From Archives Management to Knowledge Extraction: Tools and Impact Evaluation

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI **CULTURE, EDUCAZIONE E SOCIETÀ**
Laboratorio di **Doc**umentazione

Anna Rovella[1], Assunta Caruso[1], Martin Critelli[1], Armando Bartucci[2], Francesca M.C. Messiniti[1].
[1]DICES, University of Calabria, Italy, {anna.rovella, assunta.caruso, martin.critelli, francesca.messiniti}@unical.it.
[2]University of Macerata, Italy, {a.bartucci}@unimc.it.

Objective

The aim of this paper is to illustrate the application of some automatic knowledge extraction tools on archive records. These tools are based on Machine and Deep Learning technologies that facilitate the implementation process and allow the creation of scalable solutions. The purpose of the work is to automatically extract metadata, keywords, terms, phrases, entities, tables and graphs from documents which are characterized by complex informational dimensions. The ultimate goal is to measure the quality of the extraction obtained. In particular, the work aims to evaluate the results obtained from the application of the selected tools on a specific corpus of archival documents. The evaluation process will pay particular attention to the accuracy of the results on the semantic level.

Introduction

The Knowledge Society and digital transition have led to a profound transformation of archives even in their approach to the issues of access to documents and the information contained in them. The development, as well as, the application of knowledge extraction tools in archives is becoming one of the possible responses to the growing demand for information that archives face in the public and private sectors.
The diffusion of machine and deep learning techniques in many applications, also in the archival field, may represent an interesting approach for the recovery of knowledge both during the document management phase and in that of digital preservation and retrieval for cultural purposes. The knowledge extraction from this complex and varied type of documentation represents a significant test bench for collaboration between archivists, documentalists and computer scientists.

The case study and methodology

The selection of the corpus.
The first step of the work involved the selection of the corpus. Then a system training phase and metadata extraction process followed.
We selected a corpus of 78 research projects. For ease of recovery, we have extracted the projects from the EU Cordis database. They are all funded by Horizon 2020 and are related to the Mercury pollution domain. The choice of such a wide domain allows heterogeneity of contents connected not only with the pollution from Mercury but also with the impact on the health of living beings. The sample is also representative of documents belonging to public and private archives in several European countries.
From the archival point of view, each project corresponds to a hypothetical file in a creator's archive (University, Research body, Company, etc.). The project file contains all the documentation produced during the research management (deliverables, administrative documents, reports, patents, OpenAir datasets, scientific publications, etc.). Thus, we obtained a total number of 604 documents in our corpus consisting of 78 projects. For the purpose of this work we have carefully analyzed deliverables. They are particularly representative because they are both scientific and administrative documents. In this work we have considered only native digital documents in PDF format that represent 98% of deliverables related to the selected projects.

