UNIVERSIDADE DE BRASÍLIA
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, CONTABILIDADE,
CIÊNCIA DA INFORMAÇÃO E DOCUMENTAÇÃO
DEPARTAMENTO DE CIÊNCIA DA INFORMAÇÃO E DOCUMENTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

PRESERVAÇÃO DE DOCUMENTOS DIGITAIS:
O PAPEL DOS FORMATOS DE ARQUIVO

Ernesto Carlos Bodê

ORIENTADORA: Profª Dra. Miriam Paula Manini

BRASÍLIA
2008

ERNESTO CARLOS BODÊ

PRESERVAÇÃO DE DOCUMENTOS DIGITAIS: O PAPEL DOS FORMATOS DE ARQUIVO

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Ciência da Informação do Departamento de Ciência da Informação e Documentação da Universidade de Brasília como exigência parcial para a obtenção do Título de Mestre em Ciência da Informação.

ORIENTADORA: Profª Dra. Miriam Paula Manini

BRASÍLIA

2008

BODÊ, ERNESTO CARLOS

Preservação de Documentos Digitais: O Papel dos Formatos de Arquivo / Ernesto Carlos Bodê. Brasília: CID/Unb, 2008.

153 fl. (Dissertação de Mestrado). Orientadora: Profª. Drª. Miriam Paula Manini

1. Documentos digitais 2. Formatos de Arquivo 3. Preservação. I. Título

UNIVERSIDADE DE BRASÍLIA (UnB)
Faculdade de Economia, Administração, Contabilidade e Ciência da Informação e Documentação (FACE)
Departamento de Ciência da Informação e Documentação (CID)
Programa de Pós-Graduação em Ciência da Informação (PPGCInf)

FOLHA DE APROVAÇÃO

Título: "Preservação de Documentos Digitais: O papel dos Formatos de Arquivo".

Autor: Ernesto Carlos Bodê

Área de concentração: Transferência da Informação

Linha de pesquisa: Gestão da Informação e do Conhecimento

Dissertação submetida à Comissão Examinadora designada pelo Colegiado do Programa de Pós-graduação em Ciência da Informação do Departamento de Ciência da Informação e Documentação da Universidade de Brasília como requisito parcial para obtenção do título de **Mestre em Ciência da Informação.**

Dissertação aprovada em: 08 de dezembro de 2008.

Aprovado por:

Profa Dra Miriam Paula Manini
Presidente – Orientador (UnB/PPGCInf)

Profa Dra Marisa Bräscher Basílio Medeiros
Membro Interno – (UnB/PPGCInf)

Prof. Dr. Pedro Paulo Abren Funari
Membro Externo – (Universidade Estadual de Campinas)

Prof. Dr. Dibio Leandro Borges
Suplente – (UnB/CIC)

Dedico:

À minha família.

AGRADECIMENTOS

Agradeço a todos os professores, colegas e amigos que de diferentes maneiras contribuíram para o sucesso desse trabalho.

"Find more pleasure in intelligent dissent that in passive agreement, for, if you value intelligence as you should, the former implies a deeper agreement than the latter."

Um dos 10 mandamentos de **Bertrand Russel**

LISTA DE ABREVIATURAS E SIGLAS

AAF	Advanced Authoring Format
ASCII	American Standard Code for Interchange of Information
CEDARS	Exemplars in Digital Archives Project
DOC	Extensão Formato Microsoft para editor de texto
DRS	Digital Repository Services
GIF	Graphic Interchange Format
HTLM	Hyper Text Language Markup
IANA	Internet Assigned Numbers Authority
ISO	International Standard Organization
JPG	Joint Photographic Experts Group
NBR	Sigla de Normas Brasileiras
NEDLIB	Networked European Deposit Library
NLA	National Library of Australia
OAIS	Open Archival Information System
OCLC	On Line Computer Library Center
PBS	Public Broadcasting Service
PDF	Portable Document Format
PDF/A	Portable Document Format/Archiving
PUID	Pronom Unique Identification
RLG	Research Library Group
SAAI	Sistema Aberto para Arquivamento de Informação
SQL	Structure Query Language
TAR	Extensão de Formato muito utilizado em ambiente Linux
TIFF	Tagged Image File Format
UNICODE	Universal Code (tabela de códigos para armazenamento)
UPF	Universal Preservation Format
ZIP	Extensão de Formato para compactação de arquivos

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE FIGURAS

SUMÁRIO

RESUMO

A dissertação refere-se a uma pesquisa sobre preservação de documentos digitais com enfoque específico na relação entre formatos de arquivo e a efetiva preservação por longos períodos. O universo de pesquisa limita-se ao poder judiciário brasileiro. A estrutura do trabalho está dividida em basicamente três partes. A primeira delas corresponde a uma introdução ao projeto e uma revisão bibliográfica sobre temas pertinentes à preservação digital. A segunda parte corresponde à conceituação dos elementos teóricos essenciais ao desenvolvimento da pesquisa e inclui o próprio conceito de documento, preservação digital e metadados. A terceira e última parte corresponde à metodologia de coleta de dados, incluindo o universo de coleta correspondente e a análise de dados coletados. Finalmente, a dissertação apresenta uma série de conclusões e observações sobre os formatos de arquivos efetivamente utilizados no poder judiciário brasileiro em seus aspectos qualitativos para preservação digital.

Palavras-chave: Documentos digitais, formatos de arquivo, preservação digital, metadados.

ABSTRACT

The present research report is about digital preservation and it focuses on the relation between file formats and preservation for long term. The report is structured into three mains parts. The first one is made of an introduction to the report and a bibliographic revision on digital preservation issues. The second part is made of a development of related concepts used in the report, and includes the document concept, digital preservation and metadata. The third and last part explains the methodology of data collected, explaining the universe for collection and its analyze. Finally, in the last chapter we can find general conclusions about all the report.

Key-words: Digital documents, file formats, digital preservation, metadata.

1 INTRODUÇÃO

Entre tantas novidades boas e não tão boas, a contemporaneidade trouxe-nos o advento do documento digital. Nem todo registro de informações que utiliza a eletrônica para gravação e reprodução faz uso da tecnologia digital, ou seja, nem todo documento eletrônico é digital, veja-se o caso dos discos em vinil[1]. De qualquer forma, os documentos digitais vêm, cada vez mais, assumindo uma posição de destaque em vários aspectos da vida moderna: é o caso da *fotografia digital* ou dos *arquivos de imagens* gerados no processo de digitalização de documentos em suporte papel[2]. As disciplinas que utilizam documentos como matéria-prima de trabalho - como a história, a biblioteconomia, a arquivologia e tantas outras - não poderiam deixar de ser afetadas pela presença do documento digital.

Um dos problemas mais instigantes que se apresenta em função da existência do documento digital é sua preservação. Aqui cabe uma distinção entre os termos preservação, conservação e restauração. Segundo Muñoz Viñaz, o termo *conservação* pode ser entendido num sentido restrito em oposição à idéia de *restauração*, ou seja, atividades para manter (*keep*) o original ou, num sentido mais amplo, significando a soma dessa primeira idéia e outras atividades possíveis relacionadas. O mesmo autor acredita que há uma confusão terminológica:

> A confusão surge porque nas línguas latinas como o italiano, espanhol ou francês, '*conservation'* num sentido mais amplo, traduz-se por '*restauro'* (italiano), '*restauración'* (espanhol') ou '*restauration'* (Francês), de maneira que as traduções dessas línguas para o inglês e vice-versa, são freqüentemente imprecisas. As coisas ficam ainda piores porque alguns autores e organizações usam diferentes sinônimos

[1] O conceito de documento eletrônico também é utilizado em sentido amplo, significando todo tipo de documento que utiliza tecnologia eletrônica para produção e reprodução.

[2] Há que se fazer uma distinção entre documentos digitais nascidos digitais e aqueles gerados a partir da digitalização de documentos tradicionais. A digitalização, atualmente, é um processo que se aplica para praticamente todos os gêneros documentais: imagem, som e texto, etc.

> para '*conservation'* num sentido amplo, como o termo '*preservation'* e até mesmo '*restoration'*. (MUÑOZ VIÑAZ, 2005, p. 14, tradução nossa[3]).

Nesse texto, utilizaremos o termo **preservação**, preterindo o termo **conservação**, seguindo assim uma tendência entre os autores que publicam sobre **preservação digital**. O sentido do conceito de preservação que empregamos aqui é próximo ao que Muñoz Viñaz chama de sentido amplo do termo 'conservation', ou seja, diversas atividades que podem ser feitas para assegurar a **integridade** e o **acesso** aos documentos pelo maior prazo possível, idealmente para sempre. Uma excelente definição de preservação de documentos digitais foi exposta por Conway: *"Preservação* [preservation] *é a aquisição, organização e distribuição de recursos a fim de que venham a impedir posterior deterioração ou renovar a possibilidade de utilização de um seleto grupo de materiais"* (CONWAY, 2001, p. 14)

Um pesquisador atento ao problema da **preservação** de documentos digitais pode se preocupar com diferentes expectativas de vida para eles. Diferentemente de documentos em papel de boa qualidade ou o microfilme de guarda permanente, documentos digitais podem se tornar imprestáveis em uma década ou menos se os devidos cuidados não forem aplicados, sobre isso: *"Durante o século XX, a permanência, durabilidade e a resistência dos mais recentes meios de registro, com exceção do microfilme, continuaram a declinar"* (SEBERA, 1990, apud CONWAY, 2001, p.13).

Percebe-se então que mesmo documentos digitais que precisam ser mantidos por algumas décadas por motivos administrativos, contábeis ou fiscais, podem não durar o suficiente para cumprir sua função original. No entanto, o problema certamente é bem mais sério quando nos referimos aos documentos digitais que necessitam ser mantidos por séculos

[3] Todas as fontes bibliográficas utilizadas nessa pesquisa estão no idioma inglês, predominantemente, e português. Além disso, todas as traduções de textos originais em inglês foram feitas pelo autor e para simplificação omitiremos o termo "tradução nossa".

à frente, tanto quanto for possível, para as gerações futuras. Esses documentos compõem um legado cultural e histórico para a humanidade. Nessa pesquisa, nossa atenção se volta para a preservação dos documentos digitais de cunho **histórico e cultural** e que, por isso, necessitam de **guarda permanente**.

Há que se distinguir também, no que diz respeito aos documentos digitais, por um lado, os aspectos relacionados à preservação dos **suportes físicos** utilizados, como CDs e fitas magnéticas e, por outro lado, o próprio **conteúdo informacional** existente nos documentos. Tomemos como ilustração uma reportagem fotográfica histórica que utiliza a tecnologia digital: as filmagens no atentado de 11 de setembro nos EUA. Tais imagens foram gravadas e (re)gravadas em inúmeros suportes: CDs, discos em servidores de rede na Internet, fitas magnéticas, e etc. Cada um desses suportes documentais tem suas próprias necessidades de preservação, as quais, aliás, são muito relevantes, pois sua vida útil costuma ser bem pequena; sem mencionar o fato de que são suportes físicos muito mais frágeis que o papel, por exemplo. Portanto, um mesmo conteúdo informacional pode estar presente em diferentes suportes físicos, concomitantemente ou não. Esse **conteúdo informacional** - imagens no exemplo citado - também apresenta seus próprios problemas do ponto de vista da preservação por longos períodos.

Nesse projeto, nosso escopo compreende os **objetos digitais** que codificam conteúdos como imagens em movimento ou fixas, texto, som ou uma combinação desses elementos. Não estamos preocupados, nesse trabalho, portanto, com a preservação de suportes físicos[4] utilizados nos documentos digitais.

---

[4] Trataremos ainda de suportes fisicos no capítulo dedicado à revisão bibliográfica.

Por outro lado, indiretamente, nosso trabalho afeta a preservação de documentos em suportes tradicionais, aqueles nos quais não é possível uma separação entre conteúdo e suporte físico, como livros em papel, mapas tradicionais, e etc. A intersecção entre a preservação de documentos em suportes tradicionais e a preservação de objetos digitais ocorre em função do processo de **digitalização**. Em si, esse processo tem sido utilizado como vetor da preservação, pois os objetos digitais gerados atualmente podem conter uma alta fidelidade aos originais, o que permite poupar o acesso direto e o manuseio dos originais. Além disso, **caso se obtenha êxito na preservação desses objetos digitais**, é possível que esses persistam mesmo após a inevitável degradação física dos suportes utilizados nos documentos tradicionais, como o papel comum, os diferentes tipos de papel fotográfico, a película cinematográfica, e etc. Sobre o processo de digitalização e os cuidados com os objetos digitais gerados, Paul Conway observa que:

> Imagens digitais estão se tornando realmente comuns em bibliotecas e arquivos. A qualidade dos produtos de imagem digital pode ser espetacular. Há pouca dúvida de que a qualidade irá melhorar acompanhando a maturidade da tecnologia. Organizações estão reorganizando orçamentos, arrecadando dinheiro e antecipando receitas para fazer os projetos digitais acontecerem. Pode alguma instituição – bibliotecas, arquivos, sociedades históricas ou museus – arcar com o desperdício desse investimento? Sem um esforço sério que assegure o acesso por longos períodos dos arquivos digitais de imagens, porém, o risco de perdas é muito grande. (CONWAY, 2000)

Um outro aspecto que também relaciona a preservação de objetos digitais aos documentos tradicionais é a possibilidade de restauração dos últimos, tomando-se como referencial a imagem dos primeiros:

> Considerar um repositório digital de artefatos culturais não apenas como uma ferramenta educacional e de história da arte, mas também como uma poderosa ferramenta de restauração, implica que, além das informações visuais (imagens, raios-x, e etc.) e informações textuais/metadados simples, uma abundante quantidade de dados para pesquisa/restauração deveriam ser armazenados no repositório. (DELOS-NSF, 2002)

Os objetos digitais aos quais nos referimos nesse trabalho são constituídos por **dígitos binários**. Qualquer objeto digital, em última análise, independentemente do tipo de conteúdo (texto, som, imagem, e etc.) ou do tipo de suporte físico onde será gravado (disco rígido, fita

magnética, e etc.) será sempre composto por um conjunto de números binários. Esse conjunto somente é legível através de mecanismos de *hardware* e *software* apropriados. Mesmo assim, esses dois mecanismos só podem interpretar esses dígitos através de um enunciado que "explica" o significado desses *bits*. Por exemplo, é preciso indicar se um trecho de *bits* corresponde à data de gravação do arquivo, o tipo de arquivo ou parte do texto (se tratar-se de um arquivo de texto) ou parte do som (caso se trate de um arquivo de som). Esse enunciado é conhecido como **Especificação do Formato de Arquivo**, ou simplesmente **Formatos de Arquivo** (*File Formats*).

