Briefausgaben in gedruckter Form vor. Der hier aufbereitete Briefwechsel ist in Deutscher Fraktur gedruckt und kann daher von einem einfachen E-Reader nicht durchsucht werden.

Die vorliegende Studie bedient sich der Plattform Transkribus, um die Bilddateien in digitalen Text umzuwandeln. Transkribus verwendet verschiedene Modelle künstlicher

Intelligenz, darunter auch die Layout-Analyse zur näheren Identifizierung textinterner Strukturen. Den Kern von Transkribus stellen die durch „Deep-Learning“ trainierten Texterkennungsmodelle dar. Es können fertige Modelle verwendet werden oder auf Basis der eingespeisten Texte eigene Modelle trainiert werden. Für den vorliegenden Text wurde das Modell *Danish Fraktur SB 19th century PyLaia* ausgewählt (siehe Abb. 1). Zwar ist das Modell lediglich auf 390226 Wörtern trainiert worden, jedoch ist es für den Drucktyp am geeignetsten und besitzt mit 0,5% eine akzeptable Fehlerquote. Im Schnitt identifiziert das Modell also jedes zweihundertste Zeichen falsch.

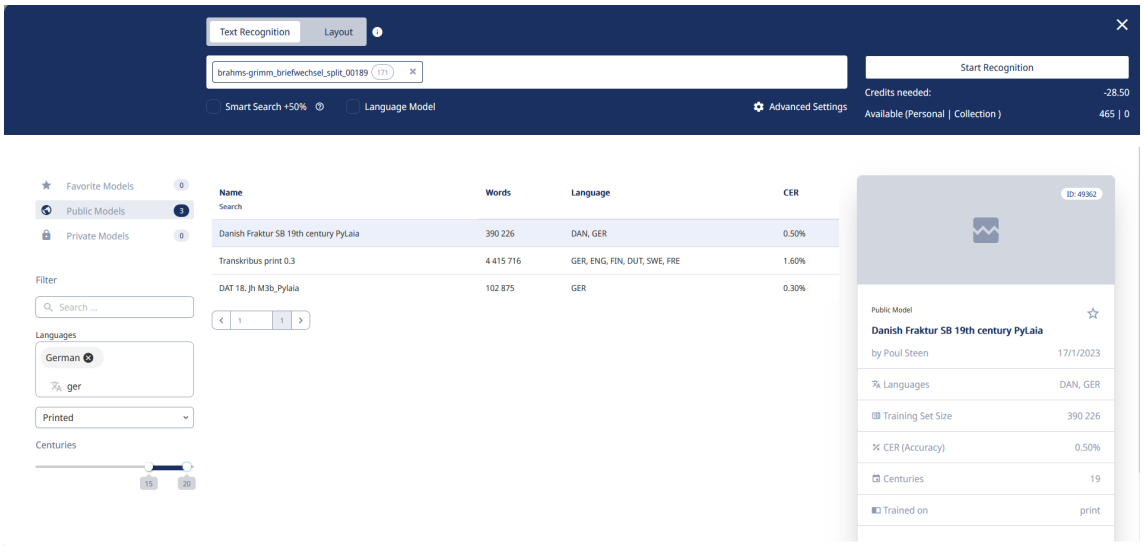



Abb. 1: Transkribus, Anwendung eines Modells

Die Erkennung des Datenbestandes erfolgt in einem Durchlauf (siehe Abb. 2). Neben der Übertragung in digitalen Text nimmt Transkribus auch eine Erkennung der Zeilen vor und ordnete Textabschnitte in sogenannte „Regionen“. Leider sind diese Regionen durchweg arbiträr gesetzt und als überflüssig zu bewerten. Die Hoffnung, bereits korrekt kodierte Absätze zu erhalten, hat sich demnach nicht erfüllt. Auch die Zeilenerkennung ist für die Brieftexte irrelevant. Der transkribierte Text wird im Anschluss im TEI-XML-Format heruntergeladen (siehe *bgbw_output.xml*).

