AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-329-1
DOI 10.57574/596532918
pag. 145-156 (luglio-dicembre 2023)

Knowledge extraction, research projects and archives management

Anna Rovella[*], Assunta Caruso[*], Martin Critelli[**], Francesca M.C. Messiniti[*]

Abstract: Archives play an important role in the knowledge society and must respond ever more quickly to information needs. For example, in the case of universities, research projects are a strategic asset for the growth of territories, the rationalization of financial resources and the development of archival science. Clearly, the documentation that characterizes the research projects has an administrative value as well. This paper, investigates the possibility of extracting knowledge from this class of documents. In particular, the purpose of this paper is to experiment with the application of some automatic metadata extraction tools on archival documents. An approach of metadata automatic extraction could provide a greater continuity between production and representation of objects. Metadata can be useful in accessing or sharing contents within digital preservation systems (i.e. ontologies, Linked Data). The chosen tools use Machine Learning technologies and supervised learning techniques together with newer Deep Learning technologies.

Keywords: Knowledge Extraction, Research Project, Metadata, Records Management, Digital Preservation.

1. Scenario

The Knowledge Society and digital transition have led to a profound transformation of archives, especially with regard to access to records and information. The development and application of knowledge extraction tools is a po-

[*] Università della Calabria, Rende (CS), Italia. anna.rovella@unical.it; assunta.caruso@unical.it; francesca.messiniti@unical.it.
[**] Istituto di Linguistica Computazionale – Consiglio Nazionale delle Ricerche, Pisa, Italia. martin.critelli@ilc.cnr.it.
All the authors contributed to the concept and design of the study, read and approved the final version of the article. Anna Rovella wrote and is responsible of: Scenario, Towards an integrated approach, and Lessons learned; Assunta Caruso revised the whole paper and wrote the section The selection of the corpus; Martin Critelli and Francesca M.C. Messiniti carried out the experiments and wrote: Selection of tools and definition of training sets, Results and evaluation, References.

tential response to the increasing demand for information that archives face in both the public and private sectors. In order to identify the correct role played in this ecosystem, archivists must extend their competence to new knowledge fields, which differ from their traditional training. In Italy, a country with a considerable archival heritage, the richness and permanent change of the legislative framework, based on digital document management and preservation, frequently creates theoretical questions and defines new training needs.

For example, the archives of (Italian) universities are increasingly populated by the production of research projects organized in archival units that show the relationship between administrative records and research documents. Research projects represent an important source of funding because Italian researchers are heavily involved in even more competitive EU programmes and calls. The data of the EU Horizon 2020 programme show about 109,383 applications by Italian researchers for a total funding of over 5.7 billion euro and in 2021, there was a 4,1% increase in the success for the new Horizon Europe programme (256 million euro allocated in grants in a single year - Horizon 2020 country profile). Research projects are an inestimable cultural and scientific heritage whose information also impact on the production system and in the evaluation processes of research quality. The archives should provide a response to the problems of quick and effective access to documents related to research projects. Within the Agenda 2022-2024, the European Union has set up a cloud infrastructure, European Open Science Cloud (EOSC), supporting both research and production systems as well as citizens. The purpose of EOSC is to develop Services for Fair Science[1] in European web of data. In this environment the stakeholders can publish, search for, and reuse data, tools and services for research, innovation and education. Research institutions are, therefore, called upon to feed the EOSC infrastructure properly and to have greater capacity for managing and accessing research documents and data[2].

Access to digital documents both in records management and long-term digital preservation systems is a much-discussed topic in the Italian and international archival community. Therefore, the spread of Machine and Deep

---

[1] «In 2016, the 'FAIR Guiding Principles for scientific data management and stewardship' were published in Scientific Data. The authors intended to provide guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets. The principles emphasize machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data» https://www.go-fair.org/fair-principles/ (accessed June 22, 2023).

