Publishing an OCR ground truth data set for reuse in an unclear copyright setting.

David Lassner, Clemens Neudecker, Julius Coburger, Anne Baillot

ZfdG

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

Titel:
Publishing an OCR ground truth data set for reuse in an unclear copyright setting. Two case studies with legal and technical solutions to enable a collective OCR ground truth data set effort

Autor*in:
David Lassner

Kontakt: lassner@tu-berlin.de
Institution: Technische Universität Berlin, Machine Learning Group | The Berlin Institute for the Foundations of Learning and Data (BIFOLD)
GND: 1246941414 ORCID: 0000-0001-9013-0834

Autor*in:
Julius Coburger

Kontakt: julius.coburger@gmx.de
Institution: Technische Universität Berlin, Machine Learning Group
GND: 124694197X ORCID: 0000-0003-4502-7955

Autor*in:
Clemens Neudecker

Kontakt: clemens.neudecker@sbb.spk-berlin.de
Institution: Staatsbibliothek zu Berlin – Preußischer Kulturbesitz
GND: 1246943069 ORCID: 0000-0001-5293-8322

Autor*in:
Anne Baillot

Kontakt: anne.baillot@univ-lemans.fr
Institution: Le Mans Université | École normale supérieure de Lyon, Interactions, Corpus, Apprentissages, Représentations - ICAR
GND: 1065904681 ORCID: 0000-0002-4593-059X

DOI des Artikels:
10.17175/sb005_006

Nachweis im OPAC der Herzog August Bibliothek:
1780168195

Erstveröffentlichung:
10.12.2021

Letzte Überprüfung aller Verweise: 02.12.2021

GND-Verschlagwortung:

Informatik | Maschinelles Lernen | Optische Zeichenerkennung | Urheberrecht |

Zitierweise:
David Lassner, Julius Coburger, Clemens Neudecker, Anne Baillot: Publishing an OCR ground truth data set for reuse in an unclear copyright setting. Two case studies with legal and technical solutions to enable a collective OCR ground truth data set effort. 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_006.

David Lassner, Julius Coburger, Clemens Neudecker, Anne Baillot

Publishing an OCR ground truth data set for reuse in an unclear copyright setting. Two case studies with legal and technical solutions to enable a collective OCR ground truth data set effort

Abstracts

In dieser Arbeit stellen wir einen OCR-Trainingsdatensatz für historische Drucke vor und zeigen, wie sich im Vergleich zu unspezifischen Modellen die Erkennungsgenauigkeit verbessert, wenn sie mithilfe dieser Daten weitertrainiert werden. Wir erörtern die Nachnutzbarkeit dieses Datensatzes anhand von zwei Experimenten, die die rechtliche Grundlage zur Veröffentlichung digitalisierter Bilddateien am Beispiel von deutschen und englischen Büchern des 19. Jahrhunderts betrachten. Wir präsentieren ein Framework, mit dem OCR-Trainingsdatensätze veröffentlicht werden können, auch wenn die Bilddateien nicht zur Wiederveröffentlichung freigegeben sind.

We present an OCR ground truth data set for historical prints and show improvement of recognition results over baselines with training on this data. We reflect on reusability of the ground truth data set based on two experiments that look into the legal basis for reuse of digitized document images in the case of 19th century English and German books. We propose a framework for publishing ground truth data even when digitized document images cannot be easily redistributed.

1. Introduction

Digital access to Cultural Heritage is a key challenge for today's society. It has been improved by Optical Character Recognition (OCR), which is the task by which a computer program extracts text from a digital image in order to draw the text from that image and present it in a machine-readable form. For historical prints, off-the-shelf OCR solutions often result in inaccurate readings. Another impediment to accessing digitized cultural heritage data consists in the fact that cultural heritage institutions provide online access to massive amounts of digitized images of historical prints that have not been (or have been poorly) OCRed. Solutions to improve this situation would benefit a wide range of actors, be they scholars or a general audience. Many actors would indeed profit greatly from methods conceived to extract high quality machine-readable text from images.

The results of an OCR method can be improved significantly by using a pre-trained model and fine-tuning it on only a few samples that display similar characteristics.[1] To that end, there has been a growing effort from the Digital Humanities community to create and publish data sets for specific historical periods, languages and typefaces aiming at enabling scholars to fine-

[1] See Liebl / Burghardt 2020; Reul et al. 2017; Springmann et al. 2018.

tune OCR models for their collection of historical documents.[2] In Germany, the DFG-funded OCR-D initiative brings together major research libraries with the goal to create an open source framework for the OCR of historical printed documents, including specifications and guidelines for OCR ground truths.[3]

In order to improve OCR results, images and the corresponding transcriptions are collected in such a way that each pair (image and text) only represents one line of text from the original page. This is called a ground truth data set and is precisely what we will focus on in the following.

Besides the fact that creating transcriptions of images manually is tedious work, another major issue arises from this type of collective effort in that the institutions that produce the scan often claim some form of copyright to it. For example, on the first page of any of their PDFs, Google Books »[...] request[s] that you use these files for personal, non-commercial purposes«[4]. As a consequence, a scholar aiming to create an OCR ground truth data set would not know with certainty whether the rights to redistribute the textline images derived from the PDF can be considered as granted.

In this paper, we present an OCR ground truth data set with an unclear copyright setting for the image data. We discuss the legal background, show the relevance of the data set and provide in-depth analysis of its constitutiq on and reuse by investigating two different approaches to overcome the copyright issues.

