Impact Analysis of OCR Quality on Research Tasks in Digital Archives

Myriam C. Traub, Jacco van Ossenbruggen, and Lynda Hardman

Centrum Wiskunde & Informatica, Amsterdam
firstname.lastname@cwi.nl

Abstract. Humanities scholars increasingly rely on digital archives for their research instead of time-consuming visits to physical archives. This shift in research method has the hidden cost of working with digitally processed historical documents: how much trust can a scholar place in noisy representations of source texts? In a series of interviews with historians about their use of digital archives, we found that scholars are aware that optical character recognition (OCR) errors may bias their results. They were, however, unable to quantify this bias or to indicate what information they would need to estimate it. This, however, would be important to assess whether the results are publishable. Based on the interviews and a literature study, we provide a classification of scholarly research tasks that gives account of their susceptibility to specific OCR-induced biases and the data required for uncertainty estimations. We conducted a use case study on a national newspaper archive with example research tasks. From this we learned what data is typically available in digital archives and how it could be used to reduce and/or assess the uncertainty in result sets. We conclude that the current knowledge situation on the users' side as well as on the tool makers' and data providers' side is insufficient and needs to be improved.

Keywords: OCR Quality, Digital Libraries, Digital Humanities

1 Introduction

Humanities scholars use the growing numbers of documents available in digital archives not only because they are more easily accessible but also because they support new research tasks, such as pattern mining and trend analysis. Especially for old documents, the results of OCR processing are far from perfect. While improvements in pre-/post-processing and in the OCR technology itself lead to lower error rates, the results are still not error-free. Scholars need to assess whether the trends they find in the data represent real phenomena or result from tool-induced bias. It is unclear to what extent current tools support this assessment task. To our knowledge, no research has investigated how scholars can be supported in assessing the data quality for their specific research tasks.

In order to find out what research tasks scholars typically carry out on a digital newspaper archive (*RQ1*) and to what extent scholars experienced OCR quality to be an obstacle in their research, we conducted interviews with humanities scholars (Section 2). From the information gained in the interviews, we were able to classify the research tasks and describe potential impact of OCR quality on these tasks (*RQ2*). With a literature study, we investigated, how digitization processes in archives influence the OCR quality, how Information Retrieval (IR) copes with error-prone data and what workarounds scholars use to correct for potential biases (Section 3). Finally, we report on insights we gained from our use case study on the digitization process within a large newspaper archive (Section 4) and we give examples of what data scholars need to be able to estimate the quality indicators for different task categories (*RQ3*).

2 Interviews: Usage of Digital Archives by Historians

We originally started our series of interviews to find out what research tasks humanities scholars typically perform on digital archives, and what innovative additions they would like to see implemented in order to provide (better) support for these research tasks. We were especially interested in new ways of supporting quantitative analysis, pattern identification and other forms of distant reading. We chose our interviewees based on their prior involvement in research projects that made use of digital newspaper archives and / or on their involvement in publications about digital humanities research. We stopped after interviewing only four scholars, for reasons we describe below. Our chosen methodology was a combination of a structured personal account and a time line interview as applied by [4] and [5]. The former was used to stimulate scholars to report on their research and the latter to stimulate reflection on differences in tasks used for different phases of research. The interviews were recorded either during a personal meeting (*P1, P2, P4*) or during a Skype call (*P3*), transcribed and summarized. We sent the summaries to the interviewees to make sure that we covered the interviews correctly.

We interviewed four experts. (*P1*) is a Dutch cultural historian with an interest in representations of World War II in contemporary media. (*P2*) is a Dutch scholar specializing in modern European Jewish history with an interest in the implications of digital humanities on research practices in general. (*P3*) is a cultural historian from the UK, whose focus is the cultural history of the nineteenth century. (*P4*) is a Dutch contemporary historian who reported to have a strong interest in exploring new research opportunities enabled by the digital humanities.

