Automatic Text Recognition (OCR/HTR)

Adi Keinan-Schoonbaert

2024-11-27

Demystifying the process behind automatic text recognition and offering practical tips on how anyone can get started in their your own library.

Introduction

Cultural heritage organisations have been digitising their historical collections for several decades, outputting large-scale sets of digitised collection items - printed books, newspapers, manuscripts, maps, and many other content types as simple JPG, PNG, or TIFF images. Though these collections are made available for users to read online, there is often an important layer of accessibility that is missing as the text appearing in these images is not always made searchable, editable, or analysable.

Automatic Text Recognition (ATR) refers to the process of using software to convert images of text, such as scanned or photographed documents, photos, or printed pages, into machine-readable text. ATR is primarily associated with Optical Character Recognition (OCR) (for print materials) and Handwritten Text Recognition (HTR) technologies (for handwritten materials) that enable computers to identify and extract text from various sources. These technologies are critical for digitising historical and cultural documents, making them searchable, accessible, and preservable for future generations.

Though ATR technologies have been around in some form or another for nearly as long as we've been digitising, results can be patchy for cultural heritage collections. This may be down to costs as these technologies are typically considered an add-on to the basic photographic or scanning services and institutions or projects don't have the funds available to cover that aspect of digitisation. Or the traditional OCR technology at any given time simply isn't developed enough to provide very accurate results to handle the anomalies of, for instance, quite worn or warped manuscript pages. Or in the case of handwritten texts, particularly low-resource languages, the existing technology simply cannot cope and text outputs are illegible. The advent of machine learning technologies applied to the challenge of Automatic Text Recognition of handwritten or print materials, and with it, the ability to train a software to recognise specific

text based on what it has already seen, however, provides huge opportunities to transform accessibility to the text contained within all manner of historical digitised images.

So what is the difference between the decades-old OCR technology and the more recent, AI/Machine Learning-based software? Traditional OCR systems rely on a set of predefined rules and templates to recognise characters. These rules are often based on the geometric properties of the text, such as the shapes and patterns of characters. On the other hand, Machine Learning-based OCR, or in other words, Automatic Text Recognition that leverages Artificial Intelligence, uses neural networks to learn to recognise characters from large datasets. These systems can be trained to automatically learn to identify the features that distinguish different characters and predict with some probability what that character could be. AI-OCR can handle a wide variety of fonts, sizes, and styles, and is robust to noisy, distorted, or low-quality images. The models can generalise well to different types of documents and text layouts; and they can continuously improve by retraining on new data, which allows them to adapt to new types of text and writing styles over time.

When diving deeper into AI-OCR, you may encounter the term 'Ground Truth'. Ground Truth refers to the accurate, manually created and verified data used as a standard or benchmark to both train and evaluate the performance of OCR/HTR systems. Ground truth data consists of the correct transcription of a document, including precise details about characters, words, and layout. It is used to 'teach' software how to recognise and transcribe text, as well as measure the accuracy of OCR/HTR outputs by comparing the machine-recognised text against this gold-standard reference.

Automatic Text Recognition typically involves multiple stages or elements, which include image pre-processing, binarisation, layout analysis, and text recognition.

1. **Image Pre-processing**: Pre-processing is the first step and focuses on enhancing the quality of the input image to improve recognition accuracy. This may include tasks such as noise reduction, skew correction (to align slanted images), and contrast adjustment. These operations are crucial when dealing with old or degraded documents, where imperfections in the image can hinder the subsequent stages.
2. **Binarisation**: Binarisation converts a grayscale or colour image into a binary image, typically using black for the text and white for the background. This simplifies the image, making it easier for the OCR/HTR system to distinguish between the text and non-text elements. This step is especially important for historical documents, where stains, fading, or other degradation can confuse the OCR/HTR engine.

