General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example

TOBIAS HODEL (iD)

DAVID SCHOCH (iD)

CHRISTA SCHNEIDER (iD)

JAKE PURCELL (iD)

Author affiliations can be found in the back matter of this article

ABSTRACT

Existing text recognition engines enables to train general models to recognize not only one specific hand but a multitude of historical hands within a particular script, and from a rather large time period (more than 100 years). This paper compares different text recognition engines and their performance on a test set independent of the training and validation sets. We argue that both, test set and ground truth, should be made available by researchers as part of a shared task to allow for the comparison of engines. This will give insight into the range of possible options for institutions in need of recognition models. As a test set, we provide a data set consisting of 2,426 lines which have been randomly selected from meeting minutes of the Swiss Federal Council from 1848 to 1903. To our knowledge, neither the aforementioned text lines, which we take as ground truth, nor the multitude of different hands within this corpus have ever been used to train handwritten text recognition models. In addition, the data set used is perfect for making comparisons involving recognition engines and large training sets due to its variability and the time frame it spans. Consequently, this paper argues that both the tested engines, HTR+ and PyLaia, can handle large training sets. The resulting models have yielded very good results on a test set consisting of unknown but stylistically similar hands.

CORRESPONDING AUTHOR:
Tobias Hodel

Digital Humanities, Walter Benjamin Kolleg, University of Bern, Switzerland

tobias.hodel@wbkolleg.unibe.ch

1 THE QUEST FOR TEXT RECOGNITION

Since the early 1990s, recognition of printed text has been based on engines for optical character recognition (OCR) (Rice et al., 1993). The results have been perfected over the last fifteen years leading to satisfying results for printed material, even for printed blackletter (Neudecker et al., 2012). However, the advent of offline handwritten text recognition (HTR) lagged behind print for several decades. It was only with the implementation of deep learning, especially the cell-based neural network architecture called long-short term memory (LSTM) in the early 2010s, that handwritten text recognition achieved a quality that made such recognition processes feasible in the humanities (Graves & Schmidhuber, 2009; Leifert et al., 2016).

This paper reports on the state-of-the-art in text recognition. Its primary focus is on general models which train recognition models that are capable of recognizing not just one specific hand but similar scripts from different hands that the model has not previously seen which is one of the remaining problems in handwritten text recognition. To build such models, it is necessary to bring together large masses of ground truthed data (transcribed text aligned with images). Simultaneously, available recognition engines need to be assessed based on their ability to produce efficient and powerful general models. The paper is consequently providing independent test sets for assessing the capability of general models.

The output of handwritten text recognition models usually leads to further processable results if the Character Error Rate (CER) is below 10%. Text generated with ten or fewer mistakes per hundred characters is generally legible and allows for skimming as well as efficient post-processing if necessary. The rate is derived from experience in manual correction processes in projects such as ANNO[1] (Muehlberger et al., 2014). We thus consider a CER of ≤10% to be good. A further improvement to ≤5% CER is considered very good, as the occurring errors can be narrowed down to rare/unknown words (Muehlberger et al., 2019, p. 965). We could even invoke a third category by stating that recognition models with a CER below 2.5% reach a certain level of "excellence". However, within the range below 3% CER, experience shows that that the regularity of scripts starts to play an important role and needs to be considered. Furthermore, some irregular hands cannot be recognized with a CER below 3% with existing engines, irrespective of the amount of available training material.

As a second step, it is necessary to assess the quality of the models not solely on validation sets containing the hands of the training set which have already been seen, but also on a test set consisting of similar hands of the same era and written in the same script type. For this purpose, we propose a test set for German Kurrent scripts of the 19th century, because enough training material is already available to test the capabilities of text recognition engines for this script. Of course, this set can only be one example among the wide variety of existing historical scripts. Consequently, we understand the production of ground truth, the model training, and the meaningful evaluation through test sets as shared tasks. Only through cooperation of large communities of stakeholders, including scholars, scientists, and the interested public, we will be able to make the handwritten material of the world better and easier accessible.

