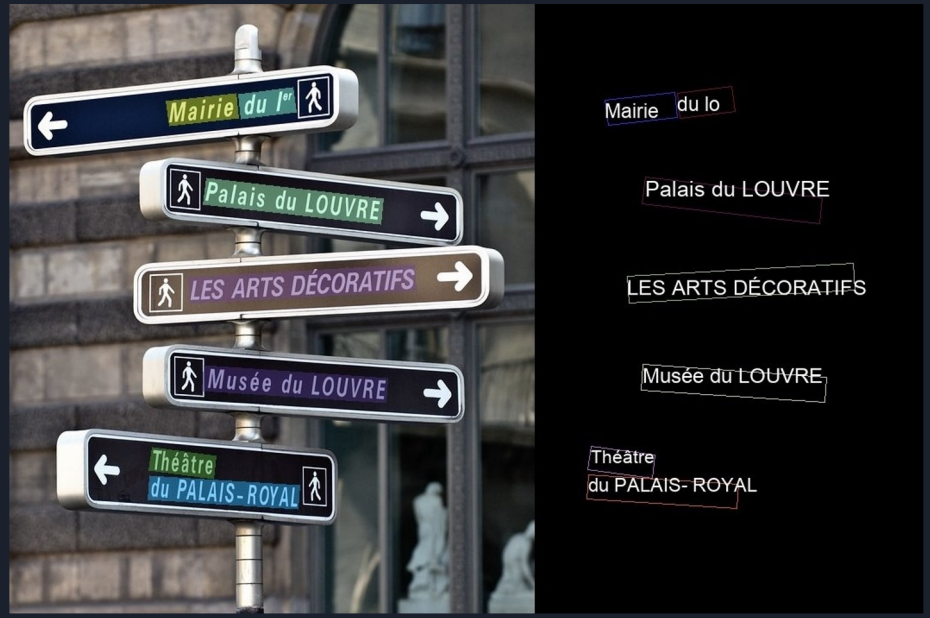
Introduction to OCR and LLM

Michal Torma, GBIF Norway, NHM, UIO

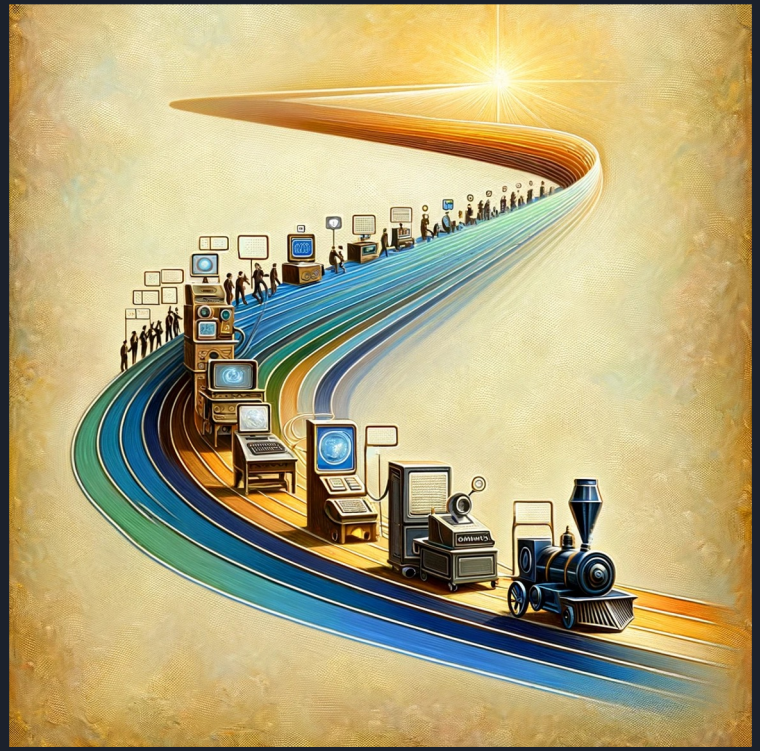
What is OCR

Optical Character Recognition (OCR) converts printed or handwritten text in images and documents into digital, editable text using advanced algorithms. This technology is widely used for digitizing paper documents, automating data entry, and enhancing accessibility for visually impaired individuals. OCR makes information processing faster, more accurate, and more efficient.