## 2.2 Datenbereinigung

Zur Aufbereitung der Output-Datei wird der Oxygen XML-Editor (Version 24) verwendet. Das Dokument enthält zunächst für für jede transkribierte Seite ein `<facsimile/>` Element, das den übertragenen Text mit der zugrundeliegenden Bilddatei verknüpft. Die ursprünglichen Zeilenumbrüche werden im Zuge dessen in `<pb>` Elementen (Page Beginning) kodiert. Diese Facsimile-Verknüpfung kann für eine spätere synoptische Darstellung nützlich sein, um stets auf die analoge Vorlage verweisen zu können. In diesem Fall werden sie zunächst belassen, dann aber nach der Textaufbereitung aus dem Dokument entfernt.

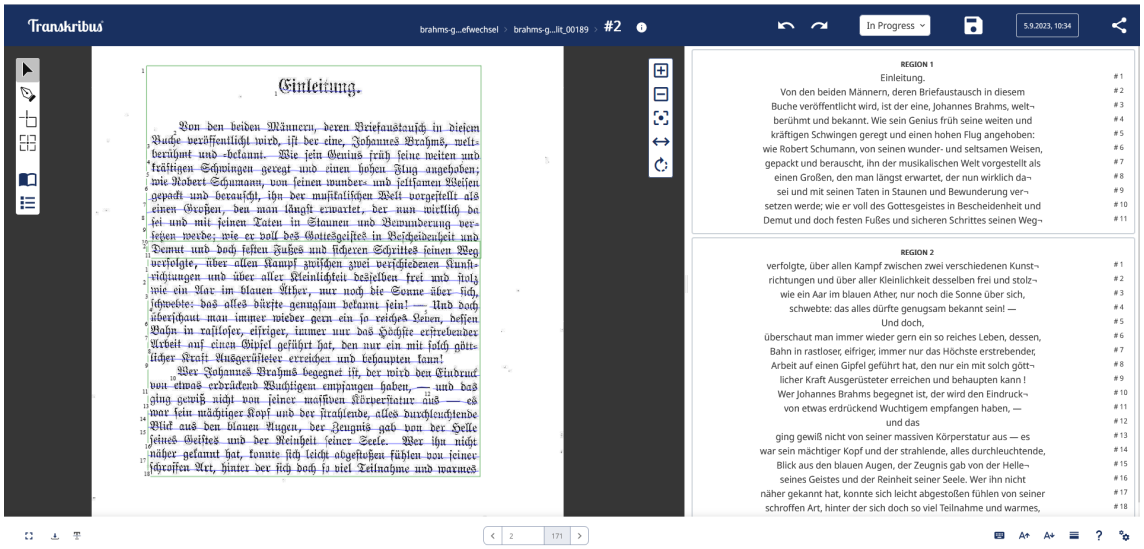



Abb. 2: Transkribus, Texterkennung

Das Dokument enthält noch weitere überflüssige Codezeilen. Für die erwähnten Regionen sowie für jede Zeile erstellt Transkribus ein eigenes `<zone/>` Element. Diese Elemente stellen allerdings eine Überfrachtung dar und können entfernt werden (siehe Abb. 3).

```
<facsimile xml:id='facs_1'>
    <surface ulx='0' uly='0' lrx='1032' lry='1530' corresp='0001_brahms-grimm_briefwechsel_split_00010.jpg'>
        <graphic url='0001_brahms-grimm_briefwechsel_split_00010.jpg' width='1032px' height='1530px'/>
        <zone points='206,230 806,230 806,878 206,878' rendition='TextRegion' xml:id='facs_1_tr_1'>
            <zone points='202,321 258,294 279,314 514,294 671,322 804,305 802,238 610,237 575,204 507,235 253,2
            <zone points='311,390 459,403 692,392 691,370 669,355 605,370 586,354 504,365 462,344 392,346 369,3
            <zone points='354,505 390,491 646,505 645,457 495,438 454,460 416,437 384,453 353,431' rendition='L
            <zone points='402,690 463,689 492,699 499,693 517,698 526,690 560,698 571,691 571,665 487,669 466,6
            <zone points='465,737 504,737 509,731 513,731 513,717 464,717' rendition='Line' xml:id='facs_1_tr_1
            <zone points='356,796 414,810 487,802 495,794 602,810 618,804 617,766 598,757 566,765 516,755 506,7
            <zone points='318,863 410,862 457,874 559,866 653,882 652,844 532,841 501,830 459,833 446,846 414,8
        </zone>
        <zone points='221,1076 740,1076 740,1352 221,1352' rendition='TextRegion' xml:id='facs_1_tr_2'>
            <zone points='371,1138 404,1136 422,1146 436,1134 487,1131 502,1146 577,1139 621,1154 640,1142 639,
            <zone points='416,1192 463,1193 475,1203 488,1202 496,1194 527,1194 540,1188 539,1170 509,1170 501,
            <zone points='216,1227 738,1241 719,1197 691,1216 675,1202 653,1215 468,1209 408,1190 350,1206 215,
            <zone points='421,1278 448,1282 456,1290 468,1287 471,1290 518,1290 518,1249 467,1250 456,1245 432,
            <zone points='438,1341 443,1341 449,1346 484,1346 494,1354 505,1354 504,1326 494,1326 487,1319 474,
        </zone>
    </surface>
</facsimile>
```

