Desenvolvimento de Ontologias para o Portal Semântico do CPDOC

Renato Rocha Souza[1], Suemi Higuchi[2], Daniela Lucas da Silva[3]

[1]Escola de Matemática Aplicada – Fundação Getúlio Vargas (EMAp – FGV).

[2] Centro de Pesquisa e Documentação de História Contemporânea do Brasil – Fundação Getúlio Vargas (CPDOC – FGV).

[3]Departamento de Biblioteconomia – Universidade Federal do Espírito Santo (UFES).

{renato.souza,suemi.higuchi}@fgv.br, danielalucas@hotmail.com

Abstract. *This paper describes the semantic portal project being developed at CPDOC – FGV, along with all the initiatives that are being undertaken in order to achieve the final goal. Among those initiatives we can highlight the domain ontology creation, in the field of Brazil's contemporary history and document description, for the proper metadata supply to the documents from the archives of interest.*

Resumo. *Este artigo descreve o projeto de criação do portal semântico do CPDOC – FGV, juntamente a todas as iniciativas que estão sendo engendradas para que este seja possível. Dentre estas, destacam-se a criação de ontologias de domínio para história contemporânea e descrição de acervos, para o adequado provisionamento de metadados para os documentos pertencentes aos acervos em questão..*

1. Introdução

O Centro de Pesquisa e Documentação de História Contemporânea do Brasil (CPDOC) é parte da Escola de Ciências Sociais e História da Fundação Getúlio Vargas. Criado em 1973, tem o objetivo de abrigar conjuntos documentais relevantes para a história recente do país, desenvolver pesquisas históricas e promover cursos de graduação e pós-graduação. Dentre estes conjuntos documentais, podem-se distinguir aqueles doados por importantes personalidades da história brasileira contemporânea e aqueles que são fruto de atividades de pesquisas do próprio CPDOC, como entrevistas, dossiês e dicionários de verbetes. Os conjuntos documentais são organizados em sistemas, com características próprias, como detalhado a seguir:

- Arquivos pessoais, doados ao CPDOC como partes do espólio de personalidades públicas constituem, atualmente, o mais importante acervo de arquivos pessoais de homens públicos do país, integrado por aproximadamente 200 fundos, totalizando cerca de 1,8 milhão de documentos, entre textos, imagens e vídeos.
- Programa de História Oral do CPDOC, que desde 1975 vem produzindo um acervo de depoimentos (em áudio e vídeo) de importância reconhecida tanto no Brasil como no exterior. No total são cerca de 1.000 entrevistas, correspondendo a mais de 5 mil horas de gravação, estando metade delas abertas à consulta na web.

- Dicionário Histórico Biográfico Brasileiro (DHBB). Começou a ser desenvolvido no CPDOC-FGV em 1974 e gerou uma primeira versão impressa em quatro volumes com 4.493 verbetes sobre conceitos da História Contemporânea Brasileira. Lançada em 2001, a segunda edição do DHBB em formato de CD-ROM atualizou os verbetes existentes e incluiu novos, atingindo um total de 6.620 entradas. A versão atual, lançada em 2010 na web, compreende 7.553 verbetes, sendo 6.584 de natureza biográfica e 969 verbetes temáticos, relativos a instituições, eventos e conceitos de interesse para a história do Brasil pós-1930.

Em 2008 o CPDOC iniciou um amplo projeto de digitalização do seu acervo, ainda em curso. Em 2010 o acervo digitalizado continha a conversão de mais de 300 mil documentos textuais, 65 rolos em película, 106 fitas (VHS, Beta e U-MATIC), 350 discos, 187 fitas cassete, 85 fitas rolo e cerca de 32.000 fotografias do acervo de Arquivos Pessoais. Além disso, foram digitalizadas 5.000 horas de entrevistas do Programa de História Oral, estando toda esta documentação disponível para consulta no CPDOC . Ao final do projeto, conta-se com cerca de 80.000 fotografias digitalizadas disponíveis para consulta através da web, dando conta de praticamente todo o acervo de imagens doado até 2010 para o Centro. Além disso, todos os verbetes do DHBB se encontram em formato digital.

A característica comum aos acervos reside no fato de conterem documentos em mídias diversificadas, como texto manuscrito, texto em formato digital, áudio com e sem transcrições, imagens e vídeos com e sem legendas, caracterizando a multimodalidade midiática que apresenta difícil tratamento para fins de recuperação, e a publicização destes acervos vem sendo realizada através de interfaces e processos distintos, apesar de serem abrigados por uma única instituição e poderem ser acessados através do mesmo portal.

