Introdução

O Centro de Pesquisa e Documentação de História Contemporânea do Brasil da Fundação Getulio Vargas[[1]](#footnote-1) (CPDOC/FGV) é uma instituição dedicada ao estudo e à preservação da memória recente do país, detentora de um rico acervo e de grande produção intelectual neste sentido. Ainda que se trate de um mesmo universo de discurso – a história contemporânea do Brasil –, sistemas distintos foram construídos para registrar e catalogar os diferentes tipos de informação existente na casa. O problema é que a forma atual como os recursos estão representados e estruturados digitalmente torna difícil e complexa a tarefa de manutenção desses sistemas, além de não garantir que seu acesso pelos usuários seja realizado de maneira centralizada, intuitiva e eficaz.

Este artigo apresenta o projeto que acena com um movimento de mudança de postura do CPDOC sobre seus dados, e as iniciativas em curso para a construção de um modelo que garanta a interoperabilidade, o compartilhamento e reuso das informações. Parte-se da reflexão sobre o uso de dados abertos/conectados em detrimento de dados relacionais*,* tomando o conceito de *Linked Open Data* em consonância com as três leis divulgadas pela W3C[[2]](#footnote-2), quais sejam: 1) se o dado não pode ser encontrado e indexado na web, ele não existe; 2) se não estiver aberto e disponível em formato compreensível por máquina, ele não pode ser reaproveitado; 3) se algum dispositivo legal não permitir sua reaplicação, ele não é útil.

Acreditamos que a ‘renovação’ buscada pelo CPDOC deve necessariamente passar por uma reorientação estratégica quanto aos esforços despendidos no trato das suas informações. O cenário atual revela certa precariedade no processo de disponibilização e acesso aos dados, limitando a consulta àquilo que os sistemas da casa permitem, que não é muito. No entanto, se mantivermos o foco na produção e distribuição desses dados e não mais nos sistemas de consulta, podemos abrir caminho para que outras formas de acesso sejam exploradas, e, por conseguinte, novos usos para estes dados, independentemente da intervenção do CPDOC.

A ideia segue o espírito da informação livre pregado pela Web Semântica no esforço de se construir uma rede global de dados interligados, onde a possibilidade de se descobrir coisas novas é ilimitada e está diretamente associada ao movimento de se tornarem abertos e disponíveis cada vez mais dados provenientes de iniciativas particulares ou de governos, empresas e organizações.

Dentre as iniciativas deste projeto, destacamos a construção de uma *triple store* em RDF a partir de dados originalmente armazenados em uma base relacional e a aplicação de tecnologias semânticas capazes de tornar o acervo do CPDOC em fonte de dados livres e interoperáveis aos moldes do conceito de *Linked Open Data[[3]](#footnote-3)*.

Este arranjo busca colocar em prática a visão da instituição como provedora de conteúdo e plataforma para a ação colaborativa de outras comunidades. Dessa forma, almeja-se emergir de uma experiência de pesquisa para um esforço concreto, produtivo e real, com espaço de aplicação e consumo em um centro de referência na área de documentação e preservação histórica.

Os sistemas de informação do CPDOC

O acervo do CPDOC encontra-se sob a tutela de três sistemas distintos desenvolvidos sob um mesmo banco de dados. Cada um desses sistemas possui gestão independente e adota critérios idiossincráticos quanto à organização e indexação das suas informações, que variam segundo as especificidades do conteúdo que abrigam: documentos de arquivos pessoais, entrevistas de história oral e verbetes do Dicionário Histórico-Biográfico Brasileiro. A seguir, uma breve descrição destes tipos documentais:

## Arquivos Pessoais

Documentos de época, em seus mais variados suportes textuais e audiovisuais, são fontes preciosas que nos ajudam a conhecer mais da nossa história. Eles podem estar na forma de manuscritos, diários, cartas, fotografias, discursos ou memorandos, e quando integram arquivos de pessoas que influenciaram o cenário político e social do país, representam muito mais que meras lembranças particulares: são registros concretos de toda uma memória coletiva.

