

**UNIVERSIDADE FEDERAL DE SANTA CATARINA
PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO**

Jairo Wensing

**PRESERVAÇÃO E RECUPERAÇÃO DE INFORMAÇÃO EM
FONTES DE INFORMAÇÕES DIGITAIS: ESTUDO DE CASO
DO GREENSTONE**

Florianópolis

2010

Jairo Wensing

**PRESERVAÇÃO E RECUPERAÇÃO DE INFORMAÇÃO EM
FONTES DE INFORMAÇÕES DIGITAIS: ESTUDO DE CASO
DO GREENSTONE**

Dissertação de mestrado apresentada à Banca Examinadora do Programa de Pós-Graduação em Ciência da Informação do Centro de Ciências da Educação da Universidade Federal de Santa Catarina, como requisito parcial para a obtenção do título de Mestre em Ciência da Informação, área de concentração Gestão da Informação, linha de pesquisa Fluxos de Informação, sob a orientação da Professora Doutora Ursula Blattmann

Florianópolis

2010

Catálogo na fonte elaborada pela biblioteca da
Universidade Federal de Santa Catarina

W476p Wensing, Jairo

Preservação e recuperação de informação em fontes de informações digitais [dissertação] : estudo de caso do GREENSTONE / Jairo Wensing ; orientadora, Ursula Blattmann. - Florianópolis, SC, 2010.
1 v.: il., grafs., tabs.

Dissertação [mestrado] - Universidade Federal de Santa Catarina, Centro de Ciências da Educação. Programa de Pós-Graduação em Ciência da Informação.

Inclui referências

1. Ciência da informação. 2. Sistemas de recuperação da informação. 3. Biblioteca Digital Greenstone - Fontes de informação. I. Blattmann, Ursula. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Informação. III. Título.

CDU 02

Jairo Wensing

**PRESERVAÇÃO E RECUPERAÇÃO DE INFORMAÇÃO EM
FONTES DE INFORMAÇÕES DIGITAIS: ESTUDO DE CASO
DO GREENSTONE**

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Informação do Centro de Ciências da Educação da Universidade Federal de Santa Catarina em cumprimento a requisito parcial para a obtenção do título de Mestre em Ciência da Informação, área de concentração Gestão da Informação, linha de pesquisa Fluxos de Informação.

APROVADA PELA COMISSÃO EXAMINADORA
Em Florianópolis, de 12 de julho de 2010.

Prof^ª. Dr^ª. Lígia Maria Arruda Café
Coordenadora PGCIN/UFSC

Banca Examinadora:

Orientadora: Prof^ª. Dr^ª. Ursula Blattmann - PGCIN/UFSC
Orientadora

Prof^ª. Dr^ª. Delsi Fries Davok - UDESC
Examinadora

Prof^ª. Dr^ª. Rosângela Schwarz Rodrigues - PGCIN/UFSC
Examinadora

Dedico este trabalho aos meus pais pela vida, educação, exemplo e amor incondicional. Aos meus filhos, a quem amo incondicionalmente, pela compreensão pela minha ausência nos seus momentos de lazer durante a realização deste trabalho. A minha amada esposa Lúcia, por tudo o que ela é na minha vida, pela sua paciência e incentivo para concluir este trabalho.

AGRADECIMENTOS

À Deus, por sempre estar presente em todos os momentos decisivos de minha vida, e ao apoio de algumas pessoas que contribuíram direta ou indiretamente para a concretização deste trabalho. Por isso, meus sinceros agradecimentos para:

A Professora Dr.^a Úrsula Blattmann, por me acolher como orientadora deste trabalho, pela competência, dedicação, amizade, apoio e motivação, e por ser este exemplo de professora.

A Professora Dr.^a Lígia Maria Arruda Café pela competência com que administra o PGCIN, e pelo incentivo para a conclusão deste trabalho, e por sua participação na fase de qualificação e suas contribuições a este trabalho.

A Professora Dr.^a Rosângela Schwarz Rodrigues, e por sua participação na fase de qualificação e defesa da dissertação, e pelas e suas contribuições a este trabalho.

A Professora Dr.^a Delsi Fries Davok, e por sua participação na defesa da dissertação, e pelas e suas contribuições a este trabalho.

À Universidade Federal de Santa Catarina, especialmente ao Programa de Pós-Graduação em Ciência da Informação, aos colegas de turma, professores e técnicos administrativos do PGCIN.

As pessoas participantes da pesquisa, pela gentileza de fornecer as informações e dispor do tempo necessário para responder os questionários.

Aos amigos da UDESC pelo incentivo para conclusão deste trabalho.

“A preservação digital não envolve a retenção do objeto informacional em si, mas também do seu significado. Assim sendo, faz-se necessário que as técnicas de preservação sejam capazes de compreender e recriar a forma original ou a função do objeto de forma que sejam asseguradas sua autenticidade e acessibilidade.”

Sayão (2007, p. 117)

Wensing, Jairo. **Preservação e recuperação de informação em fontes de informações digitais: estudo de caso do *greenstone***. 2010. 219 p. Dissertação (Mestrado em Ciência da Informação) – Programa de Pós Graduação em Ciência da Informação, Universidade Federal de Santa Catarina, Florianópolis, 2010.

RESUMO

Esta dissertação contribui para a Ciência da Informação, pois aprofunda e realiza estudos na área de fontes de informação, bibliotecas digitais, recuperação da informação e preservação digital tendo como foco o formato de arquivos digitais. O estudo teve como objetivo principal analisar os recursos disponíveis na Biblioteca Digital *Greenstone* (BDG) para preservação lógica de documentos digitais com foco no formato de arquivos e a recuperação de informação. Para atingir os objetivos específicos, adotou-se uma metodologia baseada na análise exploratória e descritiva do tema. Para análise, foi instalado o *software Greenstone 3.04*, e criada a coleção PGCIN que contempla as dissertações do programa de Mestrado em Ciência da Informação da UFSC. Após a realização da pesquisa, concluiu-se que a Biblioteca Digital *Greenstone* está preparada para atender aos requisitos de preservação lógica de arquivos digitais, pois aceita formatos de arquivos proprietários com especificação fechada, proprietários com especificação aberta, e não proprietários com especificação aberta, além de ser uma plataforma aberta e que permite o desenvolvimento de plug-ins para formatos de arquivos.

Palavras-chaves: Fontes de Informação. Biblioteca Digital *Greenstone*. Preservação Digital. Recuperação de Informação

Wensing, Jairo. **Preservation and retrieval of information from digital information Sources: a case study of *greenstone***. 2010. 219 p. Thesis (Master in Information Science) – Information Science Post-graduate Program, Santa Catarina Federal University, Florianópolis, 2010.

ABSTRACT

This dissertation contributes to the field of Information Science, deepening and conducting a study in the area of information sources, digital libraries, information retrieval and digital preservation with a focus on the format of digital files. This study aims to analyze the main features available in *Greenstone* Digital Library (GDL) for the logic preservation of digital documents focusing on file format and information retrieval. To achieve the desired goals, an exploratory methodology based on descriptive analysis of the subject was used. For analysis, *Greenstone* 3.04 was installed, and a collection that includes PGCIN and the Master's theses of the program in Information Science of UFSC was created. Upon completion of the research it was concluded that the *Greenstone* Digital Library is prepared to meet the requirements of logic preservation of digital files, because accepted proprietary file formats with file formats specifying closed owners, owners with open specification, and open specification with non-owners besides being an open platform that enables the development of plug-ins for file formats.

Keywords: Information Sources. *Greenstone* Digital Library. Digital Preservation. Information Retrieval.

Wensing, Jairo. **Conservación y recuperación de información sobre fuentes de información digital: un estudio de caso de *greenstone***. 2010. 219 p. Disertación (Máster en Ciencia de la Información) - Programa de Postgrado en Ciencia de la Información - Universidade Federal de Santa Catarina, 2010.

RESUMEN

Esta disertación contribuye a la Ciencia de la Información acentuando a medida que se lleva a cabo el estudio en el área de fuentes de información, bibliotecas digitales, recuperación de la información y la conservación digital con un enfoque en el formato de archivos digitales. El estudio tiene como objetivo analizar las principales características disponibles en la Biblioteca Digital Greenstone (BDG) para la preservación lógica de documentos digitales con un enfoque en formato de archivo y recuperación de información. Para lograr los objetivos deseados fué adoptada una metodología basada en el análisis exploratório y descriptivo de los sujetos, con la ejecución de un estudio de caso. *Software* Greenstone 3.04 fué instalado para el análisis y una colección que incluye las disertaciones PGCIN del programa de Magister en Ciencias de la Información, UFSC. Una vez terminada la investigación se concluyó que la Biblioteca Digital Greenstone está dispuesta a cumplir los requisitos de preservación lógica de archivos digitales, y por lo tanto acepta formatos de archivos con formatos de archivo propietarios cerrados, propietario abierto y no propietario con especificación abierta. Además es una plataforma abierta que permite el desarrollo de plug-ins para formatos de archivo.

Palabras claves: Fuentes de Información. Greenstone Digital Library. Preservación Digital. Recuperación de Información.

LISTA DE FIGURAS

| | |
|---|-----|
| Figura 1: 1990 <i>Windows Explorer</i> do <i>Windows Seven</i> da <i>Microsoft</i> | 41 |
| Figura 2: 1990 - O marco do <i>PostScript</i> | 52 |
| Figura 3: 1991- Projeto <i>Camelot</i> | 53 |
| Figura 4: 1992 - divulgado o formato <i>PDF</i> | 53 |
| Figura 5: 1993 - Inicia-se a geração <i>Acrobat</i> | 54 |
| Figura 6: 1994 - Lançado a versão do <i>Acrobat 2.0</i> | 54 |
| Figura 7: 1994 - <i>PDF</i> na Receita Federal USA..... | 54 |
| Figura 8: 1994 - É lançado o <i>Acrobat Reader</i> - leitor gratuito para <i>PDF</i> | 55 |
| Figura 9: 1995 - plug-in para o <i>Netscape</i> | 55 |
| Figura 10: 1996 - <i>Acrobat 3.0</i> | 55 |
| Figura 11: 1997 – Uso do byte do duplo..... | 56 |
| Figura 12: 1999 - Novos recursos de segurança para o formato <i>PDF</i> | 56 |
| Figura 13: 1999 - <i>ANSI</i> publica padrão <i>PDF</i> | 56 |
| Figura 14: 2000 - <i>PDF</i> para acesso a e-book | 57 |
| Figura 15: 2003 - Suporte a <i>XML</i> | 57 |
| Figura 16: 2005 - Ano publicação <i>PDF/A</i> | 58 |
| Figura 17: Arquivo gerado no formato <i>PDF/A</i> | 59 |
| Figura 18: 2007 - <i>Adobe</i> libera a especificação <i>PDF 1.7</i> para a <i>AIIM</i> | 60 |
| Figura 19: 2007 - Suporte para envios <i>SAFE</i> | 60 |
| Figura 20: 2007 - Liberada a especificação <i>PDF/E</i> padrão para dados de engenharia..... | 60 |
| Figura 21: 2008 - <i>PDF</i> aprovado como padrão internacional | 61 |
| Figura 22: 2008 - Orçamento dos USA são publicados em <i>PDF</i> | 61 |
| Figura 23: Uma taxonomia de modelos de Recuperação de Informação | 70 |
| Figura 24: Estrutura de arquivo invertido. | 79 |
| Figura 25: Estrutura de arquivo invertido dividido em quatro blocos..... | 79 |
| Figura 26: Tela - progresso da instalação | 112 |
| Figura 27: Tela de aviso de segurança | 113 |
| Figura 28: Tela de preparação de instalação do <i>Greenstone</i> | 113 |
| Figura 29: Tela de seleção de linguagem de preferência na instalação da <i>BDG</i> | 114 |
| Figura 30: Tela de direitos autorais | 115 |
| Figura 31: Tela de seleção do local de instalação | 115 |
| Figura 32: Tela de seleção dos componentes | 116 |
| Figura 33: Tela configuração do <i>Apache Tomcat</i> | 117 |
| Figura 34: Tela que mostra o progresso de instalação do <i>Greenstone</i> | 117 |
| Figura 35: Tela de criação da coleção do <i>PGCIN</i> | 121 |
| Figura 36: Tela download da coleção do <i>PGCIN</i> | 122 |
| Figura 37: Tela importação de documentos..... | 123 |
| Figura 38: Administração de metadados..... | 123 |
| Figura 39: Tela para adicionar plug-ins de formato de arquivo digital..... | 133 |
| Figura 40: Tela configuração de plug-ins | 134 |

| | |
|---|-----|
| Figura 41: Tela de opção de indexação | 135 |
| Figura 42: Tela de opção de indexação MGPP, MG e LUCENE | 135 |
| Figura 43: Tela de associação de língua na partição de indexação | 136 |
| Figura 44: Tela configuração de <i>browsing classifiers</i> | 137 |
| Figura 45: Tela início de importação de documentos | 138 |
| Figura 46: Tela fim de importação de documentos..... | 138 |
| Figura 47: Tela aba format - dados gerais | 139 |
| Figura 48: Tela aba format - itens de pesquisa no menu..... | 139 |
| Figura 49: Tela aba format – recursos do formato | 140 |
| Figura 50: Tela aba format – tradução de textos..... | 140 |
| Figura 51: Tela de acesso a todas as coleções instaladas..... | 141 |
| Figura 52: Tela inicial de consulta da coleção PGCIN na BDG | 141 |
| Figura 53: Tela escolha do idioma da interface e preferências de impressão .. | 142 |
| Figura 54: Tela busca pelo texto completo (search) | 142 |
| Figura 55: Tela dissertações ordenadas por título..... | 143 |
| Figura 56: Tela dissertações ordenadas por autor | 143 |
| Figura 57: Tela dissertações do aluno..... | 144 |
| Figura 58: Tela dissertações ordenadas por ano defesa dissertação..... | 144 |
| Figura 59: Tela visualização conteúdo modo texto | 145 |
| Figura 60: Tela orientadores por ordem alfabética | 145 |
| Figura 61: Tela de dissertações ao qual o orientador está vinculado. | 146 |
| Figura 62: Tela de visualiza linhas de pesquisa..... | 146 |
| Figura 63: Tela dissertações vinculadas à linha de pesquisa selecionada | 147 |
| Figura 64: Tela principal da Biblioteca Digital Greenstone | 156 |
| Figura 65: Tela inicial da Coleção PGCIN | 157 |
| Figura 66: Tela de preferências de apresentação e de busca | 158 |
| Figura 67: Tela preferências de pesquisa..... | 159 |
| Figura 68: Recuperação de Informação a partir de palavras | 159 |
| Figura 69: Indexadores do Greenstone | 160 |
| Figura 70: Tela metadados Greenstone | 161 |
| Figura 71: Tela editor de metadados..... | 161 |
| Figura 72: Tela dissertações por ordem alfabética de título | 162 |
| Figura 73: Tela seleção de visualização dissertação por título | 163 |
| Figura 74: Tela dissertações por ordem alfabética de autor | 163 |
| Figura 75: Tela seleção de visualização dissertação por autor..... | 164 |
| Figura 76: Tela dissertações por ordem alfabética de orientador..... | 164 |
| Figura 77: Tela seleção de visualização dissertação por orientador | 165 |
| Figura 78: Tela dissertações por ordem alfabética por linha de pesquisa | 165 |
| Figura 79: Tela seleção de visualização dissertação por linha pesquisa | 166 |
| Figura 80: Tela dissertações por ordem ano pesquisa..... | 166 |
| Figura 81: Tela seleção de visualização dissertação por ano..... | 167 |

LISTA DE QUADROS

| | |
|--|-----|
| Quadro 1: Cálculo de relevância | 73 |
| Quadro 2: Situações relevantes para diferentes estratégias de pesquisa | 98 |
| Quadro 3: Plug-ins especiais – nível superior | 129 |
| Quadro 4: Plug-ins especiais – nível superior | 129 |
| Quadro 5: Plug-ins Base | 130 |
| Quadro 6: Plug-ins auxiliares | 131 |
| Quadro 7: Arquivos gerados para compor a coleção PGCIN | 149 |
| Quadro 8: Arquivos que foram importados para a coleção PGCIN no Greenstone | 149 |
| Quadro 10: Plug-ins de utilizados de nível superior | 152 |
| Quadro 10: Plug-ins de utilizados de nível superior | 153 |
| Quadro 11: Plug-ins de utilizados de nível superior | 155 |

LISTA DE GRÁFICOS

| | |
|---|----|
| Gráfico 1: Relação Precisão x Revocação | 86 |
| Gráfico 2: Dimensão LSI..... | 95 |

LISTA DE ABREVIATURAS E SIGLAS

AGLS - Australian Government Locator Service
AIIM - Association for Information and Imagem Management
ANSI - American National Standards Institute
BDG – Biblioteca Digital Greenstone
CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CCSDS - Consultative Comitee for Space Data Systems
COMDEX - Computer Dealers' Exhibition
CONARQ - Conselho Nacional de Arquivos
CONSEGI - Congresso Internacional Sociedade e Governo Eletrônico –
DCMI - Dublin Core Metadata Initiative
DLF - Digital Library Federation
DOC - Extensão de nome para arquivos criados em editores de textos Word da empresa Microsoft.
DOCX – Formato de arquivo Open XML utilizado pelo Microsoft Word
DROID – Digital Record Object IDentification
e-Ping - Padrões de Interoperabilidade de Governo Eletrônico
FSF - Free *Software* Foundation
GPL - General Public License
HD – Hard Disk
IEC - International Electrotechnical Commission
ISO - International Organization for Standardization
LAN - Local Area Network
LSA – Análise Semântica Latente
LSI – Indexação Semântica Latente
MoReq - Model Requirements for the Management of Electronic Documents and Records
NARA - National Archives e Records Administration
NBR – Norma Brasileira
NISO - National Information Standards Organization
NZDL - New Zealand Digital Library Project
NZGLS -New Zealand Government Locator Service
OAIS - Open Archival Information System
OASIS – Organization for the Advancement of Structured Information Standards
OCR - Optical Character Recognition
ODF - Open Document Format

PDF - Portable Document Format
PDF/A - Portable Document Format Archive
PPT – Formato de arquivo utilizado pelo Microsoft Powerpoint
RTF – Formato de arquivo RTF Rich Text Format
SAFE - Signatures and Authentication for Everyone
SRI – Sistemas de Recuperação da Informação
SVD – Decomposição de Valores Singulares - Singular Value
Decomposition
TIC – Tecnologia da Informação e Comunicação
TREC - Text REtrieval Conference
UDESC – Universidade do Estado de Santa Catarina
UFSC - Universidade Federal de Santa Catarina
UNESCO - United National Educational Scientific and Cultural
Organization
WWW – Word Wide Web
XLS – Formato de arquivo Planilha Microsoft Excel
XMP - Extensible Metadata Platform

SUMÁRIO

| | |
|--|-----------|
| 1 INTRODUÇÃO | 29 |
| 1.1 JUSTIFICATIVAS | 31 |
| 1.1.1 <i>Justificativa Pessoal</i> | 31 |
| 1.1.2 <i>Justificativa Científica</i> | 31 |
| 1.1.3 <i>Justificativa Econômica</i> | 32 |
| 1.1.4 <i>Justificativa Social</i> | 32 |
| 1.2 PROBLEMA DE PESQUISA | 33 |
| 1.3 OBJETIVOS | 34 |
| 1.3.1 <i>Objetivos Gerais</i> | 34 |
| 1.3.2 <i>Objetivos Específicos</i> | 34 |
| 1.4 ORGANIZAÇÃO DA DISSERTAÇÃO..... | 35 |
| 2 FUNDAMENTAÇÃO TEÓRICA..... | 36 |
| 2.1 FONTES DE INFORMAÇÃO E BIBLIOTECAS DIGITAIS | 36 |
| 2.2 PRESERVAÇÃO DA INFORMAÇÃO DIGITAL..... | 42 |
| 2.3 FORMATO DE ARQUIVOS..... | 47 |
| 2.4 FORMATO DE ARQUIVOS ABERTOS..... | 48 |
| 2.5 FORMATO DE ARQUIVOS ABERTOS ODF..... | 51 |
| 2.6 FORMATO DE ARQUIVOS ABERTOS <i>PDF/A</i> | 52 |
| 2.7 METADADOS XMP..... | 62 |
| 2.8 A RECUPERAÇÃO DA INFORMAÇÃO E A CIÊNCIA DA INFORMAÇÃO..... | 64 |
| 2.9 RECUPERAÇÃO DA INFORMAÇÃO..... | 66 |
| 2.9.1 <i>Definição de Recuperação de Informação</i> | 68 |
| 2.9.2 <i>Modelos de Recuperação de informação</i> | 70 |
| 2.9.3 <i>Modelo Booleano</i> | 70 |
| 2.9.4 <i>Modelo Vetorial</i> | 72 |
| 2.9.5 <i>Modelo Probabilístico</i> | 74 |
| 2.10 INDEXAÇÃO AUTOMÁTICA DE TEXTOS..... | 77 |
| 2.10.1 <i>Arquivos Invertidos</i> | 78 |
| 2.10.2 <i>Identificação das Palavras</i> | 80 |
| 2.10.3 <i>Remoção de Stopwords</i> | 81 |
| 2.10.4 <i>Normalização Morfológica (Stemming)</i> | 81 |
| 2.10.5 <i>Identificação de Termos Compostos</i> | 82 |
| 2.11 CÁLCULO DE RELEVÂNCIA | 83 |
| 2.12 AVALIAÇÃO DA RECUPERAÇÃO DA INFORMAÇÃO..... | 84 |
| 2.13 INDEXAÇÃO SEMÂNTICA LATENTE..... | 86 |

| | |
|---|------------|
| 3 PROCEDIMENTOS METODOLÓGICOS | 96 |
| 3.1 TIPO DA PESQUISA | 96 |
| 3.2 ESTUDO DE CASO | 97 |
| 3.2 COLETA DE DADOS | 99 |
| 3.3 UNIDADE DE ANÁLISE | 101 |
| 3.5 UNIVERSO DA PESQUISA | 102 |
| 3.6 LIMITAÇÕES DA PESQUISA | 103 |
| 3.7 ETAPAS DA PESQUISA | 103 |
| 3.8 PROCEDIMENTOS PARA COLETA DE DADOS | 104 |
| 4 GREENSTONE..... | 105 |
| 4.1 OBTENDO O GREENSTONE VERSÃO 3.04..... | 110 |
| 4.2 INSTALAÇÃO DO GREENSTONE VERSÃO 3.04 | 111 |
| 4.3 CONSTRUINDO COLEÇÃO DE DISSERTAÇÕES DO PGCIN | 120 |
| 5 ANÁLISE E INTERPRETAÇÃO DOS RESULTADOS | 148 |
| 5.1 ANÁLISE DOS FORMATOS DE ARQUIVOS DA COLEÇÃO PGCIN..... | 148 |
| 5.1.1 <i>Análise dos formatos de arquivos com especificações proprietárias e fechadas no Greenstone</i> | <i>152</i> |
| 5.1.2 <i>Análise dos formatos de arquivos com especificações proprietárias e abertas no Greenstone</i> | <i>153</i> |
| 5.1.3 <i>Análise dos formatos de arquivos com especificação não-proprietária e aberta no Greenstone</i> | <i>154</i> |
| 5.2 ANÁLISE DA RECUPERAÇÃO DA INFORMAÇÃO NO GREENSTONE... | 155 |
| 6 CONCLUSÕES | 168 |
| 6.1 CONCLUSÕES | 168 |
| 6.2 SUGESTÕES | 171 |
| 6.3 RECOMENDAÇÕES | 171 |
| REFERÊNCIAS BIBLIOGRÁFICAS | 174 |
| ANEXO A: TELA DO RESULTADO ANÁLISE DO DROID | 186 |
| ANEXO B – RELATÓRIO GERADO PELO SOFTWARE DROID | 187 |
| ANEXO C – NATIONAL ARCHIVES – FORMATO FMT/111 .. | 216 |
| ANEXO D – NATIONAL ARCHIVES – FORMATO FMT/17 | 217 |
| ANEXO E – NATIONAL ARCHIVES – FORMATO FMT/18 | 218 |
| ANEXO F – NATIONAL ARCHIVES – FORMATO FMT/19..... | 219 |

1 INTRODUÇÃO

Após a Segunda Grande Guerra Mundial, o uso de computadores inicialmente restrito basicamente para fins militares expandiu a sua área de abrangência, ou seja, passou a ser utilizada em áreas como a educação, saúde, segurança, empresas públicas e privadas, e em diversos tipos de organizações. Até a década de 1970, devido a sua complexidade e ao seu alto custo, o acesso aos computadores era restrito aos profissionais que trabalhavam em Centro de Processamento de Dados - CPD, profissionais estes que se assemelhavam mais a cientistas trabalhando em laboratório, inclusive tinham um status diferenciado onde trabalhavam.

Com a disseminação dos computadores pessoais, houve uma descentralização das atividades informatizadas. Tal disseminação foi potencializada com o advento da tecnologia de rede, que evoluiu rapidamente das redes locais para as metropolitanas, nacionais e globais, sendo a Internet a maior delas. Com a popularização das Tecnologias da Informação e Comunicação, ocorreu um aumento considerável dos estoques de informação, principalmente as armazenadas em meios digitais. Com o aumento destes estoques de informação ficou mais evidente o problema da preservação e recuperação das informações de documentos no formato digital.

O interesse na preservação e recuperação de informação aumentou consideravelmente com a popularização do uso das tecnologias de informação e comunicação (TIC) ocorrida a partir dos anos 1990, onde ocorreu uma mudança considerável nos mecanismos de registro e de comunicação da informação nas instituições públicas e privadas. Os documentos produzidos no decorrer das atividades dessas instituições, até então em meio convencional, assumem novas características, isto é, passam a ser gerados em ambientes eletrônicos, armazenados em suportes magnéticos e ópticos, em formato digital.

Os documentos digitais trouxeram uma série de vantagens na produção, transmissão, armazenamento e acesso, que por sua vez, acarretaram outros problemas. A facilidade de criar e transmitir documentos traz como consequência a informalidade na linguagem nos procedimentos administrativos.

O desenvolvimento das tecnologias de informação e comunicação, e em especial o desenvolvimento da Internet, tem contribuído para um ambiente completamente novo, onde os papéis das bibliotecas tradicionais estão sendo amplamente modificados. O

potencial das redes de informação, de cooperação e de digitalização modifica substancialmente as funções de recuperação, preservação e disseminação da informação e do conhecimento.

Com o crescimento dos estoques de informação em formato digital, fontes de informação como bibliotecas, que só existiam em formato convencional, ou seja, baseado em material impresso, começaram a migrar para um suporte digital. Observa-se também uma forte dissociação entre o conteúdo informacional e o suporte de armazenamento.

Esta migração de suporte digital também ocorreu nas bibliotecas tradicionais, inclusive apareceram novos conceitos de bibliotecas como a eletrônica, virtual e por fim consolidando-se como biblioteca digital. Em decorrência dessa evolução, observa-se que também ocorreu uma mudança na função de mediação no acesso a informação.

As bibliotecas digitais são exemplos de iniciativas que contemplam os conceitos relacionados à preservação digital e recuperação da informação, bem como, podem ser vistas como grandes repositórios de produção intelectual, sobretudo no que diz respeito à disseminação intelectual de grandes campos de pesquisas científicas em diversas áreas de conhecimento.

Os conceitos de bibliotecas digitais, recuperação de informação, e preservação digital possuem um vínculo indissociável com a ciência da informação com pode-se observar nos estudos realizados por Cientistas da área da Ciência da Informação, como Barreto (1999, 2000), Blattmann (1999, 2001, 2003, 2006), Campello (2003,2005), Côte (2002), Cunha (2001, 2005), Ferneda (2003), Ferreira (2006), Kafure (2004), Kuramoto (2005), Leite (2006), Lopes (2004), Machado (2006), Marcondes (2005), Márdero Arellano (2004, 2006), Sayão (2007), entre outros.

A UNESCO promove o uso de tecnologias de informação e comunicação para o desenvolvimento econômico e social (<http://unesdoc.unesco.org/images/0014/001473/147330por.pdf>), sendo que o apoio a Biblioteca Digital Greenstone é umas destas ações.

O software *GREENSTONE* é uma ferramenta para o desenvolvimento e distribuição de coleções de bibliotecas digitais, desenvolvido pela Universidade de *Waikato* na Nova Zelândia. É um

software de código aberto, disponível através dos termos de licença pública geral do *GNU - General Public License*¹.

1.1 Justificativas

O desenvolvimento dessa pesquisa se justifica por motivos pessoais, econômicos e pela contribuição à Ciência da Informação especialmente à área de Fontes de Informação, Preservação e Recuperação de Informação, e por fim, pela contribuição à sociedade.

1.1.1 Justificativa Pessoal

A motivação pessoal para realização dessa pesquisa deve-se ao fato de o autor trabalhar como Analista de Sistemas e ocupa o cargo de Secretário de Tecnologia de Informação e Comunicação na Universidade do Estado de Santa Catarina. Cabe destacar que a motivação para realização de um projeto de pesquisa depende do contexto no qual o pesquisador está inserido visando satisfazer uma curiosidade ou uma necessidade pessoal.

Sabe-se que toda pesquisa, análise ou estudo, tem como ponto de partida uma situação percebida como problemática, ou seja, que causa desconforto e que, em consequência, exige uma explicação. Essa situação problemática surge quando há defasagem entre a concepção ou explicação de um fenômeno e a observação ou a percepção ou explicação de um fenômeno e a observação ou a percepção da realidade. Sendo assim, é a partir dessa defasagem que se origina o objeto da pesquisa.

1.1.2 Justificativa Científica

Sob o ponto de vista de contribuição a Ciência da Informação mais especificamente sobre Fontes de Informação, Bibliotecas Digitais, Preservação e Recuperação da informação, a pesquisa aprofunda estudos nesta área tendo como foco na preservação e recuperação de informação.

¹ GNU General Public License (Licença Pública Geral), GNU GPL ou simplesmente GPL, é a designação da licença para software livre idealizada por Richard Stallman no final da década de 1980, no âmbito do projeto GNU da Free Software Foundation (FSF) <<http://www.fsf.org/>>.

Os conhecimentos científicos podem se perder caso não ocorra à devida documentação da sua prática, bem como, não se adote medidas visando à preservação e recuperação de informação.

Este estudo por meio de uma pesquisa exploratória pretende buscar respostas para o problema de pesquisa. Um objeto de pesquisa é assim uma interrogação explícita em relação a um problema a ser examinado e analisado com o fim de obter novas informações (Contandriopoulos *apud* Fortin et al., 1999, p. 19).

1.1.3 Justificativa Econômica

Sob o aspecto econômico, essa pesquisa se justifica pelo fato de o processo de preservação e recuperação de informação ter alto custo para as organizações. Este custo varia de organização para organização, sendo que em alguns casos, quando a informação não é encontrada, ou se o tempo de recuperação não atende às expectativas e, principalmente se a informação não é preservada, esse processo pode trazer um forte impacto econômico para as mesmas, inclusive podendo inviabilizá-las.

1.1.4 Justificativa Social

A contribuição dessa pesquisa à sociedade se dá pela expectativa de melhorias no processo de preservação e de recuperação de informação em fontes de informação como as bibliotecas digitais, ajudando assim as pessoas, empresas, organizações a lidarem com grandes estoques de informação.

Conforme consta no seu sítio na internet, a UNESCO promove ações de apoio às bibliotecas há mais de 60 anos, pois considera que estas são essenciais ao fluxo livre de ideias e a manutenção e aumento da disseminação do conhecimento. Uma das áreas de maior prioridade da UNESCO é a promoção do uso de tecnologias de informação e comunicação para o desenvolvimento econômico e social. Os avanços tecnológicos em comunicação e informação devem ser apropriados pela sociedade para facilitar a modernização da Gestão do Estado, a participação nas decisões e a Inclusão Social, sendo que o apoio a Biblioteca Digital Greenstone é umas dessas ações.

1.2 Problema de Pesquisa

Com a disseminação do uso de computadores, houve crescimento das informações armazenadas no formato digital. Os documentos estão, de forma crescente, nascendo ou migrando de outros meios para um formato digital, e com isso está surgindo a preocupação com a preservação e recuperação de informações digitais. Em fontes de informações digitais como a Biblioteca Digital Greenstone, o problema não é diferente.

Com o presente estudo espera-se obter respostas para o seguinte problema da pesquisa:

Os recursos disponíveis na Biblioteca Digital Greenstone são suficientes para realizar a preservação lógica dos documentos digitais e a recuperação de informação?

1.3 Objetivos

Os objetivos do presente estudo estão divididos em geral e específicos.

1.3.1 Objetivos Gerais

Esta pesquisa tem o objetivo principal de analisar os recursos disponíveis na Biblioteca Digital Greenstone para preservação lógica de documentos digitais com foco no formato de arquivos e a recuperação da informação.

1.3.2 Objetivos Específicos

Os objetivos específicos são:

- a) Estudar os modelos clássicos de recuperação de informação;
- b) Identificar os recursos disponíveis para recuperação de informação na BDG;
- c) Identificar os pontos fortes e pontos fracos da BDG; e,
- d) Analisar a questão de preservação de documentos digitais sob o ponto de vista lógico na Biblioteca Digital Greenstone.

A partir dos objetivos supracitados tem-se a expectativa de identificar como o software de Biblioteca Digital Greenstone preserva as informações e recupera informação no formato digital.

1.4 Organização da Dissertação

A organização dessa dissertação utiliza a apresentação em capítulos.

O presente texto é parte do capítulo introdutório, que apresenta também a justificativa, a definição do problema e os objetivos pretendidos.

No segundo capítulo, é apresentada uma fundamentação teórica, na qual, auxiliada pela revisão de literatura, toma-se conhecimento sobre:

- a) Fontes de Informação e Bibliotecas Digitais;
- b) Preservação da Informação Digital;
- c) Formato de Arquivos;
- d) Formato de Arquivos Abertos;
- e) Formato de Arquivos Abertos *ODF* e *PDF/A*;
- f) Metadados *XMP*;
- g) Recuperação da Informação;
- h) Recuperação da Informação e a Ciência da Informação;
- i) Modelos de Recuperação da Informação;

No terceiro capítulo encontra-se a metodologia utilizada para desenvolvimento e aplicação do presente trabalho.

O quarto capítulo apresenta a Biblioteca Digital Greenstone.

O quinto capítulo apresenta a análise e a interpretação dos resultados.

O sexto capítulo apresenta as conclusões e sugestões.

Ao final estão as referências, apêndices e anexos.

2 FUNDAMENTAÇÃO TEÓRICA

A revisão da literatura pretende mostrar como o objeto da pesquisa se insere no campo dos conhecimentos sobre o tema, e como estes conhecimentos vão permitir responder as questões da pesquisa.

A atividade científica resulta de um processo cumulativo de aquisição do conhecimento. Posto isto, o processo de revisão de literatura sobre o tema iniciou com a consulta de fontes primárias e secundárias de informação, ou seja, artigos científicos, livros, teses e memorandos e documentos oficiais, verificando suas bibliografias e com isto ampliando as listas de trabalhos consultados.

Esta revisão de literatura permite compreender e ou concluir a situação em que estão inseridos os conhecimentos sobre o objeto da pesquisa apresentada, começando por Fontes de Informação, Preservação Digital, Recuperação da Informação e a Ciência da Informação, Recuperação da Informação e Modelos de Recuperação da Informação.

2.1 Fontes de Informação e Bibliotecas Digitais

Até fins dos anos de 1990, fontes de informação era sinônimo de formato impresso, sendo que a quantidade de informações disponíveis em formato de papel era extremamente maior do que as informações disponíveis em formato digital. Com a disseminação do uso de computadores para trabalho e lazer, bem como, com o aumento da capacidade de armazenamento e de recuperação de informações, e principalmente com o advento da internet, observa-se que fontes de informação viraram sinônimo de informação no formato digital.

As mídias onde estão localizadas as fontes de informação foram evoluindo com o tempo, ou seja, da pedra, papiro, papel, fotografias e microfilme, para os mais recentes dispositivos, como fitas magnéticas, fitas *K7*, discos flexíveis, fitas *VHS*, disquetes, discos rígidos (HD), *Compact Disc* (CD's), *videolaser*, *DVD's* e *pen-drives*.

Muitas informações existem somente em formato de papel, outras informações estão em papel e em formato digital, e outras somente em formato digital. Observa-se que as informações de várias formas estão migrando para o formato digital, quer seja pela digitalização de documentos, ou sendo criadas originalmente em formato digital.

De acordo com a norma NISO² há dois tipos de objetos digitais a serem considerados em bibliotecas digitais:

- a) Os objetos produzidos como representação ou substitutos de materiais em alguma forma analógica – livros impressos, manuscritos, peças de museus, entre outros; e,
- b) Os objetos originalmente “nascidos digitais”, como, por exemplo, fotografias digitais, livro eletrônico, bases de dados, websites, entre outros.

Segundo Cunha 2001, os documentos ou fontes de informação podem ser classificados da seguinte forma:

Fontes Primárias – Contêm informações originais ou, pelo menos, novas interpretações de fatos ou ideias já conhecidas e não submetidas à interpretação ou condensação.

Exemplos de fontes de informação primárias:

- a) Congressos e conferências;
- b) Legislação;
- c) Nomes e marcas comerciais;
- d) Normas técnicas;
- e) Patentes;
- f) Periódicos;
- g) Projetos e pesquisa em andamento;
- h) Relatórios técnicos;
- i) Teses e dissertações; e,
- j) Traduções.

Fontes Secundárias – Têm a função de facilitar o uso do conhecimento disperso nas fontes primárias; apresentam a informação filtrada e organizada, de acordo com o arranjo definido, dependendo da finalidade da obra. Também são considerados os produtos de análise de fontes primárias submetidas à descrição, condensação ou qualquer tipo de reorganização.

Exemplos de fontes de informação secundária:

- a) Bases de dados e bancos de dados;
- b) Bibliografias e índices;
- c) Biografias;
- d) Catálogos de bibliotecas;

² National Information Standards Organization - <http://www.niso.org/>

- e) Centros de pesquisa e laboratórios;
- f) Dicionários e enciclopédias;
- g) Dicionários bilíngues e multilíngues;
- h) Feiras e exposições;
- i) Filmes e vídeos;
- j) Fontes históricas;
- k) Livros;
- l) Manuais;
- m) Internet;
- n) Museus, herbários, arquivos e coleções científicas;
- o) Prêmios e honrarias;
- p) Redação técnica e metodologia científica;
- q) Siglas e abreviaturas; e,
- r) Tabelas, unidades, medidas e estatística.

Fontes Terciárias – Tem a função de guiar o usuário da informação para as fontes primárias e secundárias. Podem ser consideradas também como uma recompilação das informações contidas nas fontes primárias e secundárias, dentro de um critério de organização para torná-las mais acessíveis aos usuários.

Exemplos de fontes de informação terciária:

- a) Bibliografias de bibliografias;
- b) Bibliotecas e centros de informação; e,
- c) Diretórios.

Segundo Cunha (2001), as bibliotecas e centros de informação e documentação, tradicionalmente, têm sido um dos grandes responsáveis pela aquisição, armazenamento, preservação e disseminação da literatura técnico-científica. Essa literatura, porém, tem tido enorme crescimento, é cara e nem sempre possui boa cobertura por parte dos índices correntes. Assim, é necessário que as bibliotecas lancem mão, cada vez mais, dos acervos de outras bibliotecas para atender às necessidades dos usuários. O advento das bibliotecas digitais ou virtuais, fez com que esse compartilhamento de coleções fosse aprimorado e agilizado. Portanto, é importante consultar fontes que informem o endereço, o acervo e os produtos e serviços fornecidos pelas bibliotecas.

O conceito de biblioteca também por muito tempo esteve associado a fontes de informação como livros, jornais e revistas. Mas, essa visão tradicional de biblioteca tem mudado consideravelmente com a utilização de tecnologias de informação e comunicação, onde as informações passaram a residir em um formato digital. Com essa mudança, surge o conceito de biblioteca virtual. O termo biblioteca

virtual é um bom exemplo da dificuldade de conceituação das novas fontes. Basta consultar a literatura a respeito para verificar as várias definições que o termo tem assumido, causando inclusive certa confusão entre biblioteca tradicional, biblioteca eletrônica, biblioteca virtual e biblioteca digital.

De acordo com Tammaro (2008), a expressão biblioteca eletrônica (*electronic library*) tem duas décadas e o seu conceito está vinculado a equipamentos eletrônicos como computadores. Durante muito tempo, em lugar de 'biblioteca digital', foi dada preferência à expressão biblioteca virtual para definir o conceito da nova biblioteca. O primeiro a usar a expressão 'biblioteca virtual' (*virtual library*) foi o criador da Rede - Tim Berners Lee - para o sítio assim denominado e que materializa a visão de uma biblioteca como uma coleção de documentos ligados em rede, constituídos por objetos digitais e páginas Web produzidos por milhares de autores.

O adjetivo 'virtual' significa que a biblioteca não existe fisicamente. A denominação, que hoje é, no entanto, menos difundida do que 'biblioteca digital', continuou sendo usada para certas acepções, como, por exemplo, para indicar uma coleção selecionada de vínculos com sítios da Rede e também para se referir a um conceito mais amplo tanto da biblioteca eletrônica quanto da biblioteca digital, quer dizer, uma coleção de documentos fora da biblioteca como espaço físico ou lógico.

Desde o fim dos anos 1990, a expressão biblioteca digital tornou-se comum e amplamente difundida, porém as definições relativas a essa expressão continuam diferentes, e passam por constantes mudanças.

Inúmeras definições foram originadas nos últimos anos, em especial as referentes à biblioteca virtual, muito utilizada como sinônimo da biblioteca eletrônica, que promove o acesso remoto aos conteúdos e serviços tradicionais da biblioteca com a integração de recursos e serviços eletrônicos disponibilizados em redes de computadores, interagindo o usuário, a informação em formato digital e redes eletrônicas.

O conceito de biblioteca digital da *Digital Library Federation (DLF)* é uma das mais difundidas. Ela registra na sua página web (<http://www.diglib.org/about/dldefinition.htm>) uma definição abrangente que institucionaliza a visão biblioteconômica das bibliotecas digitais:

Bibliotecas digitais são organizações, que disponibilizam recursos (humanos inclusive), para a seleção, estruturação, interpretação, distribuição e disponibilização de objetos digitais, e que devem

zelar por sua integridade/autenticidade, de forma que sejam acessíveis a baixo custo para a comunidade³. (tradução nossa).

De acordo com Tammaro (2008, p.119), uma das melhores definições de biblioteca digital foi formulada pela comunidade de pesquisadores sobre biblioteca digital e empregada no *Workshop on Distributed Knowledge Work Environments*, em *Santa Fe (EUA)*, em 1997:

[...] o conceito de 'biblioteca digital' não é simplesmente o equivalente ao de uma coleção digitalizada dotada de instrumentos de gestão da informação. É, antes, um ambiente que reúne coleções, serviços e pessoas para apoiar todo o ciclo vital de criação, disseminação, uso e preservação de dados, informação e conhecimento.

Conforme Tammaro (2008, p.122), a última definição de biblioteca digital, representa a evolução de uma biblioteca tradicional para digital (talvez fosse melhor defini-la como híbrida):

[...] podemos afirmar que a biblioteca digital é o conjunto de uma ou várias coleções de objetos digitais, da descrição desses objetos, que é feita com o emprego dos chamados metadados colocadas à disposição de todos os usuários interessados graças a uma interação de tipo eletrônico que pode abranger diversos serviços, como a catalogação, indexação, recuperação de documentos e fornecimento de informações à distância. Nessa biblioteca todos os pedidos dos usuários e as respostas a eles se realizam, portanto, por meio da Rede.

A concepção de uma biblioteca digital deve ser realizada como uma ferramenta para propiciar o acesso à informação constituída em meio digital e também incluir outros meios tradicionais, mas, antes de tudo, deve constituir-se como um instrumento para a democratização do acesso ao conhecimento e inclusão social e cultural.

De acordo com Marcondes e Sayão (2003a), as bibliotecas digitais hoje são geradoras, e responsáveis pela gestão e preservação das informações digitais. Dessa forma, as bibliotecas digitais se tornam cada

³ Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities.

vez mais um elo importante na perenização dos estoques de informação digital, os quais constituem testemunhos das atividades das organizações no qual essas bibliotecas estão inseridas. Um exemplo concreto disso são as redes de bibliotecas de teses e dissertações.

Assim como vários dos conceitos da área de biblioteconomia são utilizados por outras áreas de conhecimento, o mesmo acontece com a biblioteca digital. Constata-se que o conceito de biblioteca é muito utilizado na área de tecnologia de informação e comunicação. Exemplo disso são os termos utilizados em desenvolvimento de sistemas e em sistemas operacionais. Em sistemas operacionais, como por exemplo o serviço *Windows Explorer* do *Windows Seven da Microsoft*, fica claro a utilização dos conceitos de biblioteca digital, inclusive, permite que leigos possam criar e compartilhar documentos, ou seja, fazer a gestão da sua própria biblioteca.

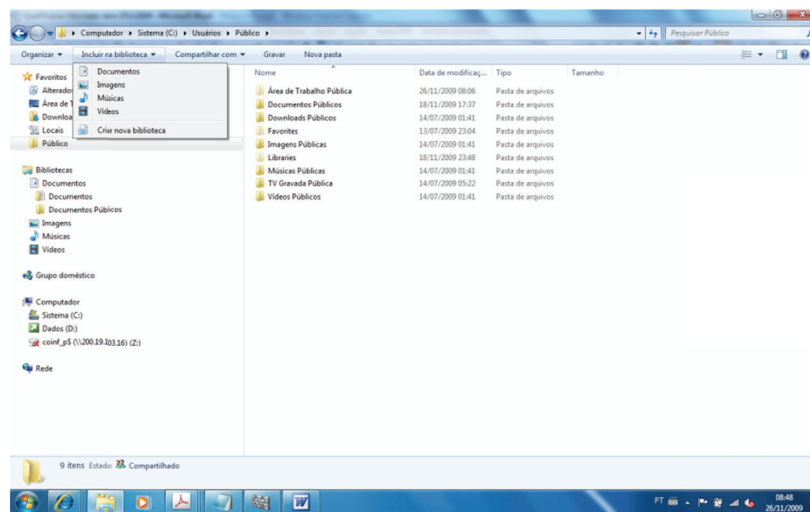


Figura 1: 1990 *Windows Explorer* do *Windows Seven da Microsoft*
Fonte: Sistema Operacional *Windows Seven Microsoft*

Uma biblioteca digital também pode ser considerada como um conjunto de tecnologias de informação e comunicação, onde inúmeras fontes de informação são acessadas a partir de um portal específico, sendo possível armazenar em formato digital grandes estoques de informação, bem como recuperar informações relevantes ao usuário de forma rápida, inclusive, permitindo reproduzir, emular, e estender os

serviços oferecidos por bibliotecas convencionais baseadas em papel e outros meios de coleção, catalogação, e disseminação da informação.