[2] «The implementation of the EOSC is based on a long-term process of alignment and coordination pursued by the Commission since 2015 with the many and diverse stakeholders of the European research landscape» https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science/european-open-science-cloud-eosc_en (accessed June 22, 2023).

Learning (ML, DL) techniques may represent an interesting approach for the knowledge retrieval for cultural purposes both during the document management phase and in that of digital preservation[3].

Due to their complex and varied type of documentation, knowledge extraction from research projects, represents a significant test bench for collaboration among archivists, documentalists and computer scientists. ML and DL studies and applications could also have a significant impact on the training of the necessary skills needed within archives.

The aim of this paper is to illustrate the application of some automatic knowledge extraction tools on archival records. These tools are based on ML and DL technologies that facilitate the extraction process and allow the creation of scalable solutions. The purpose of the work is to automatically extract metadata, keywords, terms, phrases, entities, tables and charts from documents which are characterized by complex information dimensions. The ultimate goal is to evaluate and measure the quality of the extraction obtained. We will pay particular attention to the accuracy of the results on the semantic level.

The paper is organized as follows: the next section presents related works and the research idea. In the third section we define a case study and applied methodology. Section four presents a discussion of the results of the data and metadata extractions. The final section highlights some of the 'lessons learned'.

2. Towards an integrated approach

The study of automatic metadata extraction has generated different techniques and approaches that have led to the implementation of tools and frameworks. However, before analyzing these applications, we may consider some criticalities. In the reference literature there are two different approaches to automatic metadata generation: on the one hand, research into the techniques for finding information and content in digital resources, and on the other, the software development for the creation of content. These technologies are normally managed separately, but it is not difficult to imagine, that the integrated development of such components could greatly improve automatic metadata generation and extraction allowing the original production of electronic documents rich of embedded metadata.

Another critical step, underlined by the studies, is the representation of metadata in the output phase, where normally the software is not able to rep-

[3] There is more than one definition of ML and DL. From a generic point of view the ML is a subdomain of Artificial Intelligence (AI) with which machine is able to automatically learn by reproducing human intellect capacity. In the same way the DL and the neural networks aim to simulate human brain work to support AI in many applications like, for example, chatbot, language recognition, object detection, etc.

resent metadata belonging to more than one metadata schema/metadata standard. For example, very often the Dublin Core is the standard used and in the archival field this can be seen as a limitation.

Through the use of dedicated frameworks an additional problem arises from the observation of the results. The values related to some extracted metadata (i.e. about the content of the document) may consist of a portion of the text, which could generate semantic ambiguity and may reduce the power of metadata to build structured information. We believe that linguistic resources could enhance and enrich the extraction obtained (Rovella et al. 2022). For this reason, in this paper we have tried to integrate Natural Language Processing (NLP) techniques for the analysis of marked texts derived from the metadata extraction process. The aim of this experiment is also to obtain entities description following the logic of archival standards.

NLP is a field of computer science that deals with human communication with the aim to help machines understand, interpret and generate human languages. In particular, in NLP, Deep Learning techniques allow direct learning of hierarchical language representations to perform generalized tasks (Kamath, Liu, and Whitaker 2019). As regards ML techniques in NLP applications, we know that the main limitation is the absence of sufficient data to train and refine classification models. We therefore proposed the use of *BERT* - Representations of bidirectional encoders by transformers - (Devlin et al. 2019) techniques for analysis, labelling and entity identification. Other NLP tools and packages such as NLTK (Bird, Loper, and Klein 2009), BertTopic (Grootendorst 2021), gensim (Rehurek and Sojka 2021) and spaCy (ExplosionAI GmbH spaCy 2022) have allowed us to analyze the documents attached to the research projects with greater precision.