Examples of binarisation and its effects on legibility, created by Peter Smith who worked to improve Chinese HTR processes

1. **Layout Analysis**: In layout analysis, the system identifies the structure of the document, distinguishing between various elements such as paragraphs, columns, headings, footnotes, and images. It is essential for documents with complex formatting (e.g. newspapers or tables) to ensure that the text is correctly segmented and processed. This step may also involve detecting text regions in multi-layout documents or distinguishing between handwritten and printed text.
2. **Text Recognition**: The core of the OCR/HTR process is text recognition, where the system identifies individual characters, words, and sentences within the segmented text regions. Modern OCR/HTR engines use pattern recognition techniques or machine learning algorithms, such as neural networks, to improve accuracy. Some systems also perform language modelling, where the recognised text is checked against a dictionary or corpus to ensure contextual correctness, particularly useful for older languages or scripts.
3. **Post-Processing**: OCR/HTR often produces errors, especially when dealing with degraded, old, or complex manuscripts. To improve accuracy, post-processing correction is

crucial. One effective method for refining OCR/HTR outputs is crowdsourcing text correction, where volunteers, often through online platforms, manually review and correct transcribed text. This method leverages the knowledge and dedication of the public to handle the nuances of historical documents that automated systems struggle with, such as obscure spellings or unusual handwriting styles. Recent experiments in automating text recognition have leveraged Large Language Models (LLMs) to enhance the accuracy of OCR/HTR systems. LLMs, such as GPT-based models, are particularly adept at understanding context and handling ambiguities in textual data, which makes them valuable for recognising and correcting errors in historical documents. LLMs can also be fine-tuned to specific historical corpora, allowing them to better interpret unique vocabulary, syntax, or stylistic variations. And, it should be mentioned that once you have perfectly post-processed OCR/HTR results, you can use those as Ground Truth to retrain models and improve them!

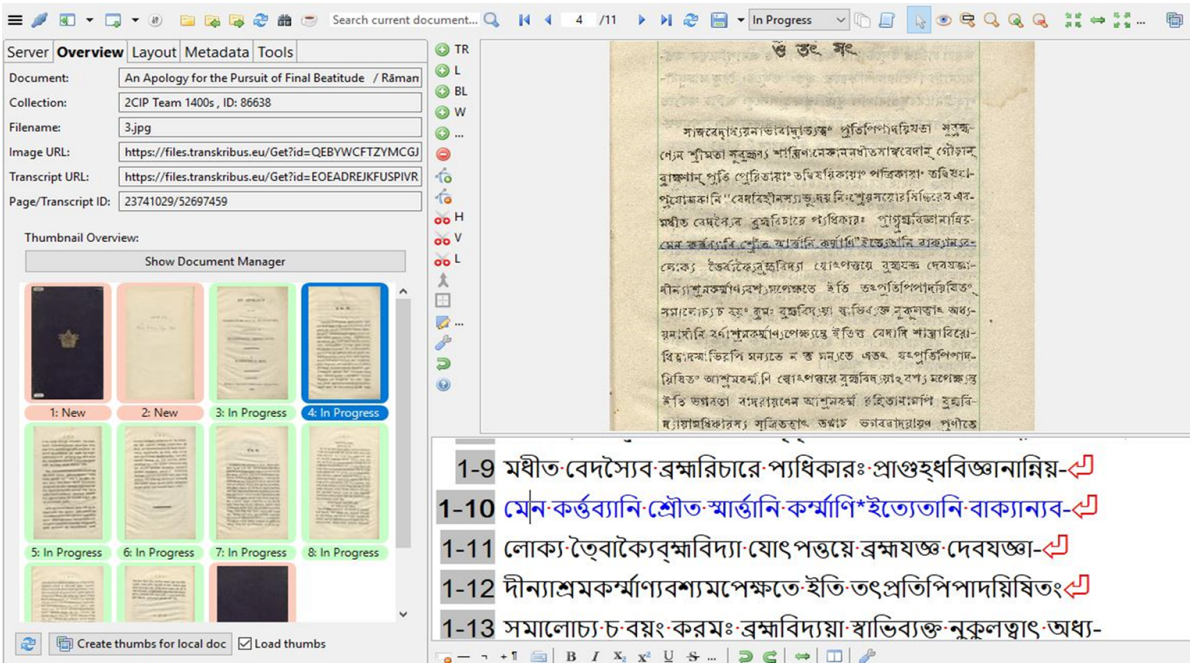
Relevance to the Library Sector (Case Studies/Use Cases)

