

Do we want AI in Archives?

Speech to text, a transcription project

Martin Rechterik - National Archives Of the Czech Republic
- University of West Bohemia in Pilsen



Moderna Arhivistika
Technical and field related problems
of tradition and electronic archiving



FAV

DEPARTMENT OF
GEOMATICS

About me

- Master's Degree in History (Faculty of Arts)
- PhD student at Geomatics' Department
- Archivist of the National Archives
- head of the team which created [The Transformation of Digital Spatial Data for Long Term Preservation in a Digital Archives](#)
- Member of [DILCIS RDB AIG](#), [DLM Forum](#) - subgroup Geoforum
- Member of International Database Preservation Group managed by [NARA](#)
- Head of the National Spatial Information Infrastructure Working Team with a focus on archiving (PT NIPI_ARCHI)
- Active at [martin.rechtorik\(at\)na.gov.cz](mailto:martin.rechtorik@na.gov.cz), [Linkedin](#), [ResearchGate](#)

Artificial Intelligence in Czech Archives

- AI use in archives and libraries is on the rise
- Libraries have an edge over archives in Czechia
- Which archives use AI in Czechia?
 - Czech Broadcasting Company Archives (Speech2Text)
 - Czech TV Archives (Speech2Text)
 - Regional Archives of Silesia (OCR/HTR)
 - Military Historical Archives (OCR/HTR)
 - National Archives (Speech2Text)

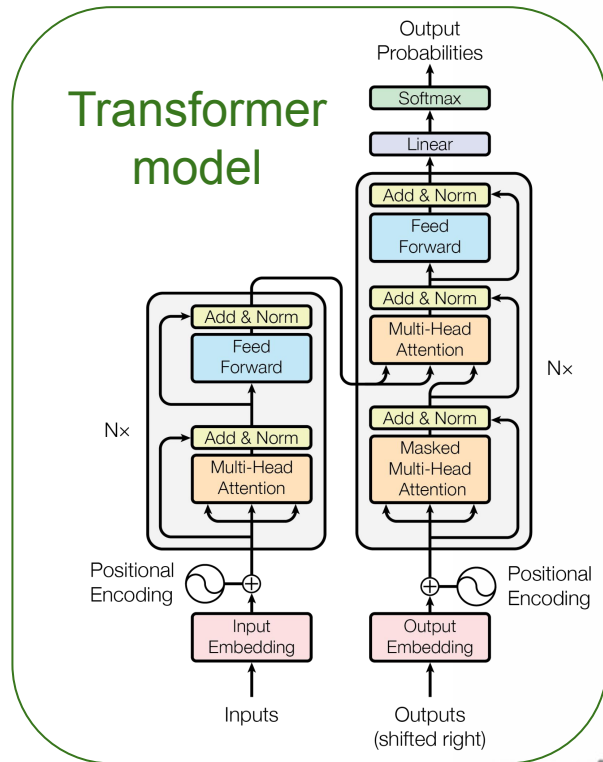
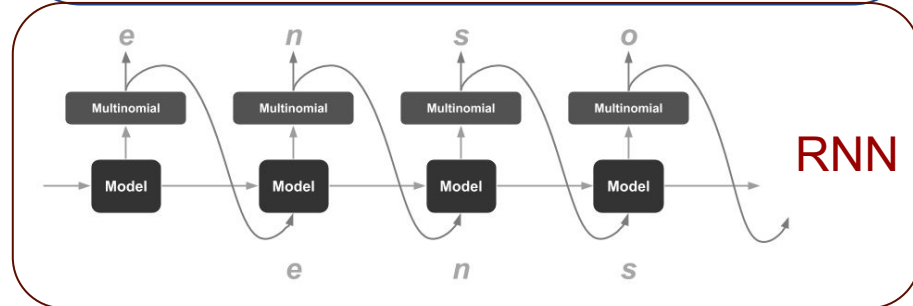
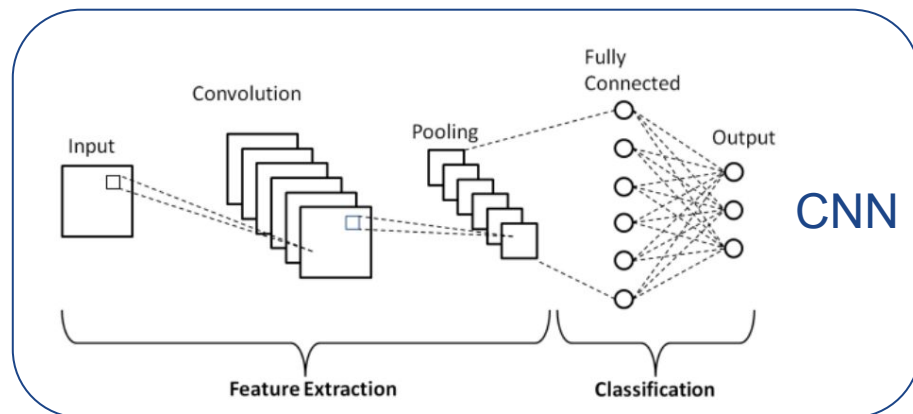
Speech to text, Transformer model