In order to address these issues, we compare in the following two ways to publish the OCR ground truth data set with image data.

- As Google Books works with cultural heritage institutions (CHIs) to digitize books, we asked permission from the CHIs to redistribute the image data.
- We published a data set formula, which consists of the transcriptions, links to the image sources, and a description on how to build the data set. For this process, we provide a fast, highly automated framework that enables others to reproduce the data set.

2. Legal background and its interpretation at CHIs

Clarifying the copyright situation for the scans of a book collection requires to take into account, for each book, the cultural heritage institution owning the book (usually a library), and, in the case of private-public partnerships, also the scanning institution (e. g. Google Books) involved in its digitization. For Google Books, there exist different contracts between CHIs and

[2] See Padilla et al. 2019. For manuscripts, just recently the Transcriptiones platform launched, see transcriptiones, ETH-Library 2020. For French texts from the 18th to the 21st century there exists HTR-United, see htr-united, Chagué / Clérice 2021. The slightly different approach of just publishing fine-tuned models for different settings is proposed by Transkribus, see Transkribus, READ-COOP 2021, or Kraken 2021 ocr_models, OCR/HTR model repository 2021.
[3] See Engl 2020.
[4] Google Inc. 2021, cited after Ruiz 2011.

Google, and not all of them are open to public inspection. However, based on comparing the ones that are available, we assume that other contracts are to some extent similar (see List of Contracts). The contracts contain information on the ›Library Digital Copy‹ for which non-profit uses are defined under Section 4.8 (cf. British Library Google Contract), which states that a

> »Library may provide all or any portion of the Library Digital Copy, that is [...] a Digital Copy of a Public Domain work to (a) academic institutions or research libraries, or (b) when requested by Library and agreed upon in writing by Google, other not-for-profit or government entities that are not providing search or hosting services substantially similar to those provided by Google.«[5]

When trying to unpack this legal information against the use case presented here, multiple questions arise. What are the legal possibilities for individual scholars regarding the use of the Library Digital Copy of a Public Domain work? How can there be limitations in the use of a Public Domain work? Is the use case of OCR model training substantially similar to any search or hosting services provided by Google? Would and can libraries act as brokers in negotiating written agreements about not-for-profit use with Google?

In the continuation of Section 4.8, additional details are specified with regard to data redistribution by ›Additional institutions‹ where

> »[a written agreement with Google] will prohibit such Additional institution from redistributing [...] portions of the Library Digital Copy to other entities (beyond providing or making content available to scholars and other users for educational or research purposes.«[6]

This brings up further questions but also opens the perspective a bit, since there appear to be exceptions for »scholars and other users for educational or research purposes«[7], which is a precise fit of the use case we present here. Now what does this mean in practice? Digital Humanities scholars are not necessarily legal experts, so how do libraries that have entered public-private-partnerships with Google for digitization of Public Domain works implement these constraints? Schöch et al. discuss a wide range of use cases in the area of text and data mining with copyright protected digitized documents, but they do not cover the creation and distribution of ground truth.[8] In other scenarios that involve copyrighted texts published in derived formats, one question typically preventing redistribution is whether it is possible to re-create the (copyright-protected) work from the derived parts. In the case of textline ground truth, it is however likely that this would constitute a violation of such a principle. In this unclear setting, scholars are in need of support and guidance by CHIs.

[5] British Library Google Books Agreement in Ruiz 2011.
[6] British Library Google Books Agreement in Ruiz 2011.
[7] British Library Google Books Agreement in Ruiz 2011.
[8] See Schöch et al. 2020.

Institution	Total # books	Total # pages	Response time (# working days)	Allowed to publish as part of the paper	Allowed to license	Alternative source	Responsible	Citation needed
Bayerische Staatsbibl.	4	12	3	yes	yes	yes	yes	yes
Biblioteca Statale Isontina Gorizia	1	3	–	–	–	–	–	–
Bodleian Library	11	20	2	yes, alternative	already CC-BY-NC	yes	yes	yes
British Library	1	35	4	no	no	no	yes	–
Harvard University, Harvard College Library	1	3	0	yes	yes	yes	no	yes
New York Public Library	5	29	3	–	–	no	no	no
Austrian National Library	2	6	10	yes, alternative	no	yes	yes	yes
Robarts – University of Toronto	2	3	–	–	–	–	–	–
University of Illinois Urbana-Champaign	6	4	0	yes	yes	no	yes	yes
University of Wisconsin – Madison	8	24	2	yes	yes	no	no	no

Tab. 1: Responses of library institutions to our request to grant permission to publish excerpts of the scans for which they were contractors of the digitization. Most institutions responded within a few working days and except for the fact that most acknowledged the public domain of the items, the responses were very diverse. Many answered that they are either not responsible or only responsible for their Library Copy of the PDF. [Lassner et al. 2021]

We have asked ten CHIs for permission to publish image data that was digitized based on their collection in order to publish them as part of an OCR ground truth data set under a CC-BY license. As shown in Table 1, the institutions gave a wide variety of responses. Many institutions acknowledged that the requested books are in the public domain because they were published before the year 1880. However, there is no general consensus on whether the CHIs are actually responsible for granting these rights, especially if one wants to use the copy from the Google Books or Internet Archive servers. Some institutions stated that they are only responsible for their Library Copy of the scan and granted permission to publish only from that source. Only two institutions, the Bayerische Staatsbibliothek and University of Illinois Urbana-Champaign stated that they are responsible and that we are allowed to also use the material that can be found on the Google Books or Internet Archive servers.