Keeping up with, and even contributing to, developments in Automatic Text Recognition technology is vital for heritage organisations and especially for the library sector, where large volumes of historical and cultural documents are preserved. ATR technologies provide several key benefits:

- **Accessibility**: ATR technologies are essential for libraries working to make their text content available for users. Once digitised and converted into searchable text, documents become accessible to a global audience. Researchers and the general public can search and access texts remotely, making previously hidden information available. Creating searchable documents allows users to quickly locate specific terms or phrases within vast collections. This enhances the research process and saves time.
- **Content Enrichment**: ATR can help enrich library records by enhancing their metadata; therefore assisting libraries in delivering a better service to their users. Different entities could be extracted from the text, such as author or place of publication, as well as subjects and descriptions. These could be used to enhance catalogue records for the benefit of library users.
- **Digital Research**: By converting historical texts into machine-readable formats, ATR technologies support digital humanities projects, where large-scale analysis, such as text mining or linguistic research, can uncover new insights into history, culture, and language development. The ability to extract text from digitised items is fundamental for any downstream tasks and enables unlocking content for large-scale analysis, e.g. text mining, Natural language processing (NLP), Named Entity Recognition (NER), sentiment analysis or topic modelling.

Case Studies

In collaboration with PRImA Research Lab, the British Library ran several OCR/HTR competitions with the aim of encouraging the development of state-of-the-art in text recognition software, facilitating dialogue around the challenges and opportunities of ATR, and creating openly licensed ground truth datasets. Competitions around early Bengali books and Quarterly Lists were run as part of the Two Centuries of Indian Print project, and the project used Transkribus to create OCR transcriptions for the Bengali books, in collaboration with the School of Cultural Texts and Records at Jadavpur University. Additional competitions were focused on finding solutions to automatically transcribe historical Arabic scientific manuscripts, using materials digitised by the Library and published on the Qatar Digital Library (QDL).



Screenshot from Transkribus, used as the ATR engine in the Two Centuries of Indian Print project

The Open Islamicate Texts Initiative (OpenITI) is a collaborative project involving researchers from Aga Khan University's Institute for the Study of Muslim Civilisations in London, the Roshan Institute for Persian Studies at the University of Maryland, College Park, and Universität Hamburg. Its goal is to establish the digital infrastructure needed for the study of Islamicate cultures. OpenITI focuses on building digital resources for Islamicate studies by enhancing optical character recognition (OCR) and handwritten text recognition (HTR) for Arabic-script texts, creating standardised OCR and HTR outputs and text encoding, and developing platforms for collaborative work on Islamicate text corpora and digital editions.

The Legacies of Curatorial Voice in the Descriptions of Incunabula Collections at the British

Library project was part of Digital Curator Dr Rossitza Atanassova's AHRC-RLUK funded Professional Practice Fellowship Project (2022–23). It aimed to explore innovative methods for working with digitised catalogues, enhancing the discoverability and usability of the collections they document. The research centred on the *Catalogue of Books Printed in the 15th Century Now at the British Museum* (BMC), published between 1908 and 2007, which describes over 12,700 volumes in the British Library's incunabula collection. It used Transkribus as the ATR software and also as a search and publishing web platform. The project could then apply computational techniques and corpus analysis of the catalogue data thanks to the availability of OCRed text and provides fresh insights into this resource!

Hands-on activity and other self-guided tutorial(s)

Here are some excellent resources to get you started:

Automatic Text Recognition: Harmonising ATR workflows: This DARIAH-EU-supported resource is a *great* place to start. This website features a set of video tutorials on ATR in English, French and German, and also includes really useful blog posts, articles and links.

Check out the Transkribus YouTube channel for a set of beginner-friendly tutorial videos, learning how to get started with Transkribus to effectively digitise and preserve historical documents.