Through the application of NLP tools, we have tried to extract entities (Persons, Organizations, Places, Events) and some other related metadata such as Rights and Responsibilities. A rapid examination of the documentation attached to the research projects has shown, on average, the presence of structured and unstructured sources, textual documents defined on several levels and which also include objects such as tables and graphs. The proposed approach aims to overcome some frequent problems in the information extraction process. The most frequent issues encountered include the extraction of content (metadata, keywords, entities, concepts, objects) but also the possibility of using experimental data often present in tables or images whose information is not immediately understandable and searchable. This need is very much felt in the contexts in which the documents, information and knowledge are essential to the decision support process.

Research projects represent an important form of knowledge, a factor of development both for the productive system and territories. In the absence of an integrated approach, the heterogeneity of the objects leads to a logic of

fragmentation and complexity which are not easy to govern. All the tools used in this work are based on ML or DL techniques and open source technologies. The evaluation of the selected tools was carried out on a set of domain files and the semantic quality of the metadata output has been highlighted.

3. The case study and methodology

3.1. The selection of the corpus

The first step of the work involved the selection of the corpus, followed by the training phase and metadata extraction process.

We selected a corpus of 78 research projects, which, for ease of recovery, were extracted from the EU Cordis database. They are all funded by Horizon 2020 and are related to the Mercury pollution domain (Fig.1).

Source	Topic	Projects	Selected projects for the corpus
CORDIS EU research results	Mercury pollution	232	78

Figure 1. Some information about corpus composition.

The choice of such a vast domain allows the heterogeneity of contents related not only to Mercury pollution but also to the impact on the health of living beings.

The selected case study is also representative because of the production of public and private documents that can characterize the management of Horizon 2020 projects.

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.). Consequently, our corpus of 78 projects contains 604 documents. For the purpose of this work we have carefully analyzed deliverables as they are particularly representative since they are both scientific and administrative documents. We considered only electronic documents, in PDF format, which make up 98% of the deliverables related to the selected projects (Fig. 2).

Programme	Period	Language	Deliverables/selected projects	File format
H2020	2016-2025	english	604/78	PDF/JPEG/XLSX

Figure 2. The deliverables in the selected corpus.

3.2. Selection of tools and definition of training sets

In the first part of our experiment, we used the Cermine (Tkaczyk et al. 2015) framework for automatically extracting metadata. This software application works according to a workflow that activates different tasks: convert documents in XML; extract metadata; identify, separate and save any objects different from text (i.e. images of the chart or table type, in our case).

In particular, Cermine analyzes the documents at five different levels:

1. reading and identification of the characters (dimensions and page coordinates) of the document;
2. recognition of the different sections of the document by the geometric analysis of the pages (page segmentation);
3. detection of characters and structures, heuristic analysis for the correct order of reading text areas;
4. text classification and, for each defined zone, assigning a metadata match;
5. separation of text from images for the creation of two different kinds of output: a file for text and metadata (in NLM JATS format) and a directory for the images (png format files).

Figure 3 represents the metadata extraction workflow in the Cermine framework.

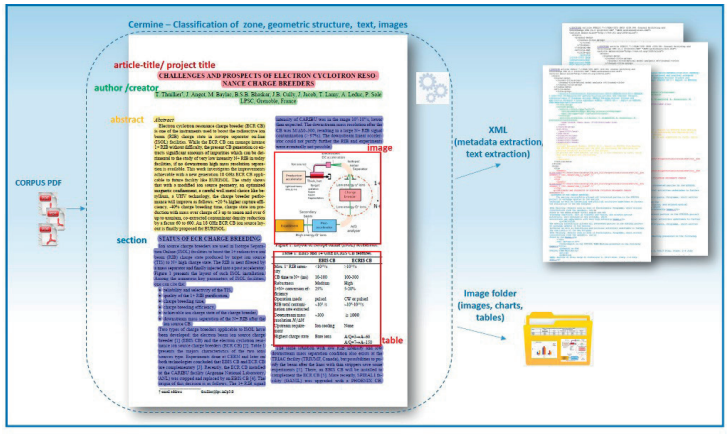

Figure 3. The Cermine framework metadata extraction workflow.

