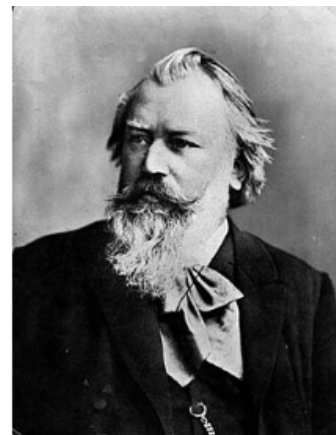


Zertifikat „Digital Humanities“

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



Vorgelegt am 25. März 2024 von:

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

Inhaltsverzeichnis

1	Einführung	3
1.1	Versuchsaufbau	5
2	Digitale Aufbereitung des Brahms-Grimm-Briefwechsels	6
2.1	Texterfassung per OCR	6
2.2	Datenbereinigung	7
2.3	Aufbereitung der TEI-Kodierung	9
2.3.1	Anwendung des Brief-Templates	11
2.4	Abschließende Schritte	13
3	Schlussbemerkungen	15
3.1	Limitationen	15
3.2	Ausblick	16
	Quellen- und URL-Verzeichnis	17

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 Werkgesamtausgaben¹, oder im Kontext der Erforschung von Komponistenbiografien.

In den letzten Jahrzehnten war das Editionswesen grundlegenden Veränderungen unterworfen. Durch die Digitalisierung 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 TEI² als Standard für digitale Briefeditionen. Für Musikalische Notentexte wurde derweil das XML-Format MEI³ 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 Drittbrieftexte und einen lexikalischen Apparat zur Auszeichnung von Personen, Werken, Orten und weiteren Aspekten. Dadurch findet eine weitreichende Verlinkung zwischen den Dokumenten statt, die so gleich ein vernetztes Erforschen der Dokumente begünstigt.

¹ Siehe beispielsweise die Haydn Briefausgabe, <https://www.haydn-institut.de/index.php/gesamtausgabe/briefausgabe>, letzter Aufruf am 25. März 2024.

² Siehe <https://tei-c.org/>.

³ Siehe <https://music-encoding.org/>.

⁴ Siehe <https://weber-gesamtausgabe.de/de/A002068.html#correspondence>.

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 CorrespSearch⁵. 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 vierten Bandes des Johannes Brahms Briefwechsels herausgegeben wurde.⁶ Da einige Ori-

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

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

ginale verschollen sind, gilt die Briefausgabe mitunter als letzter Textzeuge. Die digitale Aufbereitung dient demzufolge also auch der Sicherung der Briefinhalte.

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 erwartende 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. Der übertragene Text kann schließlich im TEI-Format ausgegeben werden. Im zweiten Schritt wird die TEI-Datei im Oxygen XML-Editor geöffnet und bereinigt - für komplexere Eingriffe wird ein XSLT-Dokument (XSL-Transformation-Sheet) angelegt. 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, welche bereits einfache Such-Funktionen unterstützt.

⁷ Siehe Editionsrichtlinien zur Ausgabe der Briefe, Tagebücher und Dokumente Webers, https://weber-gesamtausgabe.de/de/Projekt/Editionsrichtlinien_Text.html, letzter Aufruf am 25. März 2024.

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

2 Digitale Aufbereitung des Brahms-Grimm-Briefwechsels

Richard Barth war zu Beginn seiner Musikerlaufbahn in Münster als Konzertmeister tätig, wo er eine geradezu 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 und demgemäß eine wichtige Forschungsquelle. Barths Stand als Zeitgenosse der beiden Schreiber kommt durchaus in seinen Anmerkungen innerhalb der Briefausgabe zum Ausdruck. Neben einem ausführlichen Vorwort zu beiden Schreibern verfasste er 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 ohnehin bereits mit der einfachen String-Suche oder per Pfadsprachen durchsucht werden kann. 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.⁹ Transkribus verwendet verschiedene Modelle künstlicher

⁹ Siehe <https://readcoop.eu/de/transkribus/>.

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.

