

## Zeitschrift für digitale Geisteswissenschaften

---

Beitrag aus:

Fabrikation von Erkenntnis – Experimente in den Digital Humanities. Hg. von Manuel Burghardt, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels Walkowski, Joëlle Weis, Ulrike Wuttke. Wolfenbüttel 2021. (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 5) text/html Format. Teilband 2 / Sonderband 5 der ZfdG: DOI: [10.17175/sb005](https://doi.org/10.17175/sb005)

Titel:

Bomber's Baedeker – vom Text zum Bild zur Datenquelle

---

Autor\*in:

Felix Bach

Kontakt: [fbach9310@gmail.com](mailto:fbach9310@gmail.com)

Institution: Leibniz-Institut für Europäische Geschichte (IEG) | Hochschule Darmstadt

GND: [124168099X](https://nbn-resolving.org/urn:nbn:de:hbz:5:1-63868-p0011-9) ORCID: [0000-0001-9517-270X](https://orcid.org/0000-0001-9517-270X)

---

Autor\*in:

Stefan Schmunk

Kontakt: [stefan.schmunk@h-da.de](mailto:stefan.schmunk@h-da.de)

Institution: Hochschule Darmstadt

GND: [1028340028](https://nbn-resolving.org/urn:nbn:de:hbz:5:1-63868-p0011-9) ORCID: [0000-0001-9706-9757](https://orcid.org/0000-0001-9706-9757)

---

Autor\*in:

Cristian Secco

Kontakt: [stcrsecc@stud.h-da.de](mailto:stcrsecc@stud.h-da.de)

Institution: Leibniz-Institut für Europäische Geschichte (IEG) | Hochschule Darmstadt

GND: [1241301123](https://nbn-resolving.org/urn:nbn:de:hbz:5:1-63868-p0011-9) ORCID: [0000-0002-5023-015X](https://orcid.org/0000-0002-5023-015X)

---

Autor\*in:

Thorsten Wübbena

Kontakt: [wuebbena@ieg-mainz.de](mailto:wuebbena@ieg-mainz.de)

Institution: Leibniz-Institut für Europäische Geschichte (IEG)

GND: [123312396](https://nbn-resolving.org/urn:nbn:de:hbz:5:1-63868-p0011-9) ORCID: [0000-0001-8172-6097](https://orcid.org/0000-0001-8172-6097)

---

DOI des Artikels:

[10.17175/sb005\\_004](https://doi.org/10.17175/sb005_004)

Nachweis im OPAC der Herzog August Bibliothek:

[1770855890](https://nbn-resolving.org/urn:nbn:de:hbz:5:1-63868-p0011-9)

Erstveröffentlichung:

22.09.2021

Lizenz:

Sofern nicht anders angegeben 

Medienlizenzen:

Medienrechte liegen bei den Autor\*innen

Letzte Überprüfung aller Verweise: 16.09.2021

GND-Verschlagwortung:

[Bomber's Baedeker](#) | [Datentransformation](#) [«OCR»](#) [«XML»](#) [«Python»](#) | [Digitalisierung](#) | [Informatik](#) | [Optische Zeichenerkennung](#) |

Zitierweise:

Felix Bach, Stefan Schmunk, Cristian Secco, Thorsten Wübbena: Bomber's Baedeker – vom Text zum Bild zur Datenquelle. In: Fabrikation von Erkenntnis – Experimente in den Digital Humanities. Hg. von Manuel Burghardt, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels Walkowski, Joëlle Weis, Ulrike Wuttke. Wolfenbüttel 2021. (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 5) text/html Format. DOI: 10.17175/sb005\_001 PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: [10.17175/sb005\\_004](https://doi.org/10.17175/sb005_004).

Felix Bach, Stefan Schmunk, Cristian Secco, Thorsten Wübbena

## Bomber's Baedeker – vom Text zum Bild zur Datenquelle

---

### Abstracts

Das zweibändige Druckwerk *The Bomber's Baedeker. A Guide to the Economic Importance of German Towns and Cities* wurde während des Zweiten Weltkrieges vom britischen Foreign Office und dem Ministry of Economic Warfare erstellt. Darin sind Städte des Deutschen Reichs mit mehr als tausend Einwohner\*innen und Informationen zu deren kriegswichtigen Infrastrukturen, Industrie- und Produktionsanlagen aufgeführt. Weltweit existieren nur noch vier nachgewiesene Exemplare und keines davon war bislang für die wissenschaftliche Nutzung digital zugänglich. Der *Bomber's Baedeker* wurde 2019 in der Bibliothek des Leibniz-Instituts für Europäische Geschichte (IEG) (wieder-)entdeckt, in Kooperation mit der Universitätsbibliothek Mainz digitalisiert und im Rahmen einer institutionsübergreifenden Kooperation vom Bereich Digitale historische Forschung | DH Lab des IEG und von der Hochschule Darmstadt, u. a. in Lehrveranstaltungen mit Studierenden, erschlossen und aufbereitet, so dass der *Bomber's Baedeker* nun als offene, maschinenlesbare Datenquelle unter Einhaltung der FAIR-Prinzipien genutzt, analysiert und weiterbearbeitet werden kann.

The two-volume printed work *The Bomber's Baedeker. A Guide to the Economic Importance of German Towns and Cities* was produced by the British Foreign Office and the Ministry of Economic Warfare during the Second World War. It lists towns and cities of the German Reich with more than a thousand inhabitants and information on their war-related infrastructure, industrial and production facilities. Only four verified copies still exist worldwide and none of them has been digitally accessible for scholarly use until now. In 2019, *The Bomber's Baedeker* was (re-)discovered in the library of the Leibniz Institute of European History (IEG), digitised in cooperation with the University Library of Mainz and made accessible and processed in a cross-institutional cooperation between the Digital Historical Research Unit | DH Lab of the IEG and the Darmstadt University of Applied Sciences, including in courses with students, so that *The Bomber's Baedeker* can now be used, analysed and further processed as an open, machine-readable data source in compliance with the FAIR principles.

