A System for the Management, Visualization and Annotation of Digital Collections of Handwritten Historical Documents

Carlos Andrés Correa-Guillén, Ángel Gómez, and Estefanía López-Salas

Faculty of Computer Science, Universidade da Coruña, 15071 A Coruña, Spain
Centro de Investigación CITIC - Laboratorio Interdisciplinar de Aplicaciones de la Inteligencia Artificial (LIA2), Faculty of Computer Science, Universidade da Coruña, 15071 A Coruña, Spain
Grupo de Estudios Territoriales (GET), Faculty of Architecture, Universidade da Coruña, 15071 A Coruña, Spain
Correspondence: carlos.guillen@udc.es; angel.gomez@udc.es; estefania.lsalas@udc.es

DOI: *https://doi.org/10.17979/spudc.000024.27*

Abstract: Today, there are huge amounts of handwritten documents accessible in digital collections. However, working with these documents for research purposes is limited due to the lack of appropriate technological tools to read, transcribe and analyze them in the same digital environment. This paper presents a web-application that allows a dynamic interaction with a particular collection, the so-called General Answers of the Cadaster of Ensenada. The app was designed to give access and support the analysis of these historical handwritten documents that are crucial to study the natural and human environment of the 18^{th}-century Spain. The app allows parameterized searches, the visualization and annotation of digitized documents, and its integration with other applications through an API. In addition, it includes a functionality to segment and classify handwritten numbers automatically.

1 Introduction

Spanish digital archives and libraries, such as the Portal of Spanish Archives (PARES),[1] the Virtual Library of Defense (BVD),[2] the Virtual Library of Documentary Heritage (BVPB),[3] or the Hispanic Digital Library (BDH), [4] contains a huge amount of digital documents, from several time periods, that anyone interested can freely access through the corresponding online platforms. The amount of those sources is, in fact, growing rapidly in the last decades. However, its exploitation is still highly limited in the web-environment due to several technical challenges. We can read and explore each document, but it is still not possible to add information over them

[1] Ministry of Culture and Sport - Spain. Portal of Spanish Archives (PARES), accessible at: *https://pares.culturaydeporte.gob.es/inicio.html*.

[2] Ministry of Defense - Spain. Virtual Library of Defense (BVD), accessible at *https://bibliotecavirtual.defensa.gob.es/BVMDefensa/es/inicio/inicio.do*.

[3] Ministry of Culture and Sport - Spain. Virtual Library of Documentary Heritage (BVPB), accessible at: *https://bvpb.mcu.es/es/inicio/inicio.do*.

[4] National Library of Spain. Hispanic Digital Library (BDH), accessible at: *https://www.bne.es/en/catalogues/hispanic-digital-library*.

or to share it with others to favor, for instance, collaborative research based on data coming from historical documentation.

One rich collection accessible in digital archives is the General Answers of the Cadaster of Ensenada conducted in the mid-18^{th} century in the old Crown of Castile. This digital collection, which is accessible at *https://pares.mcu.es/Catastro/*, contains the complete copy of the General Answers that is preserved in the General Archive of Simancas currently, with the records gathered from 13.000 places (Ministry of Culture and Sport – Spain, 2023).

The General Answers were the first part of the works carried out within the Cadaster of Ensenada. This was a large-scale census and statistical investigation that aimed to implement a single tax in the Crown of Castile, that is, a fiscal reform to force each person living into the Castilian territory to pay according to their possessions (Alimento, 2002). In order to achieve that goal, it was necessary to make a register of all the properties as well as a deep on-site study of each of them, but firstly, the cadaster officers carried out a survey with 40 common general questions at each locality. The findings or answers to these questions were gathered in books, called General Answers. One book was generated per each locality of the old Crown of Castile. As a result, it was produced a huge volume of documentation that allow us to approach many different issues related to population, society, economy, or geography, to cite a few, from mid-18^{th}-century Spain (Camarero Bullón, 2002). In this paper we present the web-application Archive Lens developed to facilitate the analysis of that historical documentation by researchers through an effective management, dynamic interaction, advanced search and annotation capabilities that, on the whole, aims to contribute to the enrichment of shared knowledge.[5]

2 Conceptualization

The proposed system was developed to ease the work when a particular theme is investigated, but through the whole collection of the General Answers. In fact, its conceptualization was born in direct relation with the in-progress work within the project Mapping Hospitals.[6] The aim of this project is to investigate the network of hospitals that existed in Galicia in the mid-18^{th} century through the Cadaster of Ensenada.

The 30^{th} question of the cadaster survey was focused on gathering data about the existing hospitals in each locality. Therefore, if we pay attention to this particular question, we have the opportunity to approach the past network of hospitals, hospices, inns or rest houses that were used mainly for pilgrims on travel to Santiago de Compostela, but also by the poor and the ill.

However, the project has some limitations that derived mainly from the huge amount of records to be reviewed. So far, we have only approached the network of hospitals corresponding to Galicia and this work implied the reading of the 30^{th} question from a total of 2912 records (López Salas, 2021). Therefore, to have a tool to ease the management, visualization, annotation and search through these historical documents will reduce the cost and time needed to broaden the scope of the study to other territories of the old Crown of Castile, but also to proposed new fields of research that crosscut large amounts of cadaster records.