Não tentaremos desenvolver um aprofundamento técnico sobre o que são formatos de arquivo e suas especificações, pois isso foge ao escopo dessa introdução. Há um capítulo na dissertação dedicado inteiramente à definição aprofundada sobre formatos de arquivo. Por ora, podemos trabalhar com a seguinte definição operacional para esse conceito:

Uma **especificação de formato de arquivo** – normalmente chamada formato de arquivo simplesmente – é a explicação, normalmente registrada num documento formal, da disposição dos *bits* de um arquivo digital e a função desses *bits* ou grupos de *bits*. Por exemplo, uma especificação de um formato de arquivo X que gerou um arquivo digital Y onde os dezesseis primeiros *bits* gravados são **0100101101001111** orienta a quem necessitar que essa seqüência de *bits* (*bitstream*) corresponde a um cabeçalho (*filehead*) que registra o tipo e a versão do formato de arquivo em questão. Exemplos de especificações de formato de arquivos são o *Portable Document Format* (pdf) e o *Graphic Image Format* (GIF).

1.1 PROBLEMA E JUSTIFICATIVA

Nossa pesquisa orbita em torno do conceito de **formato de arquivo**, **identificando** as características mais adequadas que subsidiem a escolha de determinado formato de arquivo para a preservação de guarda permanente e efetuando um **levantamento** dos formatos de arquivo efetivamente em uso, dessa forma, **diagnosticando** o quadro atual no que diz respeito

aos efeitos da preservação de documentos digitais para as gerações futuras, pelo menos no que cabe à problemática das especificações de formatos de arquivo.

Nesta dissertação, o **problema** tratado pode ser assim apresentado: as características dos **Formatos de Arquivo** efetivamente utilizados nos documentos digitais da **Administração Pública Brasileira**, de guarda permanente, são adequados para a preservação por longos períodos?

Quando definimos o universo da pesquisa composto pela administração pública brasileira, na verdade planejamos uma amostra deste. Como delimitação do universo de pesquisa, restringir-nos-emos aos **documentos digitais** utilizados no **Poder Judiciário Brasileiro**. A escolha desta amostra está relacionada a fatores importantes. Primeiro, o poder judiciário é bem delimitado e estruturado, de forma que esperamos uma padronização maior nos procedimentos utilizados que envolvam tecnologias como formatos de arquivo. Como exemplo disto, a comunicação entre tribunais precisa ocorrer com base em padrões definidos. Com base numa pesquisa exploratória inicial, já identificamos algumas iniciativas neste sentido. Segundo, não existem motivos para acreditar que as opções tecnológicas do poder judiciário sejam consideravelmente diferentes dos outros poderes nacionais. Por último, dentre os poderes, o judiciário tem disponibilidade de **recursos e orçamento**[5] (BRASIL, 2007) que possibilita a utilização plena de tecnologia de ponta, a qual é um fator vital em nosso trabalho.

A melhor justificativa para esse trabalho está na importância da memória para uma sociedade. Como Donald Waters definiu em relação ao papel das bibliotecas e da própria universidade:

---

[5] Vide Relatório do TCU com pareceres prévios sobre as contas do governo da república referente ao ano de 2006.

> Eu afirmaria que a missão da universidade e da **biblioteca** é produzir cidadãos cultos. A função ampla da universidade dando suporte a essa missão, incluindo a preservação do conhecimento, está sendo mantida, mas os meios da comunicação acadêmica pelos quais a universidade efetua essas várias funções estão hoje em mutação. A comunidade acadêmica precisa se ajustar às mudanças nos meios de comunicação e porque os **programas de preservação** são, por definição, o principal mecanismo para renovar os ativos da universidade e da **biblioteca**, eles podem e devem ajudar nos necessários ajustes. (WATERS, 1998, p.100, grifos nossos)

Em consonância com essa linha de pensamento, as grandes bibliotecas vêm desenvolvendo programas voltados para a preservação de documentos digitais e, mais especificamente, preocupadas também com o problema dos formatos de arquivo. A *British Library* mantém um programa de preservação digital com vários projetos, muitos deles levados a cabo com outras instituições[6]. Aliás, considerando o custo de pesquisa em preservação digital, tem-se defendido o trabalho em cooperação:

> O fato de que a preservação digital é cara, os fundos são escassos e as responsabilidades são difusas sugere que as atividades de preservação digital se beneficiam da cooperação. Cooperação pode incrementar a capacidade de produtividade de um suprimento limitado de fundos de preservação digital através do compartilhamento de recursos, eliminando redundâncias e explorando a economia de escala. (LAVOIE, DEMPSEY, 2004)

Nos EUA, a *Library of Congress* também mantém diversos projetos especificamente sobre preservação digital: "Em muitos casos, materiais digitais são considerados mais frágeis que seus correspondentes físicos. Os arquivos em si podem facilmente ser destruídos ou armazenados em um formato que se torne obsoleto"[7].

Entre tantas instituições de renome mundial, a biblioteca da **Universidade de Harvard** mantém um programa específico para tratar do problema dos formatos de arquivo: o projeto JHOVE[8] que tem como objetivo propiciar, hoje, para as gerações futuras as funções

---

[6]Pode-se conhecer melhor os programas de preservação digital da Britsh Library em < http://www.bl.uk/aboutus/stratpolprog/ccare/introduction/digital/index.html >.

[7]Acessado em 15/04/2008. Disponível no sítio da Library of Congress: < http://www.digitalpreservation.gov/you/digitalmemories.html >.

[8] JHOVE, JSTOR/Harvard Object Validation Environment, "Format-Specific Digital Object Validation," 2004. Disponível em < http://hul.harvard.edu/jhove/index.html >.

de **validação**, **identificação** e **caracterização** de formatos de arquivo (*representation format*): "As ações de identificação, validação e caracterização são freqüentemente necessárias durante a operação de rotina de repositórios digitais e para a preservação digital"[9].

Com relação, especificamente, ao poder judiciário brasileiro, esse é detentor de um imenso acervo de documentos que registram uma parte significativa da memória do povo brasileiro. O tema, inclusive, já vem sendo tratado na pós-graduação em Ciência da Informação com o trabalho "Informação histórica: recuperação e divulgação da memória do poder judiciário brasileiro" (MANINI; MARQUES, 2007), apresentado no VIII Encontro Nacional de Pesquisa em Ciência da Informação. Sobre a importância da preservação da memória do poder judiciário brasileiro, Gunter Axt, muito apropriadamente, defende:

> Se a prática judicante é condição indispensável para a plenitude da cidadania no estado democrático de direito, então, dentre as missões do Poder Judiciário deve estar também a de comunicar didaticamente a função da Justiça para o povo, bem como os caminhos que estão disponíveis para, por meio da Justiça, garantir na prática os direitos da cidadania. Recomendável, portanto, que o Poder Judiciário, dentre outras estratégias, busque iniciativa no sentido de propiciar uma inserção positiva nos espaços de memória coletiva, seja criticando construtivamente os existentes ou criando novas inserções" (AXT, 2002, p. 226)

O poder judiciário brasileiro tem mostrado sinais de estar atento a essa importante parcela da memória nacional. Diversos "centros de memória" têm surgido em vários estados brasileiros como o **Memorial do Judiciário do Rio Grande do Sul** ou o **Museu do Tribunal de Justiça São Paulo**, ambos em pleno funcionamento. Além disso, podemos encontrar normas internas em órgãos do judiciário precisamente para oficializar ações de gestão e preservação de documentos, sendo que os documentos de caráter histórico estão merecendo especial atenção. Um exemplo nesse sentido é a resolução administrativa do Tribunal Regional do Trabalho da Décima Nona Região no qual fica clara a atitude de

---

[9] Disponível em < http://hul.harvard.edu/jhove/index.html >.

cuidado com documentos através de seu artigo 21 "As eliminações de processos judiciais serão decididas pelo Tribunal Pleno mediante proposta circunstanciada da Presidência deste Regional" (TRT, 2004).

Uma visita a esses centros de memória nos faz perceber que os documentos presentes são majoritariamente textuais e em suportes tradicionais como o papel. Por outro lado, tem surgido em todos os órgãos da justiça no Brasil uma série de projetos que buscam a substituição de documentos tradicionais por documentos digitais. O desejo de modernização ao lado da busca por mais eficiência e eficácia são as motivações de projetos encontrados em praticamente todos os sítios de órgãos da justiça no Brasil[10] como o acesso a Diários Eletrônicos, emissão de Certidões e até mesmo o próprio processo judicial na versão digital.

Dessa forma, esses novos documentos e suas particularidades precisam ser compreendidos para que se possa efetivar sua preservação para as gerações futuras. É nesse sentido que esta dissertação com enfoque no problema dos formatos de arquivo se insere e busca colaborar com uma solução adequada.

1.2 OBJETIVOS

Nosso objetivo principal é a verificação da adequação ou não dos formatos de arquivo, efetivamente utilizados em documentos digitais no **Poder Judiciário Brasileiro**, para as melhores práticas de preservação digital. Para esse fim, obteremos uma amostra significativa de arquivos utilizados em documentos digitais no referido poder constitucional, amostra esta constituída pelas mais diferentes especificações e versões de formatos de arquivo. Pretende-se que essa amostra contenha os formatos de arquivo utilizados por **documentos digitais** em geral, incluindo os documentos digitais que eventualmente venham a ser selecionados como

[10] Para uma relação rápida de endereços de órgão judiciais procure *links* no endereço http://www.stj.jus.br.

de **Guarda Permanente** após o correspondente processo de **seleção documental**. Cada um dos formatos de arquivo encontrados nessa amostra será avaliado em relação a um **Modelo de Formato de Arquivo** tomado como referência o que permitirá o diagnóstico e análise sobre a utilização dos **Formatos de Arquivo**.

Os objetivos específicos da pesquisa são:

a – Criação do **Modelo** de formato de arquivo referência;

b – Coleta dos dados no universo de pesquisa definido e criação da Amostra Final;

c – Comparação dos formatos de arquivo identificados (**Amostra Final**) com o **Modelo**.

1.3 ESTRUTURA DO TRABALHO

Esta dissertação está estruturada, basicamente, em três partes.

A **Parte I** compreende essa **Introdução**, o capítulo sobre os **Pressupostos Filosóficos e Científicos** e nossa **Revisão Bibliográfica** sobre o documento digital e vários outros conceitos relacionados à sua preservação.

A **Parte II** compreende os capítulos referentes à conceituação terminológica que utilizamos, desde um capítulo voltado ao conceito de **Documento**, passando por um estudo sobre o conceito de **Formatos de Arquivo**, um outro sobre **Metadados para Preservação** e finalmente o capítulo que sintetiza nosso **Modelo de Formato de Arquivo** para preservação.

A última parte compreende todos os capítulos dedicados à coleta de dados e sua análise. No capítulo **Coleta de Dados,** iniciamos introduzindo nossa **metodologia de coleta**, definição do **Universo da Coleta** e conceitos utilizados diretamente na coleta. Em seguida, expomos os **Dados Coletados** e, finalmente, a **Análise** efetuada sobre esses.

Cada capítulo inicia com um pequeno texto introdutório e finaliza com as **Últimas Considerações** sobre o capítulo em questão; além de informações que não tenham sido

desenvolvidas antes. Ao final, dedicamos um capítulo às **Conclusões**, nesse capítulo tentaremos expor nossas descobertas **dentro dos limites estabelecidos**: **pressupostos** que escolhemos, metodologia, a **teoria científica** relacionada e os **dados** por nós **coletados e analisados**.

Acredito ser importante tecer também considerações sobre o conteúdo objeto dessa pesquisa. Essa dissertação aborda um problema (formatos de arquivo) que é essencialmente um produto tecnológico, criado pela tecnologia, embora seus efeitos ultrapassem em muito os limites tecnológicos. No entanto, desenvolvemos a pesquisa no contexto de um departamento de Ciências da Informação, não no campo da Ciência da Computação, mais técnico ou tecnológico.

Em função disso, ao longo do trabalho sempre tive a preocupação de "traduzir" da melhor maneira possível os termos técnicos utilizados, até porque tenho muita experiência profissional na área de **tecnologia da informação e comunicações** (**TICs**) o que naturalmente afeta a maneira de lidar com esse texto. O capítulo mais árduo no que diz respeito a termos técnicos de tecnologia é aquele dedicado ao desenvolvimento e explicação do **Modelo** de referência.

2 DISCUSSÕES RECENTES SOBRE PRESERVAÇÃO DIGITAL

No Brasil, no âmbito acadêmico, identificamos alguns trabalhos de **Pós-Gradução**[11]. Citados por Thomaz (THOMAZ, 2004, p. 68) estão a dissertação de Anna Carla A. Mariz, de 1997 (Unirio); de 2001 a dissertação de Vanderlei B. dos Santos (PPGCI/Unb), em 2002 as dissertações de Rosely C. Rondinelli (PPGCI/UFF)[12] e Emília B. Cruz (PPCGI/UFMG) e a tese de Kátia P. Thomaz, em 2004 (PPGCI/UFMG), todos abordando o documento digital, sendo que as duas últimas dissertações e a tese abordam mais especificamente a questão da preservação digital, sob diversos aspectos. Mas nenhum trabalho teve como foco principal o aspecto **formatos de arquivos** para preservação.

Além desses trabalhos, identificamos também, no âmbito das engenharias, duas outras dissertações: a de Luis Felipe Lopes, em 2001[13] (Engenharia de Produção/UFSM) e a de Humberto C. Innarelli, em 2006 (Engenharia Mecânica Computacional/Unicamp). Estas duas dissertações tiveram como foco a preservação do documento eletrônico sob o prisma do *suporte documental* utilizado, principalmente CDs e DVDs.

Em 2007, tivemos a oportunidade de apresentar um trabalho no **Seminário Internacional de Bibliotecas Digitais,** em São Paulo, com o título **Preservação de Coleções de Documentos Digitais** (BODÊ, 2007). Com base nesse artigo, reproduzimos adiante os principais problemas que identificamos e algumas indicações bibliográficas que tratam mais detidamente desses problemas. As referências não são exaustivas, mas acreditamos que a maioria dos autores citados, são hoje referência nessa área de pesquisa. Nessa revisão

---

[11] Além dos trabalhos aqui relacionados e lançados na revisão bibliográfica, a maioria pode ser encontrada em bibliotecas de dissertações e teses, como o portal *Scielo*.