## 1. Einleitung

»Handbuch für den Feuersturm« war der Titel der Frankfurt Allgemeinen Zeitung im Juni 2019<sup>1</sup> zum zweibändigen Werk *The Bomber's Baedeker. A Guide to the Economic Importance of German Towns and Cities*.<sup>2</sup> Anlass für die Berichterstattung war die vollständige Digitalisierung der zweiten Auflage der seltenen britischen Quelle aus dem Jahr 1944 (809 Seiten, Quartformat, 6 ganzseitige Karten). Diese Ausgabe ist weltweit nur noch in vier Exemplaren nachweisbar, so dass es durchaus als Glückstreffer bezeichnet werden kann, dass dieses Exemplar in der Forschungsbibliothek des Leibniz-Instituts für Europäische Geschichte (IEG) entdeckt und in einem gemeinsamen Projekt für eine digitale Nutzung vorbereitet wurde. Der Titel *Bomber's Baedeker* verwundert zugleich ein wenig, denn inhaltlich werden keine touristischen und sehenswerten Reiseziele identifiziert und beschrieben, wie sie die namensgebenden roten Reiseführer Karl Baedekers seit dem 19. Jahrhundert<sup>3</sup> liefern, sondern es werden ausschließlich militärische, infrastrukturelle und ökonomisch-kriegswichtige Ziele der Royal Air Force für den Bombenkrieg gegen das Deutsche Reich

---

<sup>1</sup>Vgl. Burger 2019.

<sup>2</sup>The Bomber's Baedeker 1944.

<sup>3</sup>Zu einer medienkulturhistorischen Analyse von Reiseführern vgl. Müller 2012.

im Zweiten Weltkrieg identifiziert. Das zweibändige Druckwerk enthält Informationen zu Orten mit mehr als tausend Einwohner\*innen, sofern dort aus Sicht des britischen Foreign Office und des Ministry of Economic Warfare kriegswichtige Industrie- und Produktionsanlagen vorhanden waren. Die Daten zur geografischen Lage, Einwohnerzahl, Entfernung zu London sowie umfassende Beschreibungen von mehr als 500 Städten in Deutschland wurden im Bestreben zusammengestellt, eine möglichst effektive und effiziente Auswahl von potenziellen Zielen zu ermöglichen.<sup>4</sup>

Für das Lesen der Quelle ist die digitale Bereitstellung als gemeinfreies Digitalisat hinreichend, für eine Auswertung der enthaltenen Daten kann dies aus Sicht der Digitalen Geisteswissenschaften aber nur als erster Schritt betrachtet werden. Nach der Transformation zu einer digitalen Bilddatei ist die Weiterverarbeitung zu maschinenlesbaren Daten für zahlreiche Methoden der Digital Humanities ein wichtiger Schritt in der Aufbereitung. Im Fall des *Bomber's Baedeker* wurde dies durch eine Verbesserung der Zeichenerkennung und Überführung der Inhalte in ein standardisiertes Datenformat durchgeführt, so dass dadurch digitale Analysen ermöglicht werden. Seit 2019 findet daher im Rahmen einiger kleinerer Projekte eine intensive wissenschaftliche Auseinandersetzung und zugleich eine datenspezifische Aufbereitung statt, so dass die Qualität der maschinenlesbaren Daten stetig verbessert wurde.

## 2. Hintergrund

Woher genau die Bezeichnung *Bomber's Baedeker* stammt und wie deren etymologische Entwicklung zu deuten ist, lässt sich nicht mit hundertprozentiger Sicherheit belegen. Vermutlich ist diese Namensgebung eine Reaktion auf die im Frühjahr 1942 durch die Deutsche Luftwaffe ausgeführten Angriffe auf Exeter, Bath, Norwich, York und Canterbury.<sup>5</sup> Als Begründung der Auswahl der Ziele wurden von deutscher Seite in einer Pressekonferenz des Auswärtigen Amtes im April 1942 die Auszeichnung dieser Städte im Baedeker-Reiseführer für England angegeben.<sup>6</sup> Verständlicherweise sorgte dies in der britischen Bevölkerung für einen ungeheuren öffentlichen Furor, da durch diese Aussage deutlich wurde, dass eben nicht ausschließlich militärische Ziele bzw. Industriestandorte, sondern vielmehr bewusst historisch bedeutsame Städte von deutscher Seite als Ziele ausgewählt wurden.<sup>7</sup>

Diese Vorgänge und insbesondere die enorme öffentliche Empörung dürften dazu geführt haben, dass die Mitarbeiter\*innen des Foreign Office und des Ministry of Economic Warfare, welche die Analyse der militärisch und wirtschaftlich bedeutsamen Ziele im Deutschen Reich bereits 1942 – eben zum Zeitpunkt der sogenannten ›Baedeker Raids‹ – durchführten, ihrerseits wiederum den Namen *Bomber's Baedeker* für die eigenen Aufstellungen wählten. Die erste Auflage des *Bomber's Baedeker* erschien dann auch ein Jahr später und deckt sich mit den alliierten Absprachen zu einem gemeinsamen ›Bomber Command‹ zwischen Großbritannien und den USA auf der Casablanca-Konferenz im Januar 1943, auf der u. a. festgelegt wurde:

---

<sup>4</sup>Vgl. Hohn 1994, S. 213–230.

<sup>5</sup>Vgl. Rothnie 1991, S. 142f.

<sup>6</sup>Mit den Worten »Now the Luftwaffe will go for every building which is marked with three stars in Baedeker« wurde Legationsrat Gustav Braun von Stumm in der britischen Presse zitiert. Vgl. Knuth 2006, S. 165.

<sup>7</sup>Eine virtuelle Ausstellung des Imperial War Museums zu den ›Baedeker Raids‹ findet sich bei Google Arts and Culture.

