Jørgen Burchardt*

Are Searches in OCR-generated Archives Trustworthy?

Sind Recherchen in OCR-generierten Archiven vertrauenswürdig?

An Analysis of Digital Newspaper Archives

Eine Analyse digitaler Zeitungsarchive

https://doi.org/10.1515/jbwg-2023-0003

Abstract: Digitised archives are revolutionary tools for research that, in a few seconds, generate results that earlier often took years to obtain. But do they provide all results for the terms searched for? The accuracy of searches was tested by performing sample searches of leading newspaper databases. The test revealed several weaknesses in the search process, including an average 18 percent error rate for single words in body text, and a far higher error rates for advertisements. Such high error rates encourage a critical look at the 20-year-old sector. Although these errors can be reduced by a re-digitation and with new improved OCR engines and new search algorithms, searches will nevertheless return manipulated results. In response, and to identify infringed bias and skewed representation, database owners need to provide thorough metadata to ensure source criticism.

JEL-Codes: C 82

Keywords: optical character recognition, historical archive, source criticism, research methodology, Historische Archive, Quellenkritik, Forschungsmethodik, OCR

1 Introduction

Can we trust searches performed in archives generated by optical character recognition (OCR)? This article presents enough examples of flawed results

***Corresponding author: Jørgen Burchardt**, Nyborgvej 13, DK-5750 Ringe,
E-mail: jorgen.burchard@mail.dk

produced by the technology to suggest that the short answer to the question is a resounding *no*.

The long answer to the question, however, is more complicated, and this article gives interested historians and other readers an overview of OCR's development during the past 20 years as well as some predictions and aspirations for the technology's future. It also affords insight into the technological world behind OCR, knowledge that is essential for historians to provide traditional source criticism.

Today, information is born digital. As a result, the possibility of finding a desired piece of information is nearly 100 percent. Search engines such as Google Search are known the world over for returning results from an almost endless archive in fractions of a second.

A few decades ago, however, to access old newspapers historians had to visit sufficiently large libraries that were few and far between. Even once there, it could take hours for librarians to acquire the correct year and edition of a requested newspaper. Beyond that, without knowing the correct date, historians had to meticulously leaf through newspaper pages for hours, if not days.

In Britain, for instance, only *The Times* had a comprehensive index to assist researchers. However, without a specifically dated event, they would have to scan through countless newspaper pages in search of relevant material.[1] Although microfilms of newspapers afforded some relief – librarians could distribute the microfilm copies to other libraries – users of the technology can still hear the noise from the microfilm reels when searching for specific dates.

2 Years of Rapid Technological Development

For more than 130 years, many inventors tried to construct a system to translate written letters to another medium. In 1955, a commercial product helped the publisher of *Reader's Digest* to read typewritten documents. During the 1960s, more than 100 different models of commercial OCR systems were developed. The first users were companies with special needs, like banks, statistical offices, and post offices, who were often reading standardized documents with the

1 *A. Bingham*, The Digitization of Newspaper Archives: Opportunities and Challenges for Historians, in: Twentieth Century British History 21/2, 2010, pp. 225-231.

same format. Into the 1970s, page readers for multifonts and loose formats of varying pages were still expensive and large.[2]

By the end of the 1970s, the technology had reached maturity. Thanks to microcomputers, optical techniques and efficient software, the cost to process and store information started to decline sharply.

In the late 1980s, when the technology had become fairly developed, libraries saw possibilities for using the system to OCR-read old books and newspapers. One of the most useful OCR programmes at the time was FineReader, introduced by ABBYY in 1993. Since then, the development of OCR has been relatively intense, and today, we are close to obtaining acceptable results with handwritten text.

Many of the first OCR programmes were primitive and could only read unique fonts – for example, the custom-made monospace fonts OCR-A and OCR-B or typewriter fonts – which was nevertheless helpful when automating businesses. The programmes were improved significantly for use in digitising administrative tasks, which granted the programmes considerable accuracy in reading letters made with typewriters.

Many newspapers in Europe in the 19th and early 20th centuries were printed in a Gothic typeface, a font quite unlike a typewriter's typeface. However, because the demand for the OCR-reading of old texts was slight compared with the demand among big business, OCR engines were not built to handle the Gothic typeface. Overcoming that obstacle has thus been a major task for improving OCR engines, and even today, many years of development remain necessary before acceptable quality can be achieved.[3]

For many decades, historical archives had microfilmed large parts of their newspaper collections, thereby making digitisation economical. Small but successful projects in the 1990s blossomed into online archives in the 2000s and, since then, into their current state as well-organised online libraries affording 24/7 access to extensive collections.

The *Atlas of Digitised Newspapers and Metadata* offers an overview of the state of OCR-generated digitised archives as of 2020 and possible ways forward

2 *H.F. Schantz,* The History of OCR, Optical Character Recognition. Recognition Technologies Users Association, Manchester Center, 1982, p. 11.

3 *K. Kettunen/M. Koistinen,* Open Source Tesseract in re-OCR of Finnish Fraktur from 19th and early 20th Century Newspapers and Journals, in: DHN 2019: Digital Humanities in the Nordic Countries: Proceedings of the Digital Humanities in the Nordic Countries, 4th Conference, Copenhagen 2019, pp. 270-282, http://ceur-ws.org/Vol-2364/25_paper.pdf, 26.10.2022.

for ten such archives.[4] A fruitful source of information on the complicated technical, organisational, and legal world behind each digital archive, the *Atlas* was conceived as a spin-off of a research project aimed at producing a global history. It shows that even the most open, service-minded archives still cannot deliver ambitious researchers easy access to raw data in its database.

The historical archives have grown from national projects and, in some countries, become public libraries late into the business, thereby leaving room for private companies. In Britain, Gale Cengage, then Thomson Gale Publishers, debuted *The Times* Digital Archive in 2002, a database containing issues of the world-class newspaper published since 1785.[5]

The British Library established its archive from 2007 as a fully searchable resource. Gale Cengage supported the development of the British Library's infrastructure in return for a licence to commercialise the content for markets outside UK higher and continuing education.[6]

A similar development with private actors took place in the United States. NewspaperArchive came online in 1999 as the first archive and today boasts having nearly 16,000 titles from the United States and 28 other countries.[7] In its digital collection, ProQuest similarly has not only newspapers from the United States but some international papers as well. Despite offering access to only 60 newspapers, among them are some of the most prominent, including *The Wall Street Journal*, *The New York Times* and *The Washington Post*.[8] The Ancestry-owned Newspapers.com reports giving access to 21,000 papers,[9] whereas the Library of Congress has only 3,783 newspapers in its free public archive.[10]

Researchers in other countries, by contrast, enjoy greater as well as free access to national newspapers. For instance, in Australia and the Netherlands, many journals are digitised and accessible up to the last decade in the respective national archives of Trove and Delpher. At the same time, due to copyright concerns, many other countries restrict access to their archives. For example, in

4 *M. Beals/E. Bell*, The Atlas of Digitised Newspapers and Metadata. Reports from Oceanic Exchanges, Loughborough 2020.

