A computational framework for the analysis of the Uruguayan dictatorship archives

Lorena Etcheverry[1], Leopoldo Agorio[2], Virginia Bacigalupe[1], Sofía Barreiro[1], Elena Bing, Samuel Blixen[3], Daniel Calegari[1], Lautaro Cardozo[1], Fernando Carpani[1], Felipe Chavat[1], Diego Garat[1], Alvaro Gómez[2], Ernesto Fernández[1], Federico Fioritto[1], Fabián Hernández[3], Rodrigo Laguna[1], Victor Marabotto[1], Guillermo Moncecchi[1], Ignacio Ramírez[2], Aiala Rosa[1], Javier Stabile[1], Jorge Tiscornia[4], Nilo Patiño[4], Lía Rivero[1], Dina Wonsever[1], Guillermo Zorron[1], and Gregory Randall[2]

[1] Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay.
[2] Instituto de Ingeniería Eléctrica, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay.
[3] Facultad de Información y Comunicación, Universidad de la República, Montevideo, Uruguay.
[4] Madres y Familiares de Uruguayos Detenidos Desaparecidos, Montevideo, Uruguay.

Abstract. Between 1973 and 1985, a civic-military dictatorship ruled in Uruguay. Systematic violations of human rights marked this period. Project Cruzar.uy aims to develop tools and methodologies to analyze historical documents from that period. We present the advances in this ongoing project. We describe a set of tools to automatize the extraction and organization of information from the archives using computational tools including image processing, machine learning, natural language processing, information extraction and integration.

Keywords: Documents analysis and recognition, OCR · historical documents retrieval.

1 Introduction

Between 1973 and 1985, a civic-military dictatorship ruled in Uruguay, preceded by state terrorism from 1968 to 1973. Systematic violations of human rights, including the use of torture, raping, kidnapping, and the forced disappearance of hundreds of Uruguayans marked this period. The investigation of these crimes has been hampered by a complicity web involving military personnel, institutions, and politicians involved in those events, obstructing the access to the different sources of information that still exist from that period. However, over the years, access to some data sets has been achieved. One is the so-called *Archivo Berrutti*, which contains approximately 3 million pages of diverse material produced by the security agencies during the dictatorship and ensuing years. The documents in this collection consist of digital scans of microfilm reels of the original paper documents, no longer available. Other collections are partially digitized: e.g. the Historical Archive of the former National Directorate of Information and Intelligence, and the Archive of the Naval Fusiliers Corps.

The documents in these archives are heterogeneous. Among their contents are interrogation reports, press clippings, lists of people and places, personal records, pictures, passports, political affiliation, and any other background information deemed useful by the security agencies. These documents also contain codes that allow them to be related to additional files. However, without collaboration from the military, such codes' meaning is unclear and of small help for organizing them. A similar situation occurs with codenames used to designate places and targets.

These archives' contents are precious to understand this period of Uruguayan history better, assist in searching for missing persons, contribute elements to the ongoing trials, and build citizenship and awareness of this historical period. At least two projects focused on analyzing this documentation and related information of the dictatorship period [4,22]. In both cases, transcription and analysis were manually performed on a tiny subset of the archives to the best of our knowledge. Unfortunately, this methodology is not sufficient to process the millions of document pages that are yet to be analyzed.

More than two years ago, the project Cruzar.uy was formulated to develop a methodology capable of processing the massive amount of information in the military archives in a relatively short time. This interdisciplinary project involves researchers and collaborators ranging from electrical engineers and computer scientists to journalists, archivists, historians, and members of Uruguayan human rights organizations. The present paper presents the advances in this ongoing project. In particular, we describe a set of tools that we are deploying to automatize the extraction and organization of information from the archives employing computational tools and techniques including image processing, machine learning, and natural language processing (NLP), and information extraction and integration.

2 Related work

From a technical point of view, our work falls within the trans-disciplinary field of Digital Humanities, where digital technology and computational methods are applied to humanities research [7]. In this context, some works explore the design of tools that help researchers to find historical documents [24,25], following a similar approach to the one proposed in the current project. It is also worth mentioning the International Consortium of Investigative Journalists' work that analyzes the Panama Papers [14]. In the current project context, it is crucial to extract events and facts together with temporal information; related work on temporal information retrieval is collected in [8]. Latin American countries have been subject to several dictatorships during the XX century. The analysis of the documents produced during those periods has been carried on manually by Human Rights organizations, in general with little support from governments. As a consequence, the technical resources available for such analysis are minimal. Some exceptions include the *Equipo Argentino de Arqueología Forense* with extensive experience and known successes in forensic approaches in Argentina and many countries with similar situations [12]. The Comisión Provincial de Memoria analyses the Police Intelligence Directorate's archive of the Province of Buenos Aires, Argentina. In 1992, in Paraguay, Martín Almada found what is known as the "terror archives": a substantial body of documents related to the Stroessner dictatorship and the "Plan Condor" that articulated the repressive actions of several South American countries during the 1970s and 1980s[2]. The project Memoria Ojo Público has analyzed extensive documentation about the state terrorism during the dictatorships in Peru' [1]. An international effort, based in the University of Texas at Austin, analyzes about 80 million digitized documents obtained from the Guatemalan National Police Historical Archive [26].

The preceding examples share similarities with the one we are dealing with at Cruzar. In all these cases, the archives have been digitized, processed, and analyzed manually to the best of our knowledge. No tools have been developed to automatize those time-consuming tasks, which could significantly boost the capacity to extract and correlate information that could enable more in-depth and broader investigations to occur.

3 Our approach

The Cruzar Project's goal is to systematize and organize the aforementioned historical documents to maximize the quantity and quality of extracted knowledge. Hopefully, such information will allow us to understand the mechanisms and operation of the repressive system, to identify the processes that led to the disappearance of prisoners (which could,in turn, allow to find their remnants and provide solace to their relatives), and help in the process of making justice. Contributing to the understanding of that historical period may help to avoid future repetitions.

