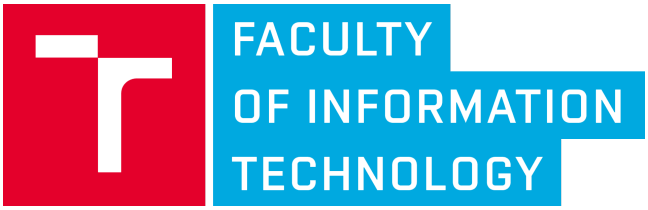


Manuál k Software pro extrakci informace z polostrukturovaných dokumentů

Michal Hradiš, Martin Kišš, Jan Kohút, Karel Beneš, Martin Kostelník

Vysoké učení technické v Brně

Brno 2022

Tento dokument byl vytvořen s finanční podporou MK ČR v rámci programu NAKI II v projektu DG18P02OVV055 (Pokročilá extrakce a rozpoznávání obsahu tištěných a rukou psaných digitalizátů pro zvýšení jejich přístupnosti a využitelnosti).

Číslo a název projektu:

DG18P02OVV055	Pokročilá extrakce a rozpoznávání obsahu tištěných a rukou psaných digitalizátů pro zvýšení jejich přístupnosti a využitelnosti

Název a popis dílčího výstupu:

Manuál k Software pro extrakci informace z polostrukturovaných dokumentů
Tento dokument popisuje funkčnost a použití software pro extrakci informace z polostrukturovaných dokumentů.

Jazyk dokumentu

Angličtina

Organizace a řešitel

Vysoké učení technické v Brně	Doc. RNDr. PAVEL SMRŽ Ph.D.

Availability

The software module is available from https://github.com/DCGM/pero-indexer.
Python module https://pypi.org/project/pero-indexer, install as “pip install pero-indexer”.

License

BSD 3-Clause License

Usage

This package provides a full pipeline for extraction of information from custom semi-structured images of text. See Fig. 1 for an actual example of extracting information from a library index card.

