1. Personal Details

Please fill in all your personal details. If you are submitting your proposal as a team, please copy the table below for each team member

Title	Dr.
Name	Giovanni Colavizza
Organisation	University of Amsterdam
Job title	Assistant Professor
E-mail address	g.colavizza@uva.nl
Phone Number	0629226788
(Envisioned) Date of PhD Defense	Completed, 28/08/2018
Availability Please fill in which period you are available	Both

Bio (max 100 words)

Please provide a short biography of each team member with a maximum of 100 words that provides evidence of your expertise in the field of your proposal.

Giovanni Colavizza is an assistant professor of digital humanities at the University of Amsterdam. He did his PhD at EPFL on the Venice Time Machine, building a prototype citation index for it (https://venicescholar.dhlab.epfl.ch). Afterwards, he has been part of the research engineering group at The Alan Turing Institute, working on AI applications to humanities and GLAM data. Giovanni is currently working on several research topics involving text mining, with a focus on historical newspapers (Living with Machines, with the British Library), the linguistic analysis of scientific publications and of Wikipedia as a source of public knowledge.

Key Publications

Please list your 3 most important publications that are relevant for your proposal. Please feel free to also include publications that are under review or in print.

Colavizza, Giovanni, Mario Infelise, and Frédéric Kaplan. 2014. "Mapping the Early Modern News Flow: An Enquiry by Robust Text Reuse Detection." *Social Informatics 2014*. In *Lecture Notes in Computer Science*, 8852:244–253. Cham: Springer. https://doi.org/10.1007/978-3-319-15168-7_31.

Colavizza, Giovanni, Matteo Romanello, and Frédéric Kaplan. 2018. "The References of References: A Method to Enrich Humanities Library Catalogs with Citation Data." *International Journal on Digital Libraries,* 19 (2–3): 151–61. https://doi.org/10.1007/s00799-017-0210-1.

Ehrmann, Maud, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. 2016. "Diachronic Evaluation of NER Systems on Old Newspapers." *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, 97–107. https://infoscience.epfl.ch/record/221391?ln=en.

2a. Title of the project

Please provide a short but specific title. Acronyms are allowed but not necessary.

Is your OCR good enough? A comprehensive assessment of the impact of OCR quality on downstream tasks.

2b. Abstract (max 250 words)

Please provide an abstract of your proposed research of maximum 250 words. Please note this abstract will be published on the KB Lab website and the KB website.

Is an average OCR quality of 70% enough for my study? What OCR quality should we ask from external suppliers? Should we re-do the OCR of our collections to bring it from 80% to 85%? Libraries and researchers alike face the same dilemma in our times of textual abundance: *when is OCR quality good enough?* User access, scientific results and the investment of limited resources increasingly depend on answering this question.

We propose to conduct a comprehensive assessment of the impact of OCR quality in Dutch newspaper, journal and book collections, comparing it with results from published corpora in English and French. This is done via *extrinsic evaluation*: assessing results from a set of representative downstream information retrieval, text mining and analysis tasks. These tasks are selected to include techniques providing advanced retrieval functionalities to library users or used by digital humanities and GLAM researchers: document ranking; topic modelling and document clustering; Named Entity Recognition (NER); document classification; vector space modelling (word embeddings).

The ultimate goal of the project is to release comprehensive guidelines detailing when OCR quality is to be considered good enough, in order to enable the informed development and use of textual collections.

3. Project description (max 1.500 words for sections 3a-c)

3a. Background and research question

Please give a short introduction on your project, building up to your research question.

The rise of digitization efforts within the library sector in recent decades has been short of "phenomenal" [30]. Digitization efforts mainly focus on materials with textual contents. The ENUMERATE report for the year 2017 states that 89% of heritage institutions part of the survey possess analogue text collections, and 55% digital ones. For libraries these numbers go up to 91% and 75% respectively. Libraries lead in terms of making their digital collections available to researchers and the general public [22]. In between the digitization and the use of textual collections, there is a critical step: Optical Character Recognition (OCR), or the extraction of text contents from images.

The importance of OCR cannot be underestimated. OCR allows digitized collections to be searched via full-text search, and mined to extract information. Most search and mining on digitized collections are performed using OCRed texts. Unfortunately, OCRed texts usually contain many errors. Particularly for historical texts and despite notable improvements over time [28], error rates can be very high, with largely unknown biasing consequences for end users [1,5,6,14,20,29,31]. Consequently, assessing and improving OCR quality has been and still is a key area for research and development [2,10,24,28].