All interviewees reported to use digital archives, but mainly in the early phases of their research. In the exploration phase the archives were used to get an overview of a topic, to find interesting research questions and relevant data for further exploration. In case they had never used an archive before, they would first explore the content the archive can provide for a particular topic (see Table 1, *E9*). At later stages, more specific searches are performed to find material about a certain time period or event. The retrieved items would later be used for close reading. For example, *P1* is interested in the representations of Anne

ID	Interview	Example	Category
E1	P1	Representation of Anne Frank in post-war media	T2
E3	P1	Contextualizing LDJ with sources used	T4
E4	P2	Comparisons of two digitized editions of a book to find differences in word use	T4
E5	P3	Tracing jokes through time and across newspapers	T3
E6	P3	Plot ngrams frequencies to investigate how ideas and words enter a culture	T1/T3
E7	P3	Sophisticated analysis of language in newspapers	T4
E8	P3	First mention of a newly introduced word	T1
E9	P3 /P4	Getting an overview of the archive's contents	T2
E11	P4	Finding newspaper articles on a particular event	T2

Table 1. Categorization of the examples for research tasks mentioned in the interviews. Task type *T1* aims to find the first mention of a concept. Tasks of type *T2* aim to find a subset with relevant documents. *T3* includes tasks investigating quantitative results over time and *T4* describes tasks using external tools on archive data.

Frank in post-war newspapers and tried to collect as many relevant newspaper articles as possible *E1*. *P3* reports on studies of introductions of new words into the vocabulary *E8*. Three of the interviewees (*P1, P3, P4*) mentioned that low OCR quality is a serious obstacle, an issue that is also reflected extensively in the literature [3, 6, 14]. For some research tasks, the interviewees reported to have come up with workarounds. *P1* sometimes manages to find the desired items by narrowing down search to newspaper articles from a specific time period, instead of using keyword search. However, this strategy is not applicable to all tasks.

Due to the higher error rate in old material and the absence of quality measures, they find it hard to judge whether a striking pattern in the data represents an interesting finding or whether it is a result of a systematic error in the technology. According to *P1*, the print quality of illegal newspapers from the WWII period is significantly worse than the quality of legal newspapers because of the conditions under which they were produced. As a consequence, it is very likely that they will suffer from a higher error rate in the digital archive, which in turn may cause a bias in search results. When asked how this uncertainty is dealt with, *P4* reported to try to explain it in the publications. The absence of error measures and information about possible preconceptions of the used search engine, however, made this very difficult. *P3* reported to have manually collected data for a publication to generate graphs tracing words and jokes over time (see *E5, E6* in Table 1) as the archive did not provide this functionality. Today, *P3* would not trust the numbers enough to use them for publications again.

P2 and *P3* stated that they would be interested in using the data for analysis independently from the archive's interfaces. Tools for text analysis, such as Voyant[1], were mentioned by both scholars (*E3, E4, E7*). The scholars could not indicate how such tools would be influenced by OCR errors. We asked the scholars whether they could point out what requirements should be met in order

[1] http://voyant-tools.org/

to better facilitate research tasks in digital archives. *P3* thought it would be impossible to find universal methodological requirements, as the requirements vary largely between scholars of different fields and their tasks.

We classified the tasks that were mentioned by the scholars in the interviews according to their similarities and requirements towards OCR quality. The first mention of a concept, such as a new word or concept would fall into category *T1*. *T2* comprises tasks that aim to create a subcollection of the archive's data, e.g. to get to know the content of the archive or to select items for close reading. Tasks that relate word occurrences to a time period or make comparisons over different sources or queries are summarized in *T3*. Some archives allow the extraction of (subsets of) the collection data. This allows the use of specialized tools, which constitutes the last category *T4*.

We asked *P1*, *P2* and *P4* about the possibilities of more quantitative tools on top of the current digital archive, and in all cases the interviewees' response was that no matter what tools were added by the archive, they were unlikely to trust any quantitative results derived from processing erroneous OCRed text. *P2* explicitly stated that while he did publish results based on quantitative methods in the past, he would not use the same methods again due to the potential of technology-induced bias.

None of our interviews turned out to be useful with respect to our quest into innovative analysis tools. The reason for this was the perceived low OCR quality, and the not well-understood susceptibility of the interviewees' research tasks to OCR errors. Therefore, we decided to change the topic of our study to better understanding the impact of OCR errors on specific research tasks. We stopped our series of interviews and continued with a literature study on the impact of OCR quality on specific research tasks.

