Transforming scholarship in the archives through handwritten text recognition

Transforming Scholarship in the Archives Through Handwritten Text Recognition: *Transkribus* as a Case Study

Introduction

Archives are increasingly investing in the digitisation of their manuscript collections but until recently the textual content of the resulting digital images has only been available to those who have the time to study and transcribe individual passages. The use of computers to process and search images of historical papers using Handwritten Text Recognition (HTR) has the potential to transform access to our written past for the use of researchers, institutions and the general public. This paper reports on the *Recognition and Enrichment of Archival Documents (READ)* European Union Horizon 2020 project which is developing advanced text recognition technology on the basis of artificial neural networks and resulting in a publicly available infrastructure: the *Transkribus* platform. Users of *Transkribus* (whether institutional or individual) are able to extract data from handwritten and printed texts via HTR, while simultaneously contributing to the improvement of the same technology thanks to machine learning principles. The automated recognition of a wide variety of historical texts has significant implications for the accessibility of the written records of global cultural heritage.

This paper uses the *Transkribus* platform as a case study, focusing on the development, application and impact of HTR technology. It demonstrates that HTR has the capacity to make a significant contribution to the archival mission by making it easier for anyone to read, transcribe, process and mine historical documents. It shows that the technology fits neatly into the archival workflow, making direct use of growing repositories of digitised images of historical texts. By providing examples of institutions and researchers who are generating new resources with *Transkribus*, the paper shows how HTR can extend the existing research infrastructure of the archives, libraries and humanities domain. Looking to the future, this paper argues that this form of machine learning has the potential to change the nature and scope of historical research. Finally, it suggests that a cooperative approach from the archives, library and humanities community is the best way to support and sustain the benefits of the technology offered through *Transkribus*.

Handwritten Text Recognition – An Overview

Handwritten Text Recognition (HTR) is an active research area in the computational sciences, dating back to the mid-twentieth century (Dimond, 1957). HTR was originally

closely aligned to the development of Optical Character Recognition (OCR) technology, where scanned images of printed text are converted into machine-encoded text, generally by comparing individual characters with existing templates (Govindan and Shivaprasad, 1990; Schantz, 1982; Ul-Hasan *et al.*, 2016). HTR developed into a research area in its own right due to the variability of different hands, and the computational complexity of the task (Bertolami and Bunke, 2008; Kichuk, 2015; Leedham, 1994; Sudholt and Fink, 2016). Statistical advances in the 1980s, and advanced pattern recognition combined with artificial intelligence in the 1990s were followed by the development of deep neural network approaches in the 2000s and 2010s.[1] This, combined with the availability of increased computer processing power, has resulted in improvements in the recognition of handwritten historical documents, as is regularly evidenced at scientific competitions in the two major conferences in this area: the International Conference of Document Analysis and Recognition (ICDAR) and the International Conference on Frontiers in Handwriting Recognition (ICFHR). Researchers originally developed this technology with handwritten materials in mind and it is widely known in the computer science field under the initials HTR. However, the technology can equally be applied to early printed texts that are too complex to be processed adequately with OCR techniques.

Most prior application of HTR has been in the financial and commercial sectors (for example for postal address interpretation (Pal *et al*., 2012), bank-cheque processing (Dimauro *et al.,* 1997), signature verification (Hafemann *et al*., 2015), and biometric writer identification (Morera *et al.,* 2018)). However, recent successes in HTR coincide with the availability of affordable, high-quality digital imaging technologies, related online systems for hosting images, and subsequent programs of mass digitisation which are being carried out by most major libraries and archives worldwide to increase access to their collections (Borowiecki and Navarrete, 2016; Ogilvie, 2016; Terras 2010). Unfortunately, it has long been the problem that

> there are growing numbers of scanned manuscripts that current OCR and handwriting recognition techniques cannot transcribe, because the systems are not trained for the scripts in which these manuscripts are written. Documents in this category range from illuminated medieval manuscripts to handwritten letters to early printed works. Without transcriptions, these documents remain unsearchable (Edwards, 2007, p.1).

Transforming Scholarship in the Archives Through Handwritten Text Recognition: *Transkribus* as a Case Study

Mass digitisation of historical material, in combination with traditional archival catalogues and finding aids, is already broadening access to document collections. Automated transcription and searching of digitised texts goes further, expanding the existing possibilities of historical enquiry for scholars, institutions, commercial providers, and other users. Successful development of HTR will improve and increase access to collections, allowing users to quickly and efficiently pinpoint particular topics, words, people, places, and events in documents, but also changing the understanding of context, and multiplying research possibilities. The generation of machine-readable textual transcripts will provide the basis for advanced semantic, linguistic, and geo-spatial computational analysis of historical primary source material (see Gregory *et al.,* 2015; Meroño-Peñuela *et al.,* 2015; Weisser, 2016 for possibilities). The research questions which can then be asked of historic manuscripts change: the way institutions can deliver and present archival material will be similarly transformed (Estill and Levy, 2016).

Commercial digitisation providers are moving into this space, undertaking digitisation on behalf of under-funded institutions and licensing back access to the resulting resources. As of early 2018, *Adam Matthew Digital* describes itself as "currently the only publisher to utilise artificial intelligence to offer Handwritten Text Recognition (HTR) for its handwritten manuscript collections" (Adam Matthew Digital 2018). At the time of writing, it offers the same software as *Transkribus*, allowing HTR-based searching across several of its themed digitised archive collections and starting to provide HTR as a part of a collection management service via its *Quartex* platform.[2] However, this commercial exercise restricts HTR to contributing organisations and means that researchers and other individuals are unable to engage with the development and application of the technology. Machine learning is not a panacea and critical appraisal of its training process and its underlying data is essential if this technology is to be integrated into archival practice and scholarly research in a meaningful way.

It is within this framework, and with open aims, that the large-scale *READ*[3] research initiative has provided *Transkribus*[4] as the platform to deliver HTR technology to institutions and individual users. Although the *READ* project has published numerous research papers on the computational aspects of HTR[5], as well as datasets[6] and other project deliverables[7], this is the first publication from the project to cover the research programme from the perspective of

the active user community. In considering examples of projects working with *Transkribus*, it indicates that the combination of HTR and digitised content has potential to extend existing methods of scholarship in significant new directions.

The *Transkribus* Platform

Various projects have undertaken work on the OCR of early printed materials and experimented with the recognition of handwritten manuscripts (Bulacu *et al.*, 2009; Edwards, 2007; Firmani, *et al.*, 2018; Fischer *et al*., 2009, Springmann and Lüdeling, 2017, Terras, 2006, Weber *et al.*, 2018; Zant *et al*., 2009). However, there has not yet been sufficient interdisciplinary work applying deep neural network models to manuscript material and more importantly, there was previously no user-friendly platform to make this technology accessible. With *Transkribus*, historical manuscripts of all dates, languages and formats can be read, transcribed and searched by means of automated recognition. The *Transkribus* research infrastructure aims to provide a complete and reliable workflow for this process. Users work with *Transkribus* to create "ground truth" [8] data that is suitable for machine learning. From submitted images and transcripts, the HTR engines[9] learn to decipher (historical) handwritten or printed text from digital images and can then automatically generate transcripts of similar material.

Transkribus was conceived and launched in 2015 with funding from the European Commission's Seventh Framework Programme (FP7), as part of the *tranScriptorium* project (2013-2015).[10] Whilst *tranScriptorium* largely focused on computer science research, the implementation and development of *Transkribus* was placed at the centre of a successor project called *Recognition and Enrichment of Archival Documents* (*READ*) (2016-2019). Funded once more by the European Commission, this time under the Horizon 2020 scheme, *READ* aims to maintain, develop and promote a functioning online research infrastructure where new technologies can feed innovation in archival research. Dr Günter Mühlberger of the University of Innsbruck coordinates the *READ* project, working with collaborators from thirteen other European universities and research institutions[11] representing ten teams of computer scientists and developers who contribute to and construct components of the *Transkribus* infrastructure[12], but also teams engaged in humanities scholarship and institutions with archival material.

Transkribus services are currently freely available online, and directed towards four intended user groups: archivists, humanities scholars, computer scientists and members of the public, all of whom are interested in the study and exploitation of historical documents. The interests of these user groups overlap and each make a vital contribution to the *Transkribus* infrastructure. Memory institutions, humanities scholars and the public can provide digitised images and transcripts as ground truth for HTR training, whilst computer scientists deliver the necessary research to sustain this technology. Each user group can also derive tangible benefits from the initiative: archives can deliver searchable digitised collections for their users, humanities scholars can conduct research efficiently and members of the public can study their family history or contribute by transcribing or correcting transcripts of historical documents. Computer scientists can also request to reuse a wealth of data, in the form of images and transcripts of historical material, for their HTR research. This growing user network is central to the success of *Transkribus*: machine learning means that HTR becomes stronger with every document processed in the platform.