2.2 Preservação da Informação Digital

Estima-se que a quantidade de informação produzida nos últimos anos é superiores a toda informação produzida pelas gerações que nos antecederam, e que grande parte dessas informações estão sendo geradas diretamente no formato digital, além do que, muitas das informações que se encontravam em outro formato também estarem migrando para formato digital.

Com o aumento da quantidade de informações produzidas em formato digital, faz-se necessário adotar critérios sobre quais e como as informações serão preservadas, pois do mesmo modo que hoje podemos recuperar informações produzidas por gerações que nos antecederam, a de se criar e adotar normas e critérios para que as futuras gerações possam acessar essas informações.

Sayão (2007, p.15) destaca que para manter os objetos digitais perenemente acessíveis para uso, se requer algo mais do que preservar simplesmente o artefato físico; é necessário considerar também várias outras dimensões como:

- a) Preservação física - foco está na preservação das mídias e na sua renovação quando se fizer necessário;
- b) Preservação lógica - foco os formatos e a dependência de hardware e software que mantenha legíveis e interpretáveis a cadeia de bits;
- c) Preservação intelectual - foco o conteúdo intelectual e sua autenticidade e integridade;
- d) Preservação do aparato – na forma de metadados - necessária para localizar, recuperar e representar a informação digital; e,
- e) Monitoramento e à instrumentalização da comunidade alvo - audiência para o qual a informação de forma privilegiada se dirige, no sentido de garantir que ele possa compreender plenamente a informação no momento do seu acesso.

Segundo Sayão (2007, p. 117), a preservação digital não envolve a retenção do objeto informacional em si, mas também do seu significado. Assim sendo, faz-se necessário que as técnicas de

preservação sejam capazes de compreender e recriar a forma original ou a função do objeto de forma que sejam asseguradas sua autenticidade e acessibilidade.

Os estudos sobre preservação digital realizados por Blattmann (1999, 2000, 2001, 2003), Ferneda (2003), Ferreira (2006), Kuramoto (2005), Marcondes (2005), Márdero Arellano (2004, 2006) e Sayão (2007), reforçam a ideia de que a mesma seja uma área em expansão, chamando a atenção de profissionais da área de Biblioteconomia, Ciências da Computação e de Ciências da Informação. De fato, a preocupação com a preservação digital está expandindo para outras áreas de conhecimento, inclusive concretizando-se em ações de governo e de órgãos não governamentais.

A UNESCO na “Carta para a Preservação do Patrimônio Digital”, publicada em 15/10/2003, mostra sua preocupação com a questão da preservação e aponta os princípios que devem nortear o assunto, em 12 artigos. O artigo 6º, “Elaborar estratégias e políticas”, aponta a necessidade de se elaborar políticas e estratégias voltadas para a preservação do patrimônio digital, indicando o grau de urgência para a discussão do assunto e a necessidade de se levar em conta às circunstâncias locais, os meios de comunicação disponíveis e as previsões de futuro. (UNESCO ..., 2003).

O Conselho Nacional de Arquivos (CONARQ) também demonstrou a sua preocupação com a preservação digital, neste sentido, em sua 34ª reunião plenária, realizada em 06/07/2004 no Rio de Janeiro, aprovou a “Carta para a Preservação do Patrimônio Arquivístico Digital: Preservar para Garantir o Acesso”, em que convoca as instituições públicas e privadas a envidarem esforços que garantam a preservação das informações digitais produzidas e armazenadas pelas mesmas, apontando também a necessidades de implementação de ações na elaboração de estratégias, políticas e normas para preservação digital, além de ações para a disseminação e compartilhamento dos conhecimentos adquiridos na área de preservação (ARQUIVO NACIONAL, 2004).

Essa preocupação faz sentido posto que tanto os softwares quanto os hardwares evoluem rapidamente, e como consequência muda a forma como a informação digital é tratada, armazenada e recuperada. Nesse panorama, os suportes que armazenam as informações digitais têm se tornados obsoletos muito rapidamente.

Juntamente com o avanço tecnológico do hardware, acontece à evolução do software, é substituído por uma versão atualizada ou uma nova tecnologia, gerando a necessidade da criação de emuladores para

sua preservação, ou seja, o problema da preservação digital está batendo a porta de todos, pois se observa que os usuários têm problemas de recuperação de informação por questões de formato de arquivos. Exemplo disso são os arquivos existentes em fontes de informação em formato digital cujos documentos foram gerados em softwares de versões ou de desenvolvedores diferentes, e ainda versões diferentes como por exemplo o Word da *Microsoft*, Fácil, Carta Certa, *Wordperfect*, *Lotus Smartsuite*, *Lótus 123*, *Supercalc* e Excel.

O termo preservação digital está muito vinculado ao meio no qual a informação está armazenada. Dessa vinculação, surgem preocupações com a obsolescência, o desgaste físico do equipamento ou das mídias, que passam por processo de desgaste em função uso e do tempo, e possuindo vida útil determinada, desde que observados alguns requisitos, como as condições ideais de temperatura, umidade relativa e iluminação. Nesta dissertação, não é aprofundado o estudo sobre a questão da preservação digital no que tange ao meio físico de armazenamento, sendo focado na questão da preservação sob o ponto de vista do formato lógico.

A preservação digital consiste na capacidade de garantir que a informação digital permaneça acessível e com qualidade de autenticidade suficiente para que possa ser interpretada no futuro, recorrendo a uma plataforma tecnológica diferente da utilizada no momento da sua criação. As informações armazenadas em meio digital de objetos digitais, e definidos como todo e qualquer objeto de informação que possa ser representado por meio de uma sequência de dígitos binários, como por exemplo, textos científicos, bancos de dados, fotos digitais, vídeos, páginas Web, imagem e software.

De acordo com Ferreira (2006), preservação digital é um conjunto de atividades ou processos responsáveis por garantir o acesso continuado a longo-prazo à informação e ao patrimônio cultural existente em formatos digitais.

Para Márdero Arellano (2004), a preservação digital compreende mecanismos que permitem o armazenamento em repositórios de dados digitais que garantam a perenidade dos seus conteúdos, integrando a preservação física, lógica e intelectual dos objetos digitais.

Com relação aos repositórios digitais, Ferreira (2006) aponta que os principais repositórios digitais (*DSpace*, *Fedora* e *Eprints*) não se propõem a implementar de políticas de preservação e nem esquemas de metainformação, mas oferecem capacidade de armazenamento, organização, descrição e disseminação do material armazenado,

possibilitando assim, em curto prazo, a incorporação de funcionalidades de preservação.

A adoção de políticas de preservação digital é a forma mais efetiva de garantir o armazenamento e uso dos recursos de informação por longos períodos de tempo. A falta dessas políticas nos projetos de repositórios digitais sugere a carência de conhecimentos técnicos sobre a importância das estratégias de preservação digital existentes (MÁRDERO ARELLANO, 2004, p. 25).

Segundo Ferreira (2006, p.33), pode-se definir como estratégias de preservação, a conservação do objeto digital no seu formato original e a conservação do conteúdo intelectual do objeto digital. Na conservação do objeto digital no seu formato original, pode se aplicar duas estratégias:

- a) Refrescamento: transferir a informação de um objeto físico de armazenamento para outro mais atual, antes que o primeiro deteriore.
- b) Emulação: técnica de criar um ambiente tecnológico que emule o ambiente original do objeto digital. Mais relevante na preservação de aplicações de software, como por exemplo, jogos de computador.

Já na conservação do conteúdo intelectual do objeto digital, segundo Ferreira (2006, p.36) pode se transferir periodicamente um objeto digital de uma tecnologia de hardware e/ou software para outra mais atual, podem ser aplicadas as seguintes formas de migração:

- a) Migração para suportes analógicos: consiste em converter um objeto digital para um suporte não digital, como, por exemplo, imprimir um texto e armazená-lo em papel.
- b) Atualização de versões: utilizado essencialmente para software, consiste em criar uma versão mais atual do mesmo.
- c) Conversão para formatos concorrentes: consiste em converter o objeto digital para outro formato concorrente, como, por exemplo, converter uma imagem para o formato jpeg.
- d) Normalização: consiste em reduzir o número de formatos de um repositório de objetos digitais, criando condições favoráveis ao processo de interoperabilidade entre sistemas distintos.
- e) Migração a pedido: consiste em aplicar processos de conversão sempre no objeto digital original, pois os diversos processos de conversão

podem gradativamente degradar o formato original do objeto.

f) Migração distribuída: consiste em aplicar remotamente a um objeto digital um conjunto de conversores, acessíveis na Internet, reduzindo assim os custos de preservação.

g) Encapsulamento: consiste em manter o objeto digital original inalterado até que a comunidade efetivamente necessite do mesmo. Nesse momento que o objeto deverá ser tratado.

h) Pedra de Rosetta digital: como no caso da Pedra de Rosetta descoberta no delta do Nilo em 1799, essa estratégia propõe preservar não as regras que permitem decodificar o objeto, mas amostras representativas desse objeto que permitam sua recuperação.

Conforme relata Ferreira (2006), em 1990 o *Consultative Comitee for Space Data Systems (CCSDS)* iniciou um esforço conjunto com a *International Organization for Standardization (ISO)* a fim de desenvolver um conjunto de normas capazes de regular o armazenamento a longo-prazo de informação digital produzida no âmbito de missões espaciais.

Desse esforço nasceu o modelo de referência *OAIS*. Trata-se de um modelo conceitual que visa identificar os componentes funcionais que deverão fazer parte de um sistema de informação dedicado à preservação digital. O modelo descreve as interfaces internas e externas do sistema e os objetos de informação que são manipulados no seu interior. O modelo foi aprovado como uma norma internacional em 2003 – ISO Standard 14721:2003.

De acordo com Ferreira (2006), para que um objeto digital possa ser preservado, é necessário definir os componentes necessários que possibilitem a recuperação das informações contidas nesse objeto. Um dos modelos mais aceitos atualmente é o *Open Archival Information System (OAIS)*, que é definido como um modelo conceitual com o objetivo de identificar os componentes funcionais que deverão fazer parte do sistema de informação dedicado à preservação digital.

2.3 Formato de Arquivos

Para decodificar um formato de arquivo, uma especificação formal deverá estar disponível. Essa especificação bem como a sua disponibilidade, tem forte impacto na vulnerabilidade e obsolescência de um arquivo.

Os formatos de arquivo podem ser proprietários com especificação fechada, proprietários com especificação aberta, e não proprietários com especificação aberta.

Os formatos de arquivos com especificações proprietárias e fechadas são encontrados nos *softwares* com grande aceitação de mercado. A Suíte de escritório Office de propriedade da Microsoft que contempla geralmente o *MS-Office*, *Excel*, *Powerpoint* e outros aplicativos. Antes da versão 2007, utilizava formatos de arquivos com especificações proprietárias e fechadas como os formatos *.doc* e *.xls*. Os formatos de arquivos com especificações proprietárias e fechadas põem em risco a preservação digital dos documentos arquivados nestes formatos.

A Microsoft adotou o formato *Ecma Office Open XML* para os seus produtos a partir da versão 2007. O formato *Open XML* que é um formato de arquivos com especificações proprietárias e abertas desenvolvida pela Microsoft, obteve em abril de 2008 a certificação internacional da Organização Internacional de Padronização (*ISO*) e da Comissão Eletrotécnica Internacional (*IEC*), conforme informação disponível no sítio da Microsoft. A partir desse reconhecimento, o *Open XML* passa a fazer parte dos padrões de formato de documentos abertos reconhecidos pela *ISO* e *IEC*, como o *HTML*, *PDF* e *ODF*. O formato de arquivo *ODF* passou a integrar a lista de formatos de arquivos neste *software*. Na versão do *Microsoft Office 2010*.

Os formatos de arquivos com especificações proprietárias e abertas, também são encontrados em *softwares* com grande aceitação de mercado, como o Adobe Acrobat. Alguns desenvolvedores disponibilizaram publicamente suas especificações, permitindo que outras empresas produzam *software* que possam utiliza-los. Existem ainda vários de formatos proprietários e abertos, que são adotados como norma, como é o caso do PDF. Existem ainda os formatos não proprietários e com especificação aberta como exemplo o *PDF/A*.

As especificações produzidas e mantidas por órgãos normatizados são mais seguras e tem mais garantia de disponibilidade por longo prazo, sendo assim são os mais recomendados para preservação digital. Pode-se citar como exemplo o *Software BrOffice*,

que suporta nativamente o formato de arquivo *ODF (Open Document Format)*. Com a crescente preocupação com a preservação digital, e por força de normas que estão sendo criadas e adotadas mundialmente, até mesmo os *softwares* proprietários nas suas versões mais atualizadas estão suportando esses formatos de arquivos, como por exemplo o *Microsoft Word*.

Um dos pontos fortes da Biblioteca Digital Greenstone, é que a mesma é uma plataforma aberta e possui a disposição, de forma livre e gratuita, *plug-ins* (programas que servem normalmente para adicionar funções a outros programas maiores) para diversas funcionalidades dentre elas os de inúmeros formatos de arquivos, além do que, o mesmo permite o desenvolvimento de *plug-ins* para qualquer formato.

2.4 Formato de Arquivos Abertos

A falta de interoperabilidade e da adoção de formatos abertos de arquivos pode afetar sobremaneira pessoas físicas e jurídicas, pois alguns documentos digitais precisam ser preservados por períodos definidos por lei ou resolução. Posto isto, faz-se necessário estudar questões relativas à preservação digital e a necessidade de integração entre as organizações, visando minimizar os esforços e recursos com preservação digital, bem como, permitir a interoperabilidade dos objetos digitais.

A criação e a adoção de padrões de interoperabilidade têm o objetivo de definir as premissas, as políticas e as especificações técnicas, as quais regulamentam a utilização da tecnologia de informação e comunicação na interoperabilidade de serviços, de forma a permitir a interação entre soluções de TIC. No Brasil, as definições referentes às tecnologias associadas à interoperabilidade são definidas pelos Padrões de Interoperabilidade de Governo Eletrônico (e-PING) (<http://www.governoeletronico.gov.br/acoes-e-projetos/e-ping-padroes-de-interoperabilidade>):

A arquitetura e-PING – Padrões de Interoperabilidade de Governo Eletrônico – definem um conjunto mínimo de premissas, políticas e especificações técnicas que regulamentam a utilização da Tecnologia de Informação e Comunicação (TIC) no governo federal, estabelecendo as condições de interação com os demais Poderes e esferas de governo e com a sociedade em geral. Políticas e

especificações claramente definidas para interoperabilidade e gerenciamento de informações são fundamentais para propiciar a conexão do governo, tanto no âmbito interno como no contato com a sociedade e, em maior nível de abrangência, com o resto do mundo – outros governos e empresas atuantes no mercado mundial. A e-PING é concebida como uma estrutura básica para a estratégia de governo eletrônico, aplicada inicialmente ao governo federal – Poder Executivo. Permite racionalizar investimentos em TIC, por meio do compartilhamento, reuso e intercâmbio de recursos tecnológicos.

A versão 4.0 do e-PING faz a seguinte consideração sobre interoperabilidade, ao avaliar os diferentes conceitos existentes:

Interoperabilidade não é somente Integração de Sistemas, não é somente Integração de Redes. Não referencia unicamente troca de dados entre sistemas. Não contempla simplesmente definição de tecnologia. É, na verdade, a soma de todos esses fatores, considerando, também, a existência de um legado de sistemas, de plataformas de *Hardware* e *Software* instaladas. Parte de princípios que tratam da diversidade de componentes, com a utilização de produtos diversos de fornecedores distintos. Tem por meta a consideração de todos os fatores para que os sistemas possam atuar cooperativamente, fixando as normas, as políticas e os padrões necessários para consecução desses objetivos.

A seguir são apresentados quatro conceitos que fundamentaram os padrões de Interoperabilidade de Governo Eletrônico (e-PING):

Conceito 1 – Governo do Reino Unido - Intercâmbio coerente de informações e serviços entre sistemas. Deve possibilitar a substituição de qualquer componente ou produto usado nos pontos de interligação por outro de especificação similar, sem comprometimento das funcionalidades do sistema.

Conceito 2 - governo da Austrália - Habilidade de transferir e utilizar informações de maneira uniforme e eficiente entre várias organizações e sistemas de informação.

Conceito 3 – *ISO* - Habilidade de dois ou mais sistemas (computadores, meios de comunicação, redes, *software* e outros componentes de tecnologia da informação) de interagir e de intercambiar dados de acordo com um método definido, de forma a obter os resultados esperados.

Conceito 4 - *Lichun Wang, Instituto Europeu de Informática – CORBA Workshops* - Interoperabilidade define se dois componentes de um sistema, desenvolvidos com ferramentas diferentes, de fornecedores diferentes, podem ou não atuar em conjunto.

De acordo com o padrão e-Ping, interoperabilidade não é somente integração de sistemas ou integração de redes. Não referencia unicamente troca de dados entre sistemas e não contempla simplesmente definição de tecnologia. É a soma de todos esses fatores, considerando também, a existência de um legado de sistemas, de plataformas de hardware e software instalado. Parte de princípios que tratam da diversidade de componentes, com a utilização de produtos diversos de fornecedores distintos. Tem por meta a consideração de todos os fatores para que os sistemas possam atuar cooperativamente, fixando as normas, as políticas e os padrões necessários para consecução desses objetivos.

Conforme consta do sitio do Ministério do Planejamento e Orçamento e Gestão do Governo do Brasil disponível em (<http://www.governoeletronico.gov.br/acoes-e-projetos/e-ping-padroes-de-interoperabilidade>) a arquitetura e-PING cobre o intercâmbio de informações entre os sistemas do governo federal – Poder Executivo e as interações com:

- a) Cidadãos;
- b) Outras esferas de governo (estadual e municipal);
- c) Outros Poderes (Legislativo, Judiciário e Ministério Público Federal);
- d) Governos de outros países;
- e) Empresas (no Brasil e no mundo);
- f) Terceiro Setor.

De acordo com e-Ping, para que se conquiste a interoperabilidade, faz-se necessário o engajamento da sociedade num esforço contínuo para assegurar que sistemas, processos e culturas de uma organização sejam gerenciados e direcionados para maximizar oportunidades de troca e reuso de informações, interna e externamente ao governo federal.

2.5 Formato de Arquivos Abertos ODF

Em 1999 foi criado e desenvolvido o formato *ODF*, sigla de *Open Document Format* ou Formato Aberto de Documentos, usado para armazenamento e troca de documentos de escritório, como textos, planilhas, bases de dados, desenhos e apresentações.

Seu desenvolvimento se iniciou em uma empresa alemã, que criou a suíte de escritório *StarOffice* e em pouco tempo se tornou um desenvolvimento aberto e acessível a todos os interessados, capitaneado por uma entidade internacional de desenvolvimento de padrões chamada *OASIS*.

O desenvolvimento do *ODF* contou e conta com a participação de diversas empresas e especialistas do mundo todo, garantindo assim a sua neutralidade tecnológica. Participam atualmente do desenvolvimento do *ODF* empresas como *IBM*, *Sun Microsystems*, *Novell*, *Adobe* e mais, recentemente, *Microsoft*.

O formato *ODF* foi desenvolvido por uma grande variedade de organizações, sendo possível usar livremente as respectivas especificações. Isto significa que o *ODF* pode ser implementado em qualquer sistema seja ele de código aberto ou não, sem ser necessário efetuar qualquer tipo de pagamento ou estar sujeito a uma licença de uso restrito. O *ODF* constitui-se uma alternativa aos formatos de documentação que são propriedade de empresas privadas, sujeitos às licenças de uso restrito ou onerosas, permitindo a organizações e indivíduos escolherem o software que mais lhes convém para lidar com os arquivos guardados nesse formato.

O *ODF* por ser um padrão aberto, também é multi-plataforma, permitindo assim a liberdade de escolha do usuário. Outra característica importante é a vantagem que se oferece em relação à guarda dos documentos digitais, pelo fato de que o mesmo não está preso a nenhuma suíte de escritório e, conseqüentemente, a suas versões. O formato é livre de *royalties* e não tem limite de reutilização.

A versão 1.0 do *ODF*, finalizada pelo *OASIS* em 2005 foi aprovada por unanimidade pela ISO, em Março de 2006, como Norma Internacional, a norma *ISO/IEC 26.300:2006*. Em Maio de 2008, o *ODF* foi aprovado e publicado pela ABNT como norma brasileira a NBR *ISO/IEC 26.300*.

Até a última versão da e-Ping, o formato *ODF* constava com o status de recomendado pelo documento, sendo facultativo aos órgãos públicos. Na versão 4.0 dos Padrões de Interoperabilidade de Governo

Eletrônico (e-PING), o *ODF* assume característica de adotado, dessa forma, torna-se obrigatório para guarda e troca de documentos eletrônicos entre todos os órgãos da administração direta, autarquias e fundações, sendo assim, deverão se enquadrar a essa regra a Universidade Federal de Santa Catarina e por ser extensão o PGCIN.

Em agosto de 2008, em Brasília, durante o Congresso Internacional Sociedade e Governo Eletrônico – CONSEGI 2008 (<http://www.consegi.gov.br>), diversas instituições assinaram o Protocolo Brasília, um documento público de intenções para adoção de formatos abertos.

2.6 Formato de Arquivos Abertos *PDF/A*

O formato *Portable Document Format (PDF)*, foi criado pela empresa *Adobe Systems* e aperfeiçoado durante os últimos 15 anos. Começou com o sonho de um escritório sem papel, como o projeto de estimulação de um dos fundadores da *Adobe*, o Sr. *John Warnock*. Inicialmente era um projeto interno da *Adobe* para criar um formato de arquivo para que documentos pudessem ser distribuídos por toda a empresa e exibidos em qualquer computador com qualquer sistema operacional.

Em 1990 o *Adobe® PostScript®* se estabelece notadamente como um padrão de impressão mundial conforme consta no sítio da *Adobe* mostrado na figura 2.



Figura 2: 1990 - O marco do *PostScript*

Fonte: <http://www.adobe.com/br/products/acrobat/adobepdf.html>

John Warnock em seu estudo chamado “projeto *Camelot*” divulgado em 1991 (figura 3), que desenvolveu o *PDF*, esboçou uma tecnologia que transformaria o modo de como as pessoas trabalhariam, pois através dessa tecnologia seria possível enviar mensagens de texto

completo e documentos gráficos (jornais, artigos de revistas, manuais técnicos.) através de redes de distribuição de correio eletrônico. Esses documentos podem ser visualizados em qualquer máquina e qualquer documento pode ser impresso localmente.

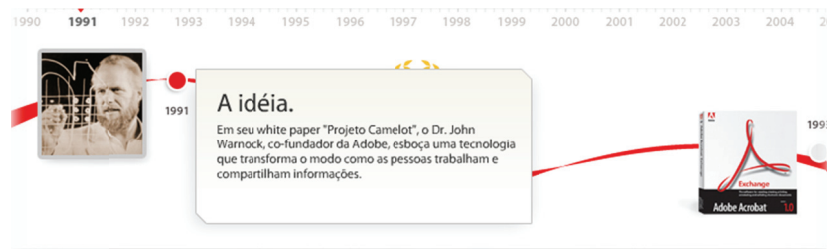


Figura 3: 1991- Projeto Camelot

Fonte: <http://www.adobe.com/br/products/acrobat/adobepdf.html>

A partir da tecnologia *postscript* que é uma linguagem de programação que define uma página e como ela deve ser impressa, e o *illustrator* era o aplicativo capaz de rodar praticamente em todos os microcomputadores disponíveis na época e capaz de abrir arquivos *postscript* mesmo que eles fossem criados em outra plataforma, os engenheiros da *Adobe* criaram o formato *PDF* que não deixa de ser um *postscript* melhorado, e um conjunto de aplicativos para criar, modificar e visualizar este formato de arquivo.

Conforme figura 4, em 1992 é divulgado o formato *Adobe PDF*, que recebeu o codinome “*Carrossel*”, e recebeu o prêmio “*Best of Comdex*” (*Comdex* é uma das maiores feiras de Tecnologia de Informação e Comunicação do mundo).



Figura 4: 1992 - divulgado o formato PDF

Fonte: <http://www.adobe.com/br/products/acrobat/adobepdf.html>

Em 1993, conforme figura 5, inicia-se a Geração do *Adobe Acrobat*, software utilizado para criar e visualizar documentos *PDF*.



Figura 5: 1993 - Inicia-se a geração Acrobat

Fonte: <http://www.adobe.com/br/products/acrobat/adobepdf.html>

Em 1994, conforme a figura 6, a *Adobe Systems* lança o *Acrobat 2.0* e oferece suporte a multimídia incorporado a *links* para arquivos externos em documentos *Adobe PDF*.



Figura 6: 1994 - Lançado a versão do Acrobat 2.0

Fonte: <http://www.adobe.com/br/products/acrobat/adobepdf.html>

Em 1994, conforme a figura 7, a Receita Federal dos Estados Unidos fornece folhetos e formulários de declaração de Imposto de renda no formato *PDF* em seu site com o objetivo de facilitar o *download*.



Figura 7: 1994 - *PDF* na Receita Federal USA

Fonte: <http://www.adobe.com/br/products/acrobat/adobepdf.html>

Em 1994, conforme figura 8, a *Adobe* começa a distribuir de forma gratuita o software *Acrobat Reader* o qual permite per arquivos *PDF*.



Figura 8: 1994 - É lançado o *Acrobat Reader* - leitor gratuito para *PDF*

Fonte: <http://www.adobe.com/br/products/acrobat/adobepdf.html>

Conforme figura 9, em 1995, foi divulgado o *plug-in* do *Acrobat* para o *Netscape*, aumentando assim a popularidade dos arquivos *PDF* no surgimento da internet.

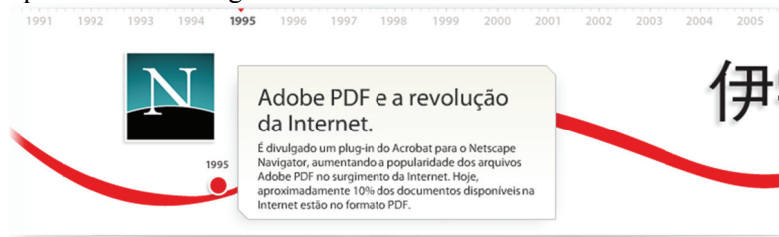


Figura 9: 1995 - *plug-in* para o *Netscape*

Fonte: <http://www.adobe.com/br/products/acrobat/adobepdf.html>

Conforme figura 10, em 1996, com o lançamento do *acrobat 3.0*, os documentos *PDF* passam a suportar fluxos de trabalho de produção e impressões completas com cores especiais, meios-tons, suporte a impressões sobrepostas e muito mais.



Figura 10: 1996 - *Acrobat 3.0*

Fonte: <http://www.adobe.com/br/products/acrobat/adobepdf.html>

Em 1997, com a vantagem do uso de *byte* duplo e o lançamento da versão em japonês, aumenta ainda mais a simpatia pelo formato *PDF* no mundo inteiro.



Figura 11: 1997 – Uso do byte do duplo

Fonte: <http://www.adobe.com/br/products/acrobat/adobepdf.html>

Em 1999, conforme figura 12, em documentos em *Adobe PDF* já é fazer anotações e revisar arquivos, restringir o acesso com o uso de senhas, incluir assinaturas digitais e capturar páginas *WEB*.



Figura 12: 1999 - Novos recursos de segurança para o formato PDF

Fonte: <http://www.adobe.com/br/products/acrobat/adobepdf.html>

Conforme figura 13, em 1999, o *American National Standards Institute (ANSI)* publica o primeiro padrão *PDF* para intercâmbio protegido de conteúdo impresso.



Figura 13: 1999 - ANSI publica padrão PDF

Fonte: <http://www.adobe.com/br/products/acrobat/adobepdf.html>

No ano de 2000, conforme a figura 14, uma versão em Adobe PDF do primeiro e-book de Stephen King, “*Riding the Bullet*”, é copiado por download 400.000 vezes em 24 horas, totalizando mais cópias do que as vendas do primeiro dia da versão impressa, demonstrando assim a popularidade do PDF.

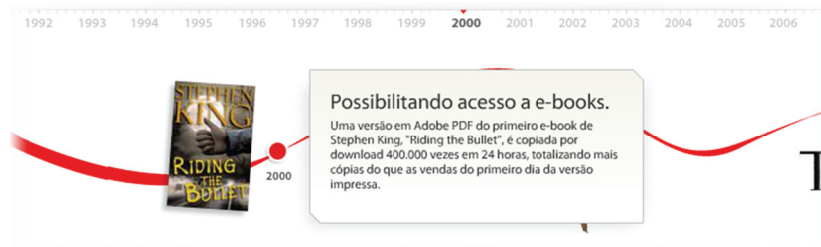


Figura 14: 2000 - PDF para acesso a e-book

Fonte: <http://www.adobe.com/br/products/acrobat/adobepdf.html>

Em 2003 conforme a figura 15, o *Adobe PDF* ganha suporte para *XML* para formulários e metadados, bem com a inclusão de recursos mais avançados, tornam o *Adobe PDF* ainda mais sofisticado.



Figura 15: 2003 - Suporte a XML

Fonte: <http://www.adobe.com/br/products/acrobat/adobepdf.html>

Em 2001, uma versão nova do *Acrobat* (aplicativo para criação de *pdf's da Adobe*) recebeu o *codinome "Brazil"* e trouxe uma série de modificações ao padrão necessárias para atender a indústria gráfica, a baixo custo, resolvendo definitivamente problemas de uniformidade de cor entre o que era mostrado no monitor vídeo e que era impresso.

Em setembro de 2005, a *Library of Congress*, a *National Archives e Records Administration (NARA)*, e várias empresas de TI, elegeram um novo formato de PDF para a preservação a longo prazo de documentos

eletrônicos, e conforme informação disponível no sítio da empresa Adobe, o formato *PDF* foi homologado pela norma *ISO 19005-1:2005* denominada de *PDF/Arquive*, ou simplesmente, *PDF/A-1* conforme consta no sítio da Adobe na internet mostrado na figura 16.



Figura 16: 2005 - Ano publicação *PDF/A*

Fonte: <http://www.adobe.com/br/products/acrobat/adobepdf.html>

O formato *PDF/A-1* baseia-se no formato *PDF 1.4 da Adobe Systems*, uma plataforma estável com formato independente de qualquer plataforma de software ou hardware que se utilize. O *PDF 1.4* foi implementado no *Adobe Acrobat 5*. Este formato teve de ser adaptado a novas exigências, que vão ao encontro de um formato de arquivo consistente e de uso generalizado. Essa norma descreve o que pode e o que não pode estar em um *PDF* para atender ao padrão, eliminando dos documentos códigos de programação, elementos externos e fontes não desejadas.

O formato *PDF/A-1* possui algumas características importantes, como armazenar no próprio documento tudo o que é necessário para visualizar e imprimir. Ele utiliza metadados *Extensible Metadata Platform (XMP)*, não admite encriptação, compressão *LZW* (por motivos de direitos de propriedade), arquivos incorporados, referências a conteúdos externos, transparências *PDF*, multimídia e *JavaScript*. A assinatura digital é suportada pelo *PDF/A-1*, desde que as fontes utilizadas estejam embutidas no formato.

A figura 17 mostra como verificar se o documento foi gerado no formato *PDF/A*.

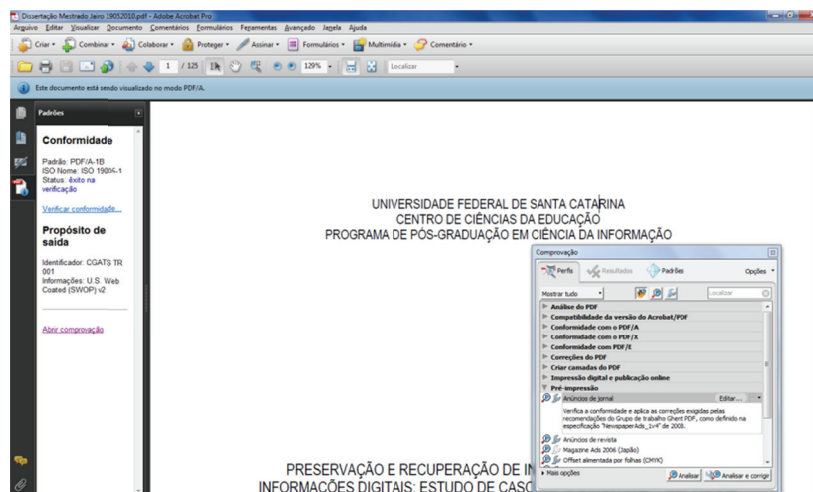


Figura 17: Arquivo gerado no formato PDF/A

O formato *PDF/A-1* divide-se em dois níveis: o *PDF/A-1a* e o *PDF/A-1b*, sendo que o *PDF/A-1a* assegura a estrutura lógica e semântica do documento e a sequência do texto, e o *PDF/A-1b* assegura apenas a aparência visual dos documentos digitais, sem garantir a coerência textual dos mesmos. Todavia, as diferenças entre estes dois níveis não têm qualquer significado para os documentos digitalizados, mas apenas para os documentos que existem somente no formato digital ou que tenham sido objeto de *Optical Character Recognition (OCR)*. Acrescente-se que nem todas as ferramentas de criação do formato *PDF/A-1* podem gerar documentos *PDF/A-1a* e *PDF/A-1b*. Na própria *Adobe Systems*, só a versão 8 do *Acrobat* é a que faz.

Um aspecto relevante no formato *PDF/A-1* é que o mesmo não constitui um sistema ou estratégia de arquivo, tampouco exclui outros formatos de arquivo, como é o caso do *TIFF*. Objetivamente, a norma internacional 19005-1 apenas identifica um perfil para documentos digitais que garante a sua inteligibilidade ao longo dos anos, ao arripio das mudanças tecnológicas, pelo que a utilização do formato *PDF/A-1* não dispensa a existência prévia de uma organização de arquivo, da qual, aliás, está dependente e é apenas parte, mas que pode ajudar a tornar mais eficaz. Ao mesmo tempo, oferece razões ao legislador para, finalmente, começar a considerar a preservação digital como uma alternativa capaz à preservação analógica.

Em 2007 a *Adobe* libera a especificação *PDF 1.7* para a *Association for Information and Image Management (AIIM)*.



Figura 18: 2007 - Adobe libera a especificação PDF 1.7 para a *AIIM*
 Fonte: <http://www.adobe.com/br/products/acrobat/adobepdf.html>

Em 2007, a *Adobe* passa a oferecer suporte ao padrão *Signatures and Authentication for Everyone (SAFE)* de assinaturas digitais.

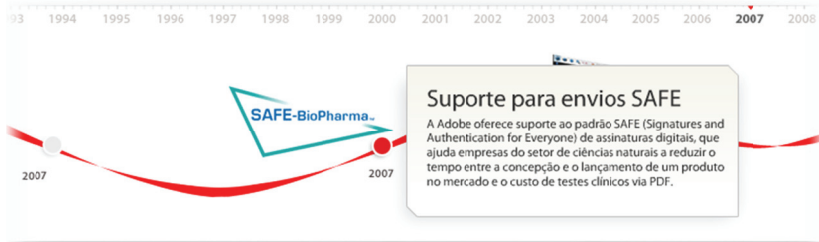


Figura 19: 2007 - Suporte para envios *SAFE*
 Fonte: <http://www.adobe.com/br/products/acrobat/adobepdf.html>

Em 2007, conforme figura, a *ISO* aprova o *PDF/E* como um formato de troca neutro e aberto para documentação técnica e de engenharia que ajuda na distribuição mais segura de informações confidenciais.



Figura 20: 2007 - Liberada a especificação *PDF/E* padrão para dados de engenharia.
 Fonte: <http://www.adobe.com/br/products/acrobat/adobepdf.html>

Com a abertura do formato *PDF* 1.7 e o reconhecimento e aceitação do formato pela *ISO*, em janeiro de 2008 conforme figura 21, o mesmo foi aprovado como padrão *ISO* 32000-1:2008. O padrão *ISO* 32000 continuar sendo desenvolvido com o objetivo de proteger a integridade e a longevidade do formato *PDF*, proporcionando um padrão aberto para mais de um bilhão de arquivos *PDF* existentes atualmente.



Figura 21: 2008 - PDF aprovado como padrão internacional

Fonte: <http://www.adobe.com/br/products/acrobat/adobe.pdf.html>

Em 2009 conforme a figura 22, o orçamento dos Estados Unidos é publicado como um documento *PDF* certificado e assinado digitalmente. O *PDF/A* é aceito para envio pela *National Archives and Records Administration dos Estados Unidos*, pelo *National Archives* da Suécia e pelo ministro Francês responsável pela energia nuclear, demonstrando a força e a aceitação deste formato de documento.



Figura 22: 2008 - Orçamento dos USA são publicados em PDF.

Fonte: <http://www.adobe.com/br/products/acrobat/adobe.pdf.html>

No Brasil não foi encontrado nenhuma norma para adoção do *PDF/A*, mas observa-se que existe um movimento crescente em diversos países que estão normatizando o *PDF/A* como padrão de arquivamento de documentos digitais.

A tendência entre as diversas bibliotecas pelo mundo estão padronizando o *PDF/A* como formato para arquivamento, dentre elas

pode-se destacar a Biblioteca Nacional Alemã (http://www.d-nb.de/eng/netzpub/ablief/np_dateiformate.htm) e a Biblioteca Nacional Austríaca (http://www.onb.ac.at/bibliothek/digitale_medien_informationen.htm).

Na França, a Direção de Modernização do Estado do Ministério do Orçamento (<http://www.modernisation.gouv.fr/>), emitiu uma recomendação no início de 2009 para o tratamento de dados eletrônicos. O documento recomenda a utilização da norma *ISO PDF/A* para arquivamento de documentos administrativos com conteúdo estático, inalterável.

O modelo *Model Requirements for the Management of Electronic Documents and Records (MoReq)*, ou seja, o modelo de requisitos para o gerenciamento eletrônico de documentos e registros é uma especificação europeia de documentos eletrônicos e gerenciamento de registros. O *MoReq* foi originalmente desenvolvido para troca de documentos padronizados entre a Comissão Europeia e os governos dos Estados-Membros. Conforme consta no site oficial do *MoReq* disponível em (<http://www.moreq2.de>), a nova versão do *MoReq2* inclui *PDF/A* na lista de formatos recomendados, por exemplo, documentos digitalizados e para arquivamento de longo prazo.

Conforme consta no site oficial da Câmara Federal de Arquitetos e Consultores de Engenharia da Áustria, disponível em (<http://www.baik-archiv.at>), exige que os documentos digitais colocados à disposição do público deverão estar de acordo com a norma *PDF/A-1b*. Além disso, a autenticidade dos documentos digitais que estão sendo adicionados ao cadastro, são assegurados através da utilização de uma assinatura digital.

O Governo Norueguês também regulamentou que, a partir de 01 de janeiro de 2009, conforme norma disponível no seu site em <http://www.regjeringen.no/en/dep/fad/pressesenter/pressemeldinger/2007/Open-document-standards-to-be-obligatory.html%3Fid%3D494810&prev=/search%3Fq%3D>, que todas as informações publicadas em sites estatais devem ser salvos em um formato de documento aberto e disponível, como *PDF/A* ou *ODF*.

O Conselho Federal Suíço em 2008, em um projeto de resolução para intercâmbio de arquivos digitais no âmbito dos processos administrativos que está disponível em <http://www.admin.ch>, determina que as comunicações eletrônicas trocadas entre o Estado e os cidadãos deverão ser feita utilizando o formato *PDF/A*.

2.7 Metadados XMP

O termo metadados significa literalmente, "dados sobre dados" e tem sido descrito como o cartão de visita dos documentos digitais. Metadados geralmente compreendem um conjunto de propriedades, onde cada propriedade tem um significado específico na *Extensible Metadata Platform - XMP*.

A especificação *XMP* inclui mais de uma dúzia de esquemas pré-definidos, com centenas de propriedades de documento comum e as características da imagem. O mais usado do esquema pré-*XMP* é o *Dublin Core* que inclui propriedades gerais, tais como Título, Criador, Assunto e Descrição. Além de esquemas pré-definidos esquemas personalizados podem ser definidos para cobrir as necessidades da empresa ou metadados específicos do setor.

XMP para documentos *PDF*, foi introduzido com o *Acrobat 5* e *PDF 1.4* em 2001, e são aplicados em todos os produtos editoriais Adobe e apoiado por dezenas de vendedores de software independentes e grupos de usuários. O *Adobe Bridge*, que faz parte do *Creative Suite*, lida com metadados *XMP* em vários formatos de arquivo.

Conforme consta do site do *PDFLIB* disponível em <http://www.pdfli.com/knowledge-base/xmp-metadata/>, informa que existem diversas normas *ISO* já publicadas ou previstas, que especificam subconjuntos *PDF* para certos domínios de aplicação, tais como a indústria das artes gráficas, de arquivamento ou a de engenharia. Exceto para os padrões pré-*PDF/X-1* e *X-3*, que foram introduzidas em 2001 e 2002, todas as normas *ISO* para arquivos *PDF* incluem o uso de metadados *XMP* (mesmo obrigatória na maioria dos casos, exceto *ISO 32000*):

- a) *PDF/A-1* in *ISO 19005-1* (publicado em 2005): formato de arquivo de documento para a preservação em longo prazo. O uso de *PDF 1.4*. *PDF/A-1* exige *XMP* para identificar arquivos conforme e suporta metadados *XMP* personalizado através de esquemas de extensão. *XMP* suporta dentro do *PDF/A-1* é baseado na especificação *XMP 2004* (ADOBE, 2004).
- b) *PDF/E-1* dentro da *ISO 24517-1* (publicado em 2008): Engenharia de formato de documento para engenharia - Uso de *PDF 1.6*. *XMP* utilizado no *PDF/E* é quase idêntico ao *PDF/A-1*, exceto que ele é baseado na mais recente especificação *XMP 2005* (ADOBE, 2005).
- c) *PDF/X-4* na *ISO 15930-7* (publicado em 2008): Troca completa dos dados de impressão (*PDF/X-4*) e troca parcial de impressão de dados com referência perfil

externo (*PDF/X-4p*) usando *PDF 1.6*. Semelhante ao *PDF/A-1*, *XMP* é necessária para expressar as normas em conformidade *PDF/X-4*. *XMP* apoio *PDF/X-4* é baseado na especificação *XMP 2005*.

- d) *PDF/X-2* na *ISO 15930-5* (publicado em 2003) e *PDF/X-5* na *ISO 15930-8* (publicado em 2008): Troca parcial dos dados de impressão utilizando *PDF 1.6* (*PDF/X-5*). *PDF/X-2* e *X-5* e outros documentos de referência documentos *PDF/X*, onde o alvo de tal referência é identificado usando várias entradas *XMP*. Isso faz com que *XMP* um componente crucial da *PDF/X-2* e *X-5*.
- e) *ISO 32000* (publicada em 2008): A gestão de documentos - *Portable Document Format - PDF 1.7*. *ISO 32000* é a versão padrão do *PDF 1.7*. O conteúdo técnico é idêntico ao *PDF 1.7* (o formato de arquivo do *Acrobat 8*), que apoia plenamente os metadados *XMP*.

O *Dublin Core* é um dos mais conhecidos esquemas de metadados *XMP*, foi padronizado como *ISO 15836* (publicada em 2003).

2.8 A Recuperação da Informação e a Ciência da Informação

De acordo com a revisão da literatura, parece haver consenso entre os autores quanto ao surgimento da Ciência da Informação. O surgimento dessa ciência foi decorrente de um “boom científico”, posterior à Segunda Guerra Mundial, que teve como marco inicial a reunião realizada em 1962 na *Georgia Institute of Technology*.

A Ciência da Informação desenvolveu-se principalmente na Rússia e nos Estados Unidos, pois nestes países, a informação foi considerada estratégica e assunto de Estado, além de ter sido necessário minimizar os custos de tratamento, operacionalização, transmissão, recuperação e aproveitamento de grandes estoques de informação.

Observa-se que a partir do surgimento da Ciência da Informação ocorreram grandes transformações na sociedade contemporânea que passou a considerar o conhecimento, a comunicação, os sistemas de significados e os usos de linguagens como objetos de pesquisa científica e domínios de intervenção tecnológica.

Durante o desenvolvimento da Ciência da Informação, houve o surgimento de correntes de pensamento que estimulou discussões teóricas, que evidenciaram a necessidade de definição da abrangência da Ciência da Informação, bem como a sua vinculação com outras ciências.

A definição de Ciência da Informação apresentada por Borko (1968, p.3) embora não seja um consenso, é uma das mais difundidas. Ele definiu a Ciência da Informação como uma ciência interdisciplinar, ou seja, um campo autônomo que tem como objetivo investigar as propriedades e o comportamento da informação, as forças que governam seu fluxo, e os meios para processá-la visando aperfeiçoar sua acessibilidade e uso.

Saracevic (1996, p. 47), definiu a Ciência da Informação como:
[...] um campo dedicado a questões científicas e a prática profissional, voltadas para os problemas da efetiva comunicação do conhecimento e de registros do conhecimento entre seres humanos, no contexto social, institucional ou individual do uso e das necessidades de informação. No tratamento dessas questões são consideradas de particular interesse as vantagens das modernas tecnologias informacionais.

A Ciência da Informação de acordo com Saracevic (1996), contem em seu núcleo a recuperação da informação como causa de seu surgimento.

De acordo com Saracevic (1996, p. 46), a Ciência da Informação tem como foco de estudo as propriedades e comportamento da informação, com as forças que regem seu fluxo e com os meios de processá-la para facilitar seu acesso e uso. Dessa forma, tem como objetivo principal investigar e mapear essas propriedades da informação pela aplicação da teoria da informação, da teoria das decisões e outros construtos da ciência cognitiva, da lógica e/ou da filosofia. Saracevic (1996) também identificou três características marcantes na evolução da Ciência da Informação: interdisciplinaridade, vinculação com a tecnologia e participação ativa na era da informação.

Saracevic constatou também que a Ciência da Informação tem mantido vínculos através dos tempos mais acentuadamente com a Biblioteconomia, Ciência da Computação, Ciência Cognitiva e Comunicação. Cabe destacar que ainda é muito comum até mesmo dentro da comunidade acadêmica, a Ciência da Informação e Biblioteconomia sejam confundidas ou consideradas como uma mesma ciência.