1.1 RECOGNIZING HANDWRITING: A RESOLVED TASK

From a computer science point of view, the recognition of handwriting seems to be a resolved task. The latest recognition engines allow for the successful recognition of specifically trained hands producing a text as reusable data. For the past decade, competitions organized around handwritten text recognition have regularly been conducted at the two most relevant conferences, the International Conference on Document Analysis and Recognition (ICDAR) and the International Conference on Frontiers in Handwriting Recognition (ICFHR). Since 2019, the emphasis on competition has decreased or become focused on thus-far sparsely researched languages (like Arabic) or specific topics (like mathematical formulas).[2] We, therefore, assume that the results achieved so far are, on the one hand, in a mature state and, on the other, currently not at the center of large research projects. The current state-of-the art will be presented briefly in 2.1 according to which a reasonable recognition rate for alphabet scripts can be achieved provided that we have enough training material at our disposal. The basis

---

1   See *https://anno.onb.ac.at/*.

2   See for ICDAR 2019: *http://icdar2019.org/competitions-2/* and for ICFHR 2020: *http://icfhr2020.tu-dortmund.de/competitions/*.

for these recent improvements is the high level of investment by the European Union, e.g., in the Recognition and Enrichment of Archival Documents project (READ), leading to the virtual research environment Transkribus (Muehlberger et al., 2019), and by national funding agencies, e.g., in eScriptorium (INRIA, 2021; Stökl Ben Ezra, 2019).

We are currently noticing an increase in regular scholarly use of platforms offering automatic transcriptions and the implementation of their project workflows. However, it is currently still problematic to predict results when applying models to unknown material, especially written by unknown or several hands. This problem arises in HTR models for one of two reasons. Some HTR models focus on one specific hand which makes prediction of results impossible for other hands since the validation set of a model consists of the same hand. The same problem arises if a multitude of hands has been part of the training: the validation sets will consist of the same hands, splitting up ground truth in the training and validation set. With more models becoming publicly available, it is even more difficult to judge which one to use. The same challenge applies to recognition engines: although it is possible to compare results from specific models, the capability of recognition engines to train large amounts of ground truth have not yet been evaluated by researchers. The lack of evaluation becomes highly problematic in real-world scenarios which sometimes need large or even general models.

1.2 REAL-WORLD SITUATION: NECESSITY FOR RECOGNITION WITH NO TRAINING

Although the recognition of handwriting is in theory trainable for any given document, we encounter the problem more often when there is just a tiny sample of a single hand available at a given institution. The training of specific handwritten text recognition models will be relatively ineffective in these cases. Due to the need to prepare and process training with sufficient material, the amount of required training material often surpasses the amount of available material. Therefore, it is not only desirable but even necessary to provide recognition models with hands of the same style and from similar periods. To achieve this goal and train capable models, it is necessary to accumulate large amounts of ground truth and design test sets that do not consist of data that has been part of training or validation sets. Only with such test sets it becomes feasible to assess the potential of (large) ground truth sets and recognition engines.

2 TOWARDS GENERAL MODELS OF RECOGNITION

As a preliminary to the description of our research strategy, we will briefly discuss the current state-of-the-art of the data model (digital format) used to prepare ground truth. We also report on models explicitly trained for one hand to indicate necessary ground truth needed to train such models (2.1). From here, we present our test set (2.2) and the results achieved using large ground truth sets trained on the available recognition engines (2.3).

One format well suited for both human- and machine-readable text is PageXML (Pletschacher & Antonacopoulos, 2010). This format brings together image linking through pixel-based layout information that points to polygons on the level of text regions and lines (see *Figure 1*). The polygons wrap around the handwritten parts, analogous to a silhouette. In addition, lines are created at the level of what are called baselines, that is, the imaginary line on which the text gets written, only crossed by descenders. Both text regions and lines within text regions are determined by reading order.

From a digital editing perspective, PageXML understands text as a set of visual signs that make up a document (Hodel, 2018; Sahle, 2013). In some sense, the goal of annotating material in PageXML is to create a documentary edition that imitates the scanned image. For the reason that it is flexible and can be adapted to describe almost any type of text layout, this format is the preferred one for many projects dealing with automatic text recognition, e.g., OCR-D (Boenig et al., 2018) and the iurisprudentia project.[3] In PageXML, the text is understood hierarchically, starting with the whole page, followed by text regions, and finally by lines. Furthermore, stand-off annotation at line level is available for indicating abbreviations or underlining and tagging named entities. Thus, the format is comparable to ALTO XML (Library of Congress, 2016). The transformation between the two formats usually runs flawlessly.

