Introduction

O Centro de Pesquisa e Documentação de História Contemporânea do Brasil da Fundação Getulio Vargas[[1]](#footnote-1) (CPDOC/FGV) is the leading historical research institute in the country and holds a major collection of personal archives, oral histories and audiovisual sources pertaining to Brazilian contemporary history. Even though belonging to the same universe – the Brazilian contemporary history, different information systems were built to store the varied historical data held in CPDOC. The problem detected recently is that the current way this data is digitally represented and stored makes it very difficult and complex to maintain them, and even more, does not ensure that the access of data is performed in a centralized, intuitive and efficient way.

This paper presents a project that aims to change the way CPDOC has been dealing with their data until now, and the ongoing initiatives are pointing to the creation of a new model that ensures interoperability, sharing and reuse of information. The project is based on the concept of using open/connected data instead of relational data, i.e. the Linked Open Data, reinforced by the three directives promoted by W3C: 1) if data cannot be found and indexed in the web, it does not exist; 2) if data is not free and available in a machine comprehensive format, it cannot be reused; 3) if any legal issue does not allow the data reuse, it is not useful.

We believe that the renovation sought by CPDOC must necessarily include a strategic reorientation concerning the effort for handling their data. The current scenario reveals some precariousness of the availability and access of data, limited by what the present information systems allow the final user to do, which is not much. On the other hand, if the focus is oriented to production and distribution instead of query systems, it is possible to allow the use of CPDOC data by the community, and all different ways of dealing with data proposed by the community, independently of CPDOC interventions. This dynamic and sustainable environment helps to keep data up-to-date and massively available.

This idea follows the directives of free information promoted by Semantic Web, where a global network of interconnected data would allow a great improvement of knowledge by connecting alike information. This is aligned with the growing initiative of opening the access of data from private, governmental and other organizations.

Among this project efforts, we highlight the creation of a triple store using RDF from the data originally stored in relational databases and the application of semantic technologies capable of transforming the CPDOC collection in a source of free and interoperable data following the Linked Open Data model.

This paper intends to reinforce CPDOC role as content provider and host of a platform for collaborative action in other communities. In this way, we aim to emerge from a research experience to a concrete and productive effort in a center of reference in the field of documentation and historical preservation. (um pouco confuso isso aqui..)

Information systems in CPDOC

The collection of CPDOC is organized in three different systems stored in the same database. Each of them has an independent management and adopts idiosyncratic criteria concerning the organization and indexing of their information, which vary depending on the specifications of the content they host: personal archives, oral history interviews and entries of the Brazilian Historical-Biographic Dictionary. Each of them is briefly described in the following subsections.

## Personal Archives

Historical documents, in the most varied textual and audiovisual supports, are precious sources of knowledge that help us to know deeper our history. They can be manuscripts, diaries, letters, photography, speeches or memos and whenever they compose personal files of people that influenced the political and social scenario of our country, they represent much more than private memories: they are concrete registry of a collective memory.

Currently, more than 200 personal archives that belonged to presidents, ministers, military personal and other important people to the Brazilian society compose the CPDOC collection. Together, they comprise nearly 1.8 million documents or 5.2 millions pages. From this, nearly 700 thousands pages are digitalized and the expectance is to digitalize the whole collection in the next years. The metadata of the collection is stored in a system called Acessus that allows the query of data using some search tools based basically on syntactic query structures.

## Oral History Interviews

The CPDOC collection of Oral History comprises currently more than 6.000 hours of recording, corresponding to nearly 2.000 interviews. More than 90%, video or audio, are in digital format. For the time being, two kinds of query are available in the database: query by subject and query by interviewed. Each interview record holds a brief technical record and a summary with the descriptions of the interview themes in the order they appear during the record. 10% of the interviews are transcribed, but to access the audio/video content the user is requested to come personally to CPDOC. Currently the institute is analyzing better ways of making this data available online, considering different aspects such as the best format, use policies, access control and copyrights linked to the interviews.

## Entries of the Brazilian Historical-Biographic Dictionary

The Brazilian Historical-Biographic Dictionary is certainly one of the main collections about Brazilian historical politicians and themes after 1930. It hosts more than 7.500 entries of biographic and thematic nature, that means, people, institutions, organizations and events carefully selected using criteria that measure the relevance of those to the political history for the given period. The dictionary database stores a few metadata concerning each entry, and the query is limited to the title or occurrence of words within the text. The user is required to log into the system in order to access data and the use of the available GUI is mandatory for querying. The problem in this kind of environment, associated to what Michael Bergman identified as part of Deep Web[[2]](#footnote-2), is the information that cannot be found by standard search tools, that means, websites that does not exist until being created dynamically as a result of an specific query. (?)

This is certainly one of the biggest challenges that CPDOC faces in what concerns their collection.

The question

Limited visibility and access (*in Google*); terminological problems; idiosyncratic indexing, low rates of recall and precision; independent systems;

Concerning relational databases:

apresentação do Alexandre

* Rigidez para mudanças frequentes
* Definições a priori
* Tabelas adicionais para “guardar” relações muitos-muitos
* Performance depende de decisões e manutenção de um DBA
* Poucas restrições sobre o domínio no modelo
* Ferramentas disponíveis para desenvolvimento de sistemas de informação. Padrões.

The solution proposed

Vantagens do modelo Graph databases

* Modelagem de diferentes tipos com diferentes propriedades
* Expansível
* Requisitos de domínio implementados por regras e axiomas, no modelo
* Queries complexas
* Protocolos e padrões: SPARQL, OWL, RDF, RDFS, etc
* Fácil interoperabilidade
* Fácil integração de vocabulários e modelos
* Fácil armazenamento de resultados (novas propriedades)
* Interoperabilidade entre sistemas

Methodology

The scope is limited, at first, to the entries of DHBB. The reason lies in the fact that the data model is the simplest of the three and also due to the unconditional support received by managers for the innovation of this model.

; Focar no que estamos fazendo no DHBB – o projeto piloto, a transformação dos dados, tratamento, interface, busca.

Expected Results

1. <http://cpdoc.fgv.br>. O CPDOC é uma unidade da Fundação Getulio Vargas, instituição de ensino e pesquisa em História e Ciências Sociais, de reconhecida importância dentro e fora do Brasil. [↑](#footnote-ref-1)
2. The Deep Web: Surfacing Hidden Value, disponível em:

   http://brightplanet.com/wp-content/uploads/2012/03/12550176481-deepwebwhitepaper1.pdf [↑](#footnote-ref-2)