O caráter interdisciplinar da Ciência da Informação pode ser constatado no perfil dos alunos do curso de mestrado em Ciência da Informação da UFSC. O curso que inicialmente era formado por profissionais oriundos da área da Biblioteconomia, todavia, com o passar

do tempo cresceu o interesse de profissionais de Ciências da Computação, Arquitetura, Engenharia, Comunicação, Administração, Arquitetura e Letras.

Para Pinheiro (1999, p.155), a Ciência da Informação tem seu próprio estatuto científico: como ciência social, é interdisciplinar por natureza, seu objeto de estudo, considerando sua característica abstrata, é de difícil apreensão; apresenta interfaces com a Biblioteconomia, a Ciência da Computação, a Ciência Cognitiva, a Sociologia da Ciência e Comunicação, entre outras áreas; e provém da bifurcação da Documentação/Bibliografia e da Recuperação da Informação.

Segundo o entendimento de Barreto (1999, p.1), a Ciência da Informação cada vez mais terá seus caminhos relacionados aos das estruturas e dos fluxos de informação.

De acordo com Gómez (2000), essa diversidade de condições epistemológicas não deve ser confundida, com uma indefinição metodológica eclética ou relativista. A Ciência da Informação recebe das Ciências Sociais seu traço identificador, que serve de princípio articulador dessa diversidade, e que corresponde ao que nos estudos metodológicos se denomina como a “dupla hermenêutica”. Sendo assim, no lugar da escolha a priori da perspectiva social, bem como pela necessidade imposta pela dupla hermenêutica, que se referem à dupla aderência as necessidades da pesquisa e ao contexto sócio-política, os temas guiarão os projetos de pesquisa.

2.9 Recuperação da Informação

Conforme matéria publicada no *The New York Times* (June 17, 2008), no século XIX, o advogado belga Paul Otlet, em parceria com Henry La Fontaine, com o intuito de disponibilizar o conhecimento existente de uma forma mais acessível, buscaram apoio do Governo da Bélgica para desenvolverem o projeto de construir uma "cidade do conhecimento". O projeto era ousado, pois consistia em criar uma grande bibliografia de todo o conhecimento publicado no mundo. O trabalho desses cientistas iniciou com a coleta de dados de todos os livros já publicados, juntamente com uma vasta coleção de revistas e artigos de jornais, fotografias e pôsteres.

Paul Otlet passou parte de sua vida desenvolvendo técnicas para registrar e recuperar informações. Um dos resultados desse trabalho são as fichas catalográficas padronizadas (12,5 x 7,5 cm), a microficha, a bibliografia universal, a classificação universal, dentre outros

instrumentos. Paul Otlet frente à dificuldade de recuperar informação em decorrência ao grande volume de papéis e livros, bem como as restrições de espaço físico, começou a pesquisar soluções nesse sentido, pensando inclusive numa solução que seria uma espécie de computador, que por meio da manipulação de rodas e raios moveria os documentos na superfície de uma mesa.

Ele também escreveu diversos documentos sobre a possibilidade do armazenamento eletrônico, dentre eles o livro “*Monde*” publicado em 1934, no qual relata a sua visão de um “cérebro mecânico coletivo” que contemplaria toda a informação do mundo, acessível instantaneamente em uma rede global de informação. Outra ideia interessante para recuperação de informação desenvolvida por Paul Otlet foi criação de um tipo de hipertexto, onde previa a ligação de documentos, com um diferencial no qual os links carregavam um significado anotado, como por exemplo, se os documentos concordavam ou discordavam entre si.

De forma semelhante, o americano *Vannevar Bush* (1945) arquitetou o “*memex*” como um sistema mecânico capaz de reproduzir as conexões mentais “*as we may think*” realizadas pelo homem para facilitar a recuperação da informação.

Apesar das pesquisas e dos conhecimentos gerados por *Paul Otlet*, *Henry La Fontaine* e *Vannevar Bush*, somente em 1951 a expressão recuperação da Informação (*information retrieval*) foi batizada pelo pesquisador americano Calvin Northrup Mooers, definindo-a como:

A Recuperação da Informação trata dos aspectos intelectuais da descrição da informação e sua especificação para busca, e também de qualquer sistema, técnicas ou máquinas que são empregadas para realizar essa operação. (MOOERS, 1951 apud SARACEVIC, 1996, p. 44).

A recuperação de informação inicialmente era considerada como objeto de interesse apenas de bibliotecários e especialistas em informação, mas com o crescimento dos estoques de informações disponibilizadas pelas Tecnologias de Informação e Comunicação (TIC's), e principalmente com o advento da Internet, despertou o interesse de profissionais das mais variadas áreas.

A recuperação da informação era considerada como um recurso estratégico para vários governos e empresas, mas foi com a expansão do uso dos computadores e o acesso a internet, bem como um crescimento muito grande dos estoques de informação no formato digital, é que aumentaram os problemas de recuperação de informação.

Com o descobrimento de que a recuperação de informação era um negócio rentável, houve um crescimento do interesse e de investimentos sem precedentes nessa área.

2.9.1 Definição de Recuperação de Informação

O termo recuperação de informação possui muitas definições, sendo que a própria palavra informação tem um conceito ambíguo, pois no contexto da recuperação de informação, o significado da palavra informação não tem uma definição exata, ou seja, em alguns os casos essa palavra pode ser substituída por documento. No entanto, o termo recuperação de informação é amplamente aceito na literatura sobre esse tema.

Segundo Baeza-Yates e Ribeiro-Neto (1999, p.1), a recuperação, a representação, o armazenamento, a organização e o acesso são processos de gestão da manipulação da informação. Pode-se definir “recuperação de informação” como o procedimento pelo qual a partir de uma necessidade de informação, busca-se uma informação em meio a um emaranhado de documentos dos mais variados tipos.

Um sistema de recuperação de informações é o responsável pelo armazenamento, recuperação e gerenciamento de informações em diferentes tipos de documentos, tendo como objetivo informar a existência e localização de documentos que possam conter a informação necessária e não necessariamente recuperar a informação.

A palavra “informação” vem do latim “*informatio*” que significa a ação de formar, representação, esboço, plano, ideia, concepção. No dicionário Aurélio, consta que informação é:

1. Ato ou efeito de informar-se; informe;
- 2- Dados sobre alguém ou algo;
3. Instrução Direção;
4. Conhecimento extraído dos dados; e,
5. Resumo dos dados.

De acordo com Le Coadic (1994, p.7), “[...] a informação é um conhecimento inscrito (gravado) sob a forma escrita (impressa), oral ou audiovisual”. Como a sugerir que os documentos inscrevem informações, mas eles são ao mesmo tempo, objetos autônomos.

Para Gonzalez (2000), a essência da recuperação de informação consiste na busca de documentos relevantes a uma dada consulta que expressa à necessidade de informação do usuário. Assim, a indexação dos conteúdos deve conter estrutura adequada e utilizar a perfeita

adequação entre linguagem natural e a linguagem controlada, onde os termos precisam obedecer à classificação prévia de forma a resultar no perfeito entendimento do sistema.

Segundo Cardoso (2000, p.1), a recuperação da informação é uma subárea da ciência da computação, que estuda o armazenamento e recuperação automática de documentos, que são objetos de dados, geralmente textos.

Para Choo (2006), recuperar uma informação é disponibilizá-la ao usuário, que a solicitou por necessidades espontâneas e/ou induzidas, objetivando construir significado, produzir novo conhecimento e tomar decisões, sejam administrativas, sejam pessoais.

A recuperação da informação é muitas vezes tratada como sinônimo de busca de informação, porque a necessidade de informação é que dispara o processo de busca da informação.

Ingwersen (1982, p.167) propõe uma sequência de nove etapas para identificar o processo mental no processo da recuperação da informação:

- 1) A necessidade de informação do usuário;
- 2) A questão sobre a informação formulada;
- 3) A negociação usuário-bibliotecário;
- 4) A formulação da estratégia de busca – análise do tópico;
- 5) A escolha das ferramentas de busca;
- 6) A procura na lista alfabética ou sistemática;
- 7) O julgamento baseado no índice (termos);
- 8) O julgamento baseado na descrição, resumos e títulos; e,
- 9) A avaliação do documento pelo usuário-bibliotecário.

A Informação é objeto de estudo tanto da Ciência da Informação como da Ciência da Computação, sendo que a Recuperação de Informação poder ser considerada como um elo entre essas duas ciências. A popularização da Internet só fez aumentar o interesse nessas ciências, não só pelo interesse científico, mas também pelo interesse comercial face as suas inúmeras aplicações.

Com a popularização dos computadores, e também com o aumento da capacidade computacional dos mesmos, tornou-se viável a execução de algoritmos complexos de recuperação e de ordenação em bases de dados. Cada vez mais os investimentos em pesquisas nesta área se disseminam tanto na iniciativa privada como no setor público. As pesquisas acadêmicas em bibliotecas digitais seguem em larga escala,

oferecendo um campo para explorar a descoberta e a recuperação em rede em ambiente controlado.

2.9.2 Modelos de Recuperação de informação

Existem vários modelos de recuperação de informação, como pode ser observado na figura 23. O modelo booleano, o modelo vetorial e o modelo probabilístico são os mais conhecidos e também são considerados como os modelos clássicos de recuperação, sendo que para cada um deles, existem modelos alternativos que visam estendê-los em funcionalidade e desempenho. Nesta dissertação serão abordados alguns dos modelos clássicos de Recuperação de Informação:

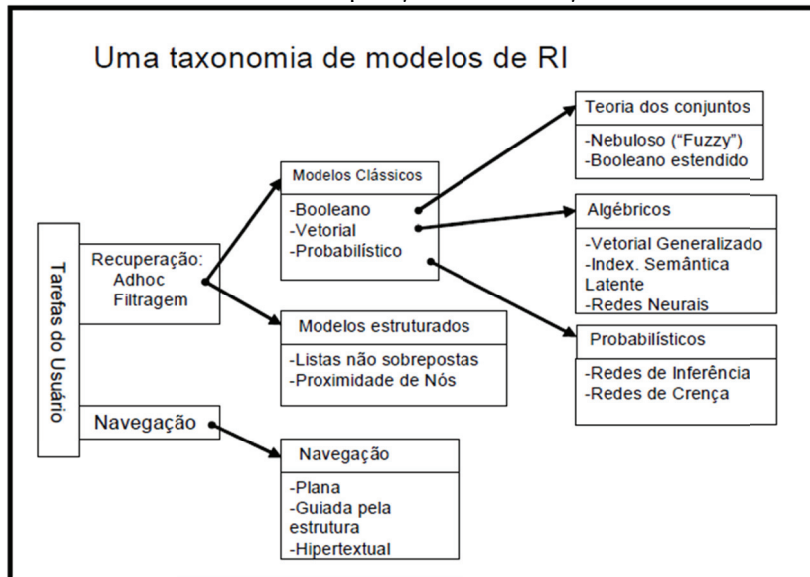


Figura 23: Uma taxonomia de modelos de Recuperação de Informação (adaptado de BAEZA-YATES; RIBEIRO NETO, 1999, p. 21).

2.9.3 Modelo Booleano

O modelo booleano foi um dos primeiros modelos de Recuperação da Informação, sendo muito utilizado até meados da década de 1990. Neste modelo, um documento é considerado relevante ou não relevante a uma consulta; não existe resultado parcial e não há informação que permita a ordenação do resultado da consulta, o que pode ser considerado uma de suas principais desvantagens. Cabe

salientar que, a ordenação por ordem de relevância é uma das características importantes dos sistemas de recuperação de informação.

A origem desse modelo tem como base a teoria dos conjuntos, e utiliza os operadores booleanos *or*, *and* e *not* para estabelecer relações específicas de ocorrência com as palavras-chave, de forma a especificar os documentos a serem recuperados. Ou seja, para cada consulta são recuperados todos os documentos que possuem os termos nas condições especificadas pelo usuário.

Pelo fato deste modelo trabalhar com operadores booleanos, requer que os usuários tenham pelo menos algum conhecimento de lógica booleana, fator este, que dificulta sobremaneira a utilização de pela grande maioria dos usuários.

Segundo Baeza-Yates e Ribeiro-Neto (1999), “a grande vantagem desse modelo é a clareza do seu formalismo e a sua simplicidade”.

Dentre os modelos alternativos ao booleano pode-se destacar a lógica difusa ou nebulosa (*fuzzy*) e o booleano estendido. *Fuzzy* em inglês significa incerto, duvidoso.

De acordo com Vanderlei Filho et al. (2002), a Lógica Difusa foi estruturada em 1965 pelo *Dr. Lofti A. Zadeh* da Universidade da Califórnia, tendo como objetivo principal tratar e representar incertezas, ou seja, possibilita inferir conclusões e dar respostas a informações incertas, imprecisas, vagas. Essa lógica permite representar valores de pertinência (grau de verdade) intermediários entre os valores de verdadeiro e falso da lógica clássica. A Lógica Difusa ou lógica *fuzzy* pode ajudar a tratar melhor as incertezas quanto à relevância dos termos dos índices em relação aos documentos e a importância dos termos de entrada para a consulta. Na teoria dos conjuntos difusos ou lógica *fuzzy* o objetivo é capturar e operar com a diversidade, a incerteza e as verdades parciais dos fenômenos da natureza de uma forma sistemática e rigorosa. A teoria dos conjuntos *fuzzy* baseia-se no fato de que os conjuntos existentes no mundo real não possuem limites precisos, isto é, o modelo *fuzzy* trabalha na possibilidade de que o resultado esteja parcialmente inserido no conteúdo consultado. Para essa operação é utilizado grau de pertinência aos conjuntos. O modelo usa uma matriz de correlação termo a termo para calcular as relações entre os termos dos conjuntos de documentos e apresentar o resultado.

O modelo booleano estendido foi introduzido em 1983 por *Salton, Fox, e Wu*. Esse modelo se diferencia do modelo booleano por usar diferentes operadores, associar pesos aos termos de cada documento, e por implementar uma função de ordenação, ou seja, tenta unir a potencialidade das expressões booleanas através da introdução do

conceito de relevância com a precisão do modelo vetorial através do uso dos operadores booleanos. Neste modelo os problemas referentes a decisões binárias do modelo clássico por meio da aferição de pesos aos termos, aproximando o modelo original do modelo vetorial.

2.9.4 Modelo Vetorial

Segundo Buckley (1985), o modelo vetorial foi idealizado por Gerard Salton, e foi inicialmente utilizado num projeto chamado *System for the Manipulation and Retrieval of Text (SMART)*. Este projeto iniciou em 1961 na *Universidade de Harvard* e mudou-se para a *Universidade de Cornell* após 1965.

De acordo com Baeza-Yates e Ribeiro-Neto (1999), o modelo vetorial baseia-se na comparação parcial entre a representação dos documentos e da consulta do usuário, onde são atribuídos pesos aos termos de indexação presentes na consulta, em função da frequência de ocorrência no documento.

Neste modelo, cada documento é representado como um vetor de termos, e cada termo possui um valor associado que indica o peso deste no documento, ou seja, cada documento possui um vetor associado que é composto por pares de elementos na forma $\{(palavra_1, peso_1), (palavra_2, peso_2), \dots, (palavra_n, peso_n)\}$, onde cada elemento deste vetor de termos é considerado uma coordenada dimensional.

Desta forma, os documentos podem ser colocados em um espaço euclidiano de n dimensões (onde n é o número de termos) e a posição do documento, são obtidas pelo seu peso em cada dimensão, ou seja, documentos que possuem os mesmos termos acabam sendo colocados em uma mesma região do espaço e, o que teoricamente tratam de assuntos similares. Os documentos mais similares à consulta podem ser considerados mais relevantes para o usuário e retornados como resposta para ela.

Uma das formas de calcular a proximidade entre os vetores é testar o ângulo entre estes vetores. No modelo original, é utilizada a função cosseno que calcula o produto dos vetores de documentos através da seguinte fórmula:

$$sim(x, y) = \frac{\sum_{i=1}^t (w_{i,x} \times w_{i,y})}{\sqrt{\sum_{i=1}^t (w_{i,x})^2} \times \sqrt{\sum_{i=1}^t (w_{i,y})^2}}$$

Onde $w_{i,x}$ é o peso do i -ésimo elemento do vetor x e $w_{i,y}$ é o peso do i -ésimo elemento do vetor y .

O objetivo do modelo vetorial consiste em estabelecer as características que melhor descrevem o documento e quais são as características que melhor distinguem o documento com relação ao restante da coleção, permitindo dessa forma uma quantificação de similaridade.

Exemplificando:

Fator TF = frequência direta de um termo dentro de um documento = contagem; fornece a medida de quão bem esse termo descreve o documento;

Fator iTF = frequência inversa = quantas vezes essa palavra aparece nos outros documentos;

Busca Vetorial = Fator TF * Fator iTF

Relevância = (Máximo de vezes que o termo aparece no documento / Máximo de vezes que um termo pode ter em um documento)

Quadro 1: Cálculo de relevância

As principais vantagens do modelo vetorial são a sua simplicidade e a facilidade que ele possui para computar as similaridades de forma eficiente através da atribuição de pesos e também o fato de que o modelo se comporta muito bem com as coleções genéricas, recuperando pelo menos documentos que se aproximam do resultado.

O modelo vetorial é amplamente utilizado por sistemas de recuperação de informações com foco na recuperação de informação na internet, embora estes também utilizem outras técnicas para determinar o ranking de documentos como resposta a uma consulta.

Modelos que se propõem a estender a funcionalidade do modelo vetorial:

a) Redes neurais - Uma rede neural consiste na representação gráfica da rede de interconexões de neurônios no cérebro humano, onde os nós dessa representação gráfica são as unidades de processo enquanto que as extremidades desempenham a função de uma conexão sináptica (região de encontro de duas células nervosas). Simular o fato que a força de uma conexão sináptica em um cérebro humano muda a todo tempo, um peso é nomeado a cada extremidade em nossa rede de neurônios. A cada instante, os estados dos nós são definidos através dos níveis de

ativação. Dependendo deste nível de ativação, o nó A pode enviar um sinal ao seu vizinho, nó B. A força deste sinal até o nó B, depende do peso associado às extremidades entre o nó A e B.

Segundo Baeza Yates e Ribeiro Neto (1999), dentro de um sistema de recuperação de informação, vetores dos documentos são comparados com vetores das consultas para o cálculo do ranking. Assim, os termos índices nos documentos e consultas têm que ser combinados e pesados para o cálculo dessa posição. O modelo de recuperação de informação baseado em redes neurais busca encontrar padrões entre as consultas e dos documentos. Cada consulta realizada envia um sinal que ativa os termos índice, que por sua vez propagam os sinais aos documentos relacionados. Estes, por sua vez, retornam os sinais a novos termos-índices, em interações sucessivas. O conjunto resposta é definido através desse processo, e pode conter documentos que não compartilhem nenhum termo índice com a consulta, mas que tenham sido ativados durante o processo.

b) Vetorial generalizado - Em 1985, Wong, Ziarko, e Wong (apud Baeza-Yates e Ribeiro Neto, 1999) propuseram uma interpretação em que os vetores de termos do índice são assumidos como linearmente independentes, mas não são ortogonais em pares. Essa interpretação é menos restritiva que a convencional, na qual os vetores de termos do índice são além de linearmente independentes, também são ortogonais. O sistema de pesos neste modelo combina o sistema tradicional de termos nos documentos com fatores de correlação entre os termos.

c) Indexação semântica latente – Alguns pesquisadores utilizam *Análise Semântica Latente - LSA e a Indexação Semântica Latente - LSI* como sinônimos, mas outros apresentam *LSI* como uma técnica que utiliza *LSA* para indexação automática de documentos textuais. A indexação semântica latente – *LSI (Latent Semantic Indexing)* é definida como uma técnica automática que analisa as coocorrências de termos em documentos textuais com vistas a descobrir relacionamentos latentes entre eles (Deerwester, Dumais et al., 1990).

2.9.5 Modelo Probabilístico

Segundo Takao (2001), o modelo probabilístico de recuperação de informação foi proposto em 1960 por Maron e Kuhns. Esse modelo tem a denominação de probabilístico porque trabalha com conceitos provenientes da área de probabilidade e estatística, tendo como base o princípio da ordenação probabilística (*Probability Ranking Principle*).

Esse princípio consiste na hipótese de que a relevância de um documento para uma determinada consulta é independente dos outros documentos, ou seja, busca-se saber a probabilidade de um documento ser ou não relevante para uma consulta. Tal informação pode ser obtida assumindo-se que a distribuição de termos na coleção seja capaz de informar a relevância provável para um documento qualquer da coleção.

Existem diversas formas de se obter estatisticamente essa informação, porém, a base matemática comumente adotada para o modelo é o teorema de *Bayes*, sendo muitas vezes chamado de modelo bayesiano. A teoria de *Bayes* auxilia a identificação em cada termo da consulta o grau de relevância e de irrelevância do documento, selecionando o mais adequado, ou seja, o que produz menor erro para o somatório final, já que o grau final de probabilidade de relevância é dado pelo somatório dos graus de relevância de cada termo.

No modelo probabilístico a função de similaridade pode aproveitar-se das informações estatísticas de distribuição dos termos contidos no índice. Com isso, determinados parâmetros podem ser ajustados de acordo com a coleção em questão, obtendo assim resultados mais relevantes.

Cada documento é modelado novamente como um vetor de características na forma $x = (x_1, x_2, \dots, x_n)$, onde cada x representa um termo e contém informação sobre sua ausência ou presença no documento (forma binária).

A identificação dos documentos relevantes a uma consulta é determinada pelo cálculo de probabilidade de cada um dos documentos da coleção ser relevante à consulta dada, onde os documentos são então listados de acordo com o seu grau provável de relevância, ou seja, na forma de um ranking.

A probabilidade de relevância de um documento é calculada através da identificação de sua relevância ou não à determinada consulta; para cada termo da consulta seu grau de relevância é identificado no documento. A informação de relevância de um termo é calculada estatisticamente com base na frequência desse termo nos documentos da coleção.

Nessa fórmula, $P(w_i)$ é a probabilidade a priori de relevância (quando $i=1$) ou de irrelevância (quando $i=2$); $P(x/w_i)$ é a aparência de relevância ou irrelevância, dado um termo x .

$$P(w_i / x) = \frac{P(x/w_i)P(w_i)}{P(x)} \quad i = 1, 2$$

Já o $P(x)$ é calculado pela seguinte fórmula.

$$P(x) = \sum_{i=1}^2 P(x/w_i)P(w_i)$$

Essa fórmula calcula a probabilidade de observação aleatória de x que pode ser tanto relevante quanto irrelevante.

Segundo (*Rijsbergen*, 1999), o modelo probabilístico é um modelo bastante próximo ao modelo difuso, porém é necessário que algumas regras probabilísticas sejam satisfeitas durante a consulta.

O modelo probabilístico é um dos poucos modelos que não necessita de algoritmos adicionais para associação de peso aos termos para serem implementados, e os algoritmos de ordenação dos resultados são completamente derivados de sua teoria.

Os modelos que procuram ampliar o escopo do modelo probabilístico são os seguintes:

a) Redes de inferência - Esse modelo amplia o modelo probabilístico tratando o processo de recuperação de informação como um processo de raciocínio baseado em evidências representadas em documentos, sendo que essas evidências devem ser utilizadas para estimar a probabilidade da informação a ser encontrada pelo usuário. Assim, redes de inferência são projetadas para incorporar diferentes fontes de evidência ao estimar a probabilidade de relevância de um documento específico para o usuário. Ao estimar probabilidades usando fontes de evidência, redes de inferências estendem o modelo probabilístico clássico.

Em uma dada consulta, se um documento específico for considerado relevante para mesma, é criada uma variável aleatória e associada a este relacionamento. Essas variáveis podem ser alteradas de acordo com os eventos futuros de forma a estabelecer relacionamentos baseados nos eventos observados.

b) Redes de crença - são representadas por um Grafo Acíclico Dirigido (GAD), o qual exhibe o relacionamento de causa e efeito entre diversas variáveis. Esse modelo, de forma similar às redes de inferência, os documentos e consultas são modelados como subconjuntos de um espaço de conceitos. Cada consulta é mapeada no espaço de conceitos, que, por sua vez, está conectado ao espaço de documentos.

2.10 Indexação Automática de Textos

Segundo Buckley (1996), o termo Indexação Automática foi introduzido por *Gerard Salton* quando na década de 1960, desenvolveu um sistema de recuperação da informação (SRI) denominado *SMART*.

A indexação, dentro de um contexto de recuperação de informação é o processo de identificação das características de um documento posteriormente inseridas em uma estrutura de índice, para o sistema de recuperação de informações possa localizar rapidamente um documento a partir de parâmetros informados em uma consulta. Esse índice é construído por meio de um processo de indexação que poderá ser manual ou automático.

No processo de indexação automática de textos busca-se identificar termos (palavras) relevantes nos documentos dentro de uma coleção de documentos, e para depois inseri-las em índice. A identificação de termos simples ou compostos, a remoção de *stopwords* (palavras irrelevantes), o *stemming* (normalização morfológica) e a seleção de termos são consideradas as etapas que compõe o processo de indexação. Para cada uma dessas etapas existem diversas técnicas. Dependendo da situação, a ordem de aplicação dessas etapas pode variar ou alguma delas pode não ser utilizada.

Normalmente os substantivos são palavras muito representativas no conteúdo de um documento, por esse motivo, geralmente, vale a pena fazer um pré-processamento do texto dos documentos contidos em uma coleção, para determinar os termos a serem usados como termos de índice.

A utilização de todos os termos de uma coleção para indexar seus documentos, pode gerar muito ruído no processo de recuperação. Uma das maneiras de redução desse ruído consiste em reduzir o conjunto de termos (palavras) que serão utilizadas para indexar os documentos. Assim, o pré-processamento dos documentos de uma coleção, poderia ser visto simplesmente como um processo de controlar o tamanho do vocabulário, ou seja, controlar o número das palavras distintas usadas como termos de um índice. Em consequência da utilização de um vocabulário controlado, um usuário pode ser surpreendido pela recuperação de alguns documentos e com a ausência de outros documentos que ele esperava ver.

A indexação de todas as palavras, apesar de apresentar um índice com mais interferência, torna a tarefa de recuperação mais simples e mais intuitiva. Além do pré-processamento do documento, outras

técnicas podem ser utilizadas com o objetivo de melhorar o desempenho da recuperação da informação.

A compressão de documentos também é uma técnica utilizada para melhorar o desempenho da recuperação da informação, reduz consideravelmente o tamanho do documento, pois um texto comprimido ocupa menos espaço de armazenamento e é transmitido mais rapidamente. A desvantagem é o tempo gasto para compressão e descompressão. Porém, as técnicas modernas de compressão estão mudando essa visão, pois as mesmas estão provendo grande velocidade de compressão, e maior ainda de descompressão e rápido acesso aleatório sem a necessidade de decodificar o texto comprimido desde o início e procura no texto comprimido sem a necessidade de descomprimir o mesmo.

A indexação é tida como o processo de mapeamento dos termos dos documentos, onde a função de similaridade irá comparar os termos da pergunta com os termos presentes nos documentos, e assim localiza os documentos que contenham o assunto desejado pelo usuário.

2.10.1 Arquivos Invertidos

De acordo com *Baeza-Yates* e *Ribeiro-Neto* (1999), A indexação automática possui quatro etapas básicas a serem seguidas: identificação de palavras, remoção de *stopwords*, *stemming* e formação de frases-termo. Após a realização dessas etapas, os termos resultantes são armazenados em um arquivo de índice utilizando uma estrutura de arquivo invertido.

A estrutura de um arquivo invertido é composta por dois elementos: o vocabulário e as ocorrências. O vocabulário é o conjunto de todos os termos (palavras) diferentes no texto. Para cada palavra, uma lista de todas as posições onde o texto aparece é criada, o conjunto de todas as listas é chamado de ocorrências. As posições podem se referir às palavras ou caracteres. Posições das palavras (posição *i* se refere a *i*-th palavras). Os termos (palavras) são convertidos para letra minúscula e algumas não são indexadas. As ocorrências apontam para as posições dos caracteres no texto.

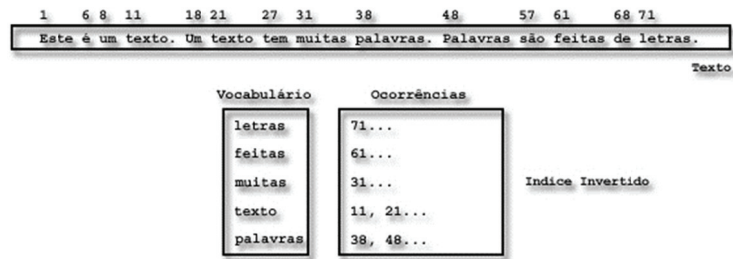


Figura 24: Estrutura de arquivo invertido.

O espaço necessário para o vocabulário é bastante pequeno, o vocabulário cresce como $O(nB)$, onde B é uma constante entre 0 e 1, estando entre 0.4 e 0.6 na prática. Por exemplo, para 1Gb de uma coleção o vocabulário terá o tamanho de apenas 5Mb. Este tamanho pode ser reduzido com a utilização de outras técnicas de normalização e *Stemming*.

Baeza-Yates e Ribeiro-Neto (1999) relatam que as ocorrências demandam muito espaço, pois cada palavra que aparece no texto é referenciada uma vez nessa estrutura, o espaço extra é $O(n)$. Mesmo não considerando as *stopwords*, na prática o espaço extra nas ocorrências é entre 30% e 40% do tamanho do texto.

Para se reduzir o espaço utilizado, pode ser utilizada uma técnica chamada endereçamento de bloco em vez de posições exatas, onde o texto é dividido em blocos, e as ocorrências apontam aos blocos onde a palavra aparece. Os índices clássicos que apontam para as ocorrências exatas são chamados de indexação total invertida (*full inverted index*).

A figura 25 apresenta um texto dividido em quatro blocos, onde as ocorrências denotam números de blocos.

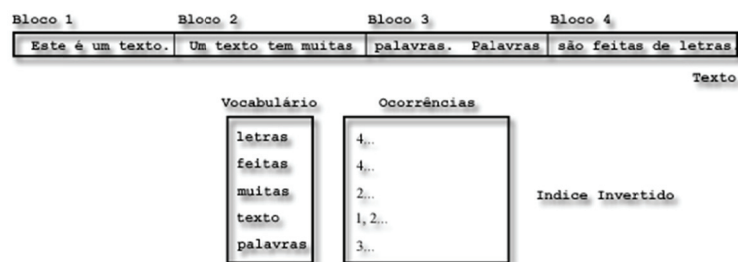


Figura 25: Estrutura de arquivo invertido dividido em quatro blocos.

2.10.2 Identificação das Palavras

A etapa de identificação das palavras consiste em realizar uma análise léxica, ou seja, converter um conjunto de caracteres em um conjunto de palavras. Neste momento as palavras são identificadas e são ignorados dígitos, hifens, marcas de pontuação e a situação das letras se maiúsculas ou minúsculas, os símbolos e caracteres de controle de arquivo ou de formatação.

Para verificar se as palavras dentro de um texto existem, a comparação de uma sequência de caracteres retirados de um texto também podem ser comparadas com dicionários. A implementação dessa comparação é realizada quando se tem como objetivo verificar se alguns documentos contêm ou não caracteres inválidos ou palavras com erros gramaticais. Uma vez identificados às sequências de caracteres inválidos, as mesmas devem ser eliminadas, e as palavras com erros gramaticais corrigidas.

A adoção de um dicionário torna-se opção muito interessante na identificação de termos específicos, ou seja, quando se deseja utilizar palavras pré-definidas no índice, evitando que palavras desconhecidas, sejam identificadas, e a utilização de vocabulário sobre o qual não tenha domínio. Para implementar essa opção, pode-se utilizar um analisador léxico.

Outro fator a ser considerado nessa etapa, é a utilização de números, geralmente não são considerados bons termos para compor o índice, pois os mesmos, não possuem um contexto de proximidade, haja vista que sua natureza é vaga. Por exemplo, considerando que um usuário está interessado nos documentos que informem o número de mestrandos em ciência da informação na UFSC entre 1992 a 2009. Essa pergunta poderia ser especificada como um conjunto de termos de índice (ciência, informação, 1992, 2009). O problema é que a presença desses números pode resultar em uma recuperação de uma variedade de documentos que tenham qualquer uma destas duas datas. Mas em alguns casos, é importante considerar que dígitos podem aparecer dentro de uma palavra. Por exemplo, 2.000 A.C. é um termo de índice claramente importante. Nestes casos, não está claro quais regras deveriam ser utilizadas, além disso, uma sequência de 11 dígitos que identificam o número de CPF pode ser altamente relevante em um determinado contexto e, nesse caso, deve ser considerado como um termo de índice.

2.10.3 Remoção de *Stopwords*

Algumas palavras não podem ser adicionadas na estrutura de índices, que quando as mesmas estão presentes em um documento texto, meramente com o intuito de conectar as frases.

Palavras que aparecem frequentemente em documentos de uma coleção, devido a sua natureza frequente ou semântica são consideradas sem valor para a recuperação. Estas palavras são denominadas palavras negativas ou *stopwords*, e dificilmente são utilizadas em uma consulta, pois sua indexação somente tornaria o índice maior do que o necessário.

Artigos, conjunções e preposições, entre outras classes de palavras cuja finalidade é auxiliar a estruturação da linguagem, não necessitam ser incluídas na estrutura de índice, pois as mesmas não são capazes de selecionar documentos que possam ser relevantes, podendo inclusive comprometer a precisão e a eficiência do sistema.

As palavras que aparecem em praticamente todos os documentos de uma coleção, não são capazes de discriminar documentos e também não devem constar na estrutura de índice.

Existem listas de *stopwords* de domínio público as quais são denominadas de *stoplists*, também chamadas de dicionários negativos. Essas listas podem ser livremente utilizadas na elaboração de ferramentas que realizem o processo de remoção de *stopwords*.

A eliminação de *stopwords* reduz consideravelmente o tamanho da estrutura do índice podendo em uma lista invertida reduzir consideravelmente o tamanho original do documento, mas também poderá apresentar falhas em alguns casos. Um exemplo disto é quando um usuário esteja realizando uma consulta em uma coleção de documentos cujo objetivo é encontrar documentos que contenham a frase “ser ou não ser”, onde a eliminação de *stopwords* poderia deixar somente o termo “ser”, tornando assim quase impossível reconhecer a frase especificada pelo usuário em algum documento. Em decorrência de problemas dessa natureza, é que a utilização de índices completos em que são indexadas todas as palavras contidas em documento, se torna mais atraente.

2.10.4 Normalização Morfológica (*Stemming*)

Durante o processo de indexação automática de texto existe uma etapa de normalização morfológica mais conhecida como *stemming*. Essa etapa consiste em eliminar as variações morfológicas de uma

palavra, por meio da identificação do seu radical. Os prefixos e os sufixos são retirados, e os radicais resultantes são adicionados à estrutura de índice. Nesse processo ocorre uma redução do tamanho da estrutura do índice, pois o número de termos distintos do índice é reduzido.

Apesar do aparente benefício da utilização dessa técnica, atualmente a maioria dos motores de busca não utiliza nenhum tipo de algoritmo de *stemming*, visto que o mesmo pode acabar utilizando palavras muito abrangentes, assim não recupera documentos que contenham termos específicos.

2.10.5 Identificação de Termos Compostos

A etapa de identificação de termos compostos também conhecida como *word-phrase formation*, busca identificar expressões compostas de dois ou mais termos. Nessa etapa, são consideradas algumas palavras que possuem significados diferentes quando utilizadas em conjunto. Isso geralmente ocorre porque existem conceitos que só podem ser descritos pela utilização de duas ou mais palavras adjacentes. Algumas vezes uma palavra é agrupada com outra a fim de modificar ou refinar seu significado, como por exemplo, o ato administrativo, no ato, ato simbólico, ato institucional. Quando isso ocorre, essas duas ou mais palavras não podem ser separadas quando indexadas. Caso sejam separadas, perde-se o conceito ou sentido da mesma.

Existem basicamente duas formas de identificar expressões, onde a primeira é realizada com base na identificação de termos que co-ocorrem com frequência em uma coleção de documentos. Nesse caso torna-se interessante que o sistema apresente ao usuário as expressões identificadas e repasse ao usuário a decisão sobre quais são as corretas. A segunda consiste na utilização de um dicionário de expressões que indique as palavras que devem ser combinadas.

Esse tipo de técnica torna a busca mais precisa, já que os termos compostos aparecem com frequência em um número menor de documentos, e tornam a consulta menos abrangente. Porém, esses termos são geralmente armazenados no índice de forma composta e, nesse caso, o usuário não pode localizá-los de forma separada. Uma solução para esse problema consiste em armazenar ambas as formas: combinada e separada.

2.11 Cálculo de Relevância

A relevância é o cerne da recuperação de informação, pois o objetivo principal de um sistema de recuperação de informação é recuperar os documentos mais relevantes para o usuário a partir de uma consulta realizada pelo mesmo. Documentos relevantes são aqueles que estão inseridos no contexto da pesquisa realizada pelo usuário, e que têm alguma relação com a informação desejada.

Em um documento, algumas palavras são mais importantes do que outras, posto que as palavras utilizadas com mais frequência com exceção das *stopwords* costumam ter um significado mais importante. As palavras constantes em títulos ou em outras estruturas, também possuem uma importância maior, pelo fato de o autor do documento por algum motivo ter considerado os mesmos muito relevantes. Os substantivos e complementos também podem ser considerados mais relevantes que os demais termos de uma oração.

Sendo assim, o cálculo de relevância de uma palavra, pode basear-se na frequência das mesmas, na análise estrutural do documento ou na posição sintática de uma palavra.

As técnicas mais comuns são baseadas na frequência com que as palavras aparecem na coleção de documentos, pois as outras necessitam de métodos adicionais (análise de linguagem natural), por exemplo que exigem maior complexidade (conhecimento).

Segundo *Rijsbergen* (1999), existem várias fórmulas que foram desenvolvidas, ou aplicadas com o intuito de calcular a importância de uma palavra baseando-se em sua frequência. Essa importância costuma ser chamada de peso e indica o grau de relação entre a palavra e os documentos em que ela aparece.

Várias fórmulas de identificação de peso, como exemplo as baseadas em cálculos de frequência absoluta, frequência relativa, frequência inversa de documentos.

A frequência absoluta, também conhecida por frequência do termo ou *term frequency* (TF), nada mais é do que a medida da quantidade de vezes que um termo aparece em um documento. Essa é a medida de peso mais simples que existe, mas não é aconselhada porque não é capaz de fazer distinção entre os termos que aparecem em poucos documentos e os termos que aparecem em vários documentos. Em alguns casos esse tipo de análise poderia ser extremamente importante, pois os termos que aparecem nos documentos não são capazes de discriminar um documento de outro.

Além disso, a frequência absoluta não leva em conta a quantidade de palavras existente no documento. Com isso, uma palavra pouco frequente em um documento pequeno pode ter a mesma importância de uma palavra muito frequente de um documento grande.

A frequência relativa busca solucionar esse último problema levando em conta o tamanho do documento, ou seja, a quantidade de palavras que ele possui, e normalizando os pesos de acordo com essa informação. Sem essa normalização, os documentos, grandes e pequenos, acabam sendo representados por valores em escalas diferentes. Com isso os documentos maiores possuem melhores chances de serem recuperados, já que receberão valores maiores no cálculo de similaridades.

A frequência relativa (Frel) de uma palavra x em um documento qualquer é calculada dividindo-se sua frequência absoluta (Fabs) pelo número total de palavras no mesmo documento (N):

$$\text{FrelX} = \frac{\text{FabsX}}{\text{N}}$$

Para solucionar o outro problema da frequência absoluta, onde a quantidade de documentos em que um termo aparece, não é considerada, torna-se então necessário obter essa informação. A frequência de documentos é que indica a quantidade de documentos em que um termo aparece.

De posse da frequência absoluta e da frequência de documentos é possível calcular a frequência inversa de documentos *inverse document frequency (IDF)*, capaz de aumentar a importância de termos que aparecem em poucos documentos e diminuir a importância de termos que aparecem nos documentos, justamente pelo fato dos termos de baixa frequência de documentos serem, em geral, mais discriminantes.

2.12 Avaliação da Recuperação da Informação

O desempenho de um sistema de recuperação de informações é avaliado de acordo com a sua capacidade em recuperar o maior número de itens relevantes, ao mesmo tempo em que filtra ao máximo os itens irrelevantes. É em cima dessa estratégia que as métricas são desenvolvidas e aplicadas.

Segundo Baeza-Yates e Ribeiro-Neto (1999, p.73), o objetivo preliminar de um sistema de recuperação de informação é recuperar todos os documentos que são relevantes a uma solicitação do usuário com uma quantidade mínima de documentos não-relevantes, sendo que as métricas mais importantes para a avaliação do resultado de um

sistema de recuperação de informações são: revocação (*recall*) e precisão (*precision*).

a) **Revocação (*Recall*)** - Revocação ou abrangência (do inglês: *Recall*) mede a habilidade do sistema em recuperar os documentos mais relevantes para o usuário.

O revocação é calculada da seguinte forma:

$$\text{Revocação} = \frac{\text{número de documentos relevantes recuperados}}{\text{total de documentos relevantes na coleção}}$$

b) **Precisão (*precision*)** - Precisão (do inglês: *Precision*) mede a habilidade do sistema de manter os documentos irrelevantes fora do resultado de uma consulta.

A precisão é calculada da seguinte forma:

$$\text{Precisão} = \frac{\text{número de documentos relevantes recuperados}}{\text{total de documentos recuperados}}$$

Segundo Baeza-Yates e Ribeiro-Neto (1999, p.77), a precisão é capaz de indicar o trabalho que o usuário teria para analisar uma determinada busca. Isso significa que, se 60% dos itens retornados fossem relevantes, o usuário teria desperdiçado 40% de seu esforço analisando itens irrelevantes.

Esses dois parâmetros estão inversamente relacionados, significando que a melhoria de um, implica na piora do outro. O gráfico ilustra essa relação:

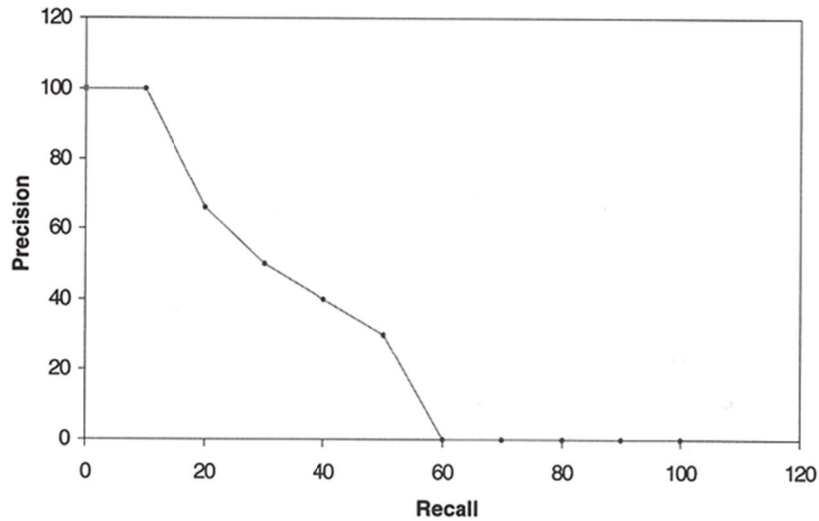


Gráfico 1: Relação Precisão x Revocação
Fonte: Baeza-Yates e Ribeiro-Neto (1999, p.77)

Com o objetivo de comparação de sistemas, principalmente acadêmicos, existem algumas coleções públicas de documentos preparadas especialmente para o processo de avaliação. Sabe-se que para sistemas diferentes possam ser avaliados e comparados, deve-se adotar uma coleção específica.

Segundo Baeza-Yates e Ribeiro-Neto (1999, p.85), a coleção mais conhecida é a *Text REtrieval Conference (TREC)* (<http://trec.nist.gov/>). Nesta conferência recebe-se um conjunto de técnicas experimentais para que seja avaliado posteriormente em sistemas de recuperação informações, por exemplo, oferecem uma série de consultas pré-definidas e conjuntos de documentos relevantes a cada uma delas.

O objetivo principal da recuperação automática de informação é encontrar todos os documentos relevantes para determinada consulta, ou maximizar o recall, evitando os erros, ou seja, deve-se reduzir o número de documentos não relevantes selecionados ao menor número possível, o que corresponde a aumentar a precisão.

2.13 Indexação Semântica Latente

A Indexação Semântica Latente – *LSI* (do inglês - *Latent Semantic Indexing*) foi desenvolvido pelo *Bellcore* (agora *Telcordia*) no final dos anos 1980 (1988), sendo patenteado em 1989 (<http://lsi.argreenhouse.com/lsi/LSI.html>).

Os primeiros documentos sobre *LSI* foram publicados por:

Dumais, ST, Furnas, GW, Landauer, TK e Deerwester, S. (1988), "Using latent semantic analysis to improve information retrieval." In *Proceedings of CHI'88: Conference on Human Factors in Computing, New York: ACM*, 281-285.

Deerwester, S., Dumais, ST, Landauer, TK, Furnas, GW e Harshman, RA (1990) "Indexação pela análise semântica latente." *Journal of the Society for Information Science*, 41(6), 391-407.

Foltz, PW (1990) "Usando Indexação Semântica Latente para Filtragem de Informação". In *RB Allen (Ed.) Proceedings of the Conference on Office Information Systems, Cambridge, MA*, 40-47.

De acordo com a teoria desenvolvida por *Deerwester, Dumais et al.* (1990), espera-se encontrar uma estrutura semântica latente em uma coleção de documentos a partir da utilização da *LSI* (do inglês *Latent Semantic Indexing*), e com isto expandir a consulta e recuperar os documentos mais relevantes para o usuário. Este modelo de recuperação de informação apesar de ter a palavra semântica em seu nome, não é semântico, mas sim baseado em métodos estatísticos.

A Indexação Semântica Latente foi idealizada com o objetivo de reduzir os problemas de sinonímia⁴, polissemia⁵ e de palavras associadas encontradas nos modelos de recuperação automática de informação. A polissemia interfere na precisão das pesquisas, pelo fato de que a mesma pode retornar documentos que não têm relação com a pesquisa desejada. Já no caso da sinonímia, são retornados poucos documentos, mesmo que existam vários documentos que poderiam ser recuperados por causa de sinônimos. Mas a situação mais crítica é das palavras associadas, pois embora existam documentos visivelmente relevantes, muitos documentos não são recuperados.