---

3   See *https://rwi.app/iurisprudentia/de*.
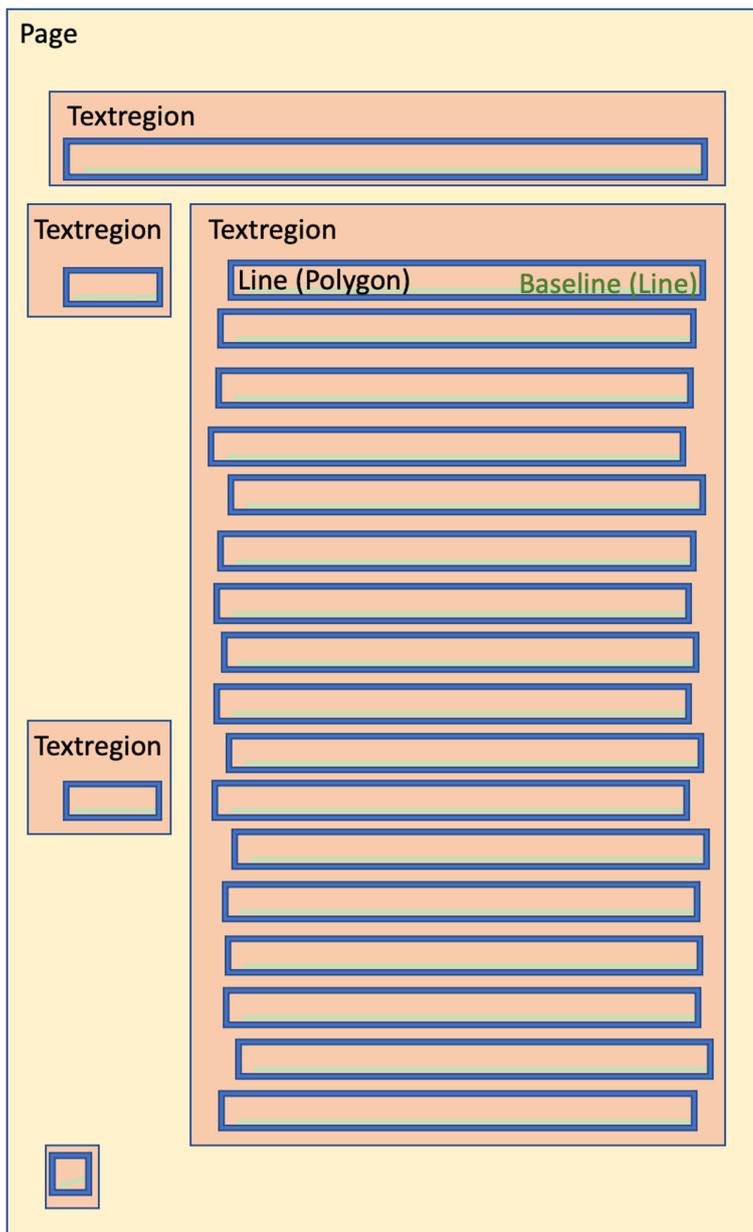
2.1 SPECIFIC MODELS

Current state-of-the-art engines and platforms propose to train specific models to recognize one hand or a set of very similar hands based on some tens of thousands of tokens. A token is a single, individual instance of a word. A ground truth, often manually produced, is split for this purpose into training and validation sets.[4] For projects focusing on one hand or a handful of different scribes, this approach is perfectly suitable. It leads to an increase in productivity when the desired outcome is a "clean" documentary transcription and when several hundreds of pages need to be processed.