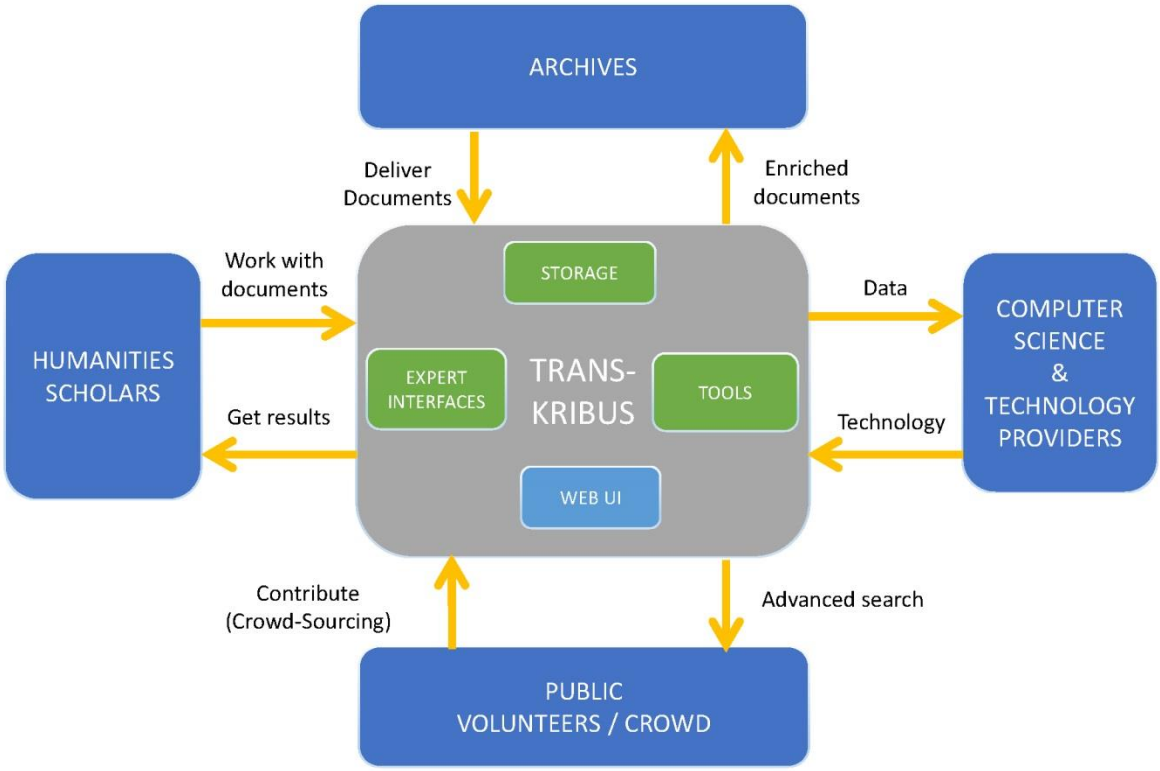


Figure 1: *Transkribus* stakeholders, representing what they can contribute, but also what they can gain, from interacting with the HTR technological platform supplied by *READ*.

Since 2015, the Digitisation and Electronic Archiving (DEA) group at the University of Innsbruck have been responsible for maintaining and developing the *Transkribus* Graphical User Interface (GUI): currently freely available via a downloadable, Java-based programme (for Windows, MacOS and Linux) that requires local installation.[13] Access to the platform is granted to users who have registered for an account at the *Transkribus* website. The *Transkribus* GUI stabilised with the release of version 1.0 at the beginning of 2017 (Kahle *et al.,* 2017): its most important features are fully integrated and functional, although new developments continue to improve the platform. In addition, all services and uploaded documents are also available in machine-readable form via a REST interface.[14] Most of the *Transkribus* software is available open source via GitHub[15], with two exceptions. The HTR module in *Transkribus* is licensed for use within the *READ* project by the Computational Intelligence Laboratory (CITlab) at the University of Rostock (one of the READ project members) and its collaborator, the German technology company PLANET.[16] An open-source neural network toolkit for HTR, known as Laia, has been created by the Pattern Recognition and Human Language Technology (PRHLT) research centre at the Polytechnic University of Valencia (another of the READ project members).[17] It will be integrated into the *Transkribus* GUI in 2019. The server component of *Transkribus* is also restricted at present in order to avoid the establishment of a competing service platform

Transkribus workflow

The latest advances in HTR research, based on deep neural networks, have been implemented in the *Transkribus* GUI. Neural networks can be trained to recognise a particular style of writing by examining and processing digitised images and transcriptions of documents. The result of the training process is what is known as a HTR "model", a computational system tailored to automatically transcribe a set of historical material. HTR technology is language-independent: the neural network training process for any type of alphabet, from any date, is the same. This means there is potential to train up models for any script, from any period (Leifert *et al.,* 2016). The technology follows a line-oriented approach where the image of a baseline (a horizontal line running underneath a line of text in a digitised image) and the corresponding correctly transcribed text represent the input for the learning algorithms of neural networks (Romero *et al.,* 2015).

Digitised images and their transcripts are the main prerequisite for working with HTR. They must be pre-processed in the *Transkribus* GUI in order to become ground truth data that can be used to train a HTR model to transcribe a specific collection of historical material, either that written by one writer or a set of similar types of writing. There are three main stages to creating ground truth in the *Transkribus* GUI. The first is uploading digitised images to the platform. The second is using Layout Analysis tools to segment the digitised images into lines. The third is accurately transcribing the text of each of the lines in a digitised image.

Transkribus accepts a range of image formats, and has sufficient server space to process large collections. When a user uploads images to *Transkribus*, these images remain private to their user account and are not made publicly available. A collection owner can allow other *Transkribus* users to view or work with their documents if they wish. Training data for HTR should be representative of the different parts of an archival collection, reflecting an appropriate variety of layouts, vocabulary and writing styles. Users can therefore select specific pages to become ground truth or simply choose pages from regular intervals within a collection (e.g. every tenth page). A ground truth dataset of 15,000 transcribed words (or around 75 pages) is generally sufficient for training a HTR engine to recognise text written in one hand. A model can be trained to recognise printed text with just 5000 transcribed words (or around 25 pages). According to the principle of machine learning, the more words of ground truth that a user submits, the more accurate the results are likely to be. Indeed, if a collection contains documents written in several hands or languages, it is recommended that users create ground truth for a higher number of transcribed words. To give an example, one of the strongest HTR models has been trained by the Bentham Project at University College London, one of the members of the *READ* project.[18] This model was trained on over 50,000 words from papers written by the English philosopher Jeremy Bentham (1748-1832) and his secretaries. In the best cases, it generates an output where around 95% of characters on similar pages from the Bentham collection are transcribed correctly by the program. This model is publicly available to all *Transkribus* users under the title "English Writing M1". The figures which appear later in this paper relate to this model.

Once images reside on the *Transkribus* server, they are ready for Layout Analysis or segmentation. Recent technological breakthroughs have enhanced the accuracy of this crucial process, making it easier for machines to identify text on archival documents which have

more complex layouts (Diem *et al.*, 2017; Grüning *et al.*, 2017; Leifert *et al.*, 2016). The *Transkribus* GUI contains both automatic and manual segmentation tools that allow users to mark their images with three segmentation elements: text regions around each block of text, line regions around each line of text and baselines running along the bottom of each line of text.[19] *Transkribus* users can commence automated segmentation on a batch of pages and the tool works in minutes to find the lines in images where words are set out relatively neatly on a page. The results of automated segmentation can sometimes be less precise when documents have a more complicated structure, such as a tabular form. In such cases, a combination of automated and manual segmentation will work to divide a page into lines. HTR engines rely most heavily on information they derive from the baselines running along the bottom of each line of text. Users can therefore check the positioning of automatically generated baselines and correct them if necessary.

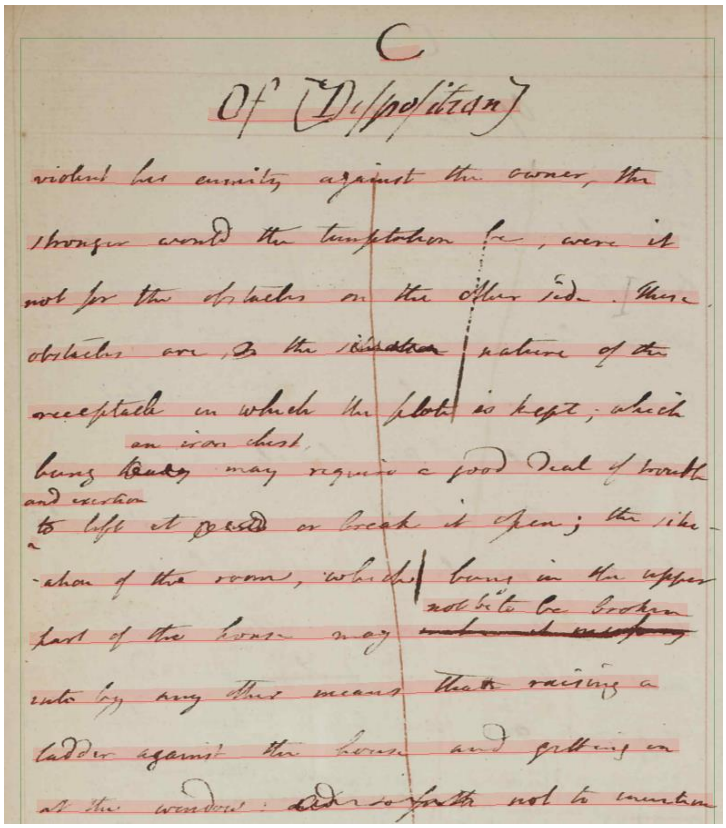


Figure 2: Screenshot showing an image of a manuscript loaded into the *Transkribus* GUI after automated baseline detection. Jeremy Bentham, "of [Disposition]", 1778, Box xxvii, Fol. 58, Bentham Papers, UCL Special Collections. Image courtesy of UCL Special Collections.