3 Literature study: Impact of OCR Quality on Scholarly Research

To find out how the concerns of the scholars are addressed by data custodians and by research in the field of computer science, we reviewed available literature.

The importance of OCR in the digitization process of large digital libraries is a well-researched topic [9, 12, 18, 19]. However, these studies are from the point of view of the collection owner, and not from the perspective of the scholar using the library or archive. User-centric studies on digital libraries typically focus on user interface design and other usability issues [8, 20, 21]. To make the entry barrier to the digital archive as low as possible, interfaces often try to hide technical details of the underlying tool chain as much as possible. While this makes it easier for scholars to use the archive, it also denies them the possibility to investigate potential tool-induced bias.

There is ample research into how to reduce the error rates of OCRed text in a post-processing phase. For example, removing common errors, such as the "long s"-to-f confusion or the soft-hyphen splitting of word tokens, has shown to improve Named Entity Recognition. This, however, did not increase the overall quality to a sufficient extent as it addressed only 12% of the errors in the chosen sample [2]. Focusing on overall tool performance or performance on representative samples of the entire collection, such studies provide little information on the impact of OCR errors on specific queries carried out on specific subsets of a collection. It is this specific type of information we need, however, to be able to estimate the impact on our interviewees' research questions. We found only one study that aimed at generating high-quality OCR data and evaluating the impact of its quality on a specific set of research questions [15]. The researchers found that the impact of OCR errors is not substantial for a task that compares two subsets of the corpus. For a different task, the retrieval of a list of the most significant words (in this case, describing moral judgement), however, recall and precision were considered too low.

Another line of research focuses on how to improve OCR tools or on using separate tools for improving OCR output in a post-processing step [11], for example by using input from the public [10]. Unfortunately, the actual extent, to which this crowdsourcing initiative has contributed to a higher accuracy has not been measured. While effective use of such studies may reduce the error rate, they do not help to better estimate the impact of the remaining errors on specific cases. Even worse, since such tools (and especially human input) add another layer of complexity and potential errors, they may also add more uncertainty to these estimates. Most studies on the impact of OCR errors are in the area of ad-hoc IR, where the consensus is that for long texts and noisy OCR errors, retrieval performance remains remarkably good for relatively high error rates [17]. On short texts, however, the retrieval effectiveness drops significantly [7, 13]. In contrast, information extraction tools suffer significantly when applied to OCR output with high error rates [16]. Studies carried out on unreliable OCR data sets often leave the OCR bias implicit. Some studies explicitly protect themselves from OCR issues and other technological bias by averaging over large sets of different queries and by comparing patterns found for a specific query set to those of other queries sets [1]. This method, however, is not applicable to the examples given by our interviewees, since many of their research questions are centered around a single or small number of terms.

Many approaches aiming at improving the data quality in digital archives have in common that they partially reduce the error rate, either by improving overall quality, or by eliminating certain error types. None of these approaches, however, can remove all errors. Therefore, even when applying all of these steps to their data, scholars still need to be able to quantify the remaining errors and assess their impact on their research tasks.

4 Use case: OCR Impact on Research Tasks in a Newspaper Archive

To study OCR impact on specific scholarly tasks in more detail, we investigated OCR-related issues of concrete queries on a specific digital archive: the historic

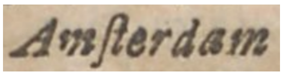

Fig. 1. Confusing the "long s" for an "f" is a common OCR error in historic texts.

newspaper archive[2] of the National Library of The Netherlands (KB). It contains over 10 million Dutch newspaper pages from the period 1618 to 1995, which are openly available via the Web. For each item, the library publishes the scanned images, the OCR-ed texts and the metadata records. Its easy access and rich content make the archive an extremely rich resource for research projects[3].