The recognition results are similar and produce Character Error Rates (CER) in the range of 2.5–8% in best cases. Comparing PyLaia (Puigcerver & Mocholí, 2018) with HTR+ (Leifert et al., 2016) shows minor differences that might depend on the randomly chosen validation pages. HTR+ usually has a slight upper hand in regards to CER.[5]

We demonstrated that different engines already exist that can train HTR models for successful recognition. All the tasks mentioned above remain, and we need to confirm whether more significant amounts of ground truth will lead to good results for recognizing unseen hands.

---

4   See for example the how-to-guide "How to Train PyLaia-Models in Transkribus" online at *https://readcoop. eu/transkribus/howto/how-to-train-pylaia-models-in-transkribus/*.

5   In OCR, Word Error Rate (WER) is usually preferred over CER. Within project READ, the concept of what a word is (in order to determine the usefulness of WER) has been internally discussed without any conclusive results. We thus prefer the measure of CER.

2.2 ENGINES AND TESTING THE ASSUMPTION OF LARGE GROUND TRUTH COLLECTIONS

The goal of recognizing more than a given set of hands has been imagined by historians for quite some time (Wettlaufer, 2016). But the groundwork for such an endeavor was laid only two or three years ago (Michael et al., 2018), and only due to the availability of large ground truth sets and the immense interest in specific time periods and script types (resulting from investments in digitization by archives and libraries). Especially for documents from the 19th century German-speaking parts of the world, enough material has been made public and formed the basis for this kind of evaluation such as Edition Humboldt.[6] We are sure that in the near future similar evaluation baselines can be made available for other scripts and time periods. Through READ, applied engines have become available, and masses of digital images of scripts alongside aligned text in PageXML have been prepared. At the State Archives of Zürich, among others, more than 100,000 images have been matched with manual transcriptions leading to an enormous ground truth set for handwritten text recognition. Subsets of the data have been used to train specific models for periods (not for single hands), models which have achieved very good results considering that several hands were part of the training and validation sets.

In 2019, the computing power of the Transkribus platform at the University of Innsbruck (now READCOOP)[7] was enhanced and the necessary storage made available. For the first time, large models based on more than 100,000 tokens were trainable within short timeframes. At the same time, the HTR neural network architecture was replaced by HTR+, both developed by the CITlab group at the University of Rostock (Michael et al., 2018) leading to improved preliminary results from smaller training sets (see *Table 1*). Consequently, the first large models were trained based on more than 140,000 tokens leading to CER rates with regards to training and validation sets similar to small models (see *Table 2* and compare to *Table 1*).

WRITING STYLE	HANDS	TOKENS	ENGINE	% CER VAL.	% CER TRAIN.
Early Modern Kurrent	1	48,277	HTR+	2.87	1.11
			PyLaia	4.2	4.3
Medieval Charter	3/4	77,353	HTR+	5.44	2.64
			PyLaia	7.80	12.30

Table 1 Results of HTR engines based on small training sets compared with a validation set of known hands.

WRITING STYLE	HANDS	TOKENS	ENGINE	% CER VAL.	% CER TRAIN.
German Kurrent 19th century (State Archives Zürich)	~12	147,608	HTR+	2.55	3.12
			PyLaia	3.31	2.90
German Kurrent 19th century (large)	unknown	26,026,908	HTR+	1.73	3.41

Table 2 Results of HTR engines based on large training sets comparing results on training set and validation set consisting of a multitude of identical hands (same hands are included in training and validation set).

The large training set included several hands (the exact number was uncertain), but the model was still reaching a CER comparable to that of smaller training sets. Thus, we could demonstrate that the engines were capable of training large amounts of data resulting in usable models. We can conclude that, given enough material, it is possible to unite different hands in one model. This first conclusion leads us to the question of whether trained models could also recognize unseen hands in similar writing. Or, put simply, if general models for specific scripts are possible with existing engines.