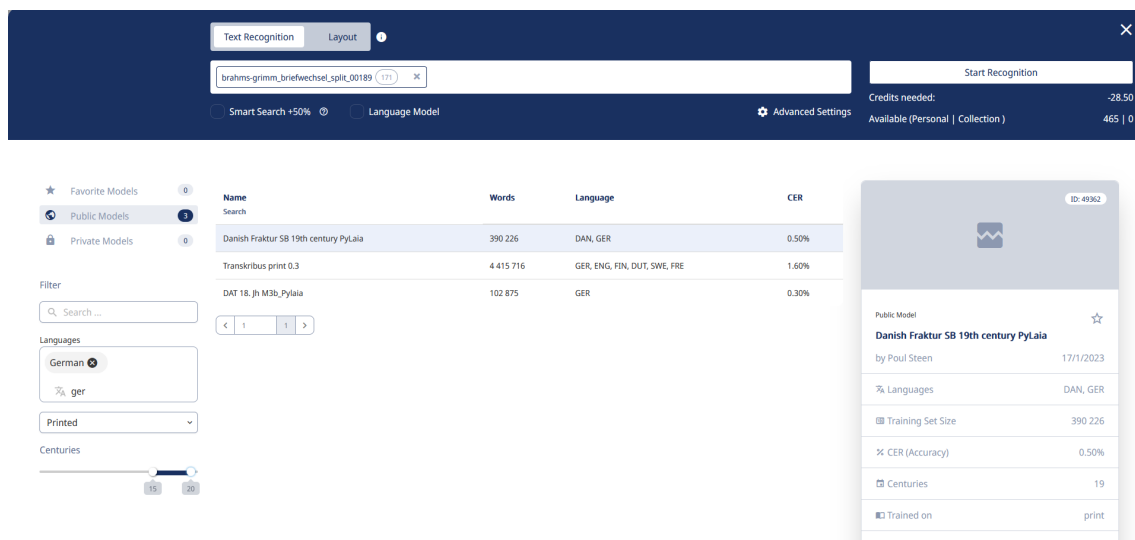


Abb. 1: Transkribus, Anwendung eines Modells

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 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).

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

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

Gleichmaßen 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 her- vorzubringen. Da die Output-Datei aus fast 15000 Codezeilen besteht, ist eine auto- matisierte 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.¹⁰ 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).¹¹ 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.“¹² Unter dieser Prämisse würde das Fehlen des Headers für die hier aufbereiteten Briefe

¹⁰ Weber Brief-Editionsrichtlinien, 3.1 Technische Vorbemerkung.

¹¹ Vgl. ebd., 4. Die Metadaten der Texte (Apparatteil).

¹² Ebd., 4.1 Die Grobgliederung des Metadatenteils.

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

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

nicht 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.¹³ 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 einzigen 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):

¹³Das `<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 des unten stehenden Brief-Templates kodiert.¹⁴ Das Muster fungiert als Datenschema und garantiert, dass sämtlicher Text korrekt „ausgezeichnet“ ist – beispielsweise das Absendedatum als solches per `<date/>` Element nachvollziehbar wird. Da ein eigener TEI-Header fehlt, 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>
```

¹⁴Für 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.¹⁵ Ähnliches gilt auch für Ergänzungen in Form von `<supplied>`-Elementen. Hier liegt einer der wesentlichen Vorteile der XML-Kodierung, da der Text jederzeit ajourniert bzw. ergänzt werden kann.

Anhand des zehnten Briefes der Ausgabe lässt sich dies exemplarisch veranschaulichen.¹⁶ Ein Teil des Dokuments scheint Barth nicht zugänglich gewesen zu sein, weshalb er in einer Fußnote „*Die Fortsetzung fehlt leider.*“ kommentiert. Das fehlende Ende findet sich allerdings in der Briefwechsel-Ausgabe zwischen Brahms und dem Geiger Joseph Joachim wieder, die 1812 von Andreas Moser herausgegeben wurde. Dort beginnt der Ausschnitt mit der Bemerkung „*Anfang fehlt*“, er wird jedoch nicht weiter besprochen. Mittels `<supplied>`-Element kann der Textteil mitsamt Quellenangabe hinzugefügt werden (siehe Abb. 6). Richard Barths Anmerkung verbleibt in auskommentierter Form. Auf diese Weise kann die Supplementierung nachvollzogen und jederzeit korrigiert werden.

```

985      gemacht, es war hübsch, daß Ihr an meine Kollektion dachtet, — ich werde ihr
986      selbst danken — grüße sie mit meiner vollen Ehrerbietung, wenn (Du) ihr
987      schreibst. Apropos — die Hauptsache: Härtels haben mir
988      die<!-- <note >Die Fortsetzung fehlt leider.</note> --><supplied
989      resp="Maximilian Greshake" reason="text missing"
990      source="Johannes Brahms im Briefwechsel mit Joseph Joachim, 1912, S. 86"
991      >Scherzi endlich geschickt, und ich lege sie bei; je ein Exemplar an Sie und
992      je eins an Dich. Willst du So gut sein, es nicht zu vergessen, sie Ihr zu
993      geben, wenn Sie zurückkommt. — Du hast wohl kein Exemplar Deines letzten
994      Liederheftes mehr? Wenn ja, und Du willst und kannst, so würdest Du mich
995      sehr erfreuen. — Lebe aber jetzt wohl. Joachim will Dich grüßen, also wende
996      dies Blatt um. </supplied>
997      </p>

```

Abb. 6: Editorische Ergänzung von Inhalten aus externen Quellen

Die Kodierung des ersten Briefes ist im Folgenden abgebildet (siehe Abb. 7); sie steht exemplarisch für das Vorgehen der Studie. Das Zwischenergebnis mit der Einleitung und den aufbereiteten Briefen wird in einer Entwurfsdatei gespeichert (siehe *bgbw_draft.xml*).¹⁷

¹⁵ 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.

¹⁶ Hierüber hinaus werden keine weiteren Ergänzungen vorgenommen; die hier besprochenen Vorgänge dienen vorrangig Veranschaulichungszwecken.

¹⁷ Zur schnellen Orientierung: Header (Z. 6–64); Faksimiles (Z. 65–1261); Textkorpus (Z. 1258–7192).