Vordringliches Ziel des Bomber Command ist die fortschreitende Zerstörung des deutschen militärischen, industriellen und wirtschaftlichen Systems, um die Moral des deutschen Volkes bis zu einem Grad zu untergraben, wo seine Fähigkeit zum bewaffneten Widerstand entscheidend geschwächt ist.<sup>8</sup>

Zeitgleich wurde im britischen Unterhaus das *Dehousing Paper* verabschiedet, in dem als strategisches Ziel der britischen Bomberverbände die gezielte Zerstörung von Wohngebieten vorgesehen wurde – einer Doktrin mit Namen ›Moral Bombing‹, an der bis zum Ende des Zweiten Weltkrieges festgehalten wurde und die bereits im Frühjahr 1942 zum 1.000-Bomber-Angriff gegen Köln führte.<sup>9</sup> Es ist allerdings festzuhalten, dass die Auswahl der tatsächlichen Ziele dem Bomber Command und dem Air Ministry oblag und deshalb nicht eindeutig nachvollziehbar ist, welche Rolle der *Bomber's Baedeker* insbesondere bei der Wahl der Angriffsziele tatsächlich besaß – vor allem, weil in diesem ausschließlich infrastrukturelle und wirtschaftliche Ziele angegeben waren. Die Datengrundlage für die Erhebung durch das Foreign Office und das Ministry of Economic Warfare bildeten – neben Informationen aus der Feindaufklärung, Adressbüchern, Luftaufnahmen, Berichten von Emigrant\*innen etc. – vor allem die Unterlagen der britischen Rückversicherer. Bei diesen waren seit Mitte der 1930er-Jahre die Brandversicherungen der deutschen Versicherungsunternehmen abgesichert. Da es sich um eine Pflichtversicherung für alle Gebäude im Deutschen Reich handelte, stellte dies eine vollständige Datenbasis dar, um eine umfassende Quartiersanalyse aller deutschen Städte zu erstellen. Darüber hinaus konnten hierüber die Standorte aller Firmen identifiziert werden und über deren Namen und Eigentümer\*innen war zugleich zu erfahren, was dort höchstwahrscheinlich produziert wurde. Auch konnte diesen Unterlagen Informationen über die Bausubstanz der Gebäude entnommen werden. Auf dieser Basis war es möglich, eine höchst detaillierte Topographie jeder einzelnen deutschen Stadt zu erstellen und zugleich die Unterschiede in der quartierbezogenen Bausubstanz zu erfassen.<sup>10</sup> De facto war genau dies der Schlüssel für die alliierten Luftangriffe gegen Deutschland, bei denen gezielt Städte bzw. Stadtteile mit leicht entzündbaren und brennbaren Baustoffen angegriffen wurden.<sup>11</sup> Folgt man dieser Argumentation, so wird daraus deutlich, dass *Bomber's Baedeker* eher für strategische Planungen genutzt wurde und weniger eine taktische Bedeutung besaß.

Während in der ersten Auflage des *Bomber's Baedeker* von 1943 nur 392 Städte mit einer Größe über 15.000 Einwohner berücksichtigt wurden, beinhaltet die zweite Auflage von 1944 insgesamt 518 Städte und umfasst auch Kleinstädte ab einer Größe von 1.000 Einwohnern. Dieser Umstand ist u. a. darauf zurückzuführen, dass ab Mitte 1943 eine Verlagerung von kriegswichtigen Produktionsstätten aus den Städten erfolgte.<sup>12</sup> Für die Datenerhebung und #aufbereitung im Rahmen des zugrunde liegenden Projektes wurde die zweite Auflage von 1944 verwendet, die neben der Nennung der jeweiligen Stadt mit entsprechenden Breiten- und Längenangaben und der Flugdistanz (in Meilen) zu London einleitend auch eine kurze Beschreibung der Stadt mit geographischen Markern (in der Nähe liegende Flüsse, Berge, Seen, Wälder etc.) sowie eine Kategorisierung der Ziele gibt.<sup>13</sup> Folgende Kategorisierung wird im *Bomber's Baedeker* für alle Städte angewandt:<sup>14</sup>

---

<sup>8</sup>TNA London, AIR 41/5, Directive 21 January 1943, International Law of the Air, 1939–1945, Confidential supplement to Air Power and War Rights, 1946, zitiert nach: Böhm 2015, S. 147.

<sup>9</sup>Vgl. Longmatte 1983.

<sup>10</sup>Vgl. Hohn 1994, S. 213–230.

<sup>11</sup>Vgl. zur Geschichte des Bombenkriegs gegen deutsche Städte: Overy 2014; Müller 2004; Boog et al. (Hg.) 2001.

<sup>12</sup>Vgl. Schmunk 2005, S. 59, 65f.

<sup>13</sup>Der Unterschied zwischen der ersten und zweiten Auflage besteht vor allem darin, dass in der zweiten Auflage weitere Städte aufgenommen und somit der Datenbestand vergrößert wurde.

<sup>14</sup>Vgl. The Bomber's Baedeker 2019 (1944), Preface.

- Transportwesen
- Infrastruktur (Wasser, Elektrizität etc.)
- Festbrennstoffe (Bergbau, Brennstofflager etc.)
- Flüssigbrennstoffe (Raffinerien, Brennstofflager etc.)
- Eisen- und Stahlindustrie
- Sonstige metallverarbeitende Betriebe
- Flugzeuge und Motoren
- Werften
- Sonstige Industrie- und Rüstungsbetriebe
- Chemie- und Munitionsbetriebe
- Textil-, Seide-, Zellstoff- und Papierbetriebe
- Gummi- und Reifenhersteller
- Lederindustrie
- Nahrungsmittelindustrie

Die aus Sicht der beiden Ministerien kriegswichtige Bedeutung dieser Kategorien wurde zudem auf einer Skala von 1 bis 3 bewertet, wobei anzumerken ist, dass nicht alle Kategorien bei allen Städten zu finden sind.

### 3. Datenerhebung – vom Bild zum XML

Die hier am *Bomber's Baedeker* durchgeführte Transformation des Inhalts eines gedruckten Buchs in maschinenlesbare Daten – auf Grundlage einer zuvor erstellten digitalen 1:1-Abbildung (Repräsentant) – ist ein wichtiger Baustein im Prozess der Datenaufbereitung und zumeist die Voraussetzung für die Anwendung zahlreicher Methoden und Verfahren der Digital Humanities.

