Transitional Justice Review

Volume 1 | Issue 4 Article 8

March 2016

Notes from the Field: The Role of Datasets in Transitional Justice Research: The Case of Brazilian Truth Commission

Glenda Mezarobba
CEDEC, glendamezarobba@gmail.com

Roberto M. Cesar Jr.
University of Sa˜o Paulo, rmcesar@usp.br

Notes from the Field: The Role of Datasets in Transitional Justice Research: The Case of Brazilian Truth Commission

Glenda Mezarobba, CEDEC
glendamezarobba@gmail.com

Roberto M. Cesar Jr., University of São Paulo
rmcesar@usp.br

Abstract

In 2012, Brazilian President Dilma Roussef installed the Brazilian Truth Commission (CNV) to address gross human rights violations that occurred from 1946-1988. One of the most important sources of information available regarding this period is the files of the agencies that comprised the Brazilian intelligence system during the dictatorship. In total, there were around 12 million pages of relevant text in the National Archives. To make effective use of this trove of information, the CNV was challenged to use some data science tools to look for useful information within this huge dataset. As a result, a prototype of a data repository with selected documents (pdfs, images, etc.) has been created, which we summarize in this note. Computational tools for searching, organizing, and visualizing potentially important documents were developed and utilized to support CNV researchers. We also reflect upon the issues that complicated the CNV's ability to gain access to reliable and comprehensive data and the limitations of analysis conducted with this type of research.

Introduction

The Brazilian Truth Commission (*Comissão Nacional da Verdade*, CNV) was established in May 2012 by President Dilma Rousseff (Workers Party, 2011- present) to examine and clarify the gross human rights violations committed from 1946-1988. It was officially concluded in December 2014. In Brazil's case, this has meant shedding light on cases of torture, killings, enforced disappearances, and concealment of corpses (even if abroad), mainly during the military dictatorship (1964-1985). The mandate of the CNV also included identifying the structures, places, institutions, and circumstances related to the practice of these human rights violations and their possible ramifications; forwarding to the relevant authorities any information that could assist them in locating and identifying the victims; recommending public policies and measures to prevent further human rights violations; and ensuring they do not recur.[1] Similar to other truth commissions, the CNV faced important methodological and operational challenges. In order to improve the success of its research and to achieve elucidative responses, the CNV was challenged to use some data science tools to look for useful information among the large set of documents available. The CNV created a prototype data repository with selected documents (e.g. PDFs and images). Computational tools for searching, organizing, and visualizing potentially important documents were developed and applied to support a group of CNV researchers.

The aims of this note from the field are to summarize how this process occurred and how the developed system works, so that other initiatives may benefit from the gained experience. The note is organized into four sections. The first briefly reconstructs the main initiatives regarding the truth effort since the end of the dictatorship. In the second and third sections, which focus on the CNV's data summarization system, we consider the different aspects that one must take into account when designing a similar knowledge discovery system. Sample results are presented and discussed to illustrate the system's capabilities. The conclusion reflects upon the relation between information management decisions and the quality of the commission's product.

[1] Law 12,528/2011, available from http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2011/Lei/L12528.htm.

Background on the Brazilian Case

To accomplish its objectives, the CNV organized public hearings; heard the accounts of victims, witnesses and people accused of involvement in gross human rights violations; requested and gathered information, data and documents including from government bodies and agencies, regardless of their degree of secrecy; conducted research and investigations; and explored several archives, some of which were outside of Brazil. Although it was created almost 30 years after the end of the military dictatorship, the commission did not start its research activities from scratch. Since the beginning, its members have been able to rely on a very large dataset accumulated over time, in initiatives undertaken by human rights defenders or resulting from public institutions previously created to deal with the legacy of the horror.

The Brazil: Never Again project (*Brasil: Nunca Mais*, BNM), which resulted in a bestselling book published in 1985, was the first reference consulted when constructing the truth behind the gross human rights violations committed during the period in question.[2] The project began to be developed with the approval of the 1979 Amnesty Law and dealt with official documents, but without the prior knowledge of or consent from the military dictatorship. It was sponsored by the World Council of Churches, coordinated by the then Archbishop of São Paulo, Cardinal Dom Paulo Evaristo Arns, and Reverend Jaime Wright, and resulted in over one million pages being catalogued and in copies being made of almost all the political processes (707 complete and dozens of other incomplete copies) by the military justice system between April 1964 and March 1979. The book revealed the repressive system, the subversion of the law, and the different forms of torture that political prisoners suffered. Data from the project indicated that there were torture allegations in about 25% of cases. The book also revealed the names of 125 people who disappeared during the dictatorship (the practice began to be registered in Brazil in 1965), most of them in the region of the Araguaia River (*Guerrilha do Araguaia*).

[2] Dom Paulo Evaristo Arns, *Brasil: Nunca Mais* (Petrópolis: Vozes, 1985).