O interesse por Indexação Semântica Latente tem crescido consideravelmente desde a sua concepção, ao ponto do maior site de busca do mundo o Google realizar investimentos nesta área, pois como foi divulgado no seu site, no 6 dia 23 de abril de 2003, o Google

⁴ Sinonímia - Várias palavras que significam a mesma coisa

⁵ Polissemia - Palavras que tem mais de um significado

⁶ <http://www.google.com/press/pressrel/applied.html>

comprou a *Applied Semantics*⁷ por considerar que é uma empresa inovadora e possuir uma experiência comprovada em indexação semântica.

No campo da recuperação da informação, um dos maiores problemas, é o casamento léxico entre as palavras digitadas pelo usuário na consulta com os documentos existentes em uma coleção.

Esse método oferece uma análise semântica dos termos em todos os documentos que foram indexados dentro uma coleção documentos, ou seja, procura superar problemas de comparações lexicais de termos ao considerar uma estrutura semântica latente implícita pela variabilidade das palavras. A estrutura semântica é definida como a estrutura que representa a correlação de termos individuais nos documentos. Nesse caso, a semântica refere-se ao fato de documentos poderem ser referenciados pelos seus próprios termos.

A indexação semântica latente tem seu funcionamento em torno da observação de que uma matriz de termos de índices por documentos é esparsa, ou seja, a maioria dos termos não aparece na maioria dos documentos, sendo assim a matriz poderia ser composta de vários valores nulos. Posto isto, essa matriz pode ser então reduzida a uma matriz menor e mais densa, através da aplicação de várias técnicas matemáticas. O quanto se deseja reduzir a matriz, é uma questão de quanta informação se está disposto a sacrificar para ganhar revocação originada pela combinação.

Se fizermos uma análise da expressão “Indexação Semântica Latente” vamos verificar que a palavra "latente" significa algo que está presente, mas obviamente não visível. Já a palavra "semântica" refere-se ao significado da língua, ou seja, em oposição ao que é realmente dito ou escrito, e a palavra "indexação" é a identificação do significado de um documento a partir do seu objeto.

A maioria dos métodos considera a ocorrência dos termos, informados na consulta, e nos documentos para realizar os cálculos de similaridade que indicarão o grau de relevância de um documento diante dessa consulta. Muitas vezes alguns termos importantes para o sentido da busca que está sendo realizada não são informados por mero desconhecimento ou mesmo esquecimento do usuário no momento de construir a consulta. Assim, na abordagem em que se considera apenas a ocorrência dos termos para se definir o grau de relevância, muitos

⁷ Applied Semantics é uma empresa produtora de softwares aplicativos situada em Santa Monica Califórnia USA.

documentos relevantes ficarão de fora (DUMAIS et. al., 1988; BLAIR; MARON, 1985).

Na tentativa de resolver essa deficiência, o método de indexação semântica latente, utiliza uma abordagem que leva em consideração à coocorrência de termos, isto é, conjuntos de termos que frequentemente são encontrados nos mesmos documentos. Pois, considerando que se estes tais termos surgem com frequência nos mesmos documentos relativos à determinada área, isto pode evidenciar que existe neste caso uma relação semântica latente, ou seja, não é explícita. Com base em técnicas estatísticas, o modelo de indexação semântica latente pode “descobrir” as possíveis correlações existentes.

Pode-se constatar que num vocabulário utilizado pelo ser humano, a utilização das palavras é caracterizada por um extenso uso de sinônimos. Portanto, uma comparação direta por termos pode ser deficiente.

De acordo *Dumais et. al.* (1988), as pessoas normalmente desejam acessar a informação baseada no seu significado, e a comparação direta de palavras não consegue realizar esse trabalho com sucesso.

Segundo *Deerwester, Dumais et. al.* (1990), a maior contribuição da técnica de indexação semântica latente é que consultas e documentos não precisam possuir termos em comum para serem considerados semelhantes. Caso o vetor de consulta e o vetor de um documento estejam próximos no espaço geométrico semântico, o documento é considerado similar à consulta.

Michael W. Berry et al. (1994) descrevem a indexação semântica latente aplicada à recuperação de informação apontando vantagens, relacionadas à sinonímia e à polissemia, entretanto, a técnica de *LSI* apresenta algumas desvantagens como:

Alto custo computacional despendido em cálculos pela utilização do modelo algébrico *SVD* manipulando matrizes esparsas.

Dificuldade para determinar a dimensão ideal do espaço conceitual reduzido - Visando determinar a dimensão ideal os pesquisadores usaram “tentativa e erro”. Em um experimento *Dumais* utilizou *k* variando de 200 a 350 nos experimentos da *TREC-3*, quando experimentava *LSA* com técnicas de expansão de consultas, mas não existe, consenso sobre o número ideal de dimensões do espaço reduzido e, devido ao custo computacional do método, o *SVD* é impraticável para utilizar “tentativa e erro”.

Problemas de escalabilidade - Quando documentos são adicionados à coleção, os novos termos pertencentes a esses documentos não são diretamente considerados na comparação dos vetores. O número

de termos não considerados cresce proporcionalmente ao número de documentos adicionados. Para que esses termos sejam considerados, uma nova execução de *SVD* torna-se necessária e o custo computacional de executá-la deve ser considerado. Contudo, se a coleção é estável, o *SVD* é executado uma única vez e o custo computacional é aceitável.

Segundo *Deerwester, Dumais et al. (1990)*, a indexação semântica latente tem como um dos objetivos de melhorar a recuperação de informação através do descobrimento de associações entre os termos em uma grande coleção de textos a fim de criar um espaço semântico. Exemplificando, através da análise de uma coleção de textos utilizando indexação semântica latente, o sistema aprenderá que quando realizamos uma consulta informando venda de carros, tem-se como retorno da consulta documentos que contenham as frases venda de carros, venda de veículos e venda de automóvel, já que carro, veículo e automóvel são sinônimos. Da mesma forma, em uma consulta por banco de dados, o resultado da consulta será somente documentos que contenham uma relação de banco de dados deixando de fora documentos que se referem a um banco como entidade financeira e banco como objeto de descanso.

Este modelo pode ser definido como uma técnica automática que analisa as coocorrências de termos em documentos textuais com vistas a descobrir relacionamentos latentes entre eles. Para identificar as relações semânticas, a indexação semântica latente utiliza o modelo de Decomposição de Valores Singulares – *SVD* (do inglês *Singular Value Decomposition*).

De acordo com *Deerweter (1990)*, a estrutura de análise da indexação semântica latente - *LSI* refere-se a uma matriz esparsa termo-documento. Posto isto, já que trabalha com vários vetores coluna, criando dessa forma uma matriz, onde nas linhas estão representados os termos indexados de cada documento e nas colunas o documento, dessa forma é criada a relação à matriz termo-documento. Explicando melhor essa relação, seja “*ti*” a linha e “*dj*” a coluna da matriz, e seja o elemento da matriz “*O_{ij}*” que representaria o número de vezes que o termo “*i*” aparece no documento “*j*”.

Uma vez indexados os termos de cada documento, e também criada a relação termo-documento, é então aplicado o *SVD*, tendo como resultado dessa decomposição três matrizes "otimizadas". Estas Matrizes recebem a denominação de “otimizadas”. Tal denominação decorre do fato que nas mesmas vão ser eliminados dados que não contribuem na matriz termo-documento. Uma vez escolhido o nível das matrizes *U'*, *S'* e *V'* as matrizes estão prontas para receber as consultas fornecidas ao

sistema. Essa matriz é analisada por um *SVD* que a decompõe em três outras matrizes:

- 1) A primeira matriz possui colunas ortogonais e representa os termos - a matriz *U* que contém os termos;
- 2) A segunda, que também possui colunas ortogonais, representa os documentos - a matriz *S* que contém os valores mais representativos da matriz termo-documento (os valores singulares da matriz);
- 3) A terceira representa a matriz diagonal de valores singulares - matriz *V* que contém os documentos.

Uma vez criadas estas três matrizes é escolhido um tamanho (nível *k*) para trabalhar com as três matrizes. Escolhido este valor, são criadas três matrizes (que serão chamadas *U'*, *S'* e *V'*) de nível *k*, a estas três novas matrizes é multiplicado o vetor *Q*, que representa uma consulta. O resultado dessa multiplicação será um vetor cujo conteúdo é uma lista dos documentos mais relevantes para a consulta fornecida. O resultado da consulta feita ao sistema será uma lista ordenada por relevância dos documentos que são mais relevantes para a consulta fornecida. Através do produto dessas três matrizes que os relacionamentos latentes são estabelecidos.

No âmbito da recuperação da informação, o *SVD* pode ser visto como uma técnica criada para derivar um conjunto de variáveis indexadas não correlacionadas (*Deerwester, Dumais et al., 1990*), em que cada termo e documento são representados por um vetor de pesos, onde o peso deve indicar a força da associação entre um termo e um documento.

A definição de peso pode ocorrer por diferentes métodos, como, por exemplo, “0” e “1”, ou seja, indicando se um termo ocorre ou não no documento, ou um valor que indica a quantidade de ocorrências de um termo em um documento. Uma consulta é representada pela soma dos vetores dos termos que compõem a consulta. O conjunto de potenciais documentos é encontrado ao se calcular, por exemplo, o cosseno ou a distância do pseudo-documento (i.e., os termos que formam a consulta) em relação ao conjunto total de documentos. (*DEERWESTER, DUMAIS et al., 1990; BAEZA-YATES e RIBEIRO-NETO, 1999*).

Um exemplo apresentado por Garcia (2006) sobre como a indexação semântica latente funciona:

d1: Shipment of gold damaged in a fire.

d2: Delivery of silver arrived in a silver truck.

d3: Shipment of gold arrived in a truck.

Garcia usa o modelo de contador de termo para pontuar o peso do termo e o peso da consulta, e que o peso local é definido como ocorrências de palavras. Na indexação dos documentos foram utilizadas as seguintes regras:

- stopwords* não são ignoradas;
- O texto é tokenizado e colocado em caixa baixa;
- Não é usado *stemming*; e,
- Os termos são classificados em ordem alfabética.

Neste exemplo, o objetivo é encontrar os documentos que possuam as palavras *gold silver truck*.

Passo 1: Contam-se as ocorrências dos termos e constrói-se a matriz A termo-documento e a consulta na seguinte matriz:

| Terms | d1 | d2 | d3 | q |
|----------|----|----|----|---|
| ↓ | ↓ | ↓ | ↓ | ↓ |
| a | 1 | 1 | 1 | 0 |
| arrived | 0 | 1 | 1 | 0 |
| damaged | 1 | 0 | 0 | 0 |
| delivery | 0 | 1 | 0 | 0 |
| fire | 1 | 0 | 0 | 0 |
| gold | 1 | 0 | 1 | 1 |
| in | 1 | 1 | 1 | 0 |
| of | 1 | 1 | 1 | 0 |
| shipment | 1 | 0 | 1 | 0 |
| silver | 0 | 2 | 0 | 1 |
| truck | 0 | 1 | 1 | 1 |

Passo 2: Decompõe-se a matriz A

$$A = USVT$$

Passo 3: Implementa-se o segundo *rank* aproximando primeiro as colunas U e V e depois as colunas e linhas de S.

$$\begin{aligned}
 \mathbf{U} \approx \mathbf{U}_k &= \begin{bmatrix} -0.4201 & 0.0748 \\ -0.2995 & -0.2001 \\ -0.1206 & 0.2749 \\ -0.1576 & -0.3046 \\ -0.1206 & 0.2749 \\ -0.2626 & 0.3794 \\ -0.4201 & 0.0748 \\ -0.4201 & 0.0748 \\ -0.2626 & 0.3794 \\ -0.3151 & -0.6093 \\ -0.2995 & -0.2001 \end{bmatrix} & \mathbf{S} \approx \mathbf{S}_k &= \begin{bmatrix} 4.0989 & 0.0000 \\ 0.0000 & 2.3616 \end{bmatrix} & \mathbf{k} = 2 \\
 \\
 \mathbf{V} \approx \mathbf{V}_k &= \begin{bmatrix} -0.4945 & 0.6492 \\ -0.6458 & -0.7194 \\ -0.5817 & 0.2469 \end{bmatrix} & \mathbf{V}^T \approx \mathbf{V}_k^T &= \begin{bmatrix} -0.4945 & -0.6458 & -0.5817 \\ 0.6492 & -0.7194 & 0.2469 \end{bmatrix}
 \end{aligned}$$

Passo 4: Procura-se o novo vetor de coordenadas do documento dentro dessa redução de duas dimensões espacial. As filas do vetor V detêm valores. Estas são as coordenadas de cada vetor de documento:

d1(-0.4945, 0.6492)

d2(-0.6458, -0.7194)

d3(-0.5817, 0.2469)

Passo 5: Encontrar o novo vetor de coordenadas da consulta dentro da redução de duas dimensões espaciais.

$$\mathbf{q} = \mathbf{q}^T \mathbf{U}_k \mathbf{S}_k^{-1}$$

Essas são as novas coordenadas da consulta vetor em duas dimensões. Note como essa matriz é agora diferente da consulta original matriz q determinada no passo 1.

$$q = q^T U_k S_k^{-1}$$

$$q = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -0.4201 & 0.0748 \\ -0.2995 & -0.2001 \\ -0.1206 & 0.2749 \\ -0.1576 & -0.3046 \\ -0.1206 & 0.2749 \\ -0.2626 & 0.3794 \\ -0.4201 & 0.0748 \\ -0.4201 & 0.0748 \\ -0.2626 & 0.3794 \\ -0.3151 & -0.6093 \\ -0.2995 & -0.2001 \end{bmatrix} \begin{bmatrix} 1 \\ 4.0989 & 0.0000 \\ 0.0000 & 2.3616 \end{bmatrix} \quad k = 2$$

$$q = \begin{bmatrix} -0.2140 & -0.1821 \end{bmatrix}$$

Passo 6: O *Rank* dos documentos é posto em ordem decrescente da consulta do cosseno de similaridade dos documentos.

$$\text{sim}(q, d) = \frac{q \bullet d}{|q| |d|}$$

$$\text{sim}(q, d_1) = \frac{(-0.2140)(-0.4945) + (-0.1821)(0.6492)}{\sqrt{(-0.2140)^2 + (-0.1821)^2} \sqrt{(-0.4945)^2 + (0.6492)^2}} = -0.0541$$

$$\text{sim}(q, d_2) = \frac{(-0.2140)(-0.6458) + (-0.1821)(-0.7194)}{\sqrt{(-0.2140)^2 + (-0.1821)^2} \sqrt{(-0.6458)^2 + (-0.7194)^2}} = 0.9910$$

$$\text{sim}(q, d_3) = \frac{(-0.2140)(-0.5817) + (-0.1821)(0.2469)}{\sqrt{(-0.2140)^2 + (-0.1821)^2} \sqrt{(-0.5817)^2 + (0.2469)^2}} = 0.4478$$

Ranking documents in descending order

$$d_2 > d_3 > d_1$$

Constata-se que a pontuação do documento d2 é maior de d3 e d1, e o seu vetor aproxima-se mais da consulta do que os outros vetores d3 e d1.

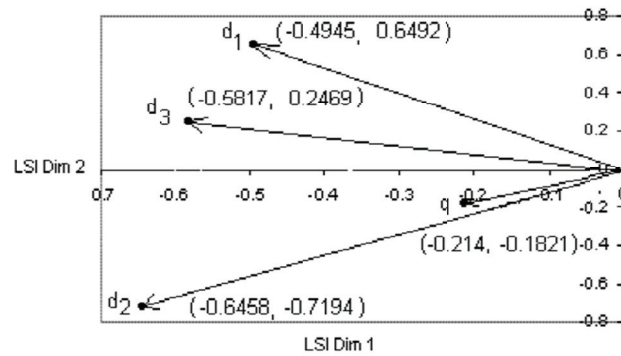


Gráfico 2: Dimensão LSI

Observa-se também que a teoria do Termo Vetor ainda é utilizada no início e no final do LSI.

A seguir apresentam-se os procedimentos metodológicos utilizados para o desenvolvimento dessa pesquisa.

3 Procedimentos Metodológicos

Na composição desta dissertação estão inseridas as definições dos procedimentos metodológicos utilizados no decorrer da pesquisa, como o tipo de pesquisa, técnica para coleta dos dados, exploração do material e por fim quanto ao tratamento e apresentação dos resultados da pesquisa, bem como, mostrar como o objeto da pesquisa se inscreve no campo dos conhecimentos sobre o tema, e como estes conhecimentos permitiram atingir os objetivos da pesquisa.

3.1 Tipo da Pesquisa

Para Marconi e Lakatos (2008), a pesquisa é um procedimento reflexivo sistemático, controlado e crítico, que permite descobrir novos fatos ou dados, relações ou leis, em qualquer campo do conhecimento. A pesquisa é um tratamento formal, com método de pensamento reflexivo, que requer um tratamento científico e se constitui no caminho para conhecer a realidade ou para descobrir verdades parciais.

Segundo Gil (2002), a pesquisa é definida por como “[...] um processo que tem por finalidade descobrir as respostas para os problemas mediante a utilização de procedimentos científicos”.

A presente pesquisa tem caráter exploratório. Segundo Gil (2002), as pesquisas exploratórias têm como objetivo proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito ou a constituir hipóteses.

Na pesquisa realizada na internet através de sites de busca como o *Google* <<http://www.google.com.br>>, *Bing* *Microsoft* <<http://www.bing.com>>, e *Author Mapper* <<http://www.authormapper.com>>, bem como o portal da CAPES, não foi encontrada nenhuma literatura sobre preservação e recuperação de informações digitais em bibliotecas digitais Greenstone, evidenciando que é uma área pouco explorada ou até inexplorada.

De acordo com Marconi e Lakatos (2008), quando uma área é pouco explorada, trata-se de uma pesquisa exploratória.

Quanto à natureza da pesquisa, trata-se de uma pesquisa aplicada, pois sua preocupação está menos voltada para o aperfeiçoamento de teorias gerais, mas sim em gerar conhecimentos para a aplicação prática sobre preservação e recuperação da informação digitais na biblioteca digital Greenstone.

Nesta pesquisa realizou-se um levantamento bibliográfico do tema proposto para respaldar a fundamentação teórica e a análise dos dados. A pesquisa bibliográfica e exploratória se valeu do levantamento dos documentos selecionados como fontes de pesquisa, proporcionando a fundamentação teórica.

A presente pesquisa utiliza-se de procedimentos qualitativos e quantitativos para a obtenção, análise e interpretação dos dados.

A análise, apesar de ocorrer desde o início do processo, se torna mais sistemática e formal após o encerramento da coleta de dados, quando se transforma em um processo indutivo, iterativo e recorrente, porque o avaliador, muitas vezes, volta às fontes para confirmar e ampliar os dados e para validar os resultados e conclusões.

De acordo com Minayo (1993, p.22), os procedimentos qualitativos se referem ao caráter subjetivo de alguns temas, ou seja, trabalha com o universo dos significados, aspirações, crenças, valores e atitudes; enquanto que os quantitativos estão relacionados ao aspecto objetivo obtido através de dados matemáticos e análises estatísticas.

Nas pesquisas qualitativas os procedimentos de coleta, interpretação e análise dos dados são mais flexíveis e podem ser construídos ao longo do processo.

O método da pesquisa foi descritivo, sendo os resultados expressos por meio de quadros, tabelas, figuras e gráficos e respectiva análise.

3.2 Estudo de caso

Segundo Gil (2002), o planejamento da pesquisa exploratória é bastante flexível, e na maioria dos casos assume a forma de pesquisa bibliográfica ou de estudo de caso.

Gil (2002) ressalta que nas ciências, durante muito tempo, o estudo de caso foi encarado como procedimento pouco rigoroso, que serviria apenas para estudos de natureza exploratória.

De acordo com Yin (2005), é encarado como o delineamento mais adequado para a investigação de um fenômeno contemporâneo dentro de seu contexto real, onde os limites entre o fenômeno e o contexto não são claramente percebidos. O estudo de caso é uma inquirição empírica que investiga um fenômeno contemporâneo dentro de um contexto da vida real, quando a fronteira entre o fenômeno e o contexto não é claramente evidente e onde múltiplas fontes de evidência são utilizadas. Essa definição, apresentada como uma definição mais técnica ajuda a

compreender e distinguir o método do estudo de caso de outras estratégias de pesquisa.

O quadro 2 exemplifica as diferentes situações para escolha correta da estratégia de pesquisa.

| Estratégia | Forma de questão de pesquisa | Exige controle sobre eventos comportamentais | Focaliza acontecimentos contemporâneos |
|---------------------|-------------------------------------|---|---|
| Experimento | Como, por que | Sim | Sim |
| Levantamento | Quem, o que, onde, quantos, quanto | Não | Sim |
| Análise de Arquivos | Quem, o que, onde, quantos, quanto | Não | Sim/não |
| Pesquisa histórica | Como, por que | Não | Não |
| Estudo de caso | Como, por que | Não | Sim |

Quadro 2: Situações relevantes para diferentes estratégias de pesquisa
Fonte: (YIN, 2005, p.24)

Segundo Yin (2005), em geral, os estudos de caso representam a estratégia preferida quando se colocam questões do tipo “como” e “porque”, quando o pesquisador tem pouco controle sobre os acontecimentos e quando o foco se encontra em fenômenos contemporâneos inseridos em algum contexto da vida real. Pode-se então complementar esses estudos de casos explanatórios com dois outros tipos – estudos exploratórios e descritivos.

Nesta pesquisa, a escolha da técnica de estudo de caso teve como base as seguintes justificativas:

- a) Trata-se de uma investigação empírica sobre a realidade contemporânea de uma determinada organização;

- b) A notória disponibilidade e facilidade de acesso do pesquisador aos dados e fatos pertinentes à realidade da organização investigada;

Martins (2006, p. 2) ressalta que:

[...] quando um Estudo de Caso escolhido é original e revelador, isto é, apresenta um engenhoso recorte de uma situação complexa da vida real, cuja análise-síntese dos achados tem a possibilidade de surpreender, revelando perspectivas que não tinham sido abordadas por estudos assemelhados, o caso poderá ser qualificado como importante, e visto em si mesmo como uma descoberta.

Martins (2006), ainda destaca que o sucesso de um estudo de caso, em muito, depende da perseverança, criatividade e raciocínio crítico do investigador para construir descrições, interpretações, enfim, explicações originais que possibilitem a extração cuidadosa de conclusões e recomendações.

Nesta pesquisa o estudo de caso é sobre a Biblioteca Digital Greenstone.

3.2 Coleta de Dados

Segundo Gil (2002), o elemento mais importante para a identificação de um delineamento é o procedimento adotado para a coleta de dados. Assim, podem ser definidos dois grandes grupos de delineamentos como aqueles que se valem das chamadas fontes de “papel”; e aqueles cujos dados são fornecidos por pessoas. No primeiro grupo estão a pesquisa bibliográfica e a pesquisa documental. No segundo estão a pesquisa experimental, a pesquisa *ex-post facto*, o levantamento e o estudo de caso.

Na composição do corpus desse trabalho, estão inseridos os procedimentos para atingir os objetivos propostos que constituem a natureza bibliográfica. Conforme Gil (2002), a natureza bibliográfica é a elaboração da pesquisa a partir de material já publicado, constituído principalmente de livros, artigos de periódicos e atualmente com material disponibilizado na Internet.

Nessa pesquisa realizou-se um estudo sobre preservação lógica e recuperação de informação digital na biblioteca digital Greenstone, pela exploração de material bibliográfico, e com isto proporcionar

embasamento teórico para obter-se maior familiaridade com o problema a fim de alcançar os objetivos da pesquisa, para tanto, na exploração do material bibliográfico foram considerados nos idiomas português, inglês e espanhol como fontes de dados os documentos em papel e em meio eletrônico.

A documentação indireta documental trata especificamente da coleta de informações de fontes primárias, tais como documentos de arquivos públicos e privados, cartas, contratos, diários e autobiografias.

De acordo com (LAKATOS; MARCONI, 2008), a coleta de dados baseada na documentação indireta consiste na leitura e análise de materiais produzidos por terceiros, os quais podem apresentar-se sob a forma de textos, jornais, gravuras, fotografias e filmes, entre outras. Essa documentação indireta bibliográfica trata especificamente de coletar informações de fontes secundárias, tais como relatórios de pesquisa baseada em trabalho de campo, estudos históricos recorrendo aos documentos originais e pesquisas utilizando correspondências de terceiros, entre outras. Essa técnica é bastante utilizada em pesquisas nas quais o foco principal é o estudo de caso e em pesquisas puramente teóricas.

De acordo com Yin (2005), a evidência para estudos de caso podem vir de seis fontes: documentos, registros arquivais, entrevistas, observação direta, observação participante e artefatos físicos. O ponto chave na coleta de dados em um estudo de caso, é que a mesma não se trata de meramente de registrar mecanicamente, como se faz em outros tipos de pesquisa, pois, pode-se interpretar as informações na medida em que estão sendo coletadas e saber imediatamente, por exemplo, se as diversas fontes de informações se contradizem e levam a necessidade de evidências adicionais.

Com relação à utilização de documentos como fonte de coleta de dados, foi também utilizada uma das coleções de demonstração da Biblioteca digital Greenstone.

De acordo com Cervo e Bervian (1983, p. 155), a coleta de dados é conceituada de forma pragmática como sendo “a tarefa importante da pesquisa, envolve diversos passos, como a determinação da população a ser estudada, a elaboração do instrumento de coleta, a programação da coleta e também os dados da própria coleta”.

3.3 Unidade de Análise

Segundo Yin (2005), a definição da unidade de análise está relacionada à maneira como são definidas as questões iniciais da pesquisa.

Yin (2005, p. 45) ressalta o que é uma unidade de análise:

O livro *The Soul of a New Machine* (1981), escrito por *Tracy Kidder*, foi vencedor do prêmio *Pulitzer*⁸. O livro, também um *Best-seller*, trata do desenvolvimento de um novo computador produzido pela *Data General Corporation*, que foi projetado para competir diretamente com outro computador desenvolvido pela *Digital Equipament Corporation*.

De fácil leitura, o livro descreve como a equipe de engenheiros da *Data General* inventou e desenvolveu o novo computador. Começa com a conceitualização inicial do computador, e termina quando a equipe entrega o controle da máquina à equipe de marketing da *Data General*.

É um exemplo excelente de estudo de caso. No entanto, o texto de *Kidder* também ilustra um problema fundamental quando se realizam estudos de caso – o de definir a unidade de análise. O estudo de caso é sobre o computador ou é sobre a dinâmica de um pequeno grupo – a equipe de engenheiros? A resposta é muito importante se pretendemos entender como o estudo de caso se relaciona com corpo mais amplo de conhecimento – ou seja, se devemos generalizar a questão à tecnologia ou à dinâmica de grupo.

A questão principal a ser esclarecida nesta pesquisa, é verificar se os recursos disponíveis na biblioteca digital *Greenstone* são suficientes para realizar a preservação lógica de documentos digitais e sua recuperação, e a se *BDG* atende a comunidade que utiliza o *Greenstone*.

A unidade de análise dessa pesquisa é uma coleção de dissertações do programa de pós-graduação em Ciência da Informação da Universidade Federal do Estado de Santa Catarina que foram criadas

⁸ N. de T. Láurea instituída em 1917 pelo jornalista norte-americano Joseph Pulitzer e outorgada anualmente pela Universidade de Colúmbia. Divide-se em oito prêmios de jornalismo, cinco de literatura, quatro bolsas de estudo e um prêmio de música.

e importadas para a BDG. Também foram criados e importados para essa coleção, vários arquivos em diferentes formatos. Faz parte da unidade de análise a comunidade lusófona do Greenstone.

3.5 Universo da Pesquisa

A pesquisa documental, segundo Gil (2008), é semelhante à bibliográfica, sendo que a única diferença está na natureza das fontes, pois na bibliográfica, se utiliza fundamentalmente as contribuições dos diversos autores sobre determinado assunto; a pesquisa documental vale-se de materiais que não receberam um tratamento analítico, ou que podem ser reelaborados conforme os objetivos da pesquisa, como o que ocorre com essa pesquisa, pois os dados coletados foram analisados conforme os objetivos. Também foram efetuadas pesquisas em manuais de utilização da Biblioteca Digital Greenstone.

Nesta pesquisa, o universo da análise é a Biblioteca Digital Greenstone.

Este estudo embasou-se na análise qualitativa para trabalhar com a realidade do processo de preservação digital e recuperação da informação na Biblioteca Digital Greenstone.

Segundo Chizzotti (1991), em uma pesquisa qualitativa todas as pessoas que participam são reconhecidas como sujeitos que elaboram conhecimentos e produzem práticas adequadas para intervir nos problemas que identificam, além de analisar e discriminar as necessidades prioritárias, e propor ações mais eficazes.

Para a análise e interpretação dos dados utilizou-se a análise qualitativa, que permite identificar e investigar os motivos que fizeram os usuários que participam da lista de discussão do Greenstone no Brasil, a estudarem ou implantarem a Biblioteca Digital Greenstone, além de identificar o grau de satisfação, os problemas, as dificuldades e as vantagens no uso da BDG e com foco na preservação digital e recuperação da informação.

Os dados também foram analisados de forma quantitativa, pois de acordo com Chizzotti (1991), algumas pesquisas qualitativas não descartam a coleta de dados quantitativos, especialmente na etapa exploratória de campo ou nas etapas em que estes dados podem mostrar uma relação mais extensa entre fenômenos particulares.

3.6 Limitações da Pesquisa

Devido à amplitude do tema referente à preservação digital, esta pesquisa está focada nos formatos de arquivos para preservação digital.

3.7 Etapas da Pesquisa

Nesta pesquisa, está previsto as etapas do estudo exploratório da BDG e da pesquisa propriamente dita:

Estudo Exploratório:

- 1) Etapa de download da versão mais atualizada da BDG;
- 2) Etapa de levantamento de plug-ins de formatos arquivos disponíveis da BDG;
- 3) Etapa de instalação da BDG;
- 4) Etapa de customização da BDG;
- 5) Etapa de criação de um protótipo de uma Biblioteca de Teses do PGCIN-UFSC
- 6) Etapa de observação direta da BDG;
- 7) Etapa de realização de testes de recuperação de informação na BDG;
- 8) Etapa de análise do Greenstone sobre o ponto de vista da preservação lógica com foco nos formato de arquivos.
- 9) Etapa de descrição da BDG com respaldo técnico científico;
- 10) Etapa de Redação.

Etapas da Pesquisa:

- 1) Revisão de literatura
- 2) Detalhamento da pesquisa
- 3) Análise do problema
- 4) Qualificação
- 5) Adequação das sugestões da qualificação
- 6) Coleta de dados
- 7) Organização dos dados
- 8) Análise e interpretação dos dados
- 9) Redação preliminar do texto
- 10) Redação final
- 11) Entrega da dissertação
- 12) Defesa da dissertação

3.8 Procedimentos para Coleta de Dados

Nesta pesquisa, para coleta de dados foi realizado a instalação, customização do Greenstone e criação de uma coleção de teses do PGCIN UFSC, e um levantamento dos plug-ins para coleta de dados disponíveis para o Greenstone, e escolha de no mínimo três plug-ins para formatos de arquivos, ou seja, um para proprietários com especificação fechada, um para proprietário com especificação aberta, e um para não proprietários com especificação aberta.

4 GREENSTONE

A Biblioteca Digital Greenstone (BDG) é um software para a criação e distribuição de coleções de bibliotecas digitais. O Greenstone é projeto de bibliotecas digitais (*New Zealand Digital Library Project* – www.nzdl.org) da Universidade de *Waikato* na Nova Zelândia, e desenvolvido e distribuído em cooperação com a UNESCO (www.unesco.org) e a *ONG Human Info* (<http://humaninfo.org/>). O objetivo do software Greenstone é disponibilizar aos usuários, especialmente nas universidades, bibliotecas e outras instituições públicas, para construir suas próprias bibliotecas digitais, principalmente nos países em desenvolvimento.

Exemplos de Bibliotecas digitais Greenstone disponíveis e acessíveis na internet:

- 1) *The New Zealand Digital Library Project* -
<<http://www.sadl.uleth.ca/nz/cgi-bin/library>>
- 2) *China: Peking University digital library* -
<<http://162.105.138.23/tapian/tp.htm>>
- 3) *Germany: Digitale Bibliothek Information und Medien* -
<<http://digbib.iuk.hdm-stuttgart.de/gsdll/cgi-bin/library>>
- 4) *Russia: Mari El Republic government information* -
<<http://gov.mari.ru/gsdll/cgi/library>>
- 5) *United States: Aladin digital library* -
<<http://www.aladin.wrlc.org/gsdll/>>
- 6) *United States: Center for the Study of Digital Libraries* -
<<http://botany.cs.tamu.edu/gsdll/cgi-bin/library>>
- 7) *Afghanistan Centre at Kabul University - ACKU* -
<<http://puka.cs.waikato.ac.nz/cgi-bin/library?a=p&p=about&c=acku>>
- 8) *Afghanistan Research and Evaluation Unit - AREU* -
<<http://puka.cs.waikato.ac.nz/cgi-bin/library?a=p&p=about&c=areu>>
- 9) *France - Agatange Collection* -
<<http://www.agatange.fr/>>
- 10) *Vietna - Agricultural Techniques for Farmers (in Vietnamese)* - <<http://icadl2007.vista.gov.vn/gsdll/cgi-bin/library.exe?site=localhost&a=p&p=about&c=cnnt&ct=1&qto=2&l=vi&w=utf-8&TARGET=>>>
- 11) *Paquistão - AHKRC Digital Library, Islamabad, Pakistan* - <<http://210.56.25.21/gsdll/cgi-bin/library.exe?a=p&p=home&l=en&w=utf-88>>

- 12) *Estados Unidos da América - Allen Park Veterans Administration Hospital Archives* -
<<http://www.dalnet.lib.mi.us/gsd/cgi-bin/library?p=about&c=va>>
- 13) *India - Archives of Indian Labour* -
<<http://www.indialabourarchives.org/>>
- 14) *Armenia - Armenian Rare Books* -
,<http://greenstone.flib.sci.am/gsd/cgi-bin/library.cgi?e=p-00000-00---off-0--00----0-10-0---0--0direct-10---4-----0-1l--10-en-50---20-home---0--1-00-0-0-01-1-0utfZz-8-00&a=p&p=about&c=Armenian>>
- 15) *Biblioteca digital de la Fundación para la Innovación Agraria - FIA* -
<<http://bibliotecadigital.innovacionagraria.cl/>>
- 16) *Biblioteca Digital Gerencia Social* <
<http://190.78.48.48/gsd/cgi-bin/library>>
- 17) *Biblioteca Digital of the Centro de Información de Recursos Naturales (CIREN)* <
<http://bibliotecadigital.ciren.cl/>>
- 18) *Bibliothèque numérique de CAMES*
<<http://www.cames.bf.refer.org/spip.php?article56>>
- 19) *Bibliothèque SIST Sénégal* <<http://www.sist.sn/cgi-bin/library>>
- 20) *Books from the Past / Llyfrau o'r Gorffennol* <
<http://www.booksfromthepast.org>>
- 21) *Catalogo de la Biblioteca Obispo Angelelli*
<<http://biblioteca.derhuman.jus.gov.ar/cgi-bin/library?site=localhost&a=p&p=about&c=angeleli&ct=0&l=es&w=utf-8>>
- 22) *Chopin Early Editions* <<http://chopin.lib.uchicago.edu/>>
- 23) *CLACSO - Latin America and the Caribbean Network of Social Science Virtual Libraries* <
<http://www.biblioteca.clacso.edu.ar/>>
- 24) *Collection Greenstone de l'Université polytechnique de Bobo-Dioulasso (UPB)* < <http://greenstone.refer.bf/cgi-bin/library?e=p-00000-00---off-0--00---0-10-0---0---0prompt-10---4-----0-1l--10-fr-50---20-home---0--1-00-0-0-01-1-0utfZz-8-00&a=q&c=upb>>
- 25) *Collection of Ecole nationale des chartes (Paris)*
<<http://catalogue.enc.sorbonne.fr/>>

- 26) *Decifrazione del V. e VI. libro de' partimenti di Fenaroli del Cav(alliere) N(iccolò) C(alichiopulo) Manzano*
<<http://dlib.ionio.gr/gsdll/cgi-bin/library?a=p&p=about&c=decifraz>>
- 27) *Detroit Public Library* -
<<http://www.thehackley.org/about.html>>
- 28) *Estela* - <<http://estela.canouvelles.cat/cgi-bin/library?l=ca&w=utf-8>>
- 29) *Freedom House Photographs* -
<<http://www.lib.neu.edu/freedomhouse>>
- 30) *Greater Cincinnati Memory Project* -
<<http://www.cincinnatiemory.org/>>
- 31) *Great Lakes Shipping Database* -
<<http://www.dalnet.lib.mi.us/gsdll/cgi-bin/library?p=about&c=shipping>>
- 32) *Human Rights in Argentina* -
<<http://conadi.jus.gov.ar/greenstone>>
- 33) *iArchives*
<http://www.iarchives.com/demos_clients/greenstone.jsp>
- 34) *Illinois Wesleyan University Argus Digital Collection* - <
<http://europa.iwu.edu/gsdll/cgi-bin/library?c=argus&p=about>>
- 35) *Illustrated London News* -
<<http://digital.liby.waikato.ac.nz/iln/library?site=localhost&a=p&p=about&c=iln&ct=0&l=en&w=utf-8>>
- 36) *Indian Institute of Management, Kozhikode* <
<http://www.iimk.ac.in/gsdll/cgi-bin/library>>
- 37) *Indian Institute of Science Publications Database* -
<<http://vidya-mapak.ncsi.iisc.ernet.in/cgi-bin/library>>
- 38) *Kazakhstan Human Rights Commission* -
<<http://www.unesco.kz/cgi-bin/library?a=p&p=about&c=HRCru&l=ru&w=windows-1251&ct=1&qto=2>>
- 39) *Library of Kazak Governmental Legal Information* -
<<http://hrc.nabr.kz/gsdll/cgi-bin/library?site=hrc.nabr.kz&a=p&p=about&c=HRCKz&ct=1&qto=2&l=kk&w=utf-8>>
- 40) *Local History Online* -
<<http://www.localhistoryonline.org.nz/>>
- 41) *Marshall Foundation Digital Library* -
<<http://www.marshallfoundation.org/Database.htm>>

- 42) *Memoria Académica* -
<<http://www.memoria.fahce.unlp.edu.ar/>>
- 43) *Mirabilia Vicomercati* -
<<http://www.mirabiliavicomercati.org/sezioni/006/index.html>>
- 44) *MOST Digital Library (UNESCO)* - <<http://digital-library.unesco.org/shs/most/gsdll/cgi-bin/library?c=most&a=p&p=about>>
- 45) *Municipal Library of Almaty City* -
<<http://hrc.nabrk.kz/gsdll/cgi-bin/library?site=localhost&a=p&p=about&c=akalkz&ct=1&qto=2&l=kk&w=utf-8>>
- 46) *Music Information Retrieval Research*
<<http://www.music-ir.org/>>
- 47) *MyManuskrip : Digital Library for Malay Manuscripts*
<<http://mymanuskrip.fsktm.um.edu.my/>>
- 48) *National University of Science and Technology (NUST)*
<<http://library.nust.ac.zw/gsdll/cgi-bin/library>>
- 49) *New York Botanical Garden*
<<http://library.nybg.org/library/page1.php>>
- 50) *Notable Women of Simmons College* <
<http://my.simmons.edu/library/notablewomen/>>
- 51) *NZ Chinese Journals*
<<http://www.nzchinesejournals.org.nz/>>
- 52) *Union of BC Indian Chiefs* -
<<http://www.ubcic.bc.ca/Resources/ourhomesare/testimonies.htm>>
- 53) *Oxford Digital Library* -
<<http://www2.odl.ox.ac.uk/gsdll/cgi-bin/library/>>
- 54) *Pacific Archive of Digital Data for Learning and Education - PADDLE* - <<http://www.paddle.usp.ac.fj/>>
- 55) *Papers Past* - <<http://paperspast.natlib.govt.nz/cgi-bin/paperspast>>
- 56) *Rwanda HIV/SIDA* -
<<http://www.cnls.gov.rw/digitallibrary.htm>>
- 57) *State Library of Tasmania Sheet Music Collection* -
<<http://greenstone.statelibrary.tas.gov.au/>>
- 58) *Sudanese Association of Libraries and Information (SALI) Digital Library* -
<<http://puka.cs.waikato.ac.nz/cgi-bin/sali/library>>
- 59) *Sudan Open Archive* - <<http://www.sudanarchive.net/>>

- 60) The Arafura Digital Archive - <<http://arada.cdu.edu.au/cgi-bin/library>>
- 61) The Black Abolitionist Archive - <<http://www.dalnet.lib.mi.us/gsdll/cgi-bin/library?p=about&c=baa>>
- 62) The Council of Independent Colleges Historic Campus Architecture Project -<<http://puka.cs.waikato.ac.nz/cgi-bin/cic/library>>
- 63) The Cushing/Whitney Medical Digital Library- <<http://cwml.dl.med.yale.edu/gsdll/cgi-bin/library?site=localhost&a=p&p=about&c=ppcdot&ct=0&l=en&w=utf-8>>
- 64) The Social Management Digital Library - <<http://200.7.107.179/gsdll/cgi-bin/library>>
- 65) The United Nations Digital Library - Islamabad <<http://library.un.org.pk/gsdll/cgi-bin/library>>
- 66) The Writing University Archive - <<http://iwp.info-science.uiowa.edu/cgi-bin/library>>
- 67) Ulukau, the Hawaiian Electronic Library - <<http://ulukau.olelo.hawaii.edu/>>
- 68) Washington Research Library Consortium Special Collections - <<http://www.aladin.wrlc.org/dl/>>

O Greenstone, com a interface completa e toda a documentação, está disponível em vários idiomas como inglês, francês, espanhol, português e russo. É software do tipo *open-source*, multilíngue, multiplataforma compatível com *Microsoft Windows*, *UNIX*, *LINUX* e *Mac OS X*. Seus programas-fonte são disponíveis sob os termos da *General Public License* (GPL).

Até a versão 2.x o software foi desenvolvido, na linguagem de programação *PERL*, ele também utiliza o serviço *Apache Webserver*. A versão 3 (três) do Greenstone, é um redesenho completo e reimplementação do software original Biblioteca Digital Greenstone versão 2 (dois). Ela mantém muitas características e compatibilidades com a versão 2.x como por exemplo, continua multilíngue, multiplataforma, e altamente configurável. O Greenstone 3 foi escrito em *Java*, e é estruturado como uma rede de módulos independentes que se comunicam usando *XML*. Assim, ele é executado de forma distribuída, e sua aplicação pode estar distribuída em diferentes servidores de acordo com a sua necessidade. Esse design modular

aumenta a flexibilidade e a extensibilidade da Biblioteca Digital Greenstone.

A opção pelo *Greenstone 3* como objeto de estudo nesta pesquisa deu-se pelo fato de que, embora essa versão esteja em desenvolvimento, é a mais atualizada, e existe a recomendação expressa nos seguintes casos:

- a) Quando for necessária uma maior flexibilidade com a interface, e uso de *XSLT*;
- b) Se o usuário tem sua própria biblioteca de front-end e necessita conversar com um servidor de coleta Greenstone usando *XML* e *SOAP*;
- c) Quando o objetivo é de criar uma biblioteca distribuída; e,
- d) Quando o usuário deseja adicionar novas características a uma biblioteca Greenstone e têm dificuldade em entender o código fonte Greenstone2 C ++.

4.1 Obtendo o Greenstone versão 3.04

A versão 3.04 da biblioteca Digital *Greenstone* pode ser obtida no sitio <http://www.greenstone.org/greenstone3-home>, estando disponível para três tipos de plataformas.

- 1) Plataforma *Windows* – Disponível para ambiente Windows 32 bits (ou seja, *Windows 2000/XP/Vista/2003/2008*). Para versões do *Windows 95/98/Me/NT*, é necessário o uso da versão 2.8 do Greenstone. Essa distribuição inclui tudo que é necessário para executar Greenstone (incluindo uma coleção de demonstração pré-construída) e para construir novas coleções dentro do Greenstone. Opcionalmente, pode-se optar por instalar o *ImageMagick* (para processamento de imagem) e o *GhostScript* (para processamento de *PostScript*). O pacote de instalação do Greenstone 3.04 para Windows, agora inclui um servidor web Apache. O arquivo de instalação do Greenstone para essa plataforma ocupa aproximadamente 93.9 Mbytes.
- 2) Plataforma *MAC-OS* - Essa distribuição contém os binários ligados dinamicamente construídos e testados em *MacOS 10.5 (Leopard)* rodando em plataforma *Intel*. Essa distribuição inclui tudo que é necessário para executar Greenstone (incluindo uma coleção de demonstração pré-construída) e de construir coleções Greenstone. Opcionalmente, você pode optar por instalar o

ImageMagick (para processamento de imagem) e *GhostScript* (para processamento de *PostScript*). Para instalar essa distribuição, faz-se necessário baixar o arquivo *dmg*, montá-lo e depois executar o programa instalador do pacote, de preferência seguindo passo-a-passo as instruções. O arquivo de instalação do Greenstone para essa plataforma ocupa aproximadamente 64.5 Mbytes.

- 3) Plataforma *LINUX/UNIX* - Essa distribuição vem com os binários linux ligado estaticamente. Para compilar em outras plataformas Unix, é necessário baixar uma versão de origem. Essa distribuição inclui tudo que é necessário para executar Greenstone (incluindo uma coleção de demonstração pré-construída) e para construir novas coleções Greenstone. *ImageMagick* com o apoio *JPEG2000* está incluído, e podem ser instalados opcionalmente. Para instalar essa distribuição, e necessário realizar o download do pacote de instalação que está disponível no site oficial do Greenstone (www.greenstone.org), e depois executá-lo da linha de comando. O arquivo de instalação do Greenstone para essa plataforma ocupa aproximadamente de 87.5 Mbytes.

4.2 Instalação do Greenstone versão 3.04

Inicialmente optou-se por instalar a versão para Linux, utilizando a distribuição Debian. Realizado o download e iniciado a instalação, o autor se deparou com uma série de dificuldades de instalação, como a complexidade de instalação no Linux e necessidade de várias intervenções para configuração. Infelizmente a complexidade de instalação da Biblioteca Digital em uma plataforma aberta faz com que alguns usuários utilizem a plataforma de software proprietário devido a sua facilidade de instalação e operação. Sendo assim, a versão escolhida foi a 3.04 para *Windows*.