[12] As dissertações de Vanderlei B. dos Santos e Rosely C. Rondinelli parecem ter sido o ponto de partida para as publicações nacionais: **Gerenciamento arquivístico de documentos eletrônicos: uma abordagem teórica da diplomática arquivística contemporânea** (RONDINELLI, 2002) e **Gestão de documentos eletrônicos: uma visão arquivística** (SANTOS, 2005).

[13] Origem do livro **A qualidade dos suportes no armazenamento de informações** (LOPES; MONTE, 2004).

focamos mais que a própria questão dos formatos de arquivo (item 5) pois os diversos problemas se inter-relacionam.

2.1 ATUALIZAÇÃO TECNOLÓGICA DE *HARDWARE* E *SOFTWARE*

Sabemos que não teremos acesso ao conteúdo dos objetos digitais senão através de máquinas (leitoras de mídias e computadores). Precisamos também de todo o *software* necessário e relacionado para que um computador possa funcionar adequadamente. Assim, podemos dizer que os objetos digitais têm uma forte dependência com todo esse aparato. Manter os primeiros em condições de uso para acesso futuro implica em cuidados com os últimos.

Porém, nenhum sistema composto de *hardware* e *software* durará mais que algumas décadas (já se fala em anos). A obsolescência tecnológica que temos verificado implica na falta de peças de reposição e técnicos capazes de reparar estes equipamentos (veja o caso das máquinas de escrever de algumas décadas atrás ou os primeiros computadores fabricados).

A melhor solução parece ser ficar atento para este processo e não ignorá-lo; as atualizações devem ocorrer permanentemente para mitigar o processo de obsolescência tecnológica. Sobre a obsolescência de *software*, falaremos mais sobre isto no item *integridade dos conteúdos*.

Pode-se ter acesso a mais informações e propostas de ação sobre *hardware* e *software* no capítulo 2 do livro *on-line* disponibilizado pela *Digital Preservation Coalision* (DPC, 2006) ou nos textos disponibilizados pelo projeto *InterPares*[14].

[14] http://www.interpares.org.

2.2 DETERIORAÇÃO DOS SUPORTES

Todo material físico passa por um processo de desgaste em função do tempo e possui uma vida útil determinada. Em condições ideais de temperatura, umidade relativa e iluminação é possível prolongar ao máximo esta vida; no entanto, todos sucumbirão. Na prática, é muito difícil manter documentos e seus suportes físicos em condições ideais de guarda, principalmente por longos períodos. Até porque, quando estes documentos estão no início de seu ciclo de vida, ainda em uso administrativo, na maioria das vezes não há como impor condições nem mesmo adequadas quanto mais ideais de armazenamento. Para reforçar estes problemas é preciso lembrar que os objetos digitais são infinitamente mais sensíveis que os documentos em suportes tradicionais. Sabemos da existência de documentos com milênios de idade, em argila, papiro ou pergaminho e estes documentos, apesar de seu péssimo estado de conservação, através da aplicação de técnicas adequadas, ainda podem ser lidos pelo homem. Já no caso dos objetos digitais, mínimas falhas em seu conteúdo podem invalidar todos os arquivos. Isto se deve às particularidades da tecnologia utilizada para criar os arquivos e agrupar os *bits*; grosso modo, o conjunto dos *bits* de um arquivo compõe uma estrutura que precisa ser respeitada integralmente.

Como danos ao suporte físico[15] podem danificar o conteúdo dos *bits* dos objetos digitais, os cuidados com o suporte físico são importantes. Os objetos digitais não guardam uma relação entre conteúdo e suporte físico indissociável (como veremos nos capítulos sobre documento e documento digital) e, portanto, podemos migrar o conteúdo para outros suportes

[15] Atente-se para o detalhe de que não necessariamente todo o espaço de armazenamento em uma mídia qualquer é utilizado. Assim, danos como arranhões em mídias podem não afetar o conteúdo dos documentos, desde que a área afetada não contenha dados gravados. Há também artifícios tecnológicos, que dentro de certos limites, podem recuperar uma parte do conteúdo (parte das seqüências de bits) perdido dos objetos digitais.

físicos. Mas isto precisa ser feito antes de haver danos ao suporte físico, pois após o conteúdo ter sido danificado não há mais o que possa ser feito: perdeu-se o documento.

Visto que não há algo que possa ser feito para impedir a deterioração dos suportes físicos dos objetos digitais, duas alternativas nos restam para mitigá-la: **1** – Estabelecer condições idéias de armazenamento e climatização e **2** – Estabelecer uma política de migração periódica de suportes.

No primeiro caso, há que se partir do levantamento dos suportes físicos utilizados, discos magnéticos, *compact discs* (CDs)[16], DVDs, fitas magnéticas, e etc. Cada uma destas categorias – e dentro delas há também variações entre diferentes fabricantes e modelos – possui suas próprias especificações que devem ser seguidas. No segundo caso, também para cada tipo de suporte, há que se determinar sua vida útil média e, claro, antes do fim da mesma, é preciso providenciar a troca deste suporte. Esta atividade deve ser feita conjuntamente com as preocupações com a atualização tecnológica do *hardware* responsável pela reprodução destes documentos, ou seja, procurar utilizar novas mídias com tecnologia atualizada, diminuindo os problemas com a falta de manutenção em equipamentos muito antigos.

Informações técnicas e detalhes específicos devem ser buscados junto aos fabricantes dos suportes físicos dos objetos digitais. Há também algumas publicações que podem ser consultadas. Sobre fitas magnéticas e material sonoro (que inclui CDs) pode-se consultar os trabalhos da coleção CPBA[17] publicados no Brasil em 2001, mas originais do início da década de 90 (Van BOGART, 2001) e (LAURENT, 2001). Há também um trabalho bem mais

[16] Observe-se que tanto os CD´s como os DVD´s possuem vários subtipos, como o CD-R ou CD-RW, por exemplo.

[17] CPBA é acrônimo de Conservação Preventiva em Bibliotecas e Arquivos, trata-se de um projeto que reúne um coleção de artigos em diferentes fascículos dedicados a vários aspectos da preservação e conservação de documentos em diferentes suportes.

atualizado especificamente para CDs e DVDs (BYERS, 2003). Nacionalmente, pode-se consultar uma dissertação de mestrado da Unicamp que apresenta uma metodologia para testes de confiabilidade em mídias to tipo CDs (INNARELLI, 2006). Há também o trabalho A qualidade dos suportes no armazenamento de informações (MONTE; LOPES, 2004).

2.3 INTEGRIDADE DOS CONTEÚDOS

O conteúdo dos objetos digitais, sejam eles do gênero textual, sonoro, imagético ou qualquer outro, será sempre gravado como seqüências de zeros e uns (*bits*). Após a transferência destes *bits* para a memória do computador (através de todo o *hardware* associado), será necessário *software* para interpretar e traduzir, num modo compreensível aos humanos, os conteúdos. Ocorre que os *softwares* também sofrem um processo de defasagem tecnológica e, assim como o *hardware,* estão em constante modernização. Esta "atualização" trás em seu bojo um problema: ler o conteúdo de um objeto digital muito antigo pode requerer o *software* antigo que foi utilizado e que pode já não estar mais disponível. Mais ainda, um determinado aplicativo, digamos um editor de texto, foi projetado para funcionar em um determinado sistema operacional, que, por sua vez, foi projetado para funcionar em determinado tipo de *hardware*. Assim, a necessidade de uso de um *software* antigo requer todo um aparato de outros *softwares* e *hardware*s específicos. Esse quadro pode inviabilizar o acesso a objetos digitais muito antigos.

A solução mais evidente é estabelecer uma política de monitoramento e constante atualização dos objetos digitais garantindo que sempre possam ser lidos no futuro. Este é o processo de migração dos conteúdos de objetos digitais.

Sobre migração de *software* ver o capítulo 4 do livro on-line editado pelo *Digital Preservation Coalition* (DPC, 2006). Também pode-se encontrar informações no trabalho Preservação no Universo Digital (CONWAY, 2001). O livro digital do português Manuel Ferreira traz uma série de estratégias contra a obsolescência de software (FERREIRA, 2006).

2.4 FIDEDIGNIDADE DOS CONTEÚDOS

É claro que não basta manter os objetos digitais intactos ao longo do tempo. É preciso lançar mão de estratégias para manter o acesso ao conteúdo dos mesmos e, desta forma, possibilitar a contínua leitura destes documentos. Falamos anteriormente de cuidados para manter o suporte físico que mantém os dados íntegros e procedimentos de migração de dados que permitirão o contínuo acesso aos mesmos pelos softwares. Mesmo com todos estes cuidados e o sucesso destas estratégias, estes objetos digitais ainda poderão não ter seu conteúdo fidedigno. Há uma diferença sutil mas de grande importância entre manter a integridade funcional dos conteúdos dos objetos digitais e garantir que estes conteúdos sejam fidedignos, ou seja, representem realmente o que originalmente foi gravado nos mesmos.

Em coleções de documentos em suportes tradicionais este problema praticamente não se evidencia pois, por comparação, é fácil verificar se o conteúdo de um documento não foi alterado; basta, por exemplo, comparar o conteúdo de dois exemplares de um mesmo livro ou dois artigos do mesmo número de um periódico. Normalmente, eventuais alterações seriam facilmente detectáveis. No mundo digital o problema é mais delicado. Alterações em documentos digitais não podem ser facilmente detectáveis. Se não houver travas de segurança que impeçam estas alterações, a princípio não será possível verificar se houve alteração ou o que foi alterado.

Para garantir a fidedignidade de objetos digitais é necessário dispor de recursos de segurança, por *hardware* ou *software*, que impeçam alterações nos documentos ou que pelo menos indique se houve alterações. Por exemplo, documentos em um CD-R não podem fisicamente ser alterados, pelo menos não enquanto estiverem gravados neste CD-R. Parece que a única maneira de confirmar se houve alteração em um documento e o que foi alterado é a comparação deste documento com outro exemplar sabidamente fidedigno.

Para compreender melhor o conceito de fidedignidade pode-se estudar os textos produzidos pelo projeto InterPares, principalmente aqueles do projeto InterPares2[18] Há também o livro *Preservation of the integrity of Electronic Records* (DURANTI, 2002), apesar de focar os documentos de arquivo, os conceitos chave aplicam-se a qualquer tipo de documento.

2.5 AUTENTICIDADE DO CONTEÚDO

A característica da autenticidade de um objeto digital refere-se à comprovação de autoria daquele documento, ou seja, confirmar quem ou qual organização criou o documento. Sem mecanismos de confirmação da autoria de um objeto digital, sua credibilidade pode ser questionada também, notadamente quando se trata de um documento que comprove ações de indivíduos ou trabalhos de cunho literário, por exemplo.

É possível verificar a autenticidade de um objeto digital através de vários mecanismos, como o *lay-out* utilizado, tipos de fontes, vocabulário de época. E há ainda recursos de assinatura digital. Em geral, a análise de autenticidade de um documento qualquer, inclusive um objeto digital, não é simples e exige um considerável estudo e esforço intelectual.

Para se compreender melhor a extensão desta importantíssima característica de um objeto digital adequadamente preservado pode-se consultar o livro *Trusting Records* (MACNEIL, 2000), também com foco nos documentos de arquivo mas com conceitos aplicáveis a qualquer documento. Há também o artigo *Can Bits and Bytes be Authentic?* (HOFMAN, 2002). Sobre assinaturas digitais pode-se consultar o artigo *Assinaturas Digitais e a Arquivologia* (BODÊ, 2006).

---

[18] http://www.interpares.org.

2.6 FORMATOS DE ARQUIVO

Para cada Formato de Arquivo produzido por determinado *software*, existirá uma especificação técnica. Na verdade, haverá também uma especificação para cada **versão** de um determinado formato, por exemplo, a especificação **TIFF 5.0** e a **TIFF 6.0**, cada uma com seu detalhamento técnico. Dependendo do Formato de Arquivo, tal especificação técnica pode ser extremamente diferente para cada versão de um mesmo formato.

As especificações de cada Formato de Arquivo são de caráter bastante técnico e estão no escopo de desenvolvedores de *software* em geral. Estas especificações técnicas explicam, detalhadamente, como as seqüências de *bits* no arquivo devem ser estruturadas e onde cada tipo de dado deve ser gravado. Para cada formato de arquivo haverá diferenças marcantes entre as especificações.

Um ponto crucial sobre Formatos de Arquivo e que está diretamente ligado aos problemas com sua preservação se refere ao fato de se tratar de um formato proprietário ou não. Os formatos abertos de arquivo (aqueles em que o público tem acesso aos detalhes técnicos) são mais adequados para a preservação futura pois as possibilidades de compreender o significado de sua estrutura de *bits* é maior.

Existem várias propostas para tentar manter, no futuro, o acesso às informações de um documento gravado através de um determinado Formato de Arquivo, como a **emulação** e a **migração**[19]. De qualquer forma, todas as propostas dependem do conhecimento sobre Formatos de Arquivo para que possam ser executadas com sucesso em maior ou menor grau.

Para se compreender melhor a questão dos formatos de arquivo pode-se consultar o artigo *The bits and bites of data formats* (ASCHENBRENNER, 2004) ou *Selecting file*

[19] Ver o trabalho de Manuel Ferreira (FERREIRA, 2006).

formats for long-term preservation (BROWN, 2003). Adiante teremos um capítulo dedicado inteiramente ao conceito de Formato de Arquivo.

3 O DOCUMENTO

3.1 O DOCUMENTO TRADICIONAL

Nesse trabalho nos interessa o conceito de documento utilizado no seguinte campo semântico: “Qualquer base de conhecimento, fixada materialmente e disposta de maneira que se possa utilizar para consulta, estudo, prova, etc.” (FERREIRA, 1986, p. 605). Este conceito amplo é fundamental na história da cultura da humanidade e, mais acentuadamente ainda, em nossa pós-modernidade. Ganha contornos específicos dependendo dos diferentes pontos de vista. Assim, nas ciências ligadas à administração de organizações, por exemplo, ressalta-se seu valor na tomada de decisões e como fator de comunicação mais ou menos eficaz. Aqui, nos interessa o enfoque ligado à **História** e às ciências **Documentais** que se preocupam com a **organização e tratamento** do documento, entre outras coisas. Mais especificamente, nos interessa a característica, que documentos podem assumir, de se tornarem instrumentos para a preservação da memória de indivíduos, organizações e, em última análise, até mesmo da humanidade.