2.3 GENERALIZING MODELS: BUILDING GROUND TRUTH

When assembling the above-mentioned large models, it turned out that they were somewhat capable of recognizing similar hands but did not result in a stable recognition of those types. The model trained on the material of the State Archives in Zürich was based on many pages, but they were all written by a relatively small number of scribes, all with comparable training. The

---

6   Online, access to transcriptions is provided at *http://edition-humboldt.de/api/v1.1/tei-xml.xql*.

7   For more information see *https://readcoop.eu/*.

result was a specialization of the model (in machine learning terms, an "overfitting"). Although the recognition of the known hands tested better, the quality of the recognition for similar but not identical hands was reduced. In contrast to specific models, we were forced to conclude that too much training material (of the same kind) could also be used in model training. Consequently, we started to assemble more training data from similar but not identical hands to balance the training (and validation) set for a specific style of handwriting (German Kurrent in our case). The model that led to the best recognition results on our specific test set (see *Table 3*) is thus built on material from a wide variety of documents, that is, the State Archives of Zürich (Regierungsratsprotokolle), the Passau Diocesan Archives (Birth, Death, and Marriage Records), lecture notes from lectures given by Alexander von Humboldt (Vorlesungsmitschriften zu Vorlesungen Alexander von Humboldts), letters by Swiss law professor Eugen Huber as well as a variety of small data sets of texts written in German Kurrent which are not named or published here due to the pending publication of editions and/or copyright ownership with regards to the images.[8]

The trained model is based on 5,100,439 tokens and reaches a CER of 6.53% against a validation set of known hands. The training set reached a CER of 4.40%, indicating that the neural network did not overfit.[9] The number of hands brought together in the model must be in the hundreds, and, despite this variety, we conclude that its performance was strong. Still, all this does not yield a conclusive articulation of quantitative and qualitative capabilities of such model with regards to the recognition of unseen hands of the same period. Due to this shortcoming, we created a test set independent of the training material.

3 COMPARING HTR ENGINES AND DISCUSSION OF RESULTS

3.1 CREATION OF SPECIFIC TEST SETS: THE MINUTES OF THE FEDERAL COUNCIL IN SWITZERLAND

In the context of creating large HTR models with the assumed capacity to recognize a variety of hands from a particular script, we tried to identify suitable approaches to measure whether a model is sufficiently "generalized". As is typical for research in machine learning, we wanted to test the generalization by application of a test set that:

- did not contain seen materials (in our case, "hands");

- consisted of a variety of hands;

- spanned more than a decade of writing;

- was large enough to be split randomly to avoid bias of any kind;

- could be published so that other models can be tested against it.

Thanks to an ongoing partnership with the Swiss Federal Archives and to previous digitization efforts, we had access to the meeting minutes of the Federal Council (see *Figure 2*), written by hand between 1848 and 1903 mostly in German (Swiss German was never used as written language in the administration).[10]
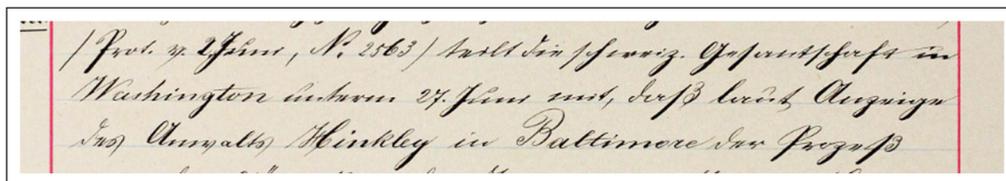
Figure 2 Visual impressions of the test set. Transcription of the sample line (middle line): *Washington unterm 27. Juni mit, daß laut Anzeige.*

---

8   Due to the varying provenance and connected copyright issues, it was unfortunately neither possible to publish the training nor the validation set as a whole.

9   Ideally, the result of the recognition process of the training set is slightly worse than the result on the validation set. This is because the set may get hindered from overfitting through added noise (visual and statistical) in the process of training. Due to the high capacity of HTR+, overfitting processes can be still observed in many cases. The reported result on the large training set is nearly ideal from our point of view (Hodel, 2020).

10   The untranscribed data set (all handwritten minutes) is available at *https://www.amtsdruckschriften.bar. admin.ch/*. The automatic recognition of all 150,000 pages is currently in preparation and the data set will be published as searchable website in due time.