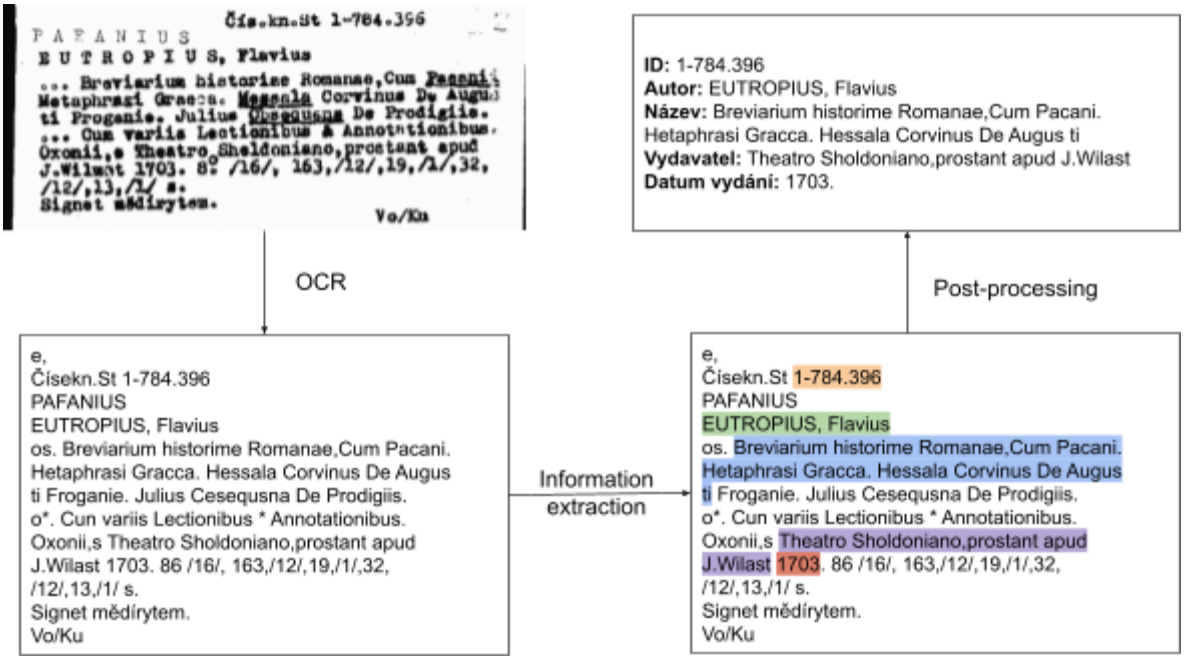


Fig 1. Application of pero-indexer on a new input document.

This package also provides the ability for the user to train their own model on data of their desire. This includes several additional steps as the process includes automatic preparation of training data for the extraction model, as illustrated in Fig. 2.