Abb. 3: XML-Dokument, Faksimile-Verknüpfung der Titelseite

Gleichermaßen müssen die Element-Tags für die Regionen und Zeilen im Text selbst entfernt werden; sie teilen sich auf in `<p/>` (paragraph), `<lg/>` (line group) und `<l/>` (line) (siehe Abb. 4). Ziel der Bereinigung ist es demnach, einen kohärenten Fließtext hervorzubringen. Da die Output-Datei aus 14848 Codezeilen besteht, ist eine automatisierte Entfernung der entsprechenden Element-Tags unabdingbar. Verwendet wird dafür ein XSLT-Dokument (siehe *bgbw_transformation.xsl*). Kurz gefasst kopieren XSL-Transformationen eine vorliegende XML-Datei und manipulieren dabei Inhalte unter festgelegten Parametern (Löschen, Ersetzen, Verschieben, Ergänzen). Die Änderungen

werden direkt im XSLT dokumentiert. Da Transkribus häufig Striche (Trenn- bzw. Bindestriche, Halbgeviert- und Langstriche) mit dem Negationszeichen (¬) transkribiert, muss dies zusätzlich korrigiert werden. Vergleichsweise einfache Probleme können auch mit der Suchen/Ersetzen-Funktion des XML-Editors gelöst werden. Das Ergebnis der Datenbereinigung wird vor der Aufbereitung zwischengespeichert (siehe *bgbw_cleaned.xml*).

```
7375        <pb facs='#facs_2' xml:id='0002_brahms-grimm_briefwechsel_split_00012.jpg' n='2'/>
7376        <p facs='#facs_2_tr_2'>
7377            <lg>
7378                <l facs='#facs_2_tr_2_tl_1'>Einleitung.</l>
7379                <l facs='#facs_2_tr_2_tl_2'>Von den beiden Männern, deren Briefaustausch in diesem</l>
7380                <l facs='#facs_2_tr_2_tl_3'>Buche veröffentlicht wird, ist der eine, Johannes Brahms, welt¬</l>
7381                <l facs='#facs_2_tr_2_tl_4'>berühmt und bekannt. Wie sein Genius früh seine weiten und</l>
7382                <l facs='#facs_2_tr_2_tl_5'>kräftigen Schwingen geregt und einen hohen Flug angehoben:</l>
7383                <l facs='#facs_2_tr_2_tl_6'>wie Robert Schumann, von seinen wunder- und seltsamen Weisen,</l>
7384                <l facs='#facs_2_tr_2_tl_7'>gepackt und berauscht, ihn der musikalischen Welt vorgestellt als</l>
7385                <l facs='#facs_2_tr_2_tl_8'>einen Großen, den man längst erwartet, der nun wirklich da¬</l>
7386                <l facs='#facs_2_tr_2_tl_9'>sei und mit seinen Taten in Staunen und Bewunderung ver¬</l>
7387                <l facs='#facs_2_tr_2_tl_10'>setzen werde; wie er voll des Gottesgeistes in Bescheidenheit und</l>
7388                <l facs='#facs_2_tr_2_tl_11'>Demut und doch festen Fußes und sicheren Schrittes seinen Weg¬</l>
7389            </lg>
7390        </p>
```