Transcription is the third and final part of creating ground truth. After segmentation, the *Transkribus* GUI displays a text editor field divided into lines which are connected to the lines drawn on the image. Users need to produce a consistent transcript of each line of the text in the image, replicating any spelling mistakes, unusual symbols or abbreviations. The neural networks can also learn from normalised transcriptions, where abbreviations have been expanded (Thöle, 2017). Users have the option to transcribe their documents in the *Transkribus Web* interface[20], a streamlined version of *Transkribus* that makes transcription simpler and quicker for larger teams or volunteers. The *Transkribus* GUI has a suite of tagging tools for those users who wish to create rich transcripts that could form part of a digital edition. At the current time, there is no benefit in marking up transcripts that are being prepared as ground truth. HTR engines are programmed to ignore tags and instead focus on recognising text. However, developments in Named Entity Recognition technology should permit the recognition of tagged content in the near future.

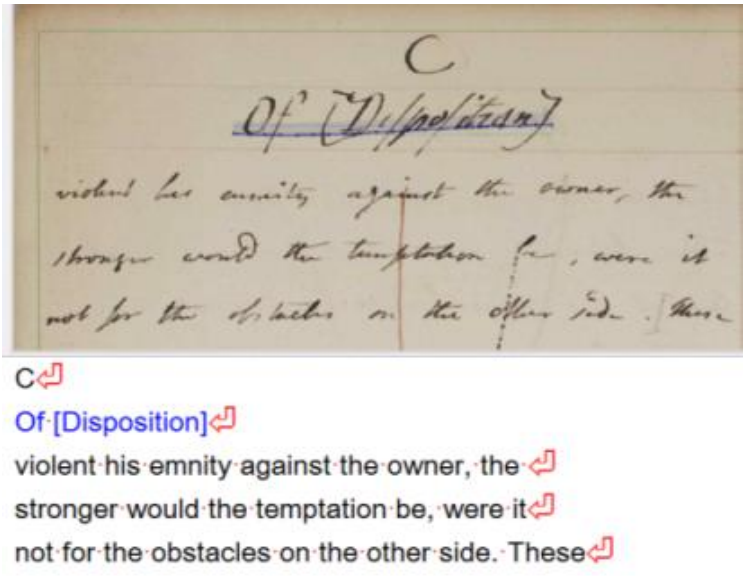


Figure 3: Screenshot showing an image of a manuscript which has been segmented automatically and then manually transcribed by a human transcriber in the *Transkribus* GUI. The blue line in the image represents a baseline and its corresponding text is shown in blue in the text editor below. Jeremy Bentham, "Of [Disposition]", 1778, Box xxvii, Fol. 58, Bentham Papers, UCL Special Collections. Image courtesy of UCL Special Collections.

Figure 4: Screenshot showing a line from an image of a manuscript transcribed manually by a human transcriber in the *Transkribus Web* interface. Jeremy Bentham, "Of [Disposition]", 1778, Box xxvii, Fol. 58, Bentham Papers, UCL Special Collections. Image courtesy of UCL Special Collections.

In summary, users upload images to the *Transkribus* GUI, segment each page into lines and then transcribe each page with a high level of consistency. With these three simple steps, ground truth creation is complete. Users who have existing transcriptions of their documents also have the option to truncate the process of creating training data, thanks to a Text2Image matching tool.[21] Once images and text file transcriptions have been uploaded to *Transkribus*, the Text2Image algorithm seeks to match the lines in the images to the lines of the transcribed text. Only lines that have been matched with a certain predefined confidence value will be included in the training data. The Text2Image matching tool therefore represents a simple and cost-efficient entry point into ground truth production and HTR for those who have collated existing transcriptions.

Users can request access to train their own HTR models or email the *Transkribus* team at the University of Innsbruck to request that a model be trained to recognise the text from their ground truth pages. At this stage, users can also send files containing relevant dictionaries or vocabulary lists which can improve the accuracy of the recognition. The process of model generation is complex: the learning effect of the HTR is achieved by adapting its respective hypotheses to the existing training data in an iterative process and thus independently finding those rules which provide the best output (the correct text) with a given input (the picture of the line), but for the user in *Transkribus* this complexity is resolved to a few parameters (Grüning, T. *et al.*, 2016; Sánchez, J. A. *et al*., 2014 and 2017; Strauß *et al*., 2016; Weidemann, M. *et al.*, 2017). It takes between several hours and several

days to train a model, depending on the size of the training data and the load on the computing infrastructure. The actual result of the training process is a model which is capable of recognising handwritten or printed documents which are similar to the ground truth. However, the output is not the transcription of the page itself, but rather a confidence matrix showing the likelihood of the appearance of each character in the alphabet at a given spot in the image of a line. With this confidence matrix further actions are possible, such as decoding the confidences into transcribed text, taking them as an input for keyword searching or in the future, using them to correct an automated transcript.

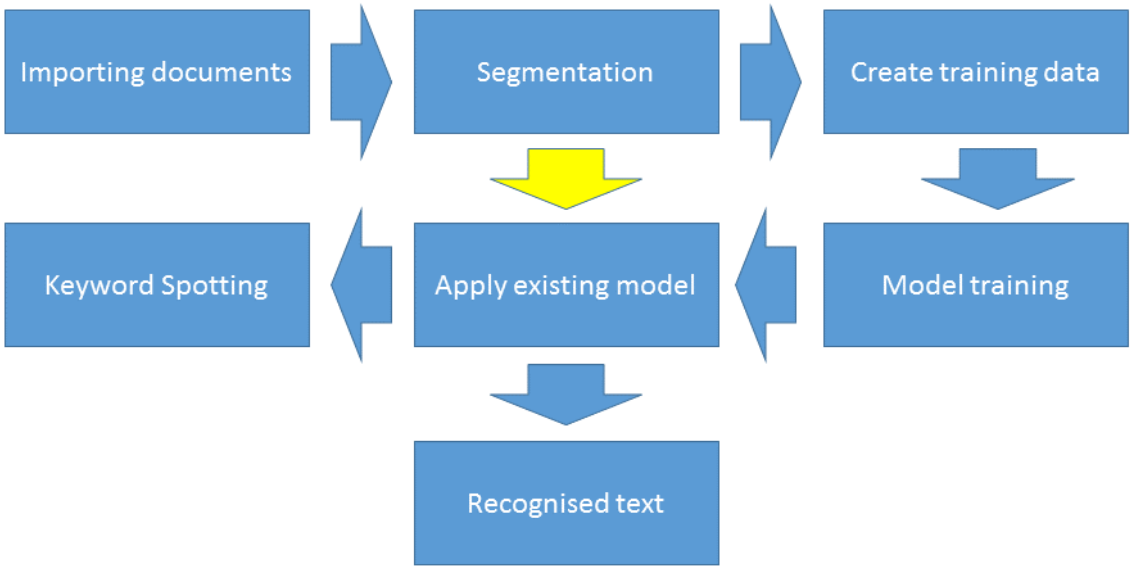
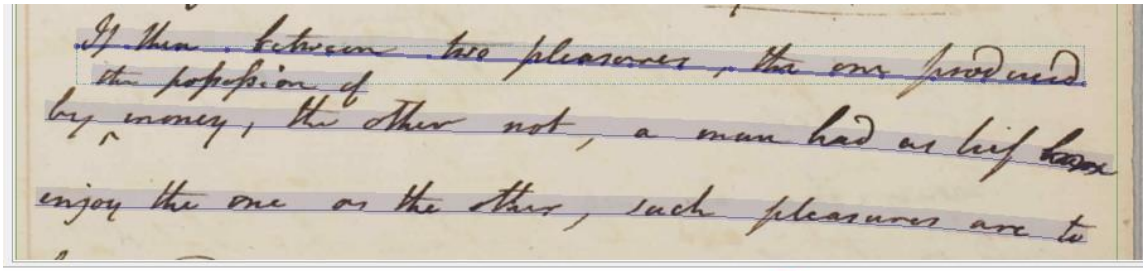


Figure 5: Workflow for HTR in *Transkribus*.

Once training is complete, users can access their model in the *Transkribus* GUI and generate an automated transcript of a page from their ground truth set. Any pages from the same collection that were not used as training data must be uploaded to the *Transkribus* GUI and then segmented into lines before they too can be automatically transcribed with HTR. In the current set-up 30 pages (with an average of 40 lines on each page) can be automatically transcribed in just over 28 minutes. It would therefore take 32 days and 18 hours to

automatically recognise 50,000 pages. New GPU servers are due to be installed at the University of Innsbruck, which will consequently improve these processing rates.