Wie oben beschrieben, besitzt die Bibliothek des IEG mit dem zweibändigen Werk eines der wenigen noch verfügbaren Exemplare. Mit dieser Situation geht auch eine entsprechende Verantwortung einher, zum einen in konservatorischer Hinsicht und zum anderen im Hinblick auf die Verfügbarmachung des Inhalts für die Forschung. Beiden Aspekten kann mithilfe der Digitalisierung begegnet werden. Deshalb wurde im Rahmen eines gemeinsamen Vorhabens zwischen IEG und der Universitätsbibliothek der Johannes Gutenberg-Universität Mainz im dortigen Servicezentrum Digitalisierung und Fotodokumentation der *Bomber's Baedeker* im Jahr 2019 digitalisiert. Die angefertigten Digitalisate (in den Formaten ›.jpg‹ und ›.pdf‹) stehen seitdem in *Gutenberg Capture*, dem Online-Portal der Universitätsbibliothek zur digitalen Erschließung und Bereitstellung von Quellenmaterial für die Wissenschaft, zur Verfügung.<sup>15</sup>

Im Zuge der Digitalisierung wurde durch die Universitätsbibliothek auch eine erste Erschließung des Textes mittels Optical Character Recognition (OCR) durchgeführt.

An dieser Stelle setzte die Machbarkeitsstudie an, die im Wintersemester 2019 / 20 an der Hochschule Darmstadt durchgeführt wurde, und in deren Rahmen das Konzept für einen automatisierten Prozess der XML-isierung und einer OCR-Verbesserung des *Bomber's Baedeker* entstand. Dieses Konzept war ein erster, methodologisch wichtiger Baustein, der dann im Rahmen einer Zusammenarbeit zwischen der

---

<sup>15</sup> The *Bomber's Baedeker* 2019 (1944).

Hochschule Darmstadt, der Bibliothek des IEG und dem Bereich Digitale historische Forschung | DH Lab (ebenfalls IEG) umgesetzt wurde. Die ersten experimentellen Ansätze der Transformation der Quelle zu Daten wurden hier realisiert und auf Grundlage dieser datafication – also der Umwandlung von Informationen in maschinenlesbare, quantifizierbare Daten zum Zweck der Aggregation und Analyse – und der entsprechenden Verfügbarmachung ist nun eine weitergehende Bearbeitung und Analyse dieser Daten möglich.

## 3.1. OCR

Die Inhalte des *Bomber's Baedeker* werden bereits im Druck stark strukturiert dargestellt und sind daher grundsätzlich sehr gut geeignet, um die darin enthaltenen Informationen in eine maschinenlesbare, objektorientierte und strukturierte Form zu bringen. Eine nähere Betrachtung ergibt, dass jeweils zwei Hauptabschnitte pro Stadt aufgeführt werden, die sich in folgende Muster aufgliedern:

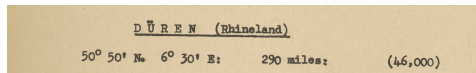


Abb. 1: »Kopfteil« (Düren), Ausschnitt [The Bomber's Baedeker. Guide to the economic importance of German towns and cities, London 1944, S. 176. [Public Domain Mark 1.0]]

### I. Kopfteil

- Name der Stadt (Großbuchstaben, Sperrsatz), dahinter in Klammern: die Verwaltungseinheit und gegebenenfalls die Region.
- In der nächsten Zeile: Geokoordinaten der Stadt im Format ›00° 00' N. 00° 00' E:‹. Hier sind Variationen zu beobachten. So tauchen auch einstellige Angaben auf und es existieren nicht immer Nachkommastellen in den Geokoordinaten, also z. B. ›00° N. 00° E:‹.
- Im Anschluss an die Koordinaten folgt die Entfernung zu London in Meilen, im Format ›000 miles:‹.
- Die letzte Information in der zweiten Zeile des Kopfteils gibt die Einwohnerzahl in Klammern wieder.

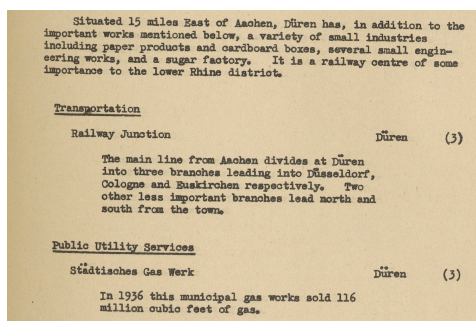


Abb. 2: »Beschreibungsteil« (Düren), Ausschnitt [The Bomber's Baedeker. Guide to the economic importance of German towns and cities, London 1944, S. 176. [Public Domain Mark 1.0]]

## II. Beschreibungsteil

- Eine knappe Beschreibung der wichtigsten Charakteristika der jeweiligen Stadt (z. B. grobe geographische Verortung, Verkehrsinfrastruktur, industrielle Bedeutung).
- Strukturierung nach Kategorien (z. B. ›Transportation‹, ›Liquid Fuels and Substitutes‹ usw.)
- Es folgt eine detaillierte Auflistung der wichtigsten Ziele innerhalb der Kategorie, pro potenzielles Ziel unterteilt in drei Punkte:
  - Name und Beschreibung des Ziels.
  - Standort.
  - Priorität (in absteigender Reihenfolge ›1‹ bis ›3‹ sowie ›–‹ für nicht erwähnenswerte bzw. nicht einschätzbare Ziele).

Ein Blick in eines der bei *Gutenberg Capture* zur Verfügung stehenden Digitalisate im PDF-Format zeigt, dass hier bereits ein OCR-Verfahren eingesetzt wurde. Jede Seite verfügt über einen Text-Layer, der es ermöglicht, im Dokument zu suchen. Die Ergebnisse des verwendeten OCR-Verfahrens sind aber insgesamt nur für eine einfache Suche verwendbar, wenig verlässlich und letztlich nicht für eine direkte Umwandlung im Sinne des Vorhabens geeignet. Nach entsprechender Analyse<sup>16</sup> kann von einer OCR-Genauigkeit von ungefähr 80 % ausgegangen werden, wobei die für die effiziente Umsetzung des Projektes benötigte Erkennungsqualität aber mindestens bei mehr als 95 % liegen sollte.<sup>17</sup> Die vorliegenden OCR-Daten waren daher nicht für eine Prozessierung nutzbar und die OCR musste erneut durchgeführt werden. Angesichts dieses neuen Zwischenschritts ergab sich auch die Situation einer erneuten, genauen Betrachtung des digitalen Ausgangsmaterials. Zwei Probleme zeigten sich direkt, zum einen der Vergilbungsgrad der Seiten und die daraus resultierende zu geringe Helligkeit des Digitalisats, so dass relevante Details nicht von der OCR erkannt werden konnten (false negative) beziehungsweise irrelevante Artefakte auf dem Papier erkannt wurden (false positive). Zum anderen zeigten sich einige Schwächen im Kontrast des Schriftbilds, was angesichts des Alters der Schreibmaschinenseiten des physischen Originals und dessen Nutzungsszenarien nicht weiter verwundert.<sup>18</sup>