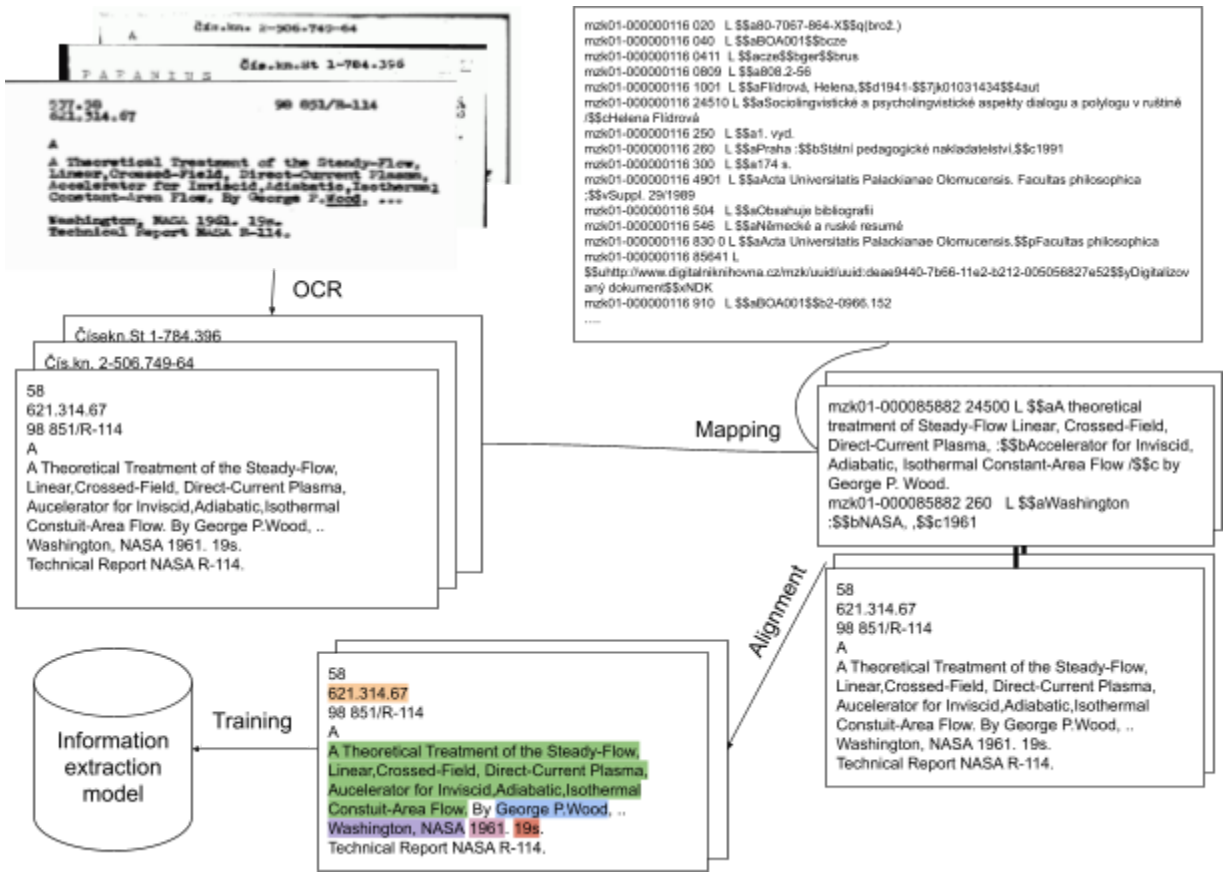


Fig. 2: Training a new model on custom data

The package is designed for usage as a command line application, but it is open for extension to support additional input and output formats.

Requirements

The core functionality is platform-independent, but full pipelines expect Linux-like environment.
Python 3.6+, pero-ocr, transformers, whoosh, and pandas.
For needs of training and faster processing: CUDA capable GPU with at least 4 GB RAM and CUDA toolkit.

Publicly available model

We publish a model trained to extract information from Czech library index cards, possibly containing non-Czech publications (see examples at the end of this document). The model is available online. Its accuracy has been measured as 86 % on unseen data.

Software organization

This software consists of a collection of individual executables, each dedicated to a single function. Together, they are orchestrated into two basic tasks: (1) using a (pre)trained model to extract information from unseen documents (i.e. inference) and (2) training a new model for processing of custom documents.

While the pipelines, themselves being shell scripts, need to be downloaded directly (cloned as part of the Git repository), they assume that pero-indexer has been installed and the individual executables are available in `$PATH`. If pero-indexer is not pip-installed, explicit paths need to be provided as per `pyproject.toml`.

Inference

The inference pipeline is invoked as:
`./pipelines/pipeline-inference.sh images align-model ocr-config work [stage]`

Here, the individual parameters stand for:

images	A folder with images containing scans of documents to be processed
align-model	A folder with a trained model for alignment of transcriptions.
ocr-config	Configuration of a PERO-OCR model which processes input pages into transcriptions (i.e. all of layout analysis, line cropping and OCR proper need to be addressed)
work	A folder for intermediate and final results of the process.
stage	Allows to start the pipeline from the middle, e.g. if a later stage fails for insufficient file permissions, the initial steps can be skipped after fixing the error. Optional.

The output of inference can be found in `$work/readable_output`, it is organized as individual files corresponding to respective files in `$images`.

There are three stages of the inference pipeline:

1. OCR, where transcriptions of the input images are obtained
2. Information extraction, where the transcriptions are labeled for information content
3. Output reformatting, where the information is presented in a user-friendly way

Training

The training pipeline is invoked as:

```
./pipelines/pipeline-training.sh images ocr-config bib-db work [stage]
```

Here, the individual parameters stand for:

images	A folder with images containing scans of documents to be processed
ocr-config	Configuration of a PERO-OCR model which processes input pages into transcriptions (i.e. all of layout analysis, line cropping and OCR proper need to be addressed)
bib-db	A path to a database of bibliographic data in the MARC 21 format, as described e.g. here. If a different information scheme is desired, the corresponding preprocessing stage has to be extended.
work	A folder for intermediate and final results of the process.
stage	Allows to start the pipeline from the middle, e.g. if a later stage fails for insufficient file permissions, the initial steps can be skipped after fixing the error. Optional.