Despite these known issues, to this day "there have been few efforts to think systematically or strategically about the problems of errorful OCR or to think creatively about what kinds of research are possible within our current paradigm". Only preliminary efforts have been made in order to assess the actual, practical impact of OCR errors on the use we make of OCRed corpora; efforts which would allow us to move beyond the dichotomy "clean transcriptions" and "dirty OCR" [28, p.10-11], and overcome the widespread lack of a quantified understanding of OCR quality [31]. Instead, many scholars still refuse to use or to trust OCRed texts in any way, while libraries strive to rise the OCR quality bar with limited empirical understanding of the benefits for doing so. Quite consequently, the very first recommendation of the 2018 research agenda on OCR by Smith and Cordell encourages statistical analyses of OCR outputs [28].

The research question we bring forth is thus the following: *when can we say that OCR quality is good enough?* The project focuses on what can be done with what we already

have, in a complementary way to ongoing OCR-related research focused on improving OCR results. Answering the proposed question is important for the following two reasons:

1) it would allow scholars and the public to use available OCRed resources in an informed way, instead of ignoring the OCR quality issue altogether or refuse to use data of less than perfect quality.
2) It would inform the strategic decisions of heritage institutions in terms of what OCR quality to aim for and if/when to re-OCR collections.

3b. Theoretical background

Please provide a brief theoretical background on your project

Most OCRed texts come with an estimation of their "quality". Typically, this quality is an average confidence metric from the OCR model which was used to perform OCR. This is an instance of *intrinsic OCR evaluation*, where we only rely on the OCR model to asses it(self). Such assessments are unsatisfactory because they might not be comparable when the software/provider changes and provide no indication on how the OCR quality influences other tasks or is related to other, external data or systems [13]. This is the broader scope of *extrinsic OCR evaluations*.

The simplest examples of extrinsic OCR evaluations are *dictionary lookup* and *word-error rates*. These methods are still widespread [26], yet they are problematic for highly volatile languages (i.e., changing a lot over time) such as Dutch.[1] More generally, extrinsic evaluations include any use of a dataset or task which takes as input the output of OCR, in order to assess the impact of OCR quality on the task itself. External evaluations are more involved but also more informative, because they allow to reason about the practical use of OCR outputs. Extrinsic evaluations require at least two versions of the same texts: one clean or high-quality, while the other is its OCRed version. Task results on the former are considered as the "ideal" result and are compared to task results on the latter version of the texts.

Some work has been published on extrinsic OCR evaluations as of late, with an almost exclusive focus on English and French. Studies have considered information access and retrieval [15,32], authorship attribution [11], Named Entity Recognition [12], topic modelling [21,23]. A recent and welcome study compared different tasks on a corpus in English: topic modelling, collocation analysis, authorship attribution and vector space modelling [13]. From this study, a critical OCR quality band between 70 and 80% emerged, where most tasks are unfeasible below it and instead provide good outcomes above it, with results within the band varying according to the task at hand.

These early results are still very preliminary. In particular, work on other languages than English and, to a lesser degree, French, basically still needs to start. *We here consider Dutch as the main language of focus, in view of comparing results with English and French corpora*. This would be the first multilingual comparison at this scale and scope, and a logical extension of the assessment of OCR quality which the KB already performs.[2] The comparison over languages will be complemented by a *comparison over time*, given that Dutch, English and French are all changing at varying degrees. Lastly, the range of tasks is still reduced in previous art, and comparisons between tasks are limited to a single paper [13]. *Comparing different tasks on the same data, as we propose here, is of critical importance to arrive at general conclusions.*

3c. Methods and techniques

[1] https://nl.wikipedia.org/wiki/Geschiedenis_van_de_Nederlandse_spelling.
[2] https://lab.kb.nl/about-us/blog/%E2%80%8Bnewspaper-ocr-quality-–-what-do-we-have-and-how-can-we-improve-it.

Please explain which methods or approach you will use to successfully complete your project

The proposed tasks to be compared in the extrinsic evaluation are:

1) *Document ranking*: considering basic keyword or Boolean queries and a standard full-text engine (e.g., Indri or Elasticsearch) as backend. The main goal would be to expand on previous related art [32], by comparing over time and language.
2) *Topic modelling and document clustering*: these unsupervised methods are used to group documents into clusters, in the absence of annotated labels [3,8,9].
3) *Document classification*: when annotated data is available, a model can be trained to predict the class of other documents [9,16,17].
4) *Named Entity Recognition (NER)*: part of the broader class of tasks named parsing or sequence labelling, NER is an enabler of the construction of knowledge bases including named entities such as persons, places and organizations [10,25].
5) *Vector space modelling*: in recent years, neural network-based methods have delivered notable advances in machine learning. Underpinning them are language models based on word embeddings: vector space representations of words and their meanings. Word embeddings can be used for similarity queries and clustering, and as features for other tasks (including document classification and NER, mentioned above). As such, this technique is widespread and of critical importance for text mining. The *impact of OCR quality on modern neural networks language models (word embeddings) remains unknown*. Methods here are based on neural networks and include the seminal Word2Vec [18,19] up to more recent context-based embedding methods such as BERT [7]. Importantly, methods which use character-level information, and which should thus be particularly resilient to OCR errors, will also be considered, such as FastText [16]. Vector space models will also be used and compared on the document ranking task as an alternative to standard full-text engines.

These tasks have been selected based on their importance for user search functionalities (e.g., in Delpher), their popularity among digital humanities practitioners and their complementarity.

There are two groups of datasets which are required for the project:

1. *Textual ground truth* including a typed and an OCRed version of the same texts (the OCRed version can be easily generated from images). The KB already provides for three datasets of the kind:
    a. https://www.esciencecenter.nl/project/deep-learning-ocr-post-correction.
    b. https://lab.kb.nl/dataset/ground-truth-impact-project.
    c. https://lab.kb.nl/dataset/dbnl-ocr-data-set.

    With respect to French and English, we plan to use a dataset from the NewsEye project [4].
2. *Annotations for the supervised tasks* of document classification (3) and NER (4). These data will need to be generated as part of the project, either by the applicant with the help of the KB team and/or via crowdsourcing using existing tools. Typically, not many annotations are required in order to run such evaluation [9]. Annotations will be assessed using inter-annotator agreement.

Once both datasets have been acquired, supervised tasks (3,4) will be compared in terms of relevant evaluation metrics (e.g., precision, recall and accuracy). Results from unsupervised tasks (1,2,5) will be compared between each other: taking the results from the clean text as "ideal", the results from the OCRed text will be compared and considered as decaying proportionally with how much the two solutions differ. A *comparison across time* for each task will be achieved by slicing the available data over time windows, in order to assess a) if the task performance on the ground truth varies purely due to language change and independently of OCR quality; b) if the impact of OCR quality also varies over time.

4. Outcomes

Please describe what the outcomes of the project will be & how you will use the KB data.

The envisioned project outcomes are the following:

- A **research paper** detailing the results for the Digital Humanities and GLAM communities. Possible venues for consideration include the Journal of Computing and Cultural Heritage or Digital Scholarship in the Humanities.
- A **conference submission** targeted to the libraries' community, e.g., the Joint Conference on Digital Libraries.
- A **KB technical report with guidelines and an associated blog post**, containing a summary of results and a set of technical guidelines for GLAM institutions dealing with OCR. The report should include practical guidance to answer questions such as what OCR quality to request to suppliers and when is OCR quality good enough.
- **Code and data** to replicate all results, under a CC-by attribution or MIT license.
- Less directly measurable but still quite tangible **knowledge exchange** between the applicant, the Living with Machines project and the KB team. All the envisioned outcomes should be ideally co-authored.

5. Link to the KB Research agenda

Please describe what the link is to the KB Research agenda. More information: https://www.kb.nl/organisatie/onderzoek-expertise/onderzoeksagenda-2018-2022

The project mainly links to the *Access and sharing* theme: it would provide key insights to inform the current and future development of KB's OCRed collections, as well as their use by general users and researchers. As testified by the fact that most previous KB Researcher-in-Residence projects have made use of OCRed data, the proposed project appears to be both timely and necessary. Furthermore, the project would improve the *Impact* of the KB by allowing it to strengthen its leading role in the research agenda on OCR and by publishing guidelines on OCR quality which would inform all GLAM institutions dealing with similar challenges in the Netherlands (which means most of them).

6. Workplan and timetable

Please describe: 1) how you will work together with the KB team, 2) where or if you would need assistance, and 3) a short overview of the work per 2 months.

The applicant already has the technical background necessary to carry out the proposed research. Constant collaboration with the KB team would be ideal, in particular on the following:

- defining the annotation tasks to collect ground truth and running them via crowdsourcing platforms and directly. Possibly, hosting an annotation tool such as INCEpTION https://inception-project.github.io.
- Interpreting and publishing results, especially with respect to writing guidelines for other libraries.
- Making code and derived datasets (e.g., ground truth) as open and reusable as possible.