Nach der Entscheidung für eine erneute OCR-Behandlung fiel die Wahl zunächst auf die freie Software Tesseract.<sup>19</sup> Allerdings erwies sich diese im operativen Einsatz zum Zeitpunkt Ende 2019 als nicht sehr geeignet bei der Behandlung größerer Textmengen, da eine Aufteilung des Konvoluts notwendig geworden wäre. Im Falle des *Bomber's Baedeker*, mit mehr als 800 Seiten, wäre hier ein erheblicher Zeitaufwand entstanden.

Als Alternative kam stattdessen das Werkzeug FineReader der Firma ABBYY zum Zuge,<sup>20</sup> mit dem sich zum Zeitpunkt der OCR-Verbesserung unkompliziert auch PDF-Dateien größerer Umfangs verarbeiten ließen. Darüber hinaus bietet diese Software die Option, mehrere Sprachen gleichzeitig verarbeiten zu lassen, was

---

<sup>16</sup>Mit einem Zufallsgenerator wurden 500 beliebige Zeichen des Gesamttexts ausgewählt und diese auf Richtigkeit überprüft. Die Auswahl des Validierungsverfahrens, dem sog. Bernoulli-Experiment, basiert auf einer Empfehlung der *DFG-Praxisregeln Digitalisierung*, DFG (Hg.) 2016.

<sup>17</sup>Dies ist u. a. darauf zurückzuführen, dass der Text selbst mit einer Schreibmaschine erstellt und dann hektographiert wurde, so dass einzelne Stellen schlecht lesbar und für die OCR nicht bzw. kaum interpretierbar waren.

<sup>18</sup>Es lassen sich auch Phänomene wie schwächer werdende Farbbänder beobachten, die z. B. auch innerhalb einer Seite für starke Kontrastschwankungen sorgen.

<sup>19</sup>Tesseract-OCR.

<sup>20</sup>ABBYY FineReader.



in der vorliegenden Situation von großem Nutzen war, da der Löwenanteil des Textes zwar auf Englisch<sup>21</sup> verfasst ist, aber aufgrund der behandelten Gegenstände natürlich auch zahlreiche deutsche Begriffe, wie z. B. Städtenamen etc., zu finden sind.

Die vollumfängliche Verarbeitung dauerte drei Stunden und das Ergebnis wurde in einer TXT-Datei gespeichert. Durch dieses Vorgehen konnte die Genauigkeit – ohne Veränderung oder Bearbeitung des Digitalisats – gesteigert werden. Letztlich war aber auch die damit erzielte Gesamtgenauigkeit nicht ausreichend, so dass eine Bildbearbeitung der digitalen Vorlage notwendig wurde, um die anvisierten Ziele in eine realistische Nähe rücken zu lassen. Um eine optimale Erkennung der gedruckten Zeichen zu erreichen, war es nötig, den Kontrast zu erhöhen. Hierzu wurde die in MacOS integrierte Applikation Fotos genutzt, mit der dann auf sehr einfache Art Bildkorrekturen erstellt, kopiert und in der Gesamtheit auf alle Seiten angewendet wurden. Nach dieser Optimierung konnten die Ergebnisse des OCR-Verfahrens erneut und sehr deutlich gesteigert werden, so dass nun eine Genauigkeit von ca. 95 % vorlag.<sup>22</sup> Da auf diesem Weg keine weiteren großen Verbesserungen der OCR-Qualität zu erwarten waren, wurde fortan zur weiteren Steigerung der Datenqualität auf ein Python-Skript gesetzt, so dass zum Zeitpunkt der Veröffentlichung des Data-Papers im September 2021 eine Zeichengenauigkeit von mehr als 99 % erreicht werden konnte.

## 3.2. Python-Skript

Die Struktur des für das weitere Vorgehen geschriebenen Python-Skripts kann in drei Schritte unterteilt werden. Schritt Eins stellt das Preprocessing dar. Hier werden wiederholt auftretende Fehler aus dem OCR-Verfahren verbessert, die in der nachfolgenden Ausführung weiterer Schritte im Skript zu Fehlern oder fehlerhafter Erfassung der Objekte führen könnten. Zusätzlich werden strukturelle Abweichungen an die im *Bomber's Baedeker* vorwiegend genutzten Normen angepasst und es werden Funktionen ausgeführt, die folgende Muster in den oben beschriebenen Kopf- und Beschreibungsteilen erkennen: Städtenamen, Land / Region, Koordinaten, Entfernung zu London, Bevölkerungszahl, Informationstext und strategische Ziele der beschriebenen Stadt. Darüber hinaus wird auf Seitenebene die referenzierende URL in *Gutenberg Capture* erzeugt.

Die oben genannten Funktionen liefern eine Liste von erkannten Inhalten und hier setzt der zweite Schritt an. Die Anzahl der Inhalte ist aufgrund der vorgegebenen Struktur immer dieselbe und jeder Listeneintrag korrespondiert mit dem entsprechenden Eintrag in der Vorlage. Der vierte Eintrag in der generierten Liste der Bevölkerungszahlen zum Beispiel ist aus dem vierten Eintrag in der Liste der erkannten Städte entstanden. Für die weitere Verarbeitung werden diese Listen nun in ein sogenanntes Dictionary umgewandelt und hier wird jedem Wert ein Schlüssel zugeordnet, wodurch die Daten nun so gespeichert werden können, dass klar ist, welche Informationen sie enthalten. Auf diese Weise haben wir ein Dictionary mit Listen erstellt, die schon alle benötigten Informationen enthalten.