In this context, we are continually building and improving information systems to help researchers find, extract and discover useful information from the documents and their relationships. Our approach combines tools and techniques from Image Processing, Computer Vision, NLP and Knowledge Graphs to deal with several use cases. These include finding all the documents that mention a specific person, place, or organization; building timelines on people, places and facts; and reconstructing the repressive corps' organization charts. Controlled vocabularies and ontologies have a crucial role in our approach, guiding the information extraction process and assisting in data integration tasks. Given the nature of this project, to guarantee the traceability of the information, it is necessary to maintain the link between the transcribed text, the information subsequently extracted, and the original documents at all times..

Figure 1 shows an overview of our strategy. The *data preparation* stage consists of several image pre-processing tasks, targeted at improving image quality, and classifying images according to different criteria (e.g., document type, image

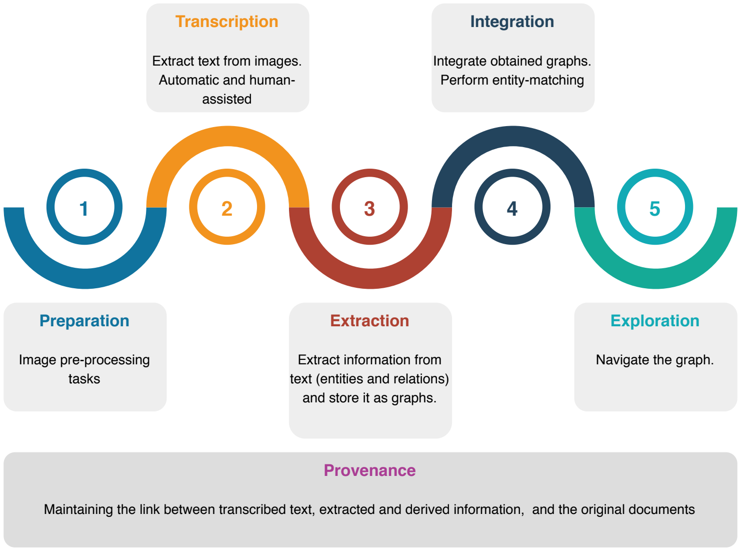
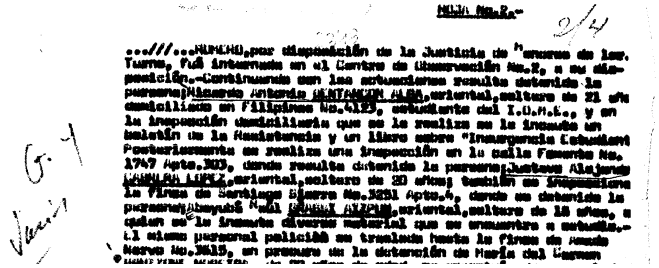

Transcription
Extract text from images. Automatic and human-assisted
Integration
Integrate obtained graphs. Perform entity-matching
1
2
3
4
5
Preparation
Image pre-processing tasks
Extraction
Extract information from text (entities and relations) and store it as graphs.
Exploration
Navigate the graph.
Provenance
Maintaining the link between transcribed text, extracted and derived information, and the original documents

Fig. 1. System overview. Our strategy involves five stages: data preparation, text transcription, information extraction, information integration, and exploration.

quality, etc.). The text contained in the digitized document images is obtained in the *transcription* stage. The resulting text is stored in a relational database, keeping track of data provenance as the relationship between the source images and the produced text. Several *information extraction* tasks are performed in a third stage, using NLP methods to extract Named Entities and relations. Ontologies and controlled vocabularies guide this process and are used to annotate and represent the extracted assertions; these are stored as Resource Description Framework (RDF)[28]] *triples*, atomic pieces of knowledge encoded as three entities (e.g., "subject-predicate-object"). Extracted triples are stored in a *triple-store:* a database capable of handling RDF data. In this stage, we also keep track of each assertion's provenance in terms of text segments. The *integration* stage provides a unified view of the sub-graphs extracted from each document. Entity Matching tasks produce integrated Knowledge Graphs required to navigate statements pulled from different documents in this stage. Finally, *exploration* tools combine different approaches as text-search, faceted-search based on the ontology, and timelines to explore documents' collection. In the following sections, we provide further details on each stage.

3.1 Data preparation

Our current work is focused on the so-called Berrutti Archive: an extensive collection of documents found on military facilities while Azucena Berrutti was the Uruguayan Defense Minister. This collection of microfilmed documents contains around 3 million pages span from 1965 to 1999, arranged in microfilm rolls of



HOJA No.2.- 2/4

...///...NUMERO, por disposición de la Justicia de Menores de 1er. Turno, fué internado en el Centro de Observación No.2, a su disposición.-Continuando con las actuaciones resulta detenida la persona: Ricardo Antonio [illegible] ALBA, oriental, soltero de 21 años domiciliado en Filipinas No.4123, estudiante del I.U.M.E., y en la inspección domiciliaria que se le realiza se le incauta un boletín de la Resistencia y un libro sobre "Insurgencia Estudiantil Posteriormente se realiza una inspección en la calle Fomento No. 1747 Apto.303, donde resulta detenido la persona: Justavo Alejandro [illegible] LOPEZ, oriental, soltero de 20 años; también se inspecciona la finca de Santiago Gómez No.3251 Apto.4, donde se detenido la persona: [illegible], oriental, soltero de 18 años, a quien se le incauta diverso material que se encuentra a estudio.- El mismo personal policial se traslada hasta la finca de Amado Nervo No.3615, en procura de la detención de María del Carmen

Fig. 2. Sample digitized image. Pixels are either black or white. Artifacts typical of typewritten material can be observed such as irregular strength due to varying amount of ink or stroke force, paper grain, ink diffusion. Text can be covered by handwritten annotations, seals, marks, dirt, etc.

about 2500 images each. There is no apparent organization within each roll, besides some rough chronological order related to each one. The original documents include handwritten text, machine written pages, pictures, and portions of printed material (as newspapers). The process of digitizing the microfilm rolls took place long before this project started; the digitized images are the sole material we have. Instead of colour, or even gray-scale, the digitized images are binary (black or white, no shades in between); this poses a challenge to any image processing or computer vision task (such as Optical Character Recognition). Figure 2 shows some examples.

