UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA

UM SISTEMA DE BIBLIOTECA DIGITAL PARA DOCUMENTOS HISTÓRICOS

TRABALHO DE GRADUAÇÃO

Aluno: Marcos Cardoso Junior (mjmcj@cin.ufpe.br)
Orientadora: Flávia de Almeida Barros (fab@cin.ufpe.br)
Co-orientador: Marcos Galindo Lima (galindo@ ufpe.br)

Recife, Março de 2005

A informação é a única matéria-prima que se reproduz quando é disseminada.

- J.F.A.K. Van Benthem

Dedicatória

Dedico este trabalho ao grande amigo Mauro Quaresma, que nos deixou desde o quarto período da graduação. Sua ausência é sentida em todos os momentos de nossas vidas, principalmente nos mais difíceis, pois com sua alegria conseguia transformar qualquer trabalho cansativo em uma grande risada.

Agradecimentos

Muitas foram as pessoas que me ajudaram na confecção deste trabalho. Agradeço a todos que contribuíram de forma direta ou indireta, e em especial a:

- Meus pais, que me incentivaram por toda a vida, dando total apoio e confiança em todos os momentos, principalmente os mais difíceis;
- Minhas irmãs, que entenderam a prioridade do uso do computador e me apoiaram no desenvolvimento do presente trabalho;
- Minha namorada, companheira e amiga, Caroline, pela compreensão da minha dedicação ao trabalho e a conseqüente ausência da minha companhia;
- Flávia Barros, que aceitou a proposta do trabalho e me orientou de forma brilhante, sempre atenciosa e disposta a colaborar;
- Marcos Galindo, co-orientador, que me orientou da melhor maneira possível este trabalho. O que aprendi com esse professor vai muito além do trabalho e carregarei pelo resto da vida.
- Ricardo Prudêncio, doutor em Ciência da Computação e professor da UFPE, que contribuiu substancialmente nas definições de diversos conceitos deste trabalho;
- Meu amigo do Centro de Informática, Marcos Pereira, que me ajudou no modelo de Recuperação de Informação e nos metadados presentes no trabalho;
- Samarone Lima, jornalista, que cedeu suas entrevistas e diversos outros materiais sobre a época da ditadura militar no Brasil, me apoiando bastante;
- Marcília Gama, do Arquivo Público - PE, que acreditou em nossa proposta e está em constante ajuda para a disponibilização dos prontuários do DOPS;
- Todos do Laboratório Liber, que me ajudaram a desenvolver a idéia da Biblioteca Digital e na digitalização dos arquivos;
- AVCIn, uma associação que só me deu alegria em todos os momentos do Centro de Informática;
- A todos os meus amigos do Centro de Informática que dividiram alguns problemas e muitas alegrias ao decorrer da graduação;
- E, sobretudo a Deus, por ter colocado todas essas pessoas em meu caminho.

Resumo

O fosso que separa as fontes históricas e os pesquisadores é enorme. Nos dias de hoje, muitas vezes para um historiador ter acesso a um arquivo histórico, ele precisa deslocar-se até a instituição detentora de acervos, ultrapassando, inclusive, barreiras continentais. Desta forma, é de imensa necessidade a criação de uma nova ponte entre o universo tecnológico e o universo documental.

Observando a carência de sistemas computacionais com esse escopo, nosso objetivo principal foi criar uma biblioteca digital para o armazenamento e recuperação de informações sobre qualquer assunto histórico e em qualquer mídia, como texto, imagem, áudio e vídeo. Para o acesso a esses dados, foi criado um módulo de Recuperação de Informação para Documentos Históricos, com a finalidade de retornar ao usuário documentos mais relevantes em relação a sua consulta.

O sistema criado segue a iniciativa do *Open Archives*, aliado ao padrão de metadados *Dublin Core*, implementando técnicas para a disseminação eficiente do conteúdo para outras instituições. Existe ainda um módulo administrativo, para a gerência dos dados contidos no banco e um módulo para a visualização dos documentos históricos multimídia.

Realizamos dois estudos de caso. O primeiro é o "Pergunte a Pereira da Costa", que possui documentos sobre a história de Pernambuco entre os anos de 1493 a 1950. Nele, utilizaremos uma base textual, adquirida pelo projeto para o desenvolvimento do módulo de recuperação e buscas em documentos históricos. O segundo estudo de caso abordado é um projeto criado a partir da proposta do presente trabalho e que está sendo abrigado pelo laboratório Liber, UFPE. O projeto, denominado de *Memórias do Golpe: o Brasil de 64 a 85,* reúne diversos documentos históricos dessa época, como entrevistas e documentos nunca antes divulgados, para o amplo acesso.

Abstract

The irrigation that separates to the historical sources and the researchers is enormous. Nowadays, many times a historian needs to dislocate until the institution to have access to a historical archive, exceeding continental barriers. Is necessary the creation of a new bridge between the technological universe and the documentary universe. Observing the computational systems lack with this target, our main objective was to create a digital library for the storage and recovery of information on any historical subject and in any media, as text, image, audio and video. For the access to these data, a Information Recovery module for Historical Documents was created, with the purpose to return to the user more excellent documents in relation its consultation.

The system follows the Open Archives Initiative, ally to the metadata standard Dublin Core, implementing techniques for the efficient content dissemination for other institutions. An administrative module still exists, for the data management in the database and a module for the historical documents multimedia visualization.

We carry through two case studies. The first one is "Ask to Pereira da Costa", that contains Pernambuco history documents between 1493 and 1950. In this first case study, we will use a text base, acquired for the recovery module development in historical documents. The second study boarded is a project created from the present work proposal and it is being sheltered for the Liber laboratory, UFPE. The project, called of *Memories of the Blow: Brazil between 64 and 85*, congregates historical documents of this time, as interviews and documents never before divulged, for the ample access.

Sumário

Lista de Figuras

Lista de Quadros

1. Introdução

De acordo com a *Digital Library Federation* (DLF) [DLF], bibliotecas digitais [BIBDIGa, BIBDIGb] são organizações que fornecem recursos para selecionar, estruturar, oferecer acesso intelectual, distribuir, preservar a integridade e garantir a permanência das coleções digitais, de tal forma que elas estejam disponíveis para uma ou várias comunidades. A maioria das bibliotecas digitais disponíveis, contudo, apresentam dificuldades quanto ao acesso das informações nelas contidas. Ora a pesquisa dá-se de forma complexa (com vários campos para o usuário preencher), ora os resultados obtidos não são relevantes. Há casos em que são retornados documentos de interesse do pesquisador, mas o sistema disponibiliza apenas sua referência, sendo esse resultado da pesquisa muitas vezes inútil, e a ida a uma biblioteca tradicional indispensável.

Tratando-se de documentos históricos, a carência é ainda maior. Visitando algumas instituições que possuem acervos históricos, como a Fundação Gilberto Freyre [FGF] e a Fundação Joaquim Nabuco [FUNDAJ], percebe-se que muitos historiadores gostariam de ter acesso aos documentos sem precisar deslocar-se à instituição detentora do material. Não é raro, inclusive, pesquisadores de outros países visitarem essas instituições para pesquisar sobre algum assunto histórico específico. Bibliotecas Digitais para Documentos Históricos disponíveis são poucas. Os problemas enfrentados pelas bibliotecas digitais nesse campo específico são os mesmos enfrentados pelas citadas acima. Espera-se, entretanto, que esses acervos documentais estejam disponíveis para o acesso de todos, eliminando a necessidade do deslocamento até a instituição detentora do material. Nas Bibliotecas Digitais pesquisadas, dificilmente isso ocorre.

Para a visualização do documento, verificou-se que é necessário um módulo robusto e eficiente, pois os poucos sistemas que disponibilizam o documento são lentos e de difícil acesso, com a necessidade de instalação de alguns *plugins*. O ideal também seríamos ter uma biblioteca digital não só de acervos com imagens ou texto, mas um acervo multimídia. Assim, o usuário poderia ter acesso arquivos históricos do tipo texto, imagem, áudio ou vídeo.

Quando se trata de acervos históricos, algumas vezes o documento é de difícil leitura, por motivos como o estado de conservação do material, a qualidade da digitalização realizada no documento original ou quando o documento é manuscrito. Assim sendo, seria de fundamental importância que juntamente com esse módulo, o

usuário tivesse condições de realizar transformações no documento para propiciar sua melhor visualização. Este visualizador teria dois detalhes imprescindíveis: a usabilidade do módulo deve ser bastante satisfatória e o sistema deve ter um desempenho muito bom.

Com o avanço das técnicas de Recuperação de Informação [Yates, 1999], podemos criar bibliotecas digitais com poderosos sistemas de busca. As pesquisas efetuadas pelos usuários devem ser simples, os resultados os mais relevantes possíveis, sendo apresentados em um tempo de resposta aceitável. Para atender todos esses requisitos, esse módulo do sistema deve atrair uma atenção especial para o desenvolvedor, com pesquisas na área e tomada de decisão da melhor técnica para o escopo de dados de documentos históricos.

Além do mais, quando falamos em bibliotecas digitais, devemos sempre pensar na disseminação eficiente do conteúdo. E isso não se dá apenas com o usuário tendo acesso a uma pesquisa. É preciso que o sistema possua recursos para disponibilizar tais informações para outras instituições, de modo a ampliar o acesso aos repositórios como um meio de aumentar a disponibilização, independente do tipo de conteúdo oferecido. Para isso, verificou-se que a utilização da iniciativa *Open Archives* (OAI) [OAIa, OAIb] é de fundamental importância.

No contexto da OAI, a disponibilização das informações é realizada através de Metadados [Galindo 2004, Milstead 1999, Weibel 1995]. Trata-se de informação estruturada sobre recursos (digitais e não-digitais). Os metadados podem ser utilizados para viabilizar uma ampla série de operações nesses acervos. A biblioteca digital deve manter em formato XML [XMLa, XMLb] as principais informações do repositório, tornando-se um provedor de dados.

Com a publicação de fundos arquivísticos em meio digital, não só teremos uma disponibilização em larga escala — o que proporcionará a qualquer pessoa ligada à Internet o acesso ao conteúdo documental —, mas sua virtual preservação [UNESCO].

Para exemplificar o que será utilizado para construir a biblioteca digital para documentos históricos, decidimos atacar um assunto do interesse da maioria da população brasileira: a época da ditadura militar.

Este trabalho consiste em apresentar uma biblioteca digital para documentos históricos, com diversas técnicas, descrevendo detalhadamente cada uma delas. O restante deste documento está organizado em alguns capítulos.

No capítulo 2 será apresentado o contexto em que o presente trabalho está inserido. Serão definidos alguns conceitos sobre bibliotecas digitais, detalhando alguns casos disponíveis na Internet. Serão também apresentadas algumas técnicas de recuperação de informação, dentre as quais algumas serão usadas no sistema. Outro tópico abordado será como se dar a disponibilização de acervos, relatando sobre alguns metadados e alguns padrões, como o *Dublin Core* [DCa, DCb, DCc] e *MARC21* [MARC21], além de descrever a iniciativa *Open Archives*.

No capítulo 3 serão detalhado os principais requisitos da Biblioteca Digital para Documentos Históricos, incluindo o módulo de indexação e busca para documentos históricos que será utilizada no trabalho e como será feita a disseminação do conteúdo.

O capítulo 4 traz os dois estudos de caso do trabalho. O primeiro estudo de caso é um projeto do Laboratório Liber [LIBER], chamado Pergunte a Pereira da Costa [PC]. Resgatamos a base de dados desse projeto para desenvolver um sistema de Recuperação de Informação para documentos históricos. O nosso segundo estudo de caso trata-se da demonstração da Biblioteca Digital para Documentos Históricos criada para este trabalho. Utilizamos diversos acervos históricos da época da ditadura militar no Brasil para exemplificar o sistema criado. O presente trabalho deu origem ao projeto: Memórias do Golpe – O Brasil de 64 a 85. O mesmo será abrigado no laboratório Liber.

O capítulo 5 trará conclusões e trabalhos futuros acerca do projeto realizado.

2. Contexto

Como dito, o presente trabalho tem por objetivo apresentar um sistema para Bibliotecas Digitais para Documentos Históricos. Pesquisas foram feitas juntas a instituições, como a Fundação Gilberto Freyre, e o sistema foi modelado para abrigar qualquer acervo histórico. O sistema ainda pode abrigar documentos em formatos de qualquer mídia, como texto, áudio, vídeo ou imagem.

A solução aqui proposta faz uso de técnicas de Recuperação de Informação para que os documentos retornados por um engenho de busca sejam relevantes ao que o usuário pesquisou. Além disso, também utiliza um sistema para a correta disseminação da informação, seguindo as diretrizes do *Open Archives Iniciative* (OAI).

O propósito desta seção é apresentar alguns conceitos sobre Bibliotecas Digitais, listando algumas de escopo geral e outras específicas para documentos históricos, apresentando os seus problemas e as possíveis soluções. Sendo a recuperação da informação uma área essencial ao presente trabalho, apresentamos algumas técnicas para o mesmo, suas vantagens e desvantagens. Ainda nessa seção apresentaremos de que forma podemos realizar a disponibilização dos acervos, com uso de metadados e explicando o que se trata a iniciativa *Open Archives*. Por fim, faremos algumas considerações finais do que foi tratado nessa seção, mostrando as melhores soluções para o escopo do projeto.

2.1. Bibliotecas Digitais

Como o próprio nome sugere, Biblioteca Digital (ou Biblioteca Virtual) pode ser descrita como uma biblioteca que dispensa um ambiente físico e com informação não mais atrelada ao suporte de papel impresso [Prudencio 2003].

Mas podemos encontrar na literatura muitos conceitos para o termo.

"Bibliotecas digitais são organizações que fornecem recursos para selecionar, estruturar, oferecer acesso intelectual, distribuir, preservar a integridade e garantir a permanência das coleções digitais, de tal forma que elas estejam disponíveis para uma ou várias comunidades" [DLF].

"Uma biblioteca que mantêm toda, ou uma parte substancial de sua coleção numa forma processável pelo computador como uma alternativa, suplemento ou complemento à

forma impressa tradicional e material em microfilme, que, atualmente, domina os acervos bibliográficos." [Saffady 1995].

"Uma coleção organizada de dados multimídia com métodos de gerenciamento da informação, que representa os dados como informação útil e conhecimento para o povo numa variedade de contextos sociais e organizacionais" [Griffin 1995].

"Coleção organizada de dados multimídia em rede" [Mosata].

Documentos que fazem parte de uma Biblioteca Digital podem ser produzidos originalmente em formatos digitais - imagens, arquivos texto produzidos através de editores – ou podem ser cópias digitalizadas de documentos originais.

Podemos encontrar na Internet Bibliotecas Digitais para os mais diversos fins, algumas com alguns problemas críticos, como a falta de um sistema de buscas, como pode ser visto na figura 2.1 [BVL].

Figura 2.1: Exemplo de uma biblioteca digital sem um sistema de busca

Apesar de ser uma excelente iniciativa (um portal de literatura portuguesa), a interface da Biblioteca Digital da figura 2.1 não dá suporte a consultas por palavras-chaves. A única possibilidade de navegar em seu conteúdo é através do índice, oferecido na interface. A solução para ampliar e facilitar o acesso aos dados seria a indexação do conteúdo dos documentos, um trabalho nem sempre trivial.

Um outro exemplo de Bibliotecas Digitais interessante é a Biblioteca Virtual do Rio Grande do Sul [MHN], mostrada na figura 2.2. Aqui existe a possibilidade da busca por conteúdo, porém tal sistema é muito complexo para usuários leigos, que desconhecem a álgebra de Boole.

Figura 2.2: Pesquisa com operadores booleanos explícitos.

Veremos a seguir exemplos de Bibliotecas Digitais para Documentos Históricos e alguns problemas específicos a esta aplicação.

2.1.1. Bibliotecas Digitais para Documentos Históricos

Quando o assunto é direcionado para bibliotecas digitais especificamente para documentos históricos, os problemas são mais agravantes. Além de existirem poucas disponíveis para este fim, elas geralmente não mostram o documento histórico, e sim a

referência sobre o mesmo (onde está localizado fisicamente) e, poucas vezes, um resumo.

A seguir, serão apresentadas duas Bibliotecas Digitais para Documentos Históricos disponíveis na Internet.

A figura 2.3 traz a página de pesquisa da Biblioteca Digital do Museu Histórico Nacional [MHN].

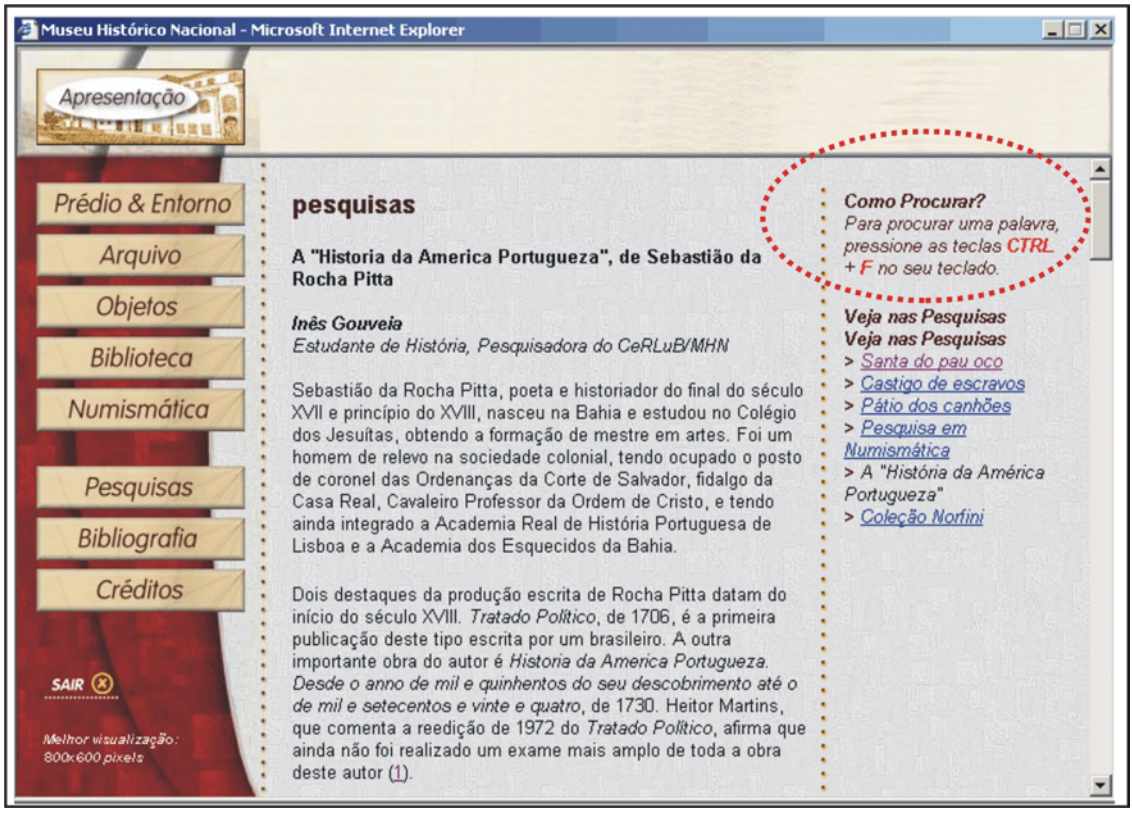


Figura 2.3: No detalhe (acima, lado direito), a forma de como proceder a "busca" na biblioteca digital.

Essa figura mostra uma Biblioteca Digital para Documentos Históricos com uma busca pouco convencional: utilizando o artifício do navegador, o CTRL+F, para proceder a "busca". Assim como o problema identificado na Biblioteca Digital da figura 2.1, este sistema necessita de um sistema de buscas.

