Automatic Text Recognition Road Map to get you started

Before using Automatic Text Recognition, you need to consider whether it will be relevant to your project. Here are a few questions to get you started and help you build a data management plan.
*Before answering the questions, you can consult step 0: introduction (*https://youtu.be/TDYD5ZMenTg*)*

I- General Information

1- What type of texts/text collections do you work with? *Question*	*Your answer*	*Advice*
a - Does your text collection consist of prints or manuscripts, or both?		Printed documents are generally easier to process with *Automated Text Recognition* (ATR).
b - Are your documents digitised? What is the quality of the digitisation?		ATR requires high-definition images to achieve the best results. We recommend an image resolution between 150 and 300 DPI. Please note that high-resolution scanning can be expensive and image rights difficult to obtain if you wish to reuse them or share training data. Finally, digitising documents takes time, which can delay the start of ATR processing. (*see step 1*)
c - Do you need to optimise the digitisation? For example, straighten, crop, or improve image quality.		Generally speaking, if you are working with collections from a heritage institution or portals such as e-codices or *Gallica*, you will not need to optimise images. But if you are working with microfilm, damaged documents, or documents with complex layouts, optimisation can improve ATR results. (*see step 2)*
d - How many images do you have in your corpus?		In the case of manuscripts, using ATR is only worthwhile for large collections. The more images you have, the more you will need to use command line tools. At some point in the process—for instance, for prediction or in training a model from scratch—the interfaces will not suffice and you will have to work in command lines. If you are dealing with 500 images or fewer, the text and image recognition platform Transkribus is available to you for free.
e - How complex is the layout of your documents? Do all the pages have the same layout? Are there additions in the margins? Do different parts of the text have different orientations on the page?		Layout analysis can still be a major issue in the automation of the work process. Depending on the complexity of the layout, you will need to spend more or less time defining and training the segmentation model and/or correcting the generated segmentation. If the segmentation is not done properly, it will create noise in the transcription. (*see step 3*)
f - What are the specificities of your documents? For instance, in which century were they written and which types of writings or fonts do they display? Is the corpus homogeneous in this regard?		Not all documents are equally difficult to read. For instance, manuscripts from the 13th century are generally easier to read than those from the 16th century.
g- Are you working with official or administrative documents, illustrated documents, personal notes, etc.? Is there more than one type of document at hand?		Official documents are easier to read than more personal ones. Be aware of the specific difficulties of the documents with which you are working.
h - What is the language of the corpus? Are you working with monolingual or multilingual collections?		ATR templates are often language-sensitive and specialise in just one language. You can either divide your collection on the basis of language or find/build multilingual templates.
i - How complex is the text itself: are there abbreviations or special characters?		You need to have a clear notion of the content of your corpus in order to be able to choose your tools and models accordingly and apply these to the transcription rules. For instance, do you decide to keep abbreviations or not? What sign will you use to symbolise a particular feature of your document such as superscript letters?