4.1 Task: First mention of a concept

One of the tasks often mentioned during our interviews was finding the first mention of a term (task *T1* in Section 2). For this task, scholars can typically deal with a substantial lack of precision caused by OCR errors, since they can detect false positives by manually checking the matches. The key requirement is recall. Scholars want to be sure that the document with the first mention was not missed due to OCR errors. This requires a 100% recall score, which is unrealistic for large digital archives. As a second best, they need to minimize the risk of missing the first mention to a level that is acceptable in their research field. The question remains how to establish this level, and to what extent archives support achieving this level. To understand how a scholar could assess the reliability of their results with currently available data, we aim to find the first mention of "Amsterdam" in the KB newspaper archive. A naive first approach is to simply order the results on the query "Amsterdam" by publication date. This returned a newspaper dated October 25, 1642 as the earliest mention. We then explore different methods to assess the reliability of this result. We first tried to better understand the corpus and the way it was produced, then we tried to estimate the impact of the OCR errors based on the confidence values reported by the OCR engine, and finally we tried to improve our results by incremental improvement our search strategy.

Understanding the digitization pipeline We started by obtaining more information on the archive's digitization pipeline, in particular details about the OCR process, and potential post-processing steps.

Unfortunately, little information about the pipeline is given on the KB website. The website warns users that the OCR text contains errors[4], and as an example mentions the known problem of the "long s" in historic documents (see Fig. 1), which causes OCR software to mistake the 's' for an 'f'. The page does not provide quantitative information on OCR error rates.

[2] `www.delpher.nl/kranten`

[3] See `lab.kbresearch.nl` for examples.

[4] `http://www.delpher.nl/nl/platform/pages/?title=kwaliteit+(ocr)`

After contacting library personnel, we learned that formal evaluation on OCR error rates or on precision/recall scores of the archive's search engine had not been performed so far. The digitization had been a project spanning multiple years, and many people directly involved no longer worked for the library. Parts of the process had been outsourced to a third party company, and not all details of this process are known to the library. We believe this practice is typical for many archives. We further learned that article headings had been manually corrected for the entire archive, and that no additional error correction or other post-processing had been performed. We concluded that for the first mention task, our inquires provided insufficient information to be directly helpful.

Uncertainty estimation: using confidence values Next, we tried to use the confidence values reported by the OCR engine to assess the reliability of our result. The ALTO XML[5] files used to publish the OCR texts do not only contain the text as it was output by the OCR engine, they also contain confidence values generated by the OCR software for each page, word and character. For example, this page[6], contains:

```
<Page ID=”P2” ... PC=”0.507”>
```

Here, *PC* is a confidence value between 0 (low) and 1 (high confidence). Similar values are available for every word and character in the archive:

```
<String ID=”P2_ST00800” ... CONTENT=”AM” ...
        SUBS_CONTENT=”AMSTERDAM.” WC=”0.45” CC=”594”/>
<String ID=”P2_ST00801” ... CONTENT=”STERDAM.” ...
        SUBS_CONTENT=”AMSTERDAM.” WC=”0.30” CC=”46778973”/>
```

Here, *WC* is the word-level confidence, again expressed as a value between 0 and 1. CC is the character-level confidence, expressed as a string of values between 0-9, with one digit for each character. In this case, 0 indicates high, and 9 indicates low confidence. This is an example for a word that was split by a hyphen. The representation of its two parts as “subcontent” of “AMSTERDAM” assures its retrieval by the search engine of delpher.

```
<String ID=”P2_ST00766” ... CONTENT=”Amfterdam,”
       WC=”0.36” CC=”0866869771”/>
```

For the last example, this means the software has lower confidence in the correct “m”, than in the incorrect “f”. Note that since the above XML data is available for each individual word, it is a huge dataset in absolute size, that could, potentially, provide uncertainty information on a very fine-grained level. For this, we need to find out what these values mean and/or how they have been computed. However, the archive's website provides no information about how the confidence values have been calculated.

[5] `http://www.loc.gov/standards/alto/`

[6] `http://resolver.kb.nl/resolve?urn=ddd:010633906:mpeg21:p002:alto`

Category	Confusion matrix	CV output	CV alternatives
available for:	sample only	full corpus	not available
T1 1st mention of x	find all queries for x, impractical	estimated precision not helpful	improve recall
T2 Selecting subset relevant to x	as above	estimated precision, requires improved UI	improve recall
T3. Pattern over time x	pattern summarized over set of alt. queries	estimates of corrected precision	estimates of corrected recall
T3.a Compare x_1 and x_2	warn for diff. susceptibility to errors	as above, warn for diff. distribution of CVs	as above
T3.b Compare $corpus_1$ and $corpus_2$	as above	as above	as above

Table 2. The different types of tasks require different levels of quality. Quality indicators can be used to generate better estimates of the quality and also (to some extent) to compensate low quality. x stands for an abstract concept that is the focus of interest in the research task.