- Transformer model relies entirely on self-attention mechanisms, discarding recurrent and convolutional layers.
- It uses self-attention to compute representations of input and output sequences, allowing it to weigh the importance of different words in a sentence.
- The architecture employs multi-head attention, which allows the model to jointly attend to information from different representation subspaces at different positions.
- Since the model does not have a built-in notion of sequence order, it incorporates positional encodings to give the model information about the position of words in a sequence.
- The Transformer consists of an encoder that processes the input and a decoder that generates the output, both composed of multiple layers of self-attention and feed-forward neural networks.

Speech to text, transformer model

- The model is trained using a variant of stochastic gradient descent and employs techniques like label smoothing and dropout to improve generalization.
- The Transformer model achieves state-of-the-art results on various machine translation benchmarks, outperforming previous models that relied on recurrent architectures.
- The architecture is highly parallelizable, making it more efficient to train on large datasets compared to traditional RNNs.
- The introduction of the Transformer has significantly influenced the field of natural language processing, leading to the development of various models like BERT and GPT.

CNN, RNN and Transformer Models



The BBC Czechoslovak editorial office 2001-2007

- Basic stats of the [Archival collection](#)
 - acquired 2013 by 5th Department, Miroslav Kunt
 - curated and preserved by National Archives
 - audio records and its structured data (MS Excel + MS Access)
 - Size 850,313 GBs, 14 068 audio files
 - estimated records length is around than 305 days in total
 - BBC archive (cca 927 recordings, more then 7 days)
 - BBC red archive (13 141 recordings, 297 days)
 - file format of recordings: MP2

BBC broadcasting in the Czech and Slovak language

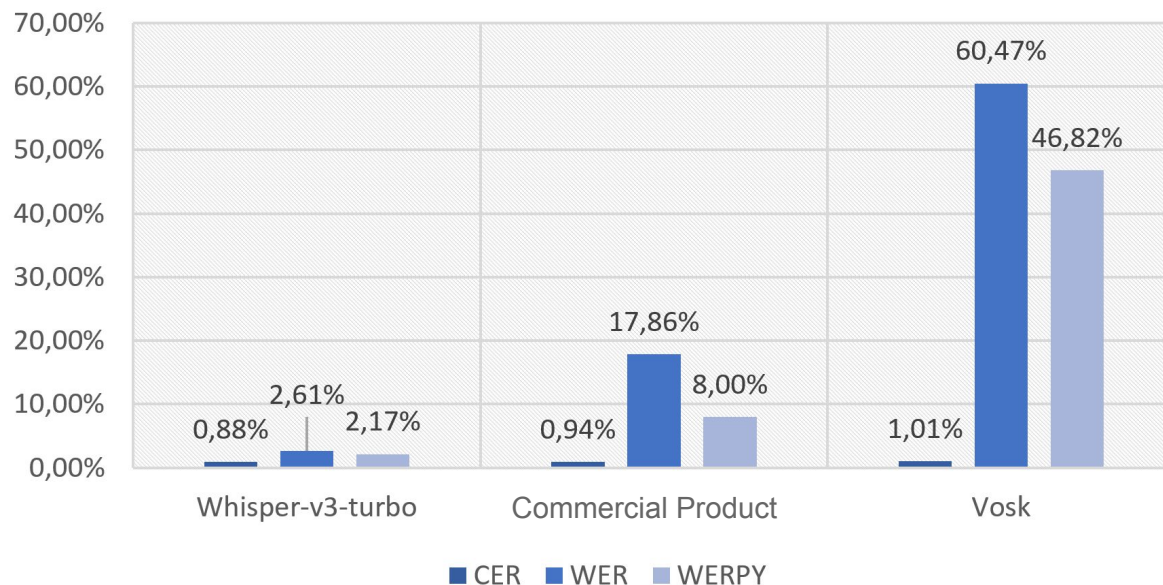
- It started on the 6th September 1939, a few days after the invasion of Nazi Germany into Poland.
- The first Czechoslovak broadcasts on the station were initiated by the Labour Party, which was the first to support the restoration of Czechoslovakia after the Munich Agreement and the subsequent occupation by Germany in March 1939
- The first speech in Czech in support of his countrymen was delivered by the Social Democrat Josef Kosina.
- Under the banner of the Overseas Service, the war years saw a great expansion of international broadcasting, including Czechoslovak broadcasting.
- Except the Czechoslovak broadcasts, there was another broadcast in Czech and Slovak, which was the Voice of the Free Republic. This was the broadcast of the Czechoslovak government in exile, represented by President Edvard Beneš or Foreign Minister Jan Masaryk.