Uma das maiores Bibliotecas Digitais para documentos históricos do mundo, a LOC (*Library of Congress*) [LOC] possui um projeto chamado "*American Memory*" [AMLC]. A biblioteca possui um sistema de busca e os resultados são retornados ao usuário rapidamente. Contudo, o sistema de visualização do documento histórico é feito de

forma bastante simples, como mostrado na figura 2.4. Mesmo assim, essa Biblioteca Digital é tomada como referência, pois, como já foi dito, muitas delas não disponibilizam o documento histórico para a visualização.

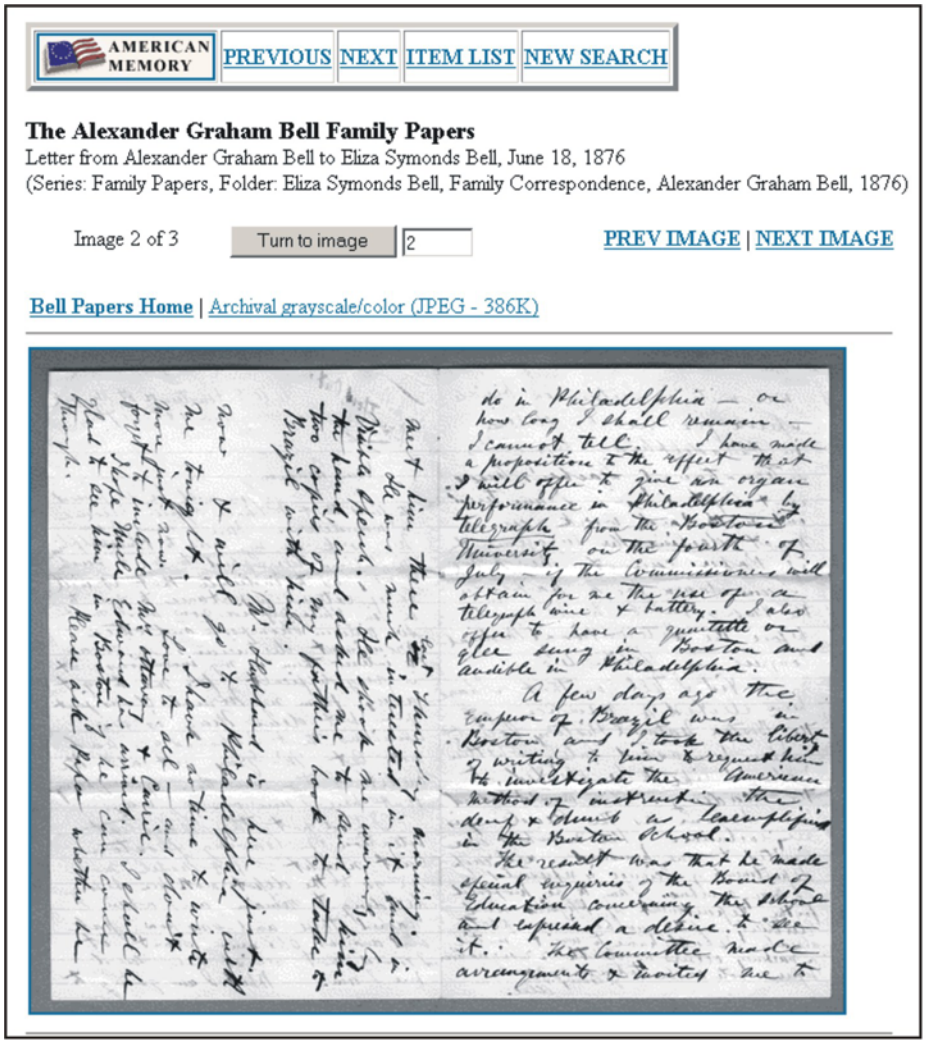


Figura 2.4: A visualização do documento histórico na LOC

Alguns documentos históricos, como mostrado na figura 2.4, são pouco visíveis. Seria bastante útil ao usuário puder realizar transformações no documento para a sua melhor visualização, como um simples zoom na imagem. Poderíamos também ter a possibilidade de clarear, escurecer ou até mesmo negativar (algumas vezes, apenas com um documento negativado é que a sua visualização pode ser realizada).

2.2. Recuperação da Informação para Bibliotecas Digitais

Recuperação de informação é a representação, armazenamento, organização e acesso aos dados contidos em uma base de dados. A representação e organização dos dados devem prover fácil acesso ao usuário às informações que ao mesmo interessa [Yates 1999].

A simples recuperação de dados, inseridos dentro do contexto de recuperação de informação, consiste em determinar que documento de uma coleção contém as palavras-chave da consulta de um usuário. Freqüentemente, isto não satisfaz o usuário, não retornando a informação realmente desejada.

Então, técnicas robustas são necessárias para que o usuário possa ter acesso aos dados de forma mais relevante possível.

Ao contrário de sistema de buscas na Web, como o *Google* e o *Yahoo!*, que processam bilhões de informações, lidando com bases com eternas modificações, mostraremos casos em que usamos as buscas em sistemas mais fechados. Tratando-se especificamente de Bibliotecas Digitais para Documentos Históricos, os dados raramente mudam (a não ser que não tenham sido inseridos na base de forma correta). Partindo desse pressuposto, iremos então investigar como se dá a Recuperação da Informação para esse escopo de projeto: corpus de documentos (itens de dados) que dificilmente irão sofrer alguma alteração.

Basicamente, um sistema de recuperação de informação para esse escopo de projeto é formado por um corpus de documentos e um sistema de consulta (representada por palavras-chave), a partir do qual o sistema encontra um conjunto ordenado de documentos que são relevantes para o usuário.

A arquitetura básica de um sistema de RI [Barros 2005] pode ser verificada na figura 2.5.

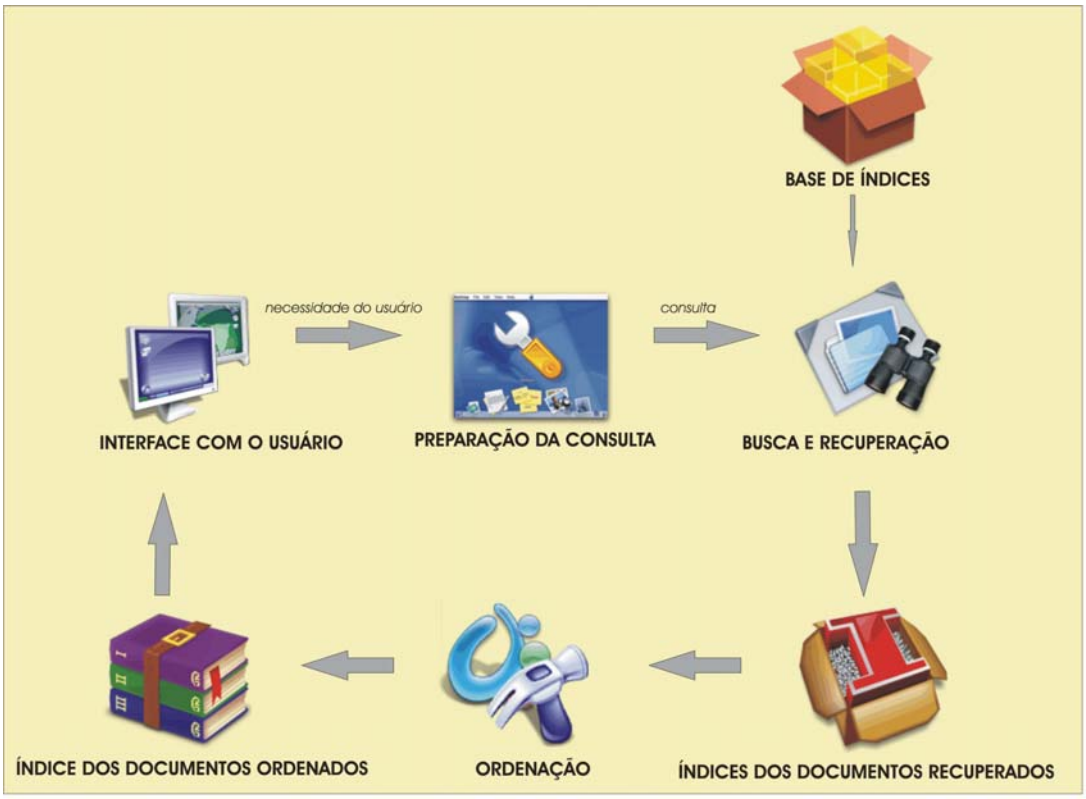


Figura 2.5: Arquitetura Básica de um sistema de RI

Um sistema de Recuperação de Informação possui basicamente cinco etapas principais, que serão descritas nas subseções que seguem.

2.2.1. Aquisição (seleção) dos documentos

Após a definição de qual será o tema de sua biblioteca digital, precisamos fazer a aquisição dos mesmos. É nessa fase que os documentos serão selecionados de acordo com a necessidade da instituição. Dependendo do tipo de Biblioteca Digital, os documentos selecionados só serão imagem ou texto, por exemplo; de uma determinada resolução ou tamanho.

A aquisição de documentos históricos comporta num trabalho difícil e custoso. Os documentos estão distribuídos em diversas instituições e disponíveis em mídias variadas, como papel, microfilme, fitas K-7, VHS, entre outros. Muitos dos documentos estão ilegíveis e/ou em estado precário de conservação e apenas com a manipulação de um especialista na área, tal tarefa pode ser realizada.

2.2.2. Preparação dos documentos

Nem todas as palavras de um texto são realmente significativas para representar a semântica de um documento. Algumas palavras possuem maior significado do que outras. Deveremos considerar então, nesta etapa, o processamento dos textos de um documento para determinar quais termos serão usados para identificar o acervo. São os chamados termos indexados. O objetivo principal desta segunda etapa é selecionar os termos do documento que melhor descrevem o seu conteúdo, reduzindo a complexidade da representação do mesmo.

A seleção desses termos pode ser feita manualmente ou automaticamente. Na primeira opção, ela geralmente é realizada por um especialista na área, como um bibliotecário, ou, no caso de uma Biblioteca Digital para Documentos Históricos, também pode ser feita por um historiador. A seleção automática é a mais comum e utilizada pela maioria dos sistemas de RI.

Vale salientar que, em casos de engenhos de busca na Web, essa seleção não existe, ou seja, a busca é feita no texto completo. Esse procedimento retorna uma visão mais lógica e mais completa do documento. Por outro lado, o custo computacional para esse procedimento é muito alto. Tratando-se de bases fechadas, como as das bibliotecas digitais, podemos realizar uma redução nos termos que ocorrem no documento sem grande prejuízos.

Um bom algoritmo para operações sobre o texto pode reduzir em até 30% o texto de seu tamanho original [Yates 1999]. Com isso, cresce a performance da busca e a relevância dos resultados alcançados.

Muitas são as operações sobre o texto e cada sistema de RI implementa uma ou mais dessas fases, dependendo de seu propósito. Dentre as operações, destacam-se:

- **Análise Léxica**

O objetivo desta operação é converter o texto original em uma lista de palavras, identificando cada palavra que ocorre no texto. Tem-se como procedimento padrão utilizar espaços como separadores de palavras, tratando pontuação, hífens, dígitos, letras maiúsculas e minúsculas de acordo com o caso abordado, já que cada caso requer tratamentos diferenciados.

- **Eliminação de Stop Words**

Algumas palavras não são bons descriminadores, ou seja, não possuem um valor semântico associado ao documento. Encontra-se nesse conjunto palavras muito freqüentes na base (não sendo relevantes na busca) ou termos como artigos, preposições, conjunções, alguns advérbios e adjetivos.

Duas grandes vantagens nessa operação são a diminuição na representação do texto e a melhora na ordenação na recuperação. A desvantagem é que diminui a cobertura na recuperação, ou seja, alguns resultados relevantes podem ser desperdiçados na consulta.

- **Stemming**

Freqüentemente, o usuário especifica uma palavra na consulta, mas apenas uma variação dessa palavra aparece nos documentos relevantes. Como exemplo, podemos citar palavras no plural, no gerúndio, verbos flexionados, aumentativo ou diminutivo.

O objetivo principal dessa operação é substituir a palavra pelo seu radical, possibilitando o casamento entre variações de uma mesma palavra. Por exemplo, quando um usuário faz uma busca pela palavra documento, o sistema pode retornar registros não só com a palavra buscada, mas com termos como documentos, documentação, documentário, documentado, etc. Muitas são as técnicas de stemming, mas nenhuma universal, pois todas elas dependem muito do idioma em questão. Especificamente citando a língua portuguesa, essa técnica é mais complexa, pelo fato do idioma ser mais complexo por possuir muitas regras.

- **Identificação de Grupos Nominais**

É nesta técnica que identificamos os grupos nominais (termos compostos) para indexar o documento. Por exemplo: Recuperação de Informação, Inteligência Artificial.

Tais operações sobre o texto criam uma visão lógica do documento, criando uma representação do mesmo, que será utilizado pelo modelo de RI escolhido.

Um exemplo desta passagem do documento original, para a sua representação, através de algumas operações sobre o texto, se encontra na figura 2.6.

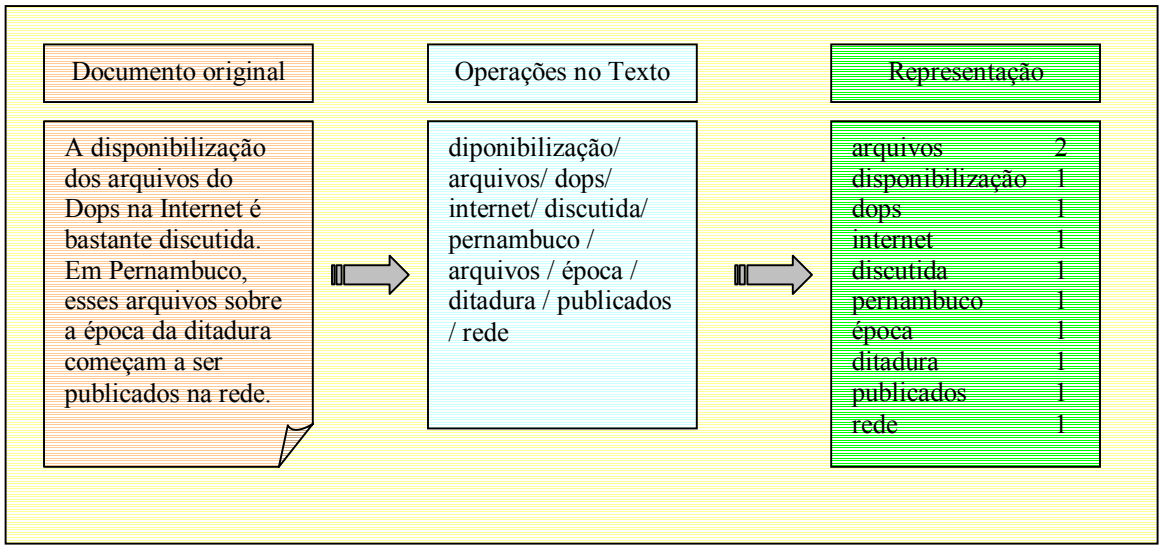


Figura 2.6: Representação de um documento original

2.3. *Disponibilização de acervos digitais*

A disseminação da informação, hoje em dia, consegue atingir diversas instituições, dentre elas, as Bibliotecas Digitais. Com o avanço da Internet, todavia, o volume de informações disponíveis cresceu substancialmente, causando alguns problemas como o grande número de detentores de informações e seus alto graus de autonomia e a falta de uma estrutura para acolher esses dados [Galindo 2004]. Com isso, o desenvolvimento de padrões que descrevem essas informações de forma exata torna-se imprescindível para aquelas instituições que desejem disponibilizar os seus dados, ou, focado no presente trabalho, para disseminar os seus acervos digitais.

Uma parte da solução do problema descrito acima é o uso de metadados. Muitos são os conceitos encontrados sobre o assunto. Segundo a definição de Tronchin: "Metadado é a descrição do dado, do ambiente onde ele reside, como ele é manipulado e para onde ele é distribuído" [Tronchin 1998]. Ou seja, trata-se de informações estruturadas sobre os recursos presentes em um repositório de dados. Tais recursos podem ser imagens, livros, músicas, artigos científicos, documentos históricos, dentre muitos outros.

Podemos considerar que há três formas possíveis de relacionar estruturas de informação e as estruturas descritas pelos metadados, que são mostrados na figura 2.7 [METAPT].

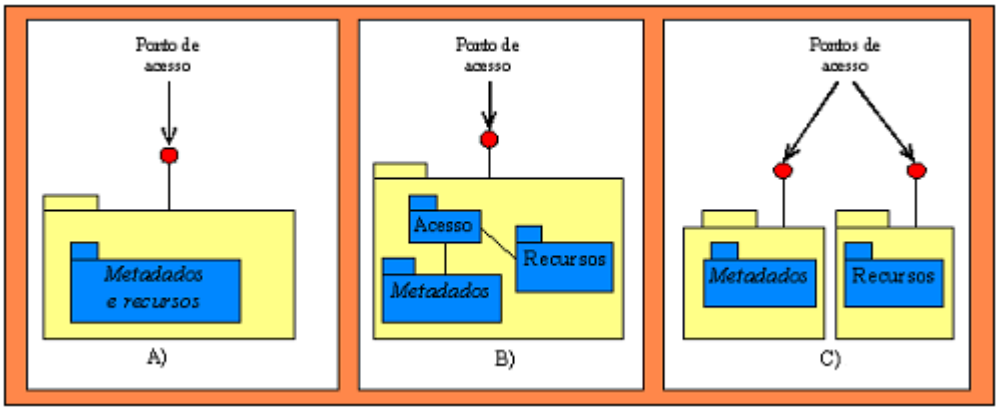


Figura 2.7: Relacionamentos entre os metadados e os recursos

Os chamados *metadados embutidos* são aqueles cujos recursos são disponibilizados em formato HTML, onde são ebutidos nos documentos das páginas, através das meta-tags (<META>). Exemplo deste tipo pode ser visto na figura 2.8.

```
<head>
<title>Exemplo de Metadados Embutidos</title>

<meta name="title" content="Exemplo de Metadados Embutidos">
<meta name="creator" content="Marcos Cardoso Junior">
<meta name="abstract" content="Este documento faz parte do
trabalho de graduação e traz um exemplo de Metadados
Emebebidos">
<meta name="format" content="text/html">
</head>
```

Figura 2.8: Exemplo de metadados embutidos

O cenário de *metadados associados* é inerente ao HTML. Utilizando a tecnologia XML, é possível definir um recurso como um grupo de registros em um arquivo separado. Esse arquivo XML referencia os conteúdos das informações de acordo com o padrão de metadados escolhido. Apesar da separação dos recursos e seus metadados, o acesso é feito em apenas um ponto. A figura 2.9 exemplifica este tipo de metadados.

```
<?xml version="1.0"?>

<metadata>
  <title>Exemplo de metadados associoados</title>
  <creator>Marcos Cardoso Junior</creator>
  <abstract>Este documento faz parte do trabalho de graduação e
traz um exemplo de Metadados Associados</subject>
  <format>text/html</format>
</metadata>
```

Figura 2.9: Exemplo de metadados associados

Finalmente, a perspectiva dos *metadados separados* é a do modelo da biblioteca tradicional, com bases de dados bibliográficas existentes em sistemas próprios e os recursos, isto é, os livros, presentes nas prateleiras.