Abb. 4: XML-Dokument, Code der ersten Einleitungsseite

## 2.3 Aufbereitung der TEI-Kodierung

Die Briefausgabe liegt vor dem finalen Aufbereitungsschritt in einem bereinigten Fließtext in XML vor, hat die 171 Seiten in `<facsimile/>` Elementen verknüpft und ist mit einem provisorischen TEI-Header versehen. Die Aufgabe besteht nun darin, die einzelnen Briefe aus dem transkribierten Fließtext herauszunehmen und in eine feste Briefstruktur – im Folgenden als „Template“ bezeichnet – zu übertragen. Grundsätzlich bestehen alle XML-Dokumente aus zwei Bereichen, den Metadaten (Informationen zur Datei und ihren Inhalten) sowie dem eigentlichen Textkorpus.[1] Standardmäßig besitzt jeder einzelne Brief einen eigenen Header, in welchem die Metadaten, also die Informationen zu Quellenmaterial, Archiven, Schreiber, Adressaten, Editoren usw. vermerkt sind. Wie eingangs besprochen, sieht der vorliegende Versuch eine digitale Aufbereitung und keine digitale Neu-Edierung der Briefe vor. Demnach muss sich auch die Kodierung an der analogen Vorlage orientieren, das heißt in diesem Fall gibt es nur einen einzigen Header für die gesamte Briefausgabe. Die Metadaten der Briefe müssen demnach weitestgehend aus dem jeweiligen Textkorpus extrahierbar sein.

Der TEI-Header ist in vier große Bereiche unterteilt (Siehe Abb. 5).[2] Die Weber Editionsrichtlinien schreiben: „Viele der in diesem Bereich notwendigen Angaben zum elektronischen Dokument beziehen sich in einer Edition auf alle integrierten Texte und sind lediglich redundant (und damit kopierbar) in allen Einzeldokumenten abgelegt.“[3] Unter dieser Prämisse würde das Fehlen des Headers für die hier aufbereiteten Briefe nicht

[1] Weber Brief-Editionsrichtlinien, 3.1 Technische Vorbemerkung.
[2] Vgl. ebd., 4. Die Metadaten der Texte (Apparatteil).
[3] Ebd., 4.1 Die Grobgliederung des Metadatenteils.

| | |
|---|---|
| <fileDesc/> | file Description: Die genaue inhaltliche Beschreibung des vorliegenden Datensatzes (Titelei, Zugänglichkeit, Quellen usw.) |
| <encodingDesc/> | encoding Description: Informationen zur Einrichtung des vorliegenden Datensatzes als digitalem Dokument (Projektbeschreibung) |
| <profileDesc/> | profile Description: Angaben zum (historischen) „Profil“ des Textes (Entstehung, Datierung, Textsorte, Sprache usw.) |
| <revisionDesc/> | revision Description: Angaben zu Überarbeitung des vorliegenden Datensatzes (Dokumentation der Änderungen) |

Abb. 5: Der <teiHeader/> (Editionsrichtlinien zur Ausgabe der Briefe, Tagebücher und Dokumente Webers)

durchweg ein Problem darstellen. Einige Angaben im Header der Briefausgabe dürften also auf die Briefe – automatisch oder sinngemäß – übertragbar sein.

Die Angaben des `<fileDesc/>` bleiben ohnehin unvollständig, da sie bereits in Barths Ausgabe fehlen. Quellennachweise und weitere Hinweise sind dort nicht vorhanden. Der `<encodingDesc/>` ist verhältnismäßig überschaubar und wird einmalig für die gesamte Briefausgabe verfasst. Schwieriger gestaltet es sich mit dem `<profileDesc/>`, welcher Daten zu den jeweiligen Briefen enthält. Viele Elemente wie beispielsweise der `<corresp desc/>`, welcher die Sender-Empfänger-Daten der Briefe beinhaltet, können aber einfach nachgereicht werden, da sich die beteiligten Schreiber im Briefwechsel mit wenigen Ausnahmen auf Grimm und Brahms beschränken. Diese Informationen können größtenteils auch direkt in den Textkorpus ausgelagert werden – in diesem Fall stellvertretend durch das `<byline/>` Element.[4] Der `<revisionDesc/>` ist wiederum unproblematisch, da er nur einmalig im Header der Briefausgabe erscheinen braucht.