5 *Ibid.*, p. 28.

6 *P. Fleming/E. King*, The British Library Newspaper Collections and Future Strategy, in: Interlending & Document Supply 37/4, 2009, pp. 223-228, and *T. Hauswedell et al.*, Of Global Reach Yet of Situated Context: an Examination of the Implicit and Explicit Selection Criteria that Shape Digital Archives of Historical Newspapers, in: Archival Science 20, 2020, pp. 139-165.

7 newspaperarchive.com/about-us, 08.08.2022.

8 about.proquest.com/en/products-services/pq-hist-news, 08.08.2022.

9 http://www.newspapers.com/papers, 08.08.2022.

10 chroniclingamerica.loc.gov/newspapers, 08.08.2022.

Austria, the national archive has only journals 70 years old in its Austrian Newspapers Online (ANNO) archive;[11] in Denmark, meanwhile, Mediestream under the Royal Library has a limitation set to 100 years.[12]

Digital archives are not only useful for interested citizens and researchers but also important heritage projects because the original sources are no longer used, and the wear and tear has stopped. Nevertheless, the development of such archives remains in its infancy, and there is still a long way to go until the total coverage of all journals is accessible to researchers. Even a well-organised archive such as Delpher has digitised only 15 percent of newspapers from the Netherlands and its colonies.[13]

3 Historians Use of Digital Archives

To be sure, digitisation has granted greater access to newspapers. No longer is visiting a physical library necessary given that an internet connection allows access from anywhere. Moreover, their limited opening hours no longer prevent immediate access, as users now enjoy access 24/7.

Among other types of users, historians have gained several new possibilities thanks to such media. Many new digital methodologies have become possible with the support of searches in large newspaper archives. Now, the first instance of an event can be dated, and the distribution of knowledge throughout a country is entirely possible.

As such, digital archives are no less than a gift to historical researchers. British historian Adrian Bingham has described the new possibilities as follows:

> "Historians can be far more confident that content will not elude them and that they will track down obscure and potentially revealing articles. Biographers can ensure that they have read every mention of their subject in the press. [...] digitization will doubtless lead to the unearthing of fresh biographical material."[14]

In particular, he has forecasted an especially rosy future for quantitative historical studies, as

11 anno.onb.ac.at/node/11, 08.08.2022.

12 www2.statsbiblioteket.dk/mediestream/info/adgang, 08.08.2022.

13 www.delpher.nl/over-delpher/wat-zit-er-in-delpher/wat-zit-er-in-delpher/kranten#cc362, 09.08.2022.

14 *Bingham,* The Digitization of Newspaper Archives, p. 228.

> "increasingly sophisticated tools are becoming available for 'text-mining' [searching and comparing very large quantities of text using algorithms, statistical formulae and language processing techniques] and many humanities scholars may find themselves increasingly working with technical experts to exploit the full potential of those archives".[15]

New types of research have indeed become possible, and, as a result, unprecedented academic cooperation has emerged. One area of study is the quantitative analysis of large databases across centuries that allows the identification of macroscopic patterns of cultural change. Some researchers in that area of study meet at conferences on computational history attended by scholars from various disciplines within history, computer science and associated disciplines, as well as cultural heritage researchers. Meanwhile, in the archives themselves, historians can study and process digitised sources, computer scientists can discuss experimental tools and evaluate methods for gauging their relevance to real-world questions and applications, and librarians can contribute with suggestions on the organisation and dissemination of historical archives.

In 1987, Roberto Franzosi predicted the significant use of quantitative data from journals as long as researchers overcame problems with reliability and data validation. 30 years later, he returned with a suggestion to study narrative history through quantitative analysis without losing track of the event itself or the people behind the numbers.[16] In that light, data from different newspapers can yield information from and about geographically disperse areas that databases on books simply cannot.[17]

Digital archives also illuminate new areas for research. For example, it is now possible to analyse long-term dynamic developments in newspaper content in connection with societal segmentation. One researcher has even examined changes in newspapers' censorship of violent deaths over time.[18] In general, because newspapers use language, the text therein can be analysed and used to

15 *Ibid.*, p. 229.

16 *R. Franzosi*, The Press as a Source of Socio-Historical Data: Issues in the Methodology of Data Collection from Newspapers, in: Historical Methods 20/1, 1987, pp. 5-16; *Idem*, A Third Road to the Past? Historical Scholarship in the Age of Big Data, in: Historical Methods 50/4, 2017, pp. 227-244.

17 *T. Lansdall-Welfare et al.*, Content Analysis of 150 Years of British Periodicals, in: Proceedings of the National Academy of Sciences 114/4, 2017, https://www.pnas.org/doi/full/10.1073/pnas.1606380114, 26.10.2022.

18 *M. Casolino*, Large Scale Analysis of Violent Death Count in Daily Newspapers to Quantify Bias and Censorship, in: Journal of Big Data 7/60, 2020, https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00338-1#citeas, 26.10.2022.

create language models suitable for investigating a given region's cultural, social, and technological transformation during a specific period.[19]

4 Ways of Counting Errors

OCR errors are a well-known problem, and counting them has been a method of measuring the quality of specific techniques. Although objective standards for evaluating have been called for,[20] a simple system for measuring the accuracy of characters has yet to be developed. Technological advanced solutions have been built by comparing dictionaries with the text read by OCR,[21] but that method is not suited for comparing several different languages with many unique characters and different periods marked by the diverse use of language. A better but more work-intensive method is comparing an OCR-read version of a text and a dataset with a 100 percent correct version, a technique called the "gold standard" or "ground-truth text" in the literature,[22] one that has been the standard for testing page-reading OCR systems.[23]

---

19 *N. Cristianini/T. Lansdall-Welfare/G. Dato,* Large-scale Content Analysis of Historical Newspapers in the Town of Gorizia 1873-1914, in: Historical Methods 51/3, 2018, pp. 139-164; *P. Bos et al.,* Quantifying "Pillarization": Extracting Political History from Large Databases of Digitized Media Collections, in: Proceedings of the 3rd HistoInformatics Workshop, 2016, pp. 57-66, http://ceur-ws.org/Vol-1632/paper_8.pdf, 26.10.2022.

20 *S. Tanner/T. Muñoz/P.H. Ros,* Measuring Mass Text Digitization Quality and Usefulness. Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive, in: D-Lib Magazine 15/7-8, 2009, https://www.dlib.org/dlib/july09/munoz/07munoz.html, 26.10.2022.; *R. Holley,* How good can it get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs, in: D-Lib Magazine 15/3-4, 2009, https://www.dlib.org/dlib/march09/holley/03holley.html, 26.10.2022; *P. Conway,* Archival Quality and long-term Preservation: a Research Framework for Validating the Usefulness of Digital Surrogates, in: Archival Science 11/3, 2011, pp. 293-309.