References to get started

- *Links to tutorials and documentation*
  - Comment faire lire des gribouillis à mon ordinateur ? par Alix Chagué. MATE-SHS.
  - Prendre en main eScriptorium. LECTAUREP, par Alix Chagué.
  - Moving from Transkribus to eScriptorium, eScripta, par Peter Stokes.
  - Chahan Vidal-Gorène, « La reconnaissance automatique d'écriture à l'épreuve des langues peu dotées »,Programming Historian en français 5 (2023), https://doi.org/10.46430/phfr0023.
  - Awesome-OCR, tools and libraries and literature related to Optical Character Recognition
- *Selection of articles to get started*
  - Chagué A, Chiffoleau F. "An accessible and transparent pipeline for publishing historical egodocuments". In: *WPIP21 - What's Past Is Prologue: The NewsEye International Conference*, 2021. https://hal.archives-ouvertes.fr/hal-03173038
  - Chagué A, Clérice T, Romary L. "HTR-United : Mutualisons la vérité de terrain !" Published online October 2021, https://hal.archives-ouvertes.fr/hal-03398740
  - Clérice T, Vlachou-Efstathiou M, Chagué A, "CREMMA Medii Aevi: Literary manuscript text recognition in Latin". *Journal of Open Humanities Data*, 2023, 9, pp.4. (10.5334/johd.97). (hal-03828353v5)6
  - Fischer, A., Liwicki, M., & Ingold, R. (2020). *Handwritten Historical Document Analysis, Recognition, and Retrieval—State of the Art and Future Trends* (Vol. 89). WORLD SCIENTIFIC. https://doi.org/10.1142/11353
  - Gabay S, Camps JB, Pinche A, Jahan C. "SegmOnto: common vocabulary and practices for analysing the layout of manuscripts (and more)". *In: 16th International Conference on Document Analysis and Recognition (ICDAR 2021)*. 2021. https://hal.archives-ouvertes.fr/hal-03336528
  - Hodel T, Schoch D, Schneider C, Purcell J. "General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example". *Journal of Open Humanities Data*. 2021;7(0):13. https://doi.org/10.5334/johd.46
  - Kahle P, Colutto S, Hackl G, Mühlberger G. "Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents". In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol 04. ; 2017:19-24. https://doi.org/10.1109/ICDAR.2017.307
  - Kiessling B, Tissot R, Stokes P, Ezra DSB. "eScriptorium: An Open Source Platform for Historical Document Analysis". In: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. Vol 2. ; 2019:19-19. https://doi.org/10.1109/ICDARW.2019.10032
  - Kiessling B. Kraken "an Universal Text Recognizer for the Humanities". *In: CLARIAH; 2019*. https://doi.org/10.34894/Z9G2EX
  - Neudecker C, Baierer K, Federbusch M, et al. "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. DATeCH2019. Association for Computing Machinery*; 2019:53-58. https://doi.org/10.1145/3322905.3322917
  - Pinche A, "Generic HTR Models for Medieval Manuscripts. The CREMMALab Project". 2023. https://hal.science/hal-03837519v3
  - Reul C, Tomasek S, Langhanki F, Springmann U. "Open Source Handwritten Text Recognition on Medieval Manuscripts Using Mixed Models and Document-Specific Finetuning". In: Uchida S, Barney E, Eglin V, eds. *Document Analysis Systems. Lecture Notes in Computer Science*. Springer International Publishing; 2022:414-428. https://doi.org/10.1007/978-3-031-06555-2_28
  - Ströbel PB, Clematide S, Volk M. "How Much Data Do You Need? About the Creation of a Ground Truth for Black Letter and the Effectiveness of Neural OCR". In: *Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association*; 2020:3551-3559. 2021. https://aclanthology.org/2020.lrec-1.436

2- How can you integrate Automated Text Recognition in your workflow?		
Question	*Your Answer*	*Advice*
a - At what stage will you begin the ATR processing for your project?		As ATR facilitates the acquisition of texts, it should be the first step in your project.
b - How much time can you dedicate to the ATR phase?		It is preferable to anticipate having to dedicate a considerable amount of time to this part of the project workflow, as it determines the quality of the textual data. Set aside at least six months in order to prepare the training data, correct the transcriptions, and anticipate segmentation problems.
c - Have you determined the type of text output you want to have and how you intend to reach it?		Your pipeline will determine your transcription rules. These rules are extremely important to ensure the homogeneity of your data.
d - What kind of analysis would you like to conduct with your corpus?		Your research objectives are key in designing your pipeline and making sure that you have integrated all the necessary textual features.
e- What type of text post-processing protocols will you need? e.g. word segmentation, lemmatisation, named entity recognition.		Text post-processing protocols are determined by the type of text you want to get—raw text for quantitative analysis or pre-edited text. (*see step 4*)
f - Have you integrated data quality checks? How would you like to proceed with them?		Do not forget to include checkpoints to verify data quality at each step in the workflow, which is likely to impact the text output in your corpus. (*see step 5*)
g - Where will you store your data? In what format? In which repositories?		It is crucial to decide what you can share and where you will do so in order to design your pipeline and data management plan. You do not have to keep everything. (*see step 6*)

Available resources (non-exhaustive list)

ATR tools

eScriptorium + kraken (free, Open Source)
OCR-D
OCR4all (free, Open Source)
Transkribus (mixed)
Teklia (commercial)
Calfa Vision (commercial)

Where to find ATR ground truth catalog

HTR-United
OCR-D catalog
other data repositories
GitHub
GitLab
Nakala

Where to find ATR models

Transkribus catalogue
HTR-United
ZENODO OCR/HTR model repository

II- Technical information

1- Why do you want to use ATR? *Answers*	*Advice*	*Remarks*
☐ To speed up the acquisition of text	Identify the tools and ATR models available in order to use a ready-to-use solution that can be directly applied to the corpus.	Limitations of ATR: - Standardisation of the representation of abbreviations, punctuation, or special characters. - Errors in the analysis of the layout of documents can introduce noise (e.g. recognition of marginal notes). - Weaker recognition of named entities.
☐ Impossibility of transcribing the corpus by hand	Identify the tools and ATR models available to you so you can use a solution that can be directly applied or fine-tuned to your s ocuments. Start collecting or creating training data to form your own template.	
☐ Others:	It is important to note that building an ATR model from scratch is expensive (costs of data creation and gathering), time-consuming, as well as energy-intensive. Moreover, the processing of complex layouts remains a major challenge.	