An den Textkorpus werden also zwei Voraussetzungen gestellt: erstens muss er zusätzliche Informationen des Metadaten-Headers enthalten, die wahlweise für einen neuen Header extrahierbar sein müssen; zweitens sollten die einzelnen Briefe ohne Komplikationen aus dem Briefwechsel-Dokument genommen und in eine eigene TEI-Datei mit eigenem `<teiHeader/>` überführt werden können. Normalerweise sieht ein TEI-XML Dokument nur ein einziges `<text/>` Element vor, in welchem nur ein einziger Brief kodiert ist. Die TEI Guidelines erlauben aber die Schachtelung von `<text/>` Elementen innerhalb des `<group/>` Elements, in welchem die Briefe einzeln abgelegt werden können. Auf diesem Wege können mehrere Briefe in einer XML-Datei ausgezeichnet werden. Das formale Muster des Briefwechsels kann demnach wie in Abb. 6 dargestellt verwendet werden (Titelblatt und Einleitung sind im `<front/>` Element enthalten):

[4] Das `<byline/>` Element ist eigentlich nur für die Angabe des Verfassers/Erstellers eines Dokuments vorgesehen, kann jedoch für diesen Fall als Alternative verwendet werden.

```
<text>
    <front>
        <titlePage/>
        <div type="preface"/>
    </front>
    <group>
        <text type="letter" n="1">
            <body>
                <div/>
            </body>
        </text>
        <text type="letter" n="2">
            <body>
                <div/>
            </body>
        </text>
        [...]
    </group>
</text>
```

### 2.3.1 Anwendung des Brief-Templates

Der transkribierte Fließtext wird im Folgenden in die jeweiligen Briefe aufgeteilt und anhand eines Brief-Templates kodiert (siehe *bgbw_templates.txt*).[5] Das Muster fungiert als Datenschema und garantiert, dass sämtlicher Text korrekt „ausgezeichnet“ ist – beispielsweise das Absendedatum als solches per `<dateline/>` Element nachvollziehbar wird. Aufgrund des fehlenden Headers muss das Template einige zusätzliche Informationen unterbringen. Das `<byline>` Element trägt so z.B. Informationen zum Sender und Empfänger des Briefes (`<byline>` *Sender* + an + *Empfänger* `</byline>`). Die Informationen können später per XSLT (oder vergleichweisen Skripten) extrahiert und in das eigentlich dafür vorgesehenen `<correspAction>` Element im Header übertragen werden. Ähnlich wird mit dem Absendeort und dem Datum verfahren, welche als leere Elemente mit im `<opener>` stehen. Die nötigen Informationen sind als Attribute kodiert (`when=`*”Datum XY”* und `key=`*”Stadt XY”*), sodass sie nicht den eigentlichen Text der `<dateline/>` interferieren. Das Brief-Template sieht wie folgt aus:

```
<text type="letter">
    <body>
        <div>
            <opener>
                <byline/>
                <date/>
                >settlement/>
                <dateline/>
                <salute/>
            </opener>
            <p/>
            <closer>
                <signed/>
            </closer>
            <postscript>
                <p/>
            </postscript>
        </div>
    </body>
</text>
```

[5] Für weitere Informationen siehe Weber Brief-Editionsrichtlinien, 3.6 Strukturelle Merkmale der Texte.