21 *Y. Adesam/D. Dannélls/N. Tahmasebi,* Exploring the Quality of the Digital Historical Newspaper Archive KubHist, in: Proceedings of the Digital Humanities in the Nordic Countries, DHN 2019, pp. 9-17.

22 *G. Nagy/T.A. Nartker/S.V. Rice,* Optical Character Recognition: An Illustrated Guide to the Frontier, in: Procs. Document Recognition and Retrieval VII, 1999, pp. 58-69; *S.V. Rice/G. Nagy/T.A. Nartker,* Optical Character Recognition: An Illustrated Guide to the Frontier, Boston 1999; *S. Pletschacher/C. Clausner/A. Antonacopoulos,* Europeana Newspapers OCR Workflow Evaluation, in: Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing, 2015, pp. 39-46.

23 *G. Chiron et al.*, 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), 2017, pp. 249-

In the study presented in this article, a low-tech solution was chosen to be investigated that has a reasonably high statistical degree of representativeness by virtue of random sampling. Although the method has been criticised for disregarding the quality of layout analysis,[24] the visual evaluation of the images of the original text and the OCR-read text should reduce that problem. A critique of another study using the same method regarding its equal weight given to important text and advertisements can also be avoided by analysing only the OCR-read text from the body text.[25]

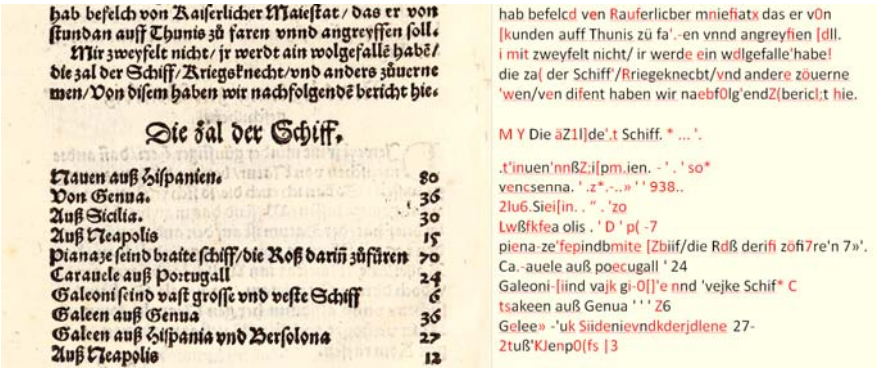
Historical newspaper archives from a wide range of countries were analysed as long as it was possible to see the raw OCR-read text together with the images of the newspaper as seen in Figure 1 with its roughly 73 percent error rate. Although not all archives automatically offer that service, the surrounding OCR-read text can nevertheless be viewed following a search for a word on the page. Ultimately, samples were taken from ANNO (Austria), Mediestream (Denmark), Digitala Samlingar (Finland), La Stampa (Italy), eluxemburgensia (Luxembourg), Delpher (the Netherlands), Digitale aviser (Norway), Premsa digitalitzada (Spain), Svenska dagstidninger (Sweden), Le Temps (Switzerland), and Newspapers.com (USA).

In each of these eleven archives, a specific newspaper was chosen, one that ranked among the largest. Given the assumption that such periodicals are printed with the greatest quality, the error rate for other newspapers was expected to be higher. On the other hand, the most prominent newspapers were probably some of the first processed by OCR, and their OCR quality was thus expected to be far from the best-possible quality due to some of the first generations of OCR engines.

252, m https://hal.archives-ouvertes.fr/hal-03025508/document, 26.10.2022; *T.A. Nartker/S.V. Rice/S.E. Lumos,* Software Tools and Test Data for Research and Testing of page-reading OCR Systems, in: Document Recognition and Retrieval XII 2005, pp. 37-47; *R.C. Carrasco*, An open-source OCR Evaluation Tool, in: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, 2014, pp. 179-184; *C. Clausner/ S. Pletschacher/A. Antonacopoulos,* Flexible Character Accuracy Measure for reading-order-independent Evaluation, in: Pattern Recognition Letters 131, 2020, pp. 390-397; *M.J. Hill/S. Hengchen*, Quantifying the Impact of Dirty OCR on Historical Text Analysis. Eighteenth Century Collections Online as a Case Study, in: Digital Scholarship in the Humanities 34/4, 2019, pp. 825-843.

24 *C. Neudecker/C. Clausner,* A Survey of OCR Evaluation Tools and Metrics, in: *M. Franke-Maier et al. (Eds.),* Qualität in der Inhaltserschließung, Berlin 2021, pp. 13-18.

25 A similar method is used in: *M. Wernersson,* Evaluation von automatisch erzeugten OCR-Daten am Beispiel der Allgemeinen Zeitung, in: ABI Technik 35/1, 2015, pp. 23-35; the critique is in: *Neudecker/Clausner,* A Survey, p. 15.


hab befelch von Kayſerlicher Maieſtat / das er von ſtundan auff Thunis zů faren vnnd angreyffen ſoll. Mir zweyfele nicht / jr werdt ain wolgefallē habē / die zal der Schiff / Kriegsknecht / vnd anders zůuerne men / Von diſem haben wir nachfolgendē bericht hie.

Die ſal der Schiff.

Nauen auß Hiſpanien.	80
Von Genua.	36
Auß Sicilia.	30
Auß Neapolis	15
Pianaze ſeind braite ſchiff / die Roß dariñ zůfüren	70
Caraucle auß Portugall	24
Galeoni ſeind vaſt groſſe vnd veſte Schiff	6
Galeen auß Genua	36
Galeen auß Hiſpania vnd Berſolona	27
Auß Neapolis	12

hab befelcd ven Rauferlicber mniefiatx das er v0n [kunden auff Thunis zů fa'.-en vnnd angreyfien [dll. i mit zweyfelt nicht/ ir werde ein wdlgefalle'habe! die za(der Schiff'/Rriegeknecbt/vnd andere zöuerne 'wen/ven difent haben wir naebf0lg'endZ(bericl:t hie.

M Y Die äZ1l]de'.t Schiff. * ... '.

.t'inuen'nnßZ:i[pm.ien. - ' . ' so*
vencsenna. ' .z*.-..» ' ' 938..
2lu6.Siei{in. . . " . 'zo
Lwßfkfea olis . ' D ' p(-7
piena-ze'fepindbmite [Zbiif/die Rdß derifi zöfi7re'n 7»'.
Ca.-auele auß poecugall ' 24
Galeoni-[iind vajk gi-0[]'e nnd 'vejke Schif* C
tsakeen auß Genua ' ' ' Z6
Gelee» -'uk Siidenievndkderjdlene 27-
2tuß'KJenp0(fs |3

Fig. 1: Sample caption Comparison Between the Original and the OCR Result. Newe Zeytung von Kayserlicher Maiestat Kriegsrüstung, 1535, p. 3. Source: ANNO Historische österreichische Zeitungen und Zeitschriften, https://anno.onb.ac.at.