Uma definição mais adequada ao nosso contexto é a seguinte:

> Documento em um sentido bem amplo e genérico é todo o registro de informação, independentemente de seu suporte físico. Abarca tudo que pode transmitir o conhecimento humano: livros, revistas, fotografias, filmes, microfilmes, microfichas, folhas, transparências, desenhos, mapas, informes, normas técnicas, patentes, fitas gravadas, discos, partituras, cartões perfurados, manuscritos, selos, medalhas, quadros, modelos, facsímiles e, de maneira geral, tudo que tenha um caráter representativo nas três dimensões e esteja submetido à intervenção de uma inteligência ordenadora. (HEREDIA HERRERA, 1991, p. 122).

Por outro lado, há uma definição ainda mais precisa e estruturada, obtida a partir de diversos autores e instituições e resumida em tese de doutorado:

> Pode-se definir documento genérico como qualquer informação registrada independentemente do suporte utilizado, a qual pode ser tratada como unidade. No primeiro nível de desdobramento, é possível distinguir-lhe dois elementos constituintes, a saber: o suporte, o meio físico sobre o qual a informação é fixada; e a mensagem ou notícia veiculada. No segundo nível, a mensagem pode ser decomposta em outros três elementos, quais sejam: a estrutura sobre a qual a informação foi registrada, envolvendo cabeçalhos e outros dispositivos para identificar e rotular partes do documento, negrito, itálico etc.; o conteúdo, propriamente dito; e o meio de fixação desse conteúdo com possibilidades para o texto, o gráfico, a figura, a tabela, etc. (THOMAZ, 2004, p. 77).

Em todas as definições acima e em várias outras que podem ser encontradas em diversas áreas da literatura científica, dois aspectos merecem destaque.

O primeiro, relaciona-se à explicitação da presença e necessidade de um **suporte físico** para a existência de um documento qualquer. De fato, este é um ponto fundamental, principalmente quando se discorre sobre documentos digitais, área na qual têm surgido algumas confusões terminológicas. Na atualidade, com o uso de tecnologias de rede e acesso a documentos em nossas telas de computador, é possível que se tenha a sensação de que alguns documentos são como que etéreos, estão em algum tipo de espaço imaterial. De fato, não é difícil encontrar termos como *documento virtual*. É preciso que se esclareça, desde já, que não existe qualquer documento – eletrônico, digital ou de qualquer outra designação – que não esteja fixado em algum tipo de suporte físico, mesmo que em algum lugar numa rede de dados ou em algum tipo de memória interna de computador.

O segundo aspecto, é que documentos contêm informações e conhecimento e estes conteúdos representam, de diferentes maneiras e formas, a **memória**, que precisa ser mantida por mais ou menos tempo para cumprir diversas finalidades, dependendo de seu valor **administrativo**, **cultural** ou **histórico**. O mais importante, é que esta memória não está apenas em um tipo específico de documento, como o documento **arquivístico histórico** ou **manuscritos** escritos em pergaminhos da Idade Média. Esta memória está também inscrita em prosaicos livros modernos, gravações de áudio, fotografias, e etc. No romance de H. G. Wells, a *Máquina do Tempo*, recentemente produzido em versão cinematográfica[20], há uma belíssima demonstração de como prosaicos livros de uma biblioteca têm seu valor para a humanidade. O personagem viajante no tempo alcança 800 mil anos após o final do século

[20] *The Time Machine*. Direção Simon Wells. DreamWorks SKG / Warner Bros. 2002.

XIX e, nas cenas finais da versão do cinema, podemos ver o que resta da humanidade voltando a receber conhecimento de uma biblioteca do passado, diga-se de passagem em versão com **Acervo Digital**.

No entanto, trata-se de ficção científica e, como veremos mais adiante, os documentos digitais, pelo menos dentro da tecnologia que dispomos hoje e com os cuidados e políticas que têm sido implementados, não são tão duráveis assim e, na verdade, estão em sério risco de perda irremediável.

3.2 O DOCUMENTO DIGITAL

Até que a humanidade obtivesse sistemas de escrita completos como os atuais, ela fez uso de símbolos gráficos e mnemônicos de vários tipos para armazenar informações, sobre os mais antigos artefatos encontrados:

> Artefatos desenterrados em Bilzingsleben, Alemanha, datados de pelo menos, 412.000 anos atrás [...] foram interpretados por seus descobridores como entalhes intencionais (algum tipo de símbolos gráficos). É evidente que os entalhes são marcas; o que significam e se significam algo, não está claro. (FISCHER, 2003, p.16)

Segundo este mesmo autor, a humanidade utilizou, então, sistemas pictográficos (como as representações em cavernas). Em um segundo momento, passou a utilizar símbolos gráficos para representar coisas reais como bens e animais, até o grande salto da **fonetização**, quando um símbolo gráfico representa um som correspondente na linguagem local. Tal invenção surgiu na Mesopotâmia entre 6.000 e 5.700 anos atrás, aproximadamente.

A humanidade utilizou os mais diversos materiais como suportes para registros de informação; segundo HUNTER (1978), foram utilizados madeira, metais, pedras, troncos, tecidos, papiro (*Cyperus papyrus*), pergaminho e, finalmente, papel.

O uso de argila em tábuas é particularmente importante, pois, ao que parece, o primeiro sistema completo de escrita (por volta de 2.500 AC) utilizou este material como suporte.

No Egito, o uso do papiro (o mais antigo conhecido tem 3.700 anos) rivalizou com o uso de tábuas de argila. Na verdade, os egípcios desenvolveram diferentes sistemas de escrita para diferentes aplicações, rituais, contabilidade, e etc.; para cada aplicação havia um sistema de escrita e suportes específicos como paredes, metais preciosos, etc.

O uso do pergaminho também foi um fato importante para o registro de informações, "O rei de Pérgamo (197-159 AC) normalmente recebe os créditos pela invenção e acredita-se que esteja relacionada com o desejo de produzir um material de escrita que rivalizasse com o papiro egípcio" (HUNTER, 1978, p. 12).

Finalmente, surge o papel, que possibilitou um grande salto na produção de documentos, já que se tratava de um material de fácil fabricação e menor custo, além da qualidade em relação a outros suportes. A data normalmente atribuída à invenção do papel é a de 105 DC, na China (HUNTER, 1978, p.50).

Vários outros suportes foram utilizados para o registro de documentos e em determinados períodos históricos alguns competiram entre si, como o papel e o pergaminho. O tipo de papel próximo do que é utilizado hoje só existiu a partir do século XIX (DOCTORS, 1999).

Temos, então, até o século XIX, uma produção documental, registrada basicamente em papiro, pergaminho e papel, documentos com conteúdo textual, de diferentes naturezas, de inventários de bens até a Literatura e a Filosofia. Em meados do século XIX surge uma invenção que acrescenta uma nova diversidade aos acervos documentais: trata-se da fotografia, *"A invenção da fotografia foi anunciada oficialmente em 19 de agosto de 1839, pelo francês Louis Jacques M. Daguerre (1787-1851), sob a forma do daguerreótipo"* (SMIT & GONÇALVES, 2005, p.9). Esta invenção passaria por um processo de evolução tecnológica que culminaria, no final do século XX com o advento da fotografia digital, a qual, por si só, tornou-se uma nova revolução. Também com tecnologia bastante próxima dos

registros fotográficos, apesar da aplicação ser diferente, encontramos o microfilme como meio para registro documental, ainda hoje bastante utilizado.

No final do século XIX, vários inventos para registro de som culminaram, no início do século XX, com os discos com gravações sonoras e, logo depois, com o uso também de fitas magnéticas. Estas últimas, após um período de evolução, passam a ser utilizadas para gravação de vídeo (os primeiros programas televisivos gravados). No final do século XX surgiriam os *Compact Discs* (CDs), inicialmente para gravações de áudio, surgindo depois os modelos específicos para vídeo (DVD's).

O próximo grande passo seria dado pelo uso de computadores pela humanidade. Os primeiros computadores modernos apareceram na década de 1940. Embora haja muitas contribuições individuais para o avanço da tecnologia, esta cresceu e se desenvolveu, na América do Norte, especialmente *"graças à associação entre militares, universidades e firmas"* (KIDDER, 1981, p. 13). O uso cada vez maior de computadores inicialmente pelas grandes corporações, mas a partir da década de 1980 do século XX, também pelo cidadão comum representou um grande salto para o **registro**, o **armazenamento** e a **recuperação** de documentos. Estas máquinas, em função da exigência de cada vez mais espaço para registro de seus *bits* (codificação digital), passaram a utilizar diferentes tecnologias, desde as fitas magnéticas, passando por discos magnéticos, ópticos e diversos outros. Hoje, no início do século XXI, as novidades incorporadas ao conjunto de mídias são os tocadores de áudio, *pen-drives* e outros.

Apresentamos, na *tabela 1* comparativa entre as características dos diferentes documentos apresentados no breve histórico acima.

Documentos		
Período	**Suporte utilizado**	**Mensagem**
Antes 6.000 AC	Material disponível na natureza: ossos, cascas de animais, madeira.	Sinais, desenhos e marcas mnemônicas.
6.000 AC e Final séc. XIX	Material elaborado para uso específico: Argila, pergaminho, papiro e o papel.	O conteúdo da mensagem se apresenta estruturado mas apresentado sob a forma de texto ou ilustrações e pinturas.
Desde o final séc. XIX até Hoje	Material elaborado para uso específico com maior grau de sofisticação: papel moderno, películas, mídias magnéticas e ópticas.	Além do conteúdo estruturado, há uma miríade enorme de formas de apresentação além do textual: imagens fixas e em movimento, bancos de dados, planilhas e etc.

Tabela 1 - Fases de evolução dos documentos

A partir da análise das informações do histórico acima exposto, podemos tecer algumas considerações. Desde o final do século XIX começam a surgir os primeiros documentos, que podemos chamar, hoje, ainda impropriamente, de eletrônicos. Impropriamente, pois, na verdade, estas primeiras tentativas de armazenar conteúdo informacional além do texto[21], como o som e imagens em movimento, merecem a designação, neste período, de documentos legíveis por máquina[22] (fonógrafos e projetores de cinema, por exemplo), já que a eletrônica propriamente dita nem mesmo existia, vindo a se desenvolver ao longo do século XX. De fato, a característica de se **necessitar de máquinas** para se ter acesso ao conteúdo destes documentos é uma transformação tecnológica importante e que traz conseqüências para a preservação destes documentos, sendo a mais óbvia a necessidade de manutenção destas máquinas juntamente com seus documentos.

[21] A fotografia foi uma bem sucedida de registro de imagens reais, com característica bem diferenciada das pinturas da época, por mais realistas que fossem. Apesar de não necessitar de equipamentos tecnológicos para sua "leitura" e somente para sua produção hoje, com a fotografia digital, este quadro está mudando . A fotografia digital requer computadores e máquinas para sua visualização adequada. Também é importante notar que o microfilme (basicamente um processo fotográfico) só é legível através de equipamento específico para sua leitura.

[22] Observamos que, na atualidade, a característica *Legível para Máquinas*, por si só, não define que um documento seja eletrônico. O acesso a manuscritos antigos (digitalizados) no outro lado do planeta, via rede de dados e computadores (legível neste lado do planeta através destas máquinas), não transforma o manuscrito antigo em documento eletrônico (apesar da existência de uma cópia deste codificada digitalmente).

Mesmo com o desenvolvimento de suportes mais sofisticados para registro de informações, como os **discos de vinil** e o desenvolvimento da eletrônica propriamente dita, numa primeira fase, estes documentos ainda têm uma característica em comum com todos os outros até então produzidos: o conteúdo da mensagem não pode ser dissociado do suporte físico utilizado, não sem danificar o documento. Somente a partir do desenvolvimento das fitas magnéticas (para áudio ou vídeo) começa a surgir o fenômeno da **independência** entre o suporte e o conteúdo do documento, característica esta que trará conseqüências importantes do ponto de vista das ações de preservação para estes documentos, como indicaremos mais adiante[23].

Através do uso dos computadores, desde meados do século XX, desenvolvem-se as tecnologias digitais aplicáveis aos documentos. Neste ponto, passamos a ter o uso de eletrônica digital e armazenamento de conteúdos sob a forma de codificação digital[24], na atualidade qualquer tipo de conteúdo além do texto, como o som e imagens. Surge então o documento **eletrônico digital**. Nós defendemos que a terminologia mais adequada seja documento eletrônico e digital, já que há documentos eletrônicos que não utilizam tecnologia digital (como as fitas magnéticas com registros de história oral, em gravações eletrônicas analógicas) e há inclusive documentos com codificação digital (legíveis por máquinas eletrônicas) mas gravados em papel, como o caso dos cartões perfurados ou *punch cards* para entrada de dados em computadores *mainframes*, já ultrapassados.

---

[23] É curioso observar que alguns tipos de bases de dados (no todo e dentro das definições propostas podem ser consideradas um único documento) podem estar armazenadas em suportes físicos diferentes, estar divididas em diferentes discos em diferentes computadores, ou parte dos dados em fitas magnéticas ou discos ópticos *off-line* por exemplo (em prateleiras onde são inseridos no equipamento conforme a demanda de informações).

[24] Inicialmente dados computacionais, números e textos e, mais recentemente (final século XX) com o advento do que se convencionou chamar multimídia: som, imagens fixas e em movimento e a combinação de todos estes elementos.

Independentemente dos rótulos terminológicos utilizados, nos parece que o mais importante é definir as características essenciais destes documentos eletrônicos e digitais. Aqui, esclarecemos que este texto tem como escopo este grupo específico de documentos, ou seja, aqueles que surgiram no final do século XX e tiveram seu grande avanço qualitativo e quantitativo com o advento da microinformática. As referidas características essenciais são: *legibilidade por máquinas*, *independência entre suporte físico e sua correspondente mensagem* e, finalmente, o fato de serem *codificados em linguagem binária digital*.

Em se tratando da correta caracterização de documentos eletrônicos e digitais, não poderíamos também deixar de mencionar uma outra informação importantíssima: a quantidade existente destes documentos em relação aos demais. Segundo levantamentos citados em artigo no *Information Management Journal*, atualmente produzimos cerca de *161 exabytes* de informações digitais. Para ilustrar o que significa esta quantidade de informações ela equivale a *"três vezes a informação contida em todos os livros já escritos"* ou *"12 pilhas de livros que alcançariam da terra, cada uma, o sol"* (IM, 2007, p. 8). Portanto, a presença dos documentos eletrônicos e digitais como representantes de nossa cultura e modo de vida hoje é muito relevante e estes dados apresentados enfatizam a importância da preocupação com políticas de preservação de, pelo menos, uma parte deste imenso patrimônio da humanidade. Para simplificar o texto trataremos esses documentos simplesmente como **Documentos Digitais**.