The *Transkribus* GUI displays standardised information about each HTR model in a particular collection, including its name, the documents on which it was trained and its accuracy level. Users are supplied with a learning curve which indicates the number of words used in the training and the best values achieved in generating the model. The platform determines the overall accuracy of the HTR model using a measurement of Character Error Rate (CER), which refers to the average percentage of characters transcribed incorrectly by the program.[22] During the training process, a small selection of pages from the ground truth is set aside as a test set and is not used to train the HTR. This means that *Transkribus* can provide CERs relating to the automated transcription of previously processed pages, as well as unknown pages from the same dataset. The platform also has a comparison function that enables users to compute and generate a visualisation of the accuracy of the computer-generated transcription of any page from the ground truth. In the best cases, HTR can produce automated transcripts of handwritten material with a CER of below 5% (meaning that 95% of the characters are correct). Outputs from models trained on printed material can be even better, reaching CERs of 1-2%. The use of dictionaries will in many cases improve the HTR results but the accuracy of neural networks on a purely visual level is high. The experience of *Transkribus* users indicates that transcripts with these accuracy rates can be proofread and corrected relatively quickly, with less effort than would be required to transcribe each page from scratch (Alvermann and Blüggel, 2017).



If·then·between·two·pleasures,·the·one·produced⏎
the·possession·of⏎
by·money,·the·other·not·a·man·had·as·his·those⏎
enjoy·the·one·as·the·other,·such·pleasures·are·to⏎

Figure 6: Screenshot showing an image of a manuscript transcribed automatically in the *Transkribus* GUI using the "English Writing M1" model and dictionary. The automated transcription of this page has a Character Error Rate (CER) of around 9%. Jeremy Bentham, "How to measure Pain and Pleasure", 1775, Box xxvii, Fol. 36a, Bentham Papers, UCL Special Collections. Image courtesy of UCL Special Collections.

If a HTR model is less accurate, with a CER of more than 10%, experiments suggest that automated transcriptions become less useful as a research resource in themselves because correcting myriad errors is more time consuming than manual transcription. However, it does not follow that less accurate results are ultimately useless. Indeed, HTR output can still be a solid foundation for searching and indexing vast collections of digitised documents. The *Transkribus* GUI provides access to a sophisticated searching technology known as Keyword Spotting.[23] This tool searches through the confidence values assigned to characters as part of the HTR process and recovers all possible matches for a given word (this is known as a "Query by String" approach). The results will return what the engine deems to be the best matches, as well other possible matches for that word based on alternative readings of each character on the page. This means that Keyword Spotting technology can find words in a collection, even if those words have been transcribed incorrectly by HTR. Moreover, it can recognise and retrieve results for words where there are historical or personal variations in spelling. Thus, this form of searching can produce useable results with HTR models that have higher error rates, up to 30% CER (Giotis, *et al.*, 2017; Puigcerver *et al.*, 2015 and 2017, Retsinas *et al*., 2016, Strauß *et al.*, 2016; Toselli *et al.*, 2017). The platform displays the results of a Keyword Spotting query as a list of transcribed words, thumbnail images of the portion of the digitised pages on which those words appear and a confidence rating for each word. In a future version of the *Transkribus* GUI, users will benefit from further research which facilitates search queries relating to partial words and graphical symbols (known as the "Query by Example" approach) (Zagoris *et al.*, 2017). Users will also be able to export their Keyword Spotting results as a data matrix for examining the contents of a document collection. A validation tool is being developed which will help users to easily eliminate incorrect results for their search term and create a controlled index of occurrences of that word.

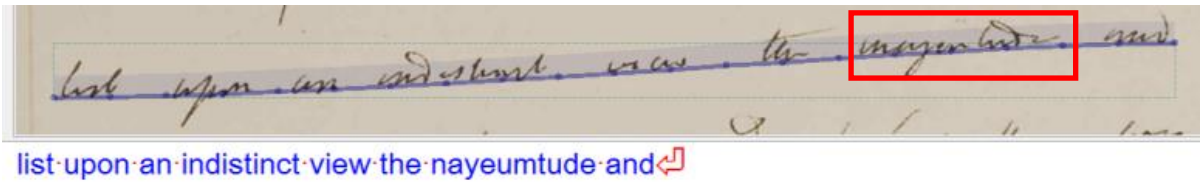


Figure 7: Screenshot showing an image of a line of a text in a manuscript transcribed automatically in the *Transkribus* GUI using the "English Writing M1" model and dictionary. The automated transcription of this page has a Character Error Rate (CER) of around 34%, which is useless for transcription. Nevertheless, the word "magnitude" is identified correctly via Keyword Spotting in *Transkribus*. Jeremy Bentham, "Annuity Notes", 1800, Box ii, Fol. 29, Bentham Papers, UCL Special Collections. Image courtesy of UCL Special Collections.

Once HTR has been completed on any given set of documents, it is up to the user to work with the resulting transcriptions in any way they feel appropriate. They can be included in digital editions, subjected to further computational analysis using semantic or linguistic techniques, or (in the case of large scale collections provided by institutions) ingested into content management systems to be used as a finding aid to locate the content of collections. There is therefore much potential to support the archival and manuscript studies community via the reliable transcription and searching of handwritten and printed texts

The *Transkribus user* community

The digitised archives and repositories that *READ* project members have provided are the primary test cases for the development of *Transkribus*. The Bentham Project at University College London has trained a succession of models with the aim of improving the automated recognition of Jeremy Bentham's handwriting.[24] Following a collaboration with the PRHLT research team at the Polytechnic University of Valencia, there is now an online platform for the Keyword Spotting of the near entirety of Bentham's papers (around 90,000 digitised images).[25] The Bentham Project is also considering how to integrate HTR technology into the workflow of *Transcribe Bentham*, its scholarly crowdsourcing initiative that asks members of the public to transcribe Bentham's writings (Causer and Terras, 2014). HTR could provide volunteers with automated transcripts of simple pages to check and correct or help them to decipher complex passages by providing suggested readings of each word on a page (Seaward, 2016). Passau Diocesan Archives[26] are utilising *Transkribus* to transcribe and

search their large collections of sacramental registers (Wurster *et al*., 2017). Experiments with a set of 1,200 images of death registers written in nineteenth-century German (around 400,000 words written by 40 different scribes) have shown that a CER between 17 and 19% can be achieved. Passau Diocesan Archives are also working to improve the automated Layout Analysis of tabular data by sorting tables from their collection into different categories that can be used as training templates. Improved table recognition and the possibility of exporting tabular data will have significant implications for the field, since many archival documents are laid out in tables and forms (Clinchant *et al.,* 2018). The National Archives of Finland[27] would like to enhance the usability of their vast collections of governmental records, many of which are digitised but not transcribed. They have worked with students and volunteers to produce training data for three collections: nineteenth-century court records written in Swedish, estate inventories of the Finnish nobility also written in nineteenth-century Swedish and diaries from the Second World War written in Finnish (Kallio, 2017). The best results came from the court records, where 75,000 words of ground truth produced a model capable of transcribing pages with a CER of around 12%. The multiplicity of writers in the other two collections meant that the results were somewhat weaker. 144,000 words of training data for the Second World War diaries led to an output with a CER of around 17%, whilst 99,000 words of the estate inventories trained a model that transcribed pages with a CER of around 24%. The National Archives of Finland will continue to engage students and volunteers in the creation of further training data in the hope of improving these accuracy rates. The State Archives of Zurich[28] had a head start in exploring the potential of HTR technology because they are in possession of nearly 200,000 pages of transcribed text relating to one of the main series of their archival collections from the nineteenth century (Hodel, 2017).[29] They have experimented with the Text2Image matching tool to pair 100,000 pages of these existing transcripts with corresponding digitised images, laying the groundwork for future training of HTR on a large scale. Training has already been undertaken on part of this data set, which comprises German language documents written between 1848 and 1853, totalling around 2,750,000 words. The output reaches a CER of around 6% when the model is applied to documents written in the same hand. A model has also been trained on a smaller subset of these documents (around 570,000 words from the years 1803-1882) and the results have a CER of around 18%. The accuracy of this model is sufficient to also recognise other texts written in nineteenth-century German and will be used as a basis for Keyword Spotting.

Transforming Scholarship in the Archives Through Handwritten Text Recognition: *Transkribus* as a Case Study

From 2015, when the *Transkribus* GUI became available, scholars, archives, volunteers and computer scientists were encouraged to download and benefit from it. At the time of writing there are now more than 17,000 registered individual users of the *Transkribus* platform, and more than 80 institutions and projects, including many libraries and archives, have signed a Memorandum of Understanding with *READ*.[30] Every month *Transkribus* users are generating about 130 HTR models trained on collections of different scripts, dates and languages: from medieval Hebrew to twentieth-century German Kurrent. Their experiences already demonstrate that the outputs of automated transcription and searching expedite connections with historical material and have the potential to reshape research practice. A detailed study of the *Transkribus* user community, which will explore the impact of HTR in specific archives and research projects, is forthcoming. The current paper presents an overview of success stories shared at the first *Transkribus* User Conference at the Technical University of Vienna, Austria, on 2-3 November 2017.