Annotation The annotation stage's goal is to enrich documents with relevant metadata such as the date, the type, the origin and eventually, a brief description. Moreover, documents may have several pages, and we want to signal these relations in the corresponding images. Most of the image annotation tools focus on the segmentation of an image. In our case, the segmentation of regions of interest (e.g. stamps, signatures) is important, but the complete annotation of a document must include other metadata. Since the addressed documents have sensitive content, web applications were not considered. Among the standalone applications, we selected LabelMe [23,29] because it has both segmenting and annotating tools and some basic global tagging. The open-source code was adapted to include the features mentioned above specific for documents.

This customized version of LabelMe allows annotating different aspects of a document, including its type, date, index of the page in a multi-page document, among other things. We have identified about 74 different types of documents (e.g.reports, letters, police files) and 52 different origins (e.g. divisions of the army or security agencies). So far, over 140000 documents have been annotated in this way by about 150 students and volunteers since 2019. Besides its direct use as valuable information about the documents, the labeled documents also constitute a reputable source of labeled data for training the various automatic classification algorithms that we are currently working on. For example, LabelMe

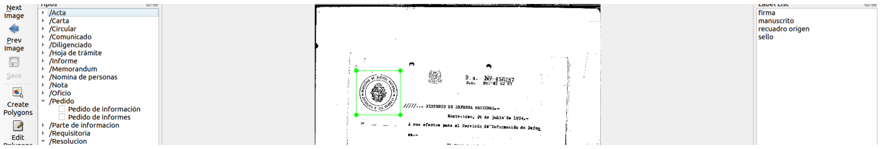


Fig. 3. Screenshot of LabelMe tool showing the stamps and signature selection feature.

lets the user select regions of the image containing stamps and signatures (see Figure 3) to test automatic signature and stamp detection algorithms.

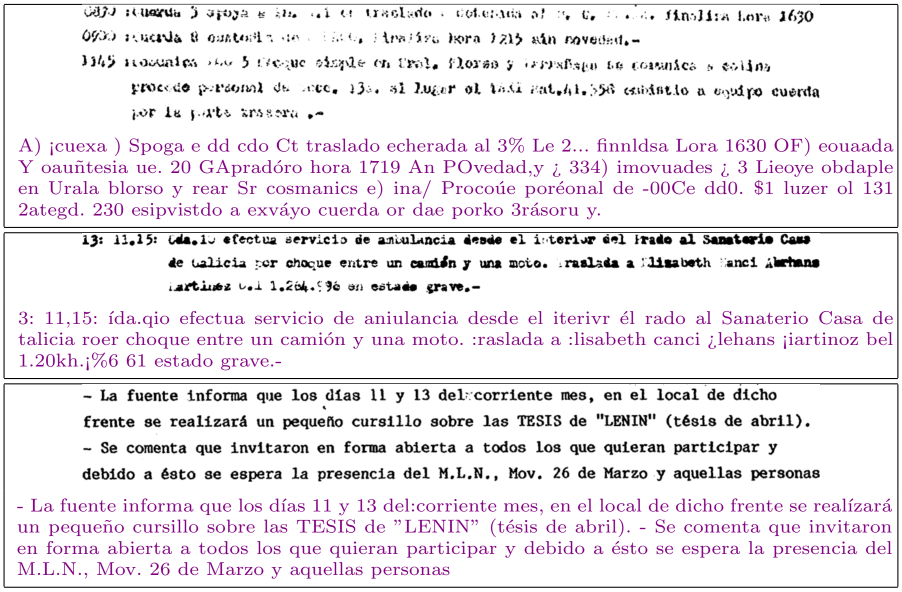
3.2 Text Transcription and Image Readability Assessment

Automatic transcription and readability assessment The low quality of the source images implies that even sophisticated commercial OCRs are ineffective for a significant part of the digitized archives. On the other hand, humans can easily read several of these problematic documents. Based on the above observation, we have implemented a hybrid manual-automatic transcription system which consists of three stages. First, an automatic transcription is obtained using the off-the-shelf, freely available Tesseract OCR (version 4.1) [5]. Then, the quality of the OCR output is estimated; this has to be done *without access to the correct transcription*. If the quality is deemed satisfactory, the OCR transcription is fed directly to the database. Otherwise, the document is queued for manual transcription in the next step. Note that this stage is critical, as documents which are incorrectly labeled as *well transcribed* will likely not be read by anyone afterwards. Thus, we put a significant effort to produce a reasonable estimation of the document readability.

Manual transcription – LUISA: If the OCR transcription of a document is deemed unsatisfactory, the image is fed to a web-based crowd-sourced transcription platform, named LUISA, which was developed by our team for this sole purpose. We will now provide more detail on the last two stages.

Readability score Given the sheer number of documents to be processed, a manual assessment of the readability of each document is out of reach. Instead, we must rely on some automatic method for estimating such readability. To achieve this, we developed a *readability score* r of the text produced by the OCR. We use a custom *enriched dictionary* D, whose quality is crucial in computing such score. Besides that, we need to ascertain that the score works as expected, at least on a number of known cases for which we do have a complete transcription. Given a text t formed by n words, $t = \{w_1, w_2, \ldots, w_n\}$, the score $r(t)$ is defined as $r(t) = \sum_{l=1}^{L} a_l p_l(t)$, where $p_l(t)$ is the fraction of words of length l in t that are present in the dictionary D, and a_l is a coefficient to be determined.