Após realizar o *download* da Biblioteca Digital Greenstone para *Windows*, iniciou-se a instalação. O computador utilizado foi um processador *Intel® core™ Duo CPU* de 2,4 GHz com 4 (quatro) GB de memória *RAM* e *HD* de 500 (quinhentos) GB. Durante a instalação no Sistema Operacional *Windows 7* (sete) *Professional 64* (sessenta e quatro) bits constatou-se a incompatibilidade com essa versão, sendo assim, para viabilizar a instalação da BDG, foi criado um ambiente virtual com sistema operacional *Windows XP* utilizando o *Windows*

virtual PC da Microsoft. O *Windows virtual PC da Microsoft* é uma ferramenta gratuita (*Free*) e pode ser obtida em diversas fontes. Para essa pesquisa foi realizado um *download* de <http://www.microsoft.com/windows/virtual-pc/>.

Para realizar a instalação da Biblioteca Digital Greenstone no *Windows XP*, basta localizar o arquivo de instalação Greenstone-3.04-win32 dar um clique duplo no mesmo que aparecerá a tela de instalação conforme figura 26.

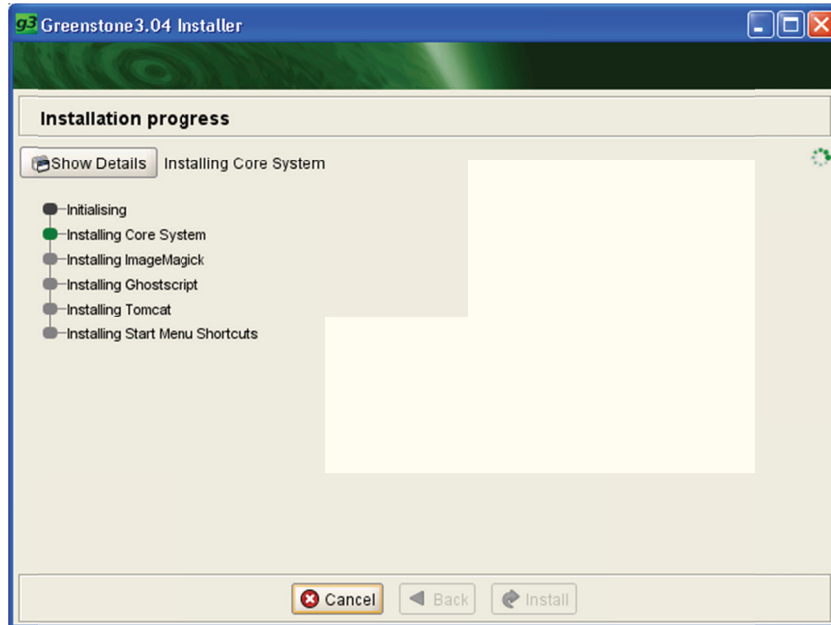


Figura 26: Tela - progresso da instalação

Na figura 27, o *Windows* mostra um aviso de segurança informando que o editor não pode ser verificado e pergunta se o usuário tem certeza que deseja executar este software. Mesmo que o software Greenstone 3.04 não possua uma assinatura digital válida que verifique o editor, pode-se executar o software posto que o mesmo é uma fonte confiável.

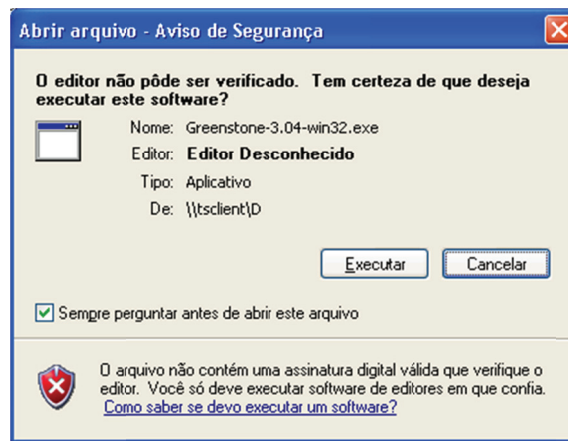


Figura 27: Tela de aviso de segurança

Na figura 28, o Greenstone mostra que está sendo inicializada a instalação.



Figura 28: Tela de preparação de instalação do Greenstone

Como a Biblioteca Digital Greenstone suporta vários idiomas, nesta etapa de instalação é possível escolher a linguagem no qual a BDG será instalada, como demonstrado na figura 29.

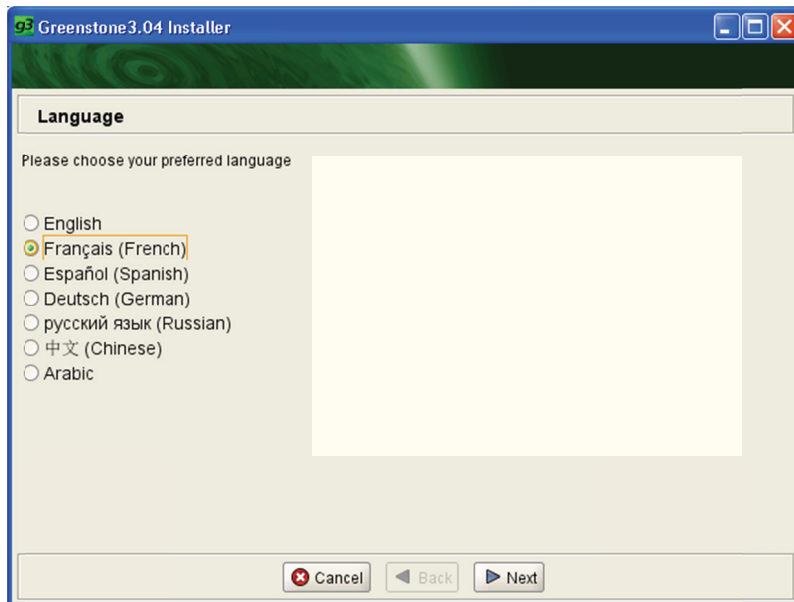


Figura 29: Tela de seleção de linguagem de preferência na instalação da BDG

A figura 30, apresenta as regras de direitos autorais, os quais são baseados no *General Public License GNU GPL* ou simplesmente *GPL*. O *GPL* (Licença Pública Geral) é a designação da licença para software livre idealizada por *Richard Stallman* no final da década de 1980, no âmbito do projeto *GNU* da *Free Software Foundation (FSF)*.

A *GPL* é a licença com maior utilização por parte de projetos de software livre, em grande parte devido à sua adoção para o projeto *GNU* e o sistema operacional *GNU/Linux*. Caso o usuário concorde com os termos, basta clicar no botão “Next” para ir para próxima etapa.

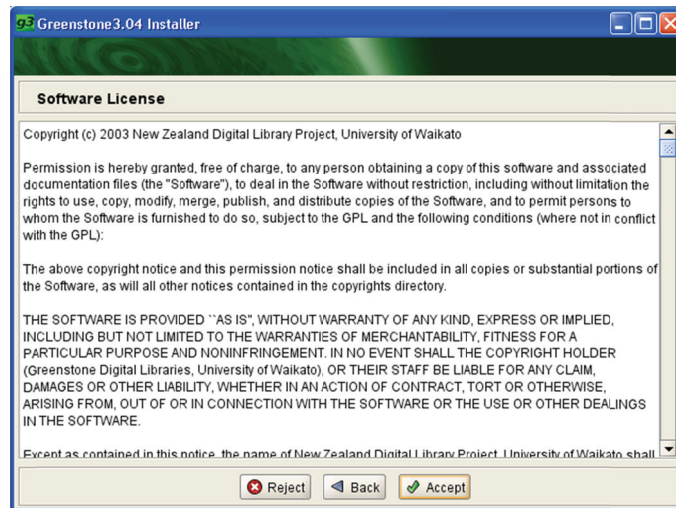


Figura 30: Tela de direitos autorais

A figura 31 mostra a tela de seleção do local de instalação. O software faz uma sugestão padrão para instalação. Caso o usuário concorde com a instalação, o mesmo deverá clicar o botão "next" para continuar a instalação.

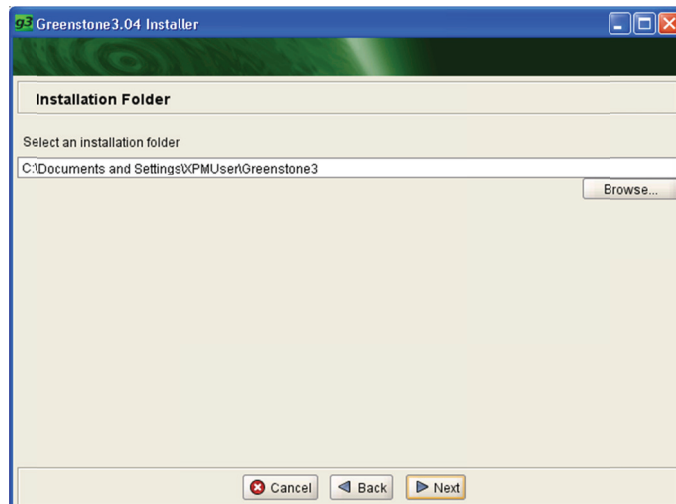


Figura 31: Tela de seleção do local de instalação

A figura 32, mostra a tela que permite a seleção dos componentes que foram instalados. A instalação do sistema principal (Core System) já vem selecionada, e não é opcional, sendo que as demais são opcionais. Recomenda-se que se instalar todas as opções de componentes.

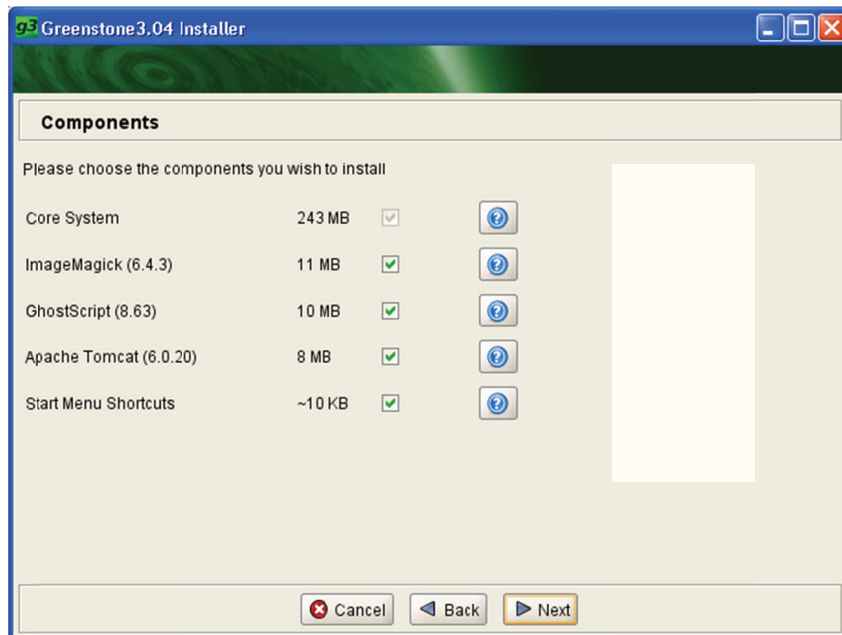


Figura 32: Tela de seleção dos componentes

A figura 33, mostra a tela que permite informar alguns parâmetros de configuração do servidor de *HTTP Apache Server*. O *Apache* é o servidor web livre mais bem sucedido no mundo. As portas 8080 e 8085 para conexão já vem configuradas inicialmente e recomenda-se manter as mesmas configurações.

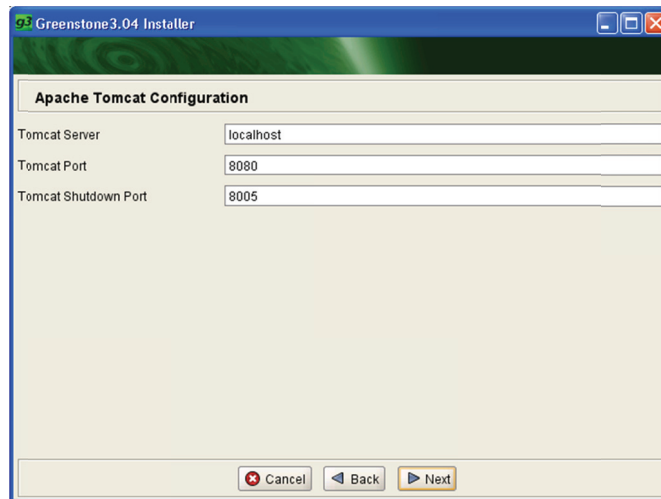


Figura 33: Tela configuração do *Apache Tomcat*

A figura 34, mostra etapas de instalação: 1 - Inicialização, 2 - Instalação do sistema principal, 3 - Instalação do *ImageMagick*, 4 - Instalação do *Ghostscript*, 5 - Instalação do *Tomcat*, 6 - Instalação e criação de atalhos do menu.

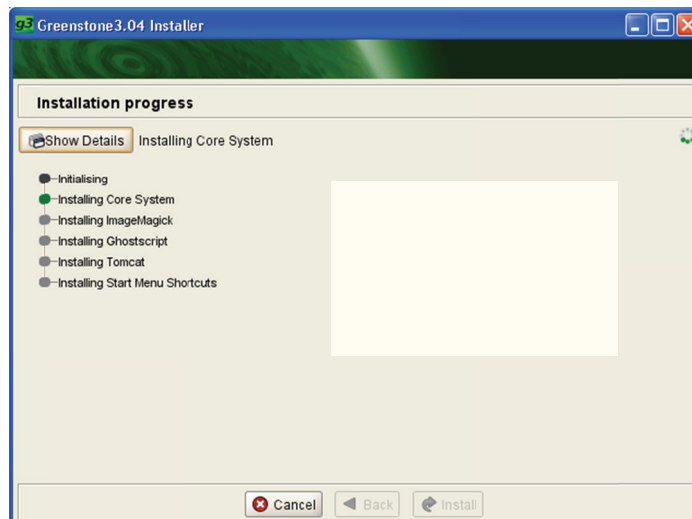


Figura 34: Tela que mostra o progresso de instalação do Greenstone

A Biblioteca Digital Greenstone pode ser iniciada a partir dos ícones disponíveis na área de trabalho do *Windows*, ou a partir do menu iniciar, selecionando a opção selecionando Greenstone. O Greenstone disponibiliza os seguintes opções:

- 1) Greenstone Editor for *Metadata Sets (GEMS)* – Software para edição de metadados
- 2) Greenstone3 Server – Inicia os softwares necessários para que o Greenstone Server funcione.
- 3) Greenstone *Librarian Interface (GLI)* – Interface para construção e configuração da Biblioteca Digital Greenstone.

Quando da instalação do Greenstone, o mesmo disponibiliza por padrão uma coleção de demonstração (coleção DEMO Greenstone), que é um pequeno subconjunto da Biblioteca de Desenvolvimento Humanitário (HDL).

No sítio www.greenstone.org estão disponíveis para *download* várias coleções de demonstração. As coleções relacionadas foram instaladas para servir de instrumento de análise de recuperação de informação, bem como para observar as funcionalidades das mesmas.

- 1) *DLS-e* - coleção Subconjunto da Biblioteca de Desenvolvimento - Da mesma maneira que o Demo Greenstone, este é um subconjunto da HDL - porém bem maior. Ela contém 250 publicações - livros, relatórios e revistas - em várias áreas do desenvolvimento humano (a completa HDL contém 1,230 publicações). Ela tem a mesma estrutura que o Demo Greenstone. É um pouco complexa, para quem está iniciando o seu aprendizado com bibliotecas digitais. O espaço requerida para instalação dessa coleção é de 150 Mb.
- 2) *WRDPDF-e* - Demonstrações *MSWord* e *PDF* – Essa coleção possui diversos documentos nos formatos PDF, *MSWord*, *RTF*, and *Postscript*, demonstrando a habilidade para construir coleções com documentos de tipos diferenciados. O espaço requerida para instalação dessa coleção é de 4 Mb.

- 3) *GSARCH-e* - A Coleção dos Arquivos Históricos do Greenstone - Uma coleção de mensagens de e-mail das listas históricas do Greenstone, que utiliza o plug-in e-mail, dividindo arquivos em formatos de e-mail. O arquivo de configuração da coleção é bem simples. O espaço requerida para instalação dessa coleção é de 5 Mb.
- 4) *CLTBIB-e* - Coleção bibliográfica com aproximadamente 4.000 entradas bibliográficas, essa coleção incorpora uma interface de busca baseada em formulário que permite a busca por campos. É bastante complexa. O espaço requerida para instalação dessa coleção é de 7 Mb.
- 5) *CLTEXT-e* - Suplemento Bibliográfico - Essa pequena coleção de 10 entradas bibliográficas ilustram os recursos da "supercoleção" que permite a busca de várias coleções ao mesmo tempo. Ela trabalha junto com a coleção Bibliografia, e os seus arquivos de configuração são quase os mesmos. O espaço requerida para instalação dessa coleção é de 1 Mb.
- 6) *MARC-e* exemplo com MARC - Baseada em arquivos MARC da Biblioteca do Congresso, essa coleção é simples (e não permite busca baseada em formulário). O espaço requerida para instalação dessa coleção é de 1 Mb.
- 7) *OAI-e* - A coleção Demo do OAI - Utilizando o Protocolo *Open Archive* e a opção *Import-From*, ele recupera o histórico dos metadados, sendo possível utilizá-los para construir uma coleção com estes registros. Neste caso eles são imagens, portanto *os plug-ins OAI e Image* são utilizados. O espaço requerida para instalação dessa coleção é de 18 Mb.
- 8) *IMAGE-e* - Coleção simples de imagens - Essa coleção bem básica de imagens não contém texto nem metadados explícitos - o que a torna não muito realística. O arquivo de configuração é o mais simples que pode haver. O

espaço requerida para instalação dessa coleção é de 1 *Mb*.

- 9) *AUTHEN-e* - A formatação e autenticação da coleção demo. Utilizando o mesmo material da coleção original Demo do Greenstone, duas características independentes podem ser mostradas: a formatação de documentos fora do padrão, e controle de acesso aos documentos utilizando a autenticação de usuário. O espaço requerida para instalação dessa coleção é de 1 *Mb*.
- 10) *GARISH - Versão Garish* da coleção demo. Essa coleção também contém o mesmo material do demo Greenstone. A sua aparência foi alterada para demonstrar como as páginas geradas podem ser configuradas de modo diferente. Ele se baseia na utilização de um arquivo macro sem um padrão definido que é fornecido pelo Greenstone. O espaço requerida para instalação dessa coleção é de 8 *Mb*.
- 11) *ISIS-e* - exemplo *CDS/ISIS* - Essa coleção é construída a partir de um banco de dados *CDS/ISIS* com aproximadamente 150 entradas bibliográficas. Utiliza o *plug-in ISISPlug*, que lê os arquivos de padrão ISIS .mst e .fdt e os converte para os metadados do Greenstone. O espaço requerida para instalação dessa coleção é de 1 *Mb*.

4.3 Construindo coleção de dissertações do PGCIN

Com o objetivo de obter mais subsídios para analisar os recursos disponíveis na biblioteca digital Greenstone na preservação lógica de documentos digitais e a recuperação da informação, foi construído uma coleção dentro da Biblioteca Digital Greenstone utilizando as dissertações de mestrado defendidas no Programa de Pós-Graduação em Ciência da Informação do Centro de Ciências da Educação, da Universidade Federal de Santa Catarina. A seguir será detalhada a construção da coleção PGCIN, demonstrando uma sequência de passos com as respectivas telas do Greenstone tendo como objetivo de situar o leitor sobre o processo de construção da coleção.

Para criar uma nova coleção, no módulo de interface de biblioteca (*Greenstone Librarian Interface - GLI*) conforme a figura 35, basta clicar na opção novo (*NEW*) do *menu* (File), onde deverá ser informado o Título da Coleção e a descrição do conteúdo da mesma. O Greenstone permite criar uma coleção tendo como base outra coleção, para tanto, basta selecionar a coleção desejada na opção “*Base this collection on*”.

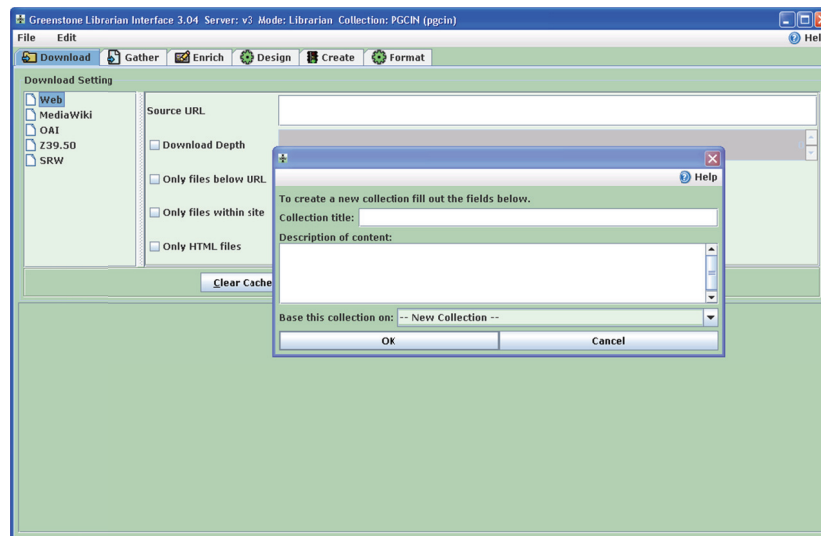


Figura 35: Tela de criação da coleção do PGCIN

Após a criação da coleção, foi realizado o download das dissertações de mestrado do PGCIN que estão disponíveis em <http://www.cin.ufsc.br/pgcin>, conforme figura 36.

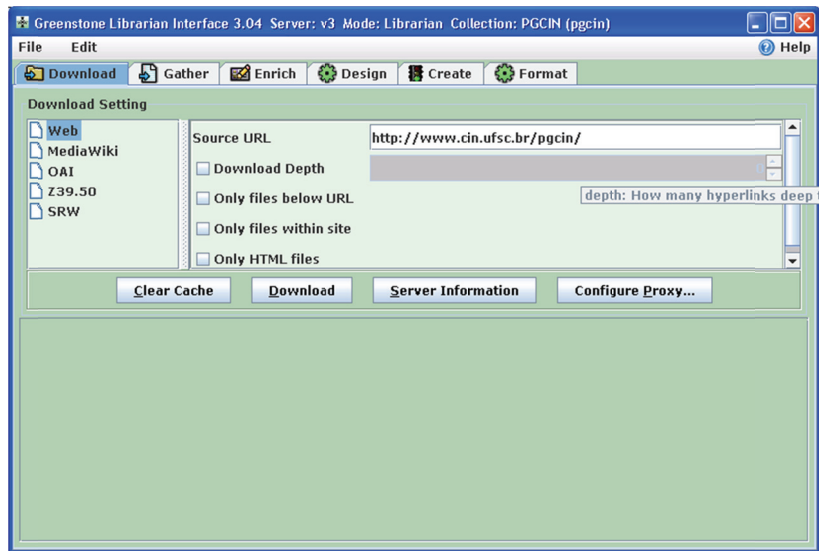


Figura 36: Tela download da coleção do PGCIN

Após o download dos arquivos, conforme figura 37, na aba *Gather* foi criada a pasta “dissertações”. Para criar a pasta, basta clicar com o botão direito na área *collection* e criar uma nova pasta (new folder). Após ter criado a pasta dissertações foram selecionados os arquivos (dissertações) no formato PDF que se encontram no lado esquerdo da aplicação e arrastado para a pasta “dissertações”.

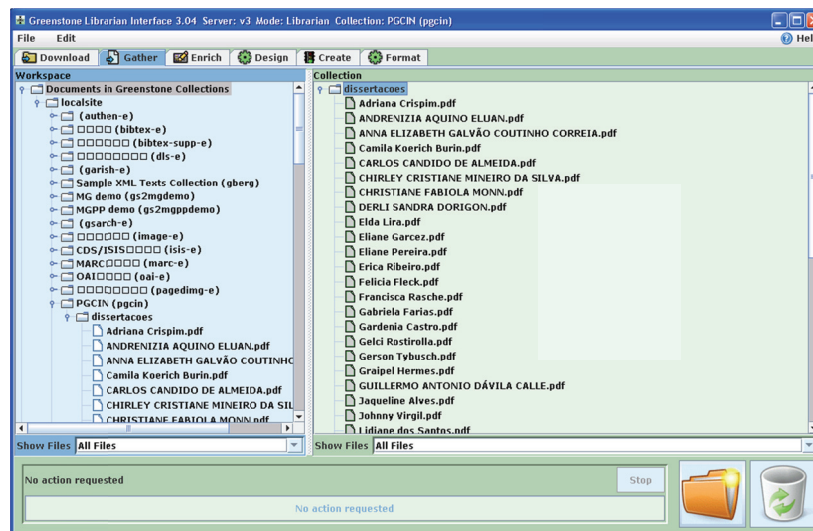


Figura 37: Tela importação de documentos

Conforme figura 38, além dos metadados que o Greenstone extrai automaticamente dos documentos adicionados a coleção, foram adicionados a coleção do PGCIN metadados padrão Dublin Core, e preenchidos alguns dos elementos como *dc.title*.

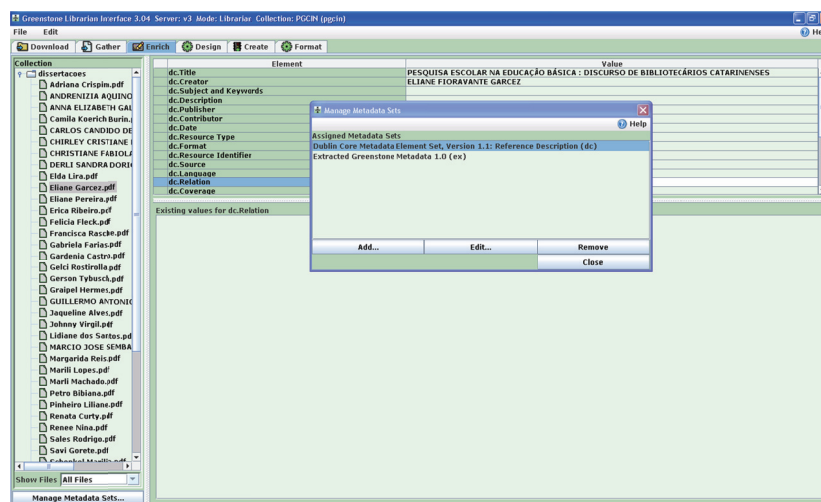


Figura 38: Administração de metadados

Foi realizado um levantamento de todos de todos os plug-ins para formatos de arquivos disponíveis para o Greenstone. No sítio do Greenstone na internet (www.greenstone.org) existe uma documentação sobre como criar plug-ins, bem como links que apontam a relação de plug-ins suportados.

De acordo informações encontradas no *Wiki* do Greenstone disponível em http://wiki.greenstone.org/wiki/index.php/Main_Page, os plug-ins estão classificados em quatro tipos:

a) *Plug-ins* especiais - nível superior

| Nome do plug-in | Description Descrição |
|---------------------|--|
| BibTexPlug-in | Plug-in para importações de arquivos BibTex. Herda SplitTextFile. |
| BookPlug-in | Plug-in que importa arquivos da coleção Biblioteca Humanidade. É uma simplificação do HBPlug-in. Herda AutoExtractMetadata. |
| CONTENTdmPlug-in | Plug-in que importa arquivos RDF de coleções exportadas. Herda ConvertBinaryFile, ReadXMLFile. |
| ConvertToRogPlug-in | Herda RogPlug-in. |
| CSVPlug-in | Plug-in que importa arquivos no formato de valores separados por vírgula. Um novo documento será criado para cada linha do arquivo. Herda SplitTextFile. |
| DatabasePlug-in | Plug-in que extrai registros de bancos de dados (requer configuração adicional Perl). Herda AutoExtractMetadata. |
| DSpacePlug-in | Plug-in que as importações de formato de arquivo DSpace Herda BasePlug-in. |

| | |
|-------------------|--|
| EmailPlug-in | Plug-in que importa arquivos de e-mail. Herda SplitTextFile. |
| ExcelPlug-in | Plug-in que importa arquivos do Microsoft Excel. Herda ConvertBinaryFile. |
| FavouritesPlug-in | Plug-in que importa arquivos favoritos do Internet Explorer. Herda ReadTextFile. |
| FOXPlug-in | Plug-in que importa arquivos de dados FOX. Herda BasePlug-in. |
| HBPlug-in | Plug-in que as importações de um diretório do livro de HTML. Utilizados pela coleção da biblioteca da humanidade. Herda BasePlug-in. |
| HTMLPlug-in | Plug-in que importa arquivos HTML. Herda ReadTextFile, HBPlug-in. |
| HTMLImagePlug-in | Plug-in que importa arquivos HTML, criando um documento Greenstone para cada imagem na página web. Herda HTMLPlug-in. |
| ImagePlug-in | Plug-in que as importações GIF, JIF, JPEG, TIFF http://www.imagemagick.org/www/formats.html . Herda BasePlug-in, ImageConverter. |
| IndexPlug-in | Plug-in que processa um arquivo index.txt, que lista todos os arquivos a serem incluídos na coleção, além de metadados adicionais para esses documentos. Herda BasePlug-in. |

| | |
|--------------------|---|
| ISISPlug-in | Plug-in que as importações CDS / arquivos de dados ISIS. Herda SplitTextFile. |
| LaTeXPlug-in | Plug-in that imports LaTeX files. plug-in que importa arquivos LaTeX. Inherits from ReadTextFile. Herda ReadTextFile. |
| LOMPlug-in | Plug-in que as importações LOM (Learning Object Metadata) arquivos. Herda ReadTextFile. |
| MARCPlug-in | Plug-in para importações de metadados MARC. Herda SplitTextFile. |
| MARCXMLPlug-in | Plug-in para importações de metadados MARC em formato XML. Herda ReadXMLFile, ReadTextFile. |
| MediaWikiPlug-in | Plug-in para importações de páginas web MediaWiki. Herda HTMLPlug-in. |
| MetadataCSVPlug-in | Plug-in para importações de metadados no formato CSV (valor separado por vírgula) formato. O campo Nome do arquivo CSV é usado para determinar quais os metadados do documento pertence. Herda BasePlug-in. |
| MP3Plug-in | Plug-in que importa arquivos de áudio MP3. Herda BasePlug-in. |

| | |
|---------------------|--|
| NulPlug-in | Plug-in que importa arquivos dummy (.Nul). Herda BasePlug-in. |
| OAIPlug-in | Plug-in para importações Open Archives Initiatives (OAI) de dados. Herda ReadXMLFile, ReadTextFile. |
| OggVorbisPlug-in | Plug-in para importações Ogg Vorbis. Herda BasePlug-in. |
| OpenDocumentPlug-in | Plug-in para importações OASIS documentos de formato OpenDocument (usado pelo OpenOffice 2.0. Herda ReadXMLFile. |
| PagedImagePlug-in | Plug-in para importações de seqüências de arquivos de imagem (formatos como para ImagePlug), com opcional de textos associados. Herda ReadXMLFile, ReadTextFile, ImageConverter. |
| PDFPlug-in | Plug-in que importa arquivos PDF. Herda ConvertBinaryFile. |
| PostScriptPlug-in | Plug-in que importa arquivos Postscript. Herda ConvertBinaryFile. |
| PowerPointPlug-in | Plug-in que importa arquivos do PowerPoint Microsoft. Herda ConvertBinaryFile. |
| ProCitePlug-in | Plug-in que importa arquivos ProCite. Herda SplitTextFile. |

| | |
|------------------------|---|
| RealMediaPlug-in | Plug-in que importa arquivos de RealMedia. Herda BasePlug-in. |
| ReferPlug-in | Plug-in que as importações Consulte os arquivos. Herda SplitTextFile. |
| RogPlug-in | Plug-in para importações. Rog ou arquivos. Mdb. Herda BasePlug-in. |
| RTFPlug-in | Plug-in que importa arquivos RTF. Herda ConvertBinaryFile. |
| SourceCodePlug-in | Plug-in que as importações do código-fonte (C / C ++, Perl, Shell). Herda ReadTextFile. |
| StructuredHTML Plug-in | Plug-in para importações de documentos HTML estruturado, dividi-los em seções com base em informações de estilo. Herda HTMLPlug-in. |
| TextPlug-in | Plug-in que importa arquivos de texto. Herda ReadTextFile. |
| UnknownPlug-in | Plug-in que importa arquivos com uma extensão de arquivo especificado pelo usuário. Nenhum processamento é feito no arquivo. Em vez de um documento fictício é criado o arquivo está anexado a esse documento. Usado para importar arquivos que Greenstone não possam lidar. Herda BasePlug-in. |

| | |
|-------------|---|
| WordPlug-in | Plug-in que as importações de documentos do Microsoft Word. Herda ConvertBinaryFile. |
| ZIPPlug-in | Plug-in que extrai arquivos comprimido ou formatos de arquivo e envia o conteúdo para baixo do plug-in pipeline. Incluem gzip (.gz, z., tgz, taz), bzip (.BZ), bzip2 (.bz2), zip (.zip, jar.) e tar (.tar). Solicita o utilitário apropriado: gunzip, bunzip, bunzip2, unzip, tar. Herda BasePlug-in. |

Quadro 3: Plug-ins especiais – nível superior

b) *plug-ins* especiais - nível superior

| | |
|---------------------|--|
| DirectoryPlugin | Processos de diretórios: através de um diretório recursivamente, passando cada arquivo que encontrado. Utilizados durante a colheita de importação e de construção. Herda PrintInfo. |
| MetadataXMLPlugin | Processos arquivos metadata.xml que são gerados por GLI. Utilizados durante a colheita de importação. Herda BasePlugin. |
| ArchivesInfPlugin | Processa o archives.inf arquivo gerado durante a importação. Utilizados durante a construção única coleção. Herda PrintInfo. |
| GreenstoneXMLPlugin | Processa os documentos de arquivo Greenstone. Utilizados durante a construção única coleção. Herda ReadXMLFile. |
| GreenstoneMETSPugin | Arquiva documentos em formato METS Greenstone. Utilizados durante a construção única coleção. Herda ReadXMLFile. |

Quadro 4: Plug-ins especiais – nível superior

c) *Plug-ins* Base

| | |
|-----------|--|
| PrintInfo | Classe base para todos os plugins e plugins auxiliar. Contém o código para gerar a saída para plugininfo.pl, e para analisar os argumentos plugin. |
|-----------|--|

| | |
|--------------------------|--|
| BasePlugin BasePlugin | Classe base para todos os plugins documento padrão. Contém o código para bloqueio de arquivos, manipulação codificação nome, associando arquivos relacionados, e atribuição de identificadores doc. Herda PrintInfo. |
| AutoExtractMetadata | Classe base para plug-ins que os processos de documentos com texto. Utiliza todos os plugins auxiliares para adicionar funcionalidade extra para BasePlugin, como a extração automática de metadados. Herda BasePlugin e todos os plugins do ajudante. |
| ReadTextFile | Classe base para plug-ins que o processo simples arquivos de texto. Contém o código para a leitura dos autos e elaboração da linguagem e codificação. Herda AutoExtractMetadata. |
| ReadXMLFile | Classe base para plug-ins que processar arquivos XML. Contém o código para gerar e executar um parser XML. Herda BasePlugin. |
| ConvertBinaryFile | Classe base para plug-ins que processar arquivos binários que são convertidos em texto / html / images executando gsConvert.pl. Contém código para chamar gsConvert.pl, a criação de plugins do secundário, que irá processar o arquivo convertido, e passar o arquivo para os plugins. Herda AutoExtractMetadata. |
| SplitTextFile | Classe base para processar arquivos de plugins que contém muitos registros. Contém o código que se divide o texto em segmentos, que depois são processadas pelo plugin de nível superior. Herda ReadTextFile. |

Quadro 5: Plug-ins Base

d) *Plug-ins* Auxiliares

| | |
|--------------------|--|
| BaseMediaConverter | plugin que fornece funcionalidade básica, como o cache de arquivos para conversão de mídia. Herda PrintInfo. |
|--------------------|--|

| | |
|---------------------|--|
| ImageConverter | plugin que converte imagens usando ImageMagick. Herda BaseMediaConverter. |
| Acronym Acrônimo | Helper plugin que localiza e marca-se siglas no texto. Herda PrintInfo. |
| Date Data | plugin que extrai informações de data histórica do texto. Herda PrintInfo. |
| EmailAddress | plugin que extrai endereços de e-mail de texto. Herda PrintInfo. |
| GIS | plugin que extrai placenames do texto. Requer a extensão GIS Greenstone. Herda PrintInfo. |
| Keyphrase | plugin que gera keyphrases do texto. Utiliza sistema de extração Kea keyphrase. Herda PrintInfo. |

Quadro 6: Plug-ins auxiliares

Os seguintes formatos de arquivos aguardam o desenvolvimento de *plug-in* para o Greenstone:

Para Documentos de escritório

- a) *AbiWord*
- b) *Gnumeric Spreadsheet*
- c) *Kword (all Koffice formats)*
- d) *OpenOffice file formats: Writer (.sxw), Calc (.sxd), Impress (.sxi), Draw (.sxd)*
- e) *StarOffice formats (.sdc, .sdw)*
- f) *Wordperfect*

Para Video:

- a) *MPEG*
- b) *Quicktime (.mov)*
- c) *AVI (Audio Video Interleave), Microsoft video*

Para Audio:

- a) *Windows Media Audio (.wma)*
- b) *Windows audio (.wav)*
- c) *Sun Audio (.au)*
- d) *Audio Interchange File Format (.aiff)*
- e) *MIDI (.mid)*
- f) *MIDI karaoke (.kar)*
- g) *CD Audio (.cda)*
- h) *Shorten (.shn)*

Anotações:

- a) *Endnote*

Images:

- a) *DjVu (.djvu)*
- b) *Photoshop (.psd)*
- c) *PaintShopPro (.psp)*

Arquivos para Macintosh:

- a) *.hqx Mac archive*
- b) *.sit*
- c) *Self extracting Archive (.sea)*

Outros:

- a) *Scalable Graphics Format (.svg)*
- b) *Synchronized Multimedia Integration Language SMIL (.smil)*
- c) *Macromedia Flash (.fla)*
- d) *Macromedia shockwave (.swf)*
- e) *OpenGL*
- f) *VRML/X3D*
- g) *TrueType Fonts (TTF)*

Neste trabalho de pesquisa foram escolhidos pelo menos um plug-ins que contemplese formatos de arquivo proprietários com especificação fechada, proprietário com especificação aberta, e não proprietários com especificação aberta.

Conforme figuras 39 e 40, a partir da aba “Design” do Greenstone, foram adicionados os seguintes plug-ins para formatos de arquivos digitais:

- a) *PDFPlug-in* – para documentos do tipo PDF
- b) *OpenDocumentPlug-in* – para documentos do tipo formado aberto
- c) *GreenstoneXMLPlug-in* – para documentos do tipo padrão XML
- d) *RTFPlug-in* – para documentos do tipo RTF
- e) *TextPlug-in* – para documentos do tipo texto
- f) *WordPlug-in* – para documentos do tipo *Microsoft Word*
- g) *PowerPoint plug-in* – para documentos do tipo *Microsoft Powerpoint*
- h) *ExcelPlug-in* – para documentos do tipo *Microsoft Excel*

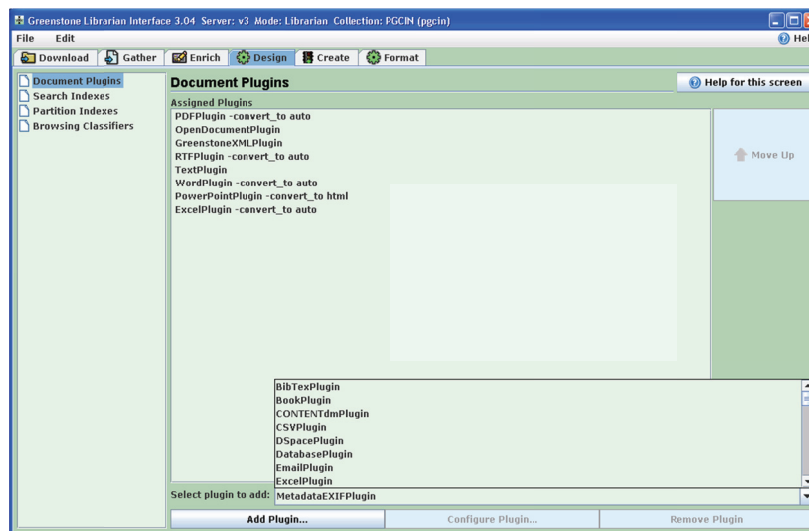


Figura 39: Tela para adicionar plug-ins de formato de arquivo digital

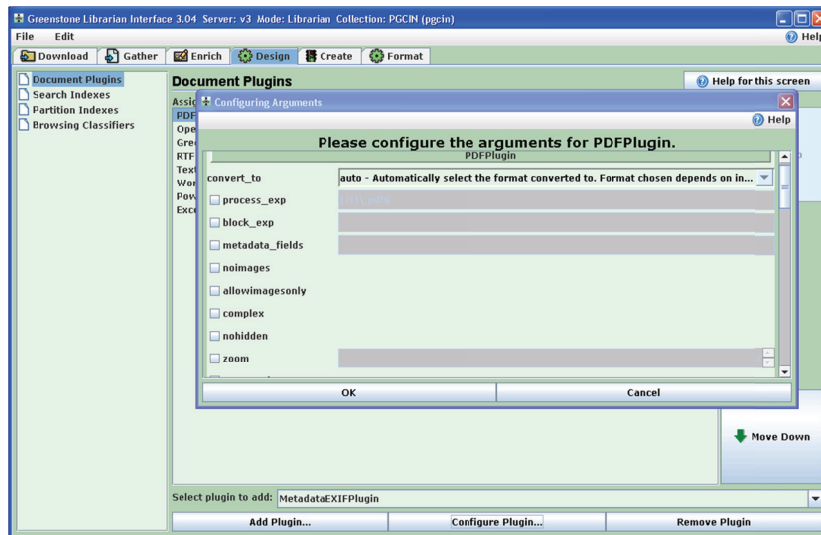


Figura 40: Tela configuração de plug-ins

Conforme figura 41, o Greenstone permite a indexação do texto inteiro (*full text*), bem como, adicionar metadados na indexação. Ainda na aba *Design – Search Indexes*, é possível selecionar as opções de indexação como:

- a) *Stem* – Gera um arquivo de stemming que consiste em eliminar as variações morfológicas de uma palavra, onde as mesmas são eliminadas através da identificação do radical de uma palavra.
- b) *Casefold* - Opção para não diferenciar letras maiúsculas e minúsculas.
- c) *Accent fold* – Opção para desconsiderar acentuação
- d) *CJK Tex Segmentation* – CJK acrescenta um espaço entre cada caractere Chinese Japanese Korean.
- e) *Indexes Levels Default* – Nível de indexação por documento ou seção do documento

No desenvolvimento dessa pesquisa foram testadas todas estas opções de indexação.

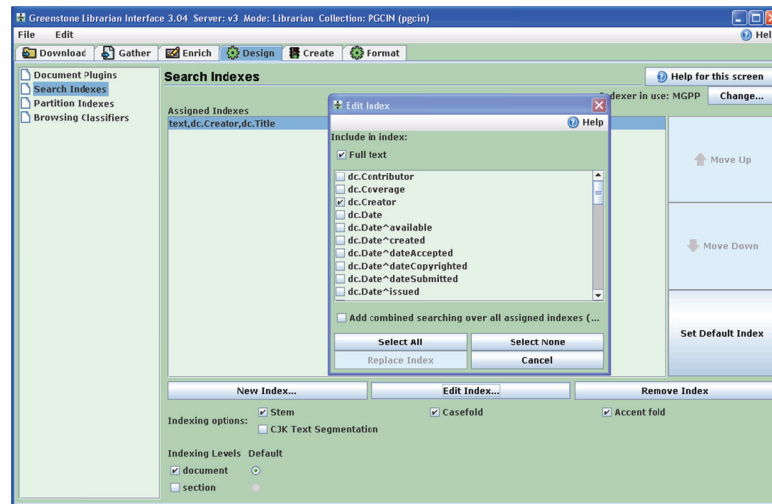


Figura 41: Tela de opção de indexação

Conforme figura 42, o Greenstone oferece três ferramentas de indexação: o *MG*, que é o indexador padrão, o *MGPP* (*MG++*) e o *Lucene* (*Apache Software Foundation*) que possuem características sofisticadas de indexação e busca.

No desenvolvimento dessa pesquisa foram testadas todas estas opções de indexação.



Figura 42: Tela de opção de indexação MGPP, MG e LUCENE

Conforme consta no site do Greenstone na internet disponível em <www.greenstone.org>, o *MG "Managing Gigabytes"* é o indexador original usado pelo Greenstone, desenvolvido principalmente por *Alistair Moffat* e baseado no livro com o mesmo nome “Gerenciando Gigabytes”. Para cada índice especificado na coleção, um arquivo de índice separado é criado. Este indexador foi testado extensivamente em coleções muito grandes, ou seja, vários *Gigabytes* de texto.

MGPP (ou *Mg ++*) é uma reimplementação de *MG*, que prevê índices de nível de documento, e compressão dos documentos originais. Uma pequena mudança na configuração de arquivo para uma coleção é tudo o que é necessário para usar *MGPP*.

O *Lucene* foi desenvolvido pela Apache Software Foundation. Ele realiza pesquisa por proximidade, mas apenas em um único nível. Foi adicionado ao Greenstone para facilitar a criação de coleções incrementais, que *MGPP* e o *MG* não oferecem.

Conforme figura 43, o Greenstone oferece a opção de associar uma linguagem na partição de indexação. Na construção da coleção do PGCIN foi selecionado o idioma português.