3.3 PÁGINAS DA *WEB* COMO DOCUMENTOS

Do ponto de vista das disciplinas que tem os documentos como talvez o mais importante objeto de trabalho, como a **Biblioteconomia** ou a **Arquivologia**, não há motivos para não considerar páginas da *web* disponíveis na rede Internet - principalmente aquelas disponibilizadas através dos protocolos do tipo HTTP - como um documento com o mesmo status que um livro de biblioteca, uma carta histórica ou um relatório financeiro contábil em

papel de uma grande empresa. Descontadas as especificidades próprias que um documento do tipo página Internet possui, todos os elementos presentes em outros tipos de documento também estão presentes no primeiro tipo. Especificidades estão presentes também em outras categorias de documentos modernos, como os filmes e a música, os quais pedem métodos próprios de tratamento. A disciplina **História**, por exemplo, vem dando atenção a esses documentos: *"As fontes audiovisuais e musicais ganham crescentemente espaço na pesquisa histórica."* (NAPOLITANO, 2006, p. 235).

No entanto, os documentos do tipo "páginas na rede Internet" podem ser vistos menos como documentos ou evidências históricas e mais como um meio de acesso a documentos históricos "reais",

> A rede mundial de computadores representa grande apoio a historiadores, sobretudo àqueles que não têm acesso às grandes instituições de coleta e preservação dos acervos audiovisuais. A Internet, no entanto, é mais um depósito de informações, um grande arquivo virtual de referência, do que um arquivo material de fontes primárias. (NAPOLITANO, 2006, p. 265).

Os conteúdos presentes em páginas da Internet têm adquirido tamanho *status* como documentos importantes que existem até mesmo instituições devotadas exclusivamente com a preocupação de sua preservação para as gerações futuras; é o caso da *International Internet Preservation Consortium* (IIPC), cuja missão, disponível em seu sítio[25], é *"[...] adquirir, preservar e tornar acessível o conhecimento e informações da Internet para as futuras gerações em qualquer lugar, promovendo o intercâmbio global e relações internacionais"*.

Uma análise nos trabalhos que vêm sendo desenvolvidos pelo IIPC, mostra um aspecto da preservação de documentos muito específico para páginas da Internet: a persistência dos *links* correspondentes ao longo do tempo. Como hoje já é largamente

[25] http://www.netpreserve.org

conhecido por todos, o acesso a uma determinada página na Internet é feito através da inserção de um endereço num programa aplicativo do tipo navegador (*browser*). Esse endereço é um *link* lógico para acessar o conteúdo daquele sítio. Dentro do próprio conteúdo das páginas nos sítios também são comumente inseridos endereços para outros sítios (*links*). Esse processo, conhecido como uso de *hipelinks*, tem apresentado um problema. Freqüentemente, o endereço ou *link* digitado num determinado sítio pode não corresponder mais ao endereço do sítio original, o qual pode não estar mais disponível ou estar disponível em um novo endereço; em outras palavras, um determinado *link* pode não ser persistente[26] ao endereço que originalmente se reporta.

Esse problema se manifesta de muitas maneiras, mas é particularmente importante no caso de citações acadêmicas e científicas. A qualidade e quantidade de textos científicos disponibilizados na Internet têm crescido a largos passos; conseqüentemente, tem crescido também o número de citações de trabalhos disponibilizados na rede. Diferentemente da citação de artigos em periódicos tradicionais em papel ou livros em bibliotecas, o acesso a determinado texto citado às vezes simplesmente passa a não ser mais possível. Esse processo pode causar certa apreensão ou até rejeição por citações de sítios em trabalhos acadêmicos. Num estudo efetuado em periódicos científicos em 2003, foi encontrado até 21% de inatividade para referências na Internet em artigos com 27 meses de idade de publicação (DELLAVALLE *et al.*, 2003, p.1).

Todas essas questões agravam-se diante do problema de possíveis alterações no conteúdo original de determinado conteúdo que foi disponibilizado anteriormente. Em certos casos, as alterações nos conteúdos de sítios são uma característica inerente ao próprio sítio,

[26] *Persistente* é o termo que tem sido utilizado, principalmente em publicações na área de tecnologia, para designar o funcionamento correto de *links* na Internet.

como no caso daqueles devotados à publicação de notícias. Podemos falar, então, sobre a **dinamicidade** da Internet.

3.3.1 A INTERNET COMO ENTIDADE DINÂMICA

Do que se trata quando falamos em **conteúdo dinâmico** na Internet ? Neils Brügger, num trabalho sobre arquivamento de páginas na Internet e abordando essa característica, afirma *"A Internet é um meio dinâmico no sentido que seu conteúdo muda ou é removido rapidamente."* (BRÜGGER, 2005, p. 21). Acrescentaríamos também, além da **mudança** e **remoção** de páginas, o **surgimento** de novas páginas e conteúdos na rede Internet. Porém, do ponto de vista do exame documental de páginas em seus sítios, certamente a mudança de conteúdo ou simplesmente o desaparecimento de páginas inteiras ou partes dessas se constitui no maior dos problemas.

As mudanças de conteúdo em páginas podem ocorrer e freqüentemente ocorrem em função de várias razões, como alterações de tecnologia que possibilitam novos recursos visuais ou por se tratar de sítios que institucionalmente têm seu conteúdo alterado; é o caso de sítios de agências de notícias. Por exemplo, durante os levantamentos que temos efetuado (que descreveremos detalhadamente em capítulo próximo) notamos a forte presença de itens dentro dos sítios dedicados a notícias relacionadas à função fim de determinado órgão. Assim, um determinado tribunal, além dos itens típicos de suas atividades judiciárias, como a busca de processos judiciários a ele submetidos, costuma conter uma área destinada a divulgar determinados julgamentos mais relevantes, como ilustra a *figura 1*:

Figura 1 - Página da Internet com notícia divulgada

A página de sítio na Internet da *figura 1* foi obtida no endereço http://www.stj.gov.br ao final de junho de 2008. Certamente, o leitor desse texto que tentar visualizar a página nesse momento não obterá acesso ao mesmo conteúdo da *figura 1*. Além disso, há alterações de conteúdos que não são facilmente perceptíveis, pelo menos visualmente, como alterações no *lay-out* de determinadas páginas pouco consultadas.

As páginas na Internet costumam funcionar como uma espécie de central de acesso a vários outros documentos, como informativos, cópias de outros documentos originais em papel, imagens fotográficas e até mesmo áudio e vídeo. Esses documentos, na forma de arquivos de computador, são acessíveis através de *hiper-links* nas páginas ou através de dispositivos que buscam documentos e exibem *hiper-links* para acesso aos arquivos. A quantidade e qualidade desses arquivos que podem ser acessados via páginas Internet, às vezes chamadas de **portais**, mudam com bastante freqüência. Particularmente, a **quantidade** desses arquivos tem crescido ao longo dos últimos anos.

O desaparecimento de sítios na Internet é outro grande problema; não é difícil encontrar endereços de sítios fora do ar. Parte do conteúdo original pode ter migrado para outros endereços ou simplesmente ter desaparecido por completo. Dentro de nosso **universo de pesquisa**, por se tratar de sítios governamentais, há uma certa estabilidade. Porém, ao longo da coleta de dados houve uma alteração com relação ao sufixo dos sítios; esse sufixo foi alterado de *.gov.br* para *.jus.br,* mantendo-se, no entanto, o mesmo conteúdo anterior[27] (BRASIL, Resolução 45)

[27] A Resolução número 45 de 17 de dezembro de 2007 do Conselho Nacional de Justiça (CNJ) dispõe sobre a padronização dos endereços eletrônicos dos órgãos do Poder Judiciário e trata também de outras alterações para endereços na Internet no âmbito do judiciário brasileiro.

3.3.2 A ESTRUTURA DE UM SÍTIO NA INTERNET

As primeiras páginas de sítios disponibilizados na rede Internet eram, se comparadas às páginas atuais, incrivelmente simples. Basicamente, o que visualizávamos era um texto fixo codificado em linguagem original **HTML**. Ao longo do tempo e da evolução tecnológica, os conteúdos têm se tornado bastante complexos, incluindo imagens fixas e em movimento, sons e muitos outros elementos como animações, planilhas, e etc. Todos esses elementos são arquivos de dados codificados nos mais diferentes **formatos de arquivo** (conceito que veremos detalhadamente adiante), como o pdf, jpg e tantos outros. Além disso, atualmente, um sítio é composto, na verdade, por várias páginas internas ao mesmo sítio ou externas, nesse caso referenciadas por *links* **externos**.

Essa estrutura complexa combinando vários documentos em um só, no entanto, não é algo tão novo assim em termos de caracterização de um documento. Um **dossiê** ou **processo**[28] tradicional, no sentido arquivístico, em suporte papel nada mais é que um documento que reúne vários outros, podendo conter documentos impressos, formulários, cartas manuscritas, fotografias, e etc.

Em 2004, um estudo foi efetuado em periódicos científicos para tentar identificar e classificar muitas das condições encontradas em sítios da Internet (MARILL, BOYKO, ASHENFELDER, 2003). A *tabela 2* resume nossa tradução e adaptação do relatório original com os elementos mais relevantes e comuns:

[28] **Dossiê**: "Conjunto de documentos relacionados entre si por assunto (ação, evento, pessoa, lugar, projeto), que constitui uma unidade de arquivamento. **Processo**: "Conjunto de documentos oficialmente reunidos no decurso de uma ação administrativa ou judicial, que constitui uma unidade de arquivamento.". Ambas definições do Dicionário Brasileiro de Terminologia Arquivística (DBT, 2005).

Classificação	Condição
Documentos HTML estáticos	Arquivos individuais HTML, GIF, JPEG
Conteúdos de tipos alternativos	FLASH, PDF, XML, formatos MS-Office
Formulários	Listas Drop-down
JavaScript	Menus de navegação e conteúdo a ser aberto em outras janelas
JavaScript em clientes	URLs geradas para interação dinâmica
Mídias sem *streaming*	*Links* diretos para áudio ou vídeo
Mídias com *streaming*	*Links* indiretos ou *plug-ins* específicos

Tabela 2 - Classificação de elementos em sítios da Internet (adaptado)

3.3.3 ÚLTIMAS CONSIDERAÇÕES

Nessa dissertação, a coleta de dados será feita por prospecção em sítios da Internet (o que será detalhado na metodologia de coleta de dados), um dos motivos pelos quais estamos tratando das relações entre **sítios na Internet** e **documentos**. No entanto, os dados que utilizaremos se referem apenas a alguns arquivos disponibilizados nos sítios e não aos sítios em si como documentos. Os arquivos que utilizaremos como amostras para identificação dos formatos de arquivo em uso são apenas uma parte de todos os arquivos presentes nos sítios pesquisados. É importante também notar que estamos interessados nos arquivos que contêm conteúdo documental, como imagens/fotografias institucionais, relatórios de trabalho, reportagens em texto, multimídia, e etc. Os elementos que puramente constituem o código e programação dos sítios (codificação HTML ou aplicativos JavaScript, por exemplo) estarão **fora de nosso escopo** de coleta e análise.

4 O QUE SÃO FORMATOS DE ARQUIVO

Sem dúvida, essa parte conceitual é a mais importante desse trabalho; pode-se dizer que se trata da alma dessa dissertação. É essa base conceitual que norteia toda a coleta de dados efetuada na pesquisa. Devemos aqui responder à pergunta fundamental: *O que são Formatos de Arquivo?*

Esse conceito parece padecer do mesmo problema que o conceito de *Documento*. Esse é um conceito prosaico e com o qual quase todas as pessoas lidam em seu dia-a-dia. Pelo mesmo motivo, ou seja, por ser largamente utilizado, apresenta vários sentidos, dependendo de quem o interpreta e utiliza. O resultado é um conceito "fácil"; todos sabem o que é, todos podem dizer o que é e, conseqüentemente, fica cada vez mais difícil defini-lo com precisão. No caso do conceito de documento, no âmbito dos pesquisadores da área de **Documentação** e **Ciência da Informação**, sabemos o quanto é difícil defini-lo precisamente.

4.1 FORMATO DE ARQUIVO: DEFINIÇÕES

Com o objetivo de definir, então, com a maior precisão e clareza possível o conceito de **Formato de Arquivo**, iniciaremos o trabalho trazendo algumas definições presentes em outros trabalhos de pesquisa. Antes, porém, vamos trazer à luz alguns conceitos ainda mais fundamentais.

4.1.1 DIGITAL E ANALÓGICO

O uso do termo digital é bastante novo, pelo menos na acepção que aqui nos interessa, ou seja, a que tem sido utilizada em **tecnologia eletrônica** e **informática**. Um aspecto fundamental desse termo se refere a uma nova maneira de registrar e representar informações.

Os primeiros artefatos eletrônicos que o homem criou utilizavam exclusivamente o que agora chamamos de tecnologias analógicas, contrapondo-se às atuais tecnologias digitais. Auto-falantes utilizados em qualquer equipamento de som, como as caixas de som do computador, são um bom exemplo de tecnologia analógica. O som produzido por esses

equipamentos é o resultado do movimento mecânico de eletroímãs; as características sonoras como os graves e agudos e a altura do som são o resultado de milhares de movimentos mais ou menos intensos.

Atualmente, apesar de ainda utilizarmos a tecnologia analógica em muitos equipamentos, como no exemplo acima, a maioria dos circuitos internos de qualquer equipamento eletrônico processa sinais no modo digital. Em oposição à miríade de opções exemplificadas no caso do alto-falante, há, no caso da tecnologia digital, um número finito de opções: **zeros e uns**. Apesar do exemplo dado no universo dos equipamentos sonoros, sem dúvida alguma, a maior aplicabilidade da tecnologia digital está no âmbito da informática: armazenar e processar informações representadas pelos números zero e um.

Um estudo aprofundado dessa tecnologia tomaria muitas e muitas páginas, mas o que nos interessa é o aspecto da codificação binária.