```

1629 <text type="letter" n="1">
1630   <body>
1631     <div type="writingSession" n="1">
1632       <pb facs="#facs_l4" n="14"/>
1633       <opener>
1634         <byline>J. O. Grimm an Johannes Brahms</byline>
1635         <date when="1853-12-21"/>
1636         <settlement key="Hannover"/>
1637         <dateline>Hannover, den 21. Dezember 53. </dateline>
1638         <salute>Mein lieber Johannes Kreisler junior!<note type="footnote"> So
1639           und auch Johannes Kreisler II hatte sich Brahms nach E. T. A.
1640           Hoffmanns verrücktem Kapellmeister Johannes Kreisler
1641           (Phantasiestücke in Callots Manier) selbst oft genannt.</note>
1642         </salute>
1643       </opener>
1644       <p> Leider ist Joachim nicht hier und kommt erst Freitag abend, — ich kann
1645         also so lange nicht warten und muß bis- auf die Rückreise meine
1646         Sehnsucht nach ihm unterdrücken. — Erst heute morgen war ich so
1647         glücklich, den Herrn Staatsminister von Schleinitz<note type="footnote">
1648           In Göttingen wurde die Stelle des Universitäts-Musikdirektors frei,
1649           um die sich Grimm bewarb. Da es nun aber nach Angabe des Herrn Dr.
1650           Georg Fischer zu jener Zeit keinen Staatsminister von Schleinitz in
1651           Hannover gab, so mag dieser Name Grimm wohl in der Zerstretheit und
1652           in Gedanken an den Leipziger Konservatoriums-Direktor, der ihn gewiß
1653           mit guten Empfehlungen ausgestattet hatte, aus der Feder geflossen
1654           sein. — Was es mit dem Referat aus Gandersheim für eine Bewandtnis
1655           hatte, ließ sich nicht mehr ermitteln.</note> zu sprechen. Der Brief
1656           von Moscheles hat famose Wirkung getan, denn ich wurde vom Schleinitz
1657           wie ein Mondkalb empfangen und habe Hoffnung; — es wird nur noch ein
1658           Referat aus Gandersheim erwartet, und wenn darin nicht zu blödsinnige
1659           Gründe gegen mich aufgeführt sind, so kann sich die Sache noch machen,
1660           und ich brauche nicht nach England zu gehen. — Wir sehen uns also erst
1661           im Mai wieder, wenn die Sachen gut gehen. — Einstweilen leb wohl, mein
1662           süßer Junge. Grüße Deine Eltern und Geschwister unbekannterweise von mir
1663           und verbringe ein schönes Fest. Beiläufig könntest Du mir einmal
1664           schreiben, wie Du die Deinigen gefunden hast, Du weißt, alles, was Du
1665           mir schreibst, ist mir wie mein Eigen — </p>
1666       <closer> Dein <signed>J. Grimm.</signed></closer>
1667     </div>
1668   </body>
1669 </text>

```

Abb. 7: Der erste Brief digital aufbereitet

2.4 Abschließende Schritte

In der Nachbereitung werden noch kleinere Redundanzen (z.B. unnötige Leerzeichen) ausgebessert und zusätzliche Angaben hinzugefügt – beispielsweise werden die entsprechenden Fußnoten an Barth attribuiert. Dafür wird erneut ein XSLT-Dokument verwendet (siehe *bgbw_completion.xsl*). Die Ausgangsdatei steht somit für eine Weiterverarbeitung bereit (siehe *bgbw_complete.xml*). Nun können weitere editorische Schritte vorgenommen werden, darunter das Tagging und Annotieren oder die Korrektur weiterer Fehlerstellen. Ebenso kann die Datei für die Nutzung exportiert werden.

Für die meisten Forschenden stellt nach wie vor das PDF-Format die zugänglichste Form eines Forschungstextes dar. Auf diesem Wege können die Briefe einfach geöffnet, gelesen und mit simplen Suchfunktionen inspiziert werden. Vorteilhafterweise 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). Der Oxygen-Editor bietet bereits zwei Transformations-Szenarien, um XML-Dateien in PDFs