It is worth noting that the importance of understanding the impact of OCR quality on text mining and analysis tasks has by now been realized by several research projects which are focusing on historical newspapers: NewsEye, Impresso, Oceanic Exchanges and Living with Machines, among others [27]. These projects are directly involving libraries, since historical newspapers make up for a large share of their textual collections. For example, the British Library hosts approximately 60M newspaper issues, of which 5%

have been digitized in ten years of work.[3] Living with Machines, a 5-year project based at The Alan Turing Institute and the British Library, is currently investing considerable efforts to assess the very same research question on newspaper data in English.

The applicant, as a former co-investigator and current visiting researcher on the project, would be uniquely positioned to broker that expertise and codebase to the proposed Dutch use-case. The applicant thus plans to directly involve and coordinate with the Living with Machines team as well as involve a student under his supervision from the Artificial Intelligence or Archives and Information masters at the University of Amsterdam.

Overview of work per two months

1. 1-2:
    a. Collect data.
    b. Design data pre-processing pipeline. Steps to consider include word segmentation and normalization, stemming and lowercasing, filtering.
    c. Pre-process data.
    d. Design and setup the annotation tasks, in coordination with KB and Living with Machines teams.
2. 3-4:
    a. Annotation tasks run.
    b. Development of code. Code developed by the Living with Machines team will be adapted where applicable.
    c. Start writing preliminary results.
3. 5-6:
    a. Run the analyses.
    b. Writeup of results.

7. References

Please add a list of references cited in the proposal, with full bibliographic details

1) Alex, Beatrice, Claire Grover, Ewan Klein, and Richard Tobin. 2012. "Digitised Historical Text: Does It Have to Be MediOCRe?" *Proceedings of the 9th Conference on Natural Language Processing (KONVENS 2012)*. http://www.oegai.at/konvens2012/proceedings/59_alex12w/59_alex12w.pdf.

2) Alex, Beatrice, and John Burns. 2014. "Estimating and Rating the Quality of Optically Character Recognised Text." In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH 2014)*. Madrid, Spain: ACM Press. https://doi.org/10.1145/2595188.2595214.

3) Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *The Journal of Machine Learning Research,* 3. http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

4) Chiron, Guillaume, Antoine Doucet, Mickael Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017. "Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information." In *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Toronto, ON, Canada: IEEE. https://doi.org/10.1109/JCDL.2017.7991582.

5) Cordell, Ryan. 2017. "'Q i-Jtb the Raven': Taking Dirty OCR Seriously." *Book History,* 20 (1): 188–225. https://doi.org/10.1353/bh.2017.0006.

6) Cordell, Ryan. 2019. "Why You (A Humanist) Should Care About Optical Character Recognition." https://ryancordell.org/research/why-ocr.

7) Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *ArXiv:1810.04805*. http://arxiv.org/abs/1810.04805.

8) Dieng, Adji B., Francisco J. R. Ruiz, and David M. Blei. 2019. "Topic Modeling in Embedding Spaces." *ArXiv:1907.04907 [Cs, Stat]*. http://arxiv.org/abs/1907.04907.

9) Eisenstein, Jacob. 2019. *Natural Language Processing*. Forthcoming for the MIT Press. https://github.com/jacobeisenstein/gt-nlp-class/tree/master/notes.

[3] https://blogs.bl.uk/thenewsroom/2019/01/heritage-made-digital-the-newspapers.html.

10) Ehrmann, Maud, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. 2016. "Diachronic Evaluation of NER Systems on Old Newspapers." *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*. https://infoscience.epfl.ch/record/221391?ln=en.

11) Franzini, Greta, Mike Kestemont, Gabriela Rotari, Melina Jander, Jeremi K. Ochab, Emily Franzini, Joanna Byszuk, and Jan Rybicki. 2018. "Attributing authorship in the noisy digitized correspondence of Jacob and Wilhelm Grimm." *Frontiers in Digital Humanities,* 5 (4). https://doi.org/10.3389/fdigh.2018.00004.

12) Hamdi, Ahmed, Axel Jean-Caurant, Nicolas Sidere, Mickael Coustaty, and Antoine Doucet. 2019. "An Analysis of the Performance of Named Entity Recognition over OCRed Documents." In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Champaign, IL, USA: IEEE. https://doi.org/10.1109/JCDL.2019.00057.

13) Hill, Mark J, and Simon Hengchen. 2019. "Quantifying the Impact of Messy Data on Historical Text Analysis." *Digital Scholarship in the Humanities*. https://researchportal.helsinki.fi/en/activities/quantifying-the-impact-of-messy-data-on-historical-text-analysis.

