Bibliography of OCR / Text Recognition Sources
2022

Dr. Željko Trbušić
Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Alex, B., et al. “Digitised Historical Text: Does It Have to Be MediOCRe?” *11th Conference on Natural Language Processing, KONVENS 2012: Empirical Methods in Natural Language Processing - Proceedings of the Conference on Natural Language Processing 2012*, ÖGAI, 2012, pp. 401–09.

Alotaibi, Faiz, et al. “Optical Character Recognition for Quranic Image Similarity Matching.” *IEEE Access*, vol. 6, no. November, 2017, pp. 554–62, https://doi.org/10.1109/ACCESS.2017.2771621.

Anderson, Niall. *Glossary for the Mass Digitisation of Text & OCR: IMPACT Workflow Resource*. British Library, 2010, pp. 1–25, https://www.digitisation.eu/download/website-files/WorkflowResources/GlossaryfortheMassDigitisationofText_OCR-ImpactWorkflowResource_01.pdf.

---. *IMPACT : Building Capability in Mass Digitisation*. 2012.

---. *Optical Character Recognition: IMPACT Best Practice Guide*. British Library, University Innsbruck, PRIMA Research Lab, 2010, pp. 1–8, https://www.digitisation.eu/download/website-files/BPG/OpticalCharacterRecognition-IBPG_01.pdf.

---. *Optical Character Recognition: IMPACT Briefing Paper*. British Library, 2010, pp. 1–4, https://www.digitisation.eu/download/website-files/BP/OpticalCharacterRecognition-BriefingPaper_01.pdf.

Antonacopoulos, A., and C. Casado Castilla. “Flexible Text Recovery from Degraded Typewritten Historical Documents.” *Proceedings - 18th International Conference on Pattern Recognition*, vol. 2, 2006, pp. 1062–65, https://doi.org/10.1109/ICPR.2006.581.

Arya, Deepak, et al. *Experiences of Integration and Performance Testing of Multilingual OCR for Printed Indian Scripts*. 2011.

Bagdanov, Andrew D., et al. *The OCR Frontiers Toolkit*. 1999.

Baird, Henry S., et al. “Robust Document Image Understanding Technologies.” *HDP 2004: Proceedings of the First ACM Hardcopy Document Processing Workshop*, 2004, pp. 9–14, https://doi.org/10.1145/1031442.1031444.

Balk, Hildelies, and Lieke Ploeger. “IMPACT: Working Together to Address the Challenges Involving Mass Digitization of Historical Printed Text.” *OCLC Systems and Services*, vol. 25, no. 4, 2009, pp. 233–48, https://doi.org/10.1108/10650750911001824.

Batawi, Yusof A., and Osama A. Abulnaja. “Accuracy Evaluation of Arabic Optical Character Recognition Voting Technique: Experimental Study.” *IJECS: International Journal of Electrical & Computer Sciences*, vol. 12, no. 1, 2012, pp. 29–33.

Björkman, Jacob. *Evaluation of the Effects of Different Preprocessing Methods on OCR Results from Images with Varying Quality*. 2019. PhD Thesis.

Blanke, Tobias, et al. “Ocropodium: Open Source OCR for Small-Scale Historical Archives.” *Journal of Information Science*, vol. 38, no. 1, 2012, pp. 76–86, https://doi.org/10.1177/0165551511429418.

---. “Open Source Optical Character Recognition for Historical Research.” *Journal of Documentation*, vol. 68, no. 5, 2012, pp. 659–83, https://doi.org/10.1108/00220411211256021.

Blostein, Dorothea, and George Nagy. “Asymptotic Cost in Document Conversion.” *Document Recognition and Retrieval XIX: Proceedings of SPIE*, edited by C. Viard-Gaudin and R. Zanibbi, SPIE, 2012, p. 82970N, https://doi.org/10.1117/12.912161.

Breuel, Thomas M. *The HOCR Microformat for OCR Workflow and Results*. 2007.

---. “The OCRopus Open Source OCR System.” *Proceedings of SPIE - The International Society for Optical Engineering*, 2008.

Büttner, Andreas. *Nasẖī – an Efficient Tool for the OCR-Aided Transcription of Printed Texts*. 2019, https://doi.org/10.20944/preprints201909.0062.v1.

Carenvall, Carl. *Adaptive Binarization of 17th Century Printed Text*. 2012.

Carrasco, Rafael C. “An Open-Source OCR Evaluation Tool.” *ACM International Conference Proceeding Series*, 2014, pp. 179–84, https://doi.org/10.1145/2595188.2595221.