The use of this software requires the definition of models and training sets, subsequently, to enrich and define the extracted knowledge set, we used NLP technologies. By applying NLTK, BERT, BERTopic, gensim and spaCy, we extracted contents and improve the quality of the information also by detecting document entities and relationships. In archival description entities and relationships are ontological keys and create interoperability between different knowledge bases. Figure 4 shows the process of extracting metadata using NLP tools.

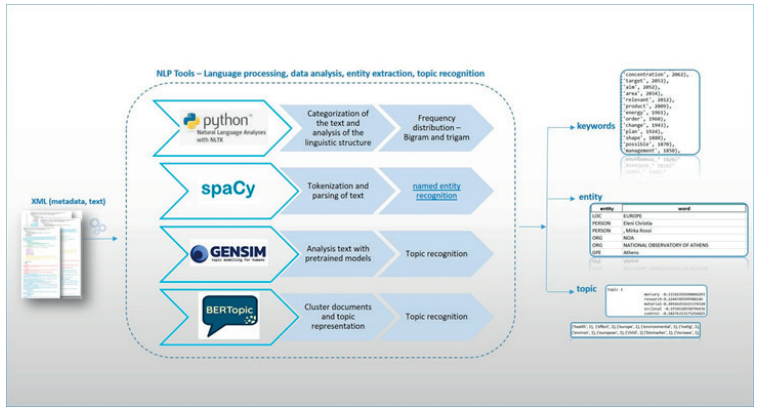

Figure 4. The process of extracting metadata using NLP tools.

In general, automatic metadata extraction produces positive interoperability effects by structuring knowledge and making the unstructured text machine readable. From the analysis of the attached documents, data and information are often in the form of tables or charts (73% of documents). Charts and tables may contain research data, accounting data, administrative information, context statistics, and so forth. The extraction of knowledge from such objects can be useful for several purposes including decision support, rationalization of resources for research, processing of statistics, information retrieval, etc. In our corpus, tables and charts are often images inserted in the text and knowledge extraction can be carried out using dedicated tools. We used several DL algorithms, such as EfficientNet CNN model (Tan and Quoc 2019), Pytesseract (Hoffstaetter and Matthias 2022) and docTR (Mindee 2022) for chart and text recognition, along with Tabula (Aristarán 2018) and ChartReader (Rane et al. 2021) to obtain complementary data.

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 metadata extraction process was carried out as illustrated in Figure 5 below.

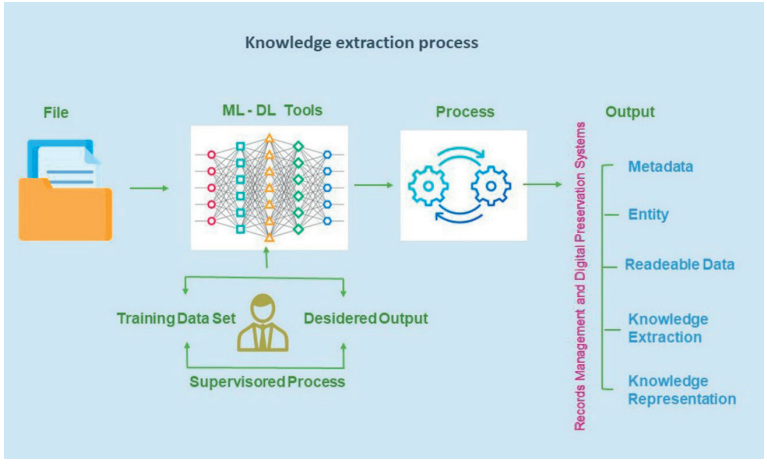

Figure 5. Knowledge extraction process.

4. Results and evaluation

The results obtained with the metadata extraction process were measured quantitatively and evaluated qualitatively through the calculation of recall and precision[4]. The extracted metadata are: ID Project, Creator, Research Program, Project Title, Record Name, Record Type, Date, File Name, Rights and Responsibility, Keyword, Topic, File Format, Entity (Persons, Organizations, Places, 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 the projects in the corpus (Fig. 6). The accuracy measurement of the recognized metadata was 98% (Fig. 7).