2.3.1. O Padrão de Metadados Dublin Core

Existem diversos padrões de metadados disponíveis. Dentre eles *MARC21* e o *SCORM* [Ghiglione 2003]. È em um terceiro padrão de metadados, contudo, que iremos concentrar os nossos esforços: o padrão de metadados *Dublin Core*. Esse padrão de metadados é o mais utilizado hoje em dia, por ser simples, de fácil entendimento, e por seus elementos se encaixarem na maioria das informações contidas em um repositório.

Esse padrão possui algumas características fundamentais para a escolha do mesmo para o presente projeto:

- A simplicidade na descrição dos recursos;
- Entendimento semântico universal dos elementos;
- Escopo internacional e extensibilidade (o que permite a adaptação às necessidades adicionais de descrição).

A Iniciativa de Metadados *Dublin Core* (*Dublin Core Metadata Initiative - DCMI*) surgiu na cidade de Dublin, Ohio. A iniciativa é uma organização dedicada a promover e difundir padrões interoperabilidade entre metadados e desenvolver vocabulários especializados para descrever os recursos que permitem aos sistemas mais inteligentes a descoberta da informação.

O conjunto de elementos *Dublin Core* (Dublin Core Metadata Element Set, Version 1.1) possui 15 elementos, descritos no quadro 2.1:

Quadro 2.1: Elementos do Dublin Core

	Nome do Elemento	Definição
01.	Title	O nome que o recurso é formalmente conhecido.
02.	Creator	É a entidade responsável por quem criou o conteúdo do recurso. Pode ser uma pessoa, uma organização ou um serviço.
03.	Subject	Pode ser o assunto ou palavras-chaves do recurso.
04.	Description	Exemplos deste elemento pode ser resumo, sumário, ou livre texto sobre o conteúdo.
05.	Publisher	Uma entidade responsável por disponibilizar o recurso. Como exemplo, pode ser uma pessoa, um serviço ou organização.
06.	Contributor	As entidades responsáveis por dar contribuições ao conteúdo do recurso.
07.	Date	Data de um evento no ciclo de vida de um recurso.
08.	Type	Natureza ou gênero do conteúdo do recurso.
09.	Format	O tipo de mídia do recurso.
10.	Identifier	Referencia de um identificador único para o recurso de um dado contexto. Pode ser uma URL.
11.	Source	A referência de um recurso da qual ele é derivado.
12.	Language	O idioma do conteúdo do recurso. Exemplo podem incluir “en”, “pt”, para inglês e português, respectivamente.
13.	Relation	Uma referência para um recurso relacionado.
14.	Coverage	Extensão ou escopo do conteúdo do recurso.
15.	Rights	Informação sobre propriedade intelectual sobre o recurso

Além do conjunto principal, o *Dublin Core* ainda dispõe de outros elementos, denominados elementos de refinamento. Tais elementos possuem a funcionalidade de complementar a descrição da base de recursos, caso as mesmas não tenham sido totalmente descritas pelo conjunto de elementos principal.

2.3.2. Open Archives Iniciative

Em inglês, *Open Archives Iniciative;* abreviado, OAI; em português: Movimento dos Arquivos Abertos.

O OAI tem o objetivo de desenvolver e promover padrões de interoperabilidade que visam facilitar a disseminação eficiente de conteúdo.

Tal movimento teve início com o objetivo de ampliar o acesso a bases de dados de artigos científicos. Os principais padrões e ferramentas desenvolvidas, contudo, não dependem do tipo de conteúdo que é oferecido.

Algumas foram as motivações que deram início ao OAI:

- Promover consolidação mundial de repositórios científicos;
- Acesso gratuito aos metadados;
- Interface entre repositórios e provedores de serviços;
- Protocolo de fácil implementação e baseado em padrões já existentes (XML, HTTP, Dublin Core).

Dois principais sistemas foram criados a partir da iniciativa: os provedores de dados e os provedores de serviços [Garcia 2003]. Mas a principal conquista do OAI foi a criação do protocolo OAI-PHM.

2.3.2.1. O Protocolo OAI-PHM

O protocolo OAI-PMH [OAIa] é um mecanismo para transferência de dados entre bibliotecas digitais. Ele serve como uma interface para que um servidor que deseje disponibilizar os metadados possa fazê-lo com facilidade.

Algumas das vantagens desse protocolo são:

- Baixo custo de implementação e manutenção;
- Fácil transferência de dados entre repositórios via Internet;
- Padrão aberto;
- Baseia-se em padrões já bastante difundidos na Internet, como o protocolo HTTP e Dublin Core.

Para um melhor entendimento do protocolo, segue a figura 2.10.

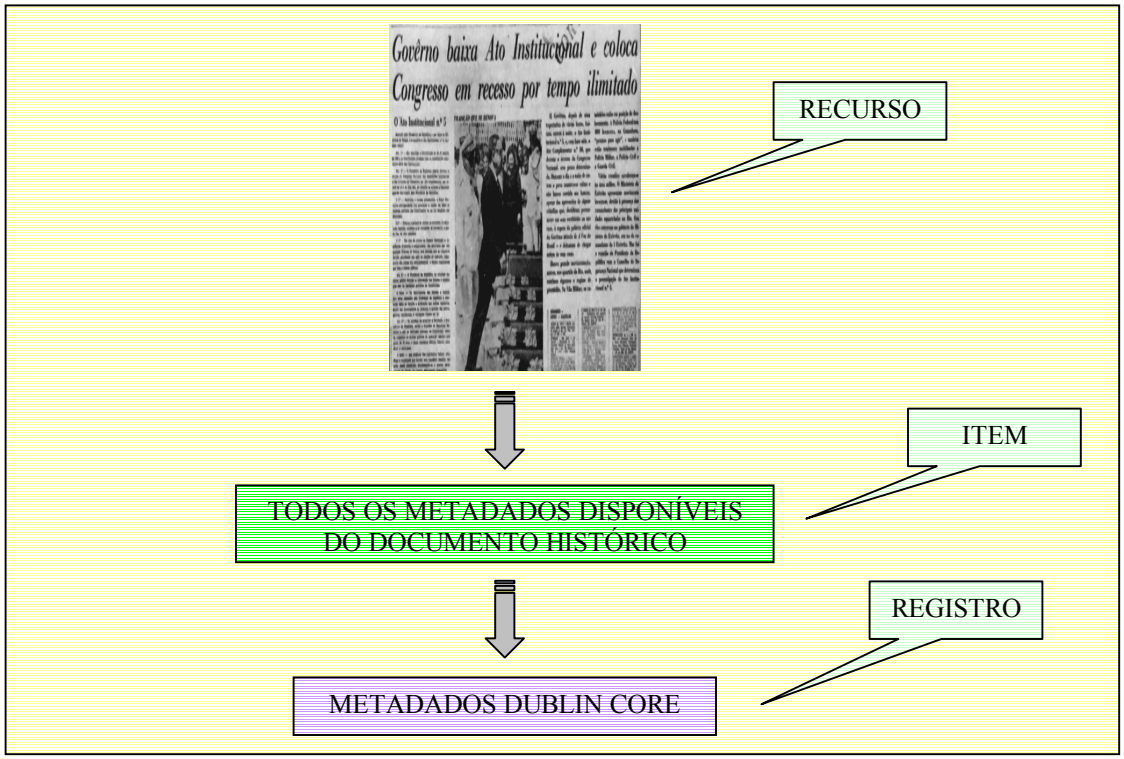


Figura 2.10: Representação gráfica do protocolo OAI-PHM

O documento mostrado acima, do Jornal do Brasil, pode ser considerado um recurso. Todos os metadados disponíveis a respeito destes recursos constituem um item. O conjunto de metadados em algum formato, por exemplo, Dublin Core, é um registro. Tal registro é o metadado de um recurso num formato específico, como o Dublin Core. Um registro tem três partes: um cabeçalho, um metadado (em XML) e opcionalmente um *about* (sobre).

2.3.2.2. Provedores de Dados e Serviços

Também chamados de Repositórios, os provedores de dados [Garcia 2003] são sistemas que utilizam o protocolo OAI-PHM para expor as informações de seus dados através dos metadados. Esses provedores também podem oferecer acesso gratuito a textos completos e a outros recursos. No Diálogo Científico [DIACIE] há um servidor de artigos digitais implementado através do software *Eprints* [EPRINTS], da Universidade Southampton.

Outro conceito importante para o OAI são os chamados *Harvesters* [Garcia 2003]. Os *Harvesters* são programas que utilizam a interface oferecida pelo protocolo OAI-PMH para coletar e armazenar metadados.

Já os provedores de serviços utilizam os metadados coletados pelos Harvesters como base para construção de novos serviços.

2.4. Considerações Finais

O propósito desta seção foi apresentar detalhadamente conceitos que serão tratados no decorrer do trabalho. Foram detalhadas diferentes técnicas de recuperação de informação para bibliotecas digitais, mostrando vantagens e desvantagens de cada uma delas.

Foi descrito também como disseminar informação de forma estruturada. Para isso, foi explicado como disseminar eficientemente o conteúdo, através do OAI e do padrão de metadados *Dublin Core*.

As próximas seções mostraram como foram utilizadas essas técnicas no desenvolvimento do presente trabalho.

3. Uma Biblioteca Digital para Documentos Históricos

Neste capítulo trataremos de explicar minuciosamente a Biblioteca Digital para Documentos Históricos, de acordo com os itens citados em todo o capítulo 2.

3.1. Concepção do Sistema

Um protótipo de uma Biblioteca Digital para Documentos Históricos foi desenvolvido em conjunto com o laboratório Liber, UFPE, sob a coordenação do professor Marcos Galindo, co-orientador do presente trabalho. O projeto foi chamado de Ultramar [ULTRAMAR], e pode-se dizer que o mesmo foi o precursor do trabalho aqui proposto.

O projeto reúne documentos da época do Brasil Colônia e foi um experimento com a visão de se construir um projeto com proporções bem maiores.

Tratando-se de um experimento, o projeto possuía diversas limitações. A busca era apenas uma aquisição de dados (ver seção 2.2), o protótipo foi implementado apenas para aceitar os documentos do Brasil Colônia, não havia a idéia da disponibilização dos dados com os metadados e o OAI, e diversas outras carências. Resumindo, o sistema se limitava a retornar os dados e apresentar o documento ao usuário. A figura 3.1 mostra a página de entrada do projeto.

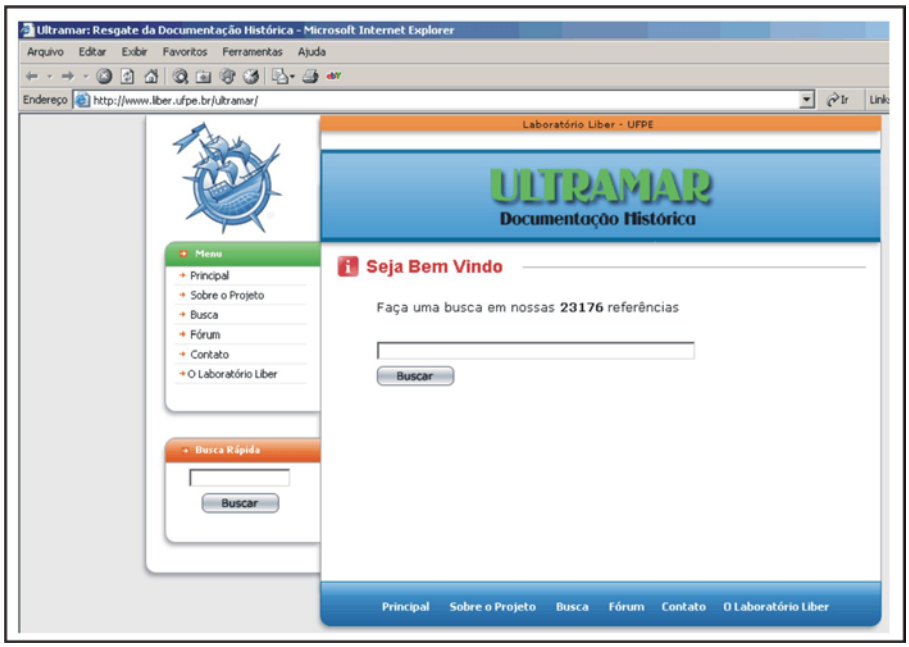


Figura 3.1: O Projeto Ultramar

Contudo, o projeto, que era simplesmente um protótipo, carregava consigo o conceito inovador de disponibilização ao alcance de todos de documentos históricos muitas vezes virgens ao olhar de um usuário comum ou, até mesmo, de um historiador. Desta maneira, o projeto atingiu interesse, inclusive, internacional. A Universidade de Salamanca (Espanha), a Universidade do Porto (Portugal) e o Instituto Real de História da Universidade de Leinden, Holanda demostraram interesse em adquirir o sistema para dar continuidade às pesquisas que foram iniciadas no laboratório. Diversas outras instituições nacionais também demostraram interesse, como a Fundação Joaquim Nabuco e a Fundação Gilberto Freyre, cedendo alguns de seus acervos para compormos e desenvolvermos a idéia do projeto Ultramar, até mesmo, fechando parcerias com o Liber.

Pode-se afirmar que, o objetivo principal do projeto Ultramar foi alcançado, ou seja, o experimento foi uma grande contribuição à discussão do uso de novas tecnologias no ambiente científico da História. Dessa forma, viu-se a necessidade da criação de um sistema completo, para a gerência e manipulação de qualquer tipo de documento histórico, seja ele da mídia da mídia que for (texto, imagem, áudio e vídeo).

Assim sendo, o presente trabalho pretende preencher essa lacuna, com uma Biblioteca Digital com diversos recursos e que a mesma irá proporcionar outras

pesquisas na área tecnológica e na área histórica. Alguns recursos que serão explicados nas seções que seguem:

- A Arquitetura do Sistema;
- A modelagem do Banco de Dados que aceita qualquer tipo de documentação histórica;
- A Visualização do documento histórico, com diversas funcionalidades para a melhor interação usuário-documento;
- A utilização do padrão de metadados *Dublin Core* a partir da Iniciativa de Open Archives, disponibilizando as informações em formato XML/RDF [RDF, Miller 1998];
- O Sistema de Administração para gerenciar as informações do sistema;

Alguns aspectos tecnológicos também foram especificados. A tecnologia utilizada para desenvolver o sistema foi o PHP [PHP, Niedarauder 2004]. O PHP é uma linguagem para criar páginas dinâmicas na Web. A curva de aprendizagem da linguagem é extremamente pequena, possui um rápido tempo de desenvolvimento e a performance dela é muito boa: essas foram as características que determinaram pela escolha da tecnologia citada.

O banco de dados escolhido para o armazenamento das informações foi o MySQL [MYSQL]. O MySQL é o mais popular banco de dados open-source do mercado. Fornece recursos simplificados e apropriados para as suas aplicações, tendo um custo extremamente reduzido. Possui ferramentas para a maioria das exigências necessárias para uma aplicação de base de dados corporativa, fornecendo uma arquitetura extremamente rápida e de fácil utilização. Muitas são as características para a escolha da tecnologia da base de dados, dentre as quais se destacam: confiabilidade e desempenho, facilidade de utilização e distribuição, recursos e suporte para as mais diversas plataformas.

Nas próximas seções serão apresentadas todas as alternativas que já foram descritas e citadas anteriormente.

3.2. Descrição das funcionalidades

Várias foram as funcionalidades que foram idealizadas para o desenvolvimento dessa Biblioteca Digital para Documentos Históricos mostrada no quadro 3.1.

Quadro 3.1: As funcionalidades da Biblioteca Digital para Documentos Históricos

Funcionalidade	Descrição
1. Busca de informações, não de dados.	De acordo com as descrições especificadas em 2.2, o sistema deveria prover um sistema de Recuperação de Informação, não de dados. Tal sistema ainda deveria ter a opção do usuário realizar uma busca avançada, limitando a sua consulta em alguns requisitos.
2. A Biblioteca Digital deve aceitar qualquer tipo de documentação histórica.	Deseja-se ter um sistema em que a sua modelagem seja totalmente adaptada para qualquer acervo histórico.
3. Biblioteca Digital Multimídia	O sistema deve aceitar arquivos do tipo texto, áudio, vídeo e imagem.
4. Disponibilização das informações para outras instituições.	A Biblioteca Digital deve ser também um servidor de dados (seção 2.3.2.2), disponibilizando as suas informações em metadados no padrão Dublin Core (2.3.1), seguindo a iniciativa do Open Archives (seção 2.3.2).
5. Deve permitir que, além de ter acesso virtual ao documento, o usuário possa manipulá-lo para a sua melhor visualização.	Na visualização do documento, o usuário pode realizar algumas manipulações no mesmo: • Aumentar ou diminuir; • Negativar, pois alguns documentos históricos são melhores vistos depois de sua negativação; • Girar / Inverter; • Clarear ou escurecer; • Desafazer essas manipulações, voltando o documento em seu formato original.
6. Excelente usabilidade do Visualizador de Documento .	Com diversas funcionalidades que podem ser feitas em um documento, é essencial que esse módulo tenha uma interface agradável e sua usabilidade seja de fácil entendimento.
7. Ser um sistema colaborativo	Quando o usuário visualizar o documento, ele deve ter a opção de deixar algum comentário sobre o mesmo. Tornando-se assim, um sistema colaborativo, pois alguns usuários, podem realizar a paleografia de documentos pouco visíveis. Ao deixar essa nota, a mesma deve ser enviada para o seu e-mail e armazenada em nossa base, para que possa ser lida por outros usuários.
8. Opção de realizar o download do documento	Ao visualizar o documento histórico, o usuário pode ter a opção de facilmente realizar o download do mesmo. Se o documento histórico for do tipo imagem

histórico	ou texto, o sistema deve gerar automaticamente o PDF dessas mídias e dar opção do usuário salvar. Caso seja áudio ou vídeo, o download é feito diretamente do arquivo disponibilizado.
9. Menu de Ajuda	Para não restar dúvidas, o sistema deve prover um módulo de ajuda para os usuários que sentirem dificuldades na manipulação do documento.
10. Os documento histórico (texto ou imagem) devem ser identificados, informando que ele foi retirado da Biblioteca Digital em questão.	Os documentos históricos devem ser representados com uma capa, informando que os mesmos foram salvos da Biblioteca Digital para documentos históricos. Em caso do documento for do tipo imagem, uma imagem que será representada pela capa deve ser gerada automaticamente, informando o título do documento histórico.
11. A Biblioteca Digital deve conter um sistema de Administração.	Para o cadastro (inserção, alteração e exclusão) de documentos históricos, a biblioteca digital deve fornecer esse serviço para o gerenciamento de seu acervo eletrônico. O sistema só deve ser acessado através de um login e uma senha.

No que diz respeito à funcionalidade 8, para a geração automática do PDF, é necessária uma biblioteca que forneça esse suporte. A mais utilizada para a linguagem PHP é a *PDFLib* [PDFLIB]. Existem muitas outras bibliotecas que realizam essa tarefa da geração de PDF na linguagem PHP, como a *ClibPDF* [CLIBPDF] e a *FPDF* [FPDF]. A escolha da *PDFLib* deu-se pela sua facilidade de uso e alta performance.

Já para a funcionalidade número 10, é necessário utilizar uma biblioteca do PHP chamada GD [GD], para gerar a imagem automática que identificará os documentos.

3.3. A Arquitetura do Sistema

A arquitetura do sistema foi criada para ser de fácil compreensão e implementação. Para tal, utilizamos uma arquitetura modularizada em que os seus principais entes são relacionados na figura 3.2.