Chiang, Yao Yi, et al. “Assessing the Impact of Graphical Quality on Automatic Text Recognition in Digital Maps.” *Computers and Geosciences*, vol. 93, 2016, pp. 21–35, https://doi.org/10.1016/j.cageo.2016.04.013.

Cojocaru, Svetlana, et al. “Optical Character Recognition Applied to Romanian Printed Texts of the 18th-20th Century.” *Computer Science Journal of Moldova*, vol. 24, no. 1, 2016, pp. 106–17.

Croft, W. B., et al. *An Evaluation of Information Retrieval Accuracy with Simulated OCR Output*. 1993.

D’Albe, Edmund Fournier. “On a Type-Reading Optophone.” *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 1914.

Drinkwater, Robyn E., et al. “The Use of Optical Character Recognition (OCR) in the Digitisation of Herbarium Specimen Labels.” *PhytoKeys*, vol. 38, 2014, pp. 15–30, https://doi.org/10.3897/phytokeys.38.7168.

Eikvil, Line. *OCR - Optical Character Recognition*. December, 1993.

Godil, Afzal, et al. *The Text Recognition Algorithm Independent Evaluation (TRAIT)*. 2017.

Gupta, Maya R., et al. “OCR Binarization and Image Pre-Processing for Searching Historical Documents.” *Pattern Recognition*, vol. 40, no. 2, 2007, pp. 389–97, https://doi.org/10.1016/j.patcog.2006.04.043.

Haaf, Susanne, et al. “Measuring the Correctness of Double-Keying: Error Classification and Quality Control in a Large Corpus of TEI-Annotated Historical Text.” *Journal of the Text Encoding Initiative*, vol. 2013, no. Issue 4, 2013, pp. 0–20, https://doi.org/10.4000/jtei.739.

Habeeb, Imad Qasim, et al. “Improving Optical Character Recognition Process for Low Resolution Images.” *International Journal of Advancements in Computing Technology*, vol. 6, no. 3, 2014, pp. 13–21.

Handley, John C., and Thomas B. Hickey. “Merging Optical Character Recognition Outputs for Improved Accuracy.” *Computer-Assisted Information Retrieval (Recherche d’Information et Ses Applications) - RIAO 1991*, 1991.

Harris, Martyn, et al. “Comparing ‘Parallel Passages’ in Digital Archives.” *Journal of Documentation*, vol. 76, no. 1, 2019, pp. 271–89, https://doi.org/10.1108/JD-10-2018-0175.

Holley, Rose. “How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs.” *D-Lib Magazine*, vol. 15, no. 3–4, 2009, pp. 1–13, https://doi.org/10.1045/march2009-holley.

---. *Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers*. March, National Library of Australia, 2009, pp. 1–28.

Hubert, Isabell, et al. *Training & Quality Assessment of an Optical Character Recognition Model for Northern Haida*. 2016, http://www.lrec-conf.org/proceedings/lrec2016/pdf/39_Paper.pdf.

Islam, Noman, et al. “A Survey on Optical Character Recognition System.” *Journal of Information & Communication Technology - JICT*, vol. 10, no. December, 2016, pp. 1–4.

Jerele, Ines, et al. *Optical Character Recognition of Historical Texts: End-User Focused Research for Slovenian Books and Newspapers from the 18th and 19th Century*. 2012.

Kanai, J., et al. “Performance Metrics for Document Understanding Systems.” *Proceedings of ICDAR’93*, 1993, pp. 424–27, https://doi.org/10.1109/icdar.1993.395703.

Kanai, Junichi, et al. “Automated Evaluation of OCR Zoning.” *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE IN^{TEL}LIGENCE*, vol. 17, no. 1, 1995.

Karpinski, R., et al. “Metrics for Complete Evaluation of OCR Performance.” *Proceedings of the 2018 International Conference on Image Processing, Computer Vision, and Pattern Recognition, IPCV 2018*, 2018, pp. 23–29.

Kettunen, Kimmo, et al. *Creating and Using Ground Truth OCR Sample Data for Finnish Historical Newspapers and Journals*. 2018.

Kettunen, Kimmo, and Mika Koistinen. “Open Source Tesseract in Re-OCR of Finnish Fraktur from 19th and Early 20th Century Newspapers and Journals – Collected Notes on Quality Improvement.” *CEUR Workshop Proceedings*, vol. 2364, 2019, pp. 270–82.

Kleiner, A., and R. C. Kurzweil. “A Description of the Kurzweil Reading Machine and a Status Report on Its Testing and Dissemination.” *Bulletin of Prosthetics Research*, vol. 10, no. 27, 1977, pp. 72–81.