[4] *Recall* is the fraction of relevant instances that were retrieved. It represents the ability to find all positive instances, i.e. the percentage of correctly extracted values with respect to the total present in the sample (where TP = correct number of extractions, TF = total number of occurrences of the metadata element in the sample). *Precision* is the fraction of relevant instances among the retrieved instances. It allows the ability of not to labeling as positive a negative instance, or of not assigning to a correct tag to a wrong label (TP / AP, where TP = number of correct extractions, AP = total extractions detected).

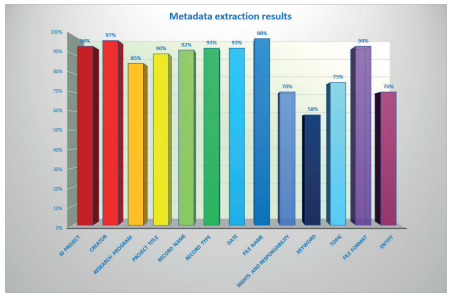

Figure 6. Metadata extraction results.

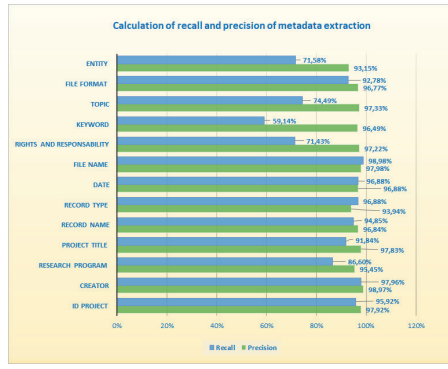

Figure 7. Calculation of recall and precision of metadata extraction.

Although the quantitative values relating to entities and rights and responsibilities, have not been optimal, we can observe a correct semantic reading. In this work we carried out a first test on methodologies and techniques that, in our opinion, might have a broad application in archival metadata. Overall, the automatic extraction of metadata obtained is positive as a consequence of the supervision process used. This has allowed a significant improvement in the performance of the analyzed tools. Regarding table extraction, in most cases, they have been recognized and data have been read. However, the results obtained with Tabula were not always optimal due to the variety of alphabets, symbols and structures used in the documents. To this end, it will be possible to improve the described results by running a further training stage. The diagram in Figure 8 represents the type of errors and the percentage of frequency.

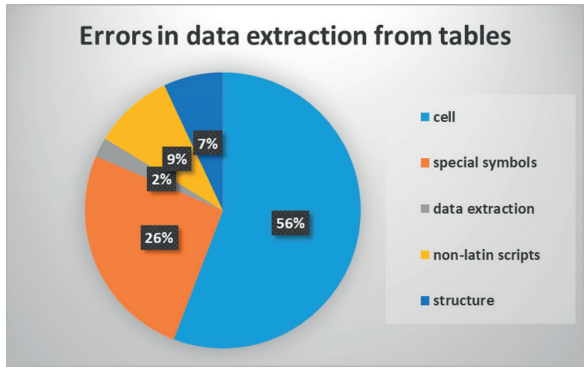

Figure 8. Errors in data extraction from tables.

Finally, the example in Figure 9 shows the result of automatic data and information extraction from charts. 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 9, are full of lines and data. The algorithms used have not been able to accurately recognize the different positions of the points on the chart, moreover the consistent association with the reference label was not always correct, despite the support provided by a supervised process.