The universe of information provided by the BNM project documents was combined with the files of the Departments of Political and Social Order (*Departamento de Ordem Política e Social* – DOPS) of Rio de Janeiro and of São Paulo, which were controlled by the Federal Police and returned to their states of origin by presidential determination in the early 1990s. In São Paulo alone, there were 34 tons of paper, including over 150,000 records on Brazilian and foreign citizens. In the states of Pernambuco and Paraná, the files were opened on the recommendation of their respective governors, which enabled the investigation of cases such as the deaths of former state Representative Paulo Stuart Wright and worker Virgilio Gomes da Silva, which were then recognized. Their names were among 17 records of politically disappeared people, found in a drawer under the label "deceased."[3]

In 2005, during Lula`s first term as president, the Brazilian government decided to identify and put together all relevant documents on human rights violations. Steel filing cabinets containing information stored by agents of the former National Intelligence Service (*Serviço Nacional de Inteligência*, SNI), the National Security Council (*Conselho Nacional de Segurança*, CSN) and the General Commission of Investigations (*Comissão Geral de Investigações*, CGI) were sent to the National Archive (*Arquivo Nacional*) in Brasília.[4] Inside them were photographs, posters, movies, books, pamphlets, magazines, 220,000 microfiche files, and 1,259 archive boxes. Only information about the honor, image, intimacy, and privacy of citizens was kept inviolate, as required by Brazilian law. Shortly after, the Foreign Ministry and the Federal Police also forwarded thousands of secret documents produced between 1964 and 1975. To improve the Brazilian process of settling accounts, in May 2009, the federal government launched the Revealed Memories (*Memórias Reveladas*) web portal.[5]

[3] See Glenda Mezarobba*, Um acerto de contas com o futuro: a anistia e suas conseqüências – um*
estudo do caso brasileiro (São Paulo: Humanitas/Fapesp, 2006).

[4] See *Arquivo Nacional*, available from http://www.arquivonacional.gov.br.

[5] See *Memórias Reveladas*, available from http://www.memoriasreveladas.gov.br.

Established with the aim of making information on the recent political history of the country available, it groups all the documentation in a national network managed by the National Archive itself. The content and whereabouts of the period's main archive—that of the Armed Forces—remain unknown. Although (important) revelations have been made from time to time by the press, successive commanders of the Armed Forces have reiterated that the archives no longer exist. This does not mean, however, that there is no awareness of the reports produced by the military. Given the circularity of the system, many documents produced by the Army, Navy, and Air Force have been located in files of other government agencies.

In addition to the information provided by the files already opened, part of the truth about the violent practices of the period also surfaced through the work of two commissions set up in Brasília. From 1995, with the establishment of the Special Commission on the Dead and Disappeared for Political Reasons (*Comissão Especial sobre Mortos e Desaparecidos Políticos*, CEMDP),[6] for example, many fanciful versions of the past promulgated by the dictatorship were disproven and new facts revealed. All this was compiled and published in *The Right to Memory and Truth* (*Direito à Memória e à Verdade*),[7] a 2007 book by the Special Secretariat for Human Rights (*Secretaria de Direitos Humanos*, SEDH). The book summarizes 11 years of the commission's activities and is the first official document to contain victims' versions of events and to attribute crimes such as torture, rape, dismemberment, decapitation, concealment of corpses, and murder of opponents of the military regime to members of the security forces. In addition, the 70,000 cases received by the Amnesty Commission (*Comissão de Anistia*)[8] since its inception in 2001 have revealed important data, much of which was supported by official documents, regarding

[6] See *Comissão Especial sobre Mortos e Desaparecidos Políticos*, available from http://cemdp.sdh.gov.br/.

[7] Brasil, *Direito à Memória e à Verdade* (Brasília: SEDH, 2007).

[8] See *Comissão de Anistia*: http://portal.mj.gov.br/anistia/.

harassment, arbitrary arrests, and torture committed by agents of the repressive state apparatus. [9]

In spite of the fact that these previous efforts had produced much information about the past, there was a perceived need for the CNV mainly because of the unsolved issue of political murders and disappearances. Thus, when the CNV began its mandate, over 12 million pages of documents on the dictatorship were available just in the National Archive. During the commission's activities, around 10 million pages were digitized. However, after six months of activity (a quarter of the mandate) CNV members were still not clear about the need to create a database or repository. In December 2012, when the commission was discussing the creation of a CNV information system to improve research on forced disappearances and deaths, a prototype of what the authors imagined would become the CNV's data repository was constituted, with all the pre-existing information and information gathered during the CNV's mandate. This CNV prototype data repository started off with information collected in the National Archive by a team of researchers. All digital information available on each case of death or forced disappearance, initially held in CDs or DVDs, was stored in the prototype repository so as to enable information and data analysis to begin. A method for searching and viewing documents potentially relevant to the aims of the research was then developed by the authors. The following section details the computational tools developed.

Data Summarization System

Three important aspects have to be taken into account while designing a knowledge discovery system like the one described in this note: (1) the original data, (2) the type of information to be extracted from the data, and (3) the way(s) researchers will interact with the system. These aspects were carefully taken into account in developing the system. The system was designed to help

[9] See Glenda Mezarobba, "Brazil," in *Encyclopedia of Transitional Justice*, edited by Lavinia Stan and Nadya Nedelsky, 67-73 (Cambridge: Cambridge University Press, 2012).

researchers search, recover, and analyze documents in an efficient way. As shown below, the huge collection of documents could not be manually analyzed, not even by a large team of researchers, hence the need for computational tools. This type of approach has been gaining increasing attention in the so-called digital humanities field.[10] Pioneering pieces in human rights, in particular, have attracted significant attention. For instance, Miller *et al.* describe a framework for making a narrative analysis of human rights violations datasets.[11] In addition, Best *et al.* discuss the importance of digital tools in the process of peace-building in countries after civil conflicts.[12]

The architecture of the data summarization system developed for the CNV is shown in Figure 1. An important aspect that should be understood is that several data sources in different formats are used. It is difficult to devise a system that can accommodate all possible data sources, data types, data quality, and other variables at the early stages of a truth commission's work. Therefore, it is important to take the aforementioned three aspects into account while designing the system. Regarding the diversified data sources and types, it was decided to create an intermediate dataset. The heterogeneous documents were integrated in a local prototype data repository. This solution is interesting because additional data sources and documents could be more easily

[10] See http://diggingintodata.org/; Gary King, "Ensuring the Data Rich Future of the Social Sciences," *Science* 331 (February 2011): 719-721.