BBC broadcasting in the Czech and Slovak language

- The Czechoslovak editorial office continued its activities after the end of the Second World War
- It represented reliable information from the democratic world during the period of communist totalitarianism until its end in 1989.
- Czechoslovak broadcasts were not only political and news, but also included cultural and spiritual programmes.
- Among the important personalities who participated in BBC programmes in the Czechoslovak language were Karel Kyncl, Br. Jan Lang S. I., Bedřich Bělohlávek, Karel Brušák, Zdena Tominová, Petr Brod, Vít Kolář and many others.
- After the Velvet Revolution, the Czech BBC newsroom became a hatchery of well-known presenters or journalists working in public or commercial medias, such as Václav Moravec, Martin Řezníček, Jan Bumba, Daniel Kaiser, Ondřej Štindl.

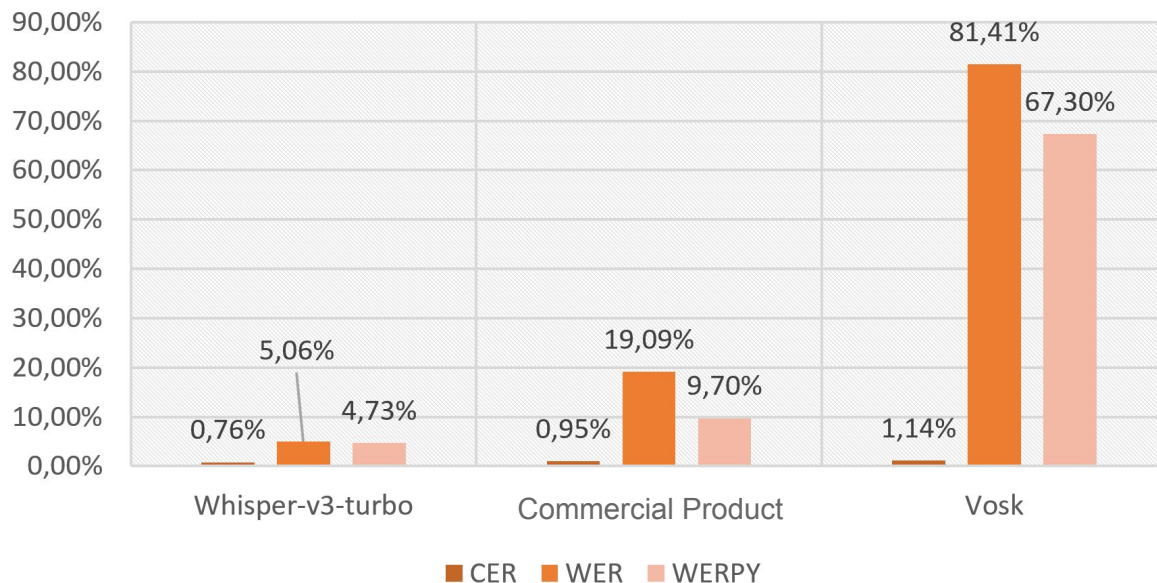
Results: Record 1

Error Rate Evaluation Record 1



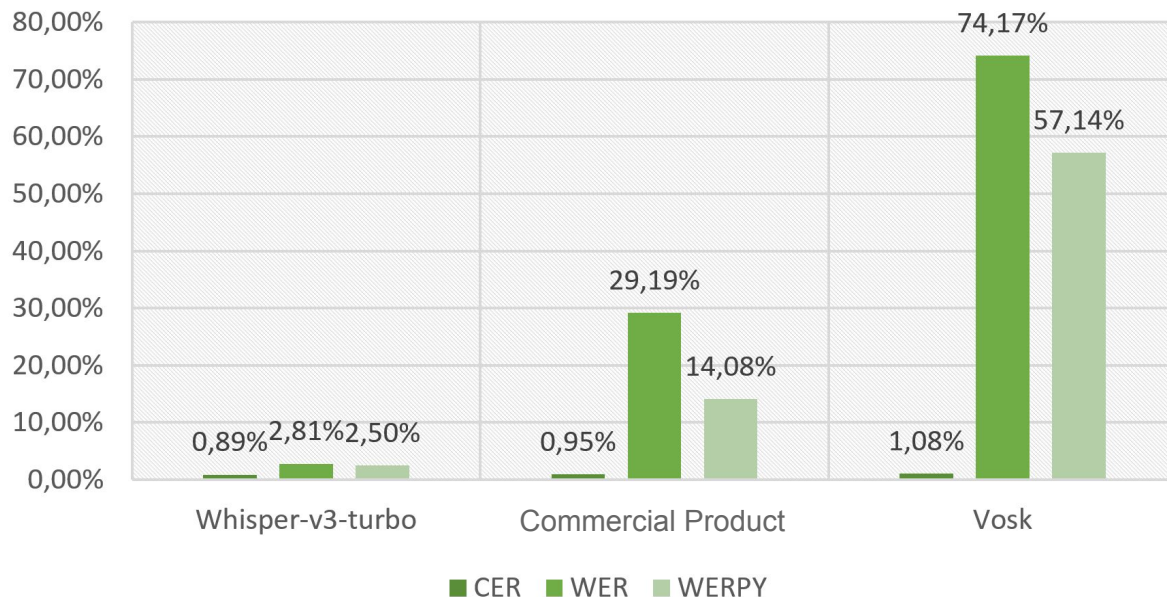
Results: Record 2

Error Rate Evaluation Record 2

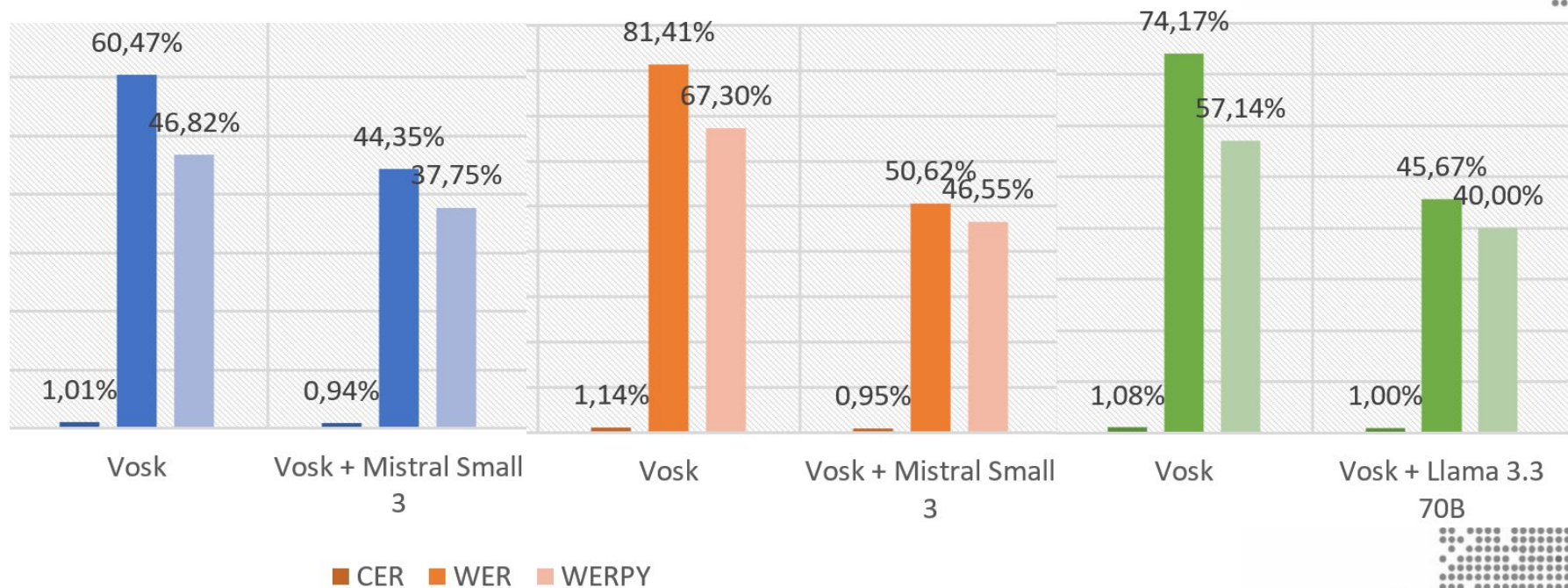


Results: Record 3

Error Rate Evaluation Record 3



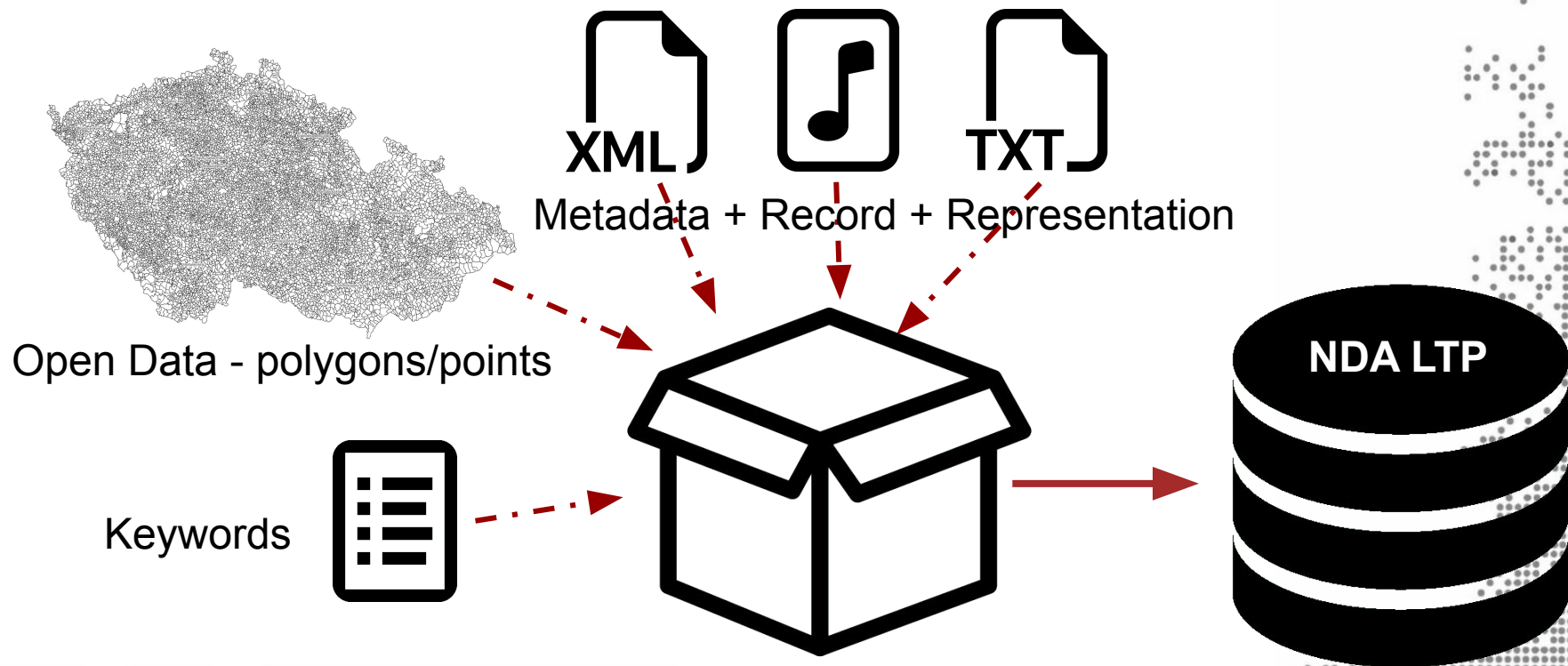
Can LLM help Vosk's outputs



SIP creation Workflow