4.1.2 CODIFICAÇÃO BINÁRIA

O princípio fundamental do uso de tecnologia digital no universo da informática é o de converter as informações utilizadas na linguagem humana – como nosso sistema de escrita e numeração – em códigos formados por grupos de números binários: somente o número zero e o número um. Naturalmente, o número e quantidade de **dígitos** (01001011...) necessários para representar essas informações dependerá da complexidade das informações a serem representadas. Assim, com 3 dígitos binários podemos representar $2^3 = 8$ códigos, conforme ilustra a *tabela 3*:

Número decimal	Código binário correspondente
0	000
1	001
2	010
3	011
4	100
5	101
6	110
7	111

Tabela 3 - Codificação binária

Os computadores atuais, além de outros dispositivos digitais, trabalham atualmente com códigos de 64 dígitos ou mais. Essa quantidade de códigos permite armazenar uma grande quantidade de informações. Muito além dos caracteres de nossa linguagem (em qualquer idioma), é possível representar as cores utilizadas numa imagem (em cada minúsculo ponto), os sons de uma música ou a fala humana. Isso sem mencionar os códigos internos, que possuem significado somente para os circuitos, como os comandos dos microprocessadores ou endereços de memória.

4.2 DEFINIÇÕES

Vamos agora trazer à luz o conceito de formato de arquivo e relacioná-lo com a representação no universo digital.

Num relatório elaborado no âmbito do projeto *The Representation and Rendering Project*[29] , da Universidade de Leeds, no Reino Unido, encontramos a seguinte definição para formato de arquivo:

> Em seu nível mais baixo, objetos digitais são seqüências de zeros e uns que representam dados codificados. Diferentes Formatos de Arquivo especificam como esses códigos representam o conteúdo intelectual criado por um autor de um objeto digital. (UNIVERSITY OF LEEDS, [s.d], p. 4).

[29] HTTP://www.leeds.ac.uk/reprend/

A definição chama a atenção para o fato de que um formato de arquivo qualquer especifica como um determinado conteúdo está estruturado.

O termo técnico associado ao "como" da definição anterior chama-se especificação. Sobre esse termo: "Uma definição completa de formato de arquivo tem de incluir o conceito de especificação (*specification*), o qual, em si, pode ser definido como os requisitos organizacionais de um arquivo" (SHEPARD; MacCARN, 1997, p. 6).

Os "requisitos organizacionais de um arquivo" referem-se à estrutura em que os códigos digitais estão organizados para cada tipo de arquivo (formatos de arquivo). Essa estrutura extrapola em muito os códigos utilizados para representar o conteúdo básico e sensível a nós humanos como texto, imagem, som e muitos outros. Além desse conteúdo, muitas outras informações são necessárias. Tomemos como exemplo um arquivo de texto simples contendo uma pequena receita. Na tela de um aplicativo editor de texto ele seria visualizado aproximadamente como na *figura 2*:

RECEITA DE BOLO

INGREDIENTES: Ovos (3), leite (1/2 litro), açúcar (150g), farinha de trigo (500g) e fermento (15g).

COMO FAZER: Misture tudo e asse em forno pré-aquecido por 40 minutos.

Figura 2 - Arquivo visualizado em editor de textos

Que informações deveriam ser gravadas no arquivo correspondente ao conteúdo do texto acima? Em primeiro lugar, o próprio texto, ou seja, os códigos binários que correspondem aos caracteres utilizados acima. Notemos também que foram utilizados caracteres com as fontes *Tahoma* e *Times New Roman*. Além disso, algumas palavras estão em negrito. Há também informações sobre os espaços entre linhas e entre caracteres, margens,

etc. Essas informações todas se referem ainda ao conteúdo visível do texto. Porém, um arquivo real necessita também de **metadados** (adiante falaremos detalhadamente sobre metadados) mínimos, como a data de criação do arquivo, o tamanho desse arquivo em *bytes*, o *software* utilizado para a criação, etc. Além desse exemplo com **texto**, quando lidamos com arquivos como **imagens fixas**, **som** ou **imagem em movimento**, o grau de complexidade aumenta consideravelmente.

Uma especificação para um formato de arquivo X nada mais é senão a determinação de quais informações (conteúdo, metadados e outros) e qual a ordem seqüencial (ou não) de gravação no arquivo físico composto por **códigos binários** (também chamados de *bitstream*).

Infelizmente, a primeira coisa a reconhecer é o quão uma especificação de formato de arquivo não é simples, desde os menos complexos arquivos de texto até formatos de arquivo específicos para imagens em movimento.

Vamos fazer uma pequena análise numa especificação real de formatos de arquivo com o objetivo de compreender ainda melhor esse conceito tão importante. Escolhemos uma especificação menos complexa tomando como parâmetro o poder de processamento e recursos do aplicativo que gera o arquivo nessa especificação: o aplicativo *WRITE*, um editor de texto da empresa *Microsoft*[30].

No início da primeira página, há uma orientação sobre características básicas dessa especificação; sabemos que esse tipo de arquivo contém, além do conteúdo propriamente dito, texto, figuras e formatação.

[30] Pode-se consultar essa especificação no Anexo I ao final dessa dissertação.

O primeiro tópico abordado tem o título de *File Header* (cabeçalho do arquivo) e descreve o conteúdo do arquivo; por exemplo, no cabeçalho está registrado o comprimento do arquivo (*length of the file*). Logo abaixo temos acesso a uma tabela com as *Word* (palavras), *Name* (nomes das palavras) e suas respectivas descrições. Cada *Word* corresponde a 16 *bits*[31]. A primeira *word* (*wIdent*) parece ser utilizada para identificar o arquivo; normalmente teria o número 0137061 (em linguagem octal), que corresponde a 1011111000110010 (em linguagem binária)[32].

Ainda na primeira página da especificação, ao final encontramos um tópico com o título *Text* (texto). Nesse tópico ficamos sabendo que o texto propriamente dito, num arquivo desse tipo inicia-se a partir da *word* 64. Mais adiante, sabemos que os caracteres ASCII[33] de números 13 e 10 têm uso especializado e correspondem respectivamente ao comando para retorno de cada linha (*carriage return*) num parágrafo e avanço para uma próxima linha (*linefeed*).

Na seqüência, temos ainda mais 6 páginas e tópicos relacionados às *Pictures* (figuras) eventualmente utilizadas no arquivo, *Formatting* (formatação), *Characters and Paragraphs* (caracteres e parágrafos), *Sections* (seções num mesmo documento) e informações sobre as fontes de caracteres utilizadas (*Font Table*). Facilmente percebemos que se trata de um conjunto de informações bastante especializadas, compreensíveis e úteis para iniciados em **Linguagem de Programação** e **Ciência da Computação**. Nosso objetivo foi de apenas exemplificar uma especificação real de formato de arquivo.

---

[31] Uma *Word* de 16 *bits* é uma convenção utilizada em linguagens de programação e significa um número com 16 dígitos binários.

[32] As representações em linguagem octal, binária ou outras como a hexadecimal e decimal (a utilizada por nós no dia-a-dia) são apenas maneiras diferentes de representar quantidades numéricas e cada uma é mais apropriada para determinado uso.

[33] ASCII, lê-se *ásqui 2,* e significa *American Standard Code for Interchange of Information*. Trata-se de uma tabela com códigos binários e seus correspondentes a caracteres comuns, especiais ou comandos específicos.

4.3 TIPOS DE FORMATOS DE ARQUIVO

Existe hoje uma grande quantidade de especificações técnicas para uma infinidade de formatos de arquivo diferentes. Muitas das especificações atualmente em uso evoluíram a partir de versões antigas de aplicativos hoje descontinuados. Além disso, *software*s novos são criados diariamente; conseqüentemente, novas especificações de formatos também. A grande explosão de novos formatos de arquivo ocorreu com o surgimento da microinformática e os computadores pessoais; mas, antes desse período, nas últimas décadas do século XX, eles já existiam no mundo dos *mainframes*[34]. Segundo Kientzle:

> Sistemas operacionais para *mainframes* tratam um arquivo como um repositório de base de dados. Cada item nessa base de dados é um *record* e, dessa forma, *mainframes* tratam arquivos como uma coleção de *records*[35]. (KIENTZLE, 1995, p. 358).

4.3.1 CLASSIFICAÇÃO DE FORMATOS DE ARQUIVO

Uma primeira classificação de formatos de arquivo pode ser feita com base no tipo de *software* utilizado para gerar os arquivos que serão gravados em algum tipo de mídia de acordo com a **especificação** do formato. O formato de arquivo ***Write*** seria do tipo Texto, pois é gerado através de um aplicativo para edição de **textos**. Essa classificação é, no entanto, problemática, pois, em geral, podemos falar em aplicativos que geram **predominantemente** textos, imagens fixas, sons, etc mas não **exclusivamente** esses tipos de conteúdos. Isso ocorre mesmo em formatos de arquivo aparentemente exclusivos para certos conteúdos. Um exemplo é o formato de arquivo MP3, feito especialmente para registro de sons em geral. Ocorre que é possível incorporar ao arquivo no formato MP3 legendas **textuais** para as músicas gravadas. Um outro exemplo nesse sentido se refere ao formato GIF, projetado para

---

[34] O termo *mainframe* é utilizado para designar computadores de grande porte, utilizados apenas por grandes corporações na era anterior à microinformática. É curioso notar que, na verdade, possuíam poder de processamento inferior aos computadores pessoais atualmente em uso.

[35] Um *record* ou registro numa base de dados corresponde a cada grupo de campos. Por exemplo, os campos nome, idade e endereço exigirão tantos registros quantos forem os nomes da relação de pessoas numa organização.

imagens fixas, apesar de existir o chamado GIF animado, que pode incorporar imagens em movimento. Assim, em geral, pode-se falar de formatos de arquivo para conteúdos predominantemente em determinado conteúdo, a *tabela 4* exemplifica o exposto:

Tipo predominante de conteúdo	Exemplos de Formatos de Arquivo
Texto	RTF, OpenOffice, ODF, DOC, AmiPro e outros
Imagens fixas	BMP, EXIF, GIF, JPG, TIFF e outros
Imagens em 3D	CAD, BIFF, X4D e outros
Sonoro	MEU, KAR, MP3, MP4 e outros
Imagens em movimento	AVI, MOV, MPEG, SWF e outros

Tabela 4 - Classificação de formatos de arquivo pelo conteúdo

Na *tabela 4*, os exemplos de formatos de arquivo são nomeados pela extensão do nome do arquivo em ambientes de computadores pessoais (*Windows*, *MacOS* e outros); discutiremos sobre extensões de formatos de arquivo na parte sobre identificação de formatos de arquivo. A *tabela 4* não é exaustiva mas apenas ilustrativa[36].

4.3.2 VERSÕES DE FORMATOS DE ARQUIVO

Nesse ponto, é necessário chamar a atenção para um detalhe técnico extremamente importante: formatos de arquivo possuem, geralmente, diferentes **versões**. Desde a primeira versão de um *software*, digamos, um editor de textos, várias modificações e aperfeiçoamentos são implementados. Por exemplo: em editor de texto pode não permitir o uso de imagens junto ao documento textual; mas, a partir de uma nova versão, esse recurso passa a ser possível. Assim, haverá modificações na especificação original do formato de arquivo para que seja possível armazenar imagens nos arquivos. Algumas novas versões de um mesmo formato de arquivo podem ser consideravelmente diferentes da versão anterior, além da própria freqüência com que surgem novos formatos. Na *figura 3*, ilustramos a capa de uma

[36] No sítio *Wotsit.org* (http://www.wotsit.org), por exemplo, é possível consultar uma relação bem mais completa de especificações de formatos.

especificação de formato de arquivo (o formato File Interchange Format) onde pode-se constatar a versão correspondente.

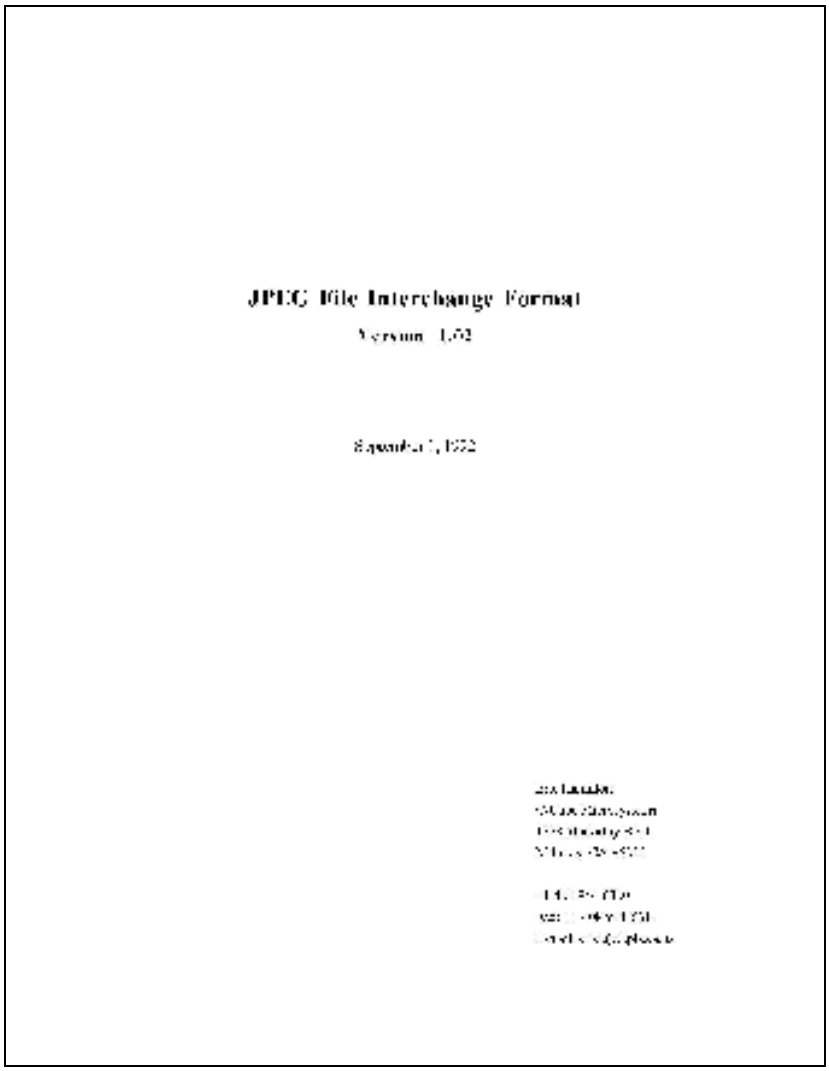


Figura 3 - Especificação com versão de formato

5 METADADOS E FORMATOS DE ARQUIVO

Por que abordar o assunto **Metadados** nessa dissertação? Porque é através do uso de **Metadados** que muitos dos procedimentos possíveis para a consecução da preservação digital – veremos mais adiante alguns deles – se tornam viáveis. Além disso, uma das aplicações mais comuns implementada através dos *dados sobre dados* é a recuperação de documentos armazenados. É claro que não faz muito sentido preservar documentos que não poderão ser recuperados de alguma maneira em algum momento no futuro.