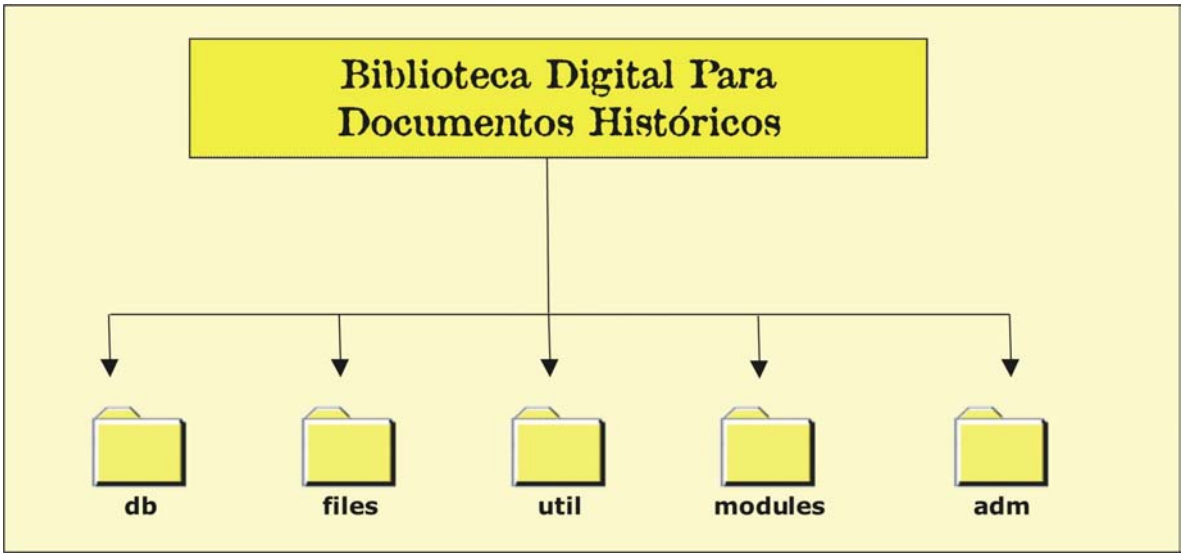


Figura 3.2: Arquitetura Básica da Biblioteca Digital para Documentos Históricos

Na área "db", encontram-se todas as consultas SQL necessárias para o sistema, além do arquivo de conexão com a base de dados MySQL.

A área "files" armazena todos os arquivos que são utilizados pela Biblioteca Digital. A figura 3.3 mostra como ela está organizada.

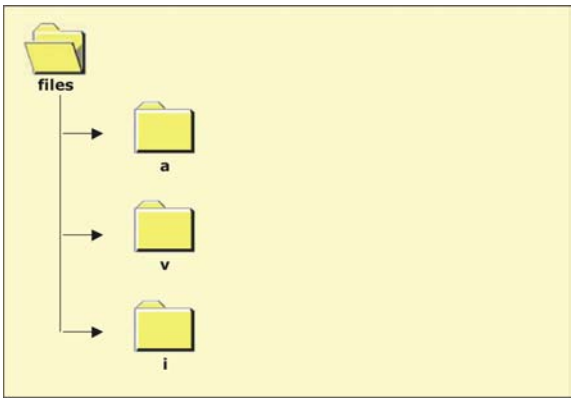


Figura 3.3: Estrutura da área files.

Como podemos ver, a área files é sub-dividida nas áreas "a" (áudio), "v" (vídeo) e "i" (imagem). Cada uma delas tem a responsabilidade de armazenar os arquivos da mídia correspondente.

Voltando à figura 3.2, podemos visualizar a área “util”. Nela são depositadas algumas funções auxiliares que ajudam na implementação.

A área “modules” é a principal do sistema e nela são depositados os seus diferentes módulos.

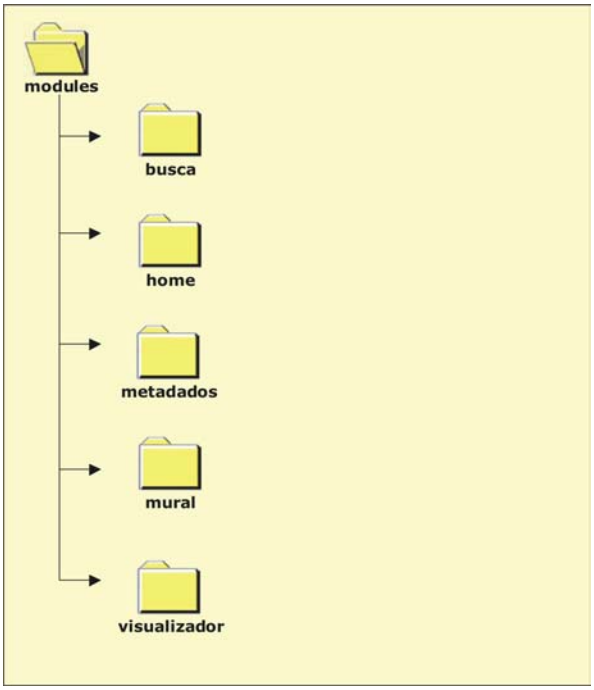


Figura 3.4: Estrutura da área modules

Cada sub área descrita na 3.4 é responsável por depositar os arquivos que integram cada módulo.

A sub-área “busca” armazena os arquivos e algoritmos do módulo de Recuperação de Informação do sistema. Na “home”, temos os arquivos das páginas principais. A “metadados” é responsável por guardar os arquivos responsáveis por disponibilizar as informações na base de dados em formato de metadados no padrão Dublin Core. A sub-área “mural” armazena os arquivos referentes ao módulo do mural da Biblioteca Digital. Por fim, na sub-área “visualizador” contém o módulo responsável pela visualização do documento histórico.

Retornando para a figura 3.2, ainda encontramos a área “adm”, responsável por conter os arquivos referentes ao sistema de gerenciamento da Biblioteca Digital.

3.4. A modelagem do Banco de Dados

Uma das funcionalidades mais relevantes descritas no quadro 3.1 foi a possibilidade do sistema de abrigar qualquer tipo de documentação histórica. Uma modelagem do banco de dados bem feita seria de fundamental importância para atingir essa meta.

Para isso, com a ajuda da Fundação Gilberto Freyre, foi feita uma modelagem da base de dados para aceitar qualquer documentação histórica. Em um trabalho minucioso, foi catalogada boa parte do acervo da instituição e verificado os principais atributos a cada um deles.

Ao final, foi determinado que um documento histórico deveria possuir os atributos listados no quadro 3.1.

Tabela 3.2: Atributos de um documento histórico

Atributo	Descrição
Titulo	Titulo do Documento
Autor	Autor do Documento
Resumo	Para descrição do documento, um resumo é sempre necessário.
Série	Todo documento possui uma série documental, como correspondências iconografia, entrevistas, etc.
Edição	Alguns documentos históricos possuem edições.
Local	Onde o documento foi produzido / impresso.
Editor	Editora, gravadora, produtora do vídeo ou áudio.
Data	A data da produção / impressão do documento.
Palavras-chave	As palavras-chaves que identificam o documento
Contribuição	Co-autor, destinatário, organizador ou colaborador.
Fonte	A fonte onde o documento foi recolhido.
Idioma	Língua em que está escrito/falado o documento.
Conservação	Estado de conservação do documento. [b] boa [m] média, [r] ruim
Propriedade	Copyright, direitos do autor, propriedade intelectual.

O restante do banco de dados foi modelado tomando como base esses atributos definidos para documentos históricos. A modelagem do banco de dados pode ser visto na figura 3.5.

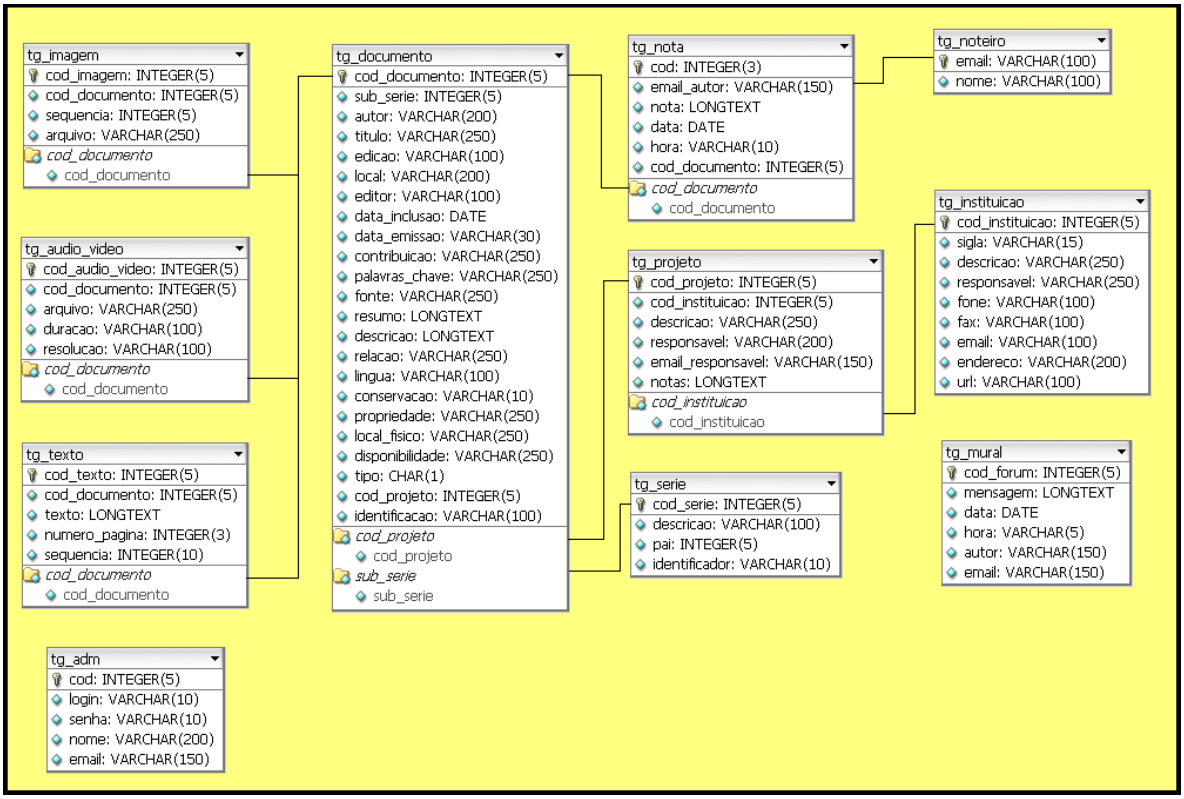


Figura 3.5: Modelagem do Banco de Dados

3.5. O Visualizador do Documento Histórico

Outra funcionalidade bastante importante no sistema é a visualização do documento histórico. É nela que o usuário irá poder realizar diversas manipulações no documento (veja tabela 3.1), salvá-lo, podendo ainda inserir e ler notas sobre o documento em questão.

A idéia básica foi construir esse módulo baseado em diversas janelas móveis, tentando assim deixá-lo com uma usabilidade já consagrada em diversos programas e sistemas operacionais no mercado. Os menus principais podem ser visualizados na figura 3.6 e a completa visualização do documento histórico pode ser vista no segundo estudo de caso, na seção 4.2.

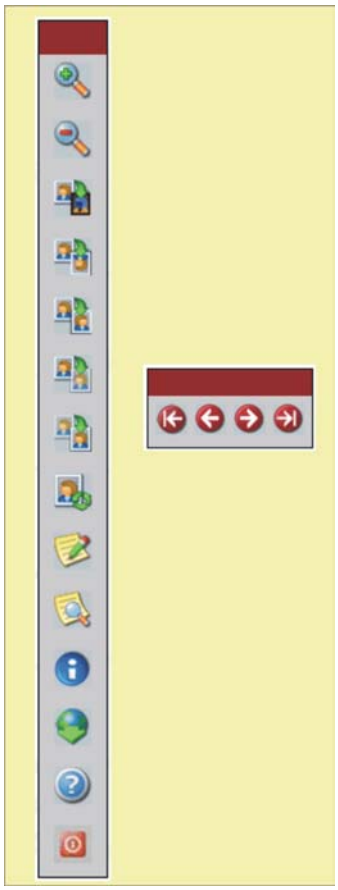


Figura 3.6: Janelas principais do Visualizador de Documentos Históricos.

Todos os efeitos encontrados neste módulo do sistema foram implementados utilizando DHTML (*Dynamic HTML)* e *JavaScript* [DHTML]. Estas tecnologias permitem que a usabilidade e a interface agradem ao usuário.

De acordo com a figura 3.6, o menu localizado à esquerda corresponde a todas as opções que o usuário pode realizar enquanto visualiza o documento. Maiores detalhes são encontrados no quadro 3.3.

Quadro 3.3: Todas as opções do visualizador de documentos

Símbolo	Funcionalidade
 **Zoom In**	Aumentar o Documento
	Diminuir o Documento

Zoom Out	
 **Negativar**	Algumas vezes, para uma melhor visualização do documento histórico, é necessário que o mesmo seja negativado.
 **Girar Vertical**	Se for necessário realizar um giro verticalmente no documento, esta opção pode ser habilitada.
 **Girar Horizontal**	Da mesma maneira, pode-se ter a necessidade de girá-lo horizontalmente.
 **Menos Opaco**	Se o usuário sentir a necessidade de clarear um pouco o documento, basta clicar nesse botão do menu e o mesmo será clareado.
 **Mais Opaco**	Caso a opção seja de escurecer o documento, essa opção pode ser habilitada.
 **Restaurar**	Após diversas manipulações, o usuário pode ter a opção de voltar às configurações originais do documento.
 **Inserir Notas**	Para inserir notas sobre o documento. A nota é enviada por e-mail ao usuário e armazenada em nossa base de dados.
 **Ler Notas**	Para o usuário ler as notas enviadas por outros visitantes do documento que está sendo visualizado.
 **Informações**	Para saber de informações sobre o documento, esta opção irá mostrar em uma janela dados como título, resumo, autor e outros do documento.
 **Download**	Realiza o download do documento. Se o mesmo for da mídia texto ou imagem, um PDF é gerado automaticamente e o usuário terá a opção de salvá-lo. Caso seja do tipo áudio ou video, o arquivo salvo será o próprio arquivo.
 **Ajuda**	Um menu de ajuda sobre como utilizar o visualizador de documentos.
 **Sair**	Sair do visualizador de documentos.

Ainda de acordo com a figura 3.6, ao lado esquerdo temos o menu de navegação do documento, com as opções de ir para a próxima página, voltar a pagina, ir para o início ou fim do documento. As demonstrações de algumas funcionalidades do visualizador de documentos serão mostradas na seção 4.2.

3.6. A disponibilização do Acervo para outras instituições

Como já foi citado anteriormente, muitas instituições estão preocupadas na interoperabilidade das informações [Rosetto]. A nossa Biblioteca Digital foi implementado para ser um servidor de dados (ver seção 2.3.2.2). Utilizamos princípios do OAI e o padrão de metadados *Dublin Core*. O mapeamento dos elementos do *Dublin Core* com os atributos de nossos documentos podem ser vistos no quadro 3.4.

Quadro 3.4: Mapeamento dos elementos Dublin Core com os atributos do documento histórico.

Elemento do Dublin Core	Atributo de nossa Base
Title	Titulo
Creator	Autor
Description	Resumo
Coverage	Local
Version	Edição
Publisher	Editor
Date	Data
Contributor	Contribução
Subject	Palavras-chave
Source	Fonte
Type	Tipo de mídia (texto, áudio, vídeo ou imagem)
Language	Idioma
Rights	Propriedade
Identifier	A URL da referência ao documento.

Para a disponibilização desses metadados no formato Dublin Core, foi utilizado o padrão XML/RDF. O RDF (Framework para Descrição de Recursos) é o formato textual XML mais utilizado para descrever recursos e aplicações de metadados.

Desta forma, a consulta para uma determinada palavra-chave retorna aos Harvesters (2.3.2.2) um conjunto de elementos no padrão de metadados Dublin Core em XML no formato RDF.

Integrando documentos, formulários e sistemas de menu, o HTML conseguiu atingir o sucesso na Web. Dessa mesma maneira, o RDF pode integrar aplicações e agentes numa Rede (Web) Semântica [Souza 2004, W3XSWEB]. As descrições formais dos termos de uma certa área (no nosso caso, documentos históricos) são denominadas ontologias e são uma parte vital da Web Semântica. RDF, ontologias e a representação formal do significado, de modo que os computadores possam ajudar as pessoas em seus trabalhos, são todos tópicos em discussão no grupo *Semantic Web Activity* [SWA].

3.7. O Sistema de Administração

O sistema de administração foi criado para a gerência das informações da Biblioteca Digital para Documentos Históricos.

Abaixo serão descritas brevemente todas as funcionalidades deste módulo.

Logar no Sistema

Por questões de segurança, é de fundamental importância que o administrador da Biblioteca Digital logue-se no sistema antes de manipulá-la.

Cadastrar Documento

Faz a inserção, alteração e exclusão dos documentos históricos no sistema.

A inserção consiste em duas etapas. A primeira, é de preencher os atributos do documento, como título, autor, resumo (ver tabela 3.2 para maiores detalhes). A segunda é responsável por fazer o upload dos arquivos para o servidor onde serão armazenados os arquivos.

Cadastrar Projetos

Responsável por realizar a inserção, alteração e exclusão dos projetos que são abrigados pela Biblioteca Digital.

Cadastrar Instituições

Esta funcionalidade tem por finalidade aplicar a inserção, alteração e exclusão das instituições que estão ligadas aos projetos. Por exemplo, o Laboratório Liber é a instituição ligada ao projeto "Resgate das vozes da resistência nos anos de chumbo". Já o Arquivo Público de Pernambuco é a instituição que está ligada ao projeto "Arquivo DOPS, Pernambuco". Para maiores detalhes dos projetos, veja a seção 4.2.

Cadastrar Administradores

Insere, altera e exclui os possíveis administradores do sistema.

Estatísticas

Seria bastante interessante para os administradores do sistema ter a opção de visualizar alguns dados dos usuários que acessam o sistema. Para isso, utilizou-se um serviço disponível gratuitamente na Internet, que passam dados como o número de visitas em seu sistema, a porcentagem do uso do navegador, em que local o sistema já foi acessado. O nome do serviço é o *Extreme Tracking* [EXT].

Esses dados servirão para cada vez mais o sistema direcionar as necessidades para o maior número de usuários possíveis.

3.8. O módulo para indexação e buscas em bases de documentos históricos

As técnicas utilizadas em cada sistema de Recuperação de Informação variam de acordo com o escopo do problema. Nos dois estudos de casos que iremos abordar, utilizamos a mesma técnica.

No primeiro estudo de caso (seção 4.1), adquirimos uma base de dados textual já existente para desenvolvermos o sistema de indexação e buscas em documentos históricos. No segundo estudo de caso (seção 4.2), a técnica utilizada foi a mesma, com a diferença de termos documentos não só textuais, mas documentos do tipo áudio, texto, vídeo e imagem.

Nas próximas seções deste capítulo, iremos descrever as técnicas escolhidas para a indexação e buscas em documentos históricos.