The output of the training is in `$work/ner_model`, it is a collection of files constituting an information extraction model, ready for usage in the inference pipeline.
Training consists of nine stages:

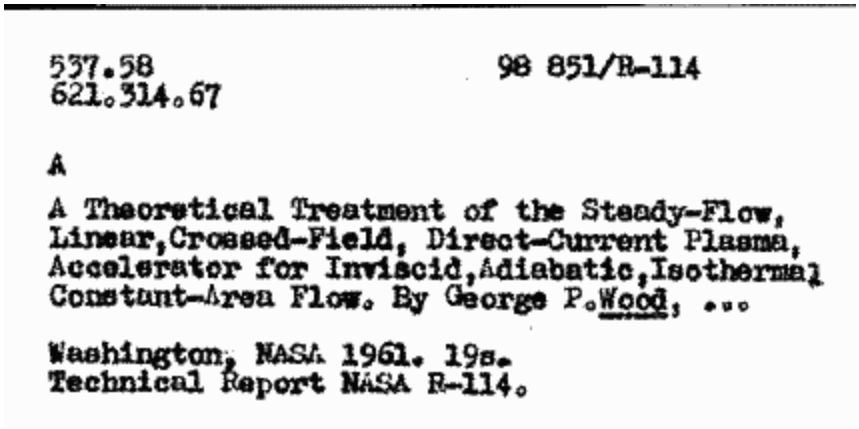
1. OCR, where transcriptions of images are obtained.
2. Preprocessing cards, where the MARC 21 database is condensed into a simple dictionary-of-dictionaries format, encoded as a Python pickle. When building a model for different data of different nature, this step needs to be adjusted accordingly.
3. Building index, where the pickle is stored in a search-efficient structure.
4. Matching images, where images (their transcriptions) are linked with the records in the index.
5. Preformatting the matches.
6. Aligning information, where information from the record is mapped to individual spans of text in the transcription of the images
7. Post-processing the alignments, preparing train/validation/test splits.
8. Downloading a pretrained model. **Requires internet access.**
9. Training the model. **Requires a GPU**.

Overall, training can take a significant amount of time, especially steps 1 and 4 which scale linearly with the number of input images with a significant multiplicative factor.

Examples

Combining the public OCR model provided with PERO OCR and the public model for information extraction from PERO Indexer, one can expect results as follows below. Please note

that the provided model is designed for Czech, thus the Czech names of categories (e.g. Název for Title etc.) and additionally, if there is a transcription mistake, it cannot possibly be recovered during information extraction (e.g. “Aucelerator” in place of Accelerator) – but information can be extracted nevertheless.



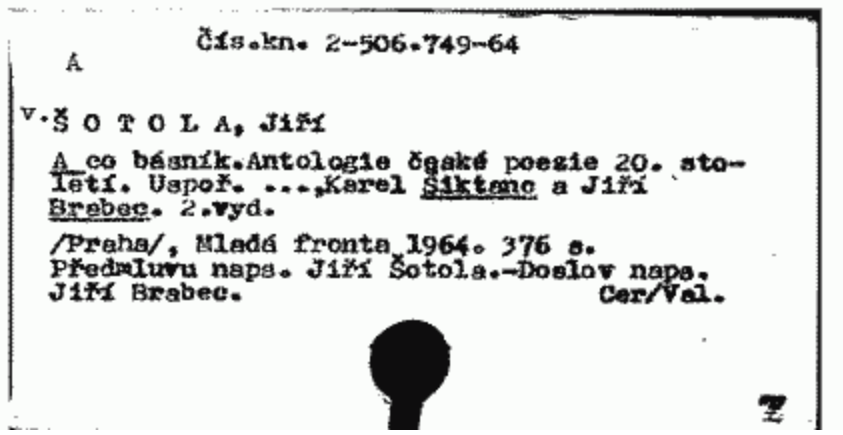
537.58 98 851/R-114
621.314.67

A

A Theoretical Treatment of the Steady-Flow, Linear,Crossed-Field, Direct-Current Plasma, Accelerator for Inviscid,Adiabatic,Isothermal Constant-Area Flow. By George P.Wood, ...

Washington, NASA 1961. 19s.
Technical Report NASA R-114.

ID: 98 851/R-114
Název: A Theoretical Treatment of the Steady-Flow, Linear,Crossed-Field, Direct-Current Plasma, Aucelerator for Inviscid,Adiabatic,Isothermal Constuit-Area Flow.
Autor: By George P.Wood,
Vydavatel: Washington, NASA
Datum vydání: 1961.
Počet stran: 19s.