3 Description of the application and its functionalities

The application Archive Lens is organized into three main levels: (1) Map; (2) Selector of localities, and (3) Visualization and annotation of documents. The first level, the Map, presents a graphic representation of the geographic area that was affected by the cadastral process in the mid-18^{th} century. In this level, the user is able to interact with the map and to select a province in which a number of historical documents were created (Figure 1).

---

[5] This paper is the result of the work done as a Final Degree Project by Carlos Andrés Correa-Guillén, under the supervision of Ángel Gómez García and Estefanía López Salas, and presented at the Faculty of Computer Science, University of A Coruña, in July 2023 (Correa-Guillén, 2023).

[6] Mapping Hospitals Project, accessible at: *https://mappinghospitals.udc.es/index_es.html*.

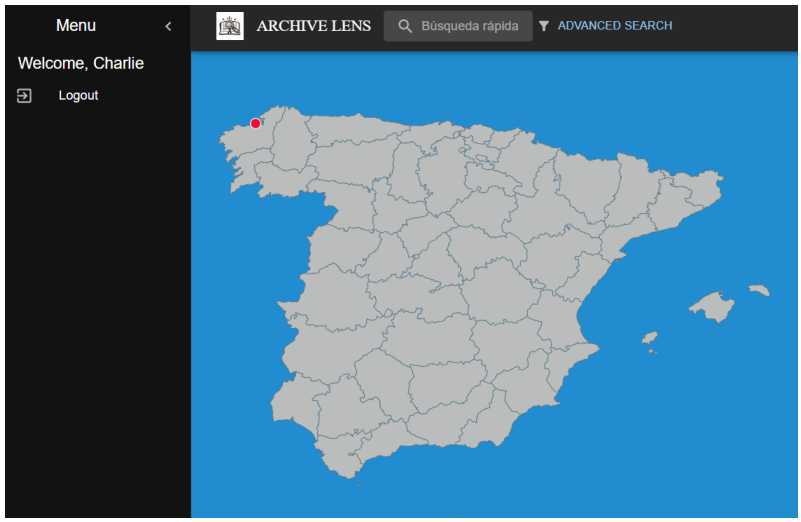

Figure 1: The Application Archive Lens: the map.

When a province is selected, the second level of the application is opened for the selection of a specific locality within the whole province (Figure 2). Each locality is presented as an independent card with information about: the present name of the locality, the name of the locality in the mid-18th century, a short description of the place, a representative image, and the total number of documents that it comprises. These documents are the digitized pages of the General Answers corresponding to each locality of the province. When a specific locality is selected by the user, he/she can access the third and last level of the application, which is devoted to a tool for the visualization and annotation of historical documents (Figure 3).

Figure 2: The Application Archive Lens: the selector of localities.

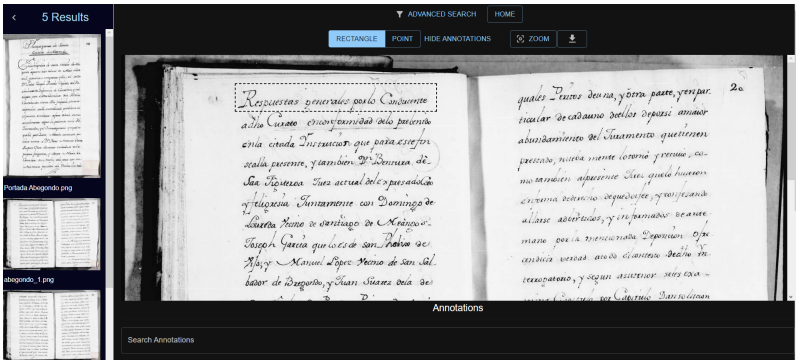

Figure 3: The Application Archive Lens: the visualization and annotation tools.

In the third and last level, the user is able to interact with each digitized page of the General Answers generated for a specific locality. We included traditional tools such as zoom in and zoom out, but also new ones to facilitate the interaction of the user with the handwritten text. In the upper part of the Visualization tool, three buttons are displayed to allow annotations over the historical documentation. The first two buttons, called Rectangle and Point, indicate the two types of annotations the user is able to create. The type called Rectangle allows the user to select a rectangular area of the document to be annotated, for instance, with the corresponding transcription of the historical text. The second type of annotations is thought to add Points to specific parts of the documents where the user may need, for instance, to include a comment for another researcher working with the same locality. Therefore, the annotation tools were designed to facilitate an effective interaction between the reader and the archival documentation while enriching the information that it already contains in a new digital layer.

Another important functionality that was explored during the development of the present work is directly related to improving the search and filter of specific data within this cadaster. Based on the fact that each book of the General Answers is organized with the same structure of 40 questions, we started with the implementation of a functionality to segment and classify handwritten numbers automatically. We applied the TrOCR model that consists of "an image Transformer encoder and an autoregressive text Transformer decoder to perform optical character recognition (OCR)" (Li et al., 2022). The results are shown in Figure 4.