Em 2008, foi criada na FGV a Escola de Matemática Aplicada (EMAp), tendo como missão atuar na aquisição e repasse do conhecimento científico e tecnológico de base matemática para utilização nas áreas de interesse da FGV e parceiros. Em contato inicial com o CPDOC, propôs-se uma parceria para aplicação das técnicas de recuperação de informação desenvolvidas no escopo da Matemática Aplicada para uso no CPDOC. A partir deste contato, foi realizado um diagnóstico nos sistemas de informação do CPDOC que apontava, de maneira geral, para a necessidade de maior integração entre os sistemas e melhoria na descrição dos dados e nas interfaces de acesso. Estes motivadores levaram à criação de projetos de parceria que, em termos gerais, buscam melhorar a integração e gestão dos sistemas de informação, e acesso externo aos acervos, aumentando a visibilidade dos arquivos salvaguardados e das produções intelectuais desenvolvidas para a sociedade.

Neste artigo descreve-se em linhas gerais o projeto do portal semântico do CPDOC, e especificamente sua vertente que envolve o desenvolvimento de ontologias. O projeto prevê a migração de todo o acervo atual para uma base de dados comum em formato *RDF triplestore*, e a unificação dos padrões de descrição entre todos os fundos e sistemas, o que envolve a criação de ontologias de descrição e de domínio. Como objetivo, pretende-se oferecer uma interface única para buscas temáticas transversais e integradas, utilizando-se conceitos e categorias de conceitos relativos ao domínio da

História Contemporânea Brasileira – como pessoas, acontecimentos e locais – através de todos os sistemas/acervos atuais.

2. O problema

O principal problema a ser enfrentado se caracteriza pelo tratamento integrado de bases heterogêneas e em formatos multimídia, e a ausência de padronização nos formatos de descrição. No âmbito do projeto, almeja-se uma interface única e um padrão unificado de metadados para descrição dos inúmeros itens dos diversos acervos.

Como foram construídos de maneira independente, os acervos, sistemas e fundos adotaram padrões idiossincráticos de descrição, ressaltando diferentes características a serem descritas e diferentes terminologias para descrevê-las. Acrescenta-se a esta dificuldade o fato de ser o acervo composto por fotografias, cartas, desenhos, periódicos, entrevistas em áudio e vídeo, gravações de rádio, de vídeo, dentre outros.

3. A solução proposta

O problema proposto demanda uma série de iniciativas razoavelmente independentes de preparação dos acervos e sistemas para a migração. Estas iniciativas são descritas a seguir:

- Projeto de reconhecimento de faces e personagens: teve como objetivo otimizar os processos de gestão do acervo fotográfico do CPDOC, a partir de técnicas de reconhecimento de faces e de personagens. Como resultado, foram desenvolvidos aplicativos para tratar os fundos organizados com legendas, realizando a detecção de faces e a combinação destas com as legendas já produzidas. Além disso, atende à demanda do CPDOC de disponibilizar ao público de maneira mais amigável a localização dos personagens em cada fotografia de nosso acervo.
- Projeto de alinhamento de som e texto: teve como objetivo produzir transcrições automáticas de voz em língua portuguesa, a serem utilizadas pelo o programa de história oral do CPDOC no tratamento de seus acervos. O material utilizado é constituído de entrevistas transcritas, entrevistas gravadas – arquivo de áudio, transcrições das entrevistas – arquivo de texto, entrevistas sem transcrição, entrevistas gravadas – arquivo de áudio, Sumário das entrevistas – arquivo de texto.
- Projeto de mineração de textos: é, na verdade, um conjunto de iniciativas de processamento de linguagem natural para oferecer, entre outras coisas, Suporte aos projetos reconhecimento de faces e personagens e de alinhamento de som e texto. Nesta iniciativa, foram coletados possíveis descritores (termos frequentes encontrados em legendas de fotos, em documentos, e em transcrições de entrevistas) com vistas à incorporação nas ontologias de domínio e também no DHBB.
- Projeto de “Wikificação” do DHBB: Foi engendrado para promover uma maior interligação das bases de dados internas do CPDOC com as externas, como a própria Wikipédia, com benefícios no sentido de aumento da publicização e estruturação de redes sociais de colaboração para contribuições e eventuais correções para o acervo. Está sendo implementada através de uma ferramenta open source de Wiki Semântico (MediaWiki com extensões semânticas), e nesta *wiki* estão sendo cadastrados verbetes do DHBB para demonstrar as possíveis funcionalidades do ambiente. Este projeto se beneficia das ontologias que estão sendo criadas.