If the chosen newspaper for an archive did not cover a sufficiently long period, then one or more newspapers were chosen as a supplement. It was impossible, however, to obtain newspapers from all countries covering all years. Sweden's archive contains the first issue of a newspaper from 1645 but stops at 1905 due to copyright issues. By contrast, Spain's archive begins with a newspaper from 1890 that published its last issue in 2010. Thus, for every ten to twenty years, an issue from each newspaper was sampled.[26] The preferred date was August 8th or the following day if the newspaper was not published on that date.

On page two of each issue, a block of approximately 100 words was taken from the body text. If that page was inconvenient or lacked such a block of text, then page four was used instead. The image of part of the chosen page was copied onto paper, and the copy was used to mark all misspelled words. Beyond that, the copy was helpful in analysing the causes of errors and used in a subsequent comparison of the different results.

5 A Significant Difference in Errors

As expected, OCR quality was low for the oldest issues but fine for issues from the most recent decades. The flaws were the same as described in the Nagy,

26 For the Danish *Politiken*, an issue was taken from each five years and with 200 words from the front page.

Nartker, and Rice's illustrated guide (Optical Character Recognition). General issues with damaged letters among others can explain many of those errors.[27] A widespread problem was the division of the text into words. In particular, when a column or page ended, the OCR engine had a problem with converting hyphenated words across lines of text into single words.

However, for nearly all of the archives, issues from one or several years had low OCR quality. Although the poor condition of the original newspapers can explain most of those flaws, other factors exist as well, including the introduction of fracture letters and the use of italics.

Issues of newspapers from several dates had error rates of 100 percent, which indicates that no proofreading had occurred. One page was totally missing from a sample from an otherwise well-processed journal with high quality. No automatic correcting system could catch that error, and no manual proofreading had been performed.

In general, the incorrect words were counted. Sometimes, two columns were mixed together, the words did not form any enlightening meaning, and such words were thus not counted. Although they should have been counted to provide an accurate picture of problems for researchers searching for long sentences, it would only be a minor problem for searches of only a few words.

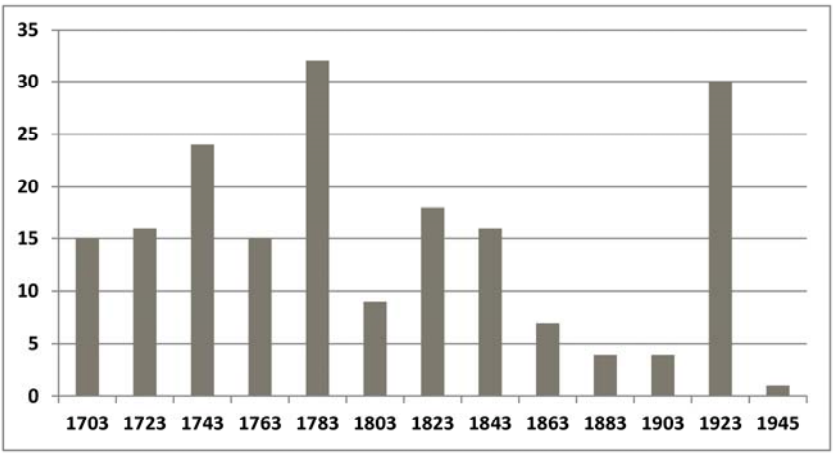


Fig. 2: Example of Error Variation Over Time, Wiener Zeitung. Source: ANNO Historische österreichische Zeitungen und Zeitschriften, https://anno.onb.ac.at.

27 *Nagy/Nartker/Rice,* Optical Character Recognition.

One example of newspaper errors appears in Figure 2. It is the newspaper *Wiener Zeitung* from ANNO (Austria) showing the general trend of increased quality. Issues from several years had high error rates, and an analysis of the images revealed flawed copying, dirty text and problems with two columns treated as one.

6 The Progress of OCR Accuracy

Altogether, the samples indicate an exciting development in improved OCR quality over the years, although such improvement did not occur smoothly as seen in Figure 3. Following primitive printing in the earliest sampled decades came an improvement caused by the enhanced quality of both paper and printing. However, beginning at around 1730, a rise in error rates surfaced that lasted nearly 100 years despite general improvements in the use of paper and print. The emergence of the broadsheet format can explain these elevated rates, for it allowed newspapers to have several columns and, as a consequence, posed problems for OCR engines.

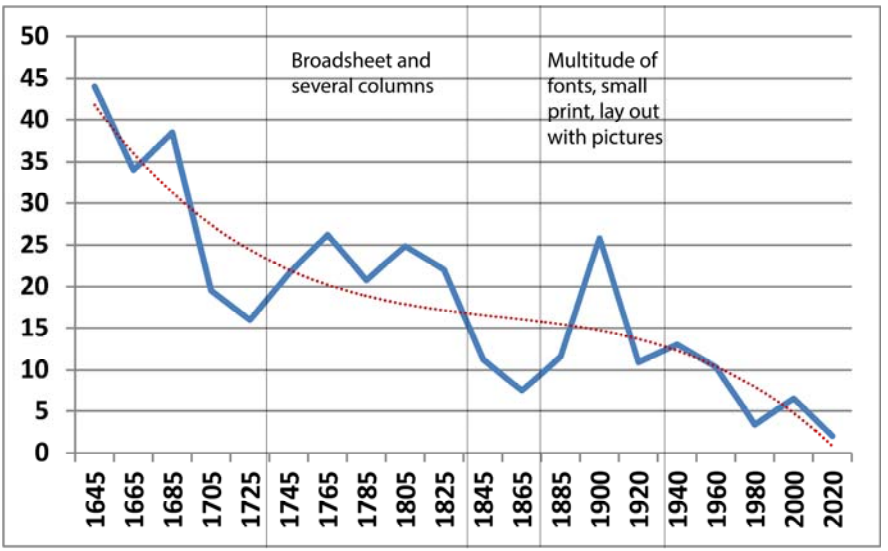


Fig. 3: OCR Word Error Rate Over Time. Accumulated Results From the Total Samples.

Around 1885, another rise in error rates occurred, albeit for a period lasting only 40 years. The period was characterised by the use of a multitude of fonts and newspaper texts filled with illustrations. In the later part of that period, automatic typesetting was introduced, which yielded more uniform letters but also expanded the possibility of using small prints.

Taken together, such uneven development implies that the likelihood of generating correct results has differed across time and not in any linear fashion. Indeed, the likelihood was greater for 1725 than for 1765, and searches yielded better results for 1865 than 1900.

7 The Average Error Rate

The average rate of errors for all archives for all years was 18 percent. Such a high rate warrants the severe consideration of the quality of searches of OCR-read text for scientific use. At the same time, the result corresponds well with the finding of a comprehensive study showing up to 10 percent wrongly detected characters in some documents and an average 4 percent error rate on characters in serial.[28] In light of that error rate, the error rate for words with an average length of approximately five letters was 20 percent. A study revealed an error rate of 8-13 percent using commercial software not tailored for older fonts,[29] and another study gives also high error rates, though it fluctuates from language to language.[30]

Such high error rates were especially a problem in searching for sentences. A search for the seven-word sentence "You speak an infinite deal of nothing" yielded a success rate well under 50 percent, with an error rate for words of ten percent. With an error rate on words on 20 percent, the success rate was as low as nearly 20 percent. For researchers aiming for a success rate of 100 percent or close to it, those rates are problematic. Claims of "first time" about incidents are not possible.