This case study underlines the lack of a clear and simple framework of reference that would be recognized and applied, and would reflect on good practices in the relationships between CHIs and digital scholarship. The lack of such a framework is addressed among others by the DARIAH initiative of the Heritage Data Reuse Charter[9] that was launched in 2017. Another approach towards such a framework is that of the ›digital data librarian‹.[10]

3. Description of the data set

In the data set that we want to publish in the context of our OCR ground truth, we do not own the copyright for the image data.[11] We therefore distinguish between the data set formula and the built data set. We publish the data set formula which contains the transcriptions, the links to the images and a recipe on how to build the data set.

The data set formula and source code are published on Github[12] and the version 1.1 we are referring to in this paper is mirrored on the open access repository Zenodo.[13] The data set is published under a CC-BY 4.0 license and the source code is published under an Apache license.

3.1 Origin

The built data set contains images from editions of books by Walter Scott and William Shakespeare in the original English and in translations into German that were published around 1830.

[9] See Baillot et al. 2016. For additional information on the DARIAH Heritage Data Reuse Charter, see data-reuse, DARIAH 2021.
[10] See Eclevia et al. 2019.
[11] The current version of the data set can be found at ocr-data/data, OCR-Data 2021.
[12] See ocr-data, OCR-Data 2021.
[13] See Lassner et al. 2021.

The data set was created as part of a research project that investigates how to implement stylometric methods that are commonly used to analyze the style of authors with the goal of analyzing that of translators. The data set was organized in such a way that other variables like authors of the documents or publication date can be ruled out as a confounder of the translator style.

We found that 1830 Germany was especially suitable for the research setting we had in mind. Due to an increased readership in Germany around 1830, there was a growing demand in books. Translating foreign publications into German turned out to be particularly profitable because, at that time, there was no copyright regulation that would apply equally across German-speaking states. There was no general legal constraint to regulate payments to the original authors of books or as to who was allowed to publish a German translation of a book. Therefore, publishers were competing in translating most recent foreign works into German, which resulted in multiple German translations by different translators of the same book at the same time. To be the first one to publish a translation into German, publishers resorted to what was later called translation factories, optimized for translation speed.[14] The translators working in such ›translation factories‹ were not specialized in the translation of one specific author. It is in fact not rare to find books from different authors translated by the same translator.

3.2 Method

We identified three translators who all translated books from both Shakespeare and Scott, sometimes even the same books. We also identified the English editions that were most likely to have been used by the translators. This enabled us to set up a book-level parallel English-German corpus allowing us to, again, rule out the confounding author signal.

As the constructed data set is only available in the form of PDFs from Google Books and the Internet Archive or the respective partner institutions, OCR was a necessary step for applying stylometric tools on the text corpus. To assess the quality of off-the-shelf OCR methods and to improve the OCR quality, for each book, a random set of pages was chosen for manual transcription.

3.2.1 Preparation

Following the OCR-D initiative's specifications and best practices,[15] for each book, we created a METS[16] file that contains the link to the source PDF as well as the chosen pages. The following example presents an excerpt from one of the METS files:

[14] See Bachleitner 1989.
[15] See ocr-d spec, OCR-D 2021.
[16] See METS, The Library of Congress 2021.

```
...
<mets:fileGrp USE="IMG">
   <mets:file ID="pdf_2jMfAAAAMAAJ_28" MIMETYPE="application/pdf">
      <mets:FLocat LOCTYPE="URL" xlink:href="http://books.google.com/books?id=2jMfAAAAMAAJ#page=28"/>
   </mets:file>
   <mets:file ID="pdf_2jMfAAAAMAAJ_103" MIMETYPE="application/pdf">
      <mets:FLocat LOCTYPE="URL" xlink:href="http://books.google.com/books?id=2jMfAAAAMAAJ#page=103"/>
   </mets:file>
   ...
</mets:fileGrp>
...
```

Fig. 1: Excerpt of a METS file as used in our data set. For each book, we created one METS file. The link to the resource contains the identifier and the page number. [Lassner et al. 2021]

The PDFs have been downloaded from the URLs in this METS file, and the page images have been extracted from the PDF, deskewed and saved as PNG files.[17]

3.2.2 Transcription

For transcription, the standard layout analyzer of Kraken 2.0.8 (depending on the layout either with black or white column separators) has been used and the transcription was pre-filled with either the German Fraktur or the English off-the-shelf model and post-corrected manually. To ensure consistency, some characters were normalized: for example, we encountered multiple hyphenation characters such as - and # which were both transcribed by -.

3.2.3 Size

In total, the data set contains 5,354 lines with 224,745 characters. It consists of German and English books from 1815 to 1852. A detailed description of the characteristics of the data set is shown in Table 2.

3.3 Reproducibility and Accessibility

The data set formula has been published as a collection of PAGE files and METS files.[18] The PAGE files contain the transcriptions on line-level and the METS files serve as the container linking metadata, PDF sources and the transcriptions. There exists one METS file per item (corresponding to a Google Books or Internet Archive id) and one PAGE file per PDF page. The following excerpt of an example PAGE file shows how to encode one line of text:

```
...
<TextLine id="textline_2">
   <Coords points="457,124 457,1712 534,1712 534,124"/>
   <TextEquiv>
      <Unicode>wenn von ſtarker Faust ein Stoß über das Schlür-</Unicode>
   </TextEquiv>
</TextLine>
...
```

Fig. 2: Excerpt from the PAGE file showing the bounding box of the line on the page image and the corresponding text string. [Lassner et al. 2021]

[17] The process is implemented in the pdfs.py submodule pdfs.py:23 and it uses the command line tools imagemagick and pdfimages, see OCR-Data 2021.
[18] See Pletschacher / Antonacopoulos 2010.