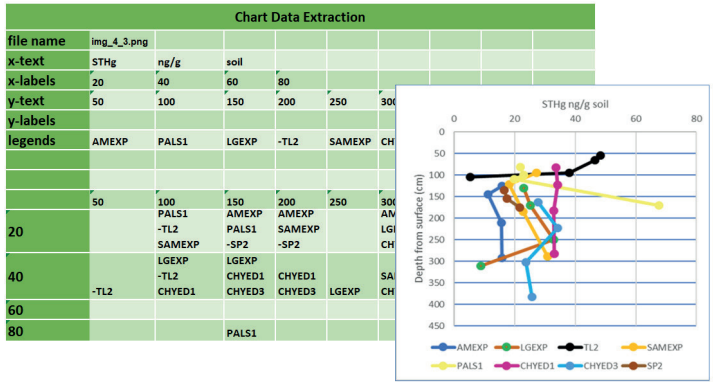

Figure 9. Chart data extraction.

5. Lessons learned

Several studies focus on the application of ML and DL technologies to archives (Colavizza et al. 2022). Similarly, in other contexts, numerous efforts aim to use automatic data and metadata extraction technologies on texts. In this paper we reported 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 possible route in such a complex and articulate journey.

References

Aristarán, Manuel. 2018. "Tabula". (Version v1.2.1). Accessed June 22, 2023. https://github.com/tabulapdf/tabula.

Bird, Steven, Edward Loper and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

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

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Tautanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of NAACL-HLT Minneapolis, Minnesota*, edited by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, 1, 4171-86. https://doi.org/10.18653/v1/N19-1423.

European Commission. n.d. "Horizon 2020 country profile." Accessed June 22, 2023. https://research-and-innovation.ec.europa.eu/statistics/framework-programme-facts-and-figures/horizon-2020-country-profiles_en.

ExplosionAI GmbH.2022. "spaCy". (Version v3.4.0). Accessed June 22, 2023. https://spacy.io/.

Grootendorst, Maarteen. 2021. "BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics (Version v0.7.0)." Accessed June 22, 2023. https://github.com/MaartenGr/BERTopic. https://doi.org/10.5281/zenodo.4381785.

Hoffstaetter, Samuel, and Matthias Lee. 2022. "Pytesseract."(Version v0.3.10) Accessed June 22, 2023. https://pypi.org/project/pytesseract/.

Kamath, Uday, Liu John, and James Whitaker. 2019. "Deep Learning for NLP and Speech Recognition." Cham: Springer Nature. https://doi.org/10.1007/978-3-030-14596-5.

Mindee. 2022. "docTR".(Version v0.5.1). Accessed June 22, 2023. https://github.com/mindee/doctr.

Rane, Chinmayee, Seshasayee M. Subramanya, Devi S. Endluri, Jian Wu, and Lee C Giles. 2021. "ChartReader: Automatic Parsing of Bar-Plots." In *IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, Las Vegas, USA, 318-25. https://doi.org/10.1109/IRI51335.2021.00050.

Rehurek, Radim, and Petr Sojka, 2021. "Software Framework for Topic Modelling with Large Corpora". (Version v4.1.0). Accessed June 22, 2023. https://github.com/piskvorky/gensim.

Rovella, Anna, Alexander Murzaku, Eugenio Cesario, Martin Critelli, Armando Bartucci, and Francesca Maria Caterina Messiniti. 2022. "Analysis, evaluation and comparison of knowledge extraction tools in the environmental and Health domain. A holistic approach." In *Proceedings of the International Knowledge Organization and Management in the Domain of Environment and Earth Observation (KOMEEO) Conference*, edited by Antonietta Folino and Roberto Guarasci. Advances in knowledge organization 18. Würzburg: Ergon Verlag, 121-46. https://doi.org/10.5771/9783956508752-121.

Tan, Mingxing, and Le V. Quoc. 2019. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." In *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, California. arXiv:1905.11946. https://doi.org/10.48550/arXiv.1905.11946.

Tkaczyk, Dominika, Pawel Szostek, Mateusz Fedoryszak, Piot Jan Dendek, and Lukasz Bolikowski. 2015. "CERMINE: automatic extraction of structured metadata from scientific literature." *International Journal on Document Analysis and Recognition* 18(4): 317-35. https://doi.org/10.1007/s10032-015-0249-8.