Atualmente o acervo do CPDOC conta com aproximadamente 200 fundos que pertenceram a presidentes, ministros, militares e outras personalidades, que, juntos, somam cerca de 1,8 milhão de documentos, ou 5,2 milhões de páginas. Deste total, aproximadamente 700 mil páginas encontram-se digitalizadas, e a expectativa é de que todo o restante seja incluído na web durante os próximos anos. A informação descritiva (metadados) do acervo encontra-se registrada em um sistema próprio chamado Accessus que permite recuperar as informações contidas nos arquivos por meio de algumas chaves de busca, todas eminentemente sintáticas.

## Entrevistas de História Oral

O acervo de História Oral do CPDOC hoje compreende mais de 6.000 horas de gravação, correspondendo a quase 2.000 entrevistas. Mais de 90% desse total encontra-se no formato digital, armazenadas em áudio ou vídeo. Por enquanto, dois tipos de consulta estão disponíveis na base de dados: a consulta por assunto e a consulta por entrevistado. Quando o pesquisador seleciona uma entrevista, tem acesso apenas a uma ficha técnica e a um sumário, que é a descrição dos temas na ordem em que aparecem no depoimento. Dez por cento delas trazem suas transcrições na própria página, porém, para ter acesso direto ao áudio ou vídeo é necessário vir consultar diretamente no CPDOC. No momento estuda-se a melhor forma de disponibilizar estas mídias no portal da instituição, considerando desde o formato mais adequado até as políticas de uso, controle de acesso e direitos autorais que cerceiam as entrevistas.

## Verbetes do Dicionário Histórico-Biográfico do Brasil

O Dicionário Histórico-Biográfico Brasileiro é, sem dúvida, uma das principais obras de referência sobre personagens e temas de história do Brasil no período pós-1930. Abrange mais de 7.500 verbetes de natureza biográfica e temática, ou seja, pessoas, instituições, organizações e eventos cuidadosamente selecionados a partir de critérios que mensuram a relevância destes para a história política do período estabelecido. A base de dados do Dicionário limita-se a armazenar alguns poucos metadados referentes a cada verbete, restringindo a busca ao título do verbete ou ocorrência de palavras no interior do texto. Só é possível recuperar as informações após o usuário autenticar-se no ambiente do banco de dados e utilizar o formulário de busca disponível. Um dos maiores problemas na adoção deste tipo de sistema, remetendo ao que Michael Bergman identificou como sendo parte da Deep Web[[4]](#footnote-4), é a informação que não pode ser encontrada pelos mecanismos de busca padrão, ou seja, páginas que não existem até serem criadas dinamicamente como resultado de uma busca específica.

Este é sem dúvida, um dos maiores desafios que o CPDOC enfrenta diante do seu próprio acervo.

O problema

Visibilidade e acesso limitados (*no-Google*); problemas terminológicos; indexação idiossincrática; baixas taxas de revocação e precisão; sistemas independentes;

Quanto aos BDs relacionais:

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.

A solução proposta

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

Metodologia

O escopo se limitará, num primeiro momento, aos verbetes do DHBB. A razão reside no fato de seu modelo de dados ser o mais simples dos três e também pelo apoio incondicional por parte dos gestores para a inovação deste modelo.

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

Resultados esperados

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. Em <http://www.w3c.br/divulgacao/pdf/dados-abertos-governamentais.pdf>. Acesso em 22/11/2011. [↑](#footnote-ref-2)
3. http://linkeddata.org/ [↑](#footnote-ref-3)
4. The Deep Web: Surfacing Hidden Value, disponível em:

   http://brightplanet.com/wp-content/uploads/2012/03/12550176481-deepwebwhitepaper1.pdf [↑](#footnote-ref-4)