[5] https://github.com/tesseract-ocr/tesseract



0830 :cuerda 3 spoga e [illegible] traslado [illegible] finaliza hora 1630
0930 :cuerda 8 custodia de [illegible], finaliza hora 1215 sin novedad.-
1145 :comunica [illegible] 3 choque simple en Gral. Flores y [illegible] se comunica a [illegible]
procede personal de [illegible]. 13a. al lugar el taxi Mat.41.558 asistido a equipo cuerda
por la parte trasera .-

A) ¡cuexa) Spoga e dd cdo Ct traslado echerada al 3% Le 2... finnldsa Lora 1630 OF) eouaada Y oauñtesia ue. 20 GApradóro hora 1719 An POvedad,y ¿ 334) imovuades ¿ 3 Lieoye obdaple en Urala blorso y rear Sr cosmanics e) ina/ Procoúe poréonal de -00Ce dd0. $1 luzer ol 131 2ategd. 230 esipvistdo a exváyo cuerda or dae porko 3rásoru y.

13: 11.15: Cda.10 efectua servicio de ambulancia desde el interior del Prado al Sanatorio Casa
de Galicia por choque entre un camión y una moto. Traslada a Elisabeth [illegible]anci Abrhans
Martinez C.I 1.264.996 en estado grave.-

3: 11,15: ída.qio efectua servicio de aniulancia desde el iterivr él rado al Sanaterio Casa de talicia roer choque entre un camión y una moto. :raslada a :lisabeth canci ¿lehans ¡iartinoz bel 1.20kh.¡%6 61 estado grave.-

- La fuente informa que los días 11 y 13 del corriente mes, en el local de dicho
frente se realizará un pequeño cursillo sobre las TESIS de "LENIN" (tésis de abril).
- Se comenta que invitaron en forma abierta a todos los que quieran participar y
debido a ésto se espera la presencia del M.L.N., Mov. 26 de Marzo y aquellas personas

- La fuente informa que los días 11 y 13 del:corriente mes, en el local de dicho frente se realízará un pequeño cursillo sobre las TESIS de "LENIN" (tésis de abril). - Se comenta que invitaron en forma abierta a todos los que quieran participar y debido a ésto se espera la presencia del M.L.N., Mov. 26 de Marzo y aquellas personas

Fig. 4. Very bad, mediocre, and very good documents and their OCR transcriptions.

To determine the coefficients a_l, 50 documents were selected and divided into five groups of 10 documents each according to our subjective assessment of readability (very bad, bad, mediocre, good, very good); this is illustrated in Figure 4. We then manually transcribed those documents as best as possible, and computed an *ideal readability score* based on the Levenshtein distance [18] between the OCR and manual transcriptions. Concretely, given the ideal score $s(t_j)$ and the vector of proportions $(p_1, \ldots, p_L)_j$ computed for each j-th of the $N = 50$ selected documents, the weights $(a_1, \ldots, a_L)$ were estimated so that $r(t_j)$ and $s(t_j)$ were as close as possible employing a linear regression. Figure 5 shows the performance of the adjusted score in terms of the best fit to the ideal score, and its relationship to our subjective quality assessment. Using a threshold of 60%, about 300000 images were deemed of enough quality to be fed directly to the following stages.

Enriched dictionary Since the dictionary determines the validity of a transcribed word, it must include all the words that may appear in a document. In our case, the list of valid words goes far beyond those found in off-the-shelf Spanish dictionaries: names and surnames derived from many countries and languages, including all possible spellings; idiosyncratic expressions; military jargon, acronyms, abbreviations, as well as common misspellings of words.

An initial dictionary contained words from a standard Spanish dictionary and lists of names and surnames found online. A list of domain-specific acronyms was then added. Then, the OCR was used to transcribe a subset of the documents, and produce a list of words which did *not* appear in the initial dictionary. This

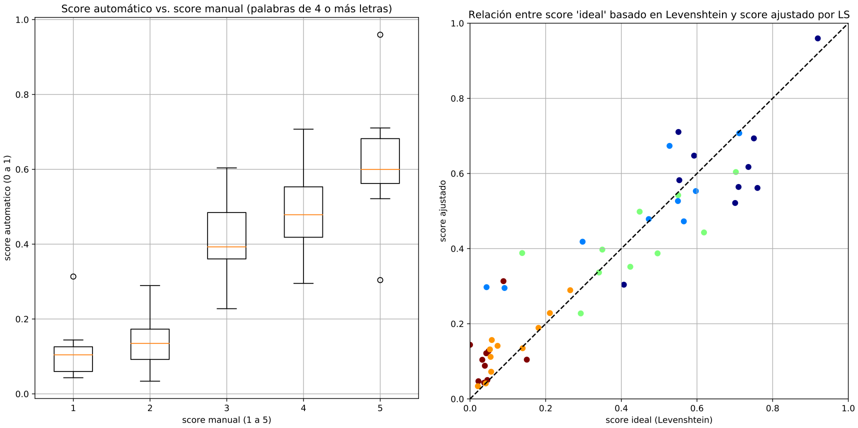


Fig. 5. Adjusted score. Left: automatic score for each of the 5 subjective qualities; a clear jump can be observed between classes 1,2 and 3,4,5. Right: least squares regression; dot colors correspond to classes 1 (red), 2 (orange), 3 (green), 4 (cyan) and 5(blue). Notice that our score does not produce very low scores on classes 4 and 5, whereas the ideal (Levenshtein) score does so for a few samples.

list, of about 200000 words, was manually filtered in search for valid terms. In this way we obtained 16000 additional words which were added to produce the final enriched dictionary.[6]

LUISA: a crowd-sourcing transcription platform The goal of LUISA (*Leyendo Unidos para Interpretar loS Archivos*) is to transcribe low quality documents through collaborative networking. The tool is named after the Uruguayan human rights activist LUISA Cuesta (Nebio Ariel Melo Cuesta's mother, disappeared). We developed a crowd-sourcing approach in order to promote social participation in the project. Users of this web application [7], which can be used on both computers and smartphones, are asked to transcribe what they see at small portions of the documents (blocks): letters, numbers, symbols, etc. Each block of text is presented at least three times randomly to different users to have a set of transcriptions for each one. Programs developed by our team automatically align the images and generate these blocks. These are then combined to obtain transcripts of the entire document using a majority vote mechanism to determine each block of text's transcription. In its design, LUISA takes care of the collaborators' privacy and considers the handled documents' sensitivity: no information that could help identify people who have collaborated is kept, neither whole documents are shown. LUISA's social impact is twofold: on the one hand, it helps to recover the documents. On the other, it creates awareness in

[6] The dictionary is freely available and can be downloaded from http://iie.fing.edu.uy/~nacho/data/luisa/luisa-dic.zip.