The <TextLine> <TextLine> contains the absolute pixel coordinates where the text is located on the preprocessed PNG image and the <TextEquiv> <TextEquiv> holds the transcription of the line.

As shown above, the METS files contain links to the PDFs. Additionally, the METS files contain links to the PAGE files as shown in the following excerpt.

```
<mets:fileGrp USE="GT">
    <mets:file ID="gt_2jMfAAAAMAAJ_28" MIMETYPE="text/xml">
        <mets:FLocat LOCTYPE="URL" xlink:href="data/xml_output/2jMfAAAAMAAJ_28.page"/>
    </mets:file>
    <mets:file ID="gt_2jMfAAAAMAAJ_103" MIMETYPE="text/xml">
        <mets:FLocat LOCTYPE="URL" xlink:href="data/xml_output/2jMfAAAAMAAJ_3103.page"/>
    </mets:file>
    <mets:file ID="gt_2jMfAAAAMAAJ_132" MIMETYPE="text/xml">
        <mets:FLocat LOCTYPE="URL" xlink:href="data/xml_output/2jMfAAAAMAAJ_3132.page"/>
    </mets:file>
    ...
</mets:fileGrp>
```

Fig. 3: Excerpt from the METS file as used in our data set. For each book, we created one METS file. This part of the METS file contains the references to the PAGE files. [Lassner et al. 2021]

As one can see, there are links from one METS file, namely the one encoding works by Walter Scott's, Volume 2, published by the Schumann brothers in 1831 in Zwickau, identified by the Google Books id 2jMfAAAAMAAJ 2jMfAAAAMAAJ , to multiple pages (and PAGE files).

Finally, the METS file contains the relationship between the URLs and the PAGE files in the <mets:structMap> <mets:structMap> section of the file:

```
<mets:structMap>
    <mets:div ID="map_001">
        <mets:fptr FILEID="gt_2jMfAAAAMAAJ_28"/>
        <mets:fptr FILEID="pdf_2jMfAAAAMAAJ_28"/>
    </mets:div>
    ...
</mets:structMap>
```

Fig. 4: Excerpt from the METS file as used in our data set. For each book, we created one METS file. Together with the links to the image resources shown in Figure 1, and the links to the PAGE files, the METS file holds the connection between the text lines and the page images. [Lassner et al. 2021]

In order to reuse the data set, a scholar may then obtain the original image resources from the respective institutions as PDFs, based on the links we provide in the METS files. Then, the pair data set can be created by running the ›make pair_output‹ command in the ›pipelines/‹ directory. For each title, it extracts the PNG images from the PDF, preprocesses them, extracts, crops and saves the line images along respective files containing the text of the line.

Although the image data needs to be downloaded manually, the data set can still be compiled within minutes.

4. Framework for creating, publishing and reusing OCR ground truth data

We have published the framework we developed for the second case study, which enables scholars to create and share their own ground truth data set formulas when they are in the same situation of not owning the copyright for the images they use. This framework offers both directions of functionality:

- Creating an XML ground truth data set from transcriptions to share it with the public (data set formula) and
- Compiling an XML ground truth data set into standard OCR ground truth data pairs to train an OCR model (built data set).[19]

As already described in the Sections 3.2 and 3.3 there are multiple steps involved in the creation, publication and reuse of the OCR data set. In this Section, we would like to show that our work is not only relevant for scholars who want to reuse our data set but also for scholars who would like to publish a novel OCR ground truth data set in a similar copyright setting.

4.1 Creation and Publication

1. Corpus construction: selection of the relevant books and pages
2. Creation of the METS files[20]
3. Transcription of the pages
4. Creation of the PAGE files[21]
5. Publication of the METS and the PAGE files

4.2 Reuse

1. Download of the METS and PAGE files
2. Download of the PDFs as found in the METS files
3. Creation of the pair data set[22]
4. Training of the OCR models[23]

In the Section 3.3, the steps listed in Reuse have been described. The download of the transcriptions and the PDFs has to be done manually but for the creation of the pair data set and the training of the models, automation is provided with our framework. We would like to also automatize the download of the PDFs; this, however, remains complicated to implement.

[19] The documentation how to create a new or reproduce an existing data set can be found at README.md, OCR-Data 2021.
[20] See mets_page_template.xml, OCR-Data 2021.
[21] See create_xml_files.py, OCR-Data 2021.
[22] See extract_pair_dataset.py, OCR-Data 2021.
[23] See train_ocr_model.py, OCR-Data 2021.

The first reason for this is a technical one: soon after starting the download, captchas appear (as early as by the 3rd image), which hinders the automatization. Another reason is the Google Books regulation itself. Page one of any Google Books PDF states explicitly:

> »Keine automatisierten Abfragen. Senden Sie keine automatisierten Abfragen irgendwelcher Art an das Google-System. Wenn Sie Recherchen über maschinelle Übersetzung, optische Zeichenerkennung oder andere Bereiche durchführen, in denen der Zugang zu Text in großen Mengen nützlich ist, wenden Sie sich bitte an uns. Wir fördern die Nutzung des öffentlich zugänglichen Materials für diese Zwecke und können Ihnen unter Umständen helfen.«[24]

Finding a way to automatize download could hence not be realized in the context of this project and will have to be addressed in future work.[25]

Additionally, we provide useful templates and automation for the creation of a novel OCR ground truth data set. As already described, we used the Kraken transcription interface to create the transcription. In Kraken, the final version of the transcription is stored in HTML files. We provide a script to convert the HTML transcriptions into PAGE files in order to facilitate interoperability with other OCR ground truth data sets.