[11] Ben Miller, Ayush Shrestha, Jason Derby, Jennifer Olive, Karthikeyan Umapathy, Fuxin Li, and Yanjun Zhao, "Digging into Human Rights Violations: Data Modelling and Collective Memory," *Proceedings of the IEEE International Conference on Big Data* (2013).

[12] Michael L. Best, William J. Long, John Etherton, and Thomas Smyth, "Rich Digital Media as a Tool in Post-Conflict Truth and Reconciliation," *Media, War & Conflict* 4.3 (2011): 231-249. Also, see Michael. L. Best, "Peacebuilding in a Networked World," *Communications of the ACM*, 56.4 (2013): 30-32; Juan Pablo Hourcade and Lisa P. Nathan, "Human Computation and Conflict," in *Handbook of Human Computation*, edited by Pietro Michelucci, 993-1009 (New York: Springer, 2013); Ben Miller, "Digital History's Relationship to Human Rights Archives and Data Analysis," *HASTAC Digital History Group Spotlight Series* (17 March 2013).

integrated into the system as the CNV's work advanced and new data discovered or developed. The developed software library takes these documents as input and can either generate more data to be stored in the repository, or produce web reports to be used by researchers.

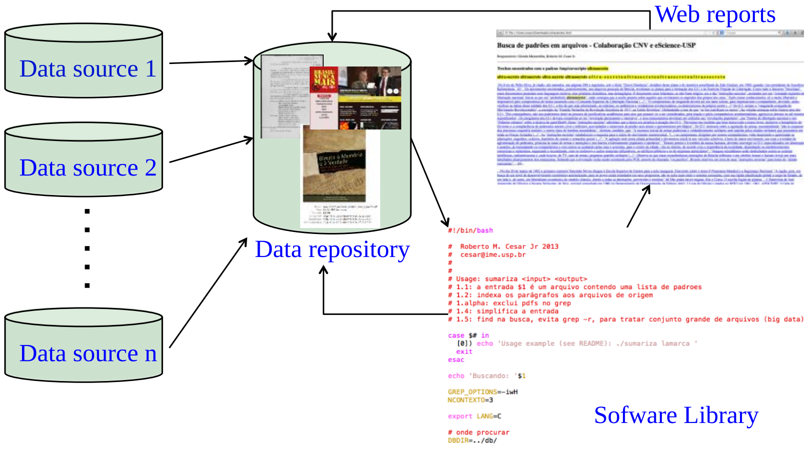
Different data types were adopted: PDF and image documents from the National Archive, published books,[13] and other documents available to the commission. For instance, while most documents from the "*Arquivo Nacional*" (a Brazilian official documents archive from the Justice Ministry) undergo a standardization procedure to be made public, the CNV obtained other types of data such as non-digital documents, images, and interviews (both text and audio). Hence, the proposed framework is designed to support multiple inputs, including some unforeseen ones. As explained below, the proposed solution to address this question is by adopting an intermediate representation in the form of text files.

Below are some of the prototype repository's features:

- Over 25,000 files[14] (each file may be composed of hundreds of documents)
- 92 Gb
- 600,000 pages

[13] Brasil, *Direito à Memória e à Verdade* (Brasília: SEDH, 2007); Licio Maciel and José Conegundes do Nascimento Orvil, *Tentativas de Tomada do Poder* (São Paulo: Editora Schoba, 2012); Elio Gaspari, *A Ditadura Envergonhada* (São Paulo: Companhia das Letras, 2002); Elio Gaspari, *A Ditadura Escancarada* (São Paulo: Companhia das Letras, 2002); Elio Gaspari, *A Ditadura Derrotada* (São Paulo: Companhia das Letras, 2003); Elio Gaspari, *A Ditadura Encurralada* (São Paulo: Companhia das Letras, 2004); Arns 1985.

[14] The current supported file formats are PDF, TIFF, PNG, JPG, DOC, XLS, and TXT. Each file stores a set of documents that have been scanned, possibly up to hundreds of pages in each file.



```
#!/bin/bash

#  Roberto M. Cesar Jr 2013
#  cesar@ime.usp.br
#
#
# Usage: sumariza <input> <output>
# 1.1: a entrada $1 é um arquivo contendo uma lista de padroes
# 1.2: indexa os parágrafos aos arquivos de origem
# 1.alpha: exclui pdfs no grep
# 1.4: simplifica a entrada
# 1.5: find na busca, evita grep -r, para tratar conjunto grande de arquivos (big data)

case $# in
  [0]) echo 'Usage example (see README): ./sumariza lamarca '
  exit
esac

echo 'Buscando: '$1

GREP_OPTIONS=-iwH
NCONTEXTO=3

export LANG=C

# onde procurar
DBDIR=../db/
```

Figure 1: Graphic illustration of the proposed framework. Data from several sources are integrated into an intermediate repository. The software library may read and write information in the repository as well as generate web reports, which were used by CNV researchers.