If you're interested in eScriptorium, have a look at their thorough documentation and tutorials on YouTube created by the OpenITI project. These are split into five parts: Part I, Part II, Part III, Part IV, and Part V.

This step-by-step guide will teach you how to use the Tesseract open-source software, and it also recommends some software that helps prepare documents for Tesseract use.

For more advanced practitioners, there's an introductory course by Dr William Mattingly teaching you how to automate OCR in Python. It includes using OpenCV (an open-source library specialising in computer vision and machine learning tasks) and Pytesseract, an OCR tool in Python. Helpfully, this course functions alongside a YouTube series of tutorials on OCR in Python.

Another one for the more confident practitioners is this Programming Historian OCR with Google Vision API and Tesseract tutorial by Isabelle Gribomont, published in March 2023.

This Programming Historian OCR and Machine Translation course by Andrew Akhlaghi (published in January 2021) uses Tesseract for OCR and takes the results to another level - translation.

Recommended Reading/Viewing

This is a great blog post by Chris Woodform, simply entitled Optical character recognition (OCR). It's a good place to start! It explains what OCR is and how it works, and even looks at the history of this technology.

Awesome OCR is a resource created by Konstantin Baierer from the OCR-D project. It hasn't been updated for a while, however, it includes really useful and comprehensive lists of software tools, libraries and literature. It also includes a list of ground truth datasets, so well worth taking a look.

The IMPACT Centre of Competence is a useful OCR resource. In their words, IMPACT is a 'not for profit organisation with the mission to make the digitisation of text "better, faster, cheaper" and to further advance the state-of-the-art in the field of document imaging, language technology and the processing of historical text.' You can find datasets, training materials, blogs and more!

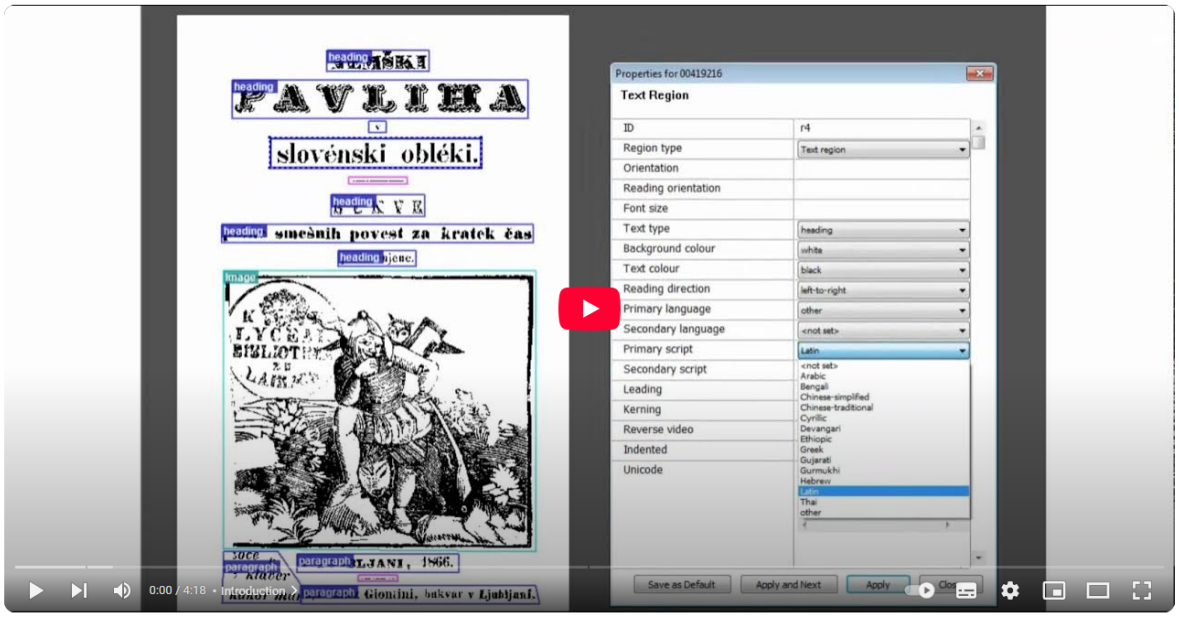


Figure 1: An Introduction to Ground Truth Production with the IMPACT Project

An excellent guide to newspaper OCR and data analysis is this Short Guide to Historical Newspaper Data, Using R by Yann Ryan.