Finally, the pair data set can be created from the PAGE transcriptions and the images of the PDFs and the OCR model can be trained.

5. Relevance of the data set

In order to evaluate the impact that the data set has on the accuracy of OCR models, we trained and tested model performance in three different settings. In the first setting, we fine-tuned an individual model for each book in our corpus using a training and an evaluation set of that book and tested the performance of the model on a held-out test set from the same book. In Table 2, we show how this data set has dramatically improved the OCR accuracy on similar documents compared to off-the-shelf OCR solutions. Especially in cases where the off-the-shelf model (baseline) shows a weak performance, the performance gained by fine-tuning is large.

In the second and third setting, we split the data set into two groups: English Antiqua, German Fraktur. There was also one German Antiqua book that we did not put into any of the two groups. For the second setting, we split all data within a group randomly into train set, evaluation set and test set and trained and tested an individual model for each group. In Table 3, the test performance of this setting is shown. For both groups, the fine-tuning improves

[24] When downloading any book PDF from Google Books one page is prepended to the document. On this page, the cited usage statement is presented. As an example, please consider Walter Scott's Werke, see Google Inc. 2006.
[25] Our progress on this topic will be documented in issue 2 of our github repository, see OCR-Data 2021.

the character accuracy by a large margin over the baseline accuracy. This experiment shows that overall, the fine-tuning within a group improves the performance of that group and that patterns are learned across individual books.

Google Books or Internet Archive identifier	baseline model	Train # lines	Test # lines	Train # chars	Test # chars	baseline character accuracy	fine-tuned character accuracy	δ
rDUJAAAAQAAJ	[illegible]best	82	11	3520	493	99.8	100.0	0.2
chroniclesofcan[illegible]	[illegible]best		3	836	97	100.0	100.0	0.0
anneofgeierstein[illegible]	[illegible]best		3	805	138	100.0	100.0	0.0
_QgOAAAAQAAJ	[illegible]best	60	8	2659	359	95.54	100.0	4.46
chroniclesofcan[illegible]	[illegible]best		5	1766	185	99.46	99.46	0.0
zviTtwEACAAJ	Fraktur_1_best		9	3396	519	98.27	99.23	0.96
quentindurward[illegible]	[illegible]best		5	1748	241	99.17	99.17	0.0
3pVMAAAAcAAJ	FraktUr_1_best		12	4830	598	96.49	99.16	2.67
2jMfAAAAMAAJ	FraktUr_1_best		20	7386	939	93.5	98.94	5.44
t88yAQAAMAAJ	FraktUr_1_best		11	3345	436	94.5	98.85	4.35
HCRMAAAAcAAJ	FraktUr_1_best		16	5100	579	92.23	98.79	6.56
zDTMtgEACAAJ	FraktUr_1_best		10	4277	560	93.93	98.75	4.82
DNUwAQAAMAAJ	FraktUr_1_best		10	4147	517	94.58	98.45	3.87
H9UwAQAAMAAJ	FraktUr_1_best		10	4017	533	97.19	98.31	1.12
AdiKyqdlp4cC	FraktUr_1_best		10	2827	405	92.84	98.27	5.43
J4knAAAAMAAJ	[illegible]best	20	3	851	104	97.12	98.08	0.96
aNQwAQAAMAAJ	FraktUr_1_best		7	2752	309	95.79	98.06	2.27
XtEyAQAAMAAJ	FraktUr_1_best		11	3489	383	94.52	97.91	3.39
D5pMAAAAcAAJ	FraktUr_1_best		12	4557	546	93.22	97.8	4.58
8AQoAAAAMAAJ	FraktUr_1_best		9	3130	434	94.93	97.7	2.77
Fy4JAAAAQAAJ	[illegible]best	20	3	743	125	96.0	97.6	1.6
anneofgeierstein[illegible]	[illegible]best		6	1747	204	98.04	97.55	-0.49
u4cnAAAAMAAJ	FraktUr_1_best		10	3936	553	91.5	97.11	5.61

Tab. 2: Performance comparison of baseline model and fine-tuned model for each document in our corpus. For almost all documents there is a large improvement over the baseline even with a very limited number of fine-tuning samples. The sum of lines and characters depicted in the table do not add up to the numbers reported in the text because during training we used an additional split of the data as an evaluation set that had the same size as the test set respectively. [Lassner et al. 2021]