Most of the data in the prototype repository is composed of scanned historical documents in PDF or another image formats (e.g. JPG or PNG). The first important issue to be addressed here is that the data should be searchable by analytical tools. Because most of the data formats come as images from digitized documents, they do not allow searching by contents like keywords, numbers, or dates.

The standard procedure to solve this problem is the application of Optical Character Recognition (OCR), which analyzes the image for text portions and then annotates them. Once a PDF or image file undergoes an OCR procedure, the resulting output may be searched by a number of different computational tools. A large proportion of the PDF files had already undergone an OCR procedure from their original databases. This was true of the National Archive database, for example. This procedure embeds text annotation in the PDF files that can be exported as text files.

These text files are directly searchable by text queries. Files that had not been annotated by OCR had to be pre-processed. The developed system includes the possible application of such OCR. In this case, the open-source software Tesseract was adopted.[15]

Tesseract, an OCR program originally developed by Hewlett-Packard and currently maintained by Google, is widely used and has some interesting qualities (starting with the fact of being open-source).[16] Such qualities explain the reason why it was adopted in the proposed system. Particularly, the developed library is composed by Linux scripts (BASH and Python) running in command-line. Tesseract also has a command-line interface where all parameters may be easily specified. Standard off-the-shelf parameters were used to run Tesseract. Although Tesseract contains some Portuguese language support, its performance in Portuguese often fails when compared to English texts. This means that there is room for improvement in the system, either by including better Portuguese language support into Tesseract or by using other types of OCR software. Tesseract has other important limitations, such as the fact that it has not been designed as a general document analysis program. It is better suited to analyze text-only documents without images in a single column. However, this is not always the case for the CNV documents, which may be presented in a large variety of layouts from the likes of official reports, digitized newspapers, and handwritten folders. This is an additional limitation that should be improved for future work.

Thus, the prototype data repository is actually composed of two layers: 1) the original files (PDFs, images, etc.), and 2) ASCII text files produced by the OCR procedures. The developed software library has tools to generate and curate the ASCII text files, which are the basis for searching the data. These tools include file and folder name standardization (because the thousands of documents are from quite diverse sources), detection and elimination of duplicated documents, and statistics extraction (number of files, folders, sizes, etc.). These tools help to create,

[15] See Tesseract, available from https://code.google.com/p/tesseract-ocr/.

[16] Open-source software means that the original source code is available, which is important in many applications.

maintain, and incorporate new documents into the repository. This is particularly important because, as mentioned, the system (prototype + software library) was developed and used during part of the work of the CNV, while new documents were being continuously collected. The CNV's work continuously generated documents to be included in the repository by different members involved in different questions. The adopted solution of incorporating new files in the repository without constraints for CNV researchers was important since it did not create any obstacles on their side. Therefore, CNV researchers could concentrate on their work without having to follow some specific protocols that could deviate from their main questions (e.g., is this file already in the repository? will it be redundant?). Repository maintenance was completely transparent for CNV researchers.

The second layer of text files are used as input for the search and summarization software. This allows keyword-based queries to recover documents where such keywords appear. Since there was no time to develop an online query-based search engine (like a Google for the truth commission), a different off-line approach was devised. The researcher prepared lists of keywords and other information for each query of interest. Linux BASH and Python tools were used to search for keywords in the files, using a keyword list stored in a separate file for each topic of interest. The researcher had to specify the file with possible keywords of interest, possible variations (such as nicknames), and historical dates and any additional information that could help recover important documents. For instance, an example of a typical query file in the prototype might be composed of the following keywords:

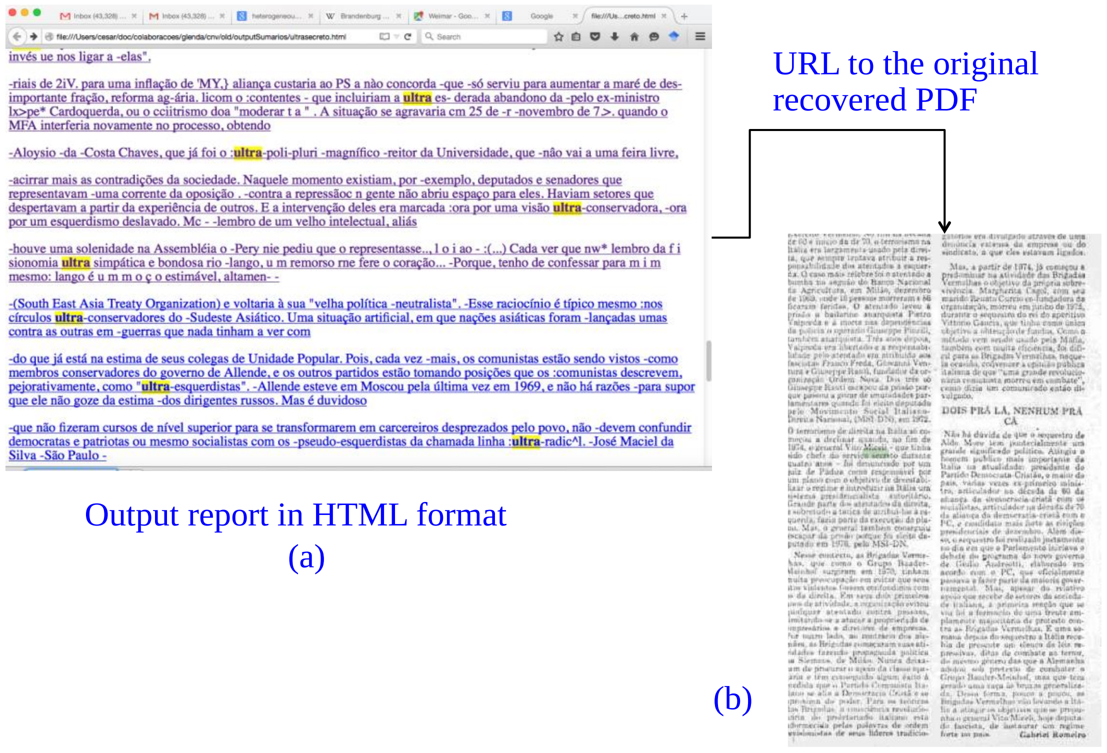
- *#Pedro Alexandrino de Oliveira Filho*
- *Pedro Alexandrino*
- *Peri*
- *04/08/1974*
- *Gameleira*
- *Tuca*
- *Maria Luiza Garlipe*
- *enfermeira HC-SP*