From this set of approximately 150,000 pages, we split the volumes into twenty-two packages. For each package, 200 lines were randomly selected and transcribed using the "sample set function" provided by Transkribus and implemented by the CITlab group.[11] The minutes were written in German, French, and Italian. Since the goal was to assess the quality of German Kurrent recognition models, text parts in French, Italian and Latinized German (starting around 1900, some scribes switched to a Latin script) were deleted. As a result, we got a test set of 2,426 lines from a period spanning more than fifty years. The data set is available in Hodel & Schoch (2021a) at *https://doi.org/10.5281/zenodo.4746341*.

3.2 ASSESSING MODELS BASED ON THE TRAINING SET

For the provided test set, we ran recognition jobs using a variety of models trained on PyLaia and HTR+ engines and measured the CER. We basically used all available models specifically aimed at German Kurrent scripts of the 19th century provided through Transkribus. Two models were trained by the authors and one by Günter Mühlberger (Transkribus German Kurrent M2). The "Transkribus-Model" was trained on an unknown training set to recognize German Kurrent in general. As a fourth model, the lines were recognized using a model called "RRB" (short for *Regierungsratsbeschlüsse*), built solely on the extensive data set from the State Archives of Zürich. As a result, we can compare in *Table 3* the capability of both, the models (against a set of unknow hands) and the engines, reducing the effect of overfitting towards specific hands. Especially with the last model, we encounter the effect of "too much" training material on general models, a type of overfitting. The model is too specialized to be of similar quality to the "broader" models that were trained on a broader variety of scripts. The Transkribus Kurrent model turned out to be too wide since it also includes material from the 17th and 18th centuries. All recognition results have been provided as combined data set. The result of the recognition is available as data set Hodel & Schoch (2021b) at *https://doi.org/10.5281/zenodo.4905560*.

HTR MODEL	HTR ENGINE	CER MEAN %	CER MEDIAN %	CER UPPER BOUND (WORST)
German Kurrent M2	HTR+	3.43	2.76	9.13
	PyLaia	18.77	13.30	51.05
Transkribus German Kurrent	HTR+	5.90	4.85	10.20
RRB	HTR+	9.15	8.13	16.28

Table 3 Comparing different large HTR models and engines, applying the introduced test set, independent of already known hands.

The experiment demonstrates that existing engines can provide recognition models that lead to good or even very good results. Simultaneously, we can conclude that there are different approaches possible from very diverse to unimaginably large and uniform data sets, provided that enough training material is available.

As an indicator, we also provide information about the worst possible result of a recognition process. When dealing with samples, a statistical possible upwards deflection (considering the interval with a 95% probability) can be calculated, in addition to an average. The "CER upper bound (worst)" thus indicates, with a 95% probability, the worst possible recognition rate of the lines. Of course, this number exaggerates the results negatively since only one of the 22 sample sets yielded this value.

3.3 PUBLICATION OF GROUND TRUTH

As we have shown, well prepared material is key to producing general recognition models. It is unthinkable that single scholars and small project teams could provide enough training material to train a general model independently. Thus, it is of utmost importance to make ground truth, when possible, available to the public in a format that allows for reuse. In addition to PageXML, ALTO XML is a valid alternative.

Alix Chaqué and Thibault Clérice have provided an excellent example of this approach by combining a multitude of French documents as ground truth on GitHub as a collection called

---

11   See as an example of using sample sets in Heigl (2020).

"HTR united" (Chaqué & Clérice, 2020/2021). Similar publication of data is going to be necessary, even to cover only a basic number of scripts and periods.

4 GROUND TRUTH AND TEST SET CREATION AS A SHARED TASK: IMPLICATIONS

In this paper, we demonstrate that the evaluation of text recognition engines or recognition models for handwritten documents is a highly complex task. It needs to take into account not only the production and provision of training material, but also ways to grasp the potential of models and the capabilities of engines. The presented data set, which is a test set for handwriting in German Kurrent for the second half of the 19th century, offers only a glimpse of what is necessary to be able to determine confidently the potential of handwritten text recognition. The preparation of ground truth for training, validation, and testing should not be based on data from just one provider, but instead assembled by a (large) group of stakeholders: Large ground truthed sets need input from archives and libraries (images), transcriptions from scholars, and preparation, usually carried out by digital humanities specialists, according to standards like PageXML.