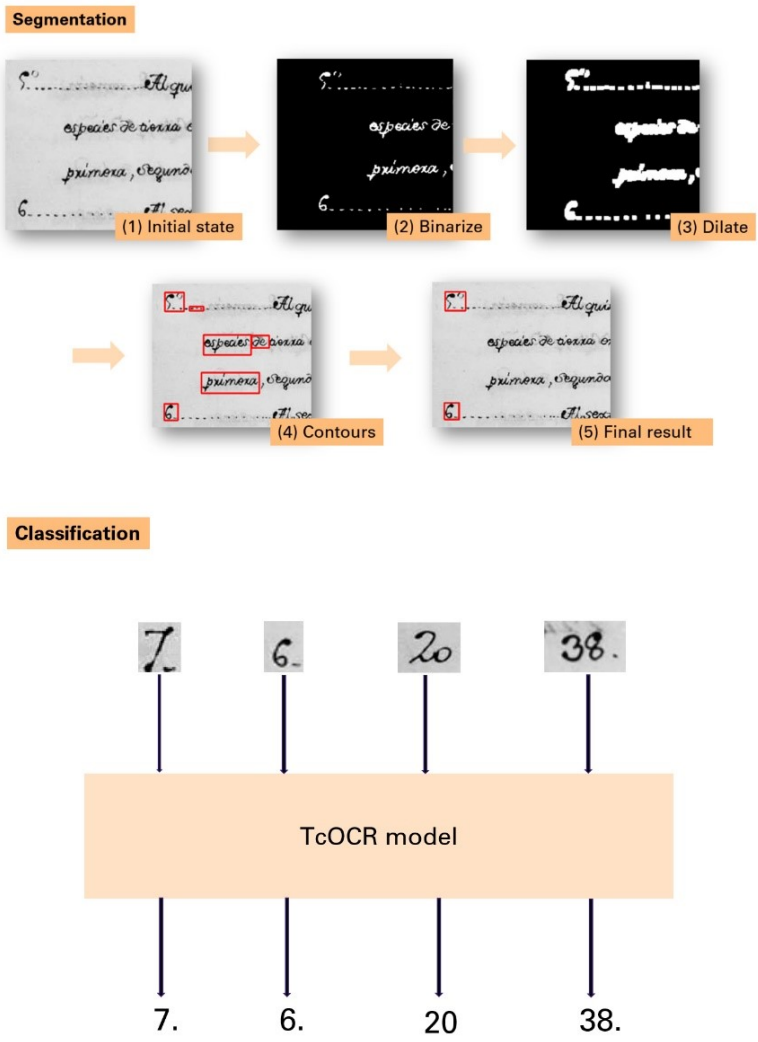

Figure 4: Summary of process for segmentation and classification of handwritten numbers in the Cadaster of Ensenada.

4 Conclusions

The application presented in the paper offers a set of advantages to work with historical handwritten documents as it integrates visualization and annotation capabilities that make easier the reading, non-automatic transcription, and collaborative research in a web-environment. Although it was designed for the Cadaster of Ensenada, it could be easily adapted to other historical documentation. We have opened new opportunities, but the word done also poses some challenges to be faced in future developments. For instance, we still need to improve the management system of users in order to be able to assign roles and verify permissions. This way, we will offer the possibility to promote discussions through annotations and also customization of each user type. In addition, the handwritten text recognition system is only focused on question numbers, but it will be really useful for historical research to explore how to apply it

to the whole text in order to allow more specific and insightful searches in the rich content of this cadaster of the mid-18^{th} century Spain.

Acknowledgements

CITIC is funded by the Xunta de Galicia through the collaboration agreement between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities for the reinforcement of the research centres of the Galician University System (CIGUS).

Bibliography

A. Alimento. Los catastros del siglo XVIII, entre tradición y modernidad. *CT: Catastro*, 46:17–26, 2002.

C. Camarero Bullón. El Catastro de Ensenada, 1745-1756 diez años de intenso trabajo y 80.000 volúmenes manuscritos. *CT: Catastro*, 46:61–68, 2002.

C. A. Correa-Guillén. *Web tool for structured viewing and online annotation of historical manuscripts*. Final Degree Project. Faculty of Computer Science, University of A Coruña, A Coruña, 2023.

M. Li, Lv, Tengchao, Chen, Jingye, Cui, Lei, Lu, Yijuan, Florencio, Dinei, Zhang, Cha, Li, Zhoujun, and Wei, Furu. TrOCR: Transformer-based Optical Character Recognition with Pretrained Models. *Arxiv. Computer and Language*, pages 1–10, 2022.

E. López Salas. Cartografía de la hospitalidad en los caminos de peregrinación: el proyecto Mapping Hospitals. *Anales de Geografía Complutense*, 2(41):389–407, 2021.

Ministry of Culture and Sport – Spain. General Answer of the Cadaster of Ensenada. *https://pares.mcu.es/Catastro/*, 2023. [Online; accessed 1-September-2023].