# Optical Character Recognition (OCR) Resources

Joel Kalvesmaki

License: CC BY-4.0

The items provided on this handout are intended to accompany the first of four 2021 North American Patristics Society workshops, devoted to optical character recognition.

## Paid products

**Adobe Acrobat**: Consumer Off-The Shelf (COTS) software, part of Adobe’s Creative Suite. Acrobat has decent OCR for everyday purposes, in major modern languages. <https://acrobat.adobe.com/us/en/>

**ABBYY Reader**: COTS with graphic user interface-driven customization of the OCR process. ABBYY normally provides much better results than Acrobat, but it lacks support for ancient languages. <https://www.abbyy.com/>

**Transkribus**. Developed at the University of Innsbruck, Transkribus is a comprehensive platform for the digitization, AI-powered recognition, transcription and searching of historical documents. The primary focus is on handwriting recognition, and so it works well for traditional OCR tasks as well. Transkribus models are currently focused on handful of languages, e.g., Arabic, English, Ethiopic, Old German, Polish, Bangla, Hebrew, and Dutch. You can try the service for free, but you will need to pay for extensive use. The organization that emerged from the academic setting is a cooperative, which means that when you become a paid member, you become an owner too. <https://readcoop.eu/transkribus>

## Free products

**Gamera**. A Python-based OCR project. Requires background knowledge of Python and system configuration. The product itself provides a graphic-user interface for minute training of OCR systems. <https://gamera.informatik.hsnr.de/addons/ocr4gamera/index.html>

**Tesseract**. Originally developed by Hewlett-Packard as proprietary software in the 1980s, it was released as open source in 2005 and development has been sponsored by Google since 2006. Includes a panoply of language training modules, including Greek, Syriac, Ethiopic, Armenian, etc. Requires familiarity with the command line. <https://github.com/tesseract-ocr>

**i2OCR**. Free online OCR service built on Tesseract. Includes the Greek training model. <https://www.i2ocr.com/>

**Kraken**. Developed at École Pratique des Hautes Études, Université PSL, this turn-key OCR system is optimized for historical and non-Latin script material. Mac or Linux only. Requires familiarity with the command line. <https://github.com/mittagessen/kraken>

**Nidaba**: A full pipeline developed by the Open Greek and Latin Project to OCR texts. Requires background knowledge of Python and system configuration. <https://openphilology.github.io/nidaba/>

## Projects using OCR to produce texts relevant to early Christian studies

**Archive.org**: The world’s archive service for the Web, Archive.org, provides OCR support for its texts. The OCR is based on ABBYY.

**Open Greek and Latin Project**: major initiative by the Universität Leipzip to represent every source text produced in Classical Greek or Latin from antiquity through the present, including texts preserved in manuscript tradition as well as on inscriptions, papyri, ostraca and other written artifacts. <https://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/>

**Open Patrologia Graeca**: Greek and Latin text generated by two open source OCR engines, OCRopus (https://github.com/tmbdev/ocropy) and Tesseract (https://github.com/tesseract-ocr). <http://tinyurl.com/nlvhy9b> <https://github.com/OGL-PatrologiaGraecaDev>

**Patrologia Latina**: Latin texts OCRed by the Open Greek and Latin project. <https://opengreekandlatin.github.io/church_fathers-dev/>

**CSEL**: Corpus Scriptorum Ecclesiasticorum Latinorum OCRed into XML by the Open Greek and Latin project. <https://github.com/OpenGreekAndLatin/csel-dev>

## Large corpora

Listed below are some major textual corpora relevant to early Christian studies. These repositories allow easy bulk download of the texts for research and development.

**First1KGreek**: <https://opengreekandlatin.github.io/First1KGreek/>

**Perseus Greek**: <https://github.com/PerseusDL/canonical-greekLit>

**Perseus Latin**: <https://github.com/PerseusDL/canonical-latinLit>

**Syriac corpus** : <https://github.com/srophe/syriac-corpus>

**Coptic Scriptorium** : <https://github.com/CopticScriptorium/corpora>

## Auxiliary Tools

**PDFtoText**: if you have a PDF with text already inside, you can extract all the plain text with this handy tool. PDFtoText is included by default with many Linux distributions, and is also available for Windows as part of the Xpdf Windows port. <https://www.xpdfreader.com/pdftotext-man.html>

**PDFtoPNG**: if your PDF does not have extractable text, and you need to get a series of images, this tool is one of the quickest and best. By default it converts PDF files to color image files in PNG. Output is one image file for each page. See also PDFtoPPM. <https://www.xpdfreader.com/pdftopng-man.html>

**Winmerge** (windows only): Convenient way to compare two or three text files. <https://winmerge.org/>