Im Zuge der digitalen Aufbereitung wechselt Barth in seiner Funktion vom Editor bzw. Herausgeber zum Author. Seine Fußnoten werden dementsprechend im `<note>` Element mit dem `type=`*"footnote"* Attribut kodiert. Das für die Editoren vorgesehene `type=`*"commentary"* Attribut ist derweil jenen Anmerkungen vorbehalten, die im Laufe der digitalen Aufbereitung aufkommen.[6] Die vollständige Kodierung des ersten Briefes ist im Folgenden abgebildet (siehe Abb. 6); sie steht exemplarisch für das weitere Vorgehen. Das Zwischenergebnis mit der Einleitung und den aufbereiteten Briefen wird in einer Entwurfsdatei gespeichert (siehe *bgbw_draft.xml*).[7]

```
<text type="letter" n="1">
    <body>
        <div type="writingSession" n="1">
            <pb facs="#facs_14" n="14"/>
            <opener>
                <byline>J. O. Grimm an Johannes Brahms</byline>
                <date when="1853-12-21"/>
                <settlement key="Hannover"/>
                <dateline> Hannover, den 21. Dezember 53. </dateline>
                <salute>Mein lieber Johannes Kreisler junior!<note type="footnote"> So
                        und auch Johannes Kreisler II hatte sich Brahms nach E. T. A.
                        Hoffmanns verrücktem Kapellmeister Johannes Kreisler
                        (Phantasiestücke in Callots Manier) selbst oft genannt.</note>
                </salute>
            </opener>
            <p> Leider ist Joachim nicht hier und kommt erst Freitag abend, — ich kann
                also so lange nicht warten und muß bis- auf die Rückreise meine
                Sehnsucht nach ihm unterdrücken. — Erst heute morgen war ich so
                glücklich, den Herrn Staatsminister von Schleinitz<note type="footnote">
                    In Göttingen wurde die Stelle des Universitäts-Musikdirektors frei,
                    um die sich Grimm bewarb. Da es nun aber nach Angabe des Herrn Dr.
                    Georg Fischer zu jener Zeit keinen Staatsminister von Schleinitz in
                    Hannover gab, so mag dieser Name Grimm wohl in der Zerstreutheit und
                    in Gedanken an den Leipziger Konservatoriums=Direktor, der ihn gewiß
                    mit guten Empfehlungen ausgestattet hatte, aus der Feder geflossen
                    sein. — Was es mit dem Referat aus Gandersheim für eine Bewandtnis
                    hatte, ließ sich nicht mehr ermitteln.</note> zu sprechen. Der Brief
                von Moscheles hat famose Wirkung getan, denn ich wurde vom Schleinitz
                wie ein Mondkalb empfangen und habe Hoffnung; — es wird nur noch ein
                Referat aus Gandersheint erwartet, und wenn darin nicht zu blödsinnige
                Gründe gegen mich aufgeführt sind, so kann sich die Sache noch machen,
                und ich brauche nicht nach England zu gehen. — Wir sehen uns also erst
                im Mai wieder, wenn die Sachen gut gehen. — Einstweilen leb wohl, mein
                süßer Junge. Grüße Deine Eltern und Geschwister unbekannterweise von mir
                und verbringe ein schönes Fest. Beiläufig könntest Du mir einmal
                schreiben, wie Du die Deinigen gefunden hast, Du weißt, alles, was Du
                mir schreibst, ist mir wie mein Eigen — </p>
            <closer> Dein <signed>J. Grimm.</signed></closer>
        </div>
    </body>
</text>
```

Abb. 6: Der erste Brief digital aufbereitet

[6] Gleichwohl die Studie keine Edition im eigentlichen Sinne vornimmt, überträgt sie doch einen bestehenden Text ins Digitale und wird deshalb im Header als `<editor/>` geführt.

[7] Zur schnellen Orientierung: Header (Z. 6–64); Faksimiles (Z. 65–1261); Textkorpus (Z. 1258–7192).

## 2.4 Abschließende Schritte

Es folgt zum Abschluss eine Nachbereitung, die dem Ziel dient, kleinere Redundanzen (z.B. unnötige Leerzeichen) zu entfernen und zusätzliche Angaben hinzuzufügen (z.B. die entsprechenden Fußnoten an Barth zu attribuieren). Dafür wird erneut ein XSLT-Dokument verwendet (siehe *bgbw_completion.xsl*). Die Ausgangsdatei steht somit für eine weitere Verarbeitung oder für die Nutzung mittels eines TEI-Readers bereit (siehe *bgbw_complete.xml*).