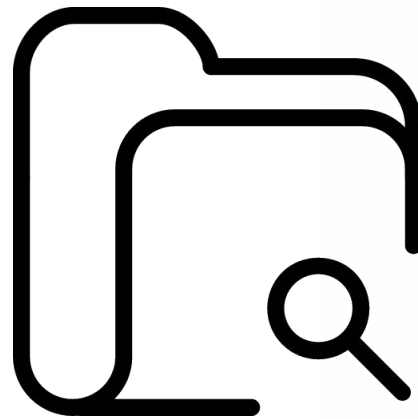
- data processing
 - walk directory tree, extract text layer from records
 - intersect extracted text with keywords list a geospatial open data
 - create requested SIP directory structure
 - create metadata (Mets, EAD cz-Profile)
 - create SIP using EARK tools
 - ingestion to the National Archival Portal
 - Data transfer to the NDA(LTP)
 - Metadata changes in AIPs will be done by Arranging Software

The role of Open Data



Access

- Traditional way
 - Finding aid - requires LLM and RAG to create abstracts
 - Create a new multi model database to search in using ElasticSearch
- Modern way
 - Vector database and search by similarity
 - Advanced ranking of results



Summary

- automation is possible
- whole process is not demanding, pipeline can be simple
- participation of students from Czech Technical University is beneficial
- videos can be processed too
- professional tool is better to split text by a speaker
- keywords and localities can be extracted without any issues
- big potential to use it in other areas like reports from meetings, interviews with data producers etc.

Summary

- insufficient metadata from data producer is limiting