3.8.1. *Visão Geral do Sistema*

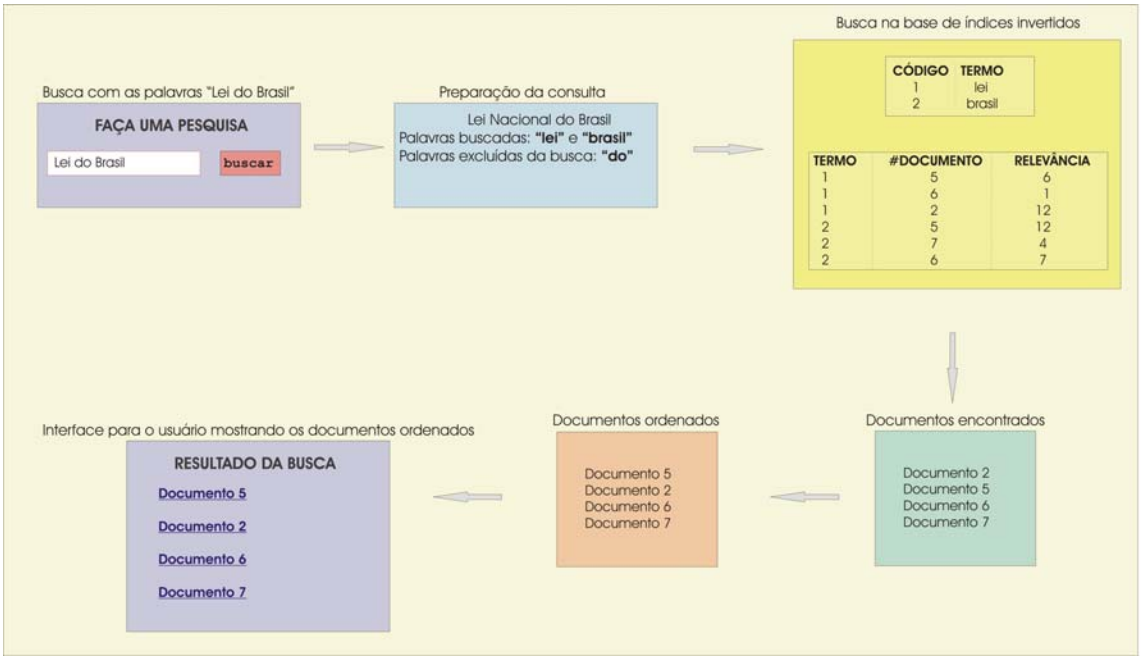


Figura 3.7: Arquitetura escolhida com fluxo de dados

Como podemos ver na figura 3.7, a arquitetura escolhida é bem parecida com a arquitetura básica para um sistema de RI, ilustrada anteriormente pela figura 2.5. Todos os passos relatados na figura 3.7 serão explicados minuciosamente nas próximas seções.

3.8.2. *Modelo de Recuperação de Informação Utilizado*

Para escolher o modelo de RI usado no sistema, foram analisados alguns modelos, como o Booleano e o Espaço Vetorial.

O primeiro modelo tem a vantagem da fácil implementação e de possuir uma teoria bem fundamentada. Contudo, tal modelo apresenta algumas desvantagens, sendo a principal delas a impossibilidade de ordenação dos documentos recuperados. Por outro lado, o modelo Espaço Vetorial permite o casamento parcial dos documentos com as consultas, associando pesos não-binários aos termos, possibilitando a ordenação dos documentos.

Para usufruir da facilidade das consultas booleanas com a ordenação do modelo Espaço Vetorial, decidiu-se optar pela implementação de um modelo chamado Booleano Estendido. Este modelo estende o modelo booleano incluindo a noção de casamento

parcial e termos com pesos, combinando características do modelo vetorial com propriedades da álgebra booleana.

3.8.3. Aquisição e preparação dos documentos

Como foi dito na seção 2.2.1, a aquisição de documentos históricos é um trabalho bastante custoso, em que na grande maioria das vezes só pode ser realizado por um especialista. Desta forma, em nosso caso, a aquisição dos documentos que serão trabalhados será feita de forma manual, com ajuda de historiadores e bibliotecários.

De acordo com o que já foi citado sobre as preparações dos documentos, decidimos trabalhar com duas das fases descritas na seção 2.2.2: a análise léxica - eliminando pontuações – e a eliminação de stop words – artigos, pronomes e palavras muito freqüentes na base. Vale salientar que, na análise léxica, também poderia ocorrer a eliminação de dígitos. Decidiu-se, entretanto, não eliminar esses termos, pois se tratando de documentos históricos, alguns dados como datas ou citação de alguns anos podem ser de extrema relevância ao documento histórico.

A análise léxica é feita de modo totalmente automático. Ou seja, os documentos são processados por um algoritmo que retira as pontuações da base textual.

Para documentos textuais, a identificação das stop-words dos documentos é semi-automática. A primeira fase é feita de modo automático. Termos que possuem uma freqüência muito alta na base são selecionados para que possamos identificar se eles possuem ou não uma relevância ao documento. Essa quantidade é definida de acordo com o tamanho de seu corpus (veja os dois estudos de casos, seções 4.1 e 4.2). Com isso, podemos identificar quais os termos da base serão identificados como stop-words, manualmente.

A preparação dos documentos se restringe a documentos textuais, uma vez que não dispomos de processadores de imagens, vídeos ou áudio. Independente da mídia, cada documento é indexado por um conjunto de atributos comuns: título, resumo, autor, palavras-chave, local de emissão, editor, data fonte e idioma. Esses atributos são armazenados no banco de dados da Biblioteca Digital. A indexação e busca é realizada com base nesses campos.

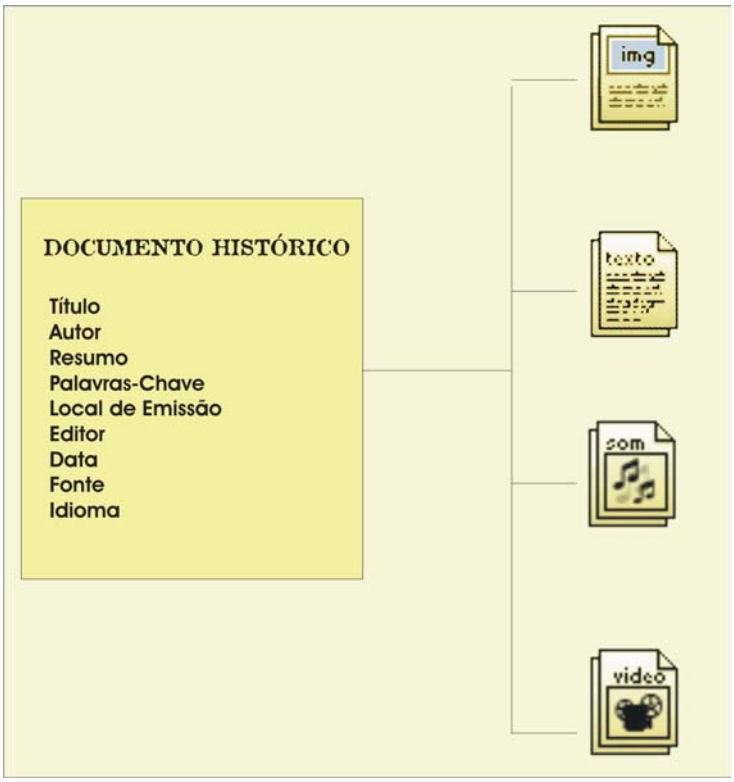


Figura 3.8: Representação interna dos Documentos Históricos.

A figura 3.8 representa os atributos comuns e os seus respectivos arquivos, de acordo com a mídia. Como a busca é textual, apenas se a mídia for do tipo texto, a consulta é feita pelos termos que estão contidos conteúdo do documento (ver figura 3.9). Para o caso restante (áudio, vídeo ou imagem), as palavras-chaves são selecionadas manualmente.

Mais detalhes sobre os atributos utilizados pode ser visto no quadro 3.2.

3.8.4. Criação da base de índices

Por tratar-se de documentos históricos, é evidente que os seus registros dificilmente sofrerão alterações. Com isso, decidiu-se calcular a relevância de cada termo em um documento específico ainda na indexação. Com isso, a indexação torna-se mais lenta, mas por outro lado, o ganho de performance do engenho de busca é notável. Maiores detalhes de como foi desenvolvido esse módulo especifico serão explicados na seção 4.1.

Após o uso da análise léxica e da identificação das stop-words, é criado um arquivo de índices invertidos, que utiliza as palavras dos dados para indexar uma coleção de documentos, facilitando a tarefa da busca. A sua estrutura é alocada em um arquivo separado e possui dois atributos: o termo e a freqüência que ele ocorre na base.

Um outro arquivo foi criado, o chamado Arquivo Invertido com TF-IDF (Term Frequency Inverse Document Frequency) [Ramos]. Esse tipo de Arquivo Invertido normalmente armazena a referência de cada termo, indicando em que documento o termo aparece e a freqüência deste termo no documento. Além do mais, como foi explicado anteriormente, a relevância de um termo no documento é pré-calculada, sendo assim, nesse arquivo invertido, além dos atributos citados, ele também armazenará a relevância que esse termo possui no documento. A escolha dos pesos foi feita com base na importância de cada atributo na descrição do documento. O cálculo dos pontos para chegar-se à relevância de um termo no documento é mostrado no quadro 3.5:

Quadro 3.5: Cálculo dos pontos para a relevância

Campo em que o Termo Aparece	Pontos Somados à Relevância Final
Título	6 pontos
Autor	6 pontos
Palavras-Chave	6 pontos
Resumo	6 pontos
Local de emissão	4 pontos
Editor	2 pontos
No conteúdo do documento (mídia do tipo texto)	1 ponto (TF-IDF)

Para um melhor entendimento da técnica utilizada, veja a figura 3.9 demonstrando um exemplo.

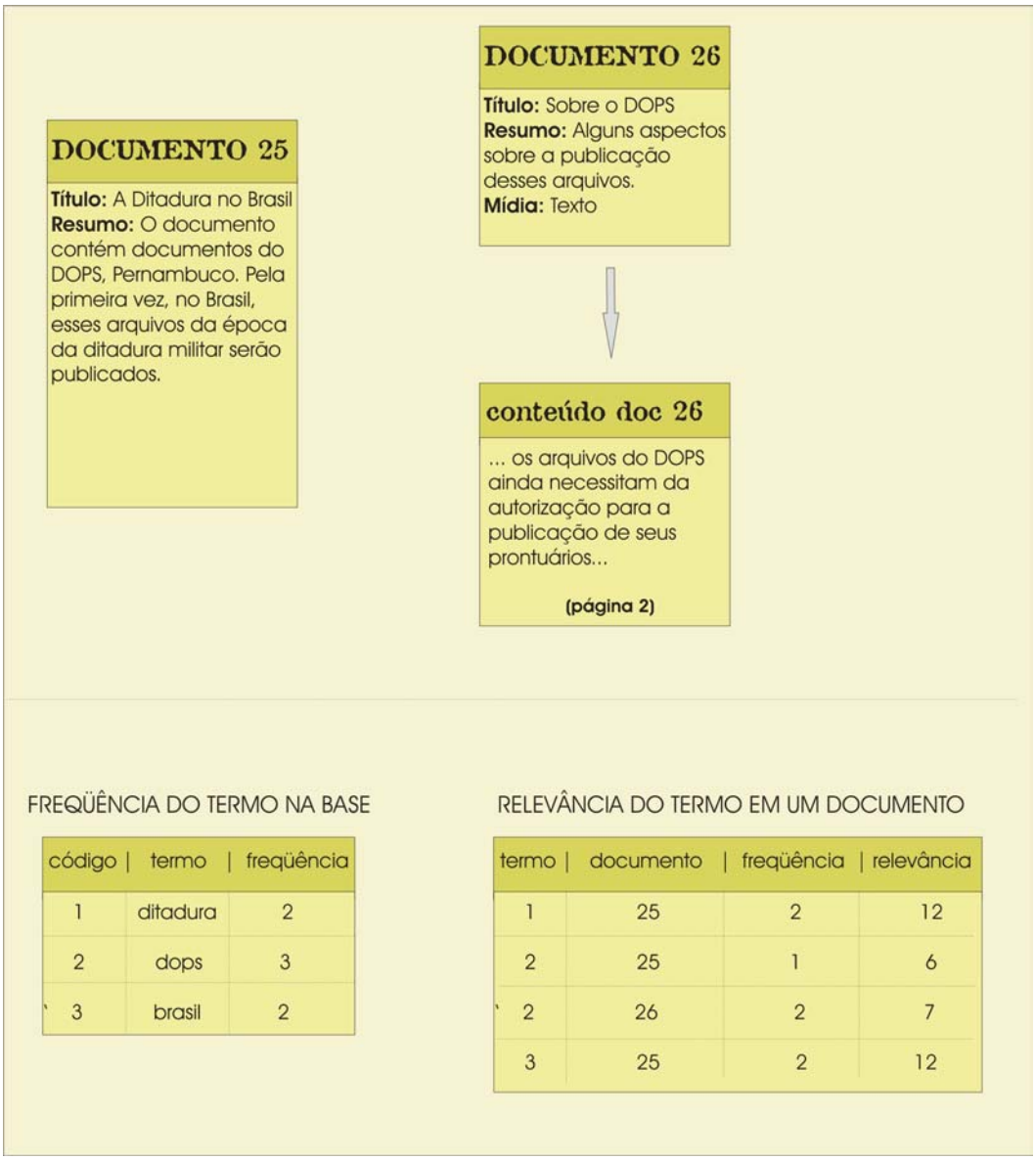


Figura 3.9: Arquivos de Índices Invertidos

3.8.5. Recuperação de documentos

A recuperação dos documentos é feita através dos arquivos de índices invertidos, eliminando a necessidade de realizar a consulta diretamente da base de textos dos documentos.

Geralmente, os documentos históricos com a mídia texto são produzidos através de uma passagem de OCR (reconhecimento óptico de caracteres) no documento. Algumas

vezes a passagem do OCR não tem 100% precisão, com alguns erros na passagem do leitor óptico. Além do mais, um idioma (seja ele qual for), sofre diversas alterações ao longo dos anos e, para documentos muito antigos, algumas e pequenas variações das palavras podem ser identificadas. Com isso, a consulta também é feita utilizando um casamento de padrões simples, permitindo a recuperação de documentos com palavras idênticas ou aproximadas a uma dada palavra. Este tipo de consulta é de fundamental importância para consultas em documentos históricos. Por exemplo, algumas vezes, podemos ter uma palavra no documento que é referida como "Pernambucodo". Sabemos que essa palavra deveria ser "Pernambuco", mas os problemas do OCR a tornaram assim. Sendo assim, o usuário que digitar a palavra-chave "Pernambuco" para a sua busca, serão retornados documentos que possuam a palavra "Pernambucodo".

Todas as consultas são feitas utilizando o auxílio de uma biblioteca do MySQL chamada *Full Text Search* [FTS]. *Full Text Search* é uma biblioteca de funções com alguns modelos de buscas já prontos, dentre elas, a consulta por similaridade de termos em um documento.

A biblioteca também permite a realização de consultas booleanas, utilizando operadores NOT, AND e OR. Aliando o *Full Text Search* com a ordenação do resultado dos documentos históricos pela sua relevância, encontramos um engenho de busca simples, com alta performance e que atende a necessidade do escopo do projeto.

A ordenação dos resultados é feita através da relevância dos termos guardados em cada documento. Se por acaso, o usuário consultar por mais de uma palavra, e essas palavras constarem em um único documento, a soma das relevâncias dos termos são somadas e a ordenação é realizada.

3.9. Considerações Finais

Neste capítulo citamos e exemplificamos como dar-se-á toda a estrutura das Bibliotecas Digitais para Documentos Históricos. No próximo, demonstraremos os resultados das técnicas escolhidas para o sistema de indexação e buscas no primeiro estudo de caso. No segundo, demostraremos o nosso produto principal, que é toda a Biblioteca Digital criada para documentos da época da ditadura militar brasileira.

4. Estudos de Casos

4.1. Estudo de Caso 1: Pergunte a Pereira da Costa

O 1º estudo de caso trata-se de um projeto realizado no laboratório Liber denominado "Pergunte a Pereira da Costa". O que mais interessou nesse projeto foi sua grande base de dados com informações sobre a história de Pernambuco. Com ela, podemos desenvolver as devidas técnicas para a criação, manipulação e testes do sistema de Recuperação de Informação para Bibliotecas Digitais para Documentos Históricos. A figura 4.1 mostra a apresentação do portal.

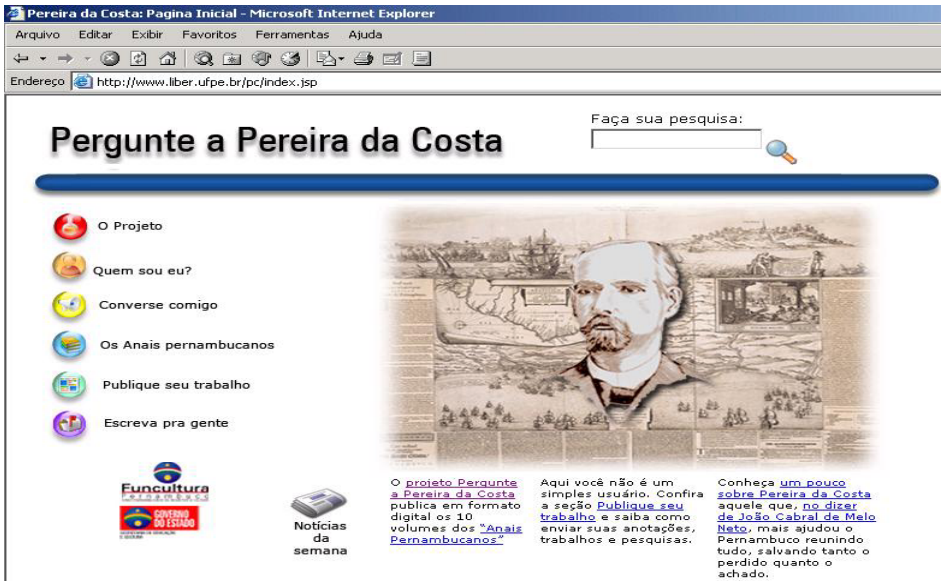


Figura 4.1: O portal Pergunte a Pereira da Costa

O presente trabalho precisava de um sistema de busca eficiente e robusto. Necessitávamos, entretanto, de um corpus de documentos para o desenvolvimento e os devidos testes do sistema de Recuperação de Informação da Biblioteca Digital para Documentos Históricos.

Para solucionar tal carência, resolvemos utilizar a base de dados do projeto Pergunte a Pereira da Costa, abrigado no laboratório Liber, UFPE. O projeto publica em formato digital os 10 volumes dos "Anais Pernambucanos" num total de 4.257 páginas sobre a História de Pernambuco.

Nesta seção, iremos relatar e demonstrar os resultados alcançados com o desenvolvimento do módulo de busca, a partir das definições vistas no capítulo 3.

4.1.1. Aquisição dos documentos

Como já foi dito, a base de dados para o desenvolvimento do sistema de busca e indexação foi a do projeto Pergunte a Pereira da Costa. Ter encontrado uma base de dados textual para documentos históricos pronta foi uma grande vantagem para o início dos trabalhos. Contudo, a base de dados estava modelada de uma forma diferente da base de dados adotada para o presente trabalho (veja na figura 3.5 como está modelada a base de dados para Bibliotecas Digitais para Documentos Históricos).

A base de dados do projeto Pergunte a Pereira da Costa estava modelada de uma forma bastante diferente da proposta pelo projeto. Basicamente, só uma tabela armazena os dados que caracterizavam um documento:

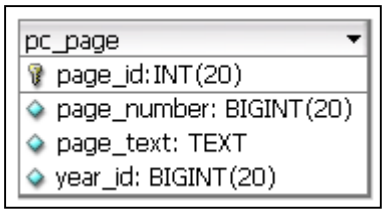


Figura 4.2: Tabela do Pereira da Costa que armazena o conteúdo dos Anais Pernambucanos

A idéia é que o sistema de busca retornasse o documento, que conteria algumas páginas. Se mantivéssemos a modelagem atual, não poderíamos realizar a busca nos atributos comuns a cada documento, como título, autor, resumo (veja quadro 3.1). Com isso, remodelamos a base de Pereira da Costa para a base proposta para a Biblioteca Digital. A grande pergunta seria como definir quais seriam os documentos. A atitude mais aceitável seria colocar cada documento como sendo um volume da série "Os Anais Pernambucanos". Contudo, seriam poucos documentos (apenas 10) e número de páginas seria muito grande, tornando praticamente impossível a verificação do desenvolvimento do sistema estava sendo realizado de forma correta. A opção mais viável foi dividir os documentos por anos. De acordo com a figura 4.2, podemos notar que cada página está referenciando um ano. Ao todo, são 310 anos que foram documentados pelo historiador Pereira da Costa. Com isso, a modelagem do banco de dados que iria servir para o desenvolvimento do sistema de indexação e recuperação teria 310 documentos, com diversas páginas cada um.