Die einfachste Form der Veröffentlichung stellt nach wie vor das PDF-Format dar. Auf diesem Wege können die Briefe gelesen und mit einfachen Suchfunktionen inspiziert werden. Auch lässt sich eine PDF direkt über den Oxygen-Editor exportieren; ein externer TEI-Reader ist dafür nicht vonnöten (siehe Anmerkungen zu den Webapplikationen in Kapitel 3.2) Da zunächst keine synoptische Darstellung der analogen und digitalen Version intendiert ist, sind die Facsimile-Verknüpfungen und Seitenumbrüche der analogen Vorlage überflüssig. Die `<facsimile>` und `<pb>` Elemente werden demnach per Regex (reguläre Ausdrücke) aus dem Dokument entfernt (siehe *bgbw_regex.txt*). Das Resultat ist eine für den PDF-Export bereitstehende XML-Datei (siehe *bgbw_final.xml*).

Der Oxygen-Editor bietet bereits ein Transformations-Szenario („TEI P5 PDF“), das für den vorliegenden Fall noch modifiziert wird.[8] Sämtliche vorgenommenen Änderungen beziehen sich nur auf die zentrale XSLT-Datei (*fo*), welche zudem umbenannt wird (siehe *bgbw_pdf.xsl*). Die Modifikationen betreffen vorrangig die Formatierung der Briefe – vor allem die Darstellung der Fußnoten – sowie die Entfernung von überflüssigen Leerseiten. Selbstverständlich erlaubt die XSLT auch weitere, spezifische Anpassungen, je nach Anwendungszweck. Für die vorliegende Studie steht aber zunächst nur die reine Lesbarkeit und die einfache Durchsuchbarkeit im Vorderung, die durch die Ausgabedatei gewährleistet wird (siehe *bgbw_final.pdf*)..

[8] Die entsprechenden Dateien sind in den Files des Oxygen Editors zu finden unter: *frameworks/tei/xml/tei/stylesheet*. Für das Transformations-Szenario werden die Dateien aus den Ordnern *common* und *fo* sowie die XML-Datei *i18n* benötigt.

# 3 Schlussbemerkungen

Fest steht, dass die digitale Aufbereitung von vornherein einen Kompromiss darstellt. Sie bietet eine Durchsuchbarkeit des Textes nach modernen Standards und ist interoperabel, das heißt sie ermöglicht eine digitale Weiterverarbeitung. Im Gegenzug muss sie jedoch einige Abstriche bei der Kodierung machen und ist inhaltlich immer noch auf dem Stand der historischen Ausgabe von 1912. Die Briefe sind möglicherweise unnvollständig, aktuelle Forschungserkenntnisse fehlen und die alte Ausgabe entspricht nicht den Standards der modernen (Brief-)Editionspraxis. Die Erfolge der digitalen Aufbereitung sind daher stets cum salo granis zu bewerten. Dennoch können solche Dokumente die Forschungspraxis bereichern, indem sie online verfügbar und maschinenlesbar sind und schnelle sowie komplexe Suchanfragen verarbeiten können.

## 3.1 Limitationen

Im Verlauf der Studie stieß der Versuch auf mehrere Probleme, die hier noch einmal gesammelt besprochen werden. Zuallererst ist ein hochauflösender Scan eines Buches keine Selbstverständlichkeit, es braucht dafür ein professionelles Aufnahmegerät und die entsprechende Einrichtung. Glücklicherweise sind moderne Handykameras zunehmend in der Lage, vergleichsweise hochauflösende Aufnahme zu machen. Mit neuen Vorrichtungen wie dem ScanTent wird das Scannen zukünftig flexibler und mobil möglich sein.[1]

Die Auflösung des hier verarbeiteten Briefwechsel-Scans war nicht optimal, aber ausreichend. Dies trifft auch auf das Transkriptionsmodell *Danish Fraktur SB 19th century PyLaia* von Transkribus zu. Ein größerer Trainings-Datensatz als 390226 Wörtern wäre zu wünschen, da doch merkbar viele Fehler im transkribierten Text gefunden wurden. Besonders häufig kamen inkorrekte Kommata und Striche an Zeilenenden sowie großgeschriebene Umlaute auf. Ferner ergab die Layout-Analyse nicht das gewünschte Ergebnis. Die erkannten Textregionen entsprachen nicht den Textabsätzen und mussten allesamt manuell korrigiert werden. Diesbezüglich ist jedoch eine zukünftige Besserung bei ausreichendem Training der Modelle vorstellbar.