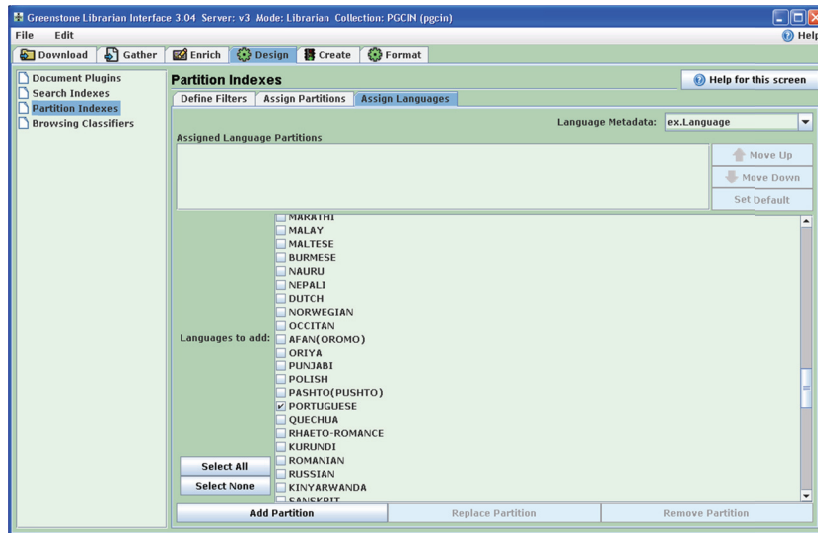


Figura 43: Tela de associação de língua na partição de indexação

Conforme figura 44, o Greenstone oferece além da recuperação pelo texto completo e ainda tem a opção de recuperar documentos utilizando filtros como por título da dissertação, autor da dissertação, orientadores, linha de pesquisa e ano dissertação.

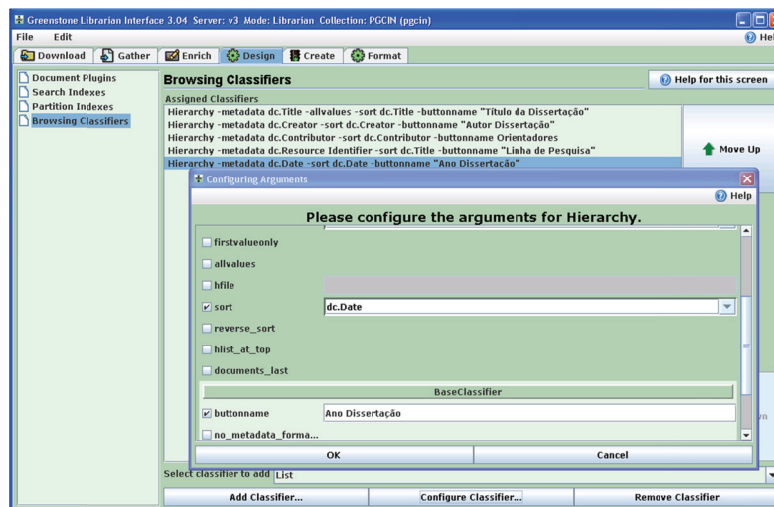


Figura 44: Tela configuração de *browsing classifiers*

A figura 45 e 46 mostra o início e fim da importação dos documentos. Foi realizado o download de 47 dissertações de mestrado no formato *PDF* do site <http://www.cin.ufsc.br/pgcin> e criado 12 (doze) arquivos de vários tipos de formato de arquivo. Dos 59 (cinquenta) documentos disponibilizados para a coleta, 44 (quarenta e quatro) documentos foram importados para a coleção do PGCIN, sendo que para 6 (seis) documentos o Greenstone não reconheceu o formato de arquivo, e outros 9 (nove) foram rejeitados pelo plug-in *PDF* disponível no Greenstone. Aparentemente os arquivos estavam íntegros, posto que os softwares nos quais foram gerados estavam abrindo normalmente os documentos, como por exemplo os arquivos com a extensão *ODT* (padrão *ODF*).

O Greenstone participa do Projeto *Open Source Trac*, que é um *wiki* melhorado, e um sistema de monitoramento para projetos de desenvolvimento de software o qual está disponível em <http://trac.greenstone.org/browser/main/trunk/greenstone2/perl/lib/plugin/s/>. Com base nas informações encontradas no *TRAC* foi alterado o plug-in "*OpenDocumentPlug-in*" e foi reconstruída a coleção do PGCIN.

Com estas alterações a BDG passou a reconhecer e processar o plug-in para arquivos no formato ODF.

A Biblioteca Digital Greenstone em um computador *Intel® core™ Duo CPU* de 2,4 GHz com 04 GB de memória RAM e HD de 500 GB, levou aproximadamente de 10 (dez) minutos para realizar o processamento da importação dos documentos, e mais 4 (quatro) minutos para realizar a compressão do texto.

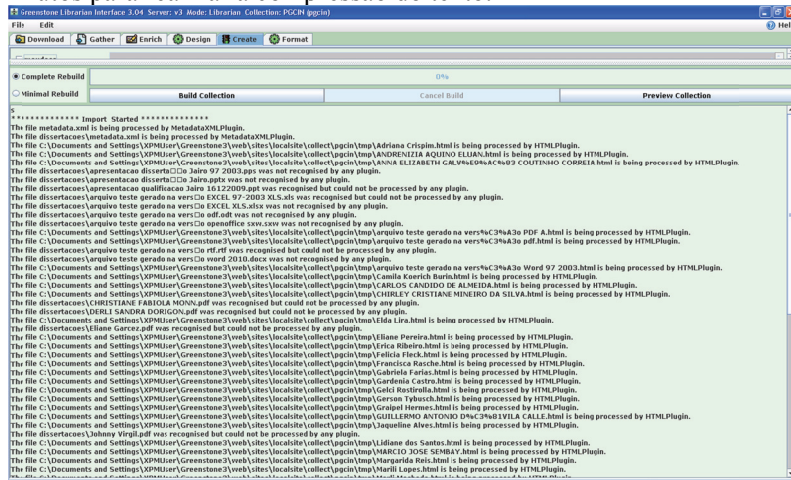


Figura 45: Tela início de importação de documentos

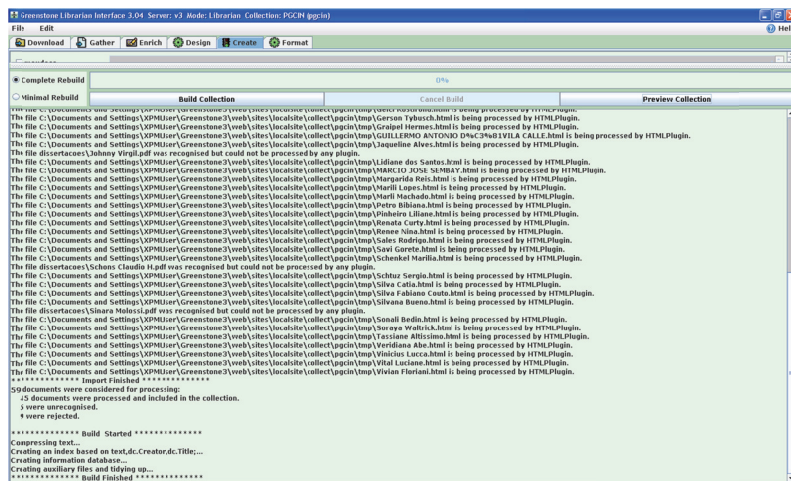


Figura 46: Tela fim de importação de documentos

Conforme mostra a figura 47, na aba *format*, é possível informar os dados gerais sobre a coleção que está sendo criada, como por exemplo o e-mail do criador e da pessoa que vai dar manutenção, o título, pasta, e o ícone da coleção, descrição da coleção.

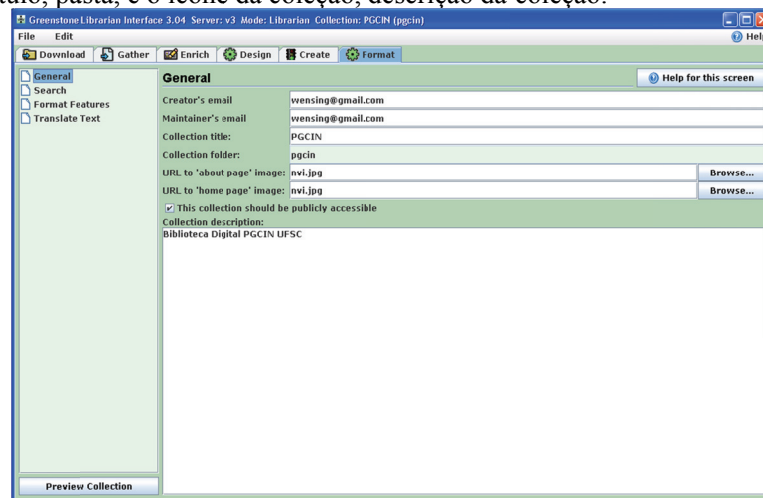


Figura 47: Tela aba format - dados gerais

Conforme mostra a figura 48, na aba *format*, é possível informar os itens de pesquisa no menu, a indexação é a para todo documento, e os índices escolhidos são o Texto integral (*full text*), *dc.creator*, *dc.title*.

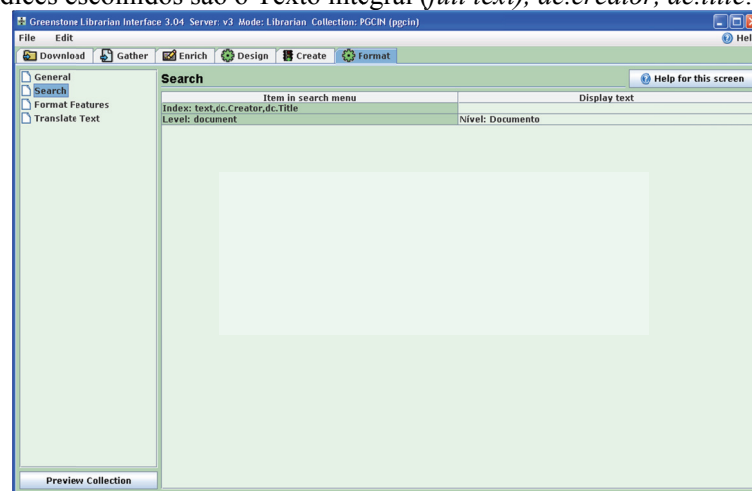


Figura 48: Tela aba format - itens de pesquisa no menu

Conforme mostra a figura 49, na aba *format*, é possível informar o formato dos recursos.

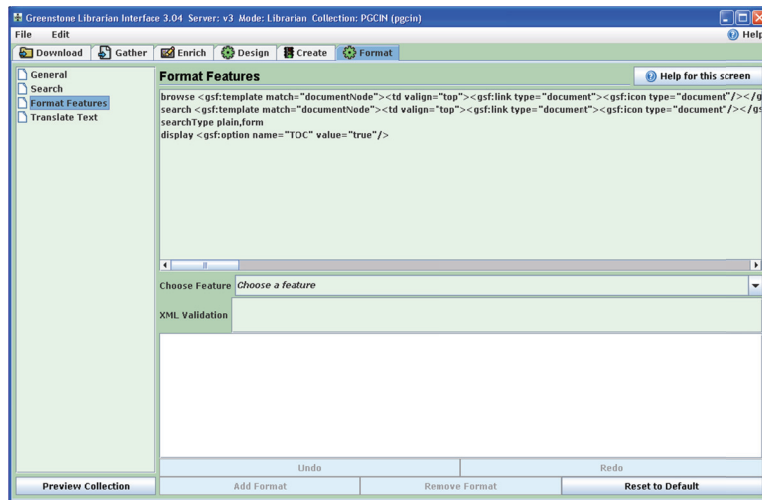


Figura 49: Tela aba *format* – recursos do formato

Conforme mostra a figura 50, na aba *format*, é possível realizar a tradução de algumas informações.

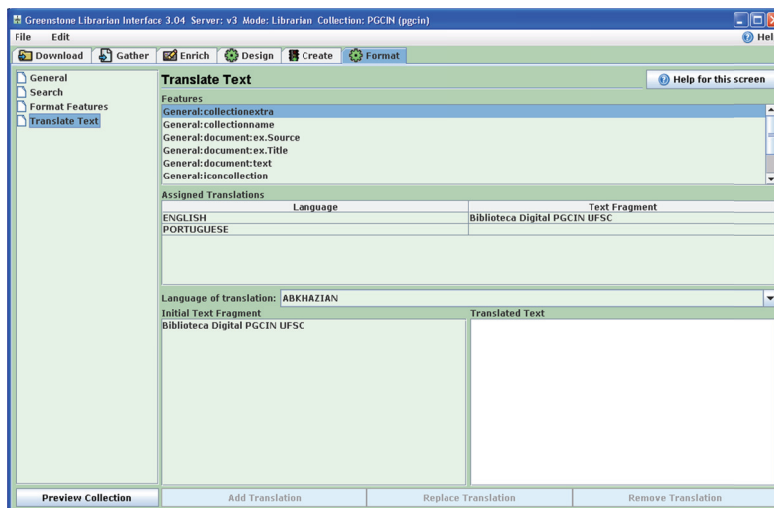


Figura 50: Tela aba *format* – tradução de textos

A coleção gerada poderá ser consultada clicando na aba “create” na opção “*preview collection*”, ou acessar um *web browser* informando o endereço onde a aplicação está instalada, que nesse caso é o seguinte <http://localhost:8080/greenstone3/library?a=p&sa=home> conforme mostra a figura 51.

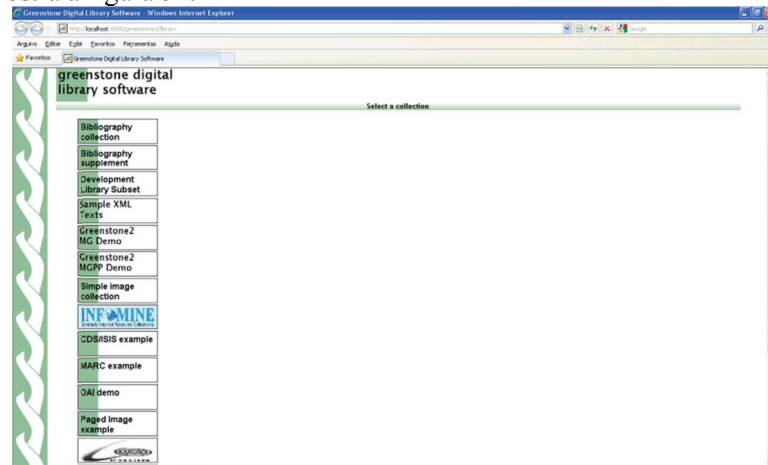


Figura 51: Tela de acesso a todas as coleções instaladas

Para selecionar a coleção gerada que no caso é a de dissertações do PGCIN, basta clicar no ícone PGCIN que a mesma ficará disponível para consulta conforme figura 52.

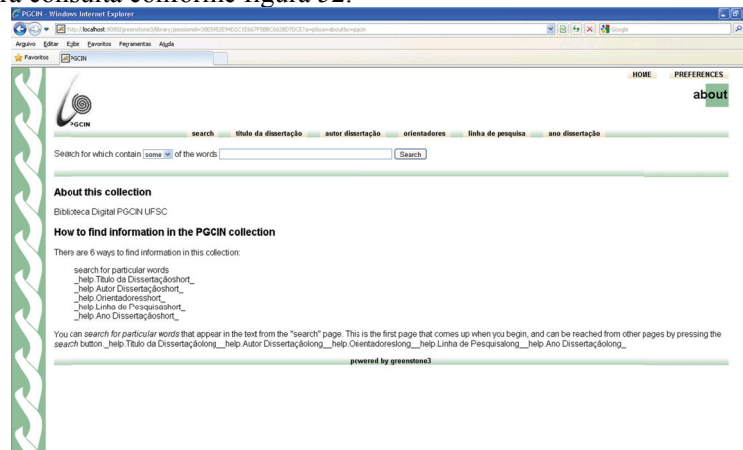


Figura 52: Tela inicial de consulta da coleção PGCIN na BDG

Conforme figura 53, ao escolher a opção preferências, é possível escolher o idioma da interface e preferências de busca.

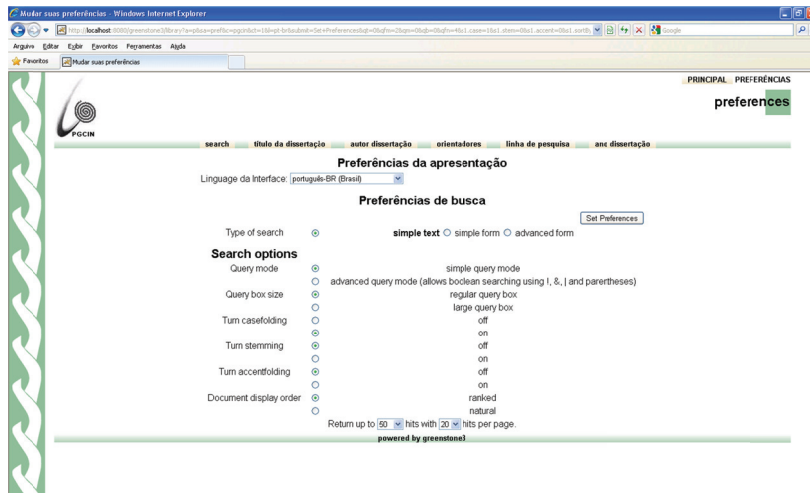


Figura 53: Tela escolha do idioma da interface e preferências de impressão

Conforme figura 54, ao escolher a opção “search”, é possível fazer consultas pelo texto indexado.

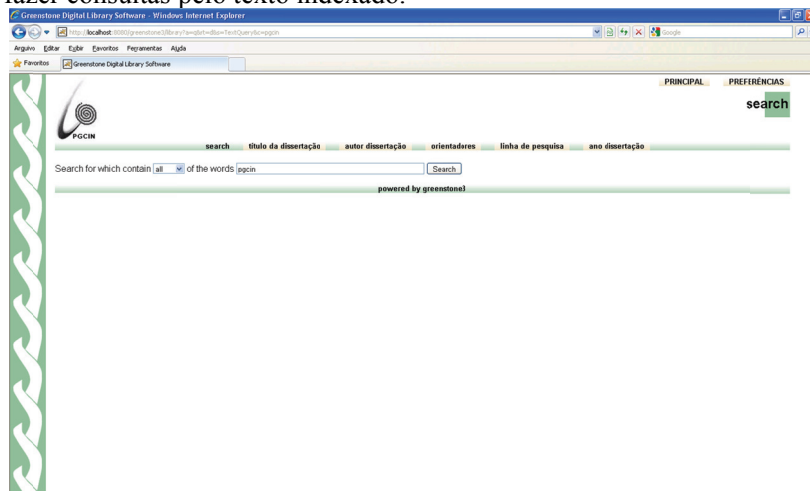


Figura 54: Tela busca pelo texto completo (search)

Conforme figura 55, ao escolher a opção “título da dissertação”, o Greenstone exhibe todos os documentos por ordem alfabética do título da dissertação.

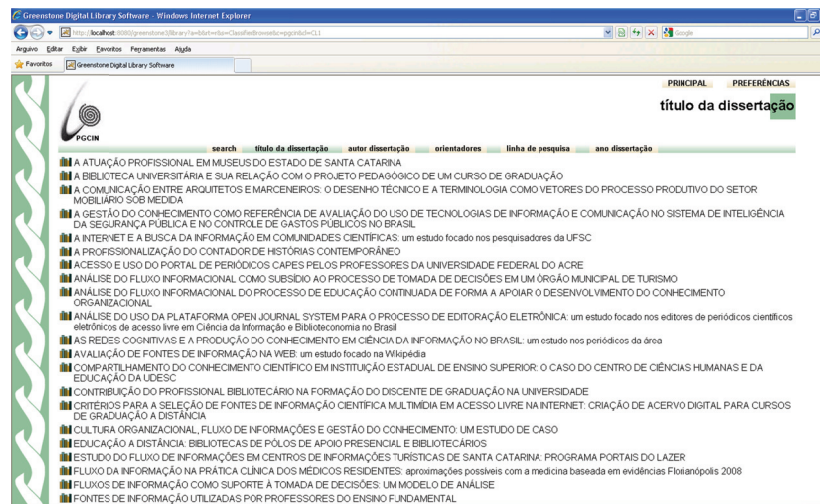


Figura 55: Tela dissertações ordenadas por título

Conforme figura 56, ao escolher a opção “Autor dissertação”, o Greenstone exhibe todos os documentos por ordem alfabética do autor da dissertação.

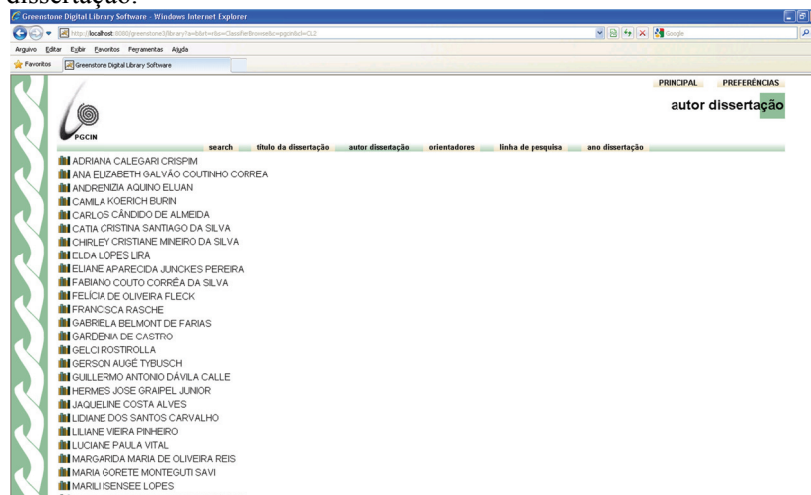


Figura 56: Tela dissertações ordenadas por autor

Conforme figura 57, ao escolher a opção “autor dissertação”, o Greenstone exhibe todos os documentos por ordem alfabética de um determinado autor. Para ver o documento, basta clicar ícone do modo texto ou no ícone modo PDF.

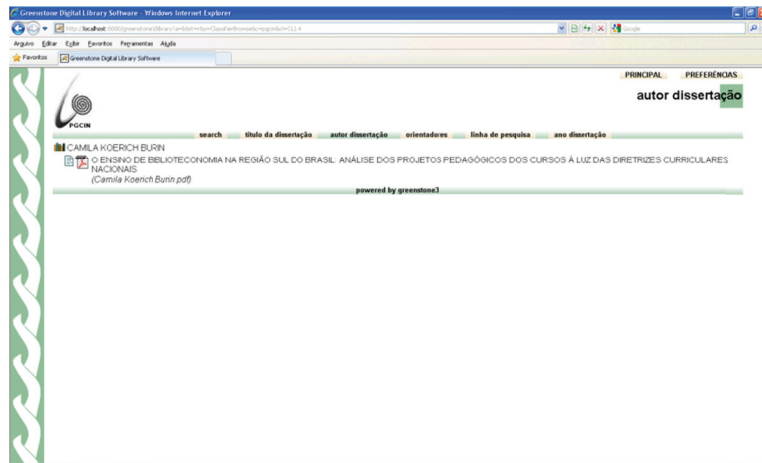


Figura 57: Tela dissertações do aluno

Conforme figura 58, ao escolher a opção “ano dissertação”, o Greenstone exhibe todos os documentos por ordem alfabética do ano de defesa dissertação.

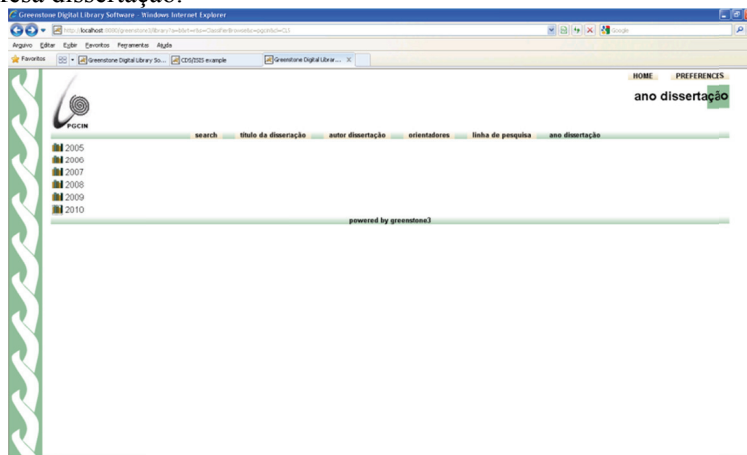


Figura 58: Tela dissertações ordenadas por ano defesa dissertação

Conforme figura 59, ao escolher a opção “ano dissertação”, o Greenstone exhibe todos os documentos por ordem alfabética do ano de defesa dissertação.

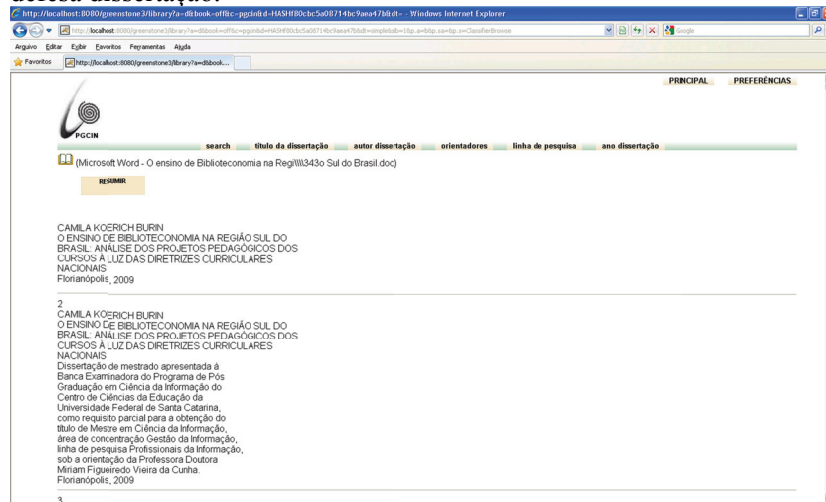


Figura 59: Tela visualização conteúdo modo texto

Conforme figura 60, ao escolher a opção “orientadores”, o Greenstone exhibe todos os documentos por ordem alfabética de orientador.

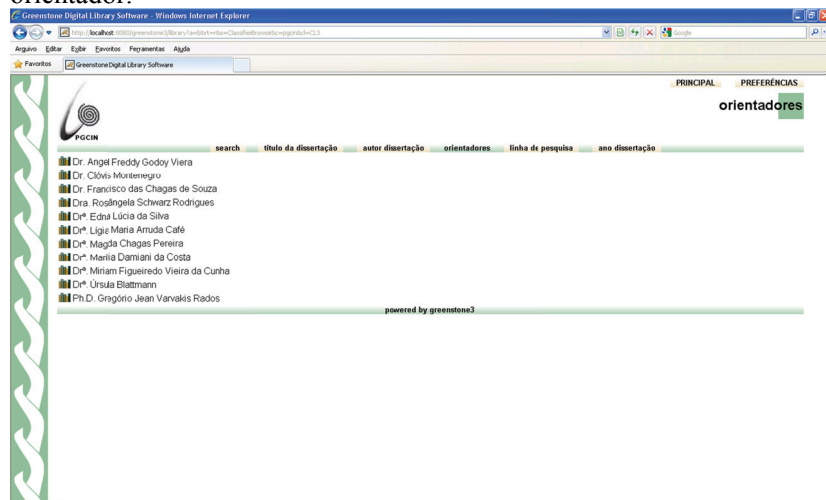


Figura 60: Tela orientadores por ordem alfabética

Conforme figura 61, ao escolher a opção “orientadores”, o Greenstone exhibe todas as dissertações ao qual o orientador está vinculado.

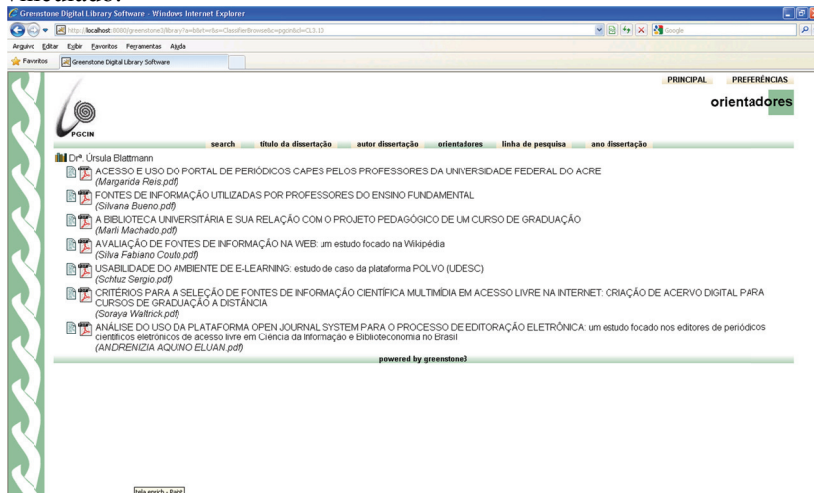


Figura 61: Tela de dissertações ao qual o orientador está vinculado.

Conforme figura 62, ao escolher a opção “linha de pesquisa”, o Greenstone exhibe todas as linhas de pesquisa.

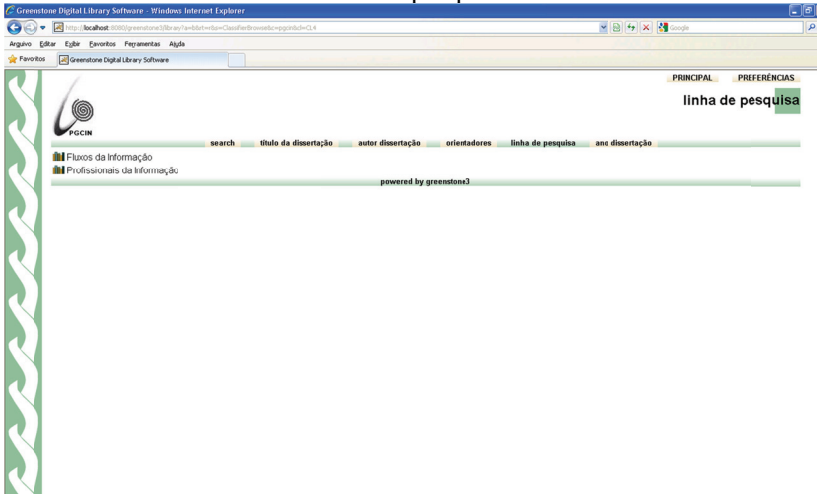


Figura 62: Tela de visualiza linhas de pesquisa.

5 ANÁLISE E INTERPRETAÇÃO DOS RESULTADOS

Nesta seção são apresentados os resultados obtidos acerca da pesquisa realizada; tais dados foram analisados e interpretados pelo pesquisador e serão descritos e representados por meio de quadros, tabelas e gráficos. Os resultados referem-se à análise da biblioteca digital Greenstone baseado na coleção construída para essa pesquisa, que é a de biblioteca de dissertações de mestrado do PGCIN UFSC.

5.1 Análise dos Formatos de Arquivos da Coleção PGCIN

Analisando a questão de preservação de documentos digitais sob o ponto de vista lógico na Biblioteca Digital Greenstone, constatou-se que um dos seus pontos fortes é a sua arquitetura, pois o mesmo foi projetado e implantado como uma plataforma aberta e possui a disposição de forma livre e gratuita, uma gama enorme de *plug-ins* (programas que servem normalmente para adicionar funções a outros programas maiores) para diversas funcionalidades, dentre elas os de inúmeros formatos de arquivos, além do que, é possível desenvolver *plug-ins* para o Greenstone para qualquer formato de arquivo.

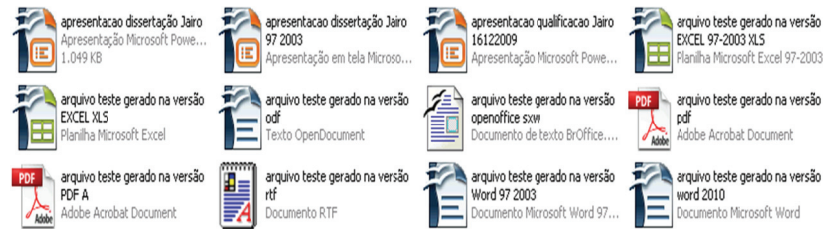
Neste trabalho de pesquisa foram selecionados os *plug-ins* para formatos de arquivo proprietários com especificação fechada, proprietário com especificação aberta, e não proprietários com especificação aberta.

Os *plug-ins* selecionados foram:

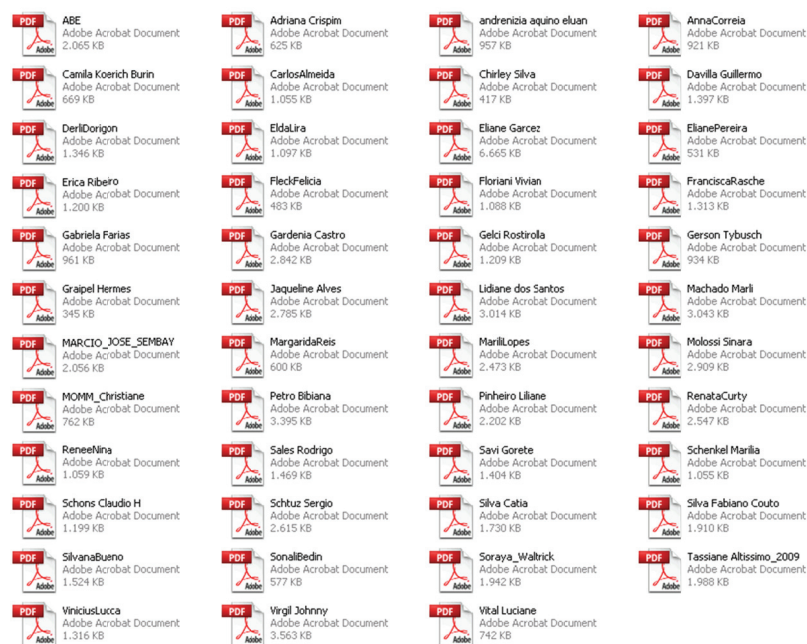
- a) *PDFPlug-in* – para documentos do tipo *PDF*
- b) *OpenDocumentPlug-in* – para documentos do tipo formado aberto
- c) *GreenstoneXMLPlug-in* – para documentos do tipo padrão XML
- d) *RTFPlug-in* – para documentos do tipo *RTF*
- e) *TextPlug-in* – para documentos do tipo texto
- f) *WordPlug-in* – para documentos do tipo *Microsoft Word*
- g) *PowerPoint plug-in* – para documentos do tipo *Microsoft Powerpoint*
- h) *ExcelPlug-in* – para documentos do tipo *Microsoft Excel*

Na coleta de dados, conforme quadro 8 foram realizados *downloads* de 47 dissertações de mestrado no formato PDF do sítio <http://www.cin.ufsc.br/pgcin>, e foram criados 12 documentos de vários tipos de formato de arquivo conforme quadro 7.

Os documentos que foram gerados com outros formatos de arquivos e inseridos na coleção PGCIN são os seguintes:



Quadro 7: Arquivos gerados para compor a coleção PGCIN



Quadro 8: Arquivos que foram importados para a coleção PGCIN no Greenstone

Dos cinquenta e nove documentos disponibilizados para a coleta, quarenta e quatro documentos foram importados para a coleção do PGCIN, sendo que o Greenstone não reconheceu o formato de arquivo de seis documentos e rejeitou outros nove documentos.

Os documentos processados corretamente e importados para a coleção do PGCIN foram os seguintes:

1. Adriana Crispim.pdf
2. ANDRENIZIA AQUINO ELUAN.pdf
3. ANNA ELIZABETH GALVÃO COUTINHO CORREIA.pdf
4. arquivo teste gerado na versão PDF A.pdf
5. arquivo teste gerado na versão odf.odt
6. arquivo teste gerado na versão pdf.pdf
7. arquivo teste gerado na versão office 97 2003.doc
8. Camila Koerich Burin.pdf
9. CARLOS CANDIDO DE ALMEIDA.pdf
10. CHIRLEY CRISTIANE MINEIRO DA SILVA.pdf
11. Elda Lira.pdf
12. Eliane Pereira.pdf
13. Erica Ribeiro.pdf
14. Felicia Fleck.pdf
15. Francisca Rasche.pdf
16. Gabriela Farias.pdf
17. Gardenia Castro.pdf
18. Gelci Rostirolla.pdf
19. Gerson Tybusch.pdf
20. Graipel Hermes.pdf
21. GUILLERMO ANTONIO DÁVILA CALLE.pdf
22. Jaqueline Alves.pdf
23. Lidiane dos Santos.pdf
24. MARCIO JOSE SEMBAY.pdf
25. Margarida Reis.pdf
26. Marili Lopes.pdf
27. Marli Machado.pdf
28. Petro Bibiana.pdf
29. Pinheiro Liliane.pdf
30. Renata Curty.pdf
31. Renee Nina.pdf
32. Sales Rodrigo.pdf
33. Savi Gorete.pdf
34. Schenkel Marilia.pdf

35. Shtuz Sergio.pdf
36. Silva Catia.pdf
37. Silva Fabiano Couto.pdf
38. Silvana Bueno.pdf
39. Sonali Bedin.pdf
40. Soraya Waltrick.pdf
41. Tassiane Altissimo.pdf
42. Veridiana Abe.pdf
43. Vinicius Lucca.pdf
44. Vital Luciane.pdf
45. Vivian Floriani.pdf

Arquivos não processados por nenhum plug-in disponível no Greenstone:

1. Sinara Molossi.pdf
2. Schons Claudio H.pdf
3. Johnny Virgil.pdf
4. Eliane Garcez.pdf
5. DERLI SANDRA DORIGON.pdf
6. apresentacao qualificação Jairo 16122009.ppt
7. arquivo teste gerado na versão EXCEL 97-2003 XLS.xls
8. CHRISTIANE FABIOLA MONN.pdf
9. arquivo teste gerado na versão rtf.rtf

Arquivos que não foram reconhecidos por nenhum plug-in:

1. arquivo teste gerado na versão word 2010.docx
2. arquivo teste gerado na versão openoffice sxw.sxw
3. apresentacao dissertação Jairo 97 2003.pps
4. apresentacao dissertação Jairo.pptx
5. arquivo teste gerado na versão EXCEL XLS.xlsx

Com o objetivo de auxiliar a análise do formato dos arquivos, foi utilizado o software *DROID – Digital Record Object Identification*, que é um software livre desenvolvido a partir do projeto denominado *PRONOM da National Archives* e que está disponível em <<http://www.nationalarchives.gov.uk/pronom>>.

Visando identificar o motivo pelo qual o Greenstone não reconheceu o formato de arquivo de cinco documentos, e rejeitou outros nove, foi refeito todo o processo de *download*, importação, adição de metadados *Dublin-core* e a realizado novamente o processo de criação da coleção. Após o processamento da criação da coleção, constatou-se que os resultados se repetiram. Concluiu-se que os cinco arquivos não foram processados por não existir *plug-ins* disponíveis. Já os nove

arquivos foram rejeitados, constatou-se que os mesmos estão protegidos por senha contra alteração, edição, impressão e cópia, mas considerando que alguns arquivos também protegidos foram processados, pode ser, portanto que o problema pode ocorrer na geração do arquivo, ou até mesmo que tenha alguma característica que gere rejeição pelos *plug-ins* do Greenstone.

5.1.1 Análise dos formatos de arquivos com especificações proprietárias e fechadas no Greenstone

Para realizar a análise dos formatos de arquivos proprietários e fechados, foram escolhidos os *plug-ins* para o *Microsoft Word* (software para editoração eletrônica de textos), *Microsoft Excel* (planilha eletrônica) e *Microsoft Powerpoint* (software para apresentação).

Os *plug-ins* utilizados:

| | |
|-------------------|--|
| WordPlug-in | Plug-in que as importações de documentos do Microsoft Word. Herda ConvertBinaryFile. |
| ExcelPlug-in | Plug-in que importa arquivos do Microsoft Excel. Herda ConvertBinaryFile. |
| PowerPointPlug-in | Plug-in que importa arquivos do PowerPoint Microsoft. Herda ConvertBinaryFile. |

Quadro 9: Plug-ins de utilizados de nível superior

Para realização dessa pesquisa, foram gerados arquivos no *Microsoft Office (Word, Excel, Powerpoint)* nas versões 97-2003, e inseridos na coleção de dissertações do PGCIN da Biblioteca Digital Greenstone.

Após o processo de construção da coleção, a biblioteca digital Greenstone apresentou as seguintes informações referentes aos arquivos cujo formato é proprietário e fechado:

- 1) Arquivo: “arquivo “apresentacao dissertação Jairo 97 2003.pps” – Este arquivo na importação não foi reconhecido por nenhum plug-in.
- 2) Arquivo: “apresentacao qualificacao Jairo 16122009.ppt” – Este arquivo na importação não foi reconhecido ou não pode ser processado por nenhum plug-in.

- 3) Arquivo: “arquivo teste gerado na versão EXCEL 97-2003 XLS.xls” – Este arquivo na importação não foi reconhecido ou não pode ser processado por nenhum plug-in.

Observa-se que os formatos de arquivo padrão Microsoft não foram importados para a coleção do PGCIN no Greenstone, porque o plug-in instalado no Greenstone não reconheceu os formatos de arquivos e quando reconheceu, não conseguiu processar os documentos.

Considerando que existiam plug-ins específicos para importação de arquivos com formatos proprietários, e mesmo repetindo todo o procedimento de geração e de importação de arquivos, os resultados foram os mesmos.

Analisando o relatório gerado pelo software *DROID* sobre os formatos de arquivos selecionados para serem inseridos na coleção, conforme Anexos A, B, C, D, E e F, observa-se que os mesmos estão dentro das normas previstas para os respectivos formatos. Além do que, os respectivos arquivos podem ser abertos nos softwares os quais foram gerados, bem como, outros similares como o BRoffice.

5.1.2 Análise dos formatos de arquivos com especificações proprietárias e abertas no Greenstone

Para realizar a análise dos formatos de arquivos proprietários e abertos, foram escolhidos os *plug-ins* para o *Microsoft Word* (software para editoração eletrônica de textos), *Microsoft Excel* (planilha eletrônica), *Microsoft Powerpoint* (software para apresentação) e *adobe PDF*.

Os *plug-ins* utilizados foram:

| | |
|-------------------|--|
| WordPlug-in | Plug-in que as importações de documentos do Microsoft Word. Herda ConvertBinaryFile. |
| ExcelPlug-in | Plug-in que importa arquivos do Microsoft Excel. Herda ConvertBinaryFile. |
| PowerPointPlug-in | Plug-in que importa arquivos do PowerPoint Microsoft. Herda ConvertBinaryFile. |
| PDFPlug-in | Plug-in que importa arquivos PDF. Herda ConvertBinaryFile. |
| RTFPlug-in | Plug-in que importa arquivos RTF. Herda ConvertBinaryFile. |

Quadro 10: Plug-ins de utilizados de nível superior

Para a realização dessa pesquisa, foram gerados arquivos no *Microsoft Office (Word, Excel, Powerpoint)* nas versões 2007 e 2010 usando formato *OpenXML* e o formato PDF no *Adobe Professional*, e inseridos na coleção de dissertações do PGCIN da Biblioteca Digital Greenstone.

Após o processo de construção da coleção, a Biblioteca Digital Greenstone processou todos os arquivos, e apresentou as seguintes restrições referentes aos arquivos cujo formato é proprietário e aberto:

- 1) Arquivo: “apresentacao dissertação Jairo.pptx” – Este arquivo na importação não foi reconhecido por nenhum *plug-in*.
- 2) Arquivo: “arquivo teste gerado na versão EXCEL XLS.xlsx” – Este arquivo na importação não foi reconhecido por nenhum *plug-in*.
- 3) Arquivo: “arquivo teste gerado na versão word 2010.docx” – Este arquivo na importação não foi reconhecido por nenhum *plug-in*.

Observa-se que os formatos de arquivo padrão Microsoft não foram importados para a coleção do PGCIN no Greenstone, posto que, o *plug-in* não reconheceu alguns dos formatos de arquivos. Considerando que existiam *plug-ins* específicos para importação de arquivos com formatos proprietários, e mesmo repetindo todo o procedimento de geração e de importação de arquivos, os resultados foram os mesmos.

Analisando relatório gerado pelo software *DROID* sobre os formatos de arquivos selecionados para serem inseridos na coleção, conforme Anexos A, B, C, D, E e F, observa-se que os mesmos estão dentro das normas previstas para os respectivos formatos. Além do que, os respectivos arquivos podem ser abertos nos softwares os quais foram gerados, bem como, outros similares como o *BRoffice*.

Constatou-se que o *plug-in* disponível é somente para arquivos gerados pelo *Microsoft Office* com versões anteriores a 2007. Ainda não está disponível o *plug-in* para o formato *OpenXML da Microsoft*.

5.1.3 Análise dos formatos de arquivos com especificação não-proprietária e aberta no Greenstone

Para realizar a análise dos formatos de arquivos proprietários e abertos, foram escolhidos os *plug-ins* para o *Microsoft Word* (software para editoração eletrônica de textos), *Microsoft Excel* (planilha

eletrônica), *Microsoft Powerpoint* (software para apresentação) e *adobe PDF*.

Os *plug-ins* utilizados foram:

| | |
|----------------------------|--|
| <i>HTMLPlug-in</i> | Plug-in que importa arquivos HTML. Herda <i>ReadTextFile</i> , <i>HBPlug-in</i> . |
| <i>OpenDocumentPlug-in</i> | Plug-in para importações <i>OASIS</i> documentos de formato <i>OpenDocument</i> (usado pelo OpenOffice 2.0. Herda <i>ReadXMLFile</i> . |
| <i>PDFPlug-in</i> | Plug-in que importa arquivos PDF. Herda <i>ConvertBinaryFile</i> . |

Quadro 11: Plug-ins de utilizados de nível superior

Para realização dessa pesquisa, foram gerados arquivos no BOffice usando formato Open Document Format – ODF e Adobe Professional usando formato *PDF-A*, e inseridos na coleção de dissertações do PGCIN da Biblioteca Digital Greenstone.

Após o processo de construção da coleção, a Biblioteca Digital Greenstone reconheceu e processou os arquivos cujos formatos são do tipo não proprietários e abertos.

A Biblioteca Digital Greenstone já disponibiliza inúmeros plug-ins de formato de arquivos, e que a mesma permite adicionar outros plug-ins de forma aberta, atendendo assim aos requisitos de preservação digital do formato lógico de arquivos.

5.2 Análise da Recuperação da Informação no Greenstone

Analisando os recursos disponíveis para recuperação de informação na Biblioteca Digital Greenstone, constatou-se que a mesma permite a recuperação de informação combinando pesquisas em textos completos, pesquisa através de navegação hierárquica, e também através de índices baseados em diferentes tipos de metadados com os padrões *Dublin Core*, *RCF 1807*, *NZGLS (New Zealand Government Locator Service)*, e *AGLS (Australian Government Locator Service)*.