Čís.kn. 2-506.749-64

A

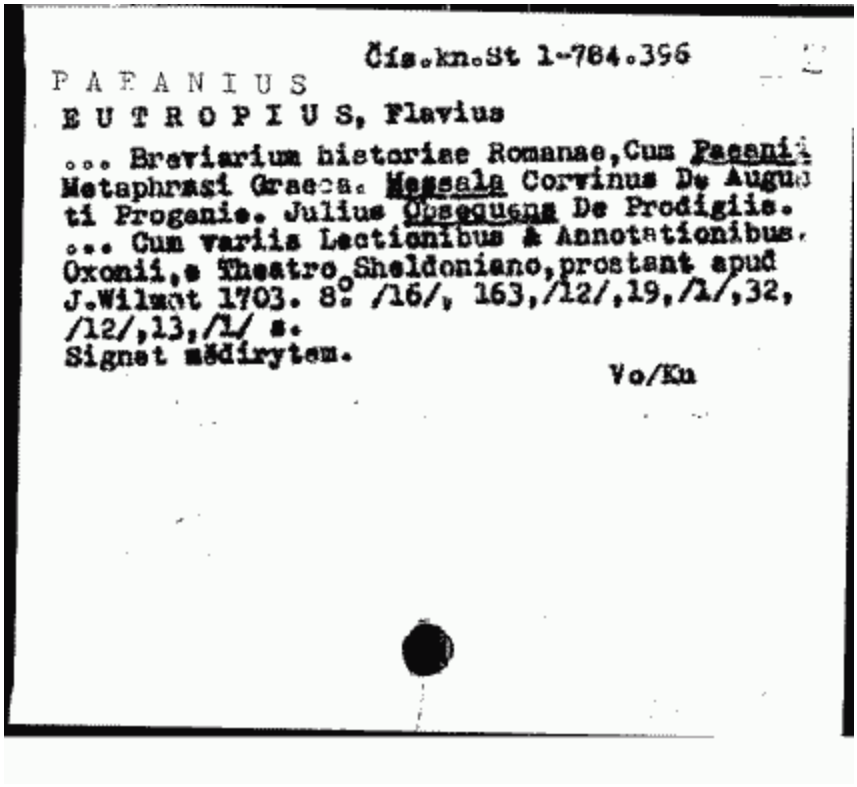
V.ŠOTOLA, Jiří

A co básník.Antologie čgaké poezie 20. století. Uspoř. ...,Karel Šiktanc a Jiří Brabec. 2.vyd.

/Praha/, Mladá fronta 1964. 376 s.
Předmluvu napsa. Jiří Šotola.-Doslov napsa. Jiří Brabec. Cer/Val.

ID: 2-506.749-64
Autor: ŠOTOLA, Jiří
Název: A co básník.Antologie
Edice: 2.vyd.

Vydavatel: Mladá fronta
Datum vydání: 1964.
Počet stran: 376 s.


Čís.kn.St 1-784.396

PACANIUS

EUTROPIUS, Flavius

... Breviarium historiae Romanae,Cum Pacanii Metaphrasi Graeca. Messala Corvinus De Augusti Progenie. Julius Obsequens De Prodigiis. ... Cum variis Lectionibus & Annotationibus. Oxonii,e Theatro Sheldoniano,prostant apud J.Wilmot 1703. 8° /16/, 163,/12/,19,/1/,32, /12/,13,/1/ s.
Signet mědirytem.

Vo/Ku

ID: 1-784.396
Autor: EUTROPIUS, Flavius
Název: Breviarium historime Romanae,Cum Pacani. Hetaphrasi Gracca. Hessala Corvinus De Augus ti
Vydavatel: Theatro Sholdoniano,prostant apud J.Wilast
Datum vydání: 1703.