[7] https://www.fing.edu.uy/mh/luisa/

the participants by confronting them with these materials' reality and allowing them to collaborate practically with the search for truth and justice.

Since its launch in May 2019, LUISA has transcribed more than 400,000 blocks. On average, the platform has 5,800 accesses per day. We estimate that over 4,000 documents have been transcribed so far by more than 10,000 collaborators. Along with the original document and the OCR generated transcription, these transcriptions constitute a third type of document that enriches the database. LUISA is also useful to collect data to develop a tailored OCR system or improve the quality of the OCR transcriptions. The texts generated reconstructing the original pages using LUISA annotated blocks, constitute a ground-truth set essential for applying natural language processing (NLP) techniques, which are discussed next.

3.3 Improving the quality of automatic transcriptions

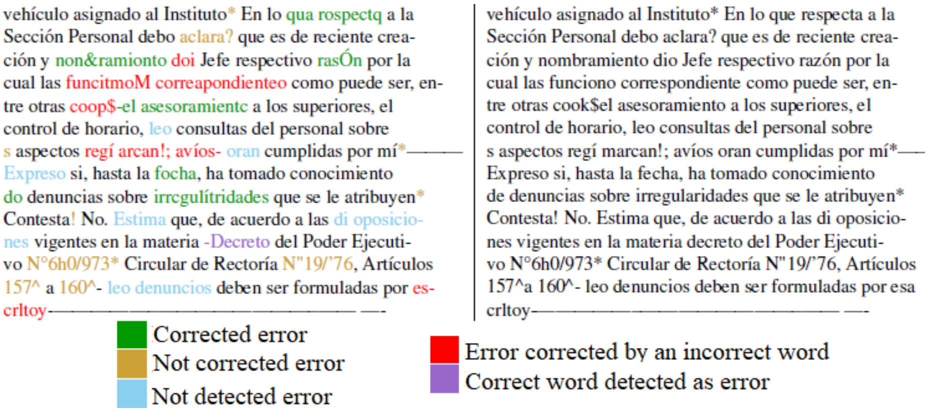
To improve the OCR output quality, we pursued two objectives: a) to correctly assemble the manual transcriptions from LUISA to generate a ground truth version of OCR documents and b) improve the outputs obtained by the OCR using advanced correction techniques.

Quality metric We use the output of the *SequenceMatcher.ratio* function of difflib, [8] which produces scores in the interval $[0, 1]$. Given a reference R text, the quality of a target text T is computed as follows. First, the *longest common subsequence* (LCS) of contiguous characters between both strings is identified. This sequence is extracted from both strings, leaving a pair of left sub-strings and a pair of right sub-strings. Then, recursively perform the same procedure between both left sub-strings and both right sub-strings. E.g., if R="la casa de su madre" and T="la caso ce eu nadre" (both of $|T| = |R| = 19$ characters), their LCS is C_1: "la cas". The remaining left sub-strings are empty and the right sub-strings are "a de su madre" and ="o ce eu nadre". The new LCS is "adre" and the process continues until no new LCS can be found. The quality is given as $(2 * sum_i |C_i| / (|R| + |T|))$, that is, the ratio of the sum of the double of the lengths of all LCSs to the sum of the lengths of the two sequences R and T. In the preceding example, this value would be $(6{+}4{+}2{+}1{+}2)/19 = 15/19 = 0.789$. For several texts, we report on the average of this metric.

Reconstruction This stage takes the block transcriptions from LUISA and rebuilds complete texts using each word's coordinates within the original document. Issues such as identifying and joining erroneously separated blocks, removing blocks belonging to stamps, and identifying line jumps between blocks are taken care of in this process to obtain a faithful transcription. A preliminary study of this procedure based on a set of 15 manually transcribed documents yielded an average score of 0.87 for the *ratio* function explained above, which is a good result.

[8] https://docs.python.org/3/library/difflib.html.

Correction of OCR output Two different approaches were used for correcting the OCR transcriptions: language models [15,3], and Statistical Machine Translation (SMT) [6]. The first case uses a Spanish language model and a dictionary of valid words. The language model chooses a set of the words which are most likely to occur within a given context, and these are further refined as those in the dictionary having a small Levenshtein [18] distance to the one being replaced. Different language models were evaluated, including various n-gram models and ELMO [21]. The best results were ultimately obtained using [19]. Figure 6 shows an example of the n-gram approach. Some incorrect words are detected and successfully corrected while others are erroneously replaced by another incorrect word. In other cases, correct words are considered errors because they do not belong to the dictionary and are therefore replaced. The OCR replaces some words by others belonging to the dictionary, so the algorithm considers them correct. Finally, some incorrect words remain undetected due to other reasons (e.g., strange symbols). We use Moses [17] to implement the SMT approach. The OCR and LUISA outputs were aligned to obtain a data set suitable for training and testing, resulting in a set of 17460 examples. Figure 7 shows an example of this approach. One advantage of SMT with respect to language models is the ability to correct erroneous words that are erroneous even if listed in the dictionary. On the other hand, STM can introduce words that are not present in the OCR output nor in the source document. The average value of the ratio mentioned above was computed on a test set of 544 documents. The result, when comparing the unaltered OCR outputs to the ground truth was 0.543. After the n-gram correction, this value was slightly improved to 0.550. However, the SMT correction resulted in a sensibly low value of 0.514.