History and progress

Optical Character Recognition (OCR) dates back to the early 1900s when Emanuel Goldberg developed a machine to read characters and convert them into telegraph code. In the 1950s, OCR technology advanced with the creation of commercial systems like the ‘Gismo’ by David H. Shepard, used for processing checks and business forms. The 1970s saw significant improvements with the introduction of intelligent character recognition, which enhanced accuracy and versatility. Today, OCR is integral to digitization efforts, enabling efficient data entry, document management, and accessibility solutions.



State of the art today - Open Source

- Tesseract OCR:
- Developed by Google.
- Supports multiple languages.
- Highly customizable and integrated into various software.
- OCRmyPDF:
- Adds OCR text layer to PDF files.
- Uses Tesseract as the backend.
- Free and open-source with command-line interface.
- gImageReader:
- Graphical front-end to Tesseract.
- User-friendly interface for text extraction.
- Supports multiple input formats.

State of the art today - Paid Local

- ABBYY FineReader:
- High accuracy OCR software.
- Supports document comparison and conversion.
- Available for Windows and Mac.
- Readiris:
- Converts documents to editable formats.
- Integrates with popular cloud storage services.
- Features advanced text recognition and layout retention.
- OmniPage Ultimate:
- High-speed OCR with batch processing.
- Supports multiple languages and document types.
- Includes advanced PDF creation and editing tools.

State of the art today - Paid Cloud

- Google Cloud Vision OCR:
  - High accuracy and reliability.
  - Supports a vast array of languages.
  - Easily integrates with other Google Cloud services.
- Amazon Textract:
  - Extracts text, forms, and tables.
  - High scalability and integration with AWS ecosystem.
  - Pay-per-use pricing model.
- Microsoft Azure Cognitive Services OCR:
  - Provides OCR as part of Cognitive Services.
  - Supports image and PDF text extraction.
  - Integrated with other Azure services and tools.

Google vision API

Try it here: https://cloud.google.com/vision/docs/drag-and-drop

Price: 1000 images/month free

Additional 1000 images/month -> $1.5

Most capable from our testing. (even for handwritten cyrilics)

LLM - Large Language Models

What are LLMs really?

- Text predictors on steroids
- Predict one word (token) at the time
- Show emergent capabilities
- Fast pace of progress
- Resource hungry

There are more of them then you think

Private

- ChatGPT (OpenAI)
- Gemini (Google)
- Claude (Anthropic)
- Grok (xAI)

Open Source

- LLama (Meta)
- Hugging Face - 680544 free models for various use cases

HW limitations on the user

Price for user: ~$20/month

Price per API request is varied. ChatGPT:
INPUT: US$5.00 / 1M tokens

OUTPUT: US$15.00 / 1M tokens

Questions

Example of our prompt (this e.g. for Uzbekistan)

You are an adept herbarium digitization system, working on OCR text extracted from the images of scanned herbarium specimens from herbaria in Uzbekistan - possible languages on the labels include Uzbek, Russian, English and Latin.
Important Note! Label OCR text contains rulers (i.e. incremental cm counts from 1 to 30), color bars and other extraneous information, which should be ignored.
First you correct any obvious OCR errors, and then you extract ONLY the following Darwin Core terms, translating into English if possible:

- scientificName: Full scientific name, not containing identification qualifications.
- catalogNumber: Unique identifier for the record in the dataset or collection.
- recordNumber: Identifier given during recording, often linking field notes and Occurrence record.
- recordedBy: List of people, groups, or organizations responsible for recording the original Occurrence.
- year: Four-digit year of the Event.

…

If there are multiple valid values for a term, you separate them with "|". IMPORTANT: If you can't identify information for a specific term, and/or the term is blank, you skip the term in my response. You respond in minified JSON.