The library of the University of Greifswald has been working productively with *Transkribus* since early 2016, building up a significant set of ground truth based on a collection of minutes from the central administrative body of their University produced by three hands during the late eighteenth and early nineteenth century (Alvermann and Blüggel, 2017). The team have used this expanding data set to train a succession of models with increasingly accurate results. In the latest experiment, 410,000 words of ground truth have created a model that can produce transcripts of parts of the collection with a CER of just 5%. With such a low error rate, even experienced archivists can correct the automated text faster than typing manually. A version of this model, trained on 250,000 words is freely available to all registered *Transkribus* users under the title of "Konzilsprotokolle v1", so they can experiment with similar documents and see the potential of HTR. The team at the University of Greifswald have also shown how HTR technology can feed into the archival workflow by using *Goobi* software to provide online access to the searchable text of 20 volumes of the collection via their library catalogue.[31] The Georgian Papers Programme[32], led by William & Mary University Libraries has a similar desire to use HTR to open up a collection of historical material. The initiative will provide an online platform of transcriptions of papers from the British Royal Collection Trust, some generated by HTR and others by scholars, students and volunteers. First experiments with some 365,000 words of ground truth from a cache of essays by the English King George III (1738-1820) led to a HTR model capable of

producing transcripts with a CER of around 16% (Cornell, 2017). The Georgian Papers Programme is now working to better these results by improving the consistency of their ground truth and combining their original model with other models already trained to recognise eighteenth- and nineteenth-century English writing.

At the time of writing, HTR tends to be strongest for Western scripts because it can draw upon a larger reserve of training data for common languages like English, French or Latin. However, *Transkribus* users are also starting to generate good results on texts written in non-Western languages. The University of Belgrade Library are working with *Transkribus* with a view to allowing users of their archive to access transcribed and searchable text. They have used *Transkribus* to train a HTR model to recognise Cyrillic handwriting from the twentieth century. A training set of some 7000 words has generated a result where the CER is as low as around 2% on material that the programme has seen before (Jerkov and Sofronijevic, 2017). With more words of training data, the recognition of previously unseen material should become stronger.

Transkribus users have also benefited from integrating the platform into the workflow of existing research projects. The *Barlach 2020* project at the University of Rostock is working on a digital edition of letters written by the German sculptor and writer Ernst Barlach (1870-1938). They have trained a HTR model with some 42,000 words of Barlach's writing, integrating an earlier edition of Barlach's letters into the training process as a dictionary (Lemke and Onasch, 2017). The resulting model transcribes pages with a CER of around 9% and the team are now using these automated transcripts as a starting point for scholarly editing. The Centre for Manuscript Genetics at the University of Antwerp[33] is working on a digital edition representing the genesis of works by the Irish writer Samuel Beckett (1906-1989).[34] The team have trained models which can recognise Beckett's writings in both English and French with CERs of around 12% and 18% respectively. The project team is interested in using these transcripts to analyse the multiple drafts, layers and noise in Beckett's personal notes (Dillen, 2017). The PROLOPE research group at the Autonomous University of Barcelona[35] are working a digital edition of plays by the Spanish playwright Félix Lope de Vega (1562-1635) (Gázquez, 2017). They have collaborated with the PRHLT centre at the Polytechnic University of Valencia to create an online resource for the Keyword Spotting of a selection of manuscripts relating to Spanish Golden Age theatre.[36] The Bavarian

Academy of Sciences and Humanities, the University of Augsburg and the Berlin-Brandenburg Academy of Sciences and Humanities are collaborating on a long-term project to create an annotated digital edition of medieval German translations of the Gospels. They use *Transkribus* as a transcription tool to manually produce rich and exportable transcripts with XML tags that will form part of this digital edition (Vetter, 2017).

Other users are establishing that HTR technology can be applied fruitfully to early printed text. As part of the OCR-D[37] project, designed to improve the automated recognition of texts printed between the sixteenth and nineteenth centuries, the Berlin-Brandenburg Academy of Sciences and Humanities are compiling a large ground truth data set of different printed sources (Boenig and Würzner, 2017). Dario Kampkaspar and colleagues at the Austrian Centre for Digital Humanities (part of the Austrian Academy of Sciences) are already in the process of training a model for a digital edition of the printed text of the eighteenth-century *Wienerisches Diarium* newspaper.[38] The team use a mixture of OCR (using the ABBY FineReader tool available in the *Transkribus* GUI) and HTR to produce transcripts, correct these transcripts and then use these corrections to retrain their HTR model in the hope of improving the accuracy of the recognition (Kampkaspar, 2017). Karen Thöle's work at the University of Göttingen shows that *Transkribus* can also cope with more challenging printed texts, in this case an incunable written in late Medieval Latin. With a ground truth set of around 35,000 words, Thöle has produced a model that is able to both recognise the text with a CER of around 5% and also acknowledge and expand frequently used abbreviations (Thöle, 2017). These diverse examples illustrate how *Transkribus* users are recognising material of different dates, languages and styles. Automated transcripts allow for an unprecedented scale of access to digitised historical material, providing a basis for scholarly editing and research work.

Moreover, it must be acknowledged that all of the statistics presented in this paper are likely to improve significantly following a major update of the technology known as HTR+, which has been developed by the CITlab team at the University of Rostock. HTR+ draws on the Tensorflow[39] software library developed by Google, which means that deep neural networks can be constructed more efficiently than ever before. Experiments on three handwritten datasets suggest that the training of HTR+ is up to ten times faster than previous versions of HTR (Michael *et al.*,2018). Most importantly, this technology can improve the

CER of automated transcriptions by between 5 and 10%. HTR+ is now available in the *Transkribus* GUI upon request, and all existing HTR models will be retrained with this technology to allow all users to benefit from this latest advance in machine learning.

Improving *Transkribus*

The expanding network of *Transkribus* users has advantages for the usability and efficacy of HTR technology. User feedback and bug-reporting feeds directly into development work on the platform. As a system of machine learning, *Transkribus* also becomes stronger as more and more data is processed (Carbonell *et al.,* 2013). All documents uploaded to the *Transkribus* GUI remain private and are not publicly shared. In the background however, the neural networks are learning from every piece of ground truth submitted in the system and consequently becoming better at recognising different hands, scripts and languages. The more *Transkribus* users there are, the stronger the HTR will be. In the computational science field, the automated recognition of manuscripts on the basis of a sufficiently large set of ground truth is now viewed as a scientifically solved problem. The next goals for computer scientists are to optimise methods so that they require less training data to achieve comparable results, and build generic models that can work on similar fonts and hands. In the future "out-of-the-box" models could make it easier for even more users to engage with and benefit from HTR, particularly those members of the public who are interested in studying historical documents. Legally, the sharing of the models is unproblematic, because HTR training does not violate any copyright or moral rights: ground truth images and transcripts are used for training but do not actually become part of the resulting neural network model.

This network effect, which is made possible by the sharing of data, will play a decisive role in the expansion of the platform in the coming years. The growth of a research community will be facilitated in two ways. Firstly, users can already exchange models among themselves or between different collections and this will be made easier in the future. Secondly, as has been suggested above, the *Transkribus* team will train global models, which will unite different sets of training data and thus cover a wide variety of document types and writing styles. It makes more sense to adapt existing models, benefiting from the training data that is already in the system rather than training every new model from scratch.

In addition, there are future developments that need to occur in the technology behind *Transkribus.* There remain problems with the recognition of documents with a layout that is tabular or otherwise complex. Research will continue into the recognition of structural elements such as marginalia, headlines, addresses, dates, salutations and signatures. Computational analysis of writing styles is making writer identification possible, with the potential to attribute authorship to previously obscure documents. With the improvement of the system, the expansion of training data, and the increasing accuracy of the models, comes new opportunities. The *READ* project team have already constructed a number of prototype tools as part of the wider *Transkribus* infrastructure that are designed to expedite digitisation, the teaching of palaeography skills and the involvement of the public in historical research. With the *Transkribus Learn*[40] platform users can practice reading historical handwriting and the *DocScan*[41] mobile app and *ScanTent*[42] device enable users to take high-quality images of documents using a mobile phone. Future work includes allowing users to make meaningful contributions to the indexing of historical holdings by helping to validate the search results delivered by Keyword Spotting, flagging false positives and creating an index of controlled search words. Moreover, it may be possible to develop automated "search agents", who browse the ever-expanding range of digitized files for specific keywords and, if there are particularly interesting occurrences and accumulations, inform the user accordingly. Central to this is, of course, the user community. The *Transkribus* team will continue to engage with users to ensure the development of an infrastructure that supports their approaches. A future study of the activities of the user community will also highlight how this new suite of tools is changing humanities research practice.

Discussion

A major benefit to *Transkribus* is the cooperative manner of working, where all workflow steps - such as loading the documents into the platform, transcribing the texts, training the models and applying them to new documents - are carried out by the user group independently and under its own responsibility. The job of the *Transkribus* team is to ensure the availability of the platform, explain the various features, provide general support, and grow the user community, while ever-improving the underlying HTR technology. However, this also raises the question of the sustainability of such a platform. Considerable resources have already been channelled into the development of the *Transkribus* infrastructure. The

high number of users and the fact that cooperation agreements have already been established with memory institutions and research groups from all over the world, show that the technology of text recognition meets with great interest and is generally perceived as a central element of the future indexing of historical documents. As detailed above, research projects from across Europe are already using the *Transkribus* GUI as a productive tool. The more users, the better the recognition of handwritten and printed text of all kinds: the platform must therefore be scalable, as well as sustainable.