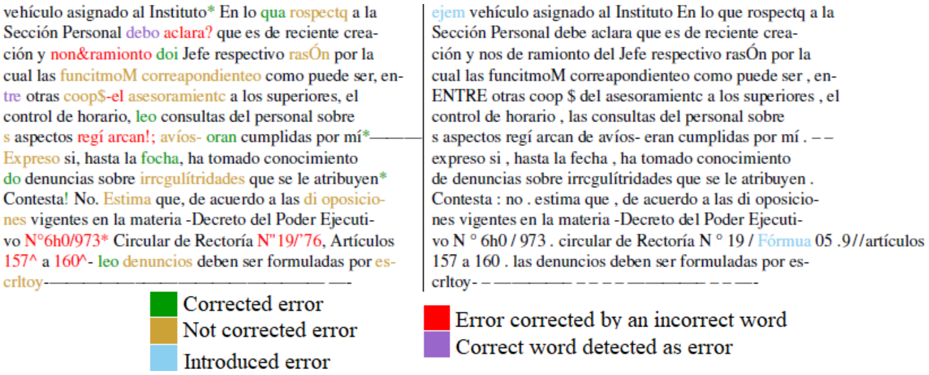


vehículo asignado al Instituto* En lo qua rospectq a la
Sección Personal debo aclara? que es de reciente crea-
ción y non&ramionto doi Jefe respectivo rasÓn por la
cual las funcitmoM correapondienteo como puede ser, en-
tre otras coop$-el asesoramientc a los superiores, el
control de horario, leo consultas del personal sobre
s aspectos regí arcan!; avíos- oran cumplidas por mí*———
Expreso si, hasta la focha, ha tomado conocimiento
do denuncias sobre irrcgulítridades que se le atribuyen*
Contesta! No. Estima que, de acuerdo a las di oposicio-
nes vigentes en la materia -Decreto del Poder Ejecuti-
vo N°6h0/973* Circular de Rectoría N"19/'76, Artículos
157^ a 160^- leo denuncios deben ser formuladas por es-
crltoy——————————————— —-

vehículo asignado al Instituto* En lo que respecta a la
Sección Personal debo aclara? que es de reciente crea-
ción y nombramiento dio Jefe respectivo razón por la
cual las funciono correspondiente como puede ser, en-
tre otras cook$el asesoramiento a los superiores, el
control de horario, leo consultas del personal sobre
s aspectos regí marcan!; avíos oran cumplidas por mí*——
Expreso si, hasta la fecha, ha tomado conocimiento
de denuncias sobre irregularidades que se le atribuyen*
Contesta! No. Estima que, de acuerdo a las di oposicio-
nes vigentes en la materia decreto del Poder Ejecuti-
vo N°6h0/973* Circular de Rectoría N"19/'76, Artículos
157^a 160^- leo denuncios deben ser formuladas por esa
crltoy——————————————— —-

🟩 Corrected error
🟫 Not corrected error
🟦 Not detected error
🟥 Error corrected by an incorrect word
🟪 Correct word detected as error

Fig. 6. OCR output (left) and its correction as produced by the n-gram model (right).



vehículo asignado al Instituto* En lo qua rospectq a la
Sección Personal debo aclara? que es de reciente crea-
ción y non&ramionto doi Jefe respectivo rasÓn por la
cual las funcitmoM correapondienteo como puede ser, en-
tre otras coop$-el asesoramientc a los superiores, el
control de horario, leo consultas del personal sobre
s aspectos regí arcan!; avíos- oran cumplidas por mí*———
Expreso si, hasta la focha, ha tomado conocimiento
do denuncias sobre irrcgulítridades que se le atribuyen*
Contesta! No. Estima que, de acuerdo a las di oposicio-
nes vigentes en la materia -Decreto del Poder Ejecuti-
vo N°6h0/973* Circular de Rectoría N"19/'76, Artículos
157^ a 160^- leo denuncios deben ser formuladas por es-
crltoy——————————— —-

ejem vehículo asignado al Instituto En lo que rospectq a la
Sección Personal debe aclara que es de reciente crea-
ción y nos de ramionto del Jefe respectivo rasÓn por la
cual las funcitmoM correapondienteo como puede ser , en-
ENTRE otras coop $ del asesoramientc a los superiores , el
control de horario , las consultas del personal sobre
s aspectos regí arcan de avíos- eran cumplidas por mí . – –
expreso si , hasta la fecha , ha tomado conocimiento
de denuncias sobre irrcgulítridades que se le atribuyen .
Contesta : no . estima que , de acuerdo a las di oposicio-
nes vigentes en la materia -Decreto del Poder Ejecuti-
vo N ° 6h0 / 973 . circular de Rectoría N ° 19 / Fórmua 05 .9//artículos
157 a 160 . las denuncios deben ser formuladas por es-
crltoy- – ———— – – – – ———— – – —-

🟩 Corrected error
🟨 Not corrected error
🟦 Introduced error
🟥 Error corrected by an incorrect word
🟪 Correct word detected as error

Fig. 7. OCR output (left) and corresponding SMT-based correction (right).

3.4 LUZ: an integrated Knowledge Base

The aim of the LUZ sub-project[9] aims to build a Knowledge Base (KB) that provides an integrated view of the information extracted from the transcriptions. This KB is designed following Linked Data principles, using Semantic Web standards such as RDF and the Web Ontology Language (OWL) [10]. The KB construction is a two-stage process: i) information extraction (IE), and ii) integration of the extracted information. Ontologies play a crucial role in both stages.