Koistinen, Mika, Jukka Kervinen, et al. “How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source Tesseract OCR Engine.” *8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, edited by Z. Vetulani and J. Mariani, Springer-Verlag, 2018, pp. 279–83.

Koistinen, Mika, Kimmo Kettunen, et al. “Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing.” *Proceedings of the 21st Nordic Conference of Computational Linguistics*, 2017, pp. 23–24.

Kordić, Vesna. *Elektronička Knjiga - Izrada El. Knjige Putem OCR Tehnologije (Optičkog Prepoznavanja Znakova)*. 2009. PhD Thesis.

Le, Daniel X., and George R. Thoma. “Automatically Creating Biomedical Bibliographic Records from Printed Volumes of Old Indexes.” *SCI 2005. Proc 9th World Multiconference on Systemics, Cybernetics and Informatics*, edited by N. Callaos and W. Lesso, International Institute of Informatics and Systemics, 2005, pp. 267–74.

Li, Ning. *An Implementation of OCR System Based on Skeleton Matching*. 1991.

Liang, Jihong, et al. “Task Design and Assignment of Full-Text Generation on Mass Chinese Historical Archives in Digital Humanities: A Crowdsourcing Approach.” *Aslib Journal of Information Management*, vol. 72, no. 2, 2020, pp. 262–86, https://doi.org/10.1108/AJIM-09-2019-0245.

Mao, Song, et al. “Design Strategies for a Prototype Electronic Preservation System for Biomedical Document.” *Archiving 2005 - Final Program and Proceedings*, vol. 2005, 2005, pp. 48–52.

Misra, Dharitri, et al. “Archiving a Historic Medico-Legal Collection: Automation and Workflow Customization.” *Archiving 2007: Final Program and Proceedings*, edited by S. A. Stovall, IS&T, 2007, pp. 157–61.

Muehlberger, Guenter, et al. “Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study.” *Journal of Documentation*, vol. 75, no. 5, 2019, pp. 954–76, https://doi.org/10.1108/JD-07-2018-0114.

Multiple Authors. *Proceedings SDIUT99 The 1999 Symposium on Document Image Understanding Technology*. 1999.

Nagy, George. “Digitizing, Coding, Annotating, Disseminating, and Preserving Documents.” *Proceedings of the 2006 International Workshop on Research Issues in Digital Libraries*, edited by P. Majumder et al., ACM, 2006.

---. “Document Analysis Systems That Improve with Use.” *International Journal on Document Analysis and Recognition*, vol. 23, no. 1, 2020, pp. 13–29, https://doi.org/10.1007/s10032-019-00344-x.

---. “Document Image Analysis: Automated Performance Evaluation.” *Document Analysis Systems*, edited by A. L. Spitz and A. Dengel, World Scientific, 1995, pp. 137–56.

---. “Optical Character Recognition: An Illustrated Guide to the Frontier.” *Procs. Document Recognition and Retrieval VII, SPIE - The International Society for Optical Engineering*, 1999, pp. 58–69, https://doi.org/10.1117/12.373511.

---. “The Lifetime Reader.” *IEEE Pervasive Computing*, vol. 17, no. 4, 2018, pp. 86–95, https://doi.org/10.1109/MPRV.2018.2873848.

Nartker, Thomas A., et al. “Software Tools and Test Data for Research and Testing of Page-Reading OCR Systems.” *Proceedings, 2005 IS&T/SPIE Symposium on ELECTRONIC IMAGING SCIENCE & TECHNOLOGY*, 2005.

Nousiainen, Sami. *Report on File Formats for Hand-Written Text Recognition (HTR) Material*. 2016.

Ntirogiannis, Konstantinos. *Document Image Binarization*. 2014, http://users.iit.demokritos.gr/$\sim$bgat/DIBCO2009/.

O’Brien, Sean, and Dhia Ben Haddej. *Optical Character Recognition*. 2012. PhD Thesis.

Philips, James P., and Nasseh Tabrizi. “Historical Document Processing : A Survey of Techniques, Tools, and Trends.” *Journal of Data Mining and Digital Humanities*, 2020, pp. 1–30.

Pletschacher, Stefan, and Apostolos Antonacopoulos. *D-TR4.2 – Typewritten OCR Prototype*. 2011, pp. 1–7.

Rakshit, Sandip, et al. “Recognition of Handwritten Textual Annotations Using Tesseract Open Source OCR Engine for Information Just In Time (IJiT).” *Proc. Int. Conf. on Information Technology and Business Intelligence*, IMT, 2009, pp. 117–25.

Rangoni, Yves, et al. “OCR Based Thresholding.” *MVA2009 IAPR Conference on Machine Vision Applications*, 2009.