- Projeto de Criação de Ontologia a partir dos Descritores de Sistemas: a descrição dos acervos do CPDOC é realizado hoje através de uma enorme lista não hierárquica de descritores, que contém, entre outras coisas, instâncias de pessoas, entidades, processos, eventos, locais e atributos. Esta lista se constitui no primeiro levantamento de conceitos para a criação da ontologia de história contemporânea, junto aos verbetes hoje presentes no DHBB.

4. O portal semântico toma forma

Todos estes projetos são fins em si, com utilidade e potencial de melhorias imediatas para os sistemas como se encontram atualmente. Mas a culminação dos projetos constitui o embrião do Portal Semântico do CPDOC. Este compreende uma solução de migração dos acervos para um sistema único, com tecnologias abertas e preconizadas pelo W3C. A proposta de Portal encerra uma solução que proporcionará:

- Acesso unificado aos acervos dos sistemas;
- Navegação e busca pautada por conceitos, independente de mídias e de sistemas;
- Buscas transversais entre sistemas (DHBB, Arquivos Pessoais, PHO, etc.);
- Interligação dos acervos através de conceitos comuns;
- Padrões únicos de descrição de itens entre os sistemas;
- Padrões de descrição adotados mundialmente, permitindo a interoperabilidade e interligação com sistemas e acervos externos;
- Integração com os repositórios da web (*Linked Data / Linked Open Data*[1]) através da utilização de uma base de dados em padrão *RDF triplestore*;
- Conceitos relevantes estruturados sob a forma de verbetes, com nome e endereço únicos, preferencialmente sob a forma de URIs;
- Maior visibilidade do acervo sob a ótica dos mecanismos de busca;
- Possibilidades aumentadas de integração dos acervos como objetos educacionais;
- Possibilidade de integração com a Biblioteca Digital da FGV;

Dentre outros aspectos. A FIG.1. A seguir exemplifica o esquema do Portal Semântico com os processos de conversão de bases:

[1] http://linkeddata.org/

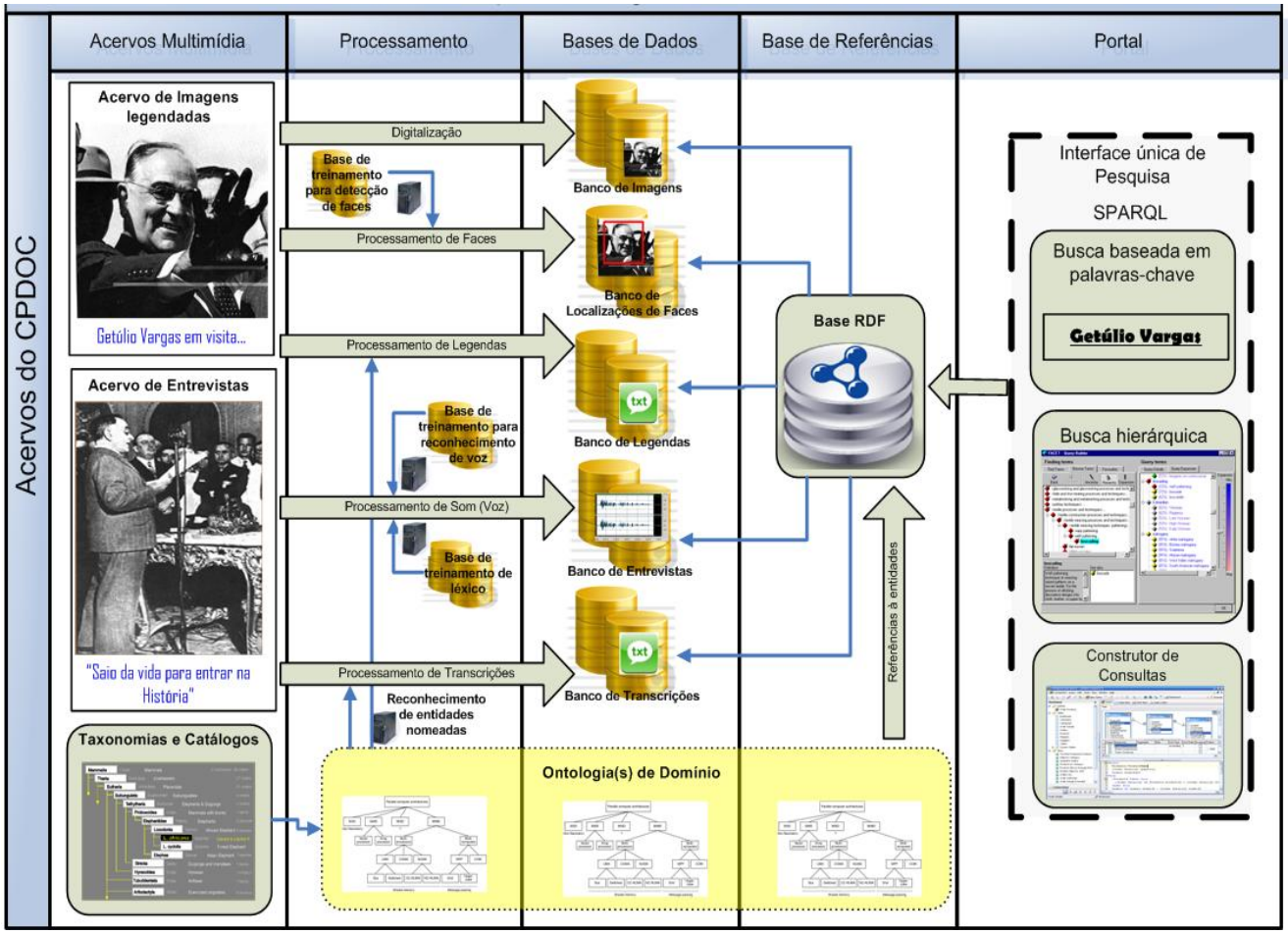

Figure 1. Esquema do Portal Semântico

Para materializar a solução integrada, em conjunto e adicionalmente aos projetos apresentados, serão desenvolvidas as seguintes ações de nível macro:

- Análise dos metadados descritivos de cada fundo/acervo;
- Criação de um formato padronizado de descrição, preferencialmente através da composição de ontologias existentes e utilizadas no escopo da web semântica (Dublin Core, Bibliographic Ontology, FOAF, dentre outros), e que sejam compatíveis com os padrões de descrição arquivística (NOBRADE, ISAAR(CPF), ISAD(G));
- Criação de Ontologias leves (light ontologies) no domínio da História Contemporânea para descrição do conteúdo dos documentos;
- Classificação dos documentos segundo os campos dos padrões de descrição adotados, e utilizando os conceitos desenvolvidos na Ontologia de História Contemporânea;
- Classificação das fotografias através da digitalização e processamento de legendas, e através de técnicas de reconhecimento de faces e de personagens;

- Classificação do material em áudio, através de processamento de transcrições e análise dos campos de metadados;
- Migração dos acervos para uma nova base de dados em formato triplestore, ou seja, um banco de dados próprio para armazenamento de dados no formato RDF;
- Interligação interna e externa dos itens dos acervos através de identificadores únicos, preferencialmente acessíveis via hipertexto (URIs e URLs);
- Criação de interfaces de pesquisa no acervo através da tecnologia SPARQL.

As ontologias a serem criadas – de descrição e no domínio da história contemporânea – serão desenvolvidas segundo a metodologia híbrida proposta em Silva, Souza e Almeida (2008), dando-se prioridade ao reuso de ontologias existentes, no caso específico das ontologias de descrição bibliográfica.

5. Discussão

Este artigo apresenta o panorama de um projeto que se encontra em plena execução, tendo sido iniciado em 2010 e com previsão de término para o final de 2012. Envolve uma série de iniciativas que estão sendo desenvolvidas em paralelo, com um horizonte de unificação, materializada na conjuminação de cinco projetos independentes, como foi apresentado, além de ações específicas do projeto do portal semântico. Constitui um projeto representativo de recuperação de informações multimodal porque lida com documentos em formatos diversificados, como áudio, vídeo, imagens, textos e modelos conceituais. Além disso, incorpora tecnologias e instrumentos oriundos do ferramental da web semântica, como triplestores RDF e ontologias. O produto final, acredita-se aumentará enormemente a publicização e acesso dos acervos e sistemas mantidos pelo CPDOC, contribuindo para seu melhor uso pela sociedade em geral.

6. Referências

SILVA, Daniela Lucas da ; SOUZA, Renato Rocha ; ALMEIDA, Maurício Barcellos de. Ontologias e vocabulários controlados: comparação de metodologias para construção. Ciência da Informação (Impresso), v. 37, p. 60-75, 2008.