2- What is your objective? *Answers*	*Advice*	*Remarks*
☐ Quantitative analysis	You need a transcription with low precision, accuracy around 85%. You can use a generic model.	If the quality of ATR predictions is not sufficient to reach your research objectives, the quality of the generated text can be improved by introducing a human proofreading phase, creating more training data, or implementing post-treatment checks.
☐ Text repository	You need a transcription with medium precision, accuracy around 90%. You can use a specific or a generic model.	
☐ Scientific Editing	You need a transcription with high precision, accuracy around 95%. You should use a specific model.	

3- Do you have technical and financial resources? *Answers*	*Advice*	*Remarks*
☐ My project is an individual project	Keep it simple: try to look for as many ready-to-use solutions as possible.	The larger the project, the easier it is to find resources and expertise.
☐ My project is an collective project	Use the expertise and resources (infrastructures, models, data) of your partners to build your project on solid ground.	
☐ My project is not funded	Keep it simple: try to look for as many ready-to-use solutions as possible.	
☐ My project is funded	The project can grow depending on your budget and project objectives. You can train your own team or call in experts. You can hire engineer(s) to create or optimise your own pipeline. You can also finance the creation of training data. Finally, you can buy or pay for access to compute servers to train new generic models.	
☐ My team cannot be trained in ATR techniques	Keep it simple: try to look for as many ready-to-use solutions as possible.	
☐ My team can be trained in ATR techniques	You can expand your use of the ATR, fine-tune, or create your own model to adapt the ATR as close as possible to your research objectives.	

III- Setting up ATR in your project

1- Choose your tool *Answers*	*Advice*	*Remarks*
☐ eScriptorium	*eScriptorium* is not a totally ready-to-use solution: you need to train or download your model. But you are completely free to train your model through your chosen parameters with *Kraken*. Contact: https://cremmacall.sciencescall.org.	Once you have chosen a tool, you need to stay in the same environment, because data and models are not totally compatible between different tools.
☐ OCR-D	https://ocr-d.de/en	
☐ Transkribus	This solution is ready to use, but it is not totally open. You have to stay within the *Transkribus* environment, so you cannot run your training on the command line and choose your training parameters. Contact: https://readcoop.eu/transkribus/	
☐ Other:	Note that the less common your solution, the fewer ready-to-use resources are available to you.	

2- Is there transcription data that can be re-used (e.g. training data)? *Answers*	*Advice*	*Remarks*
☐ Transcriptions to be aligned	You will need to spend time creating data for segmentation and text alignment, for instance.	You can find available data and models in the *HTR-united* catalog, *Github*, or in *Zenodo*. With *Transkribus*, models will be accessible via the interface. When you collect data or reuse a model, make sure that the transcription guidelines used are appropriate for your project.
☐ Training data in XML format: ALTO or PAGE	Data are ready to use. You can gather data from different projects, but be careful about compatibility.	

3- Is there an appropriate generic model? *Answers*	*Advice*	*Remarks*
☐ Yes	You can use the model directly or fine-tune it on your corpus. Most of the time, 5 images with aligned text are enough to achieve results of around 95% accuracy if the generic model is appropriate.	
☐ No	You have to create a model from scratch, you can reuse data and/or create new ones. For a simple document, 20 images with aligned text may be enough.	

4- What are your transcription rules? *Answers*	*Advices*	*Remarks*
	Examples of transcription rules: CREMMA project transcription rules: CREMMA_Transcription_Guideli... - Thibault Clérice, Malamatenia Vlachou-Efstathiou, Alix Chagué. CREMMA Medii Aevi: Literary manuscript text recognition in Latin. Journal of Open Humanities Data, 2023, 9, pp.4. https://doi.org/10.5334/johd.97 - Ariane Pinche. Guide de transcription pour les manuscrits du Xe au XVe siècle. 2022. ⟨hal-03697382⟩	Transcription rules are very important to ensure the consistency of your corpus.

5- Do you plan to share your ATR training data? What are your ATR predictions? *Answers*	*Advice*	*Remarks*
	To share training data you may need to share images, so be sure to check that you have the rights to do so.	Sharing your data will help the community to grow and build more generic models while attracting greater visibility for your project.

6- Where can you share data? In which format? *Answers*	*Advices*	*Remarks*
	You can link your repositories to *HTR-united* for greater visibility.	