1VUJAAAAQAAJ	en_best	85	11	3899	455	94.73	96.7	1.97
quentinduma00scou	en_best	20	3	708	86	95.35	95.35	0.0
4zQfAAAAMAAJ	fraktur_1_best	159	20	6817	932	87.98	94.74	6.76
7JVMAAAAcAAJ	fraktur_1_best	89	12	4604	616	65.91	94.32	28.41
YAZXAAAAcAAJ	fraktur_1_best	1752	219	66253	8327	80.17	93.61	13.44
8dAyAQAAMAAJ	fraktur_1_best	88	12	3448	380	87.11	93.42	6.31
PzMJAAAAQAAJ	en_best	61	8	2294	234	90.17	92.74	2.57
wggOAAAAQAAJ	en_best	19	3	716	94	91.49	92.55	1.06
WjMfAAAAMAAJ	fraktur_1_best	183	23	7363	814	71.62	91.52	19.9
MzQJAAAAQAAJ	en_best	36	5	1265	201	88.56	90.55	1.99
fAoOAAAAQAAJ	en_best	40	6	1675	121	86.78	87.6	0.82
kggOAAAAQAAJ	en_best	40	6	1572	243	82.72	82.72	0.0
oNEyAQAAMAAJ	fraktur_1_best	72	10	2874	386	68.39	79.02	10.63
htQwAQAAMAAJ	fraktur_1_best	78	10	3990	464	69.18	78.02	8.84

Tab. 2: Performance comparison of baseline model and fine-tuned model for each document in our corpus. For almost all documents there is a large improvement over the baseline even with a very limited number of fine-tuning samples. The sum of lines and characters depicted in the table do not add up to the numbers reported in the text because during training we used an additional split of the data as an evaluation set that had the same size as the test set respectively. [Lassner et al. 2021]

Document Group	baseline model	Train # lines	Test # lines	Train # chars	Test # chars	baseline character accuracy	fine-tuned character accuracy	δ
English Antiqua	en_best	650	82	26793	3406	94.19	96.21	2.02
German Fraktur	fraktur_1_best	3449	432	145928	17577	85.89	95.99	10.1

Tab. 3: Performance comparison of baseline model and fine-tuned model trained on a random splits of samples within the same group. [Lassner et al. 2021]

Left-out identifier	baseline model	Train # lines	Test # lines	Train # chars	Test # chars	baseline character accuracy	fine-tuned character accuracy	δ

Tab. 4: Model performance evaluated with a leave-one-out strategy. Within each group (German Fraktur and English Antiqua), an individual model is trained on all samples except from the left-out identifier on which the model is tested afterwards. The performance of the fine-tuned model is improved in each case, often by a large margin. [Lassner et al. 2021]

chroniclesofc_bec03_sc686		50	28134	2182	99.22	99.59	0.37
H9UwAQAAMAAJ_1_b3794		96	159088	5130	96.74	99.57	2.83
aNQwAQAAMAAJ_1_b3822		65	161053	3397	97.0	99.53	2.53
chroniclesofc_bec02_sc709		25	29226	1017	99.02	99.51	0.49
zDTMtgEACAAJ_1_b3794		96	159131	5430	95.05	99.43	4.38
anneofgeierstein_bec03_sc708		26	29144	1062	98.68	99.34	0.66
t88yAQAAMAAJ_1_b3786		105	160286	4181	91.13	99.28	8.15
anneofgeierstein_bec02_sc684		53	28053	2181	98.3	99.27	0.97
DNUwAQAAMAAJ_1_b3794		96	159113	5228	95.26	99.01	3.75
D5pMAAAAcAAJ_1_b3780		111	159386	5660	93.69	99.01	5.32
3pVMAAAAcAAJ_1_b3777		115	158561	6036	94.68	98.99	4.31
zviTtwEACAAJ_1_b3806		83	159741	4384	95.76	98.97	3.21
8AQoAAAAMAAJ_1_b3800		89	160966	3926	94.7	98.9	4.2
1VUJAAAAQAAJ_1_gtest	635	107	25735	4839	96.88	98.8	1.92
AdiKyqdlp4cC_1_b3793		97	160065	3736	92.34	98.47	6.13
rDUJAAAAQAAJ_1_gtest	639	103	26265	4419	97.85	98.42	0.57
quentindurwabec02_sc687_oft		49	28274	2223	97.35	98.34	0.99
HCRMAAAAcAAJ_1_b3739		157	158250	6378	91.28	98.28	7.0
J4knAAAAMAAJ_1_gtest	708	26	29219	1089	97.15	98.07	0.92
2jMfAAAAMAAJ_1_b3803		197	155342	9181	92.43	98.04	5.61
XtEyAQAAMAAJ_1_b3783		108	160349	4322	87.69	97.59	9.9
quentindurwabec01_sc708_oft		26	29284	940	96.38	97.13	0.75
wggOAAAAQAAJ_1_gtest	710	24	29362	869	92.52	96.89	4.37
_QgOAAAAQAAJ_1_gtest	664	75	27117	3320	94.43	96.66	2.23
fAoOAAAAQAAJ_1_gtest	685	51	28128	2007	94.72	96.61	1.89
4zQfAAAAMAAJ_1_b3801		199	156399	8681	88.68	96.37	7.69
PzMJAAAAQAAJ_1_gtest	662	77	27724	2817	90.7	95.49	4.79
u4cnAAAAMAAJ_1_b3795		95	159827	4889	91.31	95.21	3.9
7JVMAAAAcAAJ_1_b3780		112	159080	5816	71.35	94.62	23.27
8dAyAQAAMAAJ_1_b3780		111	159841	4271	84.45	94.24	9.79