Der größte Zeitaufwand entsteht zweifellos beim Einordnen der Briefe in die Templates. Die Textinhalte müssen manuell aus dem bereinigten Fließtext genommen und in die jeweiligen Elemente eingefügt (kodiert) werden. Je nach Länge des Textes dauert die vollständige Kodierung und Korrektur eines Briefes mehrere Minuten. Hier müssten weitere Verfahren der automatisierten Texterfassung und -verarbeitung erörtert werden, um den Zeitaufwand zu reduzieren. Inwiefern XSLTs oder LLMs (Large Language Modells) diese Aufgabe übernehmen können ist noch zu diskutieren. Inhalte, die tendenziell

[1] Siehe https://readcoop.eu/de/scantent/.

in jedem Brief vorhanden sind und sich strukturell wenig ändern – wie beispielsweise die `<dateline/>` – könnten so automatisiert erkannt und eingeordnet werden.

Zuletzt ist zum Abrufen der Briefe in ihrer digital aufbereiteten Form stets ein XML-Reader oder ein vergleichweises Programm vonnöten, welches derzeit allerdings noch nicht jedem zur Verfügung steht. Während die meisten Forschenden zwar Zugriff auf einen XML-Editor haben, können interessierte Laien im Zweifelsfall nur auf die PDF-Versionen zurückgreifen, die wiederum weniger Nutzungsmöglichkeiten bieten. Optimal wären dahingehend offizielle E-Reader bzw. TEI-Applikationen im Internet. Die digital aufbereiteten Briefwechsel könnten dann in einer für sie zugeschnittenen Applikation hochgeladen und von jedem genutzt werden.

## 3.2 Ausblick

Vorrangiges Ziel der vorliegenden Studie ist die Erprobung von Methoden zur digitalen Aufbereitung einer analogen Briefwechselausgabe. Die Veröffentlichung, Verbreitung und Verstetigung der aufbereiteten Dokumente ist dagegen separat zu diskutieren. Als geeignete Webapplikationen kämen durch Programme in Frage, die aktuell von der Forschung genutzt werden, wie etwa der TEI-Publisher oder Ediarum. Sie beide bieten umfangreiche Möglichkeiten, um digitale Texteditionen zu erstellen und zu publizieren. Besonders das Management der TEI-Schemata sowie die weitere Annotation des Editionstextes wird durch solche Programme stark vereinfacht.

Der TEI-Publisher wirbt mit einer benutzerfreundlichen Oberfläche und einer einfachen Implementierung. Das auf eXist-db basierende OpenSource-Programm ermöglicht eine einfache Visualisierung von XML-Dokumenten und deren Ausgabe in verschiedenen Formaten. Es eignet sich demnach als optimale Grundlage für die Veröffentlichung von Texteditionen. Ein zentrales Element des TEI-Publishers ist die einfache ODD-Customization. Die Möglichkeit, die TEI-Module in einer grafischen Oberfläche zu verwalten, erweist sich als besonders vorteilhaft. Die Module können so schnell und übersichtlich eingebunden werden und das Ergebnis wird direkt nachvollziehbar.

Ediarum hingegen birgt den Vorteil, dass hier die einzelnen Software-Komponenten direkt im Oxygen Editor eingebunden werden. Auch diese digitale Arbeitsumgebung basiert auf eXist-db und ermöglicht ein kollaboratives Arbeiten. Die Edition wird aus dem XML-Editor heraus erstellt und kann dann über weitere Software-Komponenten im Druck und im Web veröffentlicht werden. Leider befindet sich das Modul „ediarum.PDF“ noch in der Entwicklung, weshalb für die vorliegende Studie mit einem eigenen XSLT gearbeitet wurde.

Die im GitHub verlinkten Dokumente sollen zusammen mit dem vorliegenden Forschungsbericht zunächst einen Anhaltspunkt für weitere Schritte bieten, ohne dass sich auf die Nutzung einer bestimmten Applikation festgelegt wird.[2]

- Maximilian Greshake (22. Februar 2024)

[2] Eine Nutzung und Weiterverarbeitung ist unter der MIT-Lizenz gestattet.