umzuwandeln: zum einen per CSS, welches bereits für die eingebaute Autor-Funktion genutzt wird, und zum anderen per XSL-FO („TEI P5 PDF“). Das Verfahren per XSL-FO ist zwar komplexer, bietet im Gegenzug aber auch mehr Anpassungs-Möglichkeiten, weshalb es für den vorliegenden Fall verwendet wird. Eine vollwertige Publikation würde noch tiefgreifende Modifikationen der Transformations-Skripte erfordern. Da für die vorliegende Studie aber zunächst nur die reine Lesbarkeit und die einfache Durchsuchbarkeit im Vordergrund steht, reicht es aus, sie mit Ausnahme weniger Codezeilen in ihrer Default-Version zu belassen. Die wenigen – hier nicht näher besprochenen – Änderungen betreffen die Formatierung der Zeilenabstände und Textdivisionen sowie die Seitennummerierung und -umbrüche. Zu finden sind die entsprechenden XSLT-Dateien im Installationsordner des Oxygen Editors.¹⁸ Für den provisorischen PDF-Export wird die XML-Datei stellenweise angepasst. Das finale Dokument bedarf durchaus noch einiger Optimierungen; es soll aber auch nur einen Zwischenstand markieren (siehe *bgbw_final.pdf*). Für interne Forschungszwecke ist die PDF daher ausreichend geeignet.

¹⁸ *frameworks/tei/xml/tei/stylesheet*. Es werden die Ordner *fo* und *common* sowie die Datei *180n.xml* benötigt.

3 Schlussbemerkungen

Fest steht, dass die digitale Aufbereitung von vornherein einen Kompromiss darstellt. Das TEI/XML-Dokument bietet eine Durchsuchbarkeit des Textes nach modernen Standards und ist interoperabel, das heißt sie ermöglicht eine digitale Weiterverarbeitung. Ein daraus resultierender Vorteil ist bereits anhand der Vervollständigung des zehnten Briefes deutlich gemacht worden (siehe Kapitel 2.3). Im Gegenzug muss die Briefwechselaufbereitung jedoch einige Abstriche bei der Kodierung machen und ist inhaltlich immer noch auf dem Stand der historischen Ausgabe von 1912. Die Briefe sind möglicherweise unvollständig, 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. Jedoch sind auch 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.¹⁹

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 Models) diese Aufgabe übernehmen können ist noch zu diskutieren. Inhalte, die tendenziell

¹⁹ 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 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**²⁰ 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 für Forschende, die noch nicht so eng mit Auszeichnungssprachen vertraut sind. Die Module können so schnell und übersichtlich eingebunden werden und das Ergebnis wird direkt nachvollziehbar.

Ediarum²¹ 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 Default-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.²²

²⁰ Siehe <https://teipublisher.com/index.html>.

²¹ Siehe <https://www.ediarum.org/>.

²² Eine Nutzung und Weiterverarbeitung ist unter der MIT-Lizenz geregelt.

Quellenverzeichnis

Brahms, Johannes und Grimm, Julius Otto: Johannes Brahms im Briefwechsel mit J. O. Grimm (= Brahms-Briefwechsel IV, hrsg. von Richard Barth), Berlin, 2. Aufl. 1912.

Brahms, Johannes und Joachim, Joseph: Johannes Brahms im Briefwechsel mit Joseph Joachim (= Brahms-Briefwechsel V, hrsg. von Andreas Moser, Bd. 1), Berlin, 2. Aufl. 1912.

URL-Verzeichnis

GitHub – https://github.com/maxgreshake/bgbw_digital.

Carl Maria von Weber Gesamtausgabe – <https://weber-gesamtausgabe.de/>.

Transkribus – <https://readcoop.eu/de/transkribus/>.

TEI-Guidelines – <https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html>.

TEI-Publisher – <https://teipublisher.com/index.html>.

Ediarum – <https://www.ediarum.org/>.

Plagiatserklärung des Studierenden

Hiermit versichere ich, dass die vorliegende Arbeit „*Eine explorative Studie zur digitalen Aufbereitung analoger Briefeditionen am Beispiel des Brahms-Grimm-Briefwechsels*“ selbstständig verfasst worden ist, dass keine anderen Quellen und Hilfsmittel als die angegebenen benutzt worden sind und dass die Stellen der Arbeit, die anderen Werken – auch elektronischen Medien – dem Wortlaut oder Sinn nach entnommenen wurden, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht worden sind.

Maximilian Greshake, 25. März 2024

(Datum, Unterschrift)

Ich erkläre mich mit einem Abgleich der Arbeit mit anderen Texten zwecks Auffindung von Übereinstimmungen sowie mit einer zu diesem Zweck vorzunehmenden Speicherung der Arbeit in eine Datenbank einverstanden.

Maximilian Greshake, 25. März 2024

(Datum, Unterschrift)