Mas o que são metadados ? Primeiro, é preciso lembrar que se trata de elementos que podem ser utilizados até mesmo em documentos não eletrônicos. No nosso caso, o foco é para metadados utilizados em documentos digitais: **incorporados** (adiante falaremos mais sobre essa característica) nos objetos ou não. Quando se registra, por exemplo, em fichas de papel, os dados bibliográficos de livros em uma biblioteca ou anotações sobre artigos estudados, estamos elaborando Metadados sobre aqueles livros e esses artigos. A autores que entendem Metadados como uma *"amplificação do processo tradicional de catalogação bibliográfica"* (DAY, 1998 apud OCLC/RLG, 2001, p. 2) O prefixo *meta* é aqui empregado significando *algo (dados)* **dito/registrado** *sobre algo*. O *sobre algo* refere-se, no âmbito de nosso interesse, ao conteúdo de documentos, além de informações técnicas sobre o formato de arquivo.

Pode-se encontrar na literatura científica inúmeras tentativas de classificar Metadados, tanto no que se refere aos *tipos* como às *funções* desses. Por exemplo, *"No meu entendimento, há uma divisão clara em relação a metadados que denominei de duas categorias básicas: metadado técnico e metadado de negócios"* (IKEMATU, 2001). Em outro artigo, THOMAZ e SANTOS (2003) citando WENDLER (2001), nos reporta que Metadados estão associados a três categorias funcionais: *Descritiva*, *Administrativa* e *Funcional*. Nesses dois exemplos citados, os autores referem-se a **objetos digitais** especificamente. De fato, a própria definição

e a classificação de Metadados dependerá dos objetivos e tipo de documentos aos quais os Metadados correspondentes se referem.

Ao nosso ver, uma classificação de Metadados deve definir o tipo de documento a que se refere e categorizar em relação ao tipo de descrição efetuada, pois qualquer categoria de Metadados serve fundamentalmente para **descrever** (diferentes informações, de diferentes maneiras). Em consonância com esse raciocínio, CAMPOS (2007, p.18) registrou *"Em última instância, todo metadado descreve algum objeto. No entanto, descrevem esse objeto para fins variados."*.

Para o campo de interesse relativo a essa pesquisa, o **tipo de documento** é o **digital** e as **categorias** se referem fundamentalmente ao **conteúdo** desses documentos e às **informações técnicas** sobre esse documento. Exemplificando, tomemos um documento digital fotográfico, como o da *figura 4*:

Figura 4 - Documento Digital Fotográfico (http: //www.iptc.org)

A *tabela 5* exemplifica Metadados para o documento acima, tanto na categoria **Conteúdo** como na categoria **Informações Técnicas**.

Exemplo de MetaDado sobre conteúdo	Exemplo de Metadado técnico
Texto Descritivo: Menino afro americano com três anos de idade divertindo-se em praia durante o período de férias.37	Formato de Arquivo: JPG

Tabela 5 - Categorias de Metadados

É possível a utilização de uma grande quantidade de metadados nas duas categorias acima, como em qualquer outra categoria. Na categoria **Conteúdo**, podemos utilizar **palavras-chave** sobre o documento, ou uma **legenda curta,** etc. Na categoria de Informações Técnicas podemos registrar o **tamanho do arquivo** em *bytes*, a **versão do formato de arquivo**, além de outros dados. De todos os Metadados possíveis de serem derivados a partir de um documento digital qualquer, alguns não se encaixariam exatamente na categorização de conteúdo e informações técnicas; por exemplo, o autor do documento (o nome do fotógrafo no nosso exemplo para imagem fotográfica) refere-se ao conteúdo ou às informações técnicas? Seria o caso de criar outras categorias?

Um aprofundamento nessas discussões está além dos limites de nossa dissertação; nosso objetivo foi apenas estabelecer uma pequena delimitação terminológica para então abordar mais especificamente Metadados que têm sido utilizados para a **Preservação Digital**.

5.1 METADADOS PARA PRESERVAÇÃO

Como já registramos antes, o grande objetivo do uso de Metadados é a **Descrição** (em diferentes categorias, dependendo do foco do trabalho). De diferentes formas, todas essas descrições são úteis à preservação de um documento digital. Com relação ao **conteúdo**,

[37] A imagem e o texto descritivo foram extraídos de exemplos disponíveis no sítio da International Press Telecomunications Council (IPTC), disponível em http: //www.iptc.org.

registrar o significado original que um determinado autor quis ou não dar ao documento criado por ele ou com relação a **informações técnicas**, identificar o formato de arquivo digital utilizado e assim facilitar futuras migrações para outros formatos de arquivo. Nesse sentido, têm surgido diversas iniciativas para estabelecer padrões de Metadados especialmente para as atividades de Preservação Digital; em outras palavras, são conjuntos de elementos elaborados especialmente para o problema desse tipo de preservação. Abordaremos aqui algumas dessas iniciativas que têm sido mais citadas na literatura científica com o intuito de delimitar melhor a relação entre **Metadados** e **Preservação Digital**.

Um trabalho seminal que analisou e comparou quatro importantes conjuntos de Metadados elaborados especialmente para a Preservação Digital é o *Preservation Metadata for Digital Objects: A Review of the State of the Art*. Esse relatório, elaborado por um grupo de trabalho da OCLC/RLG em 2001, assim define a importância de Metadados no contexto da Preservação Digital:

> Todas as formas de preservação digital, exceto as mais simples, podem se beneficiar pela criação, manutenção e evolução de Metadados detalhados para apoio aos processos de preservação. Por exemplo, Metadados podem documentar o processo técnico associado com a preservação, especificar informações de direitos autorais e estabelecer a autenticidade do conteúdo digital. Eles podem registrar a cadeia de custódia de um objeto digital e identificá-lo individualmente tanto interna como externamente em relação ao arquivo em que reside. Em resumo, a criação e instalação de Metadados para Preservação parece ser um componente chave para as estratégias de preservação. (OCLC/RGL, 2001, p. 2)

Um dos objetivos do relatório da OCLC/RLG foi a busca por um padrão de Metadados para preservação. As vantagens na existência de consenso nessa área são grandes; infelizmente se constata, ainda hoje, a inexistência desse consenso.

Os conjuntos de Metadados analisados pelo relatório citado em 2001 foram: o *Exemplars in Digital Archives Project* (CEDARS), o *National Library of Australia* (NLA), o *Networked European Deposit Library* (NEDLIB) e o *Digital Repository Services* (DRS). Destaca-se no relatório a grande influência do modelo de referência *Open Archival*

Information System (OAIS)[38] em todos os conjuntos, exceto no DRS que se baseia em tecnologia XML.

Entre as conclusões mais relevantes do estudo efetuado e documentado no relatório da OCLC/RLG, além do uso do padrão OAIS (no Brasil é chamado de SAAI) estão o próprio propósito de **Metadados para Preservação**, ou seja, documentar a informação necessária para, primeiro, facilitar a tomada de decisão pelos gestores da preservação digital e, segundo, manter o acesso ao conteúdo dos objetos digitais armazenados.

Não iremos estudar detalhadamente os conjuntos de Metadados citados anteriormente, até porque se trata de um trabalho já com quase oito anos e muito provavelmente passaram por importantes alterações, o que pode ter diminuído em grande parte a própria importância do trabalho de comparação entre os modelos. Nosso primeiro objetivo era evidenciar a própria existência de Metadados específicos para a preservação[39].

O segundo objetivo era evidenciar a relação entre **Metadados** para **Preservação** e **Formatos de Arquivo**. A intersecção entre esses dois assuntos ocorre no grupo de elementos do conjunto de Metadados dedicado a descrever os objetos digitais em sua estrutura. Além da estrutura, os Metadados devem se referir a elementos administrativos como a Gestão de **Direitos Autorais** ou semânticos como a **Descrição do Conteúdo** dos documentos. A descrição da estrutura aborda informações, no nosso caso, estritamente tecnológicas. A *tabela 6* baseia-se no conjunto de elementos do padrão NEDLIB no que se refere aos elementos da estrutura mencionada.

---

[38] O modelo de referência *Open Archival Information* System (OAIS) é uma especificação de alto nível, ou seja, define elementos abstratamente e de maneira geral sem entrar em detalhes de implementação tecnológica. Esse modelo vem adquirindo grande importância no contexto da preservação digital e há, desde 2007, uma norma técnica brasileira com a tradução desse modelo, ver NBR 15472.

[39] No Anexo VIII disponibilizamos uma tabela comparativa extraída do relatório com os conjuntos completos de Metadados

Especific Hardware Requirements
Specific microprocessor req.
Specific multimedia req.
Specific peripheral req.
Operating System
Name
Version
Interpreter & Compiler
Name
Version
Instruction
Object Format
Name
Version
Aplication
Name
Version

Tabela 6 - Metadados para Preservação (Estrutura do Objeto Digital)

5.2 ÚLTIMAS CONSIDERAÇÕES

Metadados para Preservação são elementos essenciais em qualquer estratégia de preservação; apesar disso, ainda não há um conjunto único e largamente utilizado por todas as organizações. Para que os formatos de arquivos possam continuar sendo acessados ao longo do tempo, diversas ações deverão ser tomadas. Para o sucesso dessas ações algumas informações serão essenciais, como o **nome do formato de arquivo original** e sua **versão**. Essas informações e várias outras – que dependem do formato de arquivo específico, por exemplo, se esse é de áudio ou vídeo (assim necessitando de algoritmos de compressão específicos) – serão preservadas através de Metadados.

É interessante notar também que alguns formatos de arquivo, como o JPG, permitem a inserção de Metadados internamente junto ao conteúdo e demais códigos do arquivo (juntamente com as imagens, no caso desse formato específico). Esse fato nos parece vantajoso pois os Metadados de objetos digitais também são arquivos digitais e precisam ser igualmente preservados. É interessante observar que é preciso responder à pergunta *Quem preservará meus Metadados de Preservação Digital?* Quando os Metadados estão incorporados dentro dos arquivos digitais há uma vantagem na medida em que futuras

migrações (ver revisão bibliográfica sobre migração) desse arquivo poderão levar também os Metadados incorporados e haverá, então, um arquivo a menos (o de Metadados) requerendo cuidados de preservação.

6 MODELO DE FORMATO DE ARQUIVO PARA PRESERVAÇÃO

O objetivo desse capítulo é definir e expor um modelo de formato de arquivo adequado para a preservação de documentos digitais por longos períodos. É também, criar um **Mecanismo de Referência** para comparação com os **Formatos de Arquivo** efetivamente em uso nas organizações que compõem o nosso universo de coleta de dados. Ou seja, com a existência de um **Modelo** adequado será possível diagnosticar se um determinado formato de arquivo, efetivamente em uso por uma organização pesquisada, está próximo do desejável ou não.

Um formato de arquivo real é um produto de engenharia de *software* que pode ser extremamente complexo. Dependendo do conteúdo do formato de arquivo, como imagem fixa, imagem em movimento, som, texto ou combinações entre esses e outros tipos, o formato de arquivo pode possuir uma especificação bastante extensa. É possível também que estejam em uso tecnologias correlatas para compactação do tamanho em *bytes* dos arquivos ou procedimentos para criptografar arquivos.

No entanto, o **Modelo** de **Formato de Arquivo**, doravante denominado apenas **Modelo**, especificado aqui, será definido em alto nível, alienando-se de tecnologias específicas atualmente disponíveis. Queremos dizer com isso que as características do **Modelo** são abstratas e compostas de elementos com diretrizes gerais.

6.1 FORMATOS DE ARQUIVO PARA PRESERVAÇÃO

Uma primeira fonte de elementos que podem ajudar a subsidiar a escolha das características de nosso **Modelo** almejado é um outro recurso tecnológico que tem surgido no mercado: trata-se dos formatos de arquivo para preservação ou arquivamento. Um exemplo proeminente nesse sentido é o formato *Portable Document Format/Archiving* (PDF/A).

O que está implícito em formatos de arquivo como o PDF/A é a geração de arquivos digitais para documentos que já possuem um determinado *status* tal que sua preservação pelo

maior tempo possível se torna importante. Não são formatos de arquivo para utilização administrativa (uso corrente) quando esses ainda estão na fase de criação e tramitação nos ambientes de trabalho, e ainda nessa fase, podem ou não receber uma classificação como sendo de guarda permanente.

Compreender que características um formato de arquivo para arquivamento possui nos ajudará a compreender o porquê da necessidade de existência de certos elementos em nosso **Modelo**.

LeFurgy, um bibliotecário da *Library of Congress* nos Estados Unidos, escreveu um artigo (LeFURGY, 2003) sobre as possibilidades do formato de arquivo PDF/A para arquivamento e preservação de documentos por longos períodos. Naquele ano o formato pdf/a ainda estava em fase de estudos. A norma **ISO 19005-1,** em 2005, foi o resultado desses estudos levados a cabo por diversas organizações do setor público e privado.

Atente-se para o fato de que o formato de arquivo **PDF/A** (ou norma ISO 19005-1) se baseia na tecnologia do formato original da empresa Adobe: *Portable Document Format* (**PDF**). LeFurgy alertava no artigo de 2003 que o formato de arquivo **PDF** atende necessidades de produtores, usuários e instituições de guarda no que diz respeito a questões de autenticidade e confiabilidade, preservação por longos períodos e Metadados. Mas apesar disso, não é adequado para a preservação por longos períodos pois o produtor do formato a empresa *Adobe* controla sua produção e não está obrigado a continuar publicando a especificação. Além disso, não se trata de um formato que exija necessariamente todos os elementos para visualização do conteúdo dentro do arquivo final, por exemplo, ele pode não incorporar uma cópia das fontes originais utilizadas para o texto (LeFURGY, 2003). O formato **PDF/A** foi criado para aproveitar as vantagens do formato **PDF** e agregar vantagens específicas para a preservação digital.