- *garrafa com sal*
- *garrucha*

The first line contains the full name of a disappeared person. It starts with a hashtag (#), which defines a comment line that is ignored by the system. This is a useful feature because the researchers may annotate important information in the query files that will not be used by the system, including metadata like the responsible researcher from the commission, and dates and sources from where the keywords were defined. For instance, in this case, although the full name is "*Pedro Alexandrino de Oliveira Filho*", it is more difficult to find the whole string in the documents (also because of the fact that most OCR text files are quite noisy). However, "*Pedro Alexandrino*" is a less common name in Brazil and much easier to find among the documents, since it is a substring of the full name. This distinguishing feature helps to potentially recover documents of greater interest. This is the second line in the file. Hence, in this case the system will look for occurrences of *Pedro Alexandrino* which may appear alone or as part of the full name *Pedro Alexandrino de Oliveira Filho*. The third line contains his nickname and the fourth the date on which he is thought to have disappeared. The other keywords refer to clues such as information on his girlfriend or other terms found in documents associated with him. Each line is used as a seed for searching the repository documents. This solution was adopted because it is a straightforward and intuitive way for researchers to specify sets of queries that can be run offline, since there was no time available to develop an online system, as explained above. The tight deadline was an important constraint for the CNV, which had to be taken into account when developing the repository search solution.

All documents in which any of the keywords appear are selected and copied by the software into an output folder. An HTML file is created to summarize all the documents retrieved. A paragraph around each keyword found is extracted and copied into the HTML file in order to provide context to the CNV researcher, who is then able to easily browse all documents retrieved in a single file (see Figure 2). The extracted paragraph is a hyperlink to the respective recovered document. This solution allows the researcher to analyze dozens (possibly hundreds) of documents using any standard web browser. This solution was adopted for being straightforward to implement, relying only on command-line scripts without specific Graphical User Interfaces (GUI). The GUI for the researcher interaction itself was the web browser, thus allowing the researcher to analyze the output reports on different devices (computers, laptops, tablets, smartphones) and platforms (Windows, Linux, Mac OS, Android).

The search procedure uses the Linux *grep* command with heuristics for dealing with inexact text matching. The *grep* command is a program available in most Linux-based Operational Systems (including Mac OS) that looks for keywords in files. It is a very powerful and flexible tool widely used by programmers to create query systems like those described in this paper. *Grep* accepts various parameters that allow for the creation of useful searching strategies (i.e. *heuristics*). These heuristics are important because of the imperfect text typically provided by OCR procedures and bad image quality of many documents (see Figure 2). Many words are not fully complete and the output file is typically full of typos (e.g. “Registro E1lÍ'l'8d8·NRE Dintńbulçāu Inicial J APRECIAÇÃO SEMANAL UO CAMFO EXTERNO”, which is a real output from one file in the repository). Hence the search heuristics represent and attempt to circumvent these problems. Other open source tools like Linux BASH scripts and Python support the developed software in locating the files, moving them to the output folders and preparing the HTML output report.



file:///Users/cesar/doc/colaboracoes/glenda/cnv/old/outputSumarios/ultrasecreto.html

invés ue nos ligar a -elas",

-riais de 2iV. para uma inflação de 'MY,} aliança custaria ao PS a nào concorda -que -só serviu para aumentar a maré de des-importante fração, reforma ag-ária. licom o :contentes - que incluiriam a ultra es- derada abandono da -pelo ex-ministro lx>pe* Cardoquerda, ou o cciitrismo doa "moderar t a " . A situação se agravaria cm 25 de -r -novembro de 7>. quando o MFA interferia novamente no processo, obtendo

-Aloysio -da -Costa Chaves, que já foi o :ultra-poli-pluri -magnífico -reitor da Universidade, que -não vai a uma feira livre,

-acirrar mais as contradições da sociedade. Naquele momento existiam, por -exemplo, deputados e senadores que representavam -uma corrente da oposição . -contra a repressãoc n gente não abriu espaço para eles. Haviam setores que despertavam a partir da experiência de outros. E a intervenção deles era marcada :ora por uma visão ultra-conservadora, -ora por um esquerdismo deslavado. Mc - -lembro de um velho intelectual, aliás