In a sense, we should understand the process of training and evaluation as a shared task (Reiter et al., 2019). However, this sharing is not just about competing to deliver the best possible results. We instead need to consider that it will only be possible to make our textual cultural heritage accessible at scale if we combine our forces and strengths to provide training and test material.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

- Tobias Hodel: Conceptualization, Formal Analysis, Funding acquisition, Investigation, Methodology, Writing – original draft, Writing – review & editing, Supervision.

- David Schoch: Data curation, Writing – review & editing, Investigation, Validation.

- Christa Schneider: Writing – review & editing.

- Jake Purcell: Writing – review & editing.

AUTHOR AFFILIATIONS

Tobias Hodel *orcid.org/0000-0002-2071-6407*
Digital Humanities, Walter Benjamin Kolleg, University of Bern, Switzerland

David Schoch *orcid.org/0000-0002-9936-8459*
Digital Humanities, Walter Benjamin Kolleg, University of Bern, Switzerland

Christa Schneider *orcid.org/0000-0001-9741-0601*
Digital Humanities, Walter Benjamin Kolleg, University of Bern, Switzerland

Jake Purcell *orcid.org/0000-0002-7636-5669*
Digital Humanities, Walter Benjamin Kolleg, University of Bern, Switzerland

REFERENCES

Boenig, M., Federbusch, M., Herrmann, E., Neudecker, C., & **Würzner, K.-M.** (2018). Ground Truth: Grundwahrheit oder Ad-Hoc-Lösung? Wo stehen die Digital Humanities? In G. Vogeler (Ed.), *DHd 2018. Kritik der digitalen Vernunft Konferenzabstracts. Universität zu Köln 26. Februar bis 2. März 2018* (pp. 219–223). Retrieved from *http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf#page=221*

Chaqué, A., & **Clérice, T.** (2021). *HTR-United*. GitHub. Retrieved from *https://github.com/HTR-United/htr-united*

Graves, A., & **Schmidhuber, J.** (2009). Offline handwriting recognition with multidimensional recurrent neural networks. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21* (pp. 545–552). Red Hook, NY: Curran Associates. Retrieved from *http://papers.nips.cc/paper/3449-offline-handwriting-recognition-with-multidimensional-recurrent-neural-networks.pdf*

Heigl, E. (2020). "Testsamples – die unparteiische Alternative." *Rechtsetzung im Ostseeraum* (blog), 10 March 2020. Retrieved from *https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/de/test-samples-the-impartial-alternative/*

Hodel, T. (2018). Konsequenzen automatischer Texterkennung – Ein Aufriss zur Texterkennung mit Machine Learning. In G. Vogeler (Ed.), *DHd 2018. Kritik der digitalen Vernunft Konferenzabstracts. Universität zu Köln 26. Februar bis 2. März 2018* (pp. 249–251). Retrieved from *http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf#page=251*

Hodel, T. (2020). Best-practices zur Erkennung alter Drucke und Handschriften – Die Nutzung von Transkribus large- und small-scale. In C. Schöch (Ed.), *DHd 2020. Spielräume Digital Humanities zwischen Modellierung und Interpretation. Abstracts zur 7. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. in Paderborn* (pp. 84–87). DOI: *https://doi.org/10.5281/zenodo.3666689*

Hodel, T., & **Schoch, D.** (2021a). Handwritten Text Recognition Test Set: Minutes of the Swiss Federal Council (1848–1903) Data set Version 1.0. *Zenodo*. DOI: *http://doi.org/10.5281/zenodo.4746342*

Hodel, T., & **Schoch, D.** (2021b). Recognition Results for the Handwritten Text Recognition Test Set: Minutes of the Swiss Federal Council (1848–1903) Data set Version 1.0. *Zenodo*. DOI: *http://doi.org/10.5281/zenodo.4905561*

Leifert, G., Strauss, T., Grüning, T., Wustlich, W., & **Labahn, R.** (2016). Cells in multidimensional recurrent neural networks. *Journal of Machine Learning Research, 17*(97), 1–37. Retrieved from *http://jmlr.org/papers/v17/14-203.html*