---

28 The researchers made their investigations around 2004. Since then, the quality from the OCR engines is improved. *Chiron et al.*, Impact of OCR Errors.

29 *S. Drobac/K. Lindén*, Optical Character Recognition with Neural Networks and Post-correction with Finite State Methods, in: International Journal on Document Analysis and Recognition 23, 2020, pp. 279-295.

30 *C. Neudecker/H.J. Lieder/S. Kobel,* Europeana Newspapers. A Gateway to European Newspapers Online. Final Report. 2015, http://www.europeana-newspapers.eu/wp-content/uploads/2015/05/Final_Report.pdf, 26.10.2022.

Although the error rate for body text was 18 percent, the rate for advertisements was far higher. That rate was not investigated in depth, but a single sample of advertisements from one newspaper indicated a manifold higher rate as seen in Figure 4.

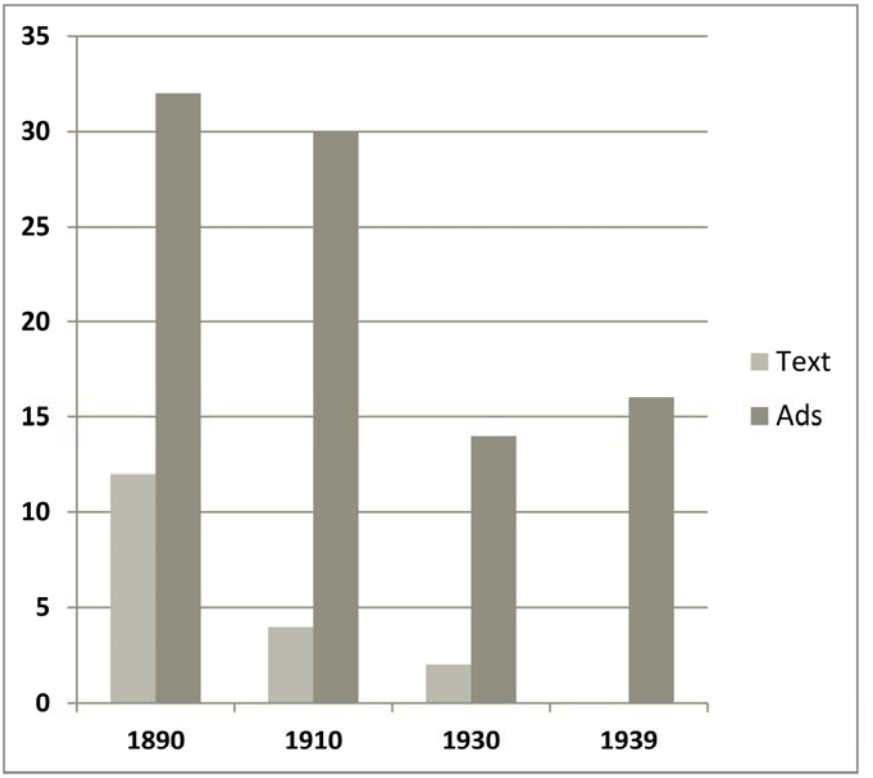


Fig. 4: Errors in Advertisements/Body Text, Päivälehti. Source: The National Library of Finland.

8 How to Improve OCR Quality

The mentioned error rates are too high for serious research, meaning that OCR quality needs to be improved. Such improvement is possible, but some efforts to that end are costly, whereas others await the development of new, enhanced technological tools. As Neudecker et al. conclude in a study on OCR evaluation

that "quality assessment or even prediction techniques based on comparatively small but representative samples for which GT [Ground Truth] is needed."[31]

It is likely that some text on microfilms cannot even be read by the human eye. The dismal prints of newspapers show many different weaknesses, often in combination at the same time, including imaging defects with heavily or lightly printed curved baselines, stray marks, and shaded backgrounds.[32] As explained by the library sector, "Usually the quality of the OCR text says more about the condition of the original materials than it does about the performance of the OCR software".[33] In that light, high error rates can be reduced by making new images of the newspaper pages.

Other solutions are more technical in nature. Although promising, they risk severe limitations for the professional historian's ambition to achieve complete source criticism. The many interpretations generated through the OCR process stand to transform the original newspaper into something else during digitisation. As Jarlbrink and Snickars have emphasised, libraries have often outsourced the digitisation process and thus run the risk of losing control over the quality of their collections.[34] The many interpretations made possible through the process are therefore unknown factors affecting the original document.

Perhaps needless to say, the process from the birth of a newspaper to the researcher's search on a computer is a long one. Many single functions can be improved except for initial steps taken long ago when printers used worn typecasts or not enough ink in the printing process.

The following discusses some of the essential elements in the OCR process as seen in Figure 5. Although some OCR programmes have other solutions for more complicated tasks, the principles between the programmes are nearly the same as illustrated in the figure.

Many problems with primarily the oldest newspapers seem to have been caused by the poor quality of the issues used. Another library may well have a better copy that has suffered less wear and tear or better storage conditions. Tracking down a usable original copy is costly but has no influence on the source criticism.

---

31 *Neudecker/Clausner*, A Survey, p. 16.

32 *Nagy/Nartker/Rice*, Optical Character Recognition.

33 *E. Klijn*, The Current State-of-Art in Newspaper Digitization. A Market Perspective, in: D-Lib Magazine 14/1-2, 2008, https://www.dlib.org/dlib/january08/klijn/01klijn.html, 26.10.2022.

34 *J. Jarlbrink/P. Snickars*, Cultural Heritage as Digital Noise: Nineteenth Century Newspapers in the Digital Archive, in: Journal of Documentation 73/6, 2017, pp. 1228-1243.

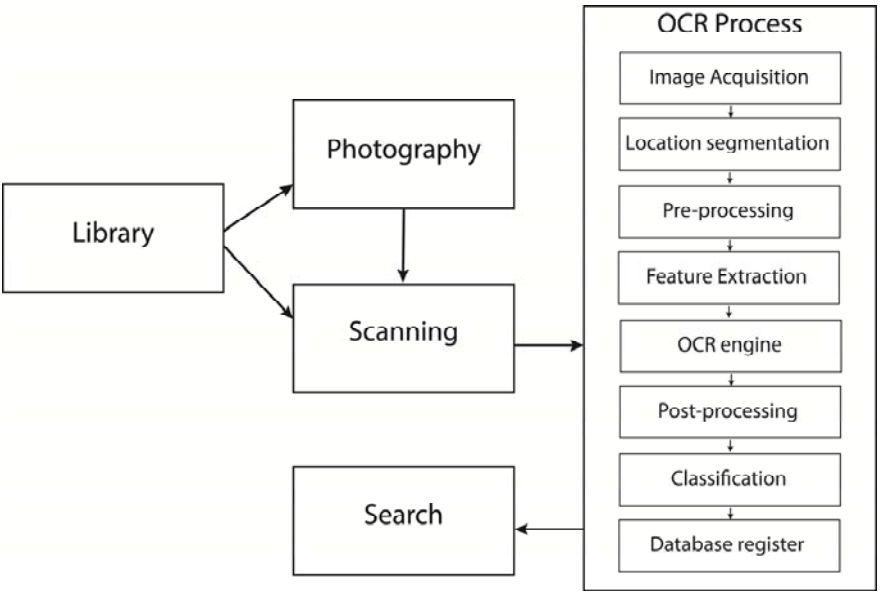


Fig. 5: Principle of the OCR Process.