-houve uma solenidade na Assembléia o -Pery nie pediu que o representasse.., l o i ao - :(...) Cada ver que nw* lembro da f i sionomia ultra simpática e bondosa rio -lango, u m remorso rne fere o coração... -Porque, tenho de confessar para m i m mesmo: lango é u m m o ç o estimável, altamen- -

-(South East Asia Treaty Organization) e voltaria à sua "velha política -neutralista". -Esse raciocínio é típico mesmo :nos círculos ultra-conservadores do -Sudeste Asiático. Uma situação artificial, em que nações asiáticas foram -lançadas umas contra as outras em -guerras que nada tinham a ver com

-do que já está na estima de seus colegas de Unidade Popular. Pois, cada vez -mais, os comunistas estão sendo vistos -como membros conservadores do governo de Allende, e os outros partidos estão tomando posições que os :comunistas descrevem, pejorativamente, como "ultra-esquerdistas". -Allende esteve em Moscou pela última vez em 1969, e não há razões -para supor que ele não goze da estima -dos dirigentes russos. Mas é duvidoso

-que não fizeram cursos de nível superior para se transformarem em carcereiros desprezados pelo povo, não -devem confundir democratas e patriotas ou mesmo socialistas com os -pseudo-esquerdistas da chamada linha :ultra-radic^l. -José Maciel da Silva -São Paulo -

DOIS PRÁ LÁ, NENHUM PRÁ CÁ

Gabriel Romeiro

Figure 2: Output example: (a) Report produced by the system. Each entry represents a passage providing the context for the keyword found, which is highlighted in yellow to help the researcher. (b) The passage is hyperlinked to the original PDF, thus helping the researcher to easily browse the relevant document.

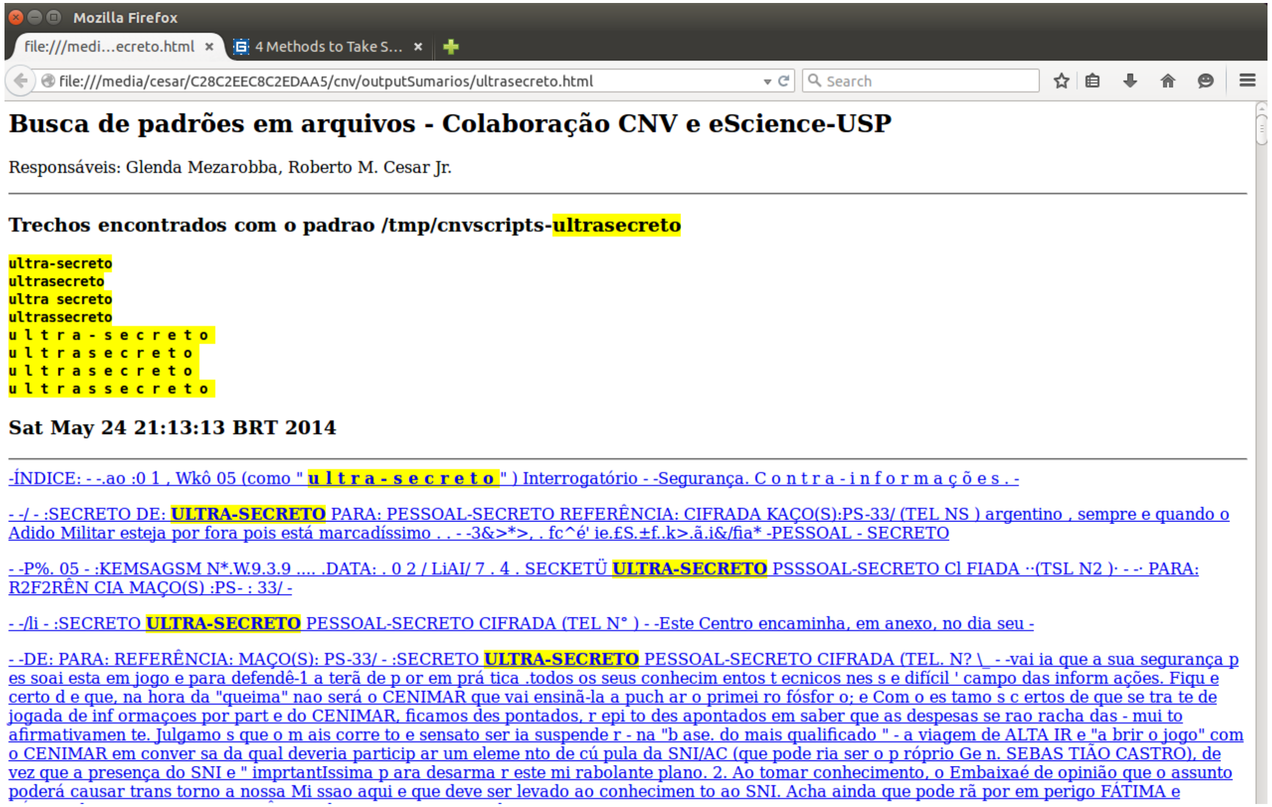
The system developed was executed on an Apple iMac Intel Core i7 quad core, 3.4GHz, 8GB SDRAM DDR3. As mentioned above, each query is composed of a set of strings representing names, nicknames, keywords, phrases, dates, and places, among other possible clues. The average search time for the system was 20 seconds/keyword to analyze the entire repository.

