# Portuguese Archives Handwritten text recognition of passport requisitions

**Dora Melo[2,3], Irene Pimenta Rodrigues[1,3] and Lígia Ferreira[1,3]**

[1] *University of Évora, Department of Informatics, Portugal*
[2] *Polytechnic University of Coimbra, Coimbra Business School—ISCAC, Portugal*
[3] *NOVA Laboratory for Computer Science and Informatics, NOVA LINCS, Portugal*

*Corresponding/Presenting author: lsf@uevora.pt*

## Talk Abstract

The DigitArq platform is the Portuguese National archive system that uses well-established description standards, namely the ISAD(G) (General International Standard Archival Description) and ISAAR(CPF) (International Standard Archival Authority Record for Corporate Bodies, Persons and Families) with a hierarchical structure adapted to the nature of archival assets. In the EPISA project, one of the tasks included the migration of the DigitArq information into a linked open data model, CIDOC-CRM [2]. This task included the representation of textual description in the ISAD(G) element 'Scope and Content' by extracting the information from natural language written text. The dataset for handwritten recognition has 1000 registers with: digital representation, a text description of the digital content, and the semantic representation in CIDOC-CRM of the text description [1]. This information enables the automatic evaluation of handwritten recognition and can be used to improve the performance of handwritten recognition through the use of semantic information. The handwritten data was selected from a set of registers with digital representation, a `jpg` file, from the Portuguese National Archive. The registers were chosen from those that have a text transcription of digital representation in the DigitArq platform. Handwritten text recognition is an important task in computer vision that has received considerable attention in recent years [3, 4]. In our approach, the open-source document processing platform ArkIndex [5, 6] (`https://teklia.com/our-solutions/arkindex/`) is used to automatize the document recognition system adapted to the passport registers with digital representation. Initially, a corpus of 100 registers was built up and a manual annotation was performed to represent the structure of the pages (text zones, pages and text zones transcriptions), producing an automatic transcription of the handwritten

text. The described approach evaluation reveals promising results that confirm that the initial annotated corpus can be used to obtain a general tool for processing the passport registers in DIGITARQ.

**Keywords:** handwritten recognition, document annotation, artificial intelligence, data analysis.

**Acknowledgements**

This work is supported by NOVA LINCS ref. UIDB/04516/2020 (`https://doi.org/10.54499/UIDB/04516/2020`) and ref. UIDP/04516/2020 (`https://doi.org/10.54499/UIDP/04516/2020`) with the financial support of FCT.IP

# References

[1] Varagnolo, D., Melo, D., Rodrigues, I.P., Rodrigues, R., Couto, P., Archives Metadata Text Information Extraction into CIDOC-CRM, *Knowledge Discovery, Knowledge Engineering and Knowledge Management. IC3K 2022. Communications in Computer and Information Science*, 1842. Springer, Cham, 2023.

[2] Melo, D., Rodrigues, I.P., and Varagnolo, D., A Strategy for Archives Metadata Representation on CIDOC-CRM and Knowledge Discovery, *Semantic Web*, 14(3), 553–584, 2023.

[3] Breuel, T., Ul-Hasan, A., Al-Azawi, M. A., and Shafait, F., High-performance OCR for printed English and Fraktur using LSTM networks. *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 1, pp. 67-72, 2017.

[4] Bluche, T., Messina, R., and Kermorvant, Scan, attend and read: End-to-end handwritten paragraph recognition with MDLSTM attention. *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 1, pp. 127-132, 2017.

[5] Boillet, M., Kermorvant, C., and Paquet, T., Multiple Document Datasets Pre-training Improves Text Line Detection With Deep Neural Networks, *International Conference on Pattern Recognition*, 2021.

[6] Hazem, A., Bonhomme, M., Maarand, M., Kermorvant, C., and Stutzmann, D., Books of Hours: the First Liturgical Corpus for Text Segmentation, *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020.