Tab. 4: Model performance evaluated with a leave-one-out strategy. Within each group (German Fraktur and English Antiqua), an individual model is trained on all samples except from the left-out identifier on which the model is tested afterwards. The performance of the fine-tuned model is improved in each case, often by a large margin. [Lassner et al. 2021]

htQwAQAAMAAJ_1_b3592	fraktur_1_best	692	98	158623	4996	88.42	94.14	5.72
YAZXAAAAcAAJ_1_b3609	fraktur_1_best	699	2190	89328	82910	80.68	92.92	12.24
MzQJAAAAQAAJ_1_best	fraktur_1_best	691	45	28714	1622	84.9	89.52	4.62
kggOAAAAQAAJ_1_best	fraktur_1_best	685	51	28216	1983	85.64	87.56	1.92
Fy4JAAAAQAAJ_1_best	fraktur_1_best	709	25	29424	943	78.9	85.15	6.25
oNEyAQAAMAAJ_1_b3593	fraktur_1_best	698	92	160955	3589	66.31	84.79	18.48

Tab. 4: Model performance evaluated with a leave-one-out strategy. Within each group (German Fraktur and English Antiqua), an individual model is trained on all samples except from the left-out identifier on which the model is tested afterwards. The performance of the fine-tuned model is improved in each case, often by a large margin. [Lassner et al. 2021]

In the third setting, we trained multiple models within each group, always training on all books of that group except one and using only the data of the left-out book for testing. In all settings, we also report the performance of the off-the-shelf OCR model on the test set for comparison.

As depicted in Table 4, the performance of fine tuning improves character accuracy each time even for the held-out book. This shows that the fine-tuned model indeed did not overfit on a specific book but captures patterns of a specific script. We should note, that in some cases of the third experiment different volumes occur as individual samples, for example, the second volume of Anne of Geierstein by Scott was not held-out when tested for the third volume of Anne of Geierstein. Scripts in different volumes are often more similar than scripts of the same font type which might improve the outcome of this experiments in some cases.

For all three experiments, the Kraken OCR engine with a German Fraktur model and an English model was used as baselines. They were provided by the maintainers of Kraken.[26]

In the context of the research project for which this data set was created, the performance gain is especially relevant as research shows that a certain level of OCR quality is needed in order to be able to obtain meaningful results on downstream tasks. For example, Hamdi et al. show the importance of OCR quality on the performance of Named Entity Recognition as a downstream task.[27] With additional cross training of sub-corpora we are confident that we will be able to push the character accuracy beyond 95% on all test sets that will enable us to perform translatorship attribution analysis.

More generally, the results show that in a variety of settings, additional ground truth data will improve the OCR results. This advocates strongly for the publication of a greater range of, and especially more diverse, sets of open and reusable ground truth data for historical prints.

[26] See Kiessling 2019. For baselines and fine-tuning version 3.0.4 of the Kraken engine was used that can be found at kraken release 3.0.4, Kiessling 2021.
[27] See Hamdi et al. 2020.

The data set we thus created and published is open and reproducible following the described framework. It can serve as a template for other OCR ground truth data set projects. It is therefore not only relevant because it shows why the community should create additional data sets: it also shows how to create the data sets and invites to new publications bound to bring Digital Humanities research a step forward.

The data pairs are compatible with other OCR ground truth data sets such as e. g. OCR-D[28] or GT4HistOCR[29]. Using the established PAGE-XML standard enables interoperability and reusability of the transcriptions. Using open licenses for the source code and the data, and publishing releases at an institutional open data repository ensures representativeness and durability.

6. Conclusion

The work we realized in order to constitute the data set we need for our stylometric research provided not only a ground truth data set, but also a systematic approach to the legal issues we encountered in the extraction of information from the scanned books we rely on as a primary source. While we have been successful at automating many work steps, improvements could still be envisioned.

In future work, we would like to enrich the links to the original resource with additional links to mirrors of the resources in order to increase the persistence of the image sources, whenever available also adding OCLC IDs as universal identifiers.[30] We would also like to look into ways to automate the download of the PDFs from Google Books, the Internet Archive or CHIs. Also, we would like to extend the framework we proposed here. It could serve for hybrid data sets with parts where the copyright for the image data is unclear (then published as data set formula), and others with approved image redistribution (which could then be published as a built data set). It could be used for example for the datasets from Bayerische Staatsbibliothek and University of Illinois Urbana-Champaign.

Finally, we would like to encourage scholars to publish their OCR ground truth data set in a similarly open and interoperable manner, thus making it possible to ultimately increase accessibility to archives and libraries for everyone.

Acknowledgements

This work has been supported by the German Federal Ministry for Education and Research as BIFOLD.

[28] See Baierer et al. 2019.
[29] See Springmann et al. 2018.
[30] OCLC is a registry of IDs referencing items in libraries, see worldcat.org, OCLC 2021.

List of contracts

The contracts between

- a number of US-based libraries and Google is available here,
- the British Library and Google is available here,
- the National Library of the Netherlands and Google is available here,
- the University of Michigan and Google is available here,
- the University of Texas at Austin and Google is available here,
- the University of Virginia and Google is available here,
- Scanning Solutions (for the Bibliotheque Municipale de Lyon) and Google is available here,
- University of California and Google is available here.