A Biblioteca Digital Greenstone realiza a indexação dos metadados, assim como do próprio conteúdo dos documentos, por meio da extração de palavras do texto. Após o tratamento dos documentos,

estes são convertidos para um formato compatível com o XML. Essa funcionalidade possibilita a criação dos índices extraídos dos textos e dos metadados, assim como a inserção de uma interface de navegação hipertextual, organizada por meio de estrutura hierárquica, permitindo assim ao usuário de realizar buscas como em um mecanismo de busca convencional, inclusive utilizando operadores booleanos, bem como, explorar o documento por meio da navegação utilizando os links inseridos e organizados hierarquicamente.

Para realização dessa pesquisa exploratória, foi instalada a Biblioteca Digital Greenstone, e também foi construída uma coleção chamada PGCIN.

Ao entrarmos no endereço onde está instalada a Biblioteca Digital Greenstone, as coleções cadastradas são listadas conforme figura 64. Neste caso a coleção escolhida foi a do PGCIN mostrada conforme figura 64.

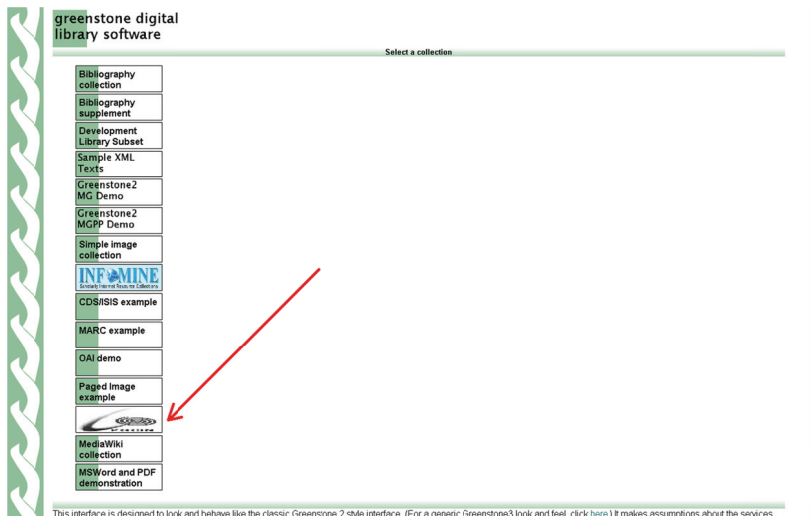


Figura 64: Tela principal da Biblioteca Digital Greenstone

Ao selecionar a coleção do PGCIN, a tela principal da mesma é mostrada conforme figura 65.

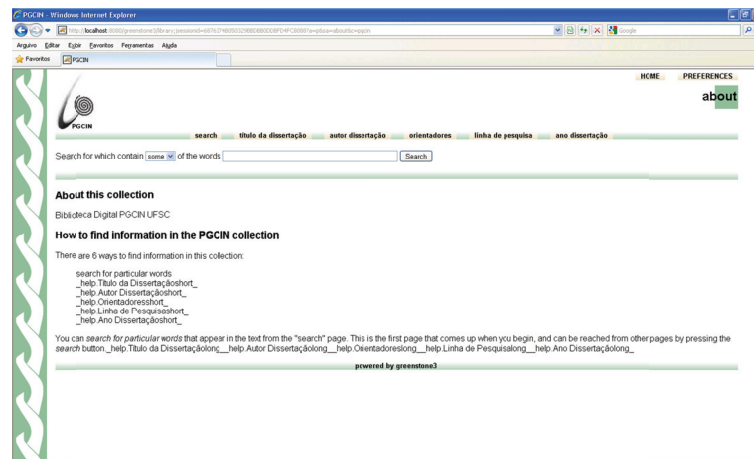


Figura 65: Tela inicial da Coleção PGCIN

Há seis maneiras de encontrar informações nesta coleção:

- 1) Busca por palavras contidas no texto
- 2) Título da Dissertação
- 3) Autor da Dissertação
- 4) Orientadores
- 5) Linha de Pesquisa
- 6) Ano da Dissertação

O Greenstone ainda permite definir as preferências de pesquisa e preferências de apresentação. Nas preferências de apresentação (linguagem da interface), é possível escolher um entre os vários idiomas disponíveis. Nas preferências de busca, conforme figuras 66 e 67 estão disponíveis as seguintes opções:

- 1) Tipo de pesquisa (*type of Search*) – Formulário Simples (*simple form*) ou avançado (*advanced form*);
- 2) Modo de consulta (*query mode*) – Modo de consulta Simples (*simple query mode*) ou modo de consulta avançada que permite pesquisas booleanas usando!, &, | e parênteses (*advanced query mode - allows boolean searching using !, &, | and parentheses*);
- 3) Tamanho da caixa de consulta (*query box size*) – caixa de consulta regular (*regular query box*) ou caixa de consulta grande (*large query box*);
- 4) Diferenciar letras maiúsculas e minúsculas (*Turn casefolding*);

- 5) Utilizar o *stemming* (*Turn stemming*) – Seleção por parte da palavra;
- 6) Diferenciar palavras acentuadas (*Turn accentfolding*);
- 7) Ordem que os documentos serão mostrados (*Document display order*);
- 8) Número de documentos que serão recuperados (*Return up to hits*); e,
- 9) Número de documentos que serão recuperados por página (*hits per page*).

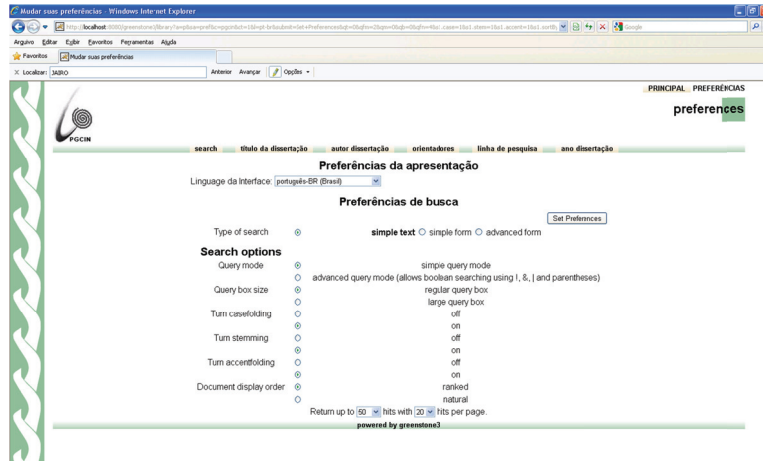


Figura 66: Tela de preferências de apresentação e de busca



Figura 67: Tela preferências de pesquisa

Ao selecionarmos a opção “*search*” conforme figura 68, o Greenstone permite realizar consultas a partir de palavras contidas nos documentos:

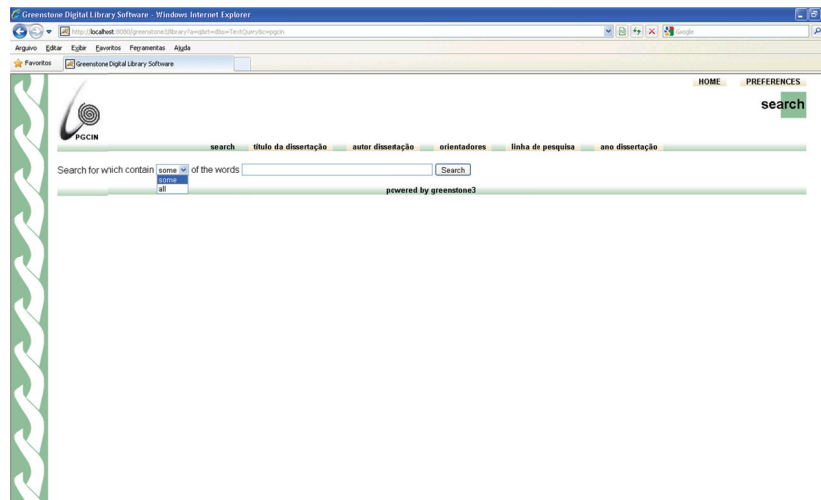


Figura 68: Recuperação de Informação a partir de palavras

Nesta pesquisa foram testadas os três tipos de indexadores disponíveis na Biblioteca Digital Greenstone, o *MG*, o *MGPP* ou *MG++* e o *Lucene*. A troca de indexador é realizada quando é apertado o botão “*change*” conforme figura 69.

O *MG* não oferece de recursos de diferenciação de acentuação, e o *LUCENE* não dispõe de recursos de *stemming*. Dos três indexadores o *MGPP* mostrou-se mais eficiente, e com mais opções para recuperação de informação. Nessa coleção foi utilizado o *MGPP* com opção de indexação *Stem*, *Casefold* e *accent fold* com nível de indexação por documento.

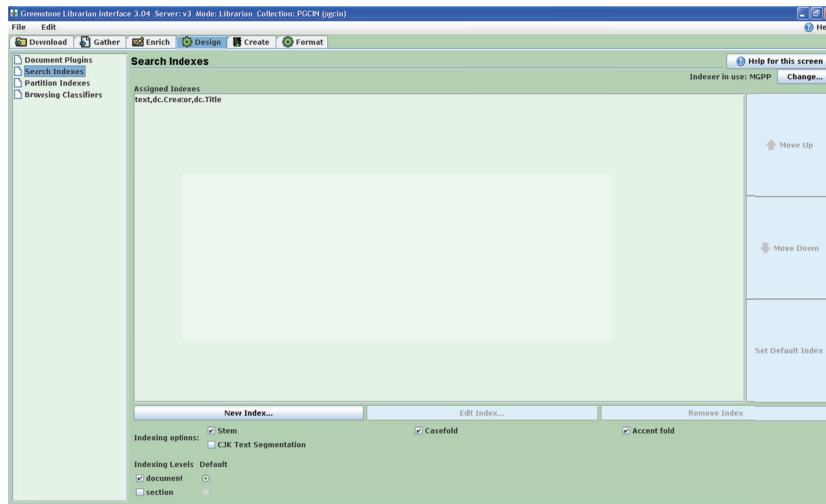


Figura 69: Indexadores do Greenstone

O Greenstone além de indexar o texto integral, também permite anexar metadados na indexação. No pacote de instalação da versão 3.04 estão disponíveis os seguintes padrões de metadados:

- 1) *Australian Government Locator Service Metadata Element Set, Version 1.3 (agls);*
- 2) *Development Library Subset Example Metadata (dls);*
- 3) *Extracted Greenstone Metadata 1.1;*
- 4) *Explode Metadata Set;*
- 5) *Greenstone metadata set (gs);*
- 6) *New Zealand Government Locator Service Metadata Standard version 2.1 (nzgls); e,*
- 7) *Qualified Dublin Core Metadata Element Set, Version 1.1: Reference Description (dc).*
- 8) *RFC 1807 Metadata Element Set, Version TR-v2.1 (rfc1807)*

Nesta coleção de teste denominada PGCIN foram utilizado os metadados padrão *Qualified Dublin Core Metadata Element Set, Version 1.1: Reference Description (dc)* e os metadados padrão do *Extracted Greenstone Metadata 1.1* conforme figura 70.

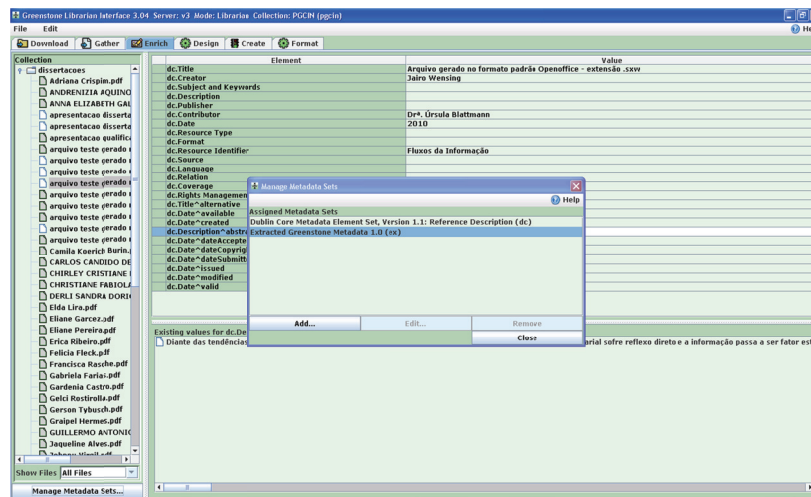


Figura 70: Tela metadados Greenstone

O Greenstone possui um editor de metadados, o qual permite incluir outros padrões ou alterar os existentes conforme figura 71. Nesta coleção não foram alterados os padrões de metadados existentes.

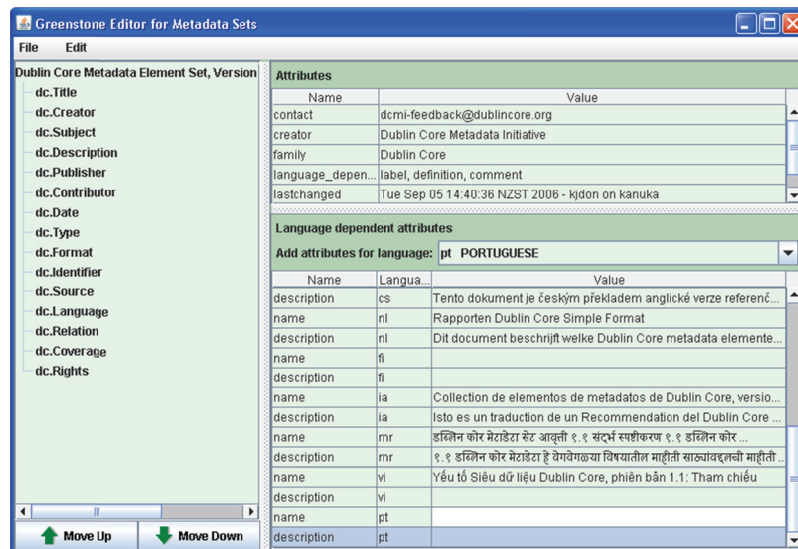


Figura 71: Tela editor de metadados

No processo de construção de uma coleção, o Greenstone obtém dos dados automaticamente das propriedades do documento no seu formato original, como por exemplo, as propriedades de um documento *Microsoft Word* e do *PDF*.

No Greenstone, é possível recuperar informações em diversos idiomas a partir dos metadados vinculados ao documento, desde que os mesmos estejam devidamente configurados e preenchidos.

Nesta coleção também é possível recuperar informações através da navegação por palavras chaves que estão organizadas hierarquicamente. A recuperação da informação pode ser feita da seguinte forma:

a) **Título da Dissertação** – Recupera todos os documentos ordenados por ordem alfabética conforme figura 72.

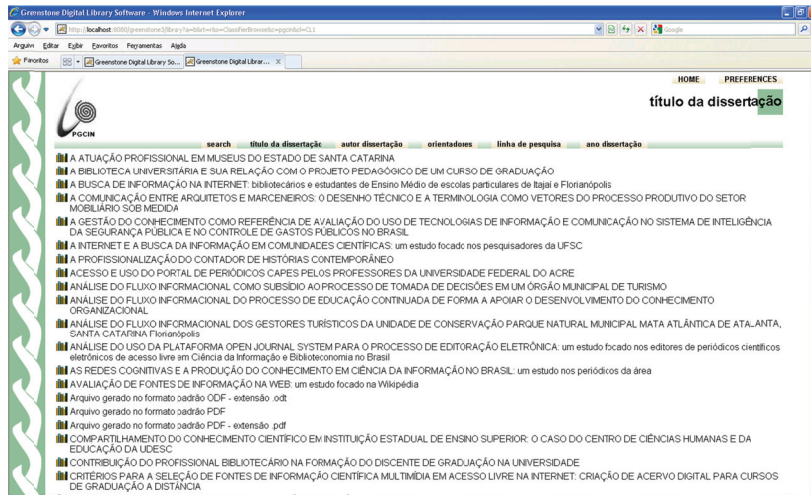


Figura 72: Tela dissertações por ordem alfabética de título

A selecionarmos o Título, o Greenstone conforme figura 73, mostra o documento selecionado que poder ser visualizado no formato HTML ou através do *software* que o mesmo foi gerado ou outro compatível com aquele formato de arquivo.

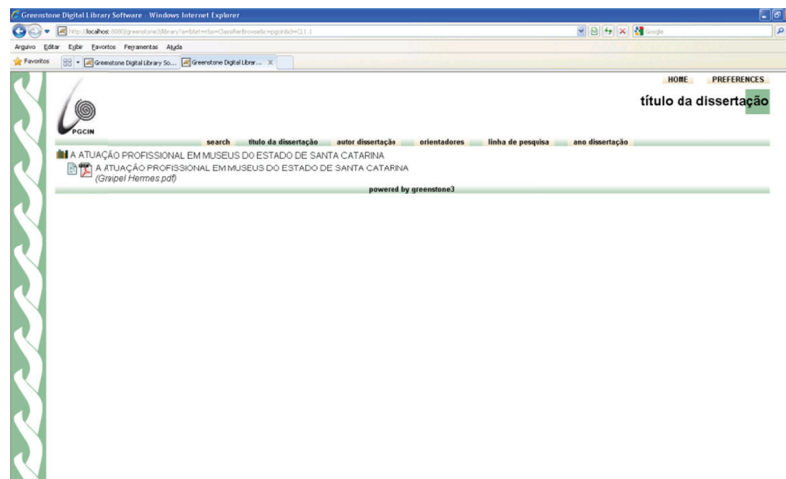


Figura 73: Tela seleção de visualização dissertação por título

b) **Autor da Dissertação** – Recupera todos os documentos conforme figura 74 e mostra-os em ordem alfabética classificada por autor da dissertação.

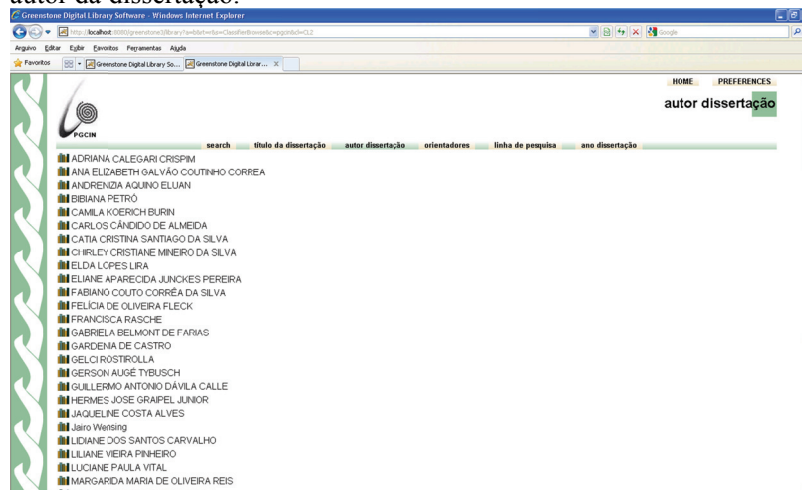


Figura 74: Tela dissertações por ordem alfabética de autor

A selecionarmos o “Autor”, o Greenstone conforme figura 75, mostra o documento selecionado que poder ser visualizado no formato HTML ou através do *software* que o mesmo foi gerado ou outro compatível com aquele formato de arquivo.

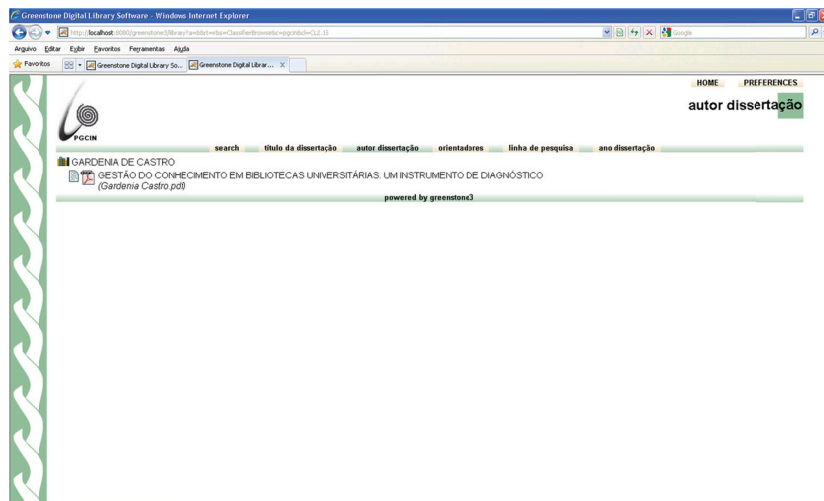


Figura 75: Tela seleção de visualização dissertação por autor

c) **Orientadores** - Recupera todos os documentos e mostra-os em ordem alfabética classificada por orientador conforme figura 76.

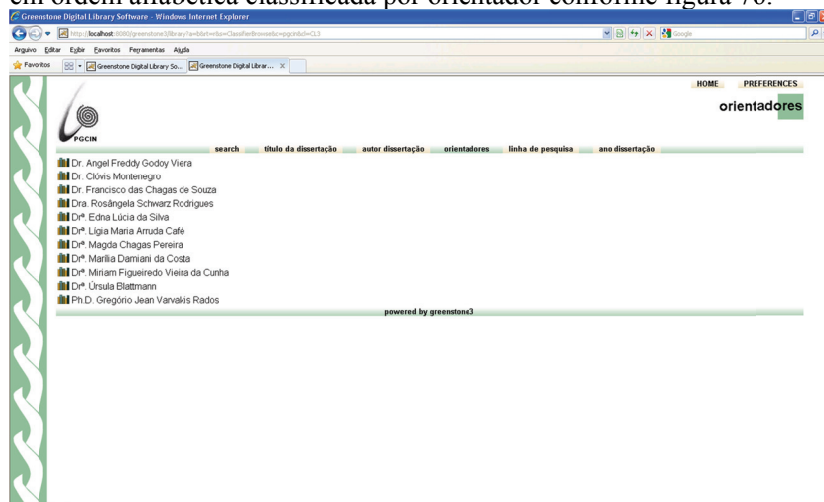


Figura 76: Tela dissertações por ordem alfabética de orientador

Ao selecionar o orientador o Greenstone exibe as dissertação que o Professor selecionado realizou a orientação de Mestrado conforme figura 77.

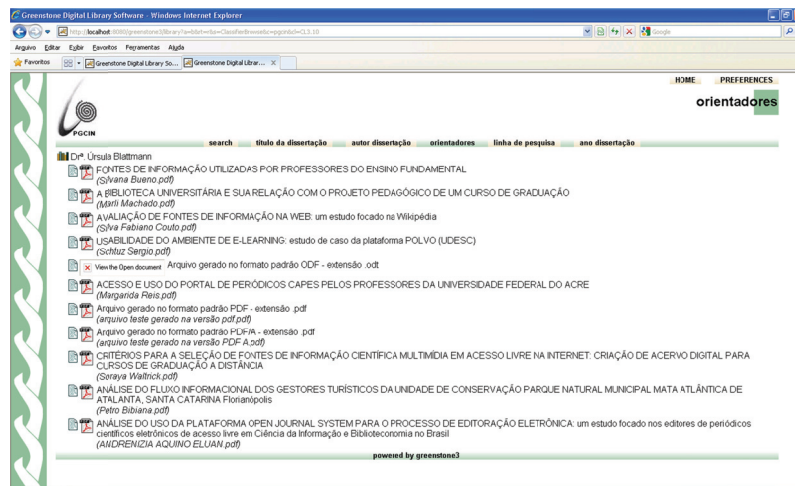


Figura 77: Tela seleção de visualização dissertação por orientador

d) Linha de Pesquisa - Recupera todos os documentos e mostra-os em ordem alfabética classificada por Linha de Pesquisa conforme figura 78.

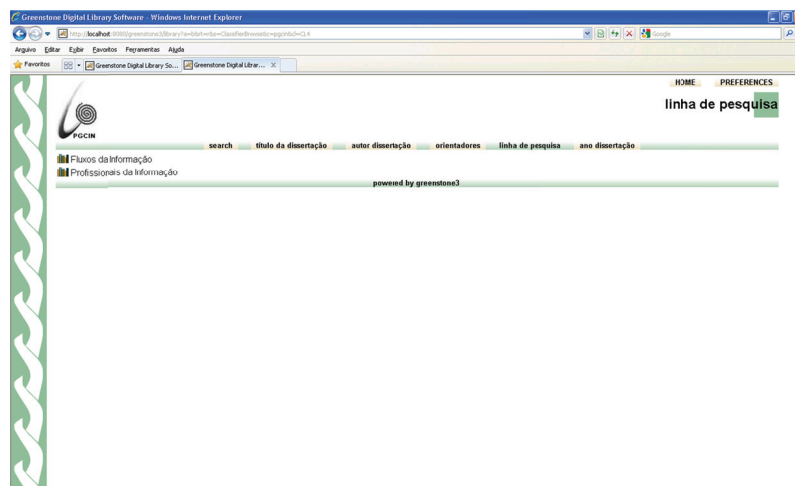


Figura 78: Tela dissertações por ordem alfabética por linha de pesquisa

Ao selecionar a linha de pesquisa, o Greentone exibe as dissertações da linha de pesquisa selecionada em ordem alfabética de título de dissertação conforme figura 79.

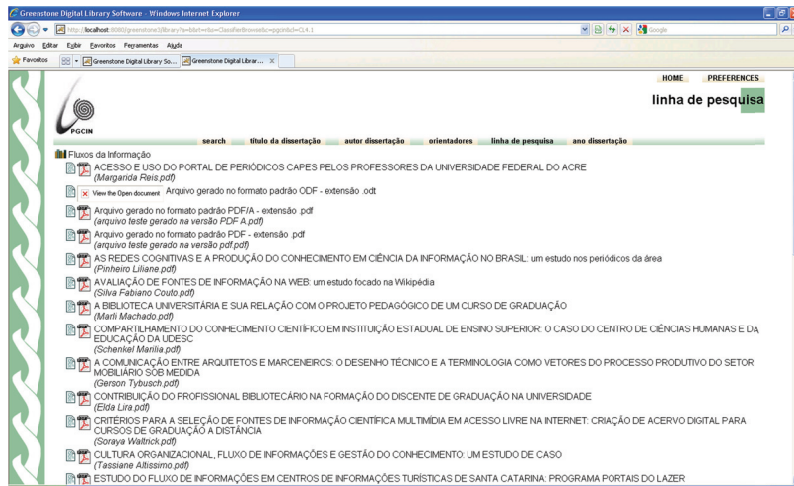


Figura 79: Tela seleção de visualização dissertação por linha pesquisa

e) **Ano Dissertação** - Recupera todos os documentos e mostra-os em ordem alfabética classificada por Ano de Dissertação conforme figura 80.

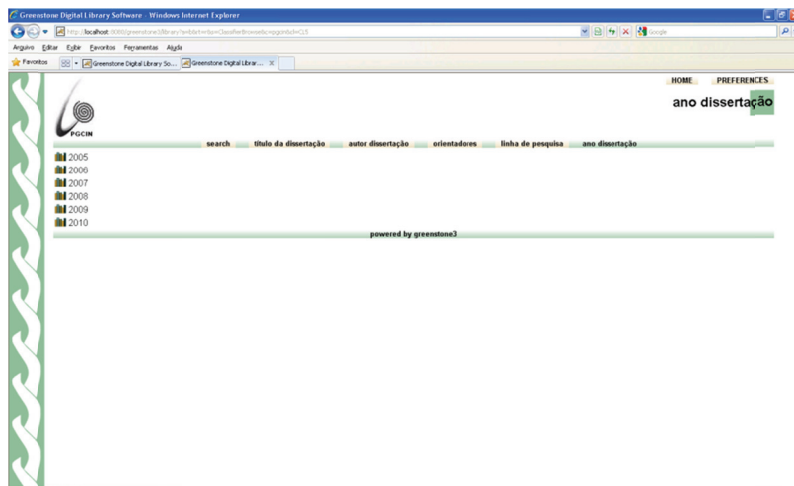


Figura 80: Tela dissertações por ordem ano pesquisa

Ao selecionar o Ano da Dissertação o Greenstone exibe as dissertações defendidas no ano selecionado em ordem alfabética por título conforme figura 81.

6 CONCLUSÕES

Neste capítulo serão apresentadas as conclusões do estudo, bem como sugestões e recomendações para futuras investigações sobre o tema abordado.

6.1 Conclusões

Sob o ponto de vista de contribuição a Ciência da Informação mais especificamente às disciplinas de Fontes de Informação, Bibliotecas Digitais, Preservação e Recuperação da informação, a pesquisa aprofundou estudos na área de recuperação de informação e preservação digital tendo como foco o formato de arquivos digitais, beneficiando as áreas correlatas como da ciência da computação, Biblioteconomia e Arquivologia.

Esta pesquisa teve como objetivo principal de analisar os recursos disponíveis na Biblioteca Digital Greenstone para preservação lógica de documentos digitais com foco no formato de arquivos e a recuperação da informação.

Para atingir o objetivo principal, foram traçados os seguintes objetivos específicos:

- a) Estudar os modelos clássicos de recuperação de informação;
- b) Identificar os recursos disponíveis para recuperação de informação na BDG;
- c) Identificar os pontos fortes e pontos fracos da BDG; e
- d) Analisar a questão de preservação de documentos digitais sob o ponto de vista lógico na Biblioteca Digital Greenstone.

A partir dos objetivos supracitados chegaram-se as seguintes conclusões:

- 1) Após estudo dos modelos clássicos de recuperação de informação e avaliação da Biblioteca Digital Greenstone, constatou-se que a mesma disponibiliza várias técnicas de recuperação de informação como *browsing*, *stemming*, pesquisa booleana e *ranking*.

2) Sobre as ferramentas de indexação dos documentos para recuperação da informação, constatou-se que dentro da BDG estão disponíveis três ferramentas para indexar as coleções: o MG, que é o indexador padrão, o *MGPP (MG++)* e o *Lucene (Apache Software Foundation)* que possuem características sofisticadas de indexação e busca. Nesta pesquisa foram testados os três modelos de indexação, onde constatou-se que o indexador MG não oferece recursos de diferenciação de acentuação, e o *LUCENE* não dispõe de recursos de *stemming*. Pela análise realizada concluiu-se que o *MGPP (ou MG++)* é o que tem mais parâmetros para recuperação de informação e que demonstrou ser mais eficiente, pois ele prevê índices de nível de documento, e compressão dos documentos originais. Constata-se que apesar dos recursos de recuperação de informação que BDG oferece, existem outros que poderiam ser disponibilizados, como a *Indexação Semântica Latente*, pois considerando que em bibliotecas digitais as coleções são mais estáveis, como no caso de uso dessa pesquisa, e com a tendência de aumento da capacidade de processamento dos computadores, a *LSI* passou a ser viável, já que o custo computacional dispendido em cálculos pela utilização do modelo algébrico *SVD* manipulando matrizes esparsas deixa de ser um fator limitante, e o problema de escalabilidade passa a não ser preocupante.

3) Segundo a pesquisa realizada, foram identificados os seguintes pontos fortes:

- a) Ser um *software* livre;
- b) Interoperabilidade de documentos;
- c) Ser altamente customizável;
- d) Instalação rápida;
- e) Disponível para vários idiomas;
- f) Disponível para várias em diversas plataformas;
- g) Interface de consulta Web;
- h) Permitir inclusão de mais de um formato de arquivo;
- i) Possibilidade de desenvolvimento de plug-ins para diversos formatos de arquivos;
- j) Permite instalação e execução em várias plataformas;
- k) Criação de coleções distintas;
- l) Inclusão de metadados obedecendo a padrões internacionais;
- m) Importação e exportação de obras/metadados;

- n) Quantidade satisfatória de documentação;
- o) Não necessitar de plug-ins na máquina do usuário final;
- p) Interface agradável e personalizável;
- q) Possibilidades de navegar pelos os campos relacionados ao documento;
- r) Realiza pesquisa utilizando campos específicos de metadados e texto-integral;
- s) Interoperabilidade automática com outros sistemas (mesma base ou bases diferentes);
- t) Exportar coleções ou partes de coleções para consulta local.

Também foram identificados os seguintes pontos fracos:

- a) Não permite a submissão de um documento pelo próprio autor, pois necessita de um sempre do intermediário administrador do sistema;
- b) Não estrutura de segurança baseada em perfis de usuários;
- c) Não possui módulo que permita a certificação digital;
- d) Não possui sistema de log indicando que fez o que no sistema;
- e) Interação com usuário (notificações por e-mail, informações na home-page);
- f) Não possui rotina de backup;
- g) Não possui estatísticas de utilização;
- h) Atualização do acervo on-line complicada,
- i) Dificil parametrização; e,
- j) Dificil customização da interface

4) Ao investigar se a Biblioteca Digital Greenstone aceita formatos de arquivos que atendem aos requisitos de preservação lógica de arquivos digitais, constatou-se que atualmente a mesma disponibiliza plug-ins para os formatos de arquivos aderentes aos padrões de preservação digital que são aceitos internacionalmente como *ODF* e *PDF/A*, e ainda permite o desenvolvimento e inserção de plug-ins para os inúmeros formatos de arquivos existentes.

6.2 Sugestões

Devido à amplitude do tema referente à preservação digital e recuperação de informação, a área de pesquisa foi limitada a recuperação de informação e a preservação digital com foco no formato de arquivos. No decorrer da pesquisa observou-se que algumas questões ficaram em aberto em decorrência das limitações impostas no escopo deste trabalho. Porém, este estudo pode ser continuado por meio de outras pesquisas, sejam de mestrado ou doutorado. As sugestões de estudos sobre o Greenstone e a seguinte:

- a) Preservação física – com foco na preservação das mídias e na sua renovação quando se fizer necessário;
- b) Preservação lógica – com foco nos formatos e a dependência de *hardware* e *software* que mantenham legíveis e interpretáveis a cadeia de bits;
- c) Preservação intelectual – com foco no conteúdo intelectual e sua autenticidade e integridade;
- d) Preservação do aparato – com foco nos metadados - necessária para localizar, recuperar e representar a informação digital;
- e) Avaliação sobre Ergonomia e usabilidade;
- f) Técnicas de recuperação de informação; e,
- g) Gestão arquivística de bibliotecas digitais.

6.3 Recomendações

Com a conclusão dessa pesquisa, observou-se que algumas questões envolvendo a Biblioteca Digital Greenstone, como preservação e recuperação de informações, formato de arquivos, podem ser recomendadas aos envolvidos nesta pesquisa:

1 – No Brasil não foi encontrado nenhuma norma obrigando a adoção do *PDF/A*, mas como demonstrado na revisão da literatura, observa-se que existe um movimento crescente em diversos países que estão normatizando o *PDF/A* como padrão de arquivamento de documentos digitais. Posto isto, faz-se necessário recomendar a administração da UFSC uma normatização sobre arquivamento de documentos, onde o padrão de formato de arquivos digitais será o '*PDF/A*'.

2 – Recomendar a Coordenação do PGCIN para que realize as seguintes alterações do sítio na internet que está disponível em <<http://www.cin.ufsc.br/pgcin/dissertacao.php>>:

a) Uma (1) cópia impressa e uma (1) cópia em *CD-ROM* no formato '*PDF/A*' deverão ser entregues na secretaria do PGCIN.

b) Uma (1) cópia impressa e uma (1) cópia em *CD-ROM* no formato '*PDF/A*' o aluno deverá entregar na Biblioteca Universitária, juntamente com o "termo de autorização para publicação eletrônica de Dissertações e Teses".

3 – Recomendar que a Coordenação do PGCIN entre em contato com os alunos para que os mesmos enviem um *CD-ROM* no formato '*PDF/A*', e que os mesmos sejam atualizados no sítio na internet do PGCIN. Tal recomendação se faz necessária pelo fato de que durante a coleta de dados observou-se que os documentos do PGCIN não estão em um formato adequado para a preservação digital, pois somente o arquivo "sales rodrigo.pdf" estava no formato *PDF/A-1b* atendendo assim aos requisitos de preservação digital. Já o documento "NelmaAraujo.pdf" está bloqueado por senha até para visualização no próprio site do PGCIN, sendo assim, não atendendo aos requisitos de preservação digital, e também não permite a sua consulta. O restante das dissertações selecionadas para compor a coleção PGCIN estão no formato PDF, mas em versões que não atendem aos requisitos de preservação digital.

4 - Recomendar a administração da UFSC dar conhecimento, ou até criar normas sobre os padrões de Interoperabilidade de Governo Eletrônico tendo como base o e-PING 4.0, posto que a partir dessa versão do e-PING, o *ODF* assumiu característica de adotado, tornando-se obrigatório para guarda e troca de documentos eletrônicos entre todos os órgãos da administração direta, autarquias e fundações, sendo assim, a Universidade Federal de Santa Catarina e por ser extensão o PGCIN deverão se enquadrar a essa regra.

5 – Recomendar aos responsáveis pelo desenvolvimento do Greenstone que considerem a utilização da técnica de indexação semântica latente aplicada à recuperação de informação pelas vantagens relacionadas aos problemas de sinonímia e à polissemia, pois a indexação semântica latente tem como objetivo de melhorar a recuperação de informação através do descobrimento de associações entre os termos em uma grande coleção de textos a fim de criar um espaço semântico.

Após a realização da pesquisa, concluiu-se que a Biblioteca Digital *Greenstone* está preparada para atender aos requisitos de preservação lógica de arquivos digitais, pois permite a inclusão de arquivos de formatos proprietários com especificação fechada, proprietário com especificação aberta, e não proprietários com especificação aberta. Além disso, é uma plataforma aberta e permite o desenvolvimento de plug-ins para inúmeros formatos de arquivos.

Como em qualquer processo de informatização, faz-se necessário antes da implantação de uma solução, o levantamento dos requisitos funcionais (o que se espera da solução) e não funcionais (recursos de infraestrutura) e considerar a perspectiva de continuidade da solução adotada. Posto isto, o *Greenstone* aparece como um forte candidato para implementação, pois possui inúmeros atributos importantes, como interoperabilidade, preservação digital lógica de arquivos digitais, recursos para recuperação de informação e ser um software livre, ainda conta com o apoio de comunidades de desenvolvimento em vários países, além do que, é um projeto de bibliotecas digitais (*New Zealand Digital Library Project* – www.nzdl.org) da Universidade de *Waikato* na Nova Zelândia, e desenvolvido e distribuído em cooperação com a UNESCO (www.unesco.org) e a *ONG Human Info* (<http://humaninfo.org/>).

REFERÊNCIAS BIBLIOGRÁFICAS

ADOBE SYSTEMS INCORPORATED. **XMP Adding Intelligent to Media**. San Jose, CA: Adobe, 2004. Disponível em: <www.aiim.org/documents/standards/xmpspecification.pdf>. Acesso em: 22 maio 2010.

ADOBE SYSTEMS INCORPORATED. **XMP Adding Intelligent to Media**. San Jose, CA: Adobe, 2005. Disponível em: <www.adobe.com/devnet/xmp/pdfs/xmp_specification.pdf>. Acesso em: 22 maio 2010.

ADOBE'S MAIN XMP. **Extensible Metadata Platform (XMP)**. Disponível em: <www.adobe.com/products/xmp/> Acesso em: 22 maio 2010.

ADOBE'S XMP DEVELOPER'S. **Adobe XMP Developer Center**. Disponível em: <partners.adobe.com/public/developer/xmp/topic.html> Acesso em: 25 nov. 2009.

ALEX WRIGHT. The Web Time Forgot. **The New York Times**. 17 Jun. 2008. Disponível em: <<http://www.nytimes.com/2008/06/17/science/17mund.html>>. Acesso em: 21 abr. 2009.

ARQUIVO NACIONAL. **Conselho Nacional de Arquivos (CONARQ)**. Carta para a preservação do patrimônio arquivístico digital. Rio de Janeiro, 2004. Disponível em: <<http://www.conarq.arquivonacional.gov.br/Media/publicacoes/cartapre-servpatrimarqdigitalconarq2004.pdf>>. Acesso em: 25 nov. 2009.

ASTI VERA, A. **Metodologia da pesquisa científica**. Porto Alegre: Globo, 1978.

BAEZA-YATES, R. A.; RIBEIRO-NETO, B. A. **Modern Information Retrieval**. Addison Wesley, 1999.

BARDIN, L. **Análise de conteúdo**. Lisboa: Edições 70, 2004.

BARRETO, A. de A. Os agregados de informação: memórias, esquecimento e estoques de informação. **DataGramaZero: Revista de Ciência da Informação**, Rio de Janeiro, v.1, n.3, ago. 2000. Disponível em: <http://www.dgz.org.br/jun00/Art_01.htm>. Acesso em: 16 abr. 2009.

BARRETO Aldo Albuquerque. Os destinos da Ciência da Informação: entre o cristal e a chama. **DataGramaZero: Revista de Ciência da Informação**, Rio de Janeiro, n. 0, p.1-9, dez. 1999. Disponível em: <http://www.dgz.org.br/dez99/Art_03.htm>. Acesso em: 16 abr. 2009.

BERRY, MICHAEL W.; DUMAIS, SUSAN T.; O'BRIEN, G.W. **Using Linear Algebra for Intelligent Information Retrieval**: Technical Report UT-CS-94-270. Tennessee, Knoxville : Computer Science Department, University of Tennessee. Disponível em: <<http://www.cs.utk.edu/~library/TechReports/1994/ut-cs-94-270.ps.Z>>. Acesso em: 16 abr. 2009.

BLATTMANN, Ursula; BOMFÁ, Cláudia Regina Ziliotto. Gestão de conteúdos em bibliotecas digitais: acesso aberto de periódicos científicos eletrônicos. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, 2006. v. 2, n.1.p. 41-56, 2006. Disponível em:<<http://143.106.108.58/seer/ojs/ojs/viewarticle.php?id=16&layout=abstract>>. Acesso em: 03 dez. 2009.

BLATTMANN, Ursula, FACHIN, Gleisy R. B.; RADOS, Gregório J.V. Recuperar a informação eletrônica pela Internet. Revista da ACB: **Biblioteconomia em Santa Catarina**, Florianópolis, v.4, n.1, 1999. Disponível em: <<http://www.ced.ufsc.br/~ursula/papers/buscanet.html>>. Acesso em: 03 dez. 2009.

BLATTMANN, Ursula. **Modelo de gestão da informação digital online em bibliotecas acadêmicas na educação à distância**: biblioteca virtual. 2001. Tese (Doutorado em Engenharia de Produção) - Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis.

BLATTMANN, Ursula; FACHIN, Gleisy R. B.; RADOS, Gregório J.V. Bibliotecário na posição do arquiteto da informação em ambiente Web. In: SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS, 10. **Anais eletrônicos**. Florianópolis, 2000. Disponível em: <<http://www.ced.ufsc.br/~ursula/papers/arquinfo.html>>. Acesso em: 03 dez. 2009.

BLATTMANN, Ursula; FRAGOSO, Graça Maria (orgs). **O zapear a informação em bibliotecas e na Internet**. Belo Horizonte: Autêntica, 2003.

BORKO, H. Information science: what is it? **American Documentation**, Chicago, v.19, n.1, p.3-5, Jan. 1968.

BUCKLEY, Chris. **SMART System Overview**. Ithaca, New York: Cornell University, 1996. 50p. (Technical Report). Disponível em: <<http://portal.acm.org/citation.cfm?id=866085&dl=GUIDE&coll=GUIDE&CFID=96052997&CFTOKEN=69269791>>. Acesso em: 30 jun. 2009.

BRASIL. **E-Ping**: padrões de interoperabilidade. Documento de referência. Brasília: Comitê Executivo de Governo Eletrônico: 2010. Disponível em: <<http://www.governoeletronico.gov.br/acoes-e-projetos/e-ping-padroes-de-interoperabilidade>>. Acesso em: 22 maio 2010.

BUSH, Vannevar. As we may think. **The Atlantic Monthly**, Jul. 1945. Disponível em: <<http://www.ps.uni-sb.de/~duchier/pub/vbush/vbush-all.shtml>>. Acesso em: 16 abr. 2009.

CAMPELLO, Bernardete; CALDEIRA, Paulo da Terra (Org.). **Introdução às fontes de informação**. Belo Horizonte : Autêntica, 2005. 181p. [Coleção Ciência da Informação]

CAMPELLO, Bernardete Santos; CENDÓN, Beatriz Valadares; KREMER, Jeannette Marguerite (orgs.). **Fontes de informação para pesquisadores e profissionais**. Belo Horizonte : Ed. UFMG, 2003.

CERVO, Amado L.; BERVIAN, Pedro A.. **Metodologia Científica : para uso dos estudantes universitários**. 3.ed. São Paulo : McGraw-Hill do Brasil, 1983.

CHIZZOTTI, Antonio. **Pesquisa em ciências humanas e sociais**. São Paulo: Cortez, 1991.

CHOO, Chun Wei, **A Organização do Conhecimento**. Tradução Eliana Rocha. São Paulo: SENAC, 2003.

CONSEGI - **Congresso Internacional Software Livre e Governo Eletrônico**. Disponível em: <<http://www.consegi.gov.br/consegi-1/historico>> - Acesso em: 15 jul. 2010.

CONSELHO NACIONAL DE ARQUIVOS. Câmara Técnica de Documentos Eletrônicos. **Modelo de requisitos para sistemas informatizados de gestão arquivística de documentos: e-ARQ Brasil**. 2006. Versão 1. Disponível em: <<http://www.conarq.arquivonacional.gov.br/Media/publicacoes/earqbrasilv1.pdf>>. Acesso em: 01 jun. 2009.

CONTANDRIOPOULOS, André-Pierre *et al*, **Saber preparar uma pesquisa**. 3. ed. São Paulo - Rio de Janeiro: HUCITEC/Abrasco, 1999.

CÔRTE, Adelaide Ramos et al.. **Avaliação de softwares para bibliotecas e arquivos : uma visão do cenário nacional**. 2. ed. rev. e ampl. São Paulo: Polis, 2002. 221 p. ISBN 8572280138 (broch.)

CUNHA, Murilo Bastos da. **Para saber mais: fontes de informação em ciências e tecnologia**. Brasília : Briquet de Lemos / Livros, 2001. 168 p.

CUNHA, Murilo Bastos da; MCCARTHY, Cavan. Estado atual das bibliotecas digitais no Brasil. In: MARCONDES, Carlos H.; KURAMOTO, Hélio; TOUTAIN, Lídia Brandão; SAYÃO; Luís (orgs.). **Bibliotecas digitais: saberes e práticas**. Salvador/Brasília: UFBA/IBICT, 2005. p. 25- 53.

DAVENPORT, Thomas H. **Reengenharia de processos: como inovar a empresa através da tecnologia da informação**. Rio de Janeiro: Campus, 1994.

DAVENPORT, Thomas H. **Ecologia da informação: por que só a tecnologia não basta para o sucesso na era da informação**. São Paulo: Futura, 1998.