For practical reasons, original newspapers are often copied to microfilm before digitisation. Today, the standard is using a reel of 35-mm film and routine photographing on an "assembly line", so to speak, in a low-income country. However, the photographing is usually not adjusted to approximate the appearance of the paper. In early years, microfilms were often made with medium-quality film (e.g. 300 PPI [Pixels Per Inch]) offering suboptimal resolution. Thus, many newspapers need to be photographed again on higher-quality film (e.g. 400 PPI).

The creation of digital images has been one of the weakest points of the OCR process, one also achieved on an assembly line in low-income countries, where workers have not attended to slight contrasts between browned paper and print. Today's technical equipment offers outstanding possibilities to complete that task, but the work can be time-consuming and thus expensive if a newspaper has varied in its appearance and if many corrections need to be made manually. For example, the Finnish newspaper archive could improve its quality by several percentage points by redoing the photography and digitisation, but that solution was rejected for financial reasons.[35]

35 Some of the oldest newspapers are reprocessed for a better quality through the new OCR platform Transkribus (press release Kansalliskirjasto 2021-12-22).

Before a digital image can be used, it needs to be pre-processed to improve its quality. Colour images need to be converted to greyscale and often into slightly mapped files to secure sufficient contrast. That task can be performed automatically if the pages do not vary too much in terms of contrast. Often, noise reduction is also included in the process. Again, however, the work is time-consuming and therefore costly.

Although the more obscure areas of the OCR process, including location segmentation, constitute a black box for most historians, they merit attention because the original information can easily be garbled. The first step is to develop a guide for the later OCR process regarding how pages are segmented, so that the OCR engine can later read pages as a human would. Several columns should be identified, and if a headline spans the top of several columns, then the programme should mark the headline to be read in its entirety before the first column.

Location segmentation has often been accomplished by using automatic programmes, as made evident during the analysis of most of the eleven newspaper archives.[36] Although many flaws are caused by insufficient segmentation, some commercial companies have achieved perfect segmentations by using manual workers, again living in low-income countries.

The next process, one far deeper into the black box, is character segmentation. A computer separates lines and single characters from each other. With a newspaper page exhibiting high quality and sharp characters, a modern programme can segment characters with rather high accuracy. However, as mentioned, OCR engines continue to require improvement before they can handle Gothic typeface.

Feature extraction with recognition is an essential step in which an OCR engine identifies each letter and classifies all letters for size and boldness, among other aspects. Although the process is automatic, most OCR engines can learn specific fonts. Again, however, human involvement is needed to give the engine examples of texts and, through experiments, indicate to it when it provides the best results. As mentioned earlier, the process can be combined with the text of the gold standard to compare the quality after a learning process. On the whole, the learning process within feature extraction with recognition is necessary for each font family, and given publishers' use of many different typefaces across time, it requires a great deal work to achieve a nearly perfect result.

Some OCR programmes perform recognition on whole words, which can cause problems when the programmes guess a solution. Although such guesses

36 The Swedish use of automatic segmentation is mentioned in *Jarlbrink/Snickars*, Cultural Heritage, p. 1238.

are often correct, they can generate bias. For instance, if the programme uses a word dictionary, then the final text favours words in the dictionary. The occurrence of all other words will be reduced, and such texts are useless for investigating the history of language.[37]

The text produced from feature extraction can be further refined as well in post-processing. For example, a programme can incorporate language models. Again, however, the programme will skew the result, and text made from those processes cannot be used to gauge the statistical occurrence of words over time and to complete similar tasks.

The result of OCR needs to be archived in a readable text format with the necessary metadata and links to illustrations and pictures. To that purpose, several formats of classification with a combination of XML formats have been developed, among which ALTO [Analyzed Layout and Text Object] is one of the most-used standards for technical metadata, often together with the standard description of the entire digitised object METS [Metadata Encoding and Transmission Standard].

Even if a text is archived in a rather complex database, a search engine can find the results quickly, and, as a result, the database becomes filled with different helpful registers waiting for questions. Some databases are built with helpful registers of keywords, while some, for practical reasons, do not use words that are not useful for the retrieval process. Although the list of these "stopwords" is not standardised, for English a list of approximately 100 words can be used and thereby impact the search process.[38]

Ultimately, users need a way to search in a database, a rather complicated piece of technology liable to skew the results. For instance, as a solution to produce better search results from a weak OCR file, a so-called fuzzy search engine can be used. As known from using Google, a search engine can give the answer to "God save the queen" even if the phrase searched for is "Got save the queen". Although such programmes can fill in the gaps and/or find related words with high accuracy, even approaching 100 percent,[39] rare occurrences are often suppressed unbeknownst to users.

37 *Nagy/Nartker/Rice*, Optical Character Recognition, p. 68.

38 www.ranks.nl/stopwords, 09.08.2022.

39 *Tanner/Muñoz/Ros*, Measuring mass Text.

9 OCR Engines

Central to the process are OCR programmes, the rapid improvement of which in recent decades will no doubt continue. Today, four programmes lead the charge according to a recent review.[40] First, the commercial programme FineReader has been described as the state of the art for contemporary material and as the best option to make manual corrections after fully automated segmentation, if necessary. The programme has strong language models that support up to 200 languages, with dictionary assistance for approximately 50 of them. It has shortcomings with early printings, however, including fracture letters.

Second, the Open Source OCR-engine Calamari has been characterised as the best option for recognition and speed.[41] It does not include pre-processing, segmentation, or post-processing capabilities, however, meaning that additional programmes are necessary.

Third, Tesseract, originally developed as proprietary software in the 1980s, became Open Source in 2005 and was later sponsored by Google. Now in its fourth version as an Open Source programme widely used by libraries, it supports the use of dictionaries and language modelling. Nevertheless, it is judged to fall short when it comes to processing historical material.

Fourth and finally, OCRopus, also known as Ocropy, is another free system, currently in its third version, the development of which was sponsored by Google and which is used for digitising books for Google Books. It has an excellent pre-processing process and makes robust line segmentation.

Despite productive work on improving parts of the OCR process, including the use of neural networks and similar technologies tested by many researchers,[42] attention to the skewing of data is seldom paid.