Library of Congress. (2016). *ALTO: Technical Metadata for Layout and Text Objects.* Retrieved from *https://www.loc.gov/standards/alto/*

Michael, J., Weidemann, M., & **Labahn, R.** (2018). *HTR engine based on neural networks P3. Optimizing speed and performance – HTR+.* Deliverable 7.9 submitted to the European Commission as part of the Recognition & Enrichment of Archival Documents (READ) project funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943. Retrieved from *https://read.transkribus.eu/wp-content/uploads/2018/12/Del_D7_9.pdf*

Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grüning, T., Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., Kahle, P., Kallio, M., Kaplan, F., Kleber, F., Labahn, R., Lang, E-M., Laube, S., Leifert, G., Louloudis, G., McNicholl, R., Meunier, J.-L., Michael, J., Mühlbauer, E., Philipp, N., Pratikakis, I., Puigcerver Pérez, J., Putz, H., Retsinas, G., Romero, V., Sablatnig, R., Sánchez, J.-A., Schofield, P., Sfikas, G., Sieber, C., Stamatopoulos, N., Strauss, T., Terbul, T., Toselli, A.-H., Ulreich, B., Villegas, M., Vidal, E., Walcher, J., Weidemann, M., Wurster, H., Zagoris, K. (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation, 75*(5), 954–976. DOI: *https://doi.org/10.1108/JD-07-2018-0114*

Muehlberger, G., Zelger, J., Sagmeister, D. (2014). User-driven correction of OCR errors: combining crowdsourcing and information retrieval technology. In Antonacopoulos, A. & Schulz, K. U. (Eds.), *DATeCH'14: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, Madrid, Spain, 19–20 May 2014 (pp. 53–56). New York, NY: Association for Computing Machinery. DOI: *https://doi.org/10.1145/2595188.2595212*

National Institute for Research in Digital Science and Technology (INRIA). (2021). *eScriptorium*. GitLab. Retrieved from *https://gitlab.inria.fr/scripta/escriptorium*

Neudecker, C., Schlarb, S., Neumann, D., & **Dogan, M.** (2012). *IMPACT: Improving Access to Text: Final Report*. Retrieved from *http://www.impact-project.eu/uploads/media/IMPACT_D-OC5.4_Final_Report.pdf*

Pletschacher, S., & **Antonacopoulos, A.** (2010). The PAGE (Page Analysis and Ground-Truth Elements) Format Framework. In *Proceedings of the 20th International Conference on Pattern Recognition*, Istanbul, Turkey, 23–26 August 2010 (pp. 257–260). Los Alamitos, CA: IEEE Computer Society. DOI: *https://doi.org/10.1109/ICPR.2010.72*

Puigcerver, J., & **Mocholí, C.** (2018). *PyLaia*. GitHub. Retrieved from: *https://github.com/jpuigcerver/PyLaia*

Reiter, N., Willand, M., & **Gius, E.** (2019). A shared task for the digital humanities. Chapter 1: Introduction to annotation, narrative levels and shared tasks. *Journal of Cultural Analytics, 2*(1), 11192. DOI: *https://doi.org/10.22148/16.048*

Rice, S. V., Kanai, J., & **Nartker, T. A.** (1993). An evaluation of OCR accuracy. In Grover, K. O. & Goetz, J. P. (Eds.), *Information Science Research Institute. 1993 Annual Research Report* (pp. 9–33). Las Vegas, NV: University of Las Vegas. Retrieved from *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.80.7878&rep=rep1&type=pdf#page=9*

Sahle, P. (2013). *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung.* PhD thesis, Universität zu Köln. Retrieved from *https://kups.ub.uni-koeln.de/5013/*

Stökl Ben Ezra, D. (2019). *eScripta*. Retrieved from *https://escripta.hypotheses.org/about*

Wettlaufer, J. (2016). Neue Erkenntnisse durch digitalisierte Geschichtswissenschaft(en)? Zur hermeneutischen Reichweite aktueller digitaler Methoden in informationszentrierten Fächern. *Zeitschrift Für Digitale Geisteswissenschaften, 1*. DOI: *https://doi.org/10.17175/2016_011*