- missing implementation of LLM incl. Retrieval Augmented Generation (RAG)
- powerful workstation with GPU is recommended
- Faster-Whisper does not support MP2 file format
- Speech2Text using Vosk model supports only one channel WAV file format and is much better for english speech

References

- ALPAYDIN, Ethem. *Machine learning*. Revised and updated edition. The MIT Press essential knowledge series. Cambridge, Massachusetts: MIT Press, [2021]. ISBN 978-0-262-54252-4.
- DRORI, Iddo. *The science of deep learning*. Cambridge: Cambridge University Press, 2023. ISBN 978-1-108-83508-4.
- FIŠER, Marek, KYKAL, T.: Digitální studovna Ministerstva obrany ČR, Využití technologií na pokročilou indexaci obsahu historických dokumentů. In Soudobé dějiny 2/2004, Available from: <https://doi.org/10.51134/sod.2024.027> [accessed 2025-04-29].
- CHOLLET, François. *Deep learning v jazyku Python: knihovny Keras, Tensorflow*. Knihovna programátora. Praha: Grada Publishing, 2019. ISBN 978-80-247-3100-1.
- KELLEHER, John D. *Deep learning*. The MIT Press essential knowledge series. Cambridge, Massachusetts: The MIT Press, [2019]. ISBN 978-0-262-53755-1.
- KOCOUREK, Milan. *Volá Londýn: historie českého a slovenského vysílání BBC : (1939-2006)*. Praha: Ottovo nakladatelství, c2013. ISBN 978-80-7451-318-3.