Bibliographic references

Norbert Bachleitner: »Übersetzungsfabriken«: das deutsche Übersetzungswesen in der ersten Hälfte des 19. Jahrhunderts. In: Internationales Archiv für Sozialgeschichte der deutschen Literatur 14 (1989), i. 1, pp. 1–50. [Nachweis im GBV]

Anne Baillot / Mike Mertens / Laurent Romary: Data fluidity in DARIAH – pushing the agenda forward. In: Bibliothek Forschung und Praxis 39 (2016), i. 3, pp. 350–357. DOI: 10.1515/bfp-2016-0039 [Nachweis im GBV]

Konstantin Baierer / Matthias Boenig / Clemens Neudecker: Labelling OCR Ground Truth for Usage in Repositories. In: Proceedings of the International Conference on Digital Access to Textual Cultural Heritage (DATeCH2019: 3, Brussels, 08.–10.05.2019) New York, NY 2019, pp. 3–8. [Nachweis im GBV]

HTR-United. In: GitHub.io. By Alix Chagué / Thibault Clérice. 2021. [online]

Marian Ramos Eclevia / John Christopher La Torre Fredeluces / Carlos Jr Lagrosas Eclevia / Roselle Saguibo Maestro: What Makes a Data Librarian? An Analysis of Job Descriptions and Specifications for Data Librarian. In: Qualitative and Quantitative Methods in Libraries 8 (2019), n. 3, pp. 273–290. [online]

Elisabeth Engl: Volltexte für die Frühe Neuzeit. Der Beitrag des OCR-D-Projekts zur Volltexterkennung frühneuzeitlicher Drucke. In: Zeitschrift für Historische Forschung 2 (2020), n. 47, pp. 223–250. [Nachweis im GBV]

Transcriptiones. A platform for hosting, accessing and sharing transcripts of non-digitised historical manuscripts. Ed. by ETH-Library. Zürich 2020. [online]

Ahmed Hamdi / Axel Jean-Caurant / Nicolas Sidère / Mickaël Coustaty: Assessing and Minimizing the Impact of OCR Quality on Named Entity Recognition. In: Digital libraries for open knowledge. International Conference on Theory and Practice of Digital Libraries. (TPDL: 24, Lyon, 25.–27.08.2020) Cham 2020, pp. 87–101. [Nachweis im GBV]

The Heritage Data Reuse Charter. In: DARIAH.eu. 2021. [online]

Informationen und Richtlinien. Ed. by Google Inc. In: Google Books. Walter Scott: Großvater's Ezählungen aus der Geschichte von Frankreich. Ed. by Georg Nicolaus Bärmann. Neue Folge. Zweiter Theil. Zwickau 1831. Digitalisiert am 15.11.2006. PDF. [online]

Benjamin Kiessling: Kraken – an Universal Text Recognizer for the Humanities. In: Digital Humanities 2019 Conference papers. (DH2019, Utrecht, 08.–12.07.2019) Utrecht 2019. [online]

Kraken 3.0.4. In: GitHub.io. Ed. by Benjamin Kiessling. 2021. [online]

David Lassner / Julius Coburger / Clemens Neudecker / Anne Baillot: Data set of the paper »Publishing an OCR ground truth data set for reuse in an unclear copyright setting«. In: zenodo.org. 2021. Version 1.1 from 07.05.2021. DOI: 10.5281/zenodo.4742068

METS. Metadata Encoding & Transmission Standard. Home. Ed. by The Library of Congress. Washington D.C. 04.10.2021. [online]

Bernhard Liebl / Manuel Burghardt: From Historical Newspapers to Machine-Readable Data: The Origami OCR Pipeline. In: Proceedings of the Workshop on Computational Humanities Research. Ed. by Folgert Karsdorp / Barbara McGillivray / Adina Nerghes / Melvin Wevers. (CHR2020, Amsterdam, 18.–20.11.2020), Aachen 2020, pp. 351–373. (= CEUR Workshop Proceedings, 2723) URN: urn:nbn:de:0074-2723-3

OCR-Data. In: GitHub.io. 2021. [online]

OCR-D. Specifications. In: OCR-D.de. Wolfenbüttel 2021. [online]

OCR/HTR model repository. In: Zenodo.org. 2021. [online]

WorldCat. Ed. by OCLC. Dublin 2021. [online]

Thomas Padilla / Laurie Allen / Hannah Frost / Sarah Potvin / Elizabeth Russey Roke / Stewart Varner: Final Report – Always Already Computational: Collections as Data. In. zenodo.org. Version 1 from 22.05.2019. DOI: 10.5281/zenodo.3152935

Stefan Pletschacher / Apostolos Antonacopoulos: The PAGE (Page Analysis and Ground-Truth Elements) Format Framework. In: Proceedings of the 20th International Conference on Pattern Recognition. Ed. by IEEE. (ICPR: 20, Istanbul, 23.–26.08.2010) Piscataway, NJ 2010, vol. 1, pp. 257–260. [Nachweis im GBV]

Public AI models in Transkribus. Ed. bsy READ-COOP. Innsbruck 2021. [online]

Christian Reul / Christoph Wick / Uwe Springmann / Frank Puppe: Transfer Learning for OCRopus Model Training on Early Printed Books. In: Zeitschrift für Bibliothekskultur 5 (2017), i. 1, pp. 32–45. In: zenodo.org. Version 1 from 22.12.2017. DOI: 10.5281/zenodo.4705364

Javier Ruiz: Access to the Agreement between Google Books and the British Library. In: Open Rights Group. Ed. by The Society of Authors. Blogpost from 24.08.2011. [online]

Christof Schöch / Frédéric Döhl / Achim rettinger / Evely Gius / Peer Trilcke / Peter Leinen / Fotis Jannidis / Maria Hinzmann / Jörg Röpke: Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. In: Zeitschrift für digitale Geisteswissenschaften 5 (2020). DOI: 10.17175/2020_006

Uwe Springmann / Christian Reul / Stefanie Dipper / Johannes Balter: Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin. In: The Journal for Language Technology and Computational Linguistics 33 (2018), i. 1, pp. 97–114. PDF. [online]

List of Figures with Captions