Information Extraction (IE The IE process comprises three sub-stages: Named Entity Recognition (NER), Co-reference resolution, and Relation Extraction (RE). Our strategy combines traditional NLP tools with the use of ontologies, following a similar approach to the one proposed in [10]. In the NER stage, persons, places, dates, organizations, and facts are recognized. We are currently evaluating various NER modules from NLP tools like spaCy[11], Stanford [11], Freeling[20], Python Natural Language Toolkit (NLTK)[12]. The output of the NER task is an RDF graph per document, based on the NLP Interchange Format (NIF) Core ontology[13]. This graph represents structural elements of the text, such as words and sentences. The annotation of these elements uses concepts from our Content Ontology (CO). This ontology defines the domain's key concepts (such as Persons, Organizations, Places, and Events) and relevant relations between them, such as Person pe *is in* Place pl *at a certain* Datetime dt, or Person pe *is member of* Organization org *during* Interval i1. Our CO is strongly

[9] Named after the Uruguayan human rights activist Luz Ibarburu. Also, luz means *light* in Spanish.

[10] Web Ontology Language https://www.w3.org/OWL/

[11] https://spacy.io/

[12] https://www.nltk.org/

[13] NIF Core Ontology https://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html

based on The Simple Event Model Ontology (SEM) [27]. Figure 8 depicts the main classes and properties in our CO. To deal with co-reference between named entities we developed a co-reference resolution module; the output of this module is also represented in terms of the NIF Core Ontology. We are currently working on a Relation Extraction module based on NLTK Relextract [14]. This module annotates existing graphs with relations from our CO. These modules have been integrated into INCEpTION [16], a semantic annotation platform that offers a collaborative environment. This tool is particularly relevant during the curation of the results from the extraction process, as different actors may interact with the output. Our approach enables to control the semantic heterogeneity that may arise during the extraction, while allowing the evolution of the underlying vocabularies. As a byproduct, a new corpus is obtained for further improving the performance of our IE modules.

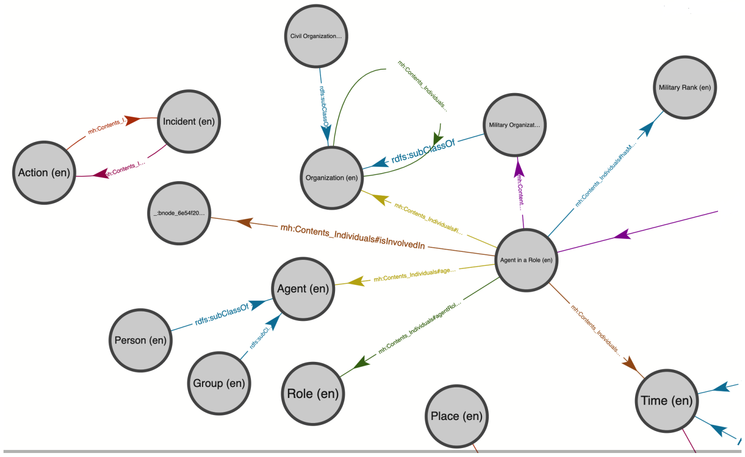


Fig. 8. Main classes and properties in our Content Ontology

We now illustrate the Information Extraction process with an example. Let us consider the following sentences taken from testimonies collected during trials [15]

- On October 5th 1976 flight 511 from Transporte Aéreo Militar Uruguayo (TAMU) flew from Buenos Aires to Montevideo, illegally transporting 22 uruguayan citizens kidnapped in Buenos Aires at"Orletti" detention center.
- Mayor Walter Pintos was the pilot of flight 511 and Soldier Ernesto Soca collaborated in Buenos Aires, at "Orletti" detention center.
- Ernesto "Drácula" Soca tortured prisoners at"Automotores Orletti", in Buenos Aires.

[14] https://www.nltk.org/_modules/nltk/sem/relextract.html

[15] Ernesto Soca Prado's judicial file and sentence available at https://sitiosdememoria.uy/index.php/causas/1242

Listing 1.1. Sample RDF triples produced by the IE process

```
:PINTOS_Walter rdf:type :Person.
:SOCA_Ernesto rdf:type :Person.
:SOCA_Ernesto_Dracula rdf:type :Person.
:MONTEVIDEO rdf:type :Place
:ORLETTI rdf:type :Place
:ORLETTI_Automotores rdf:type :Place

:AR1 rdf:type :AgentInRole ;
    :agentRoleHasAgent :PINTOS_Walter ;
    :isMemberOf :Transporte_Aereo_Militar_Uruguayo ;
    :agentRoleHasRole :Pilot .
:AR2 rdf:type :AgentInRol ;
    :agentRoleHasAgent :SOCA_Ernesto;
    :agentRoleHasRole :Collaborator .
:Inst1 rdf:type :Instant ;
    :instantDate "1976-10-05T00:00:00"^^xsd:dateTime .
:Action1 rdf:type :Actions ;
    :actionName "Flight 511 TAMU"
    :hasAgent :AR1,:AR2;
    :hasPlace :ORLETTI;
    :hasPlace :MONTEVIDEO;
    :hasTime :Inst1 .
```

A subset of the triples extracted from these sentences is depicted in Listing 1.1 using the Turtle format.

Integration and Entity Matching The use of the terms of our CO in all the graphs produced by our Information Extraction process enforces a shared schema, which reduces the integration process to Semantic Reconciliation, a.k.a. Semantic Matching, Entity Matching, and Entity Resolution. In the example, the NER process tags *Mayor Walter Pintos*, *Soldier Ernesto Soca*, and *Ernesto "Dracula" Soca* as instances of class *Person*, while *Automotores Orletti* is tagged as a *Place*. The RE process extracts the relations between these concepts and uses predicates from the ontology to represent them. Then, semantic reconciliation allows us, e.g., to identify that *Soldier Ernesto Soca* and *Ernesto "Dracula" Soca* are the same person.

The problem of Entity Resolution in the Semantic Web has received much attention in the literature [5,9]. We are currently exploring different approaches to identify corresponding entities. Using semantic web rules engines based on SWRL [13], employing triplestores with support of entailment regimes and inference forms like Stardog [16] or Virtuoso[17], or inserting owl:sameAs triples in another graph using query expressions that implement our equality conditions, are under consideration.