Sample Results

To illustrate a real system application, a query for "ultra-secreto" (top secret in Portuguese) is demonstrated below. In order to cope with the different Portuguese spellings possible, the following keywords were chosen for this query:

- ultra-secreto
- ultrasecreto
- ultra secreto
- ultrassecreto

(a)


Mozilla Firefox
file:///medi...ecreto.html ×
4 Methods to Take S... ×
file:///media/cesar/C28C2EEC8C2EDAA5/cnv/outputSumarios/ultrasecreto.html
Search

Busca de padrões em arquivos - Colaboração CNV e eScience-USP

Responsáveis: Glenda Mezarobba, Roberto M. Cesar Jr.

Trechos encontrados com o padrao /tmp/cnvscripts-ultrasecreto

```
ultra-secreto
ultrasecreto
ultra secreto
ultrassecreto
u l t r a - s e c r e t o
u l t r a s e c r e t o
u l t r a s e c r e t o
u l t r a s s e c r e t o
```

Sat May 24 21:13:13 BRT 2014

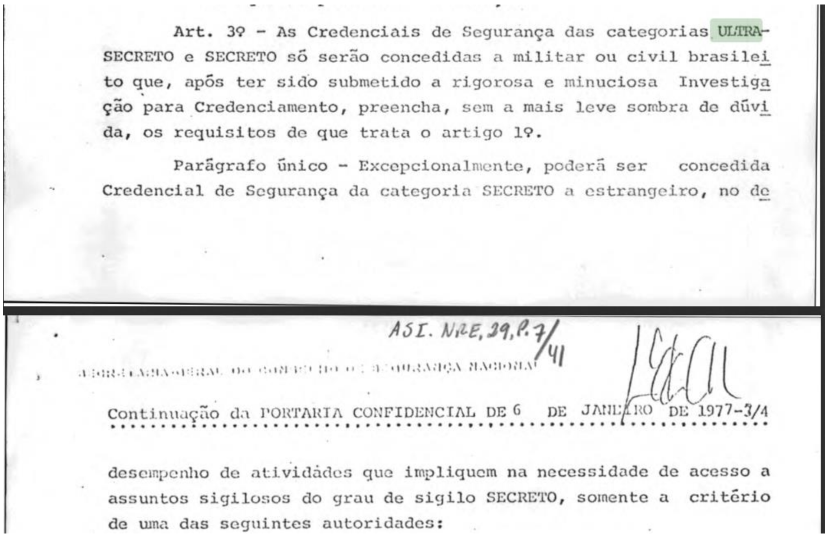
-ÍNDICE: - -.ao :0 1 , Wkô 05 (como " **u l t r a - s e c r e t o** ") Interrogatório - -Segurança. C o n t r a - i n f o r m a ç õ e s . -

- -/ - :SECRETO DE: **ULTRA-SECRETO** PARA: PESSOAL-SECRETO REFERÊNCIA: CIFRADA KAÇO(S):PS-33/ (TEL NS) argentino , sempre e quando o Adido Militar esteja por fora pois está marcadíssimo . . - -3&>*>, . fc^é' ie.£S.±f..k>.ã.i&/fia* -PESSOAL - SECRETO

- -P%. 05 - :KEMSAGSM N*.W.9.3.9 .... .DATA: . 0 2 / LiAI/ 7 . 4 . SECKETÜ **ULTRA-SECRETO** PSSSOAL-SECRETO Cl FIADA ··(TSL N2)· - -· PARA: R2F2RÊN CIA MAÇO(S) :PS- : 33/ -

- -/li - :SECRETO **ULTRA-SECRETO** PESSOAL-SECRETO CIFRADA (TEL N°) - -Este Centro encaminha, em anexo, no dia seu -

- -DE: PARA: REFERÊNCIA: MAÇO(S): PS-33/ - :SECRETO **ULTRA-SECRETO** PESSOAL-SECRETO CIFRADA (TEL. N? _ - -vai ia que a sua segurança p es soai esta em jogo e para defendê-1 a terã de p or em prá tica .todos os seus conhecim entos t ecnicos nes s e difícil ' campo das inform ações. Fiqu e certo d e que, na hora da "queima" nao será o CENIMAR que vai ensinã-la a puch ar o primei ro fósfor o; e Com o es tamo s c ertos de que se tra te de jogada de inf ormaçoes por part e do CENIMAR, ficamos des pontados, r epi to des apontados em saber que as despesas se rao racha das - mui to afirmativamen te. Julgamo s que o m ais corre to e sensato ser ia suspende r - na "b ase. do mais qualificado " - a viagem de ALTA IR e "a brir o jogo" com o CENIMAR em conver sa da qual deveria particip ar um eleme nto de cú pula da SNI/AC (que pode ria ser o p róprio Ge n. SEBAS TIÃO CASTRO), de vez que a presença do SNI e " imprtantIssima p ara desarma r este mi rabolante plano. 2. Ao tomar conhecimento, o Embaixaé de opinião que o assunto poderá causar trans torno a nossa Mi ssao aqui e que deve ser levado ao conhecimen to ao SNI. Acha ainda que pode rã por em perigo FÁTIMA e



Art. 3º - As Credenciais de Segurança das categorias ULTRA-SECRETO e SECRETO só serão concedidas a militar ou civil brasileito que, após ter sido submetido a rigorosa e minuciosa Investigação para Credenciamento, preencha, sem a mais leve sombra de dúvida, os requisitos de que trata o artigo 1º.