---

40 *C. Reul et al.*, OCR4all – An Open-source Tool Providing a (semi-)automatic OCR Workflow for Historical Printings, in: Applied Sciences 9, 2019, https://www.mdpi.com/2076-3417/9/22/4853/htm, 26.10.2022; *S. Drobac,* OCR and Post-correction of Historical Newspapers and Journals, University of Helsinki 2020, https://helda.helsinki.fi/bitstream/handle/10138/319496/OCRandpo.pdf?sequence=1&isAllowed=y, 26.10.2022.

41 *C. Wick/C. Reul/F. Puppe,* Calamari – A high-performance tensorflow-based deep learning Package for Optical Character Recognition, in: arXiv 2018, https://arxiv.org/pdf/1807.02004.pdf, 26.10.2022.

42 *Drobac/Lindén*, Optical Character Recognition; *D. Sporici/E. Cuşnir/C.-A. Boiangiu,* Improving the Accuracy of Tesseract 4.0 OCR Engine using convolution-based Preprocessing, in: Symmetry 12/715, 2020, https://www.mdpi.com/2073-8994/12/5/715/htm, 26.10.2022; *V.-N. Huynh/A. Hamdi/A. Doucet,* When to use OCR post-correction for named entity Recognition?, 22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020; *S. Rijhwani et al.,*

10 Crowdsourcing

A new way of reducing the number of errors can be achieved with contributions from volunteers via so-called crowdsourcing, which allows anyone to validate digitised information. One of the most notable successes of crowdsourcing has occurred in Australia, where more than 100,000 users of Trove, the newspaper archive at the National Library of Australia, corrected more than 100 million lines of text by 2013 – that is, the equivalent of 270 standard work years.[43] By 2022, the amount of corrected text had increased to 438,121,591 lines, more than 7 million of which had been fixed by one supremely diligent user.[44]

Crowdsourcing is rather inexpensive when the necessary platform is established. It can also supplement other mentioned tools to reduce the error rate.[45] However, because the amount of newspaper pages is typically quite high, it will remain only a supplement even despite the help of a highly active public.

11 The Future of Digital Archives and Research

The digital archives need to grow from the present stage. The reasons for necessary technical improvements of the OCR engines were explained earlier in the article.

Another critique is the exclusiveness of the archived newspapers. Only a small part of the world's newspapers is digitised. "The attractiveness of working with digital archives means that many scholars will inevitably be drawn to

---

Lexically aware semi-supervised learning for OCR post-correction, in: Transactions of the Association for Computational Linguistics 9, 2021, pp. 1285-1302; *R. Schaefer/C. Neudecker,* A two-step Approach for automatic OCR post-correction, in: Proceedings of LaTeCH-CLfL 2020, pp. 52-57; *Kettunen/Koistinen,* Open Source Tesseract, pp. 33-42.

43 *R. Holley,* Many Hands make light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers, National Library of Australia 2009; *M.-L. Ayres,* "Singing for their supper": Trove, Australian Newspapers, and the Crowd, in: IFLA WLIC 2013, https://library.ifla.org/id/eprint/245/1/153-ayres-en.pdf, 26.10.2022.

44 trove.nla.gov.au/landing/community/hallOfFame, 07.08.2022.

45 *S. Clematide/L. Furer/M. Volk,* Crowdsourcing the OCR Ground Truth of a German and French Cultural Heritage Corpus, in: Journal for Language Technology and Computational Linguistics (JLCL) 33/1, 2018, pp. 25-47; M. *Mayr,* Crowdsourcing für Bibliotheken – Best Practices und Handlungsempfehlungen, Universität Wien 2018.

those titles that they can access via their computer", explains a critical Adrian Bingham.[46]

> The digital archive – despite the obvious faults this article documents – is a useful tool. As American Dan Cohen comments about the Google Books project: "The existence of modern search technology should push us to improve historical research. It should tell us that our analog, necessarily partial methods have had hidden from us the potential of taking a more comprehensive view, aided by less capricious retrieval mechanisms which [...] are often more objective than leafing rapidly through paper folios on a time-delimited jaunt to an archive."[47]

However, a critical attitude is necessary. When everything is available online, the sheer quantity makes close reading impractical. As Thomas Moss, David Thomas, and Tim Gollins comment: "Users of necessity have to rely on search algorithms and sophisticated analytical tools to make sense of the totality of the record."[48]

As mentioned in the first chapter, use of the large databases will place demands for interdisciplinary projects with approaches from computer science, the humanities, and library workers. This creates a need in the different participants to understand the workflows and traditions of each of the disciplines involved.[49]

Naturally, future development needs to be followed by a continued critique of theory and methods. As Ted Underwood emphasises, is it necessary to create theories when we work with large-scale databases. "Humanists tend to think of computer science as an instrumental rather than philosophical discourse. The term ''datamining" makes it easy to envision the field as a collection of mining ''tools.'' But that's not an accurate picture. The underlying language of datamining – Bayesian statistics – is a way of reasoning about interpretation that can help us approach large collections in a more principled way."[50]

Humanistic scholars do not presently have established methods of assessing digital tools. As Marijn Koolen, Jasmijn van Gorp, and Jacco van Ossenbruggen emphasise, humanistic scholars do not currently have a consensus on what questions researchers should ask themselves to evaluate digital

46 *Bingham*, The Digitization of Newspaper, p. 229.

47 *D. Cohen*, Is Google good for History? 2010, hdl.handle.net/1920/6101, 25.10.2022.

48 *M. Moss/D. Thomas/T. Gollins*, The Reconfiguration of the Archive as Data to be mined, in: Archivana 86, 2018, pp. 118-151.

49 *S. Oberbichler et al.*, Integrated Interdisciplinary Workflows for Research on Historical Newspapers: Perspectives from Humanities Scholars, Computer Scientists, and Librarians, in: Journal of the Association for Information Science and Technology 73, 2022, pp. 225-239.

50 *T. Underwood*, Theorizing Research Practices we forgot to theorize twenty Years ago, in: Representations 127/1, 2014, pp. 64-72.

sources, and the three scholars suggest a long list of critical questions about the data and the tools.[51]

The rather unflattering situation of newspaper archives with different ownerships, different technologies, being closed, not allowing access to databases, etc., should be changed. One vision is that researchers could have access to archives world-wide. The people behind the mentioned project on the atlas could only use a few archives.[52] The wish for the future should be access to all newspapers in one search.