---

<sup>21</sup>Zum Zeitpunkt der Durchführung dieses Arbeitsschritts im Projekt verfügte Tesseract nicht über die entsprechenden Funktionalitäten.

<sup>22</sup>Beim Präprozessieren sollten zukünftig alternative Verfahren zur Kontrastkorrektur durchgeführt werden – Abbyy Finereader unterstützt dies ebenfalls.

Im dritten und letzten Schritt wird dieses Dictionary dann in eine XML-Datei eingefügt.<sup>23</sup> Damit hier nicht in eine komplett leere Datei geschrieben wird, wurde die Grundstruktur vorher schon erstellt. Beim Prozess des Exports werden alle Inhalte aus dem Data-Dictionary eingelesen, sowie die Seitenzahlen aus dem Originaldokument. Nun kann eine XML-Baumstruktur erzeugt werden, an die zuerst die Städtenamen angefügt werden. Anschließend wird jeder Stadt-Eintrag mit den dazugehörigen Informationen gefüllt. Als letzter Schritt wird die XML-Datei exportiert und die entstandenen HTML-Entitäten werden aufgelöst.<sup>24</sup>

### 3.3. Bereitstellung

Neben dem Python-Skript für die OCR-Optimierung finden sich in dem Projekt-Repository auch die aus seiner Anwendung entstandenen Daten: für jeden Band des *Bomber's Baedeker* liegt eine XML-Datei vor, in welcher der entsprechend formal strukturierte und angereicherte Text enthalten ist.<sup>25</sup> Zukünftige Versionen mit verbesserter Datenqualität werden dort ebenfalls publiziert.

Bei der Bereitstellung des Datensatzes und des Skripts wurden die FAIR Data Principles umfänglich berücksichtigt.<sup>26</sup> Für die Erfüllung dieser Prinzipien sorgen die Zugriffsmöglichkeiten per GitHub-Repository und die Veröffentlichung der erzeugten Daten in Zenodo, womit ein persistenter Identifikator (DOI) einhergeht.<sup>27</sup> Für die rechtliche Sicherheit bei der Wiederverwendung von Daten und Skript sorgen die gewählten Lizenzen (Creative Commons Lizenz, CC BY-SA 4.0<sup>28</sup> für die Daten sowie die GNU General Public License<sup>29</sup> für das Python-Skript).

## 4. Forschungs- und Nachnutzungspotenzial

Wie oben bereits geschildert, handelte es sich bei dieser Bereitstellung der nachnutzbaren Daten des *Bomber's Baedeker* um ein Vorhaben, welches mit einem überschaubaren Ressourceneinsatz durchgeführt wurde. Der Fokus lag dabei primär auf der persistenten Verfügbarmachung des Datenbestands für die scientific community. Eine basale, aber dennoch nicht zu vernachlässigende Verbesserung, die sich für die Forscher\*innen direkt aus dem digitalen Angebot ergibt, spiegelt sich im vereinfachten Umgang mit dem Text wider. In den bereitgestellten Daten lassen sich Such- und Analyseszenarien durchspielen, deren Umsetzung ausschließlich auf dieser digitalen Grundlage möglich sind.

Die bislang durchgeführten Arbeiten haben die Grundlagen geschaffen, um in einem nächsten Schritt ein digitales Editionsvorhaben durchzuführen. Allein im Bereich der Auszeichnung und Anreicherung mit Normdaten – z. B. bei den Städte- oder Firmennamen – besteht großes Potenzial für weitere Analysen. Neben der weiteren Aufbereitung der Daten sehen wir ein breites Feld an möglichen Forschungsfragen,

---

<sup>23</sup>Ein mit XML ausgezeichnete Text bietet sich hier als Zielformat an, da entsprechend strukturierte Dateien für weitere Möglichkeiten der Nachnutzung offenbleiben und jedes spezifische Format an dieser Stelle ohne konkrete Aufgabenstellung produziert worden wäre.

<sup>24</sup>Das hier genutzte Skript ist samt Dokumentation im GitHub-Repository des Leibniz-Instituts für Europäische Geschichte unter folgender URL aufzurufen: [https://github.com/ieg-dhr/bombers\\_baedeker/](https://github.com/ieg-dhr/bombers_baedeker/).

<sup>25</sup>Das Verzeichnis >bomber\_output\_ext< enthält das XML-Ergebnis des datengenerierenden Skripts.

<sup>26</sup>FAIR Data Principles, GOFair (Hg.) 2016–2021.

<sup>27</sup>Bach et al. 2021.

<sup>28</sup>CC BY-SA 4.0, Creative Commons – Namensnennung – Weitergabe unter gleichen Bedingungen – 4.0 International, Creative Commons (Hg.) 2021.

<sup>29</sup>GNU General Public License, Free Software Foundation, Inc. (Hg.) 2007.

die an den Text bzw. die Daten gestellt werden können. So wäre ein Abgleich der im *Bomber's Baedeker* vorgeschlagenen Ziele im Deutschen Reich mit den tatsächlich bombardierten Städten und Einrichtungen eine interessante Forschungsfrage, um einerseits Rückschlüsse auf die praktische Anwendung im Hinblick auf die Zielvorgaben der Royal Airforce zu geben. Andererseits können die detaillierten Standortinformationen der Firmen aus wirtschaftshistorischer Perspektive erstmals einen annähernd vollständigen Überblick über die Branchenverteilung in den 1930er- und 1940er-Jahren geben. In beiden Fällen wäre noch zu prüfen, wie exakt die vermerkten Geokoordinaten angegeben sind und ob sich gegebenenfalls ein geographisches Muster aus den ermittelten Abweichungen herauslesen lässt.

Unsere Arbeiten ermöglichen nun einen datengetriebenen Analyseansatz. Während mit einer traditionellen Methodik oftmals Karten ausschließlich rein visuell analysiert werden und zugleich deren Richtigkeit und Entstehungskontexte nur partiell hinterfragt werden können, z. B. ob die Ersteller\*innen – wie in Abb. 3 dargestellt – eindeutige und gültige Geoangaben verwendet haben, besitzen datengetriebene Ansätze der Visual Analytics weitaus größere Möglichkeiten und können aufzeigen, auf welcher Datenbasis die entsprechenden Angaben generiert wurden.