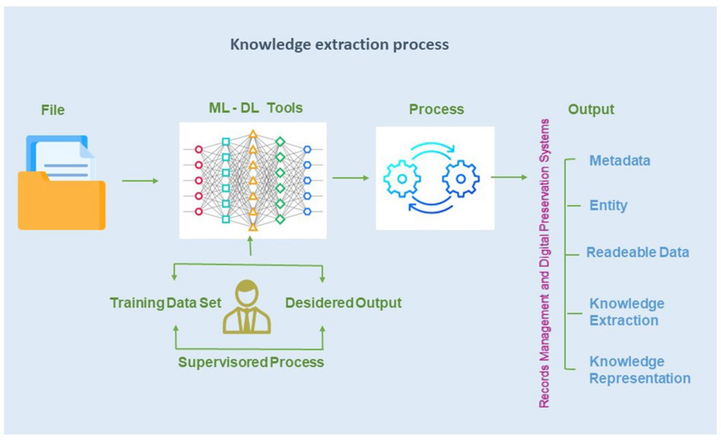

Figure 1 Knowledge extraction process

Selection of tools and definition of training sets.
For knowledge extraction activities we used the Cermine framework to translate documents in XML, to extract metadata and identify any objects of the type chart or table. The use of this software required the definition of models and training sets for recognition. Subsequently, to enrich and define the extracted knowledge set, we used NLP technologies, thus improving the quality of the information and detecting document entities and relationships. To do this we used NLTK, BERT, BERTopic, Gensim and spaCy. We consider documents in which text is one of the components, however, they may also include tables or charts (reports, decrees, projects).
From the analysis of the documents attached to the projects we have found that the representation of data and information is often in the form of tables, present in all documents or even charts that appear in 73% of documents. Very often these are images inserted in the text and the extraction of knowledge from such objects can be carried out using dedicated tools. Chart and tables may contain research data, accounting data, administrative information, context statistics, etc. The extraction of knowledge from such objects can be useful for several purposes including decision support, rationalization of resources for research, processing of statistics, retrieval of information, etc. The ability to extract usable knowledge and machine-readable data also impacts system interoperability. In particular, we used several Deep learning algorithms, such as EfficientNet CNN model, and Pytesseract and DocTR for chart and text recognition. These tools allow to obtain complementary data and are useful for the final extraction of the chart.
All the tools we analyzed and tested are open source and could be integrated both in records management and digital preservation systems to support access to content or to extract entities or data. The use of such tools could have a positive impact for: decision support, improvement of administrative efficiency, metadata process, optimization of document search performance.

Results and evaluation

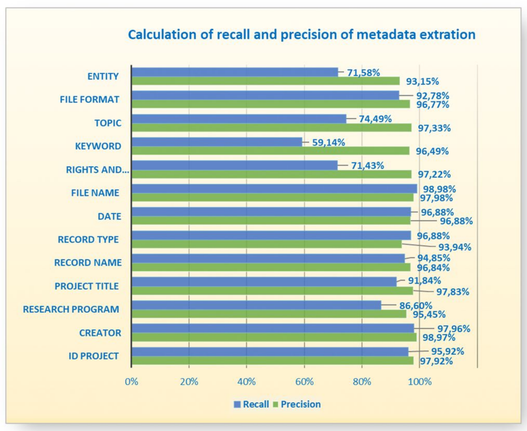

Figure 2 Calculation of recall and precision of metadata extraction

The results obtained with the metadata extraction process were measured quantitatively and evaluated qualitatively through the calculation of recall and precision. The extracted metadata are: ID Project, Creator, Research Program, Project Title, Record Name, Record Type, Date, File Name, Rights and Responsibilities, Keyword, Topic, File Format, Entity (Persons, Organizations, Places and Events). Quantitatively the average extraction of the metadata is equal to 85%. The lowest reading value is 58% (keyword) while other typically archival metadata have been recognized in over 90% of projects in the corpus.

The measurement of the accuracy of the recognized metadata was 98% (Figure 2). The quantitative values relating to entities and rights and responsibilities, are not optimized, but qualitatively significant. It will be necessary to work on these aspects. Overall, the automatic extraction of metadata obtained is positive especially due to the supervision process used that has allowed a significant improvement in the performance of the analyzed tools.

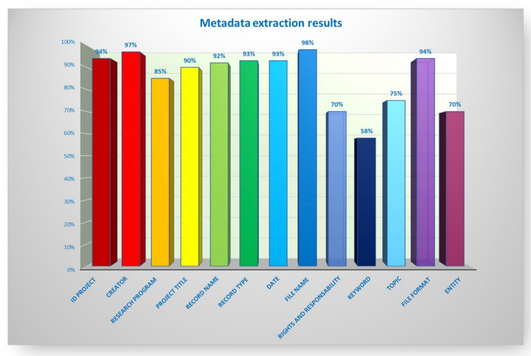

Figure 3 Metadata extraction results

In most cases the tables have been recognized and data have been read. However, the quality, accuracy of table recognition and data reading, obtained with Tabula, was not always optimized due to some errors mainly because of the variety of alphabets, symbols and structures used in documents. The diagram in Figure 4 represents the type of errors and the percentage of frequency.
The example in Figure 5 shows a result of the automatic extraction of data and information from the chart.
This process has been complex and the results obtained, although interesting, cannot yet be considered entirely satisfactory. We have found difficulties in the extraction of quantitative and qualitative data, especially when the charts, as in the case of Figure 5, are full of lines and data. The algorithms used have not been able to recognize with precision the different positions of the points on the chart, moreover the consistent association with the reference label was not always correct. This is despite the support provided by a supervised process.