The word "macroscope" has recently emerged and is similar to the astronomers' telescope and biologists' microscope. By this method the researcher is able to "selectively reduce complexity until once-obscure patterns and relationships become clear", as Shawn Graham et al. see it.[53] But, as Rik Hoekstra and Marijn Koolen emphasise: "That each step of selection, enrichment, and classification represents a choice that is based on explorations and interpretations of the data. These interactions change the data and are essential in understanding any subsequent analysis."[54]

We have been witnesses to a kind of a second digital revolution on how we practise academic history writing. In his essay on the consequences of OCR and its future, Tim Hitchcock explains that historians need to navigate the new sea of data: "Just as the primary sources we use have become digital, so the history we create has itself been turned into a new digital form. From our grant applications, to our notes and bibliographies, to our prose, to our submitted drafts, peer review reports, proofs and offprints, what was once a mechanical system marching to the rhythm of a printing press has become a semi-magical process of silent production and reproduction."[55]

Several fine projects have tried to improve the technology and make use of standards. The 2008-2011 IMPACT project improved methods, and the 2012-2015

---

51 *M. Koolen/J. van Gorp/J. van Ossenbruggen,* Toward a Model for digital Tool Criticism: Reflection as Integrative Practice, in: Digital Scholarship in the Humanities 34/2, 2019, pp. 368-385.

52 *J. Verheul et al.*, Using Word Vector Models to trace Conceptual Change over Time and Space in Historical Newspapers, 1840-1914, in: Digital Humanities Quearterly 16/2, 2022, https://eprints.whiterose.ac.uk/187445/1/DHQ_%20Digital%20Humanities%20Quarterly_%20Using%20word%20vector%20models%20to%20trace%20conceptual%20change%20over%20time%20and%20space%20in%20historical%20newspapers%2018401914.pdf, 26.10.2022.

53 *S. Graham et al.*, Exploring Big Historical Data. The Historian's Macroscope, Singapore 2016.

54 *R. Hoekstra/M. Koolen,* Data Scopes for Digital History Research, in: Historical Methods: A Journal of Wuantitative and Interdisciplinary History, 52/2, 2018, pp. 79-94.

55 *T. Hitchcock*, Confronting the Digital. Or how Academic History Writing Lost the Plot, in: Cultural and Social History 10/1, 2013, pp. 9-23, here: p. 10.

Europeana project further aimed to make European digital historic newspapers available via common heritage websites.[56]

It could be just a dream to have a similar project on a global level for the pleasure of writing international history. There are currently huge problems when researchers want to use data from a larger geographic area. Each database contains its own theoretically standardised collection of data, metadata, and images. Researchers need to know all the technical details to get access to the information and to judge its validity. The authors behind the mentioned atlas have looked behind the curtains of some archives and have a severe criticism: "The precise nature and nuance of the data is often occluded by the automatic processes that encoded it. Moreover, no true universal standard has been implemented to facilitate cross-database analysis, encouraging digital research to remain within existing institutional or commercial silos." The atlas had no ambition to provide a "better" standard, but to give information so that everyone can explore these collections "in relative safety".[57]

Use of big data gives the researcher a duty to handle new techniques as a "digital humanist". Not only do the different "silos" need to be learned, but the handling of software and data, contextualisation, methodological operationalisation, analysis, and interpretation as well.[58]

It takes into consideration the use of statistics when searching over a large time period that contains few published newspapers in the oldest time and an enormous number of publications in the present time. This concept is not specific to the OCR process itself, but normalisation is often necessary to compensate for the uneven numbers of published words. This places the demand on database owners to provide the database with information on the numbers of words in each year.[59]

---

56 *Neudecker/Lieder/Kobel*, Europeana Newspapers; *S. Pletschacher/C. Clausner/A. Antonacopoulos*, Eurpeana Newspapers OCR Workflow Evaluation, in: Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing 2015, pp. 39-46.

57 *Beals/Bell*, The Atlas, pp.1-2.

58 *H. Wijffjes*, Digital Humanities and Media History: A Challenge for Historical Newspaper Research, in: Journal for Media History 20/1, 2017.

59 The best way to explain the concept is to give an example. When a search on a text string in the largest database of digitised literature, Google Books, is made through its Ngram Viewer, the results are viewed in a time series chart showing the frequencies using a yearly count. The system is built to take care of the number of books published in specific years as explained in *J.-B. Michel et al.*, Quantitative Analysis of Culture Using Millions of Digitized Books, in: Science 14/331, 2011, pp. 176-182.

12 Conclusion

The analysis of some of the leading newspaper archives in the world shows an error rate of 18 percent on body text, with a much higher rate for text in advertisements. These high error rates allow one to look critically at this still young sector, which has many years until it reaches maturity. The number of errors needs to be reduced to meet the researchers' wish for almost 100 percent trustworthy results. Many newspapers need to be digitised again. At the very least, the oldest issues need to be photographed again, and many archives need to repeat the OCR process from its old microfilm with new and better versions of OCR engines. It is costly, but it should be the last of a repeated once-in-a-lifetime process.

Unfortunately, many archives are in the hands of private companies. However, they are often well functioning and deliver results when people search for their ancestors or for history from their hometown. Service for the relatively rare serious researcher is not their largest market and an incentive to deliver the highest service in metadata and standardised infrastructure is not strong. Their officers very rarely join academic discussions about historical archives. There is a hope that those archives will be integrated in an international network of collaborative archives with communication standards and transparency for the content.

This transparency is necessary for all archives. The technological development around the OCR engines will hopefully continue its strong energy many years from now. Higher quality will indeed be obtained from many of the actual ideas in the laboratories. Many of these ideas are based on advanced logarithms that, for the ordinary historian, remain an unknown black box. Serious users need to know how the new kind of source is created. There needs to be comprehensive information in the metadata on how the data in the OCR process has been manipulated. There is a risk of constructed skewed text, when the material has gone through the many processes. It is acceptable for a search engine to give many manipulated results thanks to its use of fuzzy logic. The researchers should bypass those skewed results by providing access to the raw database. Even here, the data have been born in a long process by different versions of programmes. To secure the important source critique for historians, the metadata needs to be accessible on its creation.

I can only repeat the advice for the knowledge situation made some years ago: "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."[60]

Many new possibilities have changed the world for historians. It has given many new possibilities for finding information and for statistical handling of really large amounts of data. We have already seen many fine studies, and more will come. Many fruitful ideas have come up from workshops with historians in cooperation with librarians and computer scientists.

Still, many universities have to deliver introductory courses on the topic. New students need to know the possibilities in this emerging field – and its pitfalls.

Bionote

Jørgen Burchardt
formerly the director of the Danish Road Museum, is now a senior researcher at the Museum Vestfyn. Initially educated as an engineer in the printing business, he studied ethnology at the University of Copenhagen and pursued further education at the Royal Institute of Technology in Sweden, Deutsches Museum in Germany, and CERN in Switzerland. He has published books, conference papers, and articles on the history of labour, technology, and organisation, including several books addressing research policy and digital publishing. He has also served as an expert advisor on digital archiving and scientific communication for government ministries, private companies, and international organisations.

60 *M.C. Traub/J.v. Ossenbruggen/L. Hardman,* Impact Analysis of OCR Quality on Research Tasks in Digital Archives, in: International Conference on Theory and Practice of Digital Libraries 2015, p. 252, https://link.springer.com/chapter/10.1007/978-3-319-24592-8_19, 26.10.2022.