Again, from the experts in the library, we learned that the default word level confidence scores were increased if the word was found in a given list with correct Dutch words. Later, this was improved by replacing the list with contemporary Dutch words by a list with historic spelling. Unfortunately, it is not possible to reproduce which word lists have been used on what part of the archive.

Another limitation is that even if we could calibrate the OCR confidence values to meaningful estimates, they could only be used to estimate how many of the matches found are likely false positives. They provide little or no information on the false negatives, since all confidence values related to characters that were considered as potential alternatives to the character chosen by the OCR engine have not been preserved in the output and are lost forever. For this research task, this is the information we would need to estimate or improve recall. We thus conclude that we failed in using the confidence values to estimate the likelihood that our result indeed represented the first mention of “Amsterdam” in the archive. We summarized our output in Table 2, where for *T1* we indicate that using the confusion matrix is impractical, using the out confidence values (CV output) is not helpful, and using the confidence values of the alternatives (CV alternatives) could have improved recall, but we do not have the data.

Incremental improvement of the search strategy We observed that the “long s” warning given on the archive’s website is directly applicable to our query. Therefore, to improve on our original query, we also queried for “Amfterdam”. This indeed results in an earlier mention: July 27, 1624. This result, however, is based on our anecdotal knowledge about the “long s problem”. It illustrates the need for a more systematic approach to deal with spelling variants. While the archive provides a feature to do query expansion based on historic spelling variants, it provides no suggestions for “Amsterdam”. Querying for known spelling variants mentioned on the Dutch history of Amsterdam Wikipedia page also did result in earlier mentions.

To see what other OCR-induced misspellings of Amsterdam we should query for, we compared a ground truth data set with the associated OCR texts. For this, we used the dataset[7] created in the context of the European IMPACT project. It includes a sample of 1024 newspaper pages, but these had not been completely finished by end of the project. This explains why this data has not been used in a evaluation of the archive's OCR quality. Because of changes in the identifier scheme used, we could only map 265 ground truth pages to the corresponding OCR text in the archive. For these, we manually corrected the ground truth for 134 pages, and used these to compute a confusion table[8]. This matrix could be used to generate a set of alternative queries based on all OCR errors that occur in the ground truth dataset. Our matrix contains a relatively small number of frequent errors, and it seems doable to use them to manually generate a query set that would cover the majority of errors. We decided to look at the top ten confusions and use the ones applicable to our query. All combinations of confusions resulted in 23 alternative spelling variations of "Amsterdam". When we queried for the misspellings, we found hits for all variations, except one, "Amfcordam". None, however, yielded an earlier result than our previous query.

This method could, however, be implemented as a feature in the user interface, the same way as historic spelling variants are supported[9]. Again, the issue is that for a specific case, it is hard to predict whether such a future would help, or merely provide more false positives.

Our matrix also contains a very long tail with infrequent errors, and for this specific task, it is essential to take all of them into account. This makes our query set very large and while this may not be a technical problem for many state of the art search engines, the current user interface of the archive does not support such queries. More importantly, the long tail also implies that we need to assume that our ground truth does not cover all OCR errors that are relevant for our task.

We conclude that while the use of a confusion matrix does not guarantee finding the first mention of a term, it would be useful to publish such a matrix on each digital archive's website. Just using the most frequent confusions can already help user to avoid the most frequent errors, even in a manual setting. Systematic queries for all known variants would require more advanced backend support.