- RASMUSSEN, Casper Hvenegaard; RYDBECK, Kerstin a LARSEN, Håkon (ed.). *Libraries, archives, and museums in transition: changes, challenges, and convergence in a Scandinavian perspective*. Online. London: New York, 2023. ISBN 978-1-032-03364-8. Available from: <https://www.taylorfrancis.com/books/oa-edit/10.4324/9781003188834/libraries-archives-museums-transition-casper-hvenegaard-rasmussen-kerstin-rydbeck-h%C3%A5kon-larsen>. [accessed 2025-04-29]
- SHAH, Chirag. *A hands-on introduction to machine learning*. Cambridge: Cambridge University Press, 2023. ISBN 978-1-009-12330-3
- Text generation with an RNN, accessed from https://www.tensorflow.org/text/tutorials/text_generation
- Van Hiep Phung, Eun Joo Rhee A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets, available from <http://dx.doi.org/10.3390/app9214500> [accessed 2025-04-29]
- VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N.; KAISER Lukasz; POLOSUKHIN, Illia. *Attention is all you need*. Cornell University 2023. Available from: <https://doi.org/10.48550/arXiv.1706.03762>. [accessed 2025-02-20].

Do we want AI in Archives?

Speech to text, a transcription project

martin.rechtorik@na.gov.cz