Rice, Stephen V., Junichi Kanai, et al. *A Report on the Accuracy of OCR Devices*. Information Science Research Institute, 1992, http://www.stephenvrice.com/images/Rice92-02.pdf.

---. *An Evaluation of OCR Accuracy*. Information Science Research Institute, 1993, pp. 1–25.

Rice, Stephen V. *Measuring the Accuracy of Page-Reading Systems*. 1996. University of Nevada, Las Vegas, PhD Thesis.

Rice, Stephen V., Frank R. Jenkins, et al. *The Fifth Annual Test of OCR Accuracy*. April, Information Science Research Institute, 1996, pp. 1–44.

---. *The Fourth Annual Test of OCR Accuracy*. Information Science Research Institute, 1995, pp. 1–39.

Rice, Stephen V., Junichi Kanai, et al. *The Third Annual Test of OCR Accuracy*. Information Science Research Institute, 1994, pp. 1–29.

Rice, Stephen V., and Thomas A. Nartker. *The ISRI Analytic Tools for OCR Evaluation: Version 5.1*. Information Science Research Institute, 1996, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.216.9427&rep=rep1&type=pdf https://github.com/eddieantonio/ocreval/blob/master/user-guide.pdf.

Santos, Eddie Antonio. "OCR Evaluation Tools for the 21 St Century." *Proceedings of the 3rd Workshop on Computational Methods for Endangered Languages Papers*, 2019.

Sauvola, J., and M. Pietikäinen. "Adaptive Document Image Binarization." *Pattern Recognition*, vol. 33, no. 2, 2000, pp. 225–36, https://doi.org/10.1016/S0031-3203(99)00055-2.

Seljan, S., et al. "From Digitisation Process to Terminological Digital Resources." *2013 36th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2013 - Proceedings*, 2013, pp. 1053–58.

Smith, David A., and R. Cordell. *A Research Agenda for Historical and Multilingual Optical Character Recognition*. Northeastern University, 2018.

Smith, Ray. "An Overview of the Tesseract OCR Engine." *Proceedings of the Ninth International Conference on Document Analysis and Recognition, ICDAR '07*, IEEE, 2007, pp. 629–33.

Smitha, et al. "Document Image Analysis Using Imagemagick and Tesseract-Ocr." *International Advanced Research Journal in Science, Engineering and Technology*, vol. 3, no. 5, 2016, pp. 108–12, https://doi.org/10.17148/iarjset.2016.3523.

Stančić, Hrvoje, and Željko Trbušić. "Evaluating and Improving OCR Efficiency." *Moderna Arhivistika*, vol. 3, no. 1, 2020.

---. "Optimisation of Archival Processes Involving Digitisation of Typewritten Documents." *Aslib Journal of Information Management*, vol. 72, no. 4, 2020, pp. 545–59, https://doi.org/10.1108/AJIM-11-2019-0326.

Strange, Carolyn, et al. "Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers." *DHQ: Digital Humanities Quarterly*, vol. 8, no. 1, 2014.

Tafti, Ahmad P., et al. "OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym." *International Symposium on Visual Computing*, 2016.

Thoma, George R., et al. "Design of a Digital Library for Early 20th Century Medico-Legal Documents." *Research and Advanced Technology for Digital Libraries: 10th European Conference, ECDL 2006*, edited by Gonzalo J. et al., Springer-Verlag, 2006, pp. 147–57.

Tomaschek, Martin. *Evaluation of Off-the-Shelf OCR Technologies*. 2017. PhD Thesis.

Traub, Myriam C., et al. "Impact Analysis of OCR Quality on Research Tasks in Digital Archives." *Research and Advanced Technology for Digital Libraries, 19th International Conference on Theory and Practice of Digital Libraries, TPDL*, edited by S. Kapidakis et al., Springer-Verlag, 2015, pp. 252–63.

Trbušić, Željko. "Mogućnosti Implementacije Sustava Za Optičko Prepoznavanje Znakova Tijekom Prihvata Gradiva u Arhivske Informacijske Sustave." *Radovi 52. Savjetovanja Hrvatskih Arhivista*, 2020.

Ul-Hasan, Adnan. *Generic Text Recognition Using Long Short-Term Memory Networks*. 2016. PhD Thesis.

U.S. Department of Commerce / National Bureau of Standards. *Guideline for Optical Character Recognition Forms*. 1976.

Vamvakas, G., et al. “A Complete Optical Character Recognition Methodology for Historical Documents.” *DAS 2008 - Proceedings of the 8th IAPR International Workshop on Document Analysis Systems*, 2008, pp. 525–32, https://doi.org/10.1109/DAS.2008.73.