DEERWETER, S. et al. Indexing by Latent Semantic Analysis. **Journal of the American Society for Information Science**, v.41, n. 6, p.391-407, 1990. Disponível em: <<http://lsi.research.telcordia.com/lsi/papers/CHI88.ps>>. Acesso em: 30 jun. 2009.

DIGITAL LIBRARY FEDERATION - DLF. **A working definition of digital library**. 1998. Disponível em: <<http://www.diglib.org/about/dldefinition.htm>>. Acesso em: 12 maio 2010.

DUMAIS, ST; FURNAS, GW; LANDAUER, TK; DEERWESTER, S. Using latent semantic analysis to improve information retrieval. In: CONFERENCE ON HUMAN FACTORS IN COMPUTING, 1988, **Proceedings...** New York: ACM, 1988. p. 281-285. Disponível em: <<http://lsi.research.telcordia.com/lsi/papers/CHI88.ps>>. Acesso em: 30 jun. 2009.

EIN-DOR, Phillip; SEGEV, Eli. **Administração de sistemas de informação**. Rio de Janeiro: Campus, 1985.

FERNEDA, E. **Recuperação da informação**: análise sobre a contribuição da ciência da computação para a ciência da informação. São Paulo: USP, 2003. 147p. Tese (Ciências da Comunicação) Escola de Comunicação e Arte da Universidade de São Paulo. Disponível em: <<http://www.teses.usp.br/teses/disponiveis/27/27143/tde-15032004-130230/publico/Tese.pdf>>. Acesso em: 21 abr. 2009.

FERREIRA, M. **Introdução à preservação digital**: conceitos, estratégias e actuais consensos. Guimarães, Portugal: Escola de Engenharia da Universidade do Minho, 2006. Disponível em: <<https://repositorium.sdum.uminho.pt/handle/1822/6411>>. Acesso em: 25 nov. 2009.

FERREIRA, Sueli Mara Soares Pinto; SOUTO, Leonardo Fernandes. Dos Sistemas de Informação Federados à Federação de Bibliotecas Digitais. **Revista Brasileira de Biblioteconomia e Documentação**, v. 2, p. 23-40, 2006. Disponível em: <<http://143.106.108.58/seer/ojs/ojs/viewarticle.php?id=17&layout=abstract>>. Acesso em: 03 dez. 2009

FOLTZ, P. W. Using latent semantic indexing for information filtering. In: CONFERENCE ON OFFICE INFORMATION SYSTEMS, 1990, **Proceedings...** Cambridge, MA, 1990. Disponível em: <<http://www-psych.nmsu.edu/~pfoltz/cois/filtering-cois.html>>. Acesso em: 21 abr.2009.

GARCIA, Edel. **Singular value decomposition (SVD) fast track tutorial**. 2006. Disponível em: <<http://www.miislita.com/information-retrieval-tutorial/singular-value-decomposition-fast-track-tutorial.pdf>>. Acesso em: 30 jun. 2009.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo : Atlas, 2002.

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.

GONZÁLEZ DE GÓMEZ, Maria Nélide. Metodologia de pesquisa no campo da ciência da informação. **DataGramaZero** – Revista de Ciência da Informação, v. 1, n. 6, dez. 2000.

GREENSTONE. **Greenstone Digital Library Software**. Disponível em: <<http://www.greenstone.org>>. Acesso em: 12 maio 2010.

GREENSTONE Digital Library Software. **WIKI do Greenstone**. Disponível em: <http://wiki.greenstone.org/wiki/index.php/Main_Page>. Acesso em: 22 maio 2010.

HARRISON, Thomas H. **Intranet data warehouse: ferramentas e técnicas para a utilização do data warehouse na intranet**. São Paulo: Berkeley, 1998.

HOLANDA, A. B. **Dicionário Aurélio Escolar da Língua Portuguesa**, 3. ed. revista e ampliada Rio de Janeiro : Editora Nova Fronteira, 1993.

KAFURE, Ivette. **Usabilidade da imagem na recuperação da informação no catálogo público de acesso em linha**. 2004. 311 p. Tese (Doutorado) - Universidade de Brasília. Departamento de Ciência da Informação e Documentação. Programa de Pós-Graduação em Ciência da Informação. Disponível em:

<http://bdtb.bce.unb.br/tesesimplificado/tde_busca/arquivo.php?codArquivo=1011>. Acesso em: 12 maio 2010.

KLEIN, David A. **A Gestão estratégica do capital intelectual: recursos para uma economia baseada em conhecimento**. Rio de Janeiro: Qualitymarke Ed. 1998.

KURAMOTO, Hélio; TOUTAIN, Lídia Brandão; SAYÃO; Luís (orgs.). **Bibliotecas digitais: saberes e práticas**. Salvador/Brasília: UFBA/IBICT, 2005.

KURAMOTO, Hélio. Ferramentas de software livre para bibliotecas digitais. In: MARCONDES, Carlos H.; KURAMOTO, Hélio; TOUTAIN, Lídia Brandão; SAYÃO; Luís (orgs.). **Bibliotecas digitais: saberes e práticas**. Salvador/Brasília: UFBA/IBICT, 2005. p. 147-164.

LEITE, Fernando César Lima; MÁRDERO ARELLANO, Miguel A. ; MORENO, Fernanda Passini. Acesso livre a publicações e repositórios digitais em Ciência da Informação no Brasil. **Perspectivas em ciência da informação**, Belo Horizonte, v. 11, n. 1, p. 82-94, jan./abr., 2006. Disponível em:

<<http://www.eci.ufmg.br/pcionline/viewarticle.php?id=443&layout=abstract>>. Acesso em: 03 dez. 2009.

LEITE, Fernando César Lima; COSTA, Sely. Repositórios institucionais como ferramentas de gestão do conhecimento científico no ambiente acadêmico. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 11, n.2, p. 206-219, maio/ago. 2006. Disponível em: <<http://www.eci.ufmg.br/pcionline/viewarticle.php?id=419>>. Acesso em: 03 dez.2009.

LANCASTER, F. W. **Information retrieval systems: Characteristics, Testing and Evaluation**. New York: Wiley, 1968.

LOPES, Ilza Leite. Novos paradigmas para avaliação da qualidade da informação em saúde recuperada na Web. **Ciência da Informação**, v. 33, n. 1, p. 81-90, jan./ abr. 2004. Disponível em: <<http://www.ibict.br/cienciadainformacao/viewarticle.php?id=54&layout=html>>. Acesso em: 21 abr. 2009.

MACHADO, Murilo Milton. **Open archives**: panorama dos repositórios. Florianópolis, 2006. 101 f. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Centro de Ciências da Educação. Programa de Pós-Graduação em Ciência da Informação. Disponível em: <<http://www.tede.ufsc.br/teses/PCIN0015.pdf>>. Acesso em: 03 dez. 2009.

MARCONDES, Carlos H.; KURAMOTO, Hélio; TOUTAIN, Lídia Brandão; SAYÃO, Luís (orgs.). **Bibliotecas digitais**: saberes e práticas. Salvador/Brasília : UFBA/IBICT, 2005.

MARCONDES, Carlos Henrique. Metadados: descrição e recuperação de informação na Web. In: MARCONDES, Carlos H.; KURAMOTO, Hélio; TOUTAIN, Lídia Brandão; SAYÃO, Luís (orgs.). **Bibliotecas digitais**: saberes e práticas. Salvador/Brasília: UFBA/IBICT, 2005. p. 97-113.

MÁRDERO ARELLANO, Miguel A.. Preservação de documentos digitais. **Ciência da Informação**, Brasília, v. 33, n. 2, p. 15-27, 2004. Disponível em: <<http://www.scielo.br/pdf/ci/v33n2/a02v33n2.pdf>>. Acesso em: 25 nov. 2009.

MÁRDERO ARELLANO, Miguel Angel; SANTOS, Regina Maria Duarte Moreira dos; FONSECA, Ramón Martins Sodoma. SEER: disseminação de um sistema eletrônico para editoração de revistas científicas no Brasil. **Arquivística.Net**, Rio de Janeiro, v. 1, n. 2, 2006. Disponível em: <<http://www.arquivistica.net/ojs/viewarticle.php?id=33&layout=abstract>>. Acesso em: 25 nov. 2009.

MARCONDES, C. H.; SAYAO, L. F. . Acesso unificado às teses eletrônicas brasileiras. **Informação & Sociedade**. Estudos, João Pessoa, v. 13, n. 1, 2003. Disponível em:

<<http://www.ies.ufpb.br/ojs2/index.php/ies/article/view/125>>. Acesso em: 03 dez. 2009.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Metodologia científica**. 5ª ed. São Paulo: Editora Atlas, 2008.

MARON, Melvin Earl; Kuhns, J. L. On relevance, probabilistic indexing, and information retrieval. **Journal of the ACM**. Disponível em: <<http://www.acm.org/pubs/citations/journals/cacm/>> Acesso em: 21 abr. 2009.

MARTINS, G. de A. **Estudo de caso**: uma estratégia de pesquisa. São Paulo: Atlas, 2006.

MICROSOFT. **Padrão Ecma Office Open XML certificação internacional ISO**. 2008. Disponível em: <<http://www.microsoft.com/latam/presspass/brasil/2008/abril/openxml.msp>> Acesso em: 29 mai. 2010.

MINAYO, M. C. de S. **O desafio do conhecimento**. 2. ed. São Paulo/Rio de Janeiro, 1993.

MIRANDA, Antonio. **Ciência da informação**; teoria e metodologia de uma área em expansão. Brasília: Tesaurus, 2003.

MOOERS, Calvin Northrup. Zatocoding applied to mechanical organization of knowledge. **American Documentation**, v.2, n.1, p.20-32, 1951.

NATIONAL ARCHIVES AND RECORDS ADMINISTRATION. **NARA**. Disponível em: <<http://www.archives.gov/index.html>>. Acesso em: 03 dez. 2009.

NATIONAL INFORMATION STANDARD ORGANIZATION. **Framework Advisory**: A framework of Guidance for Building Good Digital Collection. Bethesda, MD : National Information Standards Organization (NISO). 2004a. Disponível em: <<http://www.niso.org/framework/framework2.pdf>>. Acesso em: 25 nov. 2009.

NEW ZEALAND DIGITAL LIBRARY PROJECT - NZDL. Disponível em: <<http://nzdl.sadl.uleth.ca/cgi-bin/library.cgi>>. Acesso em: 12 maio 2010.

PAES CARDOSO, O. N. Recuperação da Informação. **Infocomp: Journal of Computer Science**, v. 2, n. 1. 2000. Disponível em: <<http://www.dcc.ufla.br/infocomp/artigos/v2.1/art07.pdf>>. Acesso em: 12 maio 2010

PDFLIB. **XMP in PDF/A**. Disponível em: <<http://www.pdfliib.com/knowledge-base/xmp-metadata/>>. Acesso em: 22 de mai. de 2010.

PINHEIRO, Lena Vânia Ribeiro. Campo interdisciplinar da Ciência da Informação: fronteiras remotas e recentes. In: PINHEIRO, Lena V. Ribeiro (org.). **Ciência da Informação, Ciências Sociais e Interdisciplinaridade**. Brasília/Rio de Janeiro, IBICT/DDI/DEP, 1999, p. 155-182.

PROPOSTA submetida pela Comissão Nacional da UNESCO dos países baixos apresentada à Conferência Geral da UNESCO e aprovada para inclusão no programa para 2002-2003. In: **BIBLIOTECA NACIONAL (Portugal)**. Manifesto para a Preservação Digital, UNESCO. Disponível em: <http://www.bn.pt/agenda/ecpa/manifesto_unesco.html>. Acesso em: 25 nov. 2009.

RIJSBERGEN, C. J. van. **Information retrieval**. 1999. Disponível em: <<http://www.dcs.gla.ac.uk/~iain/keith/>>. Acesso em: 21 abr. 2009.

ROBERTSON, S.E; Teories and models in information retrieval. **Journal of Documentation**, 33, p. 126-148. 1977.

SALTON, G.; MCGILL, M. J. **Introduction to Modern Information Retrieval**. McGraw Hill, 1983. 448p.

SALTON, G.; FOX, E.A., WU, H. Extended Boolean information retrieval. **Communications of the ACM**, v.26, n.11, p.1022-1036, Nov. 1983. Disponível em: <<http://portal.acm.org/citation.cfm?id=358466>>. Acesso em: 21 maio 2010.

SARACEVIC, T. Ciência da Informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996.

SAYÃO, Luis Fernando. Preservação digital no contexto das bibliotecas digitais. In: Marcondes, C. H.; Kuramoto, H.; Toutain, Lidia Brandão; Sayão, Luis Fernando.(Org.). **Bibliotecas digitais: saberes e práticas**. Salvador/Brasília: UFBA/IBICT, 2006, p. 115-149.

SAYAO, L. F.. Padrões para bibliotecas digitais abertas e interoperáveis. **Encontros Bibli**, v. 1, p. 2, 2007. Disponível em: <<http://www.periodicos.ufsc.br/index.php/eb/article/viewFile/461/463>>. Acesso em: 03 dez. 2009.

SILVA, E.; MENEZES, E. M. **Metodologia da pesquisa e elaboração de dissertação**. 3.ed. Florianópolis: Laboratório de Ensino a Distância da UFSC, 2001.121 p.

TAMMARO, Anna Maria; SALARELLI, Alberto. **A biblioteca digital**. Brasília: Briquet de Lemos, 2008. 377p

TAKAHASHI, T. (Org.). **Sociedade da Informação no Brasil: o livro verde**. Brasília: Ministério da Ciência e Tecnologia, 2000.

TAKAO, Eduardo Liqueo; **Uma análise de modelos de sistemas probabilísticos em recuperação de informação em bases textuais**, 2001, Dissertação (Mestrado em Ciências da Computação - UFSC). Disponível em: <<http://www.tede.ufsc.br/teses/PGCC0157.pdf>>. Acesso em: 21 abr. 2009.

TOMAÉL, M. I. et alii. Avaliação de fontes de informação na internet; critérios de qualidade. **Informação & Sociedade**; estudos, João Pessoa, v. 11, n. 2, p. 13-35, 2001.

UNESCO. UNESCO's programme aiming at preservation and dissemination of valuable archive holdings and library collections worldwide. **E-Heritage**. Disponível em: <http://portal.unesco.org/ci/en/ev.php-URL_ID=1539&URL_DO=DO_TOPIC&URL_SECTION=201.html>. Acesso em: 15 jun. 2010.

VANDERLEI FILHO, D. ; VALENCA, M. J. S. ; LUDERMIR, T. B. ; SILVA, G. P. F. . Uma Proposta Fuzzy na Avaliação de Desempenho de Bibliotecas Universitárias Brasileiras. In: XII SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS DA AMÉRICA LATINA E DO CARIBE, 12., SIMPÓSIO DE DIRETORES DE BIBLIOTECAS UNIVERSITÁRIAS DA AMÉRICA LATINA E CARIBE, 2. 2002, **Anais...** Recife, 2002. Disponível em: <<http://www.sibi.ufrj.br/snbu/snbu2002/oralpdf/38.a.pdf>>. Acesso em: 01 maio 2009.

VICENTINI, Luiz Atílio. Gestão em bibliotecas digitais. In: MARCONDES, Carlos H.; KURAMOTO, Hélio; TOUTAIN, Lídia Brandão; SAYÃO, Luís (orgs.). **Bibliotecas digitais: saberes e práticas**. Salvador/Brasília : UFBA/IBICT, 2005. p. 243-262.

XMP.**Open industry initiative**. Disponível em: <www.xmpopen.org>. Acesso em: 22 maio 2010.

YIN, Robert K. **Estudo de caso: planejamento e métodos**. 3. ed. Porto Alegre: Bookman, 2005.

ANEXO B – Relatório gerado pelo software DROID

| URI | EXT | PUID | FORMAT NAME | FOR MAT VERS ION |
|---|-----|------|----------------|---------------------------|
| arquivo teste gerado na versão openoffice sxw.sxw!/mimetype | | | | |
| arquivo teste gerado na versão openoffice sxw.sxw!/Configurations2/sta tusbar/ | | | | |
| arquivo teste gerado na versão openoffice sxw.sxw!/Configurations2/flo ater/ | | | | |
| arquivo teste gerado na versão openoffice sxw.sxw!/Configurations2/po pupmenu/ | | | | |
| arquivo teste gerado na versão openoffice sxw.sxw!/Configurations2/pro gressbar/ | | | | |
| arquivo teste gerado na versão openoffice sxw.sxw!/Configurations2/me nubar/ | | | | |
| arquivo teste gerado na versão odf.odt!/mimetype | | | | |
| arquivo teste gerado na versão odf!/mimetype | | | | |
| arquivo teste gerado na versão odf!/Configurations2/statusba r/ | | | | |
| arquivo teste gerado na versão | | | | |

| | | | | |
|--|-----|--|--|--|
| openoffice sxw.sxw!/Configurations2/toolbar/ | | | | |
| arquivo teste gerado na versão openoffice sxw.sxw!/Configurations2/images/Bitmaps/ | | | | |
| arquivo teste gerado na versão odf!/Configurations2/accelerator/current.xml | xml | | | |
| arquivo teste gerado na versão odf!/Configurations2/floater/ | | | | |
| arquivo teste gerado na versão odf!/Configurations2/popupmenu/ | | | | |
| arquivo teste gerado na versão odf!/Configurations2/progressbar/ | | | | |
| arquivo teste gerado na versão odf!/Configurations2/menubar/ | | | | |
| arquivo teste gerado na versão odf!/Configurations2/toolbar/ | | | | |
| arquivo teste gerado na versão odf!/Configurations2/images/Bitmaps/ | | | | |
| arquivo teste gerado na versão openoffice sxw.sxw!/Configurations2/accelerator/current.xml | xml | | | |
| arquivo teste gerado na versão openoffice sxw.sxw!/layout-cache | | | | |
| arquivo_teste_gerado_na_versao_openofficegoogle.odt!/mimetype | | | | |
| arquivo teste gerado na versão | | | | |

| | | | | |
|---|-----|-----------|--|---------|
| odf!/layout-cache | | | | |
| apresentacao dissertação Jairo 97 03.pps | pps | fmt/126 | Microsoft Powerpoint Presentation | 97-2002 |
| apresentacao qualificacao Jairo 161209.ppt | ppt | fmt/126 | Microsoft Powerpoint Presentation | 97-2002 |
| arquivo teste gerado na versão odf!/Pictures/000007000094D500005CD1BBE632.wmf | wmf | x-fmt/119 | Windows Metafile | |
| arquivo teste gerado na versão odf!/Pictures/00000700003448000022532A1F2EB4.wmf | wmf | x-fmt/119 | Windows Metafile | |
| arquivo teste gerado na versão odf!/Pictures/000007000044D6000026763AF4E4E4.wmf | wmf | x-fmt/119 | Windows Metafile | |
| arquivo teste gerado na versão odf.odt!/media/image46.emf | emf | x-fmt/153 | Windows Enhanced Metafile | |
| arquivo teste gerado na versão odf.odt!/media/image31.emf | emf | x-fmt/153 | Windows Enhanced Metafile | |
| arquivo teste gerado na versão odf.odt!/media/image32.emf | emf | x-fmt/153 | Windows Enhanced Metafile | |
| arquivo teste gerado na versão openoffice sxw.sxw | sxw | x-fmt/263 | ZIP Format | |
| arquivo teste gerado na versão odf.odt | odt | x-fmt/263 | ZIP Format | |
| arquivo teste gerado na versão odf | | x-fmt/263 | ZIP Format | |
| Petro Bibiana.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |

| | | | | |
|--------------------------|-----|--------|--|-----|
| Camila Koerich Burin.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |
| CarlosAlmeida.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |
| Chirley Silva.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |
| EldaLira.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |
| ElianePereira.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |
| Floriani Vivian.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |
| Francisca Rasche.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |

| | | | | |
|----------------------------|-----|--------|--|-----|
| FranciscaRasche.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |
| Gerson Tybusch.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |
| Jaqueline Alves.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |
| Lidiane dos Santos.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |
| LUCIANE PAULA VITAL.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |
| Machado Marli.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |
| MariliLopes.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |

| | | | | |
|---|-----|--------|--|-----|
| Petro Bibiana.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |
| Schenkel Marilia.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |
| Silva Catia.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |
| SilvanaBueno.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |
| ViniciusLucca.pdf | pdf | fmt/17 | Acrobat PDF 1.3 - Portable Document Format | 1.3 |
| arquivo teste gerado na versão PDF A.pdf | pdf | fmt/95 | Acrobat PDF/A - Portable Document Format | 1 |
| arquivo teste gerado na versão PDF A.pdf | pdf | fmt/18 | | 1.4 |
| arquivo teste gerado na versão pdf.pdf | pdf | fmt/95 | Acrobat PDF/A - Portable Document Format | 1 |

| | | | | |
|--|-----|--------|--|-----|
| arquivo teste gerado na versão pdf.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| Adriana Crispim.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| andrenizia aquino eluan.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| Davilla Guillermo.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| Erica Ribeiro.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| FleckFelicia.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| Gabriela Farias.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |

| | | | | |
|------------------------|-----|--------|--|-----|
| Gardenia Castro.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| Gelci Rostirolla.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| Graipel Hermes.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| MARCIO_JOSE_SEMBAY.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| Margarida Reis.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| MargaridaReis.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| MOMM_Christiane.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |

| | | | | |
|-------------------------|-----|--------|--|-----|
| ReneeNina.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| RenataCurty.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| Pinheiro Liliane.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| Sales Rodrigo.pdf | pdf | fmt/95 | Acrobat PDF/A - Portable Document Format | 1 |
| Sales Rodrigo.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| Schtuz Sergio.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| Silva Fabiano Couto.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |

| | | | | |
|---------------------------|-----|--------|--|-----|
| SonaliBedin.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| Soraya_Waltrick.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| Tassiane Altissimo_09.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| Veridiana Abe.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| AnnaCorreia.pdf | pdf | fmt/19 | Acrobat PDF 1.5 - Portable Document Format | 1.5 |
| DerliDorigon.pdf | pdf | fmt/19 | Acrobat PDF 1.5 - Portable Document Format | 1.5 |
| NelmaAraujo.pdf | pdf | fmt/19 | Acrobat PDF 1.5 - Portable Document Format | 1.5 |

| | | | | |
|---|-----|--------|--|-----|
| Savi Gorete.pdf | pdf | fmt/19 | Acrobat PDF 1.5 - Portable Document Format | 1.5 |
| Schons Claudio H.pdf | pdf | fmt/19 | Acrobat PDF 1.5 - Portable Document Format | 1.5 |
| Virgil Johnny.pdf | pdf | fmt/19 | Acrobat PDF 1.5 - Portable Document Format | 1.5 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/46 | Rich Text Format | 1.1 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/47 | Rich Text Format | 1.2 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/48 | Rich Text Format | 1.3 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/49 | Rich Text Format | 1.4 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/45 | Rich Text Format | 1.0 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/46 | Rich Text Format | 1.1 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/47 | Rich Text Format | 1.2 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/48 | Rich Text Format | 1.3 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/49 | Rich Text Format | 1.4 |

| | | | | |
|--|-----|--------|--|-----|
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/45 | Rich Text Format | 1.0 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/46 | Rich Text Format | 1.1 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/47 | Rich Text Format | 1.2 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/48 | Rich Text Format | 1.3 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/49 | Rich Text Format | 1.4 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/45 | Rich Text Format | 1.0 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/46 | Rich Text Format | 1.1 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/47 | Rich Text Format | 1.2 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/48 | Rich Text Format | 1.3 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/49 | Rich Text Format | 1.4 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/45 | Rich Text Format | 1.0 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/46 | Rich Text Format | 1.1 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/47 | Rich Text Format | 1.2 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/48 | Rich Text Format | 1.3 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/49 | Rich Text Format | 1.4 |
| arquivo teste gerado na versão rtf.rtf | rtf | fmt/45 | Rich Text Format | 1.0 |
| Eliane Garcez.pdf | pdf | fmt/20 | Acrobat PDF 1.6 - Portable Document Format | 1.6 |

| | | | | |
|--|-----|---------|--|-----|
| Molossi Sinara.pdf | pdf | fmt/20 | Acrobat PDF 1.6 - Portable Document Format | 1.6 |
| arquivo teste gerado na versão odf.odt!/META- INF/manifest.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo teste gerado na versão odf.odt!/settings.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo teste gerado na versão odf.odt!/meta.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo_teste_gerado_na_vers ao_openofficegoogle.odt!/ME TA-INF/manifest.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo teste gerado na versão openoffice sxw.sxw!/content.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo_teste_gerado_na_vers ao_openofficegoogle.odt!/con tent.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo_teste_gerado_na_vers ao_openofficegoogle.odt!/met a.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo_teste_gerado_na_vers ao_openofficegoogle.odt!/sett ings.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo_teste_gerado_na_vers ao_openofficegoogle.odt!/styl es.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo teste gerado na versão openoffice sxw.sxw!/styles.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |

| | | | | |
|--|-----|---------|----------------------------|-----|
| arquivo teste gerado na versão openoffice sxw.sxw!/settings.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo teste gerado na versão openoffice sxw.sxw!/META-INF/manifest.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo teste gerado na versão openoffice sxw.sxw!/meta.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo teste gerado na versão odf!/content.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo teste gerado na versão odf.odt!/content.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo teste gerado na versão odf!/settings.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo teste gerado na versão odf!/styles.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo teste gerado na versão odf!/META-INF/manifest.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo teste gerado na versão odf!/meta.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo teste gerado na versão odf!/manifest.rdf | rdf | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo teste gerado na versão odf.odt!/styles.xml | xml | fmt/101 | Extensible Markup Language | 1.0 |
| arquivo teste gerado na versão openoffice sxw.sxw!/Thumbnails/thumbnail.png | png | fmt/11 | Portable Network Graphics | 1.0 |
| arquivo teste gerado na versão ao openofficegoogle.odt!/Pict | png | fmt/11 | Portable Network | 1.0 |

| | | | | |
|---|-----|--------|---------------------------------|-----|
| ures/image2.png | | | Graphics | |
| arquivo teste gerado na versão odf!/Thumbnails/thumbnail.png | png | fmt/11 | Portable Network Graphics | 1.0 |
| arquivo teste gerado na versão odf!/Pictures/10000000000000000000 5A00000035C70036833.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão openoffice sxw.sxw!/Pictures/1000000000 0000500000003512DD754.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image56.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/10000000000000000000 5A0000003680E6A707F.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/10000000000000000000 444000002C58F0E4D2E.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image54.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image58.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image60.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image59.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image55.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image57.png | png | fmt/12 | Portable Network Graphics | 1.1 |

| | | | | |
|--|-----|--------|---------------------------------|-----|
| arquivo teste gerado na versão odf.odt!/media/image62.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 2D0000002270BBE2538.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image45.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image52.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 5A000000368C1EEDBB9.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 30E00000103EBAEE86C.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image61.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 444000002C5F2FE190D.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 444000002C5C11E683B.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 5A000000368FBC1B54E.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 09A00000038BE753296.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image44.png | png | fmt/12 | Portable Network Graphics | 1.1 |

| | | | | |
|--|-----|--------|---------------------------------|-----|
| arquivo teste gerado na versão odf.odt!/media/image53.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image50.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image51.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image49.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 444000002C54C841C41.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 157000001BD2FBBE90D.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image48.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image47.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image64.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 2B1000001E8F7CE931F.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image65.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 31000000FB005404EC.png | png | fmt/12 | Portable Network Graphics | 1.1 |

| | | | | |
|--|-----|--------|---------------------------------|-----|
| arquivo teste gerado na versão odf!/Pictures/1000000000000 2D0000001FA546FB008.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image76.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 30E000000FB6EDC8D96.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 310000000FC60FD80AC.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image77.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 30000021C42E3FE51.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 0EE0000004654A8BBF7.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image80.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image79.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image81.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 30A000000FB52E98CCE.png | png | fmt/12 | Portable Network Graphics | 1.1 |

| | | | | |
|--|-----|--------|---------------------------------|-----|
| arquivo teste gerado na versão odf!/Pictures/1000000000000 280000001E012988391.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image78.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 5A00000036888C04806.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 5A0000003680A98C8CE.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image74.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 444000002C5E241EDEC.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image75.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 31500000FB17F7734A.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 281000001E0FD2F9B50.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 27F000001E0580DD526.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 31400000FEBD5C0964.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image82.png | png | fmt/12 | Portable Network | 1.1 |

| | | | Graphics | |
|---|-----|--------|---------------------------------|-----|
| arquivo teste gerado na versão odf.odt!/media/image67.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image66.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image69.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 31400001011D375589.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image73.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 44400002C5B983E15D.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image63.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image72.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image71.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image68.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image41.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image12.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão | png | fmt/12 | Portable | 1.1 |

| | | | | |
|---|-----|--------|---------------------------|-----|
| odf.odt!/media/image70.png | | | Network Graphics | |
| arquivo teste gerado na versão odf!/Pictures/10000000000005A0000003680AD89B26.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image43.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/100000000000021000007CBF758F9D.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image13.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000444000002C565BD6937.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000310000000FB28596D02.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/10000000000005A00000036816906974.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image14.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000500000003512DD754.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/10000000000002D00000022D0D09EACB.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000311000001006141ECA8.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão | png | fmt/12 | Portable | 1.1 |

| | | | | |
|---|-----|--------|---------------------------|-----|
| odf.odt!/media/image15.png | | | Network Graphics | |
| arquivo teste gerado na versão odf.odt!/media/image16.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/100000000000032B000002373FE182FB.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image17.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/100000000000030C00000FC7FEE6973.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/100000000000027F000001E0085A2747.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/10000000000005A00000036828DA6DC0.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image18.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/100000000000031300000101657176C1.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image19.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/10000000000005A0000003686DD85C8F.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image4.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão | png | fmt/12 | Portable | 1.1 |

| | | | | |
|---|-----|--------|---------------------------|-----|
| odf.odt!/media/image3.png | | | Network Graphics | |
| arquivo teste gerado na versão odf!/Pictures/10000000000003100001018ADA7CE9.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image2.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000310000FC85B815D2.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/10000000000003130000100D238180D.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/10000000000005A000000368AC6F956E.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image9.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image11.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image10.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000444000002C5B8B67C12.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image5.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000444000002C5D75EF7DE.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image6.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão | png | fmt/12 | Portable | 1.1 |

| | | | | |
|---|-----|--------|---------------------------|-----|
| odf.odt!/media/image8.png | | | Network Graphics | |
| arquivo teste gerado na versão odf.odt!/media/image7.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image37.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image36.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image38.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image39.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image33.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/10000000000005A0000003687B5D6DF8.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image22.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image35.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image30.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image40.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão | png | fmt/12 | Portable | 1.1 |

| | | | | |
|---|-----|--------|---------------------------|-----|
| odf.odt!/media/image34.png | | | Network Graphics | |
| arquivo teste gerado na versão odf.odt!/media/image21.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/100000000000031500000100B476E8C1.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/10000000000002D00000022D6E9A9BF5.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000444000002C5B0E54DE9.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000280000001E0D040B124.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/10000000000001450000004CA33D723A.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/10000000000003100000FFDB372D6B.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image23.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000264000001621A15FFDF.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/10000000000002AC000001B29ED8AF06.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/10000000000005A0000003680CA9523C.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 | png | fmt/12 | Portable Network | 1.1 |

| | | | | |
|--|-----|--------|---------------------------------|-----|
| 193000001368DFF63B2.png | | | Graphics | |
| arquivo teste gerado na versão odf.odt!/media/image24.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 5A0000003684C470948.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image25.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image26.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 27F000001E04E2A4356.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image27.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 5A00000035C48AD54A7.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 5A00000036890485372.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 31400000FF723226F8.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 444000002C5D7ED3375.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 1A50000012DB0B3B394.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 5A0000003684AD65801.png | png | fmt/12 | Portable Network Graphics | 1.1 |

| | | | | |
|---|------|--------|------------------------------------|------|
| arquivo teste gerado na versão odf!/Pictures/1000000000000 281000001E0347C34C1.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 444000002C50DDBA251.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 444000002C5A03592BB.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf.odt!/media/image42.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 23E0000017D9BF4C21D.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 5A000000368530A5BAA.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 30E00000F9EB5C070C.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 31300000103856F290B.png | png | fmt/12 | Portable Network Graphics | 1.1 |
| arquivo teste gerado na versão openoffice sxw.sxw!/Pictures/100000000 00000BD00000CF94ADE4F B.jpg | jpg | fmt/43 | JPEG File Interchange Format | 1.01 |
| arquivo teste gerado na versão odf.odt!/media/image1.jpeg | jpeg | fmt/43 | JPEG File Interchange Format | 1.01 |
| arquivo teste gerado na versão odf!/Pictures/1000000000000 | jpg | fmt/43 | JPEG File Interchange | 1.01 |

| | | | | |
|---|------|---------|--|------|
| 0BD00000CF94ADE4FB.jpg | | | Format | |
| arquivo_teste_gerado_na_versao_openofficegoogle.odt!/Pictures/image1.jpeg | jpeg | fmt/44 | JPEG File Interchange Format | 1.02 |
| arquivo teste gerado na versao odfl/Pictures/1000000000000237000000E3378053C8.jpg | jpg | fmt/44 | JPEG File Interchange Format | 1.02 |
| arquivo teste gerado na versao odfl/Pictures/1000000000000237000000E381ECD7BF.jpg | jpg | fmt/44 | JPEG File Interchange Format | 1.02 |
| arquivo teste gerado na versao odfl.odt!/media/image28.jpeg | jpeg | fmt/44 | JPEG File Interchange Format | 1.02 |
| arquivo teste gerado na versao odfl.odt!/media/image29.jpeg | jpeg | fmt/44 | JPEG File Interchange Format | 1.02 |
| arquivo teste gerado na versao EXCEL 97-03 XLS.xls | xls | fmt/111 | OLE2 Compound Document Format | |
| arquivo teste gerado na versao Word 07 93.doc | doc | fmt/111 | OLE2 Compound Document Format | |
| Recuperaçao inteligente de informaçoes em portais corporativos.doc | doc | fmt/111 | OLE2 Compound Document Format | |
| arquivo teste gerado na versao PDF A.pdf | pdf | fmt/95 | Acrobat PDF/A - Portable Document Format | 1 |
| arquivo teste gerado na versao PDF A.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |

| | | | | |
|---|------|---------|--|------|
| arquivo teste gerado na versão pdf.pdf | pdf | fmt/95 | Acrobat PDF/A - Portable Document Format | 1 |
| arquivo teste gerado na versão pdf.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| Sales Rodrigo.pdf | pdf | fmt/95 | Acrobat PDF/A - Portable Document Format | 1 |
| Sales Rodrigo.pdf | pdf | fmt/18 | Acrobat PDF 1.4 - Portable Document Format | 1.4 |
| arquivo teste gerado na versão EXCEL XLS.xlsx | xlsx | fmt/189 | Microsoft Office Open XML | 2007 |
| apresentacao dissertação Jairo.pptx | pptx | fmt/189 | Microsoft Office Open XML | 2007 |
| arquivo teste gerado na versão word 10.docx | docx | fmt/189 | Microsoft Office Open XML | 2007 |

ANEXO C – National Archives – formato fmt/111

The National Archives


Search the archives

MyPage (n)

Adv

About us | Education | Records | Information management | Shop online

You are here: [Home](#) > [Services for professionals](#) > [Preservation](#) > [PRONOM](#) > [Search by format](#) > Details: Summary

 The technical registry
PRONOM

Welcome | About | Add an er
Search | Help | Informatic

Details: File format summary ? Help : detailed report on

Simple search | File format | PRONOM Unique Identifier | Software | Vendor | Lifecycles | Migration Pathways

Details for: **OLE2 Compound Document Format** Save as... XML | CSV

Go to: [Summary](#) | [Documentation](#) > | [Signatures](#) > | [Compression](#) > | [Character encoding](#) > | [Rights](#) > | [Reference files](#) > | [Properties](#) >

Summary

| | |
|-----------------------------|---|
| Name | OLE2 Compound Document Format |
| Version | |
| Other names | |
| Identifiers | PUID: fmt/111 |
| Family | |
| Classification | |
| Disclosure | None |
| Description | The OLE2 Compound Document Format is a generic document format developed by Microsoft as underlying native binary format for many of its Office applications. The format is proprietary and Microsoft does not make details of its structure public. The information here is derived primarily OpenOffice.org's reverse-engineered documentation of the format and should not therefore be regarded as definitive. A Compound Document acts as a file system in which independent data streams are organised within a hierarchy of containers, called storages. All storages and stream contained within a parent Root Storage. A Directory Stream indexes every stream and storage in file. A Compound Document begins with the Compound Document Header, including pointers to location of the Directory Stream and Master Sector Allocation Table. The remainder of the file is organised into Sectors, the positions of which are defined in the Sector Allocation Table and Master Sector Allocation Table. |
| Orientation | Binary |
| Byte order | Little-endian (Intel) and Big-endian (Motorola) |
| Related file formats | <ul style="list-style-type: none"> Has lower priority than Microsoft Powerpoint Presentation (4.0) Has lower priority than Microsoft Powerpoint Presentation (95) Has lower priority than Microsoft Powerpoint Presentation (97-2002) Has lower priority than Microsoft Excel 5.0 Workbook (xls) (5) Has lower priority than Microsoft Excel 95 Workbook (xls) (7) Has lower priority than Microsoft Excel 97 Workbook (xls) (8) Has lower priority than Microsoft Excel 2000-2003 Workbook (xls) (8X) Has lower priority than Microsoft Word for Windows Document (6.0/95) Has lower priority than Microsoft Word for Windows Document (97-2003) Has lower priority than Microsoft Outlook Email Message (97-2003) Is supertype of Microsoft Powerpoint Presentation (95) Is supertype of Microsoft Powerpoint Presentation (97-2002) Is supertype of Microsoft Excel 5.0 Workbook (xls) (5) Is supertype of Microsoft Excel 95 Workbook (xls) (7) Is supertype of Microsoft Excel 97 Workbook (xls) (8) Is supertype of Microsoft Excel 2000-2003 Workbook (xls) (8X) Is supertype of Microsoft Word for Windows Document (6.0/95) Is supertype of Microsoft Word for Windows Document (97-2003) Is supertype of Revit Family File (n/a) Is supertype of Revit Family Template (n/a) Is supertype of Revit Template (n/a) Is supertype of Revit External Group (n/a) Is supertype of Revit Project (n/a) |

Internet | Modo Protegido: Desativado


ANEXO D – National Archives – formato fmt/17

MyPage (n)

The National Archives Search the archives

About us Education Records Information management Shop online

You are here: [Home](#) > [Services for professionals](#) > [Preservation](#) > [PRONOM](#) > [Search by format](#) > Details: Summary



The technical registry
PRONOM

[Welcome](#) [About](#) [Add an entry](#)
[Search](#) [Help](#) [Information](#)



Details: File format summary ? Help : detailed report on

Simple search File format **PRONOM Unique Identifier** Software Vendor Lifecycles Migration Pathways

Details for: Acrobat PDF 1.3 - Portable Document Format 1.3 [Save as...](#) [XML](#) | [CSV](#)

Go to: [Summary](#) | [Documentation](#) > | [Signatures](#) > | [Compression](#) > | [Character encoding](#) > | [Rights](#) > | [Reference files](#) > | [Properties](#) >

Summary

| | |
|------------------------------|---|
| Name | Acrobat PDF 1.3 - Portable Document Format |
| Version | 1.3 |
| Other names | PDF (1.3) |
| Identifiers | MIME: application/pdf Apple Uniform Type Identifier: com.adobe.pdf PUID: fmt/17 |
| Family | |
| Classification | Page Description |
| Disclosure | Full |
| Description | Portable Document Format is a platform-independent format for representing formatted documents developed by Adobe Systems Incorporated. It is the native format of Adobe's Acrobat family of software products, version 1.3 corresponding to the release of Acrobat 4.0. PDF is based on, and shares the same imaging model as, the PostScript page description language. A PDF file comprises a Header section, a Body section containing the objects which make up the document, a Cross Reference Table, and a Trailer section. PDF files can contain a wide variety of content, including images, video and audio. |
| Orientation | Binary |
| Byte order | Big-endian (Motorola) |
| Related file formats | Has lower priority than Acrobat PDF 1.4 - Portable Document Format (1.4) Has lower priority than Acrobat PDF 1.5 - Portable Document Format (1.5) Has lower priority than Acrobat PDF 1.6 - Portable Document Format (1.6) Is previous version of Acrobat PDF 1.4 - Portable Document Format (1.4) Is subsequent version of Acrobat PDF 1.2 - Portable Document Format (1.2) Is supertype of Acrobat PDF/X - Portable Document Format - Exchange 1:2001 Is supertype of Acrobat PDF/X - Portable Document Format - Exchange 1a:2001 Is supertype of Acrobat PDF/X - Portable Document Format - Exchange 3:2002 |
| Technical Environment | |
| Released | 11 Mar 1999 |
| Supported until | |
| Format Risk | |
| Developed by |  Adobe Systems Incorporated |
| Supported by | None. |
| Source |  Digital Preservation Department / The National Archives |
| Source date | 11 Mar 2005 |
| Source description | |
| Last updated | 22 Oct 2009 |

Internet | Modo Protegido: Desactivado 100%

ANEXO E – National Archives – formato fmt/18

The screenshot displays the National Archives website interface. At the top, the logo for 'The National Archives' is visible, along with a search bar and navigation tabs for 'About us', 'Education', 'Records', 'Information management', and 'Shop online'. The breadcrumb trail indicates the user is in the 'PRONOM' section, specifically under 'Search by format' and 'Details: Summary'.

The main content area is titled 'The technical registry PRONOM' and shows 'Details for: Acrobat PDF 1.4 - Portable Document Format 1.4'. A navigation menu includes 'Simple search', 'File format', 'PRONOM Unique Identifier', 'Software', 'Vendor', 'Lifecycles', and 'Migration Pathways'. There are also links for 'Save as...', 'XML', and 'CSV'.

The 'Summary' section provides the following details:

| | |
|------------------------------|---|
| Name | Acrobat PDF 1.4 - Portable Document Format |
| Version | 1.4 |
| Other names | PDF (1.4) |
| Identifiers | MIME: application/pdf Apple Uniform Type Identifier: com.adobe.pdf PUID: fmt/18 |
| Family | |
| Classification | Page Description |
| Disclosure | Full |
| Description | Portable Document Format is a platform-independent format for representing formatted documents developed by Adobe Systems Incorporated. It is the native format of Adobe's Acrobat family of software products, version 1.4 corresponding to the release of Acrobat 5.0. PDF is based on, and shares the same imaging model as, the PostScript page description language. A PDF file comprises a Header section, a Body section containing the objects which make up the document, a Cross Reference Table, and a Trailer section. PDF files can contain a wide variety of content, including images, video and audio. |
| Orientation | Binary |
| Byte order | Big-endian (Motorola) |
| Related file formats | Has lower priority than Acrobat PDF 1.5 - Portable Document Format (1.5) Has lower priority than Acrobat PDF 1.6 - Portable Document Format (1.6) Has priority over Acrobat PDF 1.0 - Portable Document Format (1.0) Has priority over Acrobat PDF 1.1 - Portable Document Format (1.1) Has priority over Acrobat PDF 1.2 - Portable Document Format (1.2) Has priority over Acrobat PDF 1.3 - Portable Document Format (1.3) Is previous version of Acrobat PDF 1.5 - Portable Document Format (1.5) Is subsequent version of Acrobat PDF 1.3 - Portable Document Format (1.3) Is supertype of Acrobat PDF/X - Portable Document Format - Exchange 1a:2003 Is supertype of Acrobat PDF/X - Portable Document Format - Exchange 2:2003 Is supertype of Acrobat PDF/X - Portable Document Format - Exchange 3:2003 |
| Technical Environment | |
| Released | 01 Dec 2001 |
| Supported until | |
| Format Risk | |
| Developed by | Adobe Systems Incorporated |
| Supported by | None. |
| Source | Digital Preservation Department / The National Archives |
| Source date | 11 Mar 2005 |

The browser's status bar at the bottom indicates 'Internet | Modo Protegido: Desativado' and a zoom level of 100%.

ANEXO F – National Archives – formato fmt/19

The screenshot shows the National Archives website interface. At the top, there is a search bar and navigation tabs for 'About us', 'Education', 'Records', 'Information management', and 'Shop online'. The main content area is titled 'The technical registry PRONOM' and displays 'Details: File format summary' for 'Acrobat PDF 1.5 - Portable Document Format 1.5'. The page includes a breadcrumb trail, a search bar, and a navigation menu with options like 'Summary', 'Documentation', 'Signatures', 'Compression', 'Character encoding', 'Rights', and 'Reference files'. The main content is organized into sections: Summary, Technical Environment, and Source information.

Summary

| | |
|------------------------------|--|
| Name | Acrobat PDF 1.5 - Portable Document Format |
| Version | 1.5 |
| Other names | PDF (1.5) |
| Identifiers | MIME: application/pdf Apple Uniform Type Identifier: com.adobe.pdf PLUD: fmt/19 |
| Family | |
| Classification | Page Description |
| Disclosure | Full |
| Description | Portable Document Format is a platform-independent format for representing formatted documents developed by Adobe Systems Incorporated. It is the native format of Adobe's Acrobat family of software products, version 1.5 corresponding to the release of Acrobat 6.0. PDF is based on, and shares the same imaging model as, the PostScript page description language. A PDF file comprises: Header section, a Body section containing the objects which make up the document, a Cross Reference Table, and a Trailer section. PDF files can contain a wide variety of content, including images, video and audio. |
| Orientation | Binary |
| Byte order | Big-endian (Motorola) |
| Related file formats | Has lower priority than Acrobat PDF 1.6 - Portable Document Format (1.6) Has priority over Acrobat PDF 1.0 - Portable Document Format (1.0) Has priority over Acrobat PDF 1.1 - Portable Document Format (1.1) Has priority over Acrobat PDF 1.2 - Portable Document Format (1.2) Has priority over Acrobat PDF 1.3 - Portable Document Format (1.3) Has priority over Acrobat PDF 1.4 - Portable Document Format (1.4) Is previous version of Acrobat PDF 1.6 - Portable Document Format (1.6) Is subsequent version of Acrobat PDF 1.4 - Portable Document Format (1.4) |
| Technical Environment | |
| Released | 01 Jan 2003 |
| Supported until | |
| Format Risk | |
| Developed by | Adobe Systems Incorporated |
| Supported by | None. |
| Source | Digital Preservation Department / The National Archives |
| Source date | 11 Mar 2005 |
| Source description | |
| Last updated | 22 Oct 2009 |

Internet | Modo Protegido: Desativado