[16] https://www.stardog.com/

[17] http://vos.openlinksw.com/owiki/wiki/VOS

4 Conclusion and Future Work

This ongoing project has created an environment that will facilitate the analysis of millions of documents and therefore contribute to the clarification of different issues of the last dictatorship period and the advancement of the justice in Uruguay. It is essential to use all the existent technological capabilities to advance further and help the judges, journalists, historians, and other specialists do their job and contribute to the development of citizen consciousness.

Several work lines are open, among them: to use the data collected by LUISA (blocks of images with their human transcripts) to build a tailored OCR; an automatic method to detect stamps and signatures; the use of LabelMe annotations to train an automatic classification algorithm. The integration of these utilities in a system capable of managing a vast amount of documents and an adequate human interface is crucial to facilitate the use by interdisciplinary teams devoted to analyzing the data. Some of this work is being carried on today but is in its early stages. Finally, all the methods and software developed for this project are available to other communities dealing with similar problems searching for truth and justice.

References

1. Amancio, N.L.: Proyecto memoria: Datos contra el olvido (2017), https://memoria.ojo-publico.com/
2. Los archivos del terror, contra la desmemoria y la cultura autoritaria de hoy (2018), https://www.nodalcultura.am/2018/12/archivos-del-terror/
3. Bengio, Y., Schwenk, H., Senecal, J., Morin, F., Gauvain, J.: Neural Probabilistic Language Models, vol. 194, pp. 137–189. Springer Berlin Heidelberg (2006)
4. Blixen, S.: Propuesta de Proyecto: Sistematización, tratamiento y difusión de la información digital vinculada con las investigaciones en materia de graves violaciones a los derechos humanos en el pasado reciente y terrorismo de Estado. (2017)
5. Böhm, C., De Melo, G., Naumann, F., Weikum, G.: Linda: distributed web-of-data-scale entity matching. In: Proc. of the 21st ACM Int. Conf. on Information and knowledge management. pp. 2104–2108 (2012)
6. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics **19**(2), 263–311 (1993)
7. Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., Schnapp, J.: Digital_Humanities. Mit Press (2012)
8. Campos, R., Dias, G., Jorge, A.M., Jatowt, A.: Survey of temporal information retrieval and related applications. ACM Computing Surveys **47**(2), 1–41 (2014)
9. Christophides, V., Efthymiou, V., Stefanidis, K.: Entity resolution in the web of data. Synthesis Lectures on the Semantic Web **5**(3), 1–122 (2015)
10. Erekhinskaya, T., Tatu, M., Balakrishna, M., Patel, S., Strebkov, D., Moldovan, D.: Ten ways of leveraging ontologies for rapid natural language processing customization for multiple use cases in disjoint domains. Open Journal of Semantic Web **7**(1), 33–51 (2020)
11. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proc. of the 43rd Annual Meeting of ACL. p. 363–370. USA (2005)

12. Goldschmidt, M.: El equipo argentino de arqueología forense, orgullo nacional (2015), https://www.revistacitrica.com/el-eaaf-orgullo-nacional.html
13. Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B., Dean, M.: SWRL: A Semantic Web Rule Language Combining OWL and RuleML. Tech. rep., W3C (2004), http://www.w3.org/Submission/SWRL
14. ICIJ: The Panama Papers: Exposing the Rogue Offshore Finance Industry (2016), https://www.icij.org/investigations/panama-papers/
15. Jurafsky, D., Martin, J.H.: Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2nd Edition, chap. 4: N-grams. Prentice Hall (2008)
16. Klie, J.C., Bugert, M., Boullosa, B., de Castilho, R.E., Gurevych, I.: The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In: Proc. of the 27th Int. Conf. on Computational Linguistics: System Demonstrations. pp. 5–9. Association for Computational Linguistics (June 2018)
17. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Proc. of the 45th meeting of the ACL. pp. 177–180. Prague, Czech Republic (Jun 2007)
18. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady **10**(8), 707–710 (1966)
19. Lin, Y., Michel, J.B., Aiden Lieberman, E., Orwant, J., Brockman, W., Petrov, S.: Syntactic annotations for the Google Books NGram corpus. In: Proc. of the ACL 2012 System Demonstrations. pp. 169–174. Association for Computational Linguistics (Jul 2012)
20. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: Proc. of the Language Resources and Evaluation Conf. ELRA (May 2012)
21. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proc. of the 2018 Conf. of the North American Chapter of the ACL. pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018)
22. Rico, A. (ed.): Investigación Histórica sobre Detenidos Desaparecidos - Tomo I, Investigación Histórica sobre Detenidos Desaparecidos, vol. 1. IMPO, Montevideo, Uruguay, 1 edn. (2007)
23. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: LabelMe: a database and web-based tool for image annotation. Int. journal of Computer Vision **77**(1-3), 157–173 (2008)
24. Singh, J., Nejdl, W., Anand, A.: Expedition: a time-aware exploratory search system designed for scholars. In: Proc. of the 39th Int. ACM SIGIR Conf. pp. 1105–1108 (2016)
25. Singh, J., Nejdl, W., Anand, A.: History by diversity: Helping historians search news archives. In: Proc. of the 2016 ACM on Conf. on Human Information Interaction and Retrieval. pp. 183–192 (2016)
26. of Texas, U.: Del silencio a la memoria: Revelaciones del archivo histórico de la policía nacional
27. van Hage, W.R., Malaisé, V., Segers, R., Hollink, L., Schreiber, G.: Design and use of the simple event model (sem). Journal of Web Semantics **9**(2), 128 – 136 (2011)
28. W3C: Resource Description Framework (RDF 1.1) (2014), https://www.w3.org/2001/sw/wiki/RDF
29. Wada, K.: labelme: Image Polygonal Annotation with Python. https://github.com/wkentaro/labelme (2016)