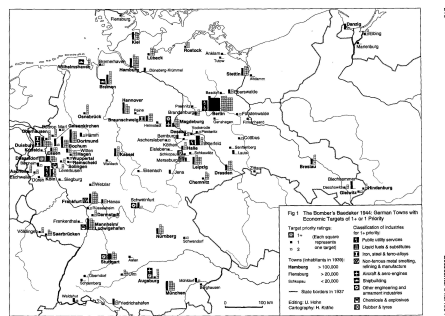


Abb. 3: Klassische Visualisierung der Ziele nach höchster Priorität aus dem *Bomber's Baedeker*. [Uta Hohn: *The Bomber's Baedeker – Target Book for Strategic Bombing in the Economic warfare against German Towns 1943–1954*. In: [Hohn 1994, S. 216.]

Dies ist allerdings erst dann möglich, wenn die entsprechenden Angaben maschinenlesbar vorliegen und so aufbereitet werden, dass sie visuell interpretiert werden können. Die Darstellung in Abb. 4 wurde – basierend auf den im *Bomber's Baedeker* gemachten Angaben der Geokoordinaten – mittels Tableau<sup>30</sup> und dem DARIAH-DE Geo-Browser<sup>31</sup> angefertigt. Hier ist erkennbar, dass die Ersteller\*innen des *Bomber's Baedeker* eine klassische (Raster-)Papierkarte als Vorlage benutzt haben müssen, mit deren Hilfe die entsprechenden Geokoordinaten in die Textfassung übertragen wurden. Auffällig ist, dass die validen Geokoordinaten aufgrund des damaligen technischen Standes der Flugzeug-Leitsysteme keine Bedeutung besaßen, da die Flugzeuge mittels Radio-Leitstrahlen navigierten. Die Geokoordinaten hatten dementsprechend nur einen untergeordneten Informationsgehalt. Auch können auf diese Weise fehlerhafte Angaben im *Bomber's*

<sup>30</sup> Tableau, Tableau Software (Hg.) 2003–2021.

<sup>31</sup> DARIAH-DE Geo-Browser, DARIAH-DE (Hg.) 2021.

*Baedeker* schneller identifiziert werden – wie beispielsweise, dass Mannheim versehentlich mit den Geokoordinaten von Kiew versehen wurde. Alleine durch dieses Beispiel wird deutlich, dass die Angabe der Geokoordinaten nur eine zusätzliche faktisch rein kartographische Information darstellt.

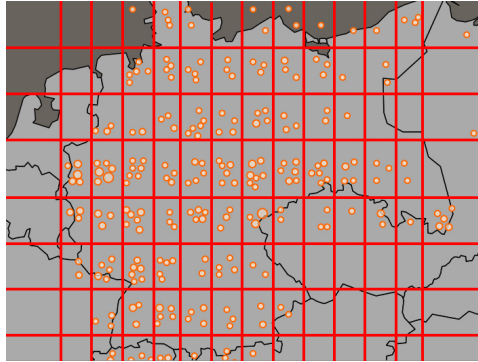


Abb. 4: Visualisierung der im Bomber's Baedeker benannten Zielstädte basierend auf den angegebenen zeitgenössischen Geokoordinaten mittels DARIAH-DE Geo-Browser und Tableau. [Bach / Secco 2021]

Wichtiger war zum Zeitpunkt der Erstellung in den 1940er-Jahren also vielmehr die Entfernung zu London und über welche Planquadrate die Flugzeuge fliegen mussten, um ihre Ziele zu erreichen. Wetterbedingungen, die Abwehrmaßnahmen der Deutschen Luftwaffe, aber auch welche Ziele in der Nähe lagen, die alternativ angefliegen werden konnten, wenn das eigentliche Ziel nicht erreicht werden konnte, spielten eine größere Rolle, als die richtige Angabe der Geokoordinaten.

Allein durch dieses Beispiel wird deutlich, dass die basale Erschließung dieses Datensatzes in XML und die dadurch hergestellte Maschinenlesbarkeit ungeheures Forschungspotenzial verspricht. Momentan arbeiten wir daran, dieses Beispiel auf die Ebene der einzelnen Städte herunterzubrechen, um zu analysieren und einzuschätzen, wie gut und valide die Alliierten tatsächlich über die Standorte von Infrastrukturen und Industrieanlagen informiert waren. Der Abgleich mit zeitgenössischen Stadtplänen und Adressbüchern ermöglicht diesbezüglich die Durchführung valider Datenanalysen.

## 5. Ausblick

Wie nicht zuletzt an diesem Beispiel der Verfügbarmachung des *Bomber's Baedeker* nachzuvollziehen ist, stellen wissenschaftliche Bibliotheken bereits eine große Anzahl von Digitalisaten bereit. In einigen Fällen wurden bereits OCR-Verfahren eingesetzt, um den Nutzer\*innen weitergehende Möglichkeiten zu bieten. Leider liegen die angebotenen OCR-Daten nicht immer in einer Qualität vor, die eine digitale Bearbeitung ohne weiteren Aufwand für die Vorverarbeitung erlaubt. Diese Situation belastet die ohnehin knappen Ressourcen in Forschungsprojekten.

Es wäre wünschenswert, wenn die hier festgehaltene Herausforderung der Erstellung von maschinenlesbaren Datensets zukünftig von Bibliotheken, Forschungsdateninfrastrukturinitiativen und Wissenschaftler\*innen gemeinsam adressiert werden könnte. Auf diese Weise kann eine höhere, leichter verarbeitbare Datenqualität in den Angeboten der Bibliotheken zur Unterstützung der digital forschenden Wissenschaftler\*innen generiert werden.