Fortunately, it lies in the nature of our task that with every earlier mention we can confirm, we can also narrow the search space by defining a new upper bound. In our example, the dataset with pages published before our 1624 upper bound is sufficiently small to allow manual inspection. The first page in the archive of the same title as the 1624 page, is published in 1619, and has a mention

[7] `lab.kbresearch.nl/static/html/impact.html`

[8] available on `http://dx.doi.org/10.6084/m9.figshare.1448810`

[9] `http://www.delpher.nl/nl/platform/pages/?title=zoekhulp`

of "Amsterdam". It is on the very bottom of the page in a sentence that is completely missing in the OCR text. This explains why our earlier strategy has missed it. The very earliest page in the archive at the time of writing is from June 1618. Its OCR text contains "Amfterftam". Our earlier searches missed this one because it is a very rare variant which did not occur in the ground truth data. While we now have found our first mention in the archive with 100% certainty, we found it by manual, not automatic means. Our strategy would not have worked when the remaining dataset would have been too large to allow manual inspection.

4.2 Analysis of other tasks

We also analyzed the other tasks in the same way. For brevity, we only report our findings to the extent they are different from task *T1*. For *T2*, selecting a subset on a topic for close reading, the problem is that a single random OCR error might cause the scholar to miss a single important document as in *T1*. In addition, a systematic error might result in a biased selection of the sources chosen for close reading, which might be an even bigger problem. Unfortunately, using the confusion matrix is again not practical. The CV output could be useful to improve precision for research topics where the archive contains too many relevant hits, and selecting only hits above a certain confidence threshold might be useful. This requires, however, the user interface to support filtering on confidence values. For the CV alternatives, they again could be used to improve recall, but it is unclear against what precision.

For task *T3*, plotting frequencies of a term over time, the issue is no longer whether or not the system can find the right documents, as in *T1* and *T2*, but if the system can provide the right counts of term occurrences despite the OCR errors. Here, the long tail of the confusion matrix might be less of a problem, as we may choose to only query for the most common mistakes, assuming that the pattern in the total counts will not be affected much by the infrequent ones. CV output could be used to lower counts for low precision results, while CV alternatives could be used to increase counts for low recall matches. For *T3.a*, a variant of *T3* where the occurrence over time of one term is compared to another, the confusion matrix could also be used to warn scholars if one term is more susceptible to OCR errors than the other. Likewise, a different distribution of the CV output for the two terms might be flagged in the interface to warn scholars about potential bias. For *T3.b*, a variant where the occurrence of a term in different newspapers is analyzed, the CV values could likely be used to indicate different distributions in the sources, for example to warn for systematic errors caused by differences in print quality or fonts between the two newspapers.

For task *T4* (not in the table), the use of OCRed texts in other tools, our findings are also mainly negative. Very few text analysis tools can, for example, deal with different confidence values in their input, apart from the extensive standardization these would require for the input/output formats and interpretation of these values. Additionally, many tools suffer from the same limitation that only their overall performance on a representative sample of the data has been evaluated, and little is known about their performance on a specific use case outside that sample. By stacking this uncertainty on top of the uncertain OCR errors, predicting its behavior for a specific case will be even harder.

5 Conclusions

Through interviews we conducted with scholars, we learned that while the uncertain quality of OCRed text in archives is seen as a serious obstacle to wider adaption of digital methods in the humanities, few scholars can quantify the impact of OCR errors on their own research tasks. We collected concrete examples of research tasks, and classified them into categories. We analyzed the categories for their susceptibility to OCR errors, and illustrated the issues with an example attempt to assess and reduce the impact of OCR errors on a specific research task. From our literature study, we conclude that while OCR quality is a widely studied topic, this is typically done in terms of tool performance. We claim to be the first to have addressed the topic from the perspective of impact on specific research tasks of humanity scholars.

Our analysis shows that for many research tasks, the problem cannot be solved with better but still imperfect OCR software. Assessing the impact of the imperfections on a specific use case remains important.

To improve upon the current situation, we think the communities involved should begin to approach the problem from the user perspective. This starts with understanding better how digital archives are used for specific tasks, by better documenting the details of the digitization process and by preserving all data that is created during the process. Finally, humanity scholars need to transfer their valuable tradition of source criticism into the digital realm, and more openly criticize the potential limitations and biases of the digital tools we provide them with.

Acknowledgements

We would like to thank our interviewees for their contributions, the National Library of The Netherlands for their support and the reviewers for their helpful feedback. This research is funded by the Dutch COMMIT/ program.