Como só havíamos os conteúdos textuais dos documentos, não os dados de seus atributos (título, autor, resumo e outros) só podemos realizar a consulta nos termos presentes em seu conteúdo. Contudo, o algoritmo de busca foi criado para aceitar os atributos da documentação histórica e o mesmo será demonstrada na seção 4.2.

Com a base criada e modelada, atingimos a fase da preparação dos documentos.

4.1.2. Preparação dos documentos

Como já foi dito anteriormente, foram utilizados duas fases na preparação dos documentos, a análise léxica e a eliminação das stop-words.

A eliminação das stop-words aconteceu em duas fases: primeiro, foi criado um script para percorrer toda a base textual e identificar as palavras que apareceriam mais de 500 vezes na base. A maioria dos termos que foram listados nessa primeira etapa foram artigos, preposições e algumas palavras que aparecem diversas vezes na base, como Recife e Pereira. Após essa amostragem, foi feita a avaliação manual dos termos, decidindo quem estaria no conjunto dos termos stop-words. Abaixo segue todos os termos que foram classificados como stop-words, na figura 4.3.

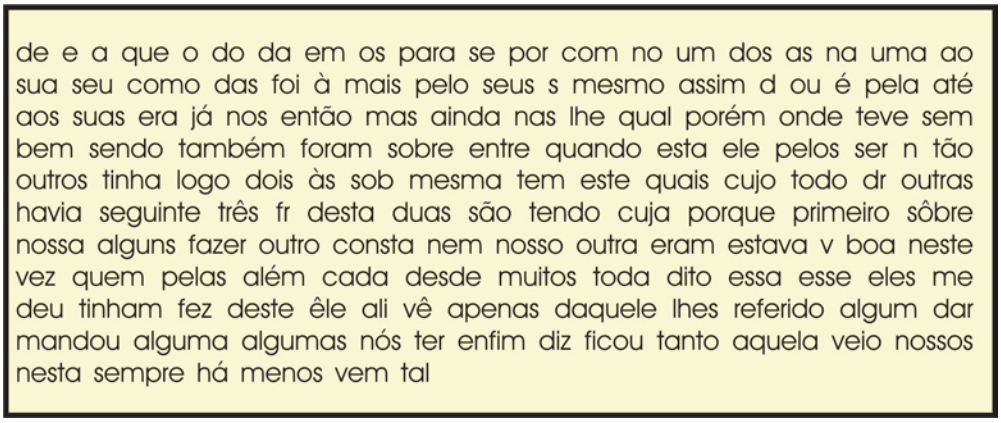

de e a que o do da em os para se por com no um dos as na uma ao sua seu como das foi à mais pelo seus s mesmo assim d ou é pela até aos suas era já nos então mas ainda nas lhe qual porém onde teve sem bem sendo também foram sobre entre quando esta ele pelos ser n tão outros tinha logo dois às sob mesma tem este quais cujo todo dr outras havia seguinte três fr desta duas são tendo cuja porque primeiro sôbre nossa alguns fazer outro consta nem nosso outra eram estava v boa neste vez quem pelas além cada desde muitos toda dito essa esse eles me deu tinham fez deste êle ali vê apenas daquele lhes referido algum dar mandou alguma algumas nós ter enfim diz ficou tanto aquela veio nossos nesta sempre há menos vem tal

Figura 4.3: Os 148 termos classificados como stop-words

Este processo de identificação das stop-words foi feito juntamente com a análise léxic, eliminando possíveis pontuações.

4.1.3. Criação da base de índices

Após essa primeira etapa de identificação das stop-words e da análise léxica, foi criado um arquivo de índices invertidos. A estrutura do mesmo, já foi mostrada, e pode

ser visto na figura 3.9. Essa estrutura armazena a freqüência do termo na base de dados. Um outro arquivo foi criado, o chamado Arquivo Invertido com TF-IDF. Este arquivo armazena a referência de cada termo, indicando em que documento o termo aparece, a freqüência do mesmo no documento e a sua relevância no documento. Como já foi dito anteriormente, esta relevância é calculada no momento em que este arquivo de índices invertidos está sendo preenchido. Veja a tabela 3.5 para rever qual foi a definição da pontuação. Esses dois arquivos foram estruturados em um banco de dados MySQL (detalhes sobre a escolha da tecnologia utilizada serão vistos na seção 3.1).

Desta forma, esta etapa foi subdividida em duas fases. Na primeira, foi criado um script em PHP para percorrer a base textual, e, obedecendo a lista de stop-words, armazenar todos os termos e suas freqüências no primeiro arquivo invertido.

Após este primeiro arquivo de índices invertidos ter sido criado, foi feito um segundo script que referenciava um termo no documento, calculando a sua freqüência e a relevância.

Com os dois arquivos invertidos criados, agora passamos para a fase da recuperação dos documentos.

4.1.4. Recuperação de Documentos

Realizando a consulta através dos arquivos invertidos, retiramos a necessidade da consulta através da base textual completa. Desta forma, e, armazenando a relevância do termo na base, o processamento da ordenação dos resultados, há um ganho de performance no processamento da busca. Como foi dito anteriormente, a consulta também permite realizar um casamento de padrões com a palavra buscada pelo usuário.

Nesta fase voltamos a relatar com mais detalhes a respeito do problema das diferentes modelagens entre a base de Pereira da Costa e a base criada para abrigar qualquer documentação histórica. Como foi relatado na seção 3.2., foi criado uma base muito pequena, com apenas 10 documentos, seguindo a modelagem da base de dados para Documentos Históricos. Alguns testes foram feitos nela e os resultados foram ordenados de maneira correta. Obviamente, seria precipitado tirar alguma conclusão com um corpus tão reduzido. Com isso, utilizamos a base completa de Pereira da Costa (os 310 documentos, num total de 10 volumes sobre a história de Pernambuco), realizando as consultas em seu conteúdo. Os resultados obtidos foram satisfatórios. As próximas figuras mostram os passos para visualizar o conteúdo dos textos.

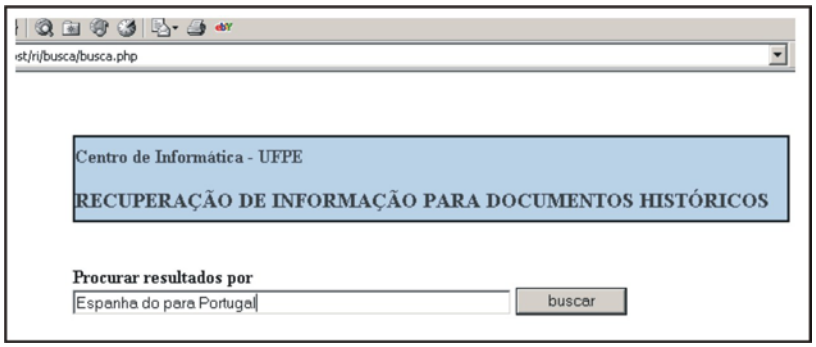


Figura 4.4: Uma busca no protótipo de RI para documentos históricos.

A figura 4.4 mostra a busca pelas palavras-chave "Espanha do para Portugal".

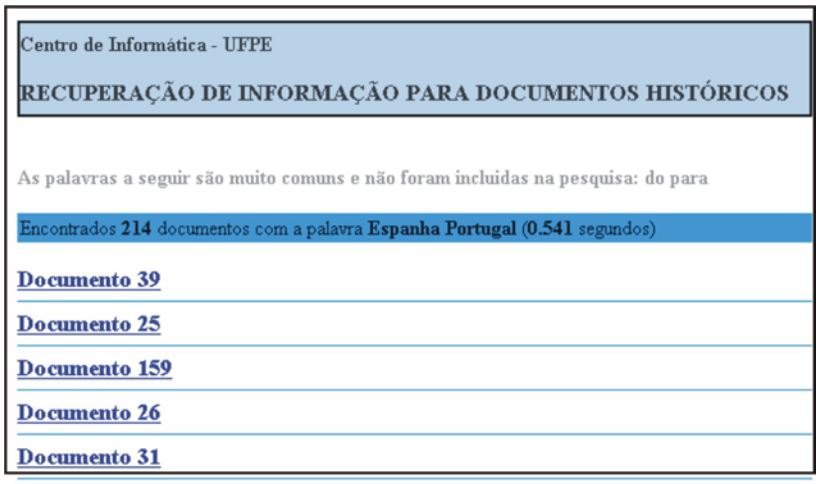
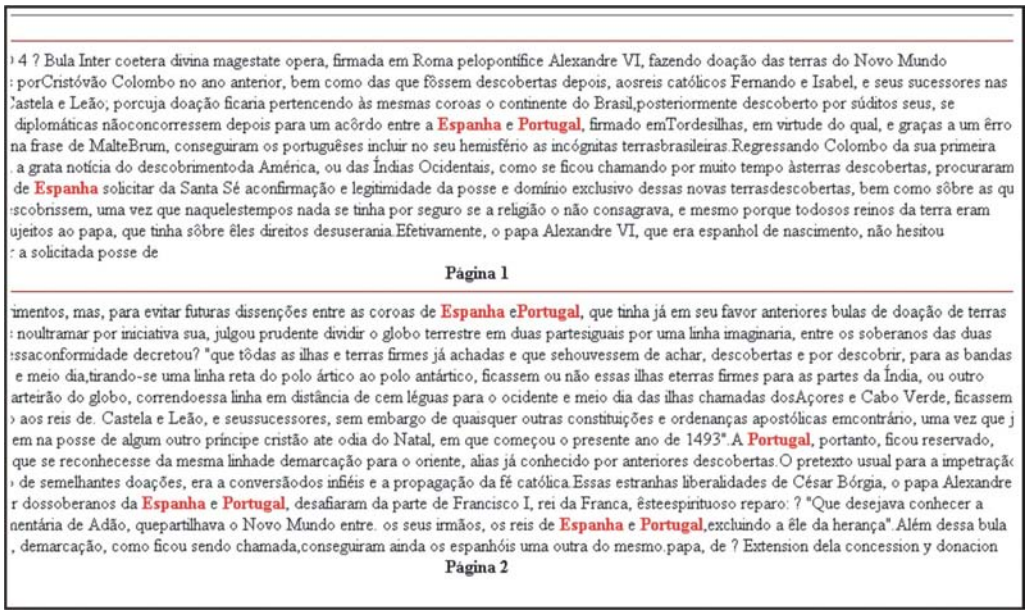


Figura 4.5: O Resultado ordenado da busca.

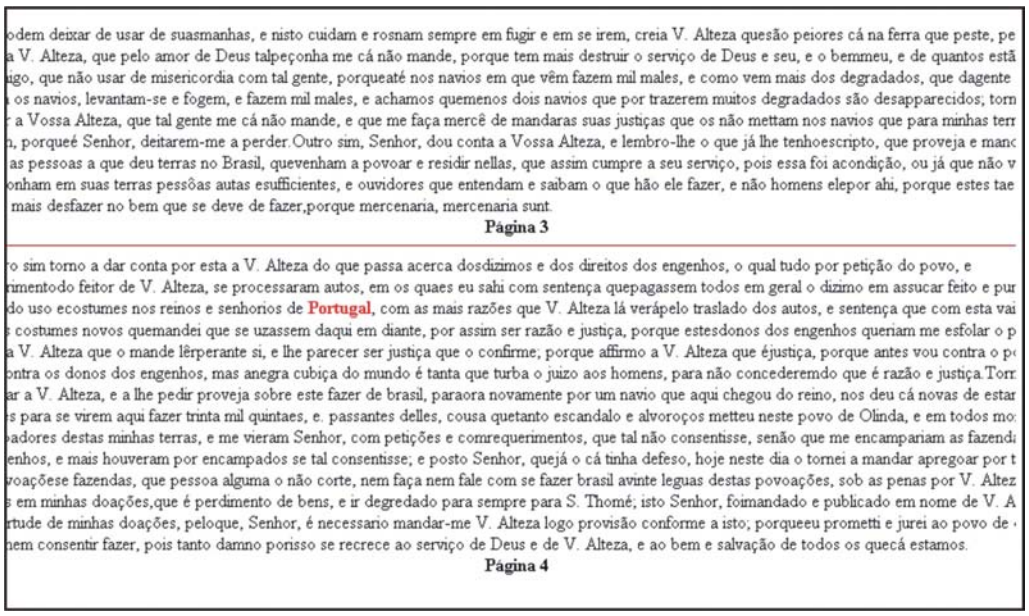
A figura 4.5 demonstra o resulta da busca. Note que os temos "do" e "para" não foram incluídos na pesquisa por serem classificados como stop-words. O tempo de resposta da busca também é mostrado.


) 4 ? Bula Inter coetera divina magestate opera, firmada em Roma pelopontifice Alexandre VI, fazendo doação das terras do Novo Mundo
: porCristóvão Colombo no ano anterior, bem como das que fôssem descobertas depois, aosreis católicos Fernando e Isabel, e seus sucessores nas
Castela e Leão; porcuja doação ficaria pertencendo às mesmas coroas o continente do Brasil,posteriormente descoberto por súditos seus, se
diplomáticas nãoconcorressem depois para um acôrdo entre a Espanha e Portugal, firmado emTordesilhas, em virtude do qual, e graças a um êrro
na frase de MalteBrum, conseguiram os portuguêses incluir no seu hemisfério as incógnitas terrasbrasileiras.Regressando Colombo da sua primeira
. a grata notícia do descobrimentoda América, ou das Índias Ocidentais, como se ficou chamando por muito tempo àsterras descobertas, procuraram
de Espanha solicitar da Santa Sé aconfirmação e legitimidade da posse e domínio exclusivo dessas novas terrasdescobertas, bem como sôbre as qu
scobrissem, uma vez que naquelestempos nada se tinha por seguro se a religião o não consagrava, e mesmo porque todosos reinos da terra eram
ujeitos ao papa, que tinha sôbre êles direitos desuserania.Efetivamente, o papa Alexandre VI, que era espanhol de nascimento, não hesitou
a solicitada posse de

Página 1

imentos, mas, para evitar futuras dissenções entre as coroas de Espanha ePortugal, que tinha já em seu favor anteriores bulas de doação de terras
: noultramar por iniciativa sua, julgou prudente dividir o globo terrestre em duas partesiguais por uma linha imaginaria, entre os soberanos das duas
ssaconformidade decretou? "que tôdas as ilhas e terras firmes já achadas e que sehouvessem de achar, descobertas e por descobrir, para as bandas
e meio dia,tirando-se uma linha reta do polo ártico ao polo antártico, ficassem ou não essas ilhas eterras firmes para as partes da Índia, ou outro
arteirão do globo, correndoessa linha em distância de cem léguas para o ocidente e meio dia das ilhas chamadas dosAçores e Cabo Verde, ficassem
aos reis de. Castela e Leão, e seussucessores, sem embargo de quaisquer outras constituições e ordenanças apostólicas emcontrário, uma vez que j
em na posse de algum outro príncipe cristão ate odia do Natal, em que começou o presente ano de 1493".A Portugal, portanto, ficou reservado,
que se reconhecesse da mesma linhade demarcação para o oriente, alias já conhecido por anteriores descobertas.O pretexto usual para a impetração
de semelhantes doações, era a conversãodos infiéis e a propagação da fé católica.Essas estranhas liberalidades de César Bórgia, o papa Alexandre
r dossoberanos da Espanha e Portugal, desafiaram da parte de Francisco I, rei da Franca, êsteespirituoso reparo: ? "Que desejava conhecer a
nentária de Adão, quepartilhava o Novo Mundo entre. os seus irmãos, os reis de Espanha e Portugal,excluindo a êle da herança".Além dessa bula
, demarcação, como ficou sendo chamada,conseguiram ainda os espanhóis uma outra do mesmo.papa, de ? Extension dela concession y donacion

Página 2


odem deixar de usar de suasmanhas, e nisto cuidam e rosnam sempre em fugir e em se irem, creia V. Alteza quesão peiores cá na ferra que peste, pe
a V. Alteza, que pelo amor de Deus talpeçonha me cá não mande, porque tem mais destruir o serviço de Deus e seu, e o bemmeu, e de quantos está
igo, que não usar de misericordia com tal gente, porqueaté nos navios em que vêm fazem mil males, e como vem mais dos degradados, que dagente
os navios, levantam-se e fogem, e fazem mil males, e achamos quemenos dois navios que por trazerem muitos degradados são desapparecidos; torn
a Vossa Alteza, que tal gente me cá não mande, e que me faça mercê de mandaras suas justiças que os não mettam nos navios que para minhas terr
h, porqueé Senhor, deitarem-me a perder.Outro sim, Senhor, dou conta a Vossa Alteza, e lembro-lhe o que já lhe tenhoescripto, que proveja e manc
as pessoas a que deu terras no Brasil, quevenham a povoar e residir nellas, que assim cumpre a seu serviço, pois essa foi acondição, ou já que não v
onham em suas terras pessôas autas esufficientes, e ouvidores que entendam e saibam o que hão ele fazer, e não homens elepor ahi, porque estes tae
mais desfazer no bem que se deve de fazer,porque mercenaria, mercenaria sunt.

Página 3

o sim torno a dar conta por esta a V. Alteza do que passa acerca dosdizimos e dos direitos dos engenhos, o qual tudo por petição do povo, e
rimentodo feitor de V. Alteza, se processaram autos, em os quaes eu sahi com sentença quepagassem todos em geral o dizimo em assucar feito e pur
do uso ecostumes nos reinos e senhorios de Portugal, com as mais razões que V. Alteza lá verápelo traslado dos autos, e sentença que com esta vai
costumes novos quemandei que se uzassem daqui em diante, por assim ser razão e justiça, porque estesdonos dos engenhos queriam me esfolar o p
a V. Alteza que o mande lêrperante si, e lhe parecer ser justiça que o confirme, porque affirmo a V. Alteza que éjustiça, porque antes vou contra o p
ontra os donos dos engenhos, mas anegra cubiça do mundo é tanta que turba o juizo aos homens, para não concederemdo que é razão e justiça.Torr
ar a V. Alteza, e a lhe pedir proveja sobre este fazer de brasil, paraora novamente por um navio que aqui chegou do reino, nos deu cá novas de estar
s para se virem aqui fazer trinta mil quintaes, e. passantes delles, cousa quetanto escandalo e alvoroços metteu neste povo de Olinda, e em todos mo
adores destas minhas terras, e me vieram Senhor, com petições e comrequerimentos, que tal não consentisse, senão que me encampariam as fazendi
enhos, e mais houveram por encampados se tal consentisse; e posto Senhor, quejá o cá tinha defeso, hoje neste dia o tornei a mandar apregoar por t
voaçõese fazendas, que pessoa alguma o não corte, nem faça nem fale com se fazer brasil avinte leguas destas povoações, sob as penas por V. Altez
em minhas doações,que é perdimento de bens, e ir degredado para sempre para S. Thomé; isto Senhor, foimandado e publicado em nome de V. A
rtude de minhas doações, peloque, Senhor, é necessario mandar-me V. Alteza logo provisão conforme a isto; porqueeu prometti e jurei ao povo de
hem consentir fazer, pois tanto damno porisso se recrece ao serviço de Deus e de V. Alteza, e ao bem e salvação de todos os quecá estamos.