Nachdem die Bereitstellung des Volltextes für die Nutzer\*innen mehr oder weniger fest im Portfolio der Bibliotheken verankert ist, wäre der nächste Schritt, das Bewusstsein für die Notwendigkeit einer hohen Datenqualität (über die Metadaten hinaus) zu schärfen. Trotz vieler guter Beispiele für Ansätze und Umsetzungen sind hier weitere Aktivitäten notwendig, um das Mindset und die aktuelle Praxis zu verändern. Die geisteswissenschaftlichen Initiativen in der Nationalen Forschungsdateninfrastruktur (NFDI) und seine relevanten Teilnehmenden wären die idealen Multiplikator\*innen, um die Standards der Datenbereitstellung für digitale Texte voranzutreiben.

## Bibliographische Angaben

Felix Bach / Cristian Secco / Stefan Schmunk / Thorsten Wübbena: The Bomber's Baedeker. A Guide to the Economic Importance of German Towns and Cities. In: zenodo.org. Data set vom 26.07.2021. DOI: [10.5281/zenodo.5138504](https://doi.org/10.5281/zenodo.5138504)

Das Deutsche Reich in der Defensive – Strategischer Luftkrieg in Europa, Krieg im Westen und in Ostasien 1943 bis 1944/45. Hg. von Horst Boog / Gerhard Krebs / Detlef Vogel. Stuttgart u. a. 2001. (= Das Deutsche Reich und der Zweite Weltkrieg, 7) [[Nachweis im GBV](#)]

Martin Böhm: Die Royal Air Force und der Luftkrieg 1922–1945. Personelle, kognitive und konzeptionelle Kontinuitäten und Entwicklungen. Paderborn 2015. [[Nachweis im GBV](#)]

Rainer Burger: Handbuch für den Feuersturm. In: Frankfurter Allgemeine Zeitung. Artikel vom 26.06.2019. [[online](#)]

Creative Commons-Lizenzen. Hg. von Creative Commons. Mountain View, CA 2021. [[online](#)]

DARIAH-DE Geo-Browser. Hg. von DARIAH-DE. Version 3.6.7. Göttingen 2021. [[online](#)]

DFG-Praxisregeln ›Digitalisierung‹. Hg. von Deutsche Forschungsgemeinschaft. Bonn 2016. PDF. [[online](#)]

FAIR Data Principles. Hg. von GOFair. In: go-fair.org. Leiden u. a. 2016–2021. [[online](#)]

GNU General Public License. Hg. von Free Software Foundation, Inc. Version 3 vom 29.06.2007. [[online](#)]

Uta Hohn: The Bomber's Baedeker-target book for strategic bombing in the Economic Warfare against German Towns 1943–45. In: GeoJournal 34 (1994), H. 2, S. 213–230. [[Nachweis im GBV](#)]

Rebecca Knuth: Burning Books and Leveling Librarians. Extremist Violence and Cultural Destruction. Westport, CT 2006. [[Nachweis im GBV](#)]

Norman Longmatte: The Bombers: The RAF offensive against Germany 1939–1945. London u. a. 1983. [[Nachweis im GBV](#)]

Rolf-Dieter Müller: Der Bombenkrieg 1939–1945. Berlin 2004. [[Nachweis im GBV](#)]

Susanne Müller: Die Welt des Baedeker. Eine Medienkulturgeschichte des Reiseführers 1830–1945. Frankfurt/Main u. a. 2012. [[Nachweis im GBV](#)]

Richard Overy: Der Bombenkrieg: Europa 1939–1945. Berlin 2014. [[Nachweis im GBV](#)]

Niall Rothnie: The Baedeker Blitz. Hitler's Attack on Britain's Historic Cities. Shepperton 1992. [[Nachweis im GBV](#)]

Stefan Schmunk: Entweder KZ oder ordentliche Deutsche. Die Luftwaffe und der Arbeitseinsatz 1942–1944. Darmstadt 2005. In: researchgate.net. DOI: [10.13140/rg.2.2.20030.08003](https://doi.org/10.13140/rg.2.2.20030.08003)

Tableau. Hg. von Tableau Software. In: tableau.com. Seattle, WA 2003–2021. [[online](#)]

The Bomber's Baedeker. Guide to the economic importance of German towns and cities. (Foreign Office & Ministry of Economic Warfare). 2 Bände. London 1944. In: Gutenberg Capture. Hg. von Universitätsbibliothek Mainz. Online-Ausgabe. Mainz 2019. URN: [urn:nbn:de:hebis:77-vcol-20056](https://nbn-resolving.org/urn:nbn:de:hebis:77-vcol-20056)

The Bomber's Baedeker Guide to the economic importance of German towns and cities. London 1944. Softwareskripte und Dokumentation. Hg. von IEG Mainz. In: github.com. 2021. [[online](#)]

TNA London, AIR 41 / 5, Directive 21 January 1943, International Law of the Air, 1939–1945, Confidential supplement to Air Power and War Rights, 1946, zitiert nach: Böhm, Martin: Die Royal Air Force und der Luftkrieg 1922–1945. Paderborn 2015, S. 147.

Baedeker Raids. The story of the historic towns and cities in Britain targeted by the German Air Force in Spring 1942. Hg. von Imperial War Museums. London 05.07.2019–05.01.2020. In: Google Arts and Culture. Virtuelle Ausstellung. 2021. [[online](#)]

## Abbildungsverzeichnis

Abb. 1: ›Kopfteil‹ (Düren), Ausschnitt [The Bomber's Baedeker. Guide to the economic importance of German towns and cities, London 1944, S. 176. Public Domain Mark 1.0; online].

Abb. 2: ›Beschreibungsteil‹ (Düren), Ausschnitt [The Bomber's Baedeker. Guide to the economic importance of German towns and cities, London 1944, S. 176. Public Domain Mark 1.0; online].

Abb. 3: Klassische Visualisierung der Ziele nach höchster Priorität aus dem Bomber's Baedeker. [Uta Hohn: The Bomber's Baedeker – Target Book for Strategic Bombing in the Economic warfare against German Towns 1943–1954. In: Hohn 1994, S. 216.]

Abb. 4: Visualisierung der im Bomber's Baedeker benannten Zielstädte basierend auf den angegebenen zeitgenössischen Geokoordinaten mittels DARIAH-DE Geo-Browser und Tableau. [Bach / Secco 2021]