Legal and business models for the continued operation of the platform are currently being developed to prepare for the end of the EU-funded phase of the project in mid-2019. From this point onwards, *Transkribus* services will be provided as part of a European Cooperative Society (SCE) based at the University of Innsbruck. This is a legal entity founded with the objective of fulfilling the needs of its members, where profit is shared between members and used to improve services.[43] The working title for the initiative is *READ-COOP*. This legal basis is intended to promote cooperation between archives, libraries, universities and the general public. At the time of writing, a freemium service model is planned, with a mixture of free and paid-for services. The availability of documents, tools and data for the members of the network will continue to be a central element of the platform, promoting open research which allows confidence in the results generated from the system. Preparations for the implementation of *READ-COOP* were presented at the second *Transkribus* User Conference in November 2018 (Dellinger, 2018). Regular business operations will begin on 1 July 2019. To become viable in the long term, the platform needs to continue to support its research community, while generating enough resources to cover staffing and infrastructure costs. The alternative, of course, is that commercial digitisation providers will act as gatekeepers to HTR technologies that they can afford to underpin: restricting access to particular collections, and subscribing users (and seldom making their computational methods transparent).

These concerns come at a time when HTR is ready to bring potential change to the wider archival environment. Training data currently available in *Transkribus* can already be used to create models that provide the basis for Keyword Spotting to search substantial parts of the archive stock in English and German. For other languages, such as Dutch or Finnish, there is also sufficient training data now available to achieve useful results at least for parts of

the document stock. In a few years' time, it can be predicted that sufficient training data will be available to make the majority of digitised archival holdings in Europe searchable with this technology. It is therefore imperative that HTR remains accessible to libraries, archives, and individuals who would benefit from it, to allow vastly improved access to our written cultural heritage. *Transkribus* users will always be able to access and export their ground truth data. This data allows users to analyse the assumptions upon which their results were built and consider possible limitations and biases of automated transcriptions. Open discussion of algorithmic provenance and dependencies is important to develop trust in the reliability of research resources generated with artificial intelligence (Dayhoff and DeLeo, 2001; Samek *et al.*, 2017). Forthcoming new metrics for assessing the accuracy of HTR in the *Transkribus* GUI will also help users to gain a more assured understanding of the strengths and possible constraints of the technology. There is no doubt that the approaches of historians and genealogists will be heavily affected as this technology becomes embedded into available research methods. A future study of *Transkribus* users will be needed to examine the ramifications of this technology, establishing how it is challenging and extending the scope of historical analysis and how these new approaches can be best conceptualised, taught and supported.

Conclusion

This paper has provided the first published overview of research undertaken in the *tranScriptorium* and *READ* EU-funded projects, which has resulted in the establishment of the *Transkribus* platform for the automated recognition of historical documents. For over three years, the platform has been providing free access to HTR technology that can be applied to banks of images of digitised manuscripts, allowing useful transcripts of the material to be generated and improving the underlying technology for current and future users via machine learning. Such a project is only possible with an interdisciplinary collaboration of computer scientists, developers, humanities scholars, archivists and librarians. The resulting infrastructure has the potential to change the reach and scope of research questions that depend on handwritten primary historical sources and this paper has supplied evidence of a range of research projects that are already successfully engaging with the *Transkribus* GUI to transcribe and study archival documents. There are benefits to be realised if HTR can be integrated into the digitisation cycle of manuscript material: using the

results from this technology as a finding aid across mass-digitised content will vastly improve access to images of historical digitised manuscripts for all. Questions of sustainability now become pressing, just as the technology has become operational: raising issues of ownership, access, and collegiate working across the heritage sector. There is no doubt that improving access to historical texts in this way will change the use and analysis of archival material: it is now time to create and implement sustainable support structures that will allow this technology to be available to as wide a research community as possible.

Bibliography

Adam Matthew Digital (2018), "Handwritten Text Recognition: Artificial intelligence transforms discoverability of handwritten manuscripts." https://www.amdigital.co.uk/products/handwritten-text-recognition (accessed 13 December 2018).

Alvermann, D. and Blüggel, B. (2017), "Transkribus at Greifswald. Idea, practice, results, perspective", paper presented at Transkribus User Conference 2017, 2-3 November 2017, Technical University of Vienna, Vienna, slides available at: https://read.transkribus.eu/wp-content/uploads/2017/07/Alvermann_Bluegel_Greifswald.pdf (accessed 13 December 2018).

Bertolami, R. and Bunke, H. (2008), "Hidden Markov model-based ensemble methods for offline handwritten text line recognition", *Pattern Recognition,* Vol. 41 No. 11, pp. 3452-460.

Boenig, M. and Würzner, K.-M. (2017), "Compilation of a Large Ground-Truth Data Set Using Transkribus", paper presented at Transkribus User Conference 2017, 2-3 November 2017, Technical University of Vienna, Vienna, slides available at: https://read.transkribus.eu/wp-content/uploads/2017/07/Boenig_Wuerzner_Groundtruth-1.pdf (accessed 13 December 2018).

Borowiecki, K. J. and Navarrete, T. (2016), "Digitization of heritage collections as indicator of innovation", *Economics of Innovation and New Technology*, Vol. 26 No. 3, pp. 227-46.

Bulacu, M., Brink A., van der Zant, T. and Schomaker, L. (2009), "Recognition of Handwritten Numerical Fields in a Large Single-Writer Historical Collection", in *2009 10th International Conference on Document Analysis and Recognition*, IEEE, pp. 808-812.

Carbonell, J. G., Michalski, R. S. and Mitchell, T. M. (2013), "An Overview of Machine Learning", in Michalski, R. S., Carbonell, J. G. and Mitchell, T. M. (Eds.), *Machine Learning: An Artificial Intelligence Approach*, Springer-Verlag Berlin Heidelberg, pp. 3-23.

Causer, T., and Terras, M. (2014), "'Many Hands Make Light Work. Many Hands Together Make Merry Work": Transcribe Bentham and Crowdsourcing Manuscript Collections", Ridge, M. (Ed.), *Crowdsourcing Our Cultural Heritage*, Ashgate, Farnham, pp. 57-88.

Clinchant, S., Déjean, H., Meunier, J., Lang, E. M. and Kleber, F. (2018), "Comparing Machine Learning Approaches for Table Recognition in Historical Register Books", in *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, IEEE, pp.133-38.

Cornell, D. (2017), "Georgian Papers Programme", paper presented at Transkribus User Conference 2017, 2-3 November 2017, Technical University of Vienna, Vienna, slides available at: https://read.transkribus.eu/wp-content/uploads/2017/07/Cornell_Georgian_Papers.pdf (accessed 13 December 2018).

Dayhoff, J. E. and DeLeo, J. M. (2001), "Artificial neural networks: opening the black box", *Cancer*, Vol. 91 No. 91, pp. 1615-635.

Dellinger, M. (2018) "The READ-COOP: working together for the future of digital cultural heritage", paper presented at Transkribus User Conference 2018, 8-9 November 2018, Technical University of Vienna, Vienna, slides available at: https://read.transkribus.eu/wp-content/uploads/2018/11/DELLINGER-SCE.pdf (accessed 13 December 2018).

Diem, M., Kleber, F., Fiel, S., Grüning, T. and Gatos, B. (2017), "cBAD: ICDAR2017 Competition on Baseline Detection", in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, pp.1355-360.

Dillen, W. (2017), "Transkribus in Practice", paper presented at Transkribus User Conference 2017, 2-3 November 2017, Technical University of Vienna, Vienna, slides available at: https://read.transkribus.eu/wp-content/uploads/2017/07/Dillen_Beckett_edition.pdf (accessed 13 December 2018).

Dimauro, G., Impedovo, S., Pirlo, G. and Salzo, A. (1997), "Automatic Bankcheck Processing: A New Engineered System", *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 11 No. 4, pp. 467-504.

Dimond, T. L. (1957), "Devices for Reading Handwritten Characters", in *Papers and discussions presented at the December 9-13, Washington, 1957, Eastern Joint Computer Conference: Computers with Deadlines to Meet*, ACM, pp. 232-37.

Edwards, J. A. (2007), "Easily Adaptable Handwriting Recognition in Historical Manuscripts", PhD Dissertation, University of California, Berkeley, https://www2.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-76.pdf (accessed 13 December 2018).

Estill, L. and Levy, M. (2016), "Chapter 12 Evaluating digital remediations of women's manuscripts", *Digital Studies/Le champ numérique*, Vol. 6 http://doi.org/10.16995/dscn.12 (accessed 13 December 2018).

European Commission (2015), "tranScriptorium", Community Research and Development Information Service, https://cordis.europa.eu/project/rcn/106843_en.html (accessed 13 December 2018).

European Commission (2016). "Recognition and Enrichment of Archival Documents". Community Research and Development Information Service, https://cordis.europa.eu/project/rcn/198756_en.html (accessed 13 December 2018).

Firmani, D., Maiorino, M., Merialdo, P. and Nieddu, E. (2018), "Towards Knowledge Discovery from the Vatican Secret Archives. In Codice Ratio - Episode 1: Machine

Transcription of the Manuscripts", arXiv preprint https://arxiv.org/abs/1803.03200 (accessed 13 December 2018).

Fischer, A., Wüthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G. and Stolz, M. (2009), "Automatic Transcription of Handwritten Medieval Documents", in *2009 15th International Conference on Virtual Systems and Multimedia*, IEEE, pp. 137-42.