Parágrafo único - Excepcionalmente, poderá ser concedida Credencial de Segurança da categoria SECRETO a estrangeiro, no de

ASI. NRE, 29, P. 7/41

SECRETARIA-GERAL DO CONSELHO DE SEGURANÇA NACIONAL

Continuação da PORTARIA CONFIDENCIAL DE 6 DE JANEIRO DE 1977-3/4

desempenho de atividades que impliquem na necessidade de acesso a assuntos sigilosos do grau de sigilo SECRETO, somente a critério de uma das seguintes autoridades:

(b)

Figure 3: (a) Example of an HTML report generated by the system.
(b) Example of a digitized document retrieved by the system, in which the query keyword is highlighted.

Figure 3(a) shows a portion of the HTML report generated. This query is interesting because documents, descriptions, folders and other texts could be retrieved. For instance, Figure 3(b) shows a context paragraph containing an instance of "ultra-secreto." The output produced by the system consists of the HTML report together with copies of all retrieved documents. This solution was adopted because of being self-contained, i.e. all information is stored in a single folder. This makes it easy for commission researchers to use since it is enough to compress this folder and to transfer it by email, upload, copying it in a pen drive or making it available for download online. Since the commission researchers were typically working in different cities (eventually countries), such a strategy becomes quite useful. The commission researcher could then browse and read the output HTML report looking for useful information. For each entry of potential interest, it is enough to click on the hyperlink and the browser is able to open the original PDF document where one of the keywords occurs. This procedure allows the researcher to analyze dozens of documents in minutes.

A current limitation of the system is that it is able to automatically link the retrieved document, but not the inner pages where the searched keywords appear. Hence, the researcher has to search the document once it is opened by the browser (e.g. by edit->search). In the future, it would be important to improve the system to allow this direct indexation of the inner portions. However, it is important to emphasize that even this simpler indexation scheme is very useful, allowing researchers to quickly browse large sets of documents, which could not be done otherwise. This is an important aspect that may help groups analyzing large sets of documents: the proposed framework represents a straightforward solution to deal with such sets of documents, with very good potential to help researchers.

Conclusion

This note from the field has described a text search and summarization system developed to support the document analysis of researchers working on the Brazilian Truth Commission. The system developed was implemented using off-the-shelf open source software and was able to search a prototype data repository of more than 600,000 PDF document pages. The system is composed of a folder of document directories and a library of scripts that are able to manage and search the repository, generating HTML reports with links to retrieved documents. The system can be improved by extracting document pages in which keywords appear (for the moment, the whole PDF file is indexed by the HTML summary for each query). Also, it is important to include more powerful OCR procedures in future research, either with other types of software or by fine tuning Tesseract with customized training sets. There is also room for improvement in the text matching, which is currently done by exact matching.

Inexact text matching distances[17] (e.g., the Levenshtein distance) may lead to better document retrieval, as there are misspellings in the original documents produced by the OCR step (See Figure 3a for an example).

In spite of the speed in which the tool was developed and its great potential to render the search for information on the political dead and disappeared of the Brazilian military dictatorship more effective and, therefore, possibly more successful, its use was not widely disseminated by the Commission. In fact, the data repository was never structured, staying in its prototype form, which constituted the main obstacle to use the text search and summarization system. As there were no budget constraints, resistance to building a data repository and to use the mentioned system appears to demonstrate, at least, a lack of knowledge of its possibilities. Because of the breadth of its work and the nature of its responsibilities, a truth commission requires a wide range of expertise. That includes computer and information-systems specialists, but the CNV could not understand that. The Brazilian case is not the only one; as Priscilla Hayner points out, "[f]ew commissioners have had prior experience in data management and analysis, and they may misgauge the task."[18]

This leads us to reflect on the insufficiency of access to and geographical distribution of archives of this type for elucidating cases of grave human rights violations. In the case of the CNV, a majority of the documents were ignored not because they do not exist or because the commission did not have access to them. Rather, they were not considered for the opposite reason: the huge amount of information available is humanly impossible to read. If a computer database is mandatory to record and analyze all the information, the preceding detailed work of coding and entering information into a database is also very important. In the Brazilian case, there was also a strong resistance from the staff in using a

[17] William Cohen, Pradeep Ravikumar, and Stephen Fienberg, "A Comparison of String Metrics for Matching Names and Records," *Kdd Workshop On Data Cleaning And Object Consolidation*, 3 (2003): 73-78.

[18] Priscilla Hayner, *Unspeakable Truths: Transitional Justice and the Challenge of Truth Commissions*, (New York: Routledge, 2nd Ed., 2011), 221.

standardized questionnaire for victims, survivors, and witnesses. There is no doubt that the absence of these tools shaped the results of the commission. Although the CNV had a huge information collection, it did not have an adequate data entry and processing system in order to proceed with an adequate analysis. This certainly influenced the quality of the commission's products, especially the Final Report.

Either way, in the Brazilian case, the documents are now available at the National Archive, where all the data gathered during the CNV's activities were sent, according to the law that created the Commission. So, it seems that the search for truth regarding the crimes of the dictatorship is just beginning.