Many excellent data analysis tools such as NLP can also be found on this GitHub page, which collated materials used for the Text to Tech workshop at the Digital Humanities Oxford Summer School, by Kaspar von Beelen, Mariona Coll Ardanuy and Federico Nanni (and this is another brilliant introduction into NLP!).

There are many papers on OCR/HTR but here's a small selection:

- Khan, R., Gupta, N., Sinhababu, A., & Chakravarty, R. (2023). Impact of Conversational and Generative AI Systems on Libraries: A Use Case Large Language Model (LLM). *Science & Technology Libraries*, 1–15. https://doi.org/10.1080/0194262X.2023.2254814
- Neudecker, C., Baierer, K., Federbusch, M., Boenig, M., Würzner, K. M., Hartmann, V., & Herrmann, E. (2019). OCR-D: An end-to-end open source OCR framework for historical printed documents. In *Proceedings of the 3rd international conference on digital access to textual cultural heritage*, 53-58. https://dl.acm.org/doi/10.1145/3322905.3322917
- Nguyen, T.T.H., Jatowt, A., Coustaty, M., & Doucet, A. (2021). Survey of Post-OCR Processing Approaches. *ACM Computing Surveys (CSUR)*, 54(6), 1-37. https://doi.org/10.1145/3453476
- Smith, D.A. & Cordell, R. (2018). A Research Agenda for Historical and Multilingual Optical Character Recognition.
- Van Strien, D., Beelen, K., Ardanuy, M., Hosseini, K., McGillivray, B., & Colavizza, G. (2020). *Assessing the impact of OCR quality on downstream NLP tasks*. SCITEPRESS - Science and Technology Publications. https://doi.org/10.17863/CAM.52068

Some journals to look out for include, for example, the International Journal on Document Analysis and Recognition (IJDAR), Pattern Recognition, or the IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI).

The team behind the Hypotheses 'Automatic Text Recognition: Harmonising ATR workflows' resource created this insightful road map to help you get started with Automatic Text Recognition. The road map is divided into three sections, and asks you to consider the following questions:

I - General Information

1. What type of texts/text collections do you work with?
2. How can you integrate Automated Text Recognition in your workflow?

II - Technical information

1. Why do you want to use ATR?
2. What is your objective?
3. Do you have technical and financial resources?

III - Setting up ATR in your project

1. Choose your tool
2. Is there transcription data that can be reused (e.g. training data)?
3. Is there an appropriate generic model?
4. What are your transcription rules?

5. Do you plan to share your ATR training data? What are your ATR predictions?
6. Where can you share data? In which format?

This document also includes links to tutorials and documentation, a selection of articles, ATR tools and where to find ATR ground truth datasets.

When it comes to ground truth datasets, a good place to look is the HTR-United resource, which includes training datasets used for both transcription or segmentation models, for different periods and styles of writing.

OCR-D is a good place to visit too, and it has a ground truth repository available here.

Finding Communities of Practice

Some tools and platforms have solid communities around them which you can get in touch with - or be a part of. Transkribus, for example, has an active Facebook page and a user group, where Transkribus users can ask questions and get support from the Transkribus team and from each other.

Several conference series discuss advances in OCR/HTR, and would be a good place to meet colleagues and hear about interesting ATR projects. For example, the International Conference on Document Analysis and Recognition (ICDAR) and the International Conference on Frontiers in Handwriting Recognition (ICFHR) have been well established. Alternatively, many Digital Humanities conferences will have papers on this topic, e.g. the ADHO DH conference serie or AI4LAM Fantastic Futures conferences.