Gázquez, R. V. (2017), "La edición del teatro de Lope", paper presented at Transkribus User Conference 2017, 2-3 November 2017, Technical University of Vienna, Vienna, slides available at: https://read.transkribus.eu/wp-content/uploads/2017/07/Valdes_PROLOPE.pdf (accessed 13 December 2018).

Giotis, A. P., Sfikas, G., Gatos, B. and Nikou, C. (2017), "A survey of document image word spotting techniques", *Pattern Recognition*, Vol. 68, pp. 310–32.

Govindan, V. K. and Shivaprasad, A. P. (1990), "Character recognition – A review", *Pattern recognition*, Vol. 23 No. 7, pp. 671-83.

Gregory, I., Donaldson, C., Murrieta-Flores, P. and Rayson, P. (2015), "Geoparsing, GIS and Textual Analysis: Current Developments in Spatial Humanities Research, *International Journal of Humanities and Arts Computing*, Vol. 9 No. 1, pp.1-14.

Grüning, T., Leifert, G., Strauß, T. and Labahn, R. (2017) "A Robust and Binarization-Free Approach for Text Line Detection in Historical Documents", in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, pp. 236-41.

Grüning, T., Leifert, G., Strauß, T. and Labahn, R. (2016), "Deliverable 7.7, HTR Engine based on Neural Networks P1", Deliverable submitted to the European Commission, http://read.transkribus.eu/wp-content/uploads/2017/01/READ_D7.7_HTRbasedonNN.pdf (accessed 13 December 2018).

Hafemann, L. G., Sabourin, R. and Oliveira, L. S. (2015), "Offline Handwritten Signature Verification – Literature review", arXiv preprint https://arxiv.org/abs/1507.07909 (accessed 13 December 2018).

Hodel, T. (2017), "Deliverable 8.4, Large Scale Demonstrators. Evaluation and Bootstrapping", Deliverable submitted to the European Commission, https://read.transkribus.eu/wp-content/uploads/2017/12/READ_8.5-StAZH-v1.pdf (accessed 13 December 2018).

Jerkov, A. and Sofronijevic, A. (2017), "University of Belgrade library, Transkribus User Experience", paper presented at Transkribus User Conference 2017, 2-3 November 2017, Technical University of Vienna, Vienna, slides available at: https://read.transkribus.eu/wp-content/uploads/2017/07/Jerkov_Uni_Belgrade_library.pdf (accessed 13 December 2018).

Kahle, P., Colutto, S., Hackl, G. and Mühlberger, G. (2017), "Deliverable 4.2, READ Platform and Service Maintenance", Deliverable submitted to the European Commission, https://read.transkribus.eu/wp-content/uploads/2017/12/D4.2.pdf (accessed 13 December 2018).

Kallio, M. (2017), "Deliverable 8.8, Layout Analysis and Crowdsourcing", Deliverable submitted to the European Commission, https://read.transkribus.eu/wp-content/uploads/2017/12/Deliverable_8.8.pdf (accessed 13 December 2018).

Kampkaspar, D. (2017), "Digital edition of the »Wienerisches Diarium«", paper presented at Transkribus User Conference 2017, 2-3 November 2017, Technical University of Vienna, Vienna, slides available at: https://read.transkribus.eu/wp-content/uploads/2017/07/Kampkaspar_Wienerisches_Diarium.pdf (accessed 13 December 2018).

Kichuk, D. (2015), "Loose, Falling Characters and Sentences: The Persistence of the OCR Problem in Digital Repository E-Books", *Libraries and the Academy*, Vol. 15 No. 1, pp. 59-91.

Leedham, C. G. (1994), "Historical perspectives of handwriting recognition systems", in *IEE Colloquium on Handwriting and Pen-Based Input,* IEEE, pp. 1-3.

Leifert, G., Strauß, T., Grüning, T. and Labahn, R. (2016), "CITlab ARGUS for historical handwritten documents", arXiv preprint http://arxiv.org/abs/1605.08412 (accessed 13 December 2018).

Lemke, K. and Onasch, P. (2017), "»I do not find hot punch as nice as writing letters. « Transcribing Ernst Barlach's letters (1870-1938) with Transkribus", paper presented at Transkribus User Conference 2017, 2-3 November 2017, Technical University of Vienna, Vienna, slides available at: https://read.transkribus.eu/wp-content/uploads/2017/07/Onasch_Lemke_Barlach.pdf (accessed 13 December 2018).

Meroño-Peñuela, A., Ashkpour, A., Van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S. and Van Harmelen, F. (2015), "Semantic Technologies for Historical Research: A Survey", *Semantic Web*, Vol. 6 No. 6, pp. 539-64.

Michael, J., Weidemann, M. and Labahn, R. (2018), "Deliverable 7.9, HTR Engine based on Neural Networks P3", Deliverable submitted to the European Commission, https://read.transkribus.eu/wp-content/uploads/2018/12/Del_D7_9.pdf (accessed 13 December 2018).

Morera, Á., Sánchez, Á., Vélez, J. F. and Moreno, A. B. (2018), "Gender and Handedness Prediction from Offline Handwriting Using Convolutional Neural Networks", *Complexity,* Vol. 2018, https://doi.org/10.1155/2018/3891624 (accessed 13 December 2018).

Mühlberger, G., (2015), "Die automatisierte Volltexterkennung historischer Handschriften als gemeinsame Aufgabe von Archiven, Geistes- und Computerwissenschaftlern. Das Modell einer zentralen Transkriptionsplattform als virtuelle Forschungsumgebung", in Becker, I. C. and Oertel, S. (Eds.), *Digitalisierung im Archiv. Neue Wege der Bereitstellung des Archivguts. Beiträge des 18. Archivwissenschaftlichen Kolloquiums am 26. und 27. November 2013,* Archivschule Marburg, Marburg, pp. 87-116.

Ogilvie, B. (2016), "Scientific Archives in the Age of Digitization", *Isis*, Vol. 107 No. 1, pp. 77-85.

Pal, U., Roy, R. K. and Kimura, F. (2012), "Multi-lingual City Name Recognition for Indian Postal Automation", in *ICFHR '12 Proceedings of the 2012 International Conference on Frontiers in Handwriting Recognition*, IEEE, pp. 169-73.

Puigcerver, J., Toselli, A. H., and Vidal, E. (2015), "ICDAR2015 Competition on Keyword Spotting for Handwritten Documents", in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, pp. 1176-180.

Puigcerver, J., Toselli, A. H. and Vidal, E. (2017), "Querying out-of-vocabulary words in lexicon-based keyword spotting", *Neural Computing and Applications*, Vol. 28 No. 9, pp. 2373-382.

Retsinas, G., Louloudis, G., Stamatopoulos, N. and Gatos, B. (2016), "Keyword Spotting in Handwritten Documents Using Projections of Oriented Gradients", *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, IEEE, pp. 411-16.

Romero, V., Sánchez, J. A., Bosch, V., Depuydt, K. and De Does, J. (2015), "Influence of text line segmentation in Handwritten Text Recognition", *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 536-40.

Samek, W., Wiegand, T. and Müller, K.-R. (2017), "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models", arXiv preprint, https://arxiv.org/abs/1708.08296 (accessed 13 December 2018).

Sánchez, J. A., Romero, V., Toselli, A. H. and Vidal, E. (2014), "ICFHR2014 Competition on Handwritten Text Recognition on tranScriptorium Datasets (HTRtS)", in *2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, IEEE, pp. 785-90.

Sánchez, J. A., Romero, V. Toselli, A. H, Villegas, M. and Vidal, E. (2017), "ICDAR2017 Competition on Handwritten Text Recognition on the READ Dataset", in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR),* IEEE, pp. 1383-388.

Schantz, H. F. (1982), The *History of OCR (Optical Character Recognition),* Recognition Technologies Users Association, Manchester Center, VT.

Seaward, L. (2016), "Deliverable 4.10, Transcribe Bentham", Deliverable submitted to the European Commission, http://read.transkribus.eu/wp-content/uploads/2017/01/READ_D4.10_TranscribeBentham.pdf (accessed 13 December 2018).

Springmann, U. and Lüdeling, A. (2017), "OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus", *Digital Humanities Quarterly*, Vol. 11 No. 2 http://digitalhumanities.org/dhq/vol/11/2/000288/000288.html (accessed 13 December 2018).

Strauß T., Grüning, T., Leifert, G. and Labahn. R. (2016), "CITlab ARGUS for Keyword Search in Historical Handwritten Documents - Description of CITlab's System for the ImageCLEF 2016 Handwritten Scanned Document Retrieval Task", in *CLEF Working Notes*, pp. 399-412.

Sudholt, S. and Fink, G. A. (2016), "PHOCNet: A Deep Convolutional Neural Network for Word Spotting in Handwritten Documents", in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR),* IEEE, pp. 277-82.

Terras, M. (2006), *Image to Interpretation: An Intelligent System to Aid Historians in the Reading of the Vindolanda Texts*, Oxford University Press, Oxford.

Terras, M. (2010), "The Rise of Digitisation: An Overview", Rukowski, R. (Ed.), *Digitisation Perspectives*, Sense Publishers, Netherlands, pp. 3-20.