14) Jarlbrink, Johan and Pelle Snickars. 2017. "Cultural heritage as digital noise: Nineteenth century newspapers in the digital archive," *Journal of Documentation,* 73 (6): 1228–43. https://doi.org/10.1108/JD-09-2016-0106.

15) Joulain-Jay, Amelia T. 2017. *Corpus Linguistics for History: The Methodology of Investigating Place-name Discourses in Digitised Nineteenth-century Newspapers*. PhD thesis, Lancaster University.

16) Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. "Bag of Tricks for Efficient Text Classification." In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. https://aclweb.org/anthology/E17-2068.

17) Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. New York: Cambridge University Press.

18) Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *ArXiv:1301.3781*. https://arxiv.org/abs/1301.3781.

19) Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." In *Advances in Neural Information Processing Systems (NIPS)*. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.

20) Milligan, Ian. 2013. "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010." *Canadian Historical Review,* 94 (4): 540–69. https://doi.org/10.3138/chr.694.

21) Mutuvi, Stephen, Antoine Doucet, Moses Odeo, and Adam Jatowt. 2018. "Evaluating the Impact of OCR Errors on Topic Modeling." In *Maturity and Innovation in Digital Libraries*, edited by Milena Dobreva, Annika Hinze, and Maja Žumer, 11279. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-04257-8_1.

22) Nauta, Gerhard Jan, Wietske van den Heuvel, and Stephanie Teunisse. 2017. *Survey Report on Digitisation in European Cultural Heritage Institutions.* Technical Report, Europeana/ENUMERATE. For the Netherlands see, https://www.den.nl/uploads/5bdafec5e7395339faf97256b2ce41bc87cebed3a3dae.pdf.

23) Nelson, Laura K. 2017. "Computational grounded theory." *Sociological Methods & Research,* 19 (3). http://doi.org/10.1177/0049124117729703.

24) Nguyen, Thi-Tuyet-Hai, Adam Jatowt, Mickael Coustaty, Nhu-Van Nguyen, and Antoine Doucet. 2019. "Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing." In *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Champaign, IL, USA: IEEE. https://doi.org/10.1109/JCDL.2019.00015.

25) Lafferty, John, Andrew McCallum, and Fernando CN Pereira. 2001. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data." *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*. https://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers.

26) Pletschacher, Stefan, Christian Clausner, and Apostolos Antonacopoulos. 2014. "Europeana Performance Evaluation report." http://www.europeana-newspapers.eu/wp-content/uploads/2015/05/D3.5_Performance_Evaluation_Report_1.0.pdf.

27) Ridge, Mia, and Giovanni Colavizza. 2019. "The Past, Present and Future of Digital Scholarship with Newspaper Collections". *Digital Humanities Conference*. https://dev.clariah.nl/files/dh2019/boa/0377.html.

28) Smith, David A., and Ryan Cordell. 2018. "A Research Agenda for Historical and Multilingual Optical Character Recognition." http://hdl.handle.net/2047/D20297452.

29) Strange, Carolyn, Daniel McNamara, Josh Wodak, and Ian Wood. 2014. "Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers." *Digital Humanities Quarterly,* 8 (1). http://www.digitalhumanities.org/dhq/vol/8/1/000168/000168.html.

30) Terras, Melissa M. 2011. "The Rise of Digitization." In *Digitisation Perspectives*, edited by Ruth Rikowski, 39: 3–20. Rotterdam: SensePublishers. https://doi.org/10.1007/978-94-6091-299-3_1.

31) Traub, Myriam C., Jacco van Ossenbruggen, and Lynda Hardman. 2015. "Impact Analysis of OCR Quality on Research Tasks in Digital Archives." In *Research and Advanced Technology for Digital Libraries*, edited by Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla. Lecture Notes in Computer Science. Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-319-24592-8_19.

32) Traub, Myriam C., Thaer Samar, Jacco van Ossenbruggen, and Lynda Hardman. 2018. "Impact of Crowdsourcing OCR Improvements on Retrievability Bias." In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (JCDL)*. Fort Worth, Texas, USA: ACM Press. https://doi.org/10.1145/3197026.3197046.

8. Terms & Conditions

Please leave in 'Yes' when you agree with the terms and conditions

Yes, I agree to the Terms and Conditions of the program, including the condition that the KB may publish the title and abstract of my proposal on the KB Lab and KB.nl at the KB's sole discretion.

Name: Giovanni Colavizza

Place: Amsterdam, NL

Date: 27-09-2019