Página 4

Figura 4.6: Visualização do documento 39.

A figura 4.6 demonstra o documento 39, o primeiro da ordenação. As palavras pesquisadas estão destacadas em vermelho. Podemos verificar a grande ocorrência das palavras consultadas pelo usuário.

Figura 4.7: Um dos últimos documentos retornados.

A figura 4.7 nos traz a visualização de um dos últimos documentos retornados na busca. Percebemos que só há ocorrência de um dos termos procurados uma única vez, indicando a baixa relevância desse termo no documento.

4.1.5. Testes Realizados

Indexação

Nos testes realizados o sistema foi instruído a indexar um total de 10 volumes, em formato texto, da série "Anais Pernambucanos", do historiador Pereira da Costa, em um total de 4257 páginas. A primeira base dos arquivos invertidos, que contém o termo e a freqüência em que ele ocorre no documento, obteve as seguintes características, a partir da análise do log:

- Quantidade de Páginas percorridas: 4.257;
- Total de Termos indexados: 76.458;
- Tempo total de Indexação desta base: 318.13 segundos;

A outra base de arquivos invertidos com TD-IDF foi mais custosa para ser feita. A razão é que cada termo foi indexado e referenciado ao um documento. Evidentemente, existem termos que aparecem em vários documentos. Além disso, além de referenciar cada termo, o script criado para essa indexação ainda calculava a relevância do termo no documento. Esta operação foi um pouco demorada, mas o ganho seria notado no momento da busca, pelo fato de já termos os pesos calculados. A partir da análise do log desta indexação, foram obtidos alguns dados.

- Quantidade de Páginas percorridas: 4.257;
- Total de Termos indexados, com seus respectivos pesos: 391.999;
- Tempo total de Indexação desta base: 1002.54 segundos;

Buscas

Foram realizadas 50 consultas ao sistema, com o número de palavras buscadas variando entre 1 e 10. O principal objetivo deste teste foi avaliar o tempo de resposta do módulo de busca. O tempo médio para as busca girou em torno de 0.50 segundos, o que se mostrou bastante satisfatório.

4.2. Estudo de Caso 2: Memórias do Golpe – O Brasil de 64 a 85

Para a Biblioteca Digital para Documentos Históricos, teríamos que definir algum escopo de documentação histórica para apresentar no presente trabalho. Vale salientar o apoio que a Fundação Gilberto Freyre nos proporcionou, colocando a disposição alguns de seus acervos para fazer parte do projeto. Os primeiros documentos históricos, contudo, que foram avaliados para entrar em nossa base de dados foi os que mostravam a época da ditadura militar no Brasil. Além de atingir um imenso domínio da população interessada, esses acervos específicos eram um dos poucos que poderíamos recolher arquivos das mídias que serão trabalhadas em nossa Biblioteca Digital: áudio, texto, imagem e vídeo.

O segundo estudo de caso iremos tratar do projeto principal do presente trabalho: a implementação da Biblioteca Digital para Documentos Históricos. Além de demonstrarmos como foi implementado o módulo de Recuperação de Informação desenvolvido a partir da base de dados de Pereira da Costa, explicaremos como foi realizado o desenvolvimento de diversos outros artefatos.

A grande dificuldade em ter acesso a esses acervos documentais era um dos agravantes em continuarmos com essa idéia. Diversos contatos foram produzidos e duas importantes parcerias foram firmadas. Na primeira delas, o jornalista pernambucano Samarone Lima disponibilizou para o presente trabalho centenas de entrevistas que tinha realizado com militantes que sofreram a época da repressão. A segunda parceria firmada foi com o Arquivo Público de Pernambuco. Através da diretora dos arquivos do DOPS (Departamento de Ordem Política e Social), Marcília Gama, conseguimos acesso a esse acervo que, ultimamente, está gerando diversas discussões na mídia nacional sobre a publicação do mesmo. Com o presente trabalho, será a primeira vez no Brasil que esses arquivos serão publicados na Internet.

O presente trabalho deu origem a um projeto ainda maior, que será abrigado pelo laboratório Liber, denominado: “Memórias do Golpe: O Brasil de 64 a 85”. A figura 4.8 mostra a entrada da Biblioteca Digital criada.

Figura 4.8: Página de entrada do sistema.

4.2.1. O módulo de Busca

O módulo de busca da informação criado foi explicado na seção 3.8 e exemplificado na seção 4.1. Com um sistema desenvolvido e testado, adaptamos o que foi feito com a base de dados do projeto Pergunte a Pereira da Costa para o sistema em questão.

A aquisição dos documentos na base de dados desse segundo estudo de caso foi feita totalmente manual. Recortes de jornais da época foram adquiridos. Os mesmo foram digitalizados e indexados. Os prontuários do DOPS, também presentes no nosso trabalho, também foi digitalizado e indexado por especialistas da área, como bibliotecários e historiadores. Os arquivos de áudio, presentes em fitas K-7, foram digitalizados e transformados para o formato mp3. Depois disso, foi feita a indexação dos mesmos na base.

Assim como na seção 4.1, foi realizada uma análise léxica automática, retirando as pontuações. Como a quantidade de documentos nessa base ainda é pequena, a seleção

dos stop-words foi realizada apenas com conjunções, pronomes e artigos. A partir daí, a base de índices criada foi da mesma maneira como descrita na seções 3.8 e 4.1.

A consulta da seção 4.1 era realizada em mídias do tipo texto, porque só tínhamos esse tipo disponível. Contudo, conforme descrito, o módulo de RI estava preparado para aceitar a busca de qualquer mídia. E assim foi implementado. As figuras 4.9 e 4.10 mostram a busca e o seu resultado.

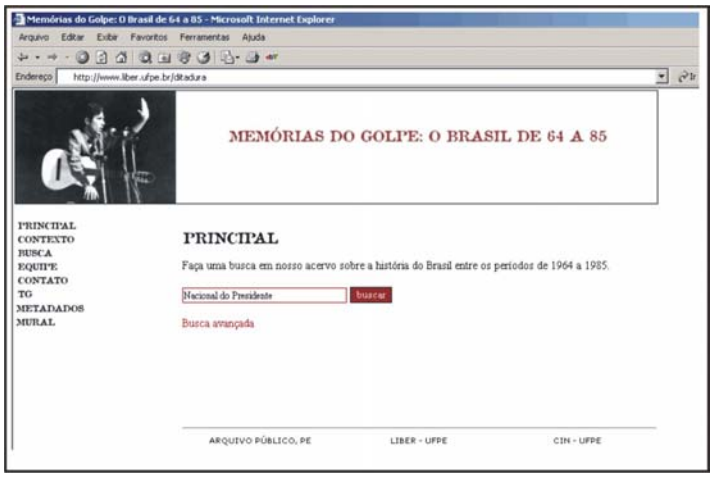


Figura 4.9: Busca por um Documento Histórico

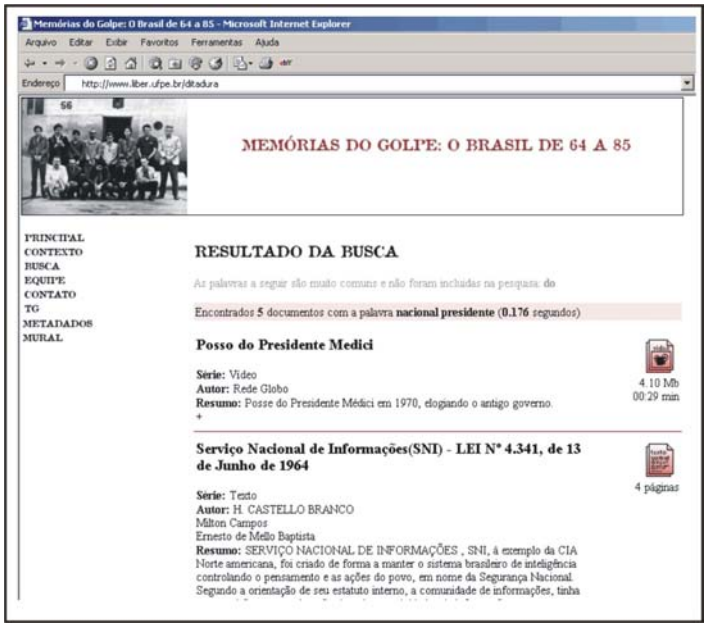


Figura 4.10: Resultado da busca pelas palavras “Nacional”, “do” e “Presidente”

Para este projeto específico, foi criado um módulo para busca avançada. O usuário pode especializar a sua consulta, por projetos específicos contidos na base de dados, por mídias específicas (texto, áudio, vídeo ou texto) e com algumas peculiaridades a mais, como pode ser visto figura 4.11.

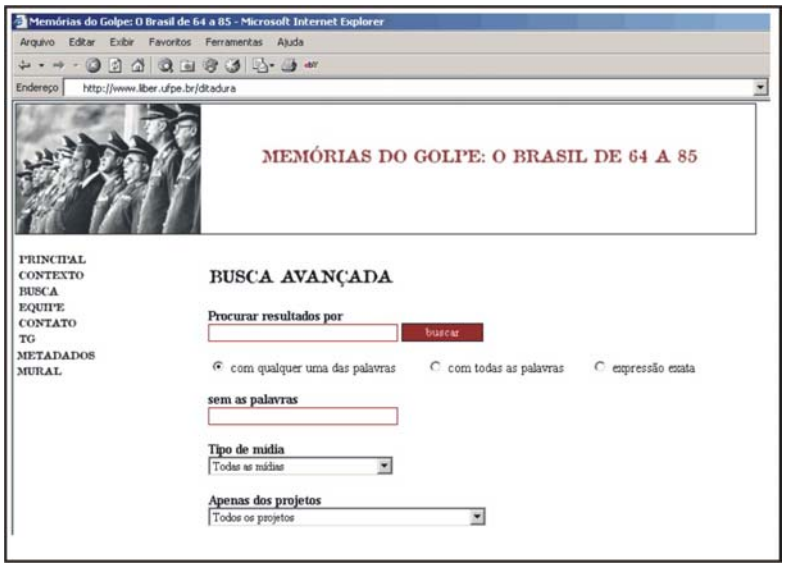


Figura 4.11: Busca Avançada

Algumas vezes o usuário deseja listar todos os documentos de um determinado projeto, onde cada um possui documentos específicos relativos a época do regime militar. Esta opção também encontra-se disponível e o este estudo de caso iniciou os seus trabalhos com três projetos:

- **Memórias do Golpe: O Brasil de 64 a 85**

O projeto é o mesmo do Estudo de Caso. Possui diversas notícias de jornais da época, matérias dos jornais atuais sobre alguns fatos ocorridos na época e alguns documentos jurídicos. Alguns documentos também são coletados na Web, em sua maior parte, através do Wikipedia [WIKI].

- **Resgate das vozes da resistência nos anos de chumbo**

O segundo projeto são centenas de entrevistas realizadas pelo jornalista pernambucano Samorone Lima com militantes que sofrerão a época da repressão. Tais arquivos estão

em formato áudio. Uma dificuldade para a publicação dessas entrevistas é a autorização das mesmas. Aos poucos, elas estão sendo adquiridas.

- **Arquivo DOPS, Pernambuco**

Através de uma parceria com o Arquivo Público de Pernambuco, estamos disponibilizando os prontuários do DOPS (Departamento de Ordem Política e Social). Será a primeira vez que esses arquivos serão publicados no Brasil, via Internet. A grande dificuldade no momento, assim como o segundo projeto, é a autorização das pessoas que estão nos prontuários. Algumas autorizações já foram conseguidas e estamos providenciando a dos demais.

4.2.2. O Visualizador do Documento Histórico

Outra funcionalidade bastante importante no sistema é a visualização do documento histórico. É nela que o usuário irá poder realizar diversas manipulações no documento (veja quadro 3.3), salvá-lo, podendo ainda inserir e ler notas sobre o documento em questão.

A idéia básica foi construir esse módulo baseado em diversas janelas móveis, tentando assim deixá-lo com uma usabilidade já consagrada em diversos programas e sistemas operacionais no mercado.

Todos os efeitos encontrados neste módulo do sistema foram implementados utilizando DHTML e JavaScript. Estas tecnologias foram importantes aliadas para que a usabilidade e a interface agradassem o usuário.

Figura 4.12: Visualizador do Documento Histórico.

De acordo com a figura 4.12, o menu localizado à esquerda corresponde a todas as opções que o usuário pode realizar enquanto visualiza o documento. Maiores detalhes são encontrados na tabela 3.3. Vale salientar também que todas as janelas são moveis, ou seja, o usuário pode arrastá-las com o mouse e a capa está sendo mostrada na figurara 4.12 é gerada automaticamente pelo sistema.

Algumas das funcionalidades do visualizador de documentos serão mostradas nas figuras 4.13 e 4.14.

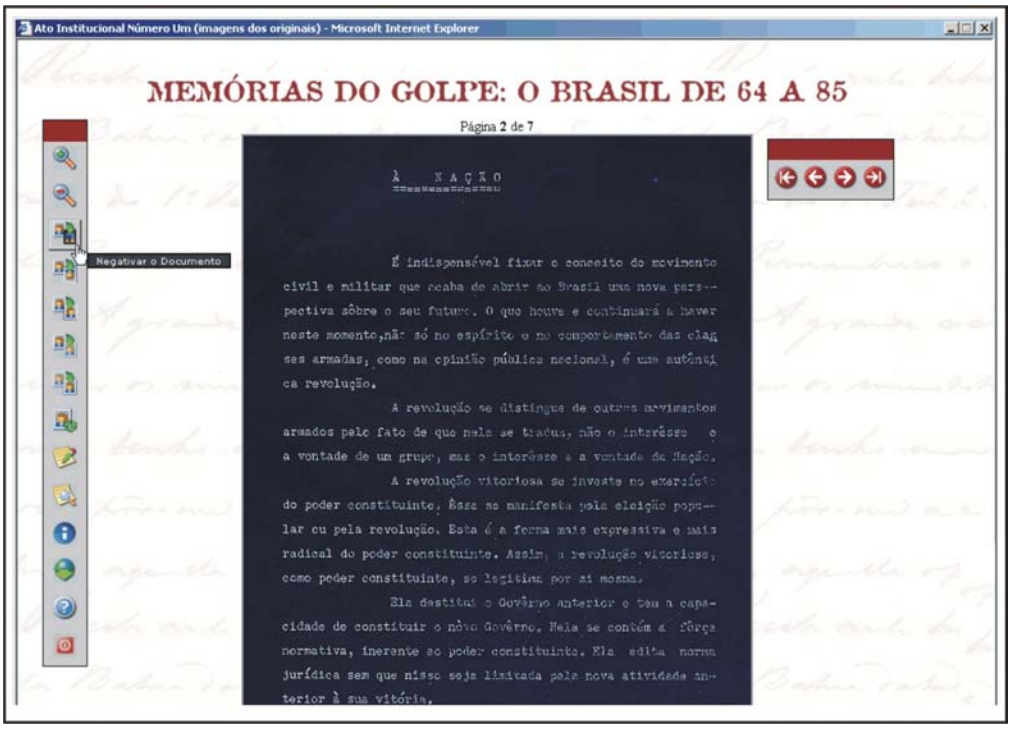


Figura 4.13: Documento Negativado

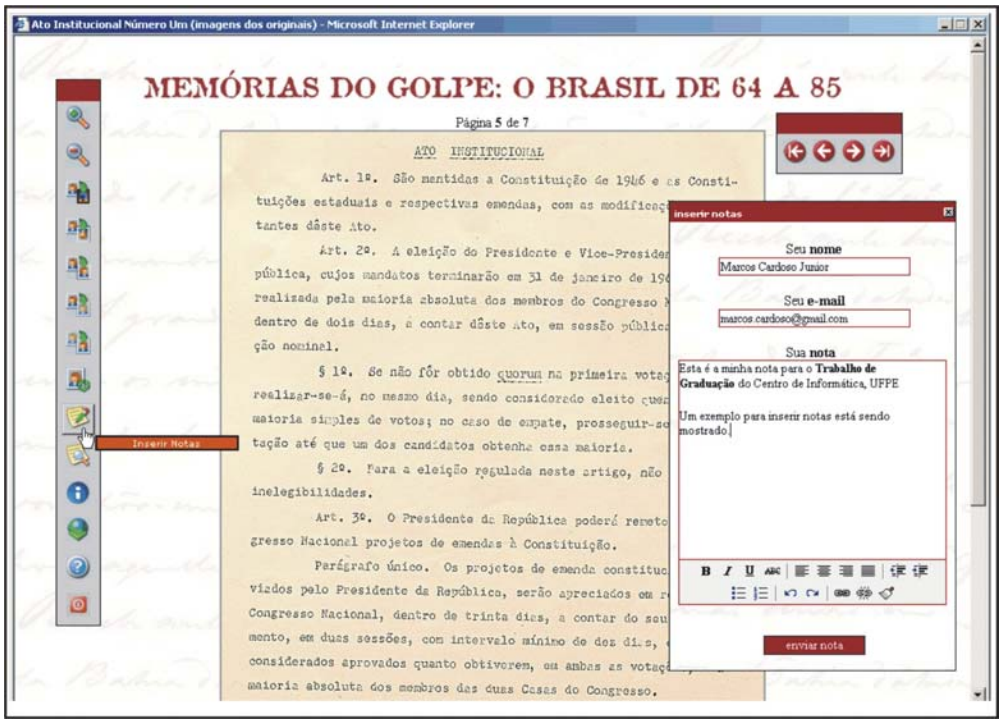


Figura 4.14: A opção de Inserir Notas sobre o documento.

A figura 4.13 demonstra um efeito de negativação do documento. Muitas vezes, apenas com esse efeito o documento histórico pode ser visualizado. Já a figura 4.14 nos mostra a inserção de notas sobre um documento. Atenção no mini-editor de texto, cuja finalidade é prover ao usuário uma confecção mais personalizada da sua nota, com a possibilidade de inserir termos em negrito, justificado, com links e outras características.

Na figura 4.15 iremos mostrar o visualizador com a demonstração da passagem de um vídeo.

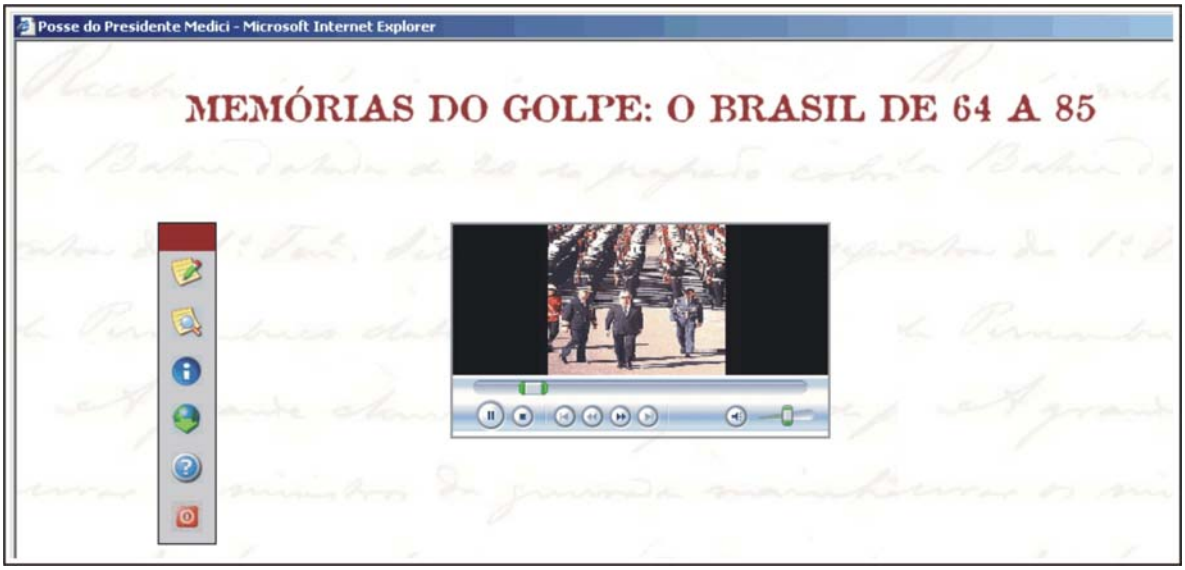


Figura 4.15: Visualizador disponibilizando um vídeo.