Thöle, K. (2017), "Transcribing a Highly Abbreviated Incunable (and Some More Manuscript Sources)", paper presented at Transkribus User Conference 2017, 2-3 November 2017, Technical University of Vienna, Vienna, slides available at: https://read.transkribus.eu/wp-content/uploads/2017/07/Thoele_Incunable.pdf (accessed 13 December 2018).

Toselli, A. H., Romero, V. and Vidal, E. (2017), "Word graphs size impact on the performance of handwriting document applications", *Neural Computing and Applications*, Vol. 28 No. 9, pp. 2477-487.

Ul-Hasan, A., Bukhari, S. S. and Dengel, A. (2016), "OCRoRACT: A Sequence Learning OCR System Trained on Isolated Characters", in *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, IEEE, pp. 174–79.

Van der Zant, T., Schomaker, L., Zinger, S. and Van Schie, H. (2009), "Where are the Search Engines for Handwritten Documents?", *Interdisciplinary Science Reviews*, Vol. 34 No. 2-3, pp. 224-35.

Vetter, A. (2017), "The Austrian Bible Translator – The Word of God in German. Annotated Critical Hybrid Edition", paper presented at Transkribus User Conference 2017, 2-3 November 2017, Technical University of Vienna, Vienna, slides available at: https://read.transkribus.eu/wp-content/uploads/2017/07/Vetter_Austrian_Bible.pdf (accessed 13 December 2018).

Weber, A., Ameryan, M., Wolstencroft, K., Stork, L., Heerlien, M. and Schomaker, L. (2018), "Towards a Digital Infrastructure for Illustrated Handwritten Archives", in Ioannides, M. (Ed.), *Digital Cultural Heritage*, Springer International, pp. 155-66.

Weidemann, M., Michael, J., Grüning, T. and Labahn, R. (2017), "Deliverable 7.8, HTR Engine based on Neural Networks P2", Deliverable submitted to the European Commission. https://read.transkribus.eu/wp-content/uploads/2017/12/Del_D7_8.pdf (accessed 13 December 2018).

Weisser, M. (2016), *Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis,* Wiley-Blackwell.

Wurster, H. W., Putz, H., Lang, E. M., Fronhöfer, W., Fronhöfer, A. and Mühlbauer, E. (2017), "Deliverable 8.11, Large Scale Demonstrators. Keyword Spotting in Registry Books P2", Deliverable submitted to the European Commission, https://read.transkribus.eu/wp-content/uploads/2017/12/READ_D8_11_LSD_Passau.pdf (accessed 13 December 2018).

Zagoris, K., Pratikakis, I. and Gatos, B. (2017), "Unsupervised Word Spotting in Historical Handwritten Document Images Using Document-Oriented Local Features", in *IEEE Transactions on Image Processing,* Vol. 26 No. 8, pp. 4032-4041.

[1] Neural networks are computational systems of hardware and software that are loosely modelled on the biological networks found in animal brains. They learn and improve their performance by training on a series of examples. Deep neural networks are a class of machine learning algorithms that use multiple layers of processing to learn and analyse their task.
[2] https://www.quartexcollections.com/ (accessed 13 December 2018).
[3] Recognition and Enrichment of Archival Documents (READ), Project ID 674943, H2020-EINFRA-2015-1 European Commission (2016), https://read.transkribus.eu/ (accessed 13 December 2018).
[4] https://transkribus.eu/ (accessed 13 December 2018).
[5] Computational science research generated by the READ project includes that on Layout Analysis (Diem *et al.* 2017, Grüning *et al.,* 2017), Keyword Spotting (Puigcerver *et al.,* 2017; Toselli *et al.,* 2017) and word spotting (Zagoris *et al*., 2017; Toselli *et al*., 2016). For other research see https://read.transkribus.eu/research-publications/ (accessed 13 December 2018).
[6] https://zenodo.org/communities/scriptnet/?page=1&size=20 (accessed 13 December 2018).
[7] https://read.transkribus.eu/deliverables/ (accessed 13 December 2018).
[8] "Ground truth" is a term commonly used in machine learning to refer to accurate, objective information provided by empirical, direct processes, rather than that inferred from sources via the statistical calculation of uncertainty. In the case of Transkribus, "ground-truth" information is gathered by training the system with enough data (usually around 15,000 words or 75 pages) relating an individual script, which can then be used to create a model that can be applied successfully to large volumes of the same script.
[9] The current implementation of the HTR engine comes from the CITlab group of the University of Rostock (see Leifert *et al.,* 2016).
[10] TranScriptorium, Project ID 600707, FP7-ICT-2011-9, European Commission (2015), http://transcriptorium.eu/ (accessed 13 December 2018).
[11] Institutions contributing to READ include the University of Innsbruck (coordinator / Austria), the Polytechnic University of Valencia (Spain), University College London (United Kingdom), National Center for Scientific Research "Demokritos" (Greece), Democritus University of Thrace (Greece), CoSector – University of London (United Kingdom), Technical University of Vienna (Austria), University of Rostock (Germany), University of Leipzig (Germany), NAVER LABS Europe (France), École Polytechnique Féderale de Lausanne (Switzerland), National Archives of Finland (Finland), State Archives of Zurich (Switzerland), Passau Diocesan Archives (Germany), and more recently the University of Edinburgh (United Kingdom).
[12] Essential components of the technology come from the CITlab group of the University of Rostock, and the PRHLT group of the Polytechnic University of Valencia, with additional contributions from the wider consortium.
[13] Downloadable from http://transkribus.eu/ (accessed 13 December 2018).
[14] https://transkribus.eu/wiki/index.php/REST_Interface (accessed 13 December 2018).
[15] https://github.com/Transkribus/ (accessed 13 December 2018).
[16] http://planet.de/ (accessed 13 December 2018).
[17] https://github.com/jpuigcerver/Laia (accessed 13 December 2018).

[18] For UCL's contribution to READ see, http://www.ucl.ac.uk/bentham-project and https://read.transkribus.eu/network/university-college-london/ (accessed 13 December 2018).
[19] Transkribus includes Layout Analysis tools by researchers at the National Center for Scientific Research "Demokritos", the CITlab group at the University of Rostock and the Computer Vision Lab at the Technical University of Vienna. The CITlab tool currently generates the best results on complex historical documents.
[20] http://transkribus.eu/r/read/ (accessed 13 December 2018).
[21] The CITlab group of the University of Rostock developed the Text2Image matching tool. For more information see, https://transkribus.eu/wiki/images/6/6f/HowToUseExistingTranscriptions.pdf (accessed 13 December 2018).
[22] The *Transkribus* GUI also allows users to compute accuracy according to a Word Error Rate (WER). Additional metrics for measuring the accuracy of automated transcriptions will be made available in the platform over the coming months. The CERs mentioned in this paper are rounded off to the nearest whole number.
[23] Various approaches to Keyword Spotting are being researched by groups at the University of Rostock, the Polytechnic University of Valencia, Democritus University of Thrace and the National Center for Scientific Research "Demokritos". The *Transkribus* GUI currently provides access to a "Query by String" method by the CITlab group at the University of Rostock.
[24] http://blogs.ucl.ac.uk/transcribe-bentham/2018/11/28/project-update-automated-recognition-bentham-handwriting/ (accessed 13 December 2018).
[25] http://prhlt-carabela.prhlt.upv.es/bentham/ (accessed 13 December 2018).
[26] http://www.bistum-passau.de/bistum/archiv and https://read.transkribus.eu/network/passau-diocesan-archives/ for their contribution to READ (accessed 13 December 2018).
[27] http://www.arkisto.fi/en/frontpage and https://read.transkribus.eu/network/national-archives-finland/ for their contribution to READ (accessed 13 December 2018).
[28] https://staatsarchiv.zh.ch/internet/justiz_inneres/sta/en/home.html and https://read.transkribus.eu/network/zurich-state-archives/ for their contribution to READ (accessed 13 December 2018).
[29] http://www.archives-quickaccess.ch/search/stazh/krp (accessed 13 December 2018).
[30] https://read.transkribus.eu/network/ (accessed 13 December 2018).
[31] http://www.digitale-bibliothek-mv.de/viewer/image/PPNUAG_0_1_St_666/1/LOG_0003/ (accessed 13 December 2018).
[32] http://georgianpapersprogramme.com/ (accessed 13 December 2018).
[33] https://www.uantwerpen.be/en/research-groups/centre-for-manuscript-genetics/ (accessed 13 December 2018).
[34] http://www.beckettarchive.org/ (accessed 13 December 2018).
[35] http://prolope.uab.cat/ (accessed 13 December 2018).
[36] http://prhlt-carabela.prhlt.upv.es/tso/ (accessed 13 December 2018).
[37] http://ocr-d.de/eng (accessed 13 December 2018).
[38] https://www.oeaw.ac.at/en/acdh/projects/wiennerisches-diarium-digital/ (accessed 13 December 2018).
[39] https://www.tensorflow.org/ (accessed 13 December 2018).
[40] https://learn.transkribus.eu (accessed 13 December 2018).
[41] https://play.google.com/store/apps/details?id=at.ac.tuwien.caa.docscan (accessed 13 December 2018).
[42] https://scantent.cvl.tuwien.ac.at/en/ (accessed 13 December 2018).
[43] https://ec.europa.eu/growth/sectors/social-economy/cooperatives/european-cooperative-society_en (accessed 13 December 2018).