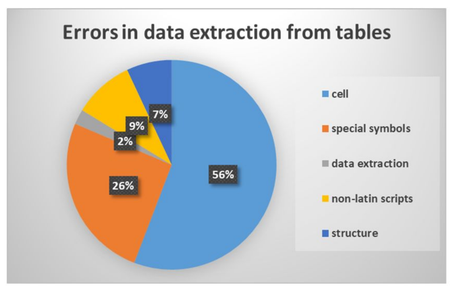

Figure 4 Errors in data extraction from tables

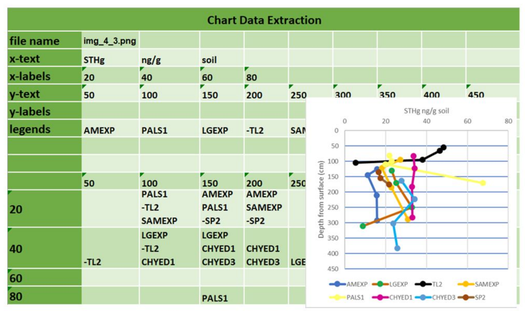

Figure 5 Chart data extraction

Lessons learned

Several works focus on the application of Machine Learning and Deep Learning technologies to archives. Similarly, in other contexts, numerous efforts aim to use the automatic data and metadata extraction technologies from texts. In this poster we report our experience and a case study of knowledge extraction from archival documents. The results are encouraging and there is value in the development, integration and optimization of tools. New archives are part of a digital ecosystem that emphasizes data, demands ease and speed of access to information and requires a constant ability to represent knowledge in all its relationships through interoperable systems. We believe that knowledge extraction can be a likely route in such a complex and articulate journey.

References

Aristarán, Manuel, "Tabula." Tabulapdf/tabula. https://github.com/tabulapdf/tabula.

Colavizza, Giovanni, Tobias Blanke, Charles Jeurgens, and Julia Noordegraaf, "Archives and AI: An Overview of Current Debates and Future Perspectives." *Journal on Computing and Cultural Heritage* 15, n.1, 2022. https://doi.org/10.1145/3479010.

Devlin, Jacob, Ming-Wei Chng, Kenton Lee, and Kristina Tautanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." 2018. http://arxiv.org/abs/1810.04805.

Grootendorst, Maarten, "BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics (Version v0.7.0)." 2020. https://doi:10.5281/zenodo.4381785.

Rane, Chinmayee, Seshasayee Mahadevan Subramanya, Devi Sandeep Endluri, Jian Wu, and C. Lee Giles, "ChartReader: Automatic Parsing of Bar-Plots." *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, 2021. doi:10.1109/IRI51335.2021.00050.

Rehurek, Radim, and Petr Sojka, "Gensim--python framework for vector space modelling." NLP Centre, Faculty of Informatics, Masaryk University, Brno CZ, n.2, 2011.

Rovella, Anna, Alexander Murzaku, Eugenio Cesario, Martin Critelli, Armando Bartucci, and Francesca Maria Caterina Messiniti, "Analysis, evaluation and comparison of knowledge extraction tools in the environmental and Health domain. A holistic approach." *Knowledge Organization and Management in the Domain of Environment and Earth Observation (KOMEEO)* 18, 2022. https://dx.doi.org/10.5771/9783956508752-121.

Tkaczyk, Dominika, Pawel Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Lukasz Bolikowski, "CERMINE: automatic extraction of structured metadata from scientific literature." *International Journal on Document Analysis and Recognition* 18, n.4, 2015. https://doi.org/10.1007/s10032-015-0249-8.