References

1. A. Acerbi, V. Lampos, P. Garnett, and R. A. Bentley. The expression of emotions in 20th century books. *PLoS ONE*, 8(3):e59030, 03 2013.
2. B. Alex, C. Grover, E. Klein, and R. Tobin. Digitised historical text: Does it have to be mediOCRe? In J. Jancsary, editor, *Proceedings of KONVENS 2012*, pages 401–409. ÖGAI, September 2012. LThist 2012 workshop.
3. A. Bingham. The digitization of newspaper archives: Opportunities and challenges for historians. *Twentieth Century British History*, 21(2):225–231, 2010.

4. M. Bron. *Exploration and Contextualization through Interaction and Concepts.* PhD Thesis. 2013.
5. C. D. Brown. Straddling the humanities and social sciences: The research process of music scholars. *Library & Information Science Research*, 24(1):73 – 94, 2002.
6. D. J. Cohen and R. Rosenzweig. *Digital history: A guide to gathering, preserving, and presenting the past on the web*, volume 28. University of Pennsylvania Press Philadelphia, 2006.
7. W. B. Croft, S. Harding, K. Taghva, and J. Borsack. An evaluation of information retrieval accuracy with simulated OCR output. Technical report, Amherst, MA, USA, 1993.
8. N. Fuhr, P. Hansen, M. Mabe, A. Micsik, and I. Slvberg. Digital libraries: A generic classification and evaluation scheme. In P. Constantopoulos and I. Slvberg, editors, *Research and Advanced Technology for Digital Libraries*, volume 2163 of *Lecture Notes in Computer Science*, pages 187–199. Springer Berlin Heidelberg, 2001.
9. R. Holley. How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine*, 15(3/4), 2009.
10. R. Holley. Many hands make light work: Public collaborative OCR text correction in Australian Historic Newspapers. Technical report, National Library of Australia, Mar. 2009.
11. K. Kettunen, T. Honkela, K. Lindén, P. Kauppinen, T. Pääkkönen, J. Kervinen, et al. Analyzing and improving the quality of a historical news collection using language technology and statistical machine learning methods. In *IFLA World Library and Information Congress Proceedings 80th IFLA General Conference and Assembly*, 2014.
12. E. Klijn. The current state-of-art in newspaper digitization a market perspective, January 2008.
13. E. Mittendorf and P. Schäuble. Information retrieval can cope with many errors. *Inf. Retr.*, 3(3):189–216, Oct. 2000.
14. B. Nicholson. Counting culture; or, how to read Victorian newspapers from a distance. *Journal of Victorian Culture*, 17(2):238–246, 2012.
15. C. Strange, D. McNamara, J. Wodak, and I. Wood. Mining for the meanings of a murder: The impact of OCR quality on the use of digitized historical newspapers. *Digital Humanities Quarterly*, 8(1), 2014.
16. K. Taghva, R. Beckley, and J. Coombs. The effects of OCR error on the extraction of private information. In H. Bunke and A. Spitz, editors, *Document Analysis Systems VII*, volume 3872 of *Lecture Notes in Computer Science*, pages 348–357. Springer Berlin Heidelberg, 2006.
17. K. Taghva, J. Borsack, A. Condit, and S. Erva. The effects of noisy data on text retrieval. *J. Am. Soc. Inf. Sci.*, 45(1):50–58, Jan. 1994.
18. S. Tanner, T. Muñoz, and P. H. Ros. Measuring mass text digitization quality and usefulness. *D-Lib Magazine*, 15(7/8):1082–9873, 2009.
19. A. Weymann, R. A. Luna Orozco, C. Mueller, B. Nickolay, J. Schneider, and K. Barzik. *Einführung in die Digitalisierung von gedrucktem Kulturgut - Ein Handbuch für Einsteiger.* Ibero-American Institute (Berlin), 2010.
20. H. I. Xie. Evaluation of digital libraries: Criteria and problems from users' perspectives. *Library & Information Science Research*, 28(3):433 – 452, 2006.
21. H. I. Xie. Users' evaluation of digital libraries (dls): Their uses, their criteria, and their assessment. *Inf. Process. Manage.*, 44(3):1346–1373, May 2008.