Iremos finalizar essa seção demonstrando o visualizador para documentos históricos do tipo texto. O interessante deste módulo é a chamada "referência cruzada" que o mesmo pode fazer. Passando o mouse por cima de qualquer palavra do texto, pode-se fazer uma pesquisa com o termo selecionado com apenas um click, de acordo com a figura 4.16.

Figura 4.16: Visualizador da mídia Texto. No exemplo, a busca pelo termo "Presidente"

4.2.3. A disponibilização do acervo para outras instituições

Com já foi mencionado, dá-se no formato Dublin Core, utilizando o padrão XML/RDF. Nas figuras 4.17 e 4.18 mostra como disponibilizamos esse material. O usuário realiza uma busca e o resultado das informações é mostrado no formato especificado.

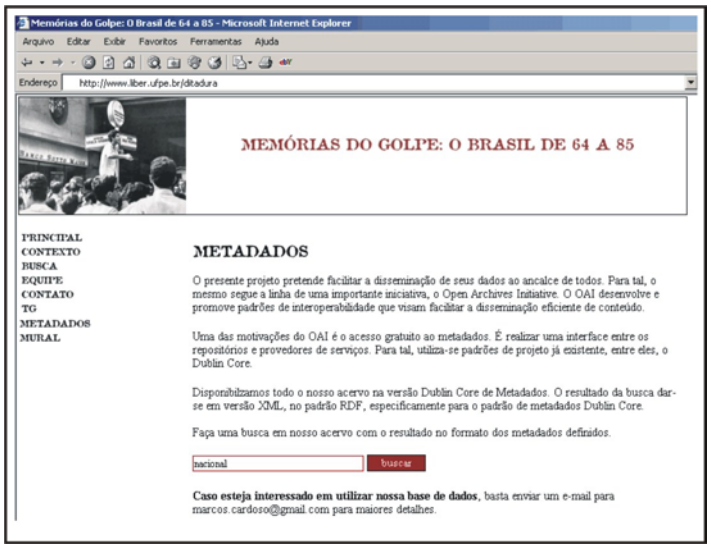


Figura 4.17: Consulta da Base para retornar os metadados

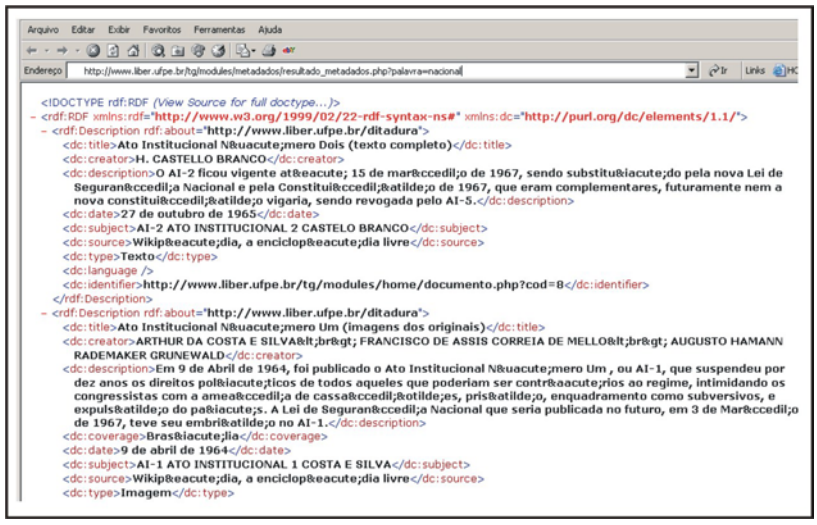


Figura 4.18: Resultado da busca pela palavra "nacional" com os metadados disponíveis no padrão Dublin Core e formato XML/RDF.

4.2.4. O Sistema de Administração

Todos as funcionalidades do sistema de administração foram explicadas na seção 3.7 e agora iremos demonstrar algumas dessas especificações criadas na Biblioteca Digital para Documentos Históricos.

- **Logar no Sistema**

A figura 4.19 mostra a tela de abertura da administração do sistema.

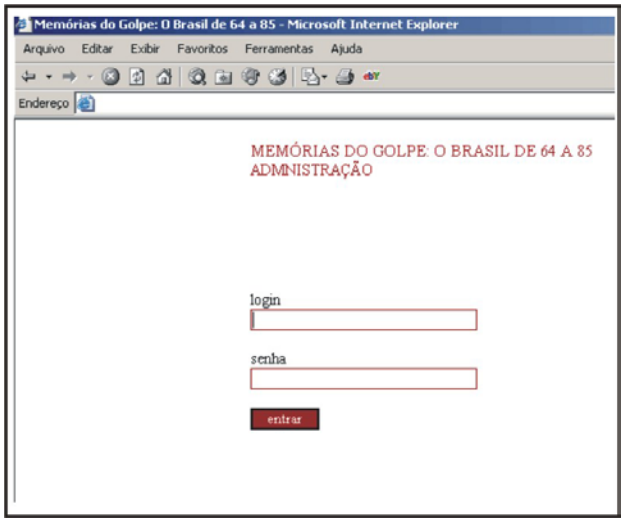


Figura 4.19: Tela de abertura do sistema de administração

- **Cadastrar Documento**

O cadastro de documento inclui as atividades de inserção, alteração e exclusão do documento. A figura 4.20 exemplifica as duas etapas do procedimento de inserção de um documento histórico do tipo imagem. Na primeira, são inseridos os dados descritos no quadro 3.2. Na segunda, são inseridas as páginas do documento histórico de forma bastante simples.

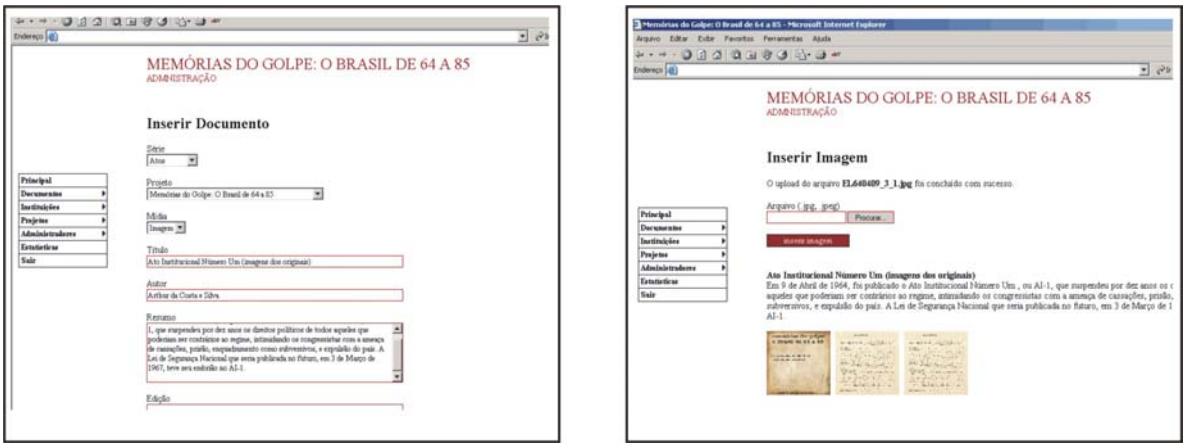


Figura 4.20: As duas etapas para inserir um documento na base.

5. Conclusão e Trabalhos Futuros

Neste trabalho, foi detalhado a construção de uma Biblioteca Digital para Documentos Históricos. Foram apresentadas as principais carências intrínsecas às bibliotecas digitais, relatando as possíveis soluções.

O objetivo principal deste trabalho é estudar e aplicar o desenvolvimento de técnicas que propiciem que uma Biblioteca Digital forneça diversos serviços interessantes para que os usuários possam usufruir da melhor forma possível suas informações. Analisando uma carência ainda maior, atacamos com detalhes o escopo das Bibliotecas Digitais para Acervos Históricos.

Para que fosse possível a implementação dessas técnicas, muitos conceitos tiveram que ser estudados.

Primeiramente, necessitava-se de um módulo de busca eficiente, que retornasse ao usuário informações relevantes a sua pesquisa. Foi realizado um estudo na área de Recuperação de Informação, para decidir qual seriam as melhores técnicas para atacar o escopo do projeto em questão. A implementação produzida foi avaliada contra um corpus com o seu desempenho computado. Consultas booleanas e do tipo Espaço Vetorial foram estudadas, decidindo-se por um modelo denominado Booleano Estendido. Este modelo foi testado e o seu desempenho foi acima do esperado, no que se diz respeito à performance e relevância dos resultados. Os termos da base textual foram indexados na base de arquivos invertidos juntos com a sua relevância no documento. Isso, apesar de deixar a indexação mais lenta, ajudou a tornar o desempenho do sistema eficiente.

Outro fator de grande relevância no presente trabalho foi a forma de disseminação da informação. Utilizando os preceitos da Iniciativa do Open Archives, juntamente com o padrão de metadados Dublin Core, a Biblioteca Digital para Documentos Históricos é também um servidor de dados, ou seja, outras instituições que queiram coletar nossos dados, podem assim fazê-lo facilmente, utilizando de algumas técnicas do OAI, como o protocolo OAI-PHM.

Ainda foram definidas diversas outras funcionalidades para o presente trabalho, todas elas concluídas. Dentre elas, o módulo de visualização do documento históricos. Diversas técnicas de DHTML foram utilizadas para que o módulo pudesse ter uma interface amigável e uma boa usabilidade para o usuário. O visualizador permite que o

visitante trabalhe com arquivos históricos do tipo texto, áudio, imagem ou vídeo. Com a publicação de fundos arquivísticos em meio digital, ter-se-á uma disponibilização dos documentos em larga escala e a sua virtual preservação.

Para o presente trabalho precisávamos de um tema para a demonstração dos resultados da Biblioteca Digital. Decidiu-se aplicar temas relativos a época da ditadura militar no Brasil. Nossa intenção não é julgar ou condenar, e sim mostrar alguns fatos que marcaram a época para que usuários possam realizar pesquisas, facilitando a produção de novos trabalhos sobre temáticas humanas e sociais tão variadas quanto importantes.

Como trabalhos futuros pode-se citar: (1) estudo e implementação de novas técnicas de RI para documentos históricos (por exemplo, alguns algoritmos hierárquicos); (2) construir um Thesaurus [Gonzales] de informações sobre a época da ditadura, afim de que os documentos possam ter uma relação entre eles; (3) realizar testes numa base multimídia maior de forma que se possam obter dados mais precisos acerca do desempenho dos algoritmos de busca.

Além dos trabalhos futuros citados acima, especificamente sobre o projeto Memórias do Golpe: O Brasil de 64 a 85, temos a intenção de cada vez mais unir diferentes projetos para que o mesmo seja uma referência para pesquisas sobre a época.

6. Referências Bibliográficas

[AMLC] América Memory from the Library of Congress. Disponível em <http://memory.loc.gov/ammem/>. Acesso em 04 mar. 2005.

[Barros, 2005] BARROS, F. Modelo de Recuperação de Documentos. Disponível em <http://www.cin.ufpe.br/~fab/cursos/ri>. Acesso em 02 mar. 2005.

[BIBDIGa] A Biblioteca Digital. Disponível em <http://www.cg.org.br/gt/gtbv/YorkatIBICT/index.htm>.Acesso em 09 fev. 2005.

[BIBDIGb] Bibliotecas Brasileiras na Internet, 2002. Disponível em <http://www.cg.org.br/gt/gtbv/bibliotecas.htm>. Acesso em 05 fev. 2005.

[BVL] Biblioteca Virtual de Literatura. Disponível em <http://www.biblio.com.br>. Acesso em 15 fev. 2005.

[CLIBPDF] ClibPDF functions. Disponível em <http://www2.stack.ru/ /PHP4/ref.cpdf.html >. Acesso em 16 fev. 2005.

[DIACIE] Diálogo Científico. Disponível em <http://dici.ibict.br/>. Acesso em 12 fev. 2005.

[DLF] Digital Libray Federantion. Disponível em <http://www.diglib.org/>. Aceso em: 05 fev. 2005.

[DCa] Dublin Core Metadada Iniciative. Disponível em <http://www.dublincore.org>. Acessado em 02 mar. 20005.

[DCb] Dublin Core Metadados, Guia de Uso. Disponível em <http://wwwbases.cnptia.embrapa.br/RuralMidia/guia.htm>. Acesso em 09 fev. 2005.

[DCc] An Introduction to Dublin Core. Disponível em < http://www.xml.com/pub/a/2000/10/25/dublincore/>. Acesso em 26 fev. 2005.

[DHTML] Dynamic Drive DHTML. Disponível em <http://www.dynamicdrive.com >. Acesso em 12 fev. 2005.

[EPRINTS] GNU EPrints Archive Software. Disponível em < http://software.eprints.org/>. Acesso em 12 de fev. 2005.

[EXT] Extreme Tracking. Disponível em <http://www.extreme-dm.com/tracking/>. Acesso em 14 jan. 2005.

[FGF] Fundação Gilberto Freyre. Disponível em <http://www.fgf.org.br>. Acesso em 02 fev. 2005

[Fico, 2004] FICO, Carlos. Além do Golpe. Versões e controvérsias sobre 1964 e Ditadura Militar. Record. 2004.

[FTS] MySQL Reference Manual - Full-Text Search Functions. Disponível em <http://dev.mysql.com/doc/mysql/en/fulltext-search.html>. Acesso em 28 jan. 2005.

[FPDF] FPDF Library – PDF Generator. Disponível em <http://www.fpdf.org/>. Acesso em 19 fev. 2005.

[FUNDAJ] Fundação Joaquim Nabuco. Disponível em <http://www.fundaj.gov.br>. Acesso em 02 fev. 2005

[Galindo, 2004] GALINDO, M; PEREIRA, M; VIEIRA, C. Bibliotecas Digitais e Metadados: Uma abordagem intregadora. 2004.

[GD] GD Graphics Library. Disponível em <http://www.boutell.com/gd/>. Acesso em 06 fev. 2005.

[Gonzales] GONZALES, M; STRUBE, L. Recuperação de Informação e Expansão Automática de Consulta com Thesaurus: uma avaliação.

[Griffin, 1995] S, Griffin. The Almaden Distributed Digital Library System. 1995.

[Ghiglione 2003] GHIGLIONE, E. SCORM in .LRN Implementation, 2003.

[Garcia 2003] GARCIA, P. Provedores de Dados de Baixo Custo: Publicação Digital ao Alcance de Todos, 2003.

[LOC] Library of Congress. Disponível em <http://www.loc.gov/>. Acesso em 26 fev. 2005.

[LIBER] Laboratório Liber, UFPE. Disponível em <http://www.liber.ufpe.br>. Acesso em 25 fev. 2005.

[Martins] MARTINS, J.; MOREIRA, E. Classificação de páginas na Internet.

[Milstead, 1999] MILSTEAD, J.; FELDMAN, S. Metadata: Cataloging by Any Other Name. On Line Magazine, 1999.

[MARC21] VOSGRAU, S. et al. Formato Marc 21 Holdings para publicações seriadas.

[Mosata] J. Mosata, Indiana Univ.

[METAPT] Introdução a Metadados na Biblioteca Nacional de Portugal. Disponível em <http://metadados.bn.pt/>. Acesso em 05 fev. 2005.

[Miller, 1998] MILLER, E; IANNELLA, R. Dublin Core Examples in RDF. 1998

[MYSQL] Página Oficial do MySQL. Disponível em <http://www.mysql.com>. Acesso em 19 fev. 2005.

[MHN] Museu Histórico Nacional. Disponível em <http://www.museuhistoriconacional.com.br>.

Acesso em 11 fev. 2005.

[Niedarauder, 2004] NIEDERAUER, J. PHP para quem sabe PHP. Novatec. 2004.

[OAIa] Página Oficial do Open Archives Iniciative. Disponível em <http//www.oai.org>. Acesso em 20 fev. 2005

[OAIb] Clube OAI Brasil. Disponível em <http://clube-oai.incubadora.fapesp.br/portal>. Acesso em 21 fev. 2005

[Pedrosa, 2001] PEDROSA, J. A Grande Barreira. Biblioteca do Exército Editora. Segunda Edição. 2001.

[PHP] Página Oficial do PHP. Disponível em <http://www.php.net>. Acesso em 19 fev. 2005.

[PDFLIB] A library for processing PDF. Disponível em <http://www.pdflib.com/>. Acesso em 11 fev. 2005.

[PC] Pergunte a Pereira da Costa. Disponível em <http://www.liber.ufpe.br/pc>. Acesso em 04 mar. 2005.

[Prudencio, 2003] PRUDÊNCIO, R. Bibliotecas Digitais: conceitos, histórico, planejamento, estruturação, operacionalização. 2003.

[RDF] RDF/XML Syntax Specification, W3C Recommendation 10 February 2004. Disponível em <http://www.w3.org/TR/rdf-syntax-grammar/>

[Rosetto] ROSETTO, M; NOGUEIRA, A. Aplicação dos Elementos Metadados Dublin Core para descrição de dados bibliográficos On-Line da Biblioteca Digital de Teses da USP.

[Ramos] RAMOS, J.Using TF-IDF to Determine Word Relevance in Document Queries.

[Souza 2004] SOUZA, R; ALVARENGA, L. A Web Semântica e

suas contribuições para a ciência da informação, 2004.

[Saffady 1995] W. Saffady. Digital Library Concepts and Technologies for the Management of Library Collections. 1995.

[SWA] Semantic Web Activity Statement. Disponível em <http://www.w3.org/2001/sw/Activity>. Acesso em 19 fev. 2005.

[Tronchin 1998] TRONCHIN, Valsoir, Análise, Modelagem e Implementação de Data Warehouses – São Paulo: Fenasoft/98 em 20/07/98.

[ULTRAMAR] Projeto Ultramar de Resgate da Documentação Histórica. Disponível em <http://www.liber.ufpe.br/ultramar>. Acesso em 19 fev. 2005.

[UNESCO] Manifesto para a Preservação Digital, UNESCO. Disponível em <http://www.bn.pt/agenda/ecpa/manifesto_unesco.html >. Acesso em 02 fev. 2005.

[W3CSWEB] W3C Semantic Web. Disponível em <http://www.w3.org/2001/sw/>. Acesso em 04 fev. 2005

[WIKI] Wikipedia. The Encyclopedia. Disponível em <http://en.wikipedia.org/wiki/Main_Page>. Acesso em 26 fev. 2005.

[Weibel, 1995] Weibel, S. Metadata: The foundations of resource description. D-Lib Magazine, July. 1995.

[XMLa] Extensible Markup Language (XML). Disponível em <http://www.w3.org/XML/>. Acesso em 15 fev. 2005.

[XMLb] XML em 10 pontos. Disponível em < http://paginas.terra.com.br/informatica/mja/W3C/XM

L-in-10-points.pt-BR.html>. Acesso em 05 fev. 2005.

[Yates, 1999] YATES, B.; NETO, R. Modern Information Retrieval. 1999

Assinaturas

_____

Flávia Almeida Barros

_____

Marcos Galindo Lima

_____

Marcos Cardoso Junior