O grupo de trabalho que desenvolveu o formato PDF/A tinha como objetivo que ele possuísse certas propriedades que o qualificariam como um formato para a preservação: segundo Susan Sullivan tais propriedades seriam: *Independência de dispositivo, auto conteúdo, auto descrição, transparência, acessibilidade, abertura da especificação e adoção* (SULLIVAN, 2006, p. 54)

Detalhando melhor essas propriedades citadas pela autora, a **Independência de Dispositivo** significa que a aparência estática do documento deve permanecer a mesma independentemente do *software* ou *hardware* utilizado para a visualização ou impressão do material. **Auto-Conteúdo** significa que tudo que for necessário para visualizar ou imprimir um documento deve estar incluído dentro do arquivo (um problema comum nesse sentido é a não incorporação dos arquivos das fontes (tipos de caracteres numéricos, textuais e outros símbolos) originais utilizadas. **Auto Descrição** implica no uso extenso de recursos de Metadados para descrever o máximo possível todos os aspectos de um arquivo. A propriedade **Transparência** significa que o conteúdo textual do arquivo deve poder ser extraído e lido independentemente da existência de um aplicativo especial para leitura de documentos no formato PDF/A. **Acessibilidade** é uma propriedade associada ao uso de criptografia e senhas de proteção; nesse caso, esses recursos são proibidos, habilitando assim o acesso livre ao conteúdo dos documentos. A **Abertura da Especificação** implica na autorização legal para uso público das informações técnicas do formato de arquivo: no caso, a detentora legal do formato PDF autorizou a publicação dessa especificação, indefinidamente, no que cabe ao subconjunto que compõe o formato PDF/A. Por último, a propriedade **Adoção** implica que o formato seja flexível o suficiente para poder ser largamente adotado no mercado: quanto mais popular for o formato, maiores serão suas chances de preservação futura. (SULLIVAN, 2006, p. 53-54). A *tabela 7* resume as características. Nela tabela, optamos por manter o termo no original em inglês; isso facilitará futuras comparações com outros termos. Além disso, a

tradução de termos diferentes eventualmente pode gerar um mesmo termo em nosso vernáculo.

Item	Característica
1	Device independent
2	Self-containment
3	Self-describing files
4	Transparency
5	Accessibility
6	Disclosure
7	Adoption

Tabela 7 - Características formato PDF/A

Há um projeto criado pela *Library of Congress*[40] nos Estados Unidos, com o intuito de dar suporte a decisões sobre preservação digital no que cabe ao uso de formatos de arquivo. Esse projeto objetiva*"a elaboração de um inventário de informações sobre formatos de arquivo em ascensão"* e *"identificar e descrever formatos que sejam promissores para a preservação por longos períodos e desenvolver estratégias para sustentar esses formatos"* (ARMS; FLEISHHAUER, 2005, p. 1).

Segundo um artigo dos responsáveis pelo referido projeto, é possível definir **Fatores de Sustentabilidade** (*Sustainability factors*) sobre os formatos de arquivo mais adequados para a preservação: nas palavras do autor:

> Fatores de sustentabilidade aplicam-se em formatos digitais em todas as categorias de informação. Nós identificamos sete fatores que influenciam a viabilidade e o custo da preservação do conteúdo. Nós acreditamos que esses fatores serão significantes se estratégias de preservação necessitem no futuro migração para novos formatos, emulação do software atualmente disponível em computadores do futuro, um híbrido de migração e emulação ou a normalização no recebimento. (ARMS; FLEISHHAUR, 2005, p. 3)

Os sete fatores acima citados são: Abertura da Especificação (*Disclosure*), Adoção (*Adoption*), Transparência (*Transparency*), Auto-Documentação (*Self-documentation*), Dependências Externas (*External Dependencies*), Impacto de Patentes (*Impact of Patents*),

[40] http://www.digitalpreservation.gov/formats

Mecanismos de Proteção Técnica (*Technical protection mechanisms*). A *tabela 8* lista os "fatores de sustentabilidade" listados no artigo em idioma original.

Item	Fatores
1	Disclosure
2	Adoption
3	Transparency
4	Self-documentation
5	External dependencies
6	Impact of patents
7	Technical protection mechanisms

Tabela 8 - Fatores de sustentabilidade para preservação

6.2 OUTRAS PROPOSTAS DE PRESERVAÇÃO

Há uma proposta de formato de arquivo universal para preservação, o qual além de acomodar qualquer tipo de formato de arquivo, resolvendo assim o problema de ter que lidar com vários formatos diferentes para preservação, é também uma proposta de formato para preservação digital.

Uma das primeiras propostas de um Formato Universal para a Preservação – *Universal Preservation Format* (UPF) – parece ser aquela proposta a partir da organização *WGBH* nos Estados Unidos. Trata-se de uma organização do tipo *Public Broadcasting Service* (PBS), voltada a programas educacionais em diversos meios, como rádio ou TV. Por se tratar de uma organização já em atividade há mais de meio século (iniciou suas atividades em 1951[41]), possui hoje um considerável acervo de documentos em diversos tipos de suportes tecnológicos, desde os primeiros tipos de fitas magnéticas até as atuais fitas digitais. Segundo David MacCarn, um dos diretores da **WGBH** na época do início do projeto UPF:

> As enormes e rápidas mudanças que ocorrem na tecnologia digital resultaram numa acentuada explosão de formatos. Treze formatos de fitas digitais estão disponíveis no momento (D-1, D-1SP, D-2, D-3, D-5, D-6, Digital Betacam, Betacam SX, Ampex DCT, Consumer DV, DVCAM, DVCPRO and Digital S) com vários outros em desenvolvimento (para a televisão de alta definição). (MacCARN, 1997).

[41] Veja-se sobre a instituição em http://main.wgbh.org/wgbh/about

Como uma possível solução ao problema dos diferentes formatos em mídias e a necessidade de arquivamento com a necessária preservação adequada desses materiais, surge a proposta UPF. De acordo com o líder do projeto,

> O *Universal Preservation Format* é um mecanismo de arquivo de dados que utiliza um *container* ou uma estrutura do tipo *wrapper*. Sua estrutura fundamental incorpora Metadados que identificam seu conteúdo dentro de um registro de tipos de dados padronizados e serve como a fonte para o mapeamento do código ou a tradução do código binário para formas acessíveis ou usáveis. (SHEPARD, 1998).

Uma análise mais detida no projeto UPF pela WGBH mostra claramente que se trata de uma proposta para criação de um formato para encapsular outros formatos de arquivo, o que fica claro na definição de Thomaz Shepard acima. O processo de encapsulamento e seus sinônimos (*wrapper*, *bundling*) são basicamente uma maneira de agregar, em um único arquivo, vários outros arquivos não necessariamente nos mesmos formatos: dependendo da tecnologia, esse encapsulamento agrega mais ou menos Metadados sobre os arquivos encapsulados.

Há alguns trabalhos teóricos sobre o processo de encapsulamento, como as propostas de Jeff Rothenberg *"Um encapsulamento é, afinal de contas, nada mais que o agrupamento lógico de itens"* (ROTHENBERG, 1999, p.28). Além da proposta do UPF citada acima, existem várias outras como os formatos AAF e MXF (utilizados na indústria de produção de vídeo e cinema) e inclusive alguns populares como o formato TAR ou ZIP.

Dentro de nossos objetivos nesse capítulo, sobre o processo de encapsulamento – exemplificado pela proposta UPF – é importante notar duas características fortemente presentes para reforçar a possibilidade de utilizar esse método como um método de preservação digital: primeiro, a possibilidade de um único formato ser utilizado para acomodar vários outros tipos de conteúdo como som e imagem em movimento, que vamos chamar de **Multiconteúdo** e segundo, a importância de **Metadados** extensivamente aplicados.

Outros artigos que abordam a relação entre **Formatos de Arquivo** e a **Preservação Digital** salientam a importância de algumas características já recorrentes nas propostas anteriores desse capítulo, notadamente a questão dos formatos proprietários e não proprietários (*Standards*). Cokie Anderson na **Universidade de Oklahoma** nos Estados Unidos, por exemplo, relata:

> Quando se escolhe um formato de arquivo, a escolha mais segura para os propósitos da preservação é o uso de *standards*. Mesmo que não existam garantias absolutas – *bits* e *bytes* podem se degradar ao longo do tempo – *standards* são a melhor garantia que temos. Se você tem que usar um formato proprietário, prefira um com especificação aberta (ANDERSON, 2005, p. 9).

Seguindo a mesma lógica, Andrew Williamson da **Universidade de Strathclyde** (Glasgow, Reino Unido), relata:

> Orientações de organismos de financiamento e serviços de consultoria geralmente recomendam atualmente, e em alguns casos exigem, uma abordagem baseada em *Standards* em todo o processo, argumentando que o conteúdo eletrônico deveria ser criado, armazenado, mantido e disseminado utilizando *Open Standards* sempre que possível. (WILLIANSON, 2005, p.508, grifos nossos)

Recentemente, foi apresentada à comunidade preocupada e envolvida com a problemática da preservação digital uma metodologia de análise dos possíveis riscos presentes em formatos digitais (tanto formatos de arquivo como formatos de mídias como o DVD por exemplo) que constituem acervos documentais. A referida metodologia busca subsidiar os tomadores de decisões e responsáveis por acervos digitais, de maneira que possam tomar decisões baseadas em dados concretos, principalmente no que cabe à migração de formatos digitais. Andreas Stanescu reporta a metodologia batizada de **INFORM** em um artigo de 2005: *"A metodologia INFORM define ferramentas, processos e métrica necessária para selecionar formatos mais aptos a suportar a passagem do tempo"* (STANESCU, 2005, p. 78). A metodologia, na seção sobre **Formatos Digitais**, relata **alguns possíveis** riscos que podem estar presentes e comprometer a preservação digital. Na *tabela 9* esses riscos estão relacionados (STANESCU,2005, p. 75). É importante frisar que se trata de uma **sugestão inicial**: a metodologia prevê uma análise de cada acervo por uma equipe especializada a fim

de determinar quais são os riscos efetivamente presentes. Observamos que alguns riscos aplicam-se exclusivamente a formatos de mídias (item 7) e até mesmo às pessoas envolvidas (recursos humanos) na equipe de trabalho correspondente (item 10 e 11).

Item	Fatores de Risco
1	Taxas de licença ou *royalties* podem ser necessárias
2	Especificação não disponível para inspeção independente
3	Versões anteriores da especificação são incompatíveis umas com as outras
4	Especificação muito complexa, extensa, ambígua ou pouco documentada
5	Especificação não é largamente aceita, a *de jure* ou a *de facto*
6	A especificação é única em sua classe e não pode ser mapeada para outra ou Metadados embutidos não podem ser mapeados para outros formatos
7	Especificação não permite cópias idênticas, tornando a operação de *refresh* impossível.
8	Especificação utiliza esquemas DRM, envelopes assinados, seções criptografadas ou marcas d´agua
9	Especificação permite extensões ou características largamente suportadas como *JavaScript* e outras
10	Equipe de pessoas com o conhecimento necessário não está disponível
11	Procedimentos de teste e equipe são rapidamente superados por mudanças de especificações.

Tabela 9 - Riscos de Formatos Digitais (adaptada)

Nosso objetivo nesse capítulo foi o de coletar as **características** importantes que devem estar presentes em formatos de arquivo para que esses possam ser preservados adequadamente pelo maior período de tempo possível. No próximo capítulo, iremos expor um resumo dessas **características**, bem como uma análise comparativa que originará os elementos de nosso **Modelo**.

6.3 ELEMENTOS DO MODELO DE FORMATO

O objetivo de um modelo de formato de arquivo com características mais adequadas, dentro do possível, para a preservação de documentos digitais é o de possibilitar, no nosso caso específico, a comparação e conseqüente diagnóstico das características dos formatos de arquivo efetivamente em uso com o **Modelo**. Essa comparação entre características possibilita analisar o quanto um formato de arquivo específico se aproxima do **Modelo**. Suponhamos por exemplo que nosso modelo ideal possua X características. Comparando esse modelo com um formato Z1, que possui apenas 80% das características do modelo X, teremos que o formato Z1 está no nível próximo (falta 20%) do ideal. Repetindo o procedimento para um formato Z2

que possua apenas 30% das características do modelo X temos então que Z2 está distante (falta 70%) do modelo ideal.

Num segundo momento então, podemos estender as análises para o nível individual dos formatos ou para o nível de acervos que utilizem determinados formatos identificados e comparados ao modelo ideal. No primeiro caso, uma conseqüência lógica seria que o formato Z1 é o mais adequado para a preservação de um documento digital específico. No segundo caso, analisando um acervo de documentos, se hipoteticamente a maioria dos formatos utilizados fossem do tipo Z1, então o acervo estaria em melhores condições para a preservação do que se a maioria utilizasse um formato do tipo Z2.

É muito arriscado falar em modelo ideal se considerarmos que estamos falando de produtos tecnológicos que sofrem avanços rápidos em suas características. Em função desse avanço "diário", novas características e necessidades podem surgir durante a próxima década e exigir a alteração no que chamamos hoje de modelo ideal. Assim, preferimos o termo "mais próximo do ideal", ou seja, ideal nas condições atuais de desenvolvimento tecnológico. É claro que um modelo assim necessita de constantes buscas de aprimoramento da evolução tecnológica e conseqüentes ajustes à realidade.

Nossa primeira tarefa então será a definição de quais são as características que tornam o **modelo de formato de arquivo** mais próximo do ideal. Uma comparação dos fatores e características relacionados nas tabelas de número 7, 8 e 9 evidencia sobreposição de elementos. Será necessário, então, primeiro filtrar todos os elementos que se referem exclusivamente a formatos de arquivo para então identificar aqueles que individualmente recebem designações diferentes, mas fundamentalmente, se referem ao mesmo conceito básico.

Em primeiro lugar, podemos fazer uma comparação direta entre a *tabela 7* (PDF/A) e a *tabela 8* (fatores de sustentabilidade). Ambas as tabelas são o resultado de estudos sobre

formatos digitais e por isso pode-se, claramente, verificar que se trata de grupos muito similares. Assim, há correspondência, respectivamente, entre as tabelas de número sete e oito da seguinte forma: **elemento 1 ≈ elemento 5**, **elemento 3 ≈ elemento 4**, **elemento 4 ≈ elemento 3**, **elemento 5 ≈ elemento 7**, **elemento 6 ≈ elemento 1**, **elemento 7 ≈ elemento 2**; os elementos de número dois na *tabela 7* e de número seis na *tabela 8* não possuem correspondência. A *tabela 10* resume essa análise:

Tabela 7	Tabela 8	Elemento comum
Elemento 1	Elemento 5	Independência de dispositivos externos
Elemento 3	Elemento 4	Metadados incorporados
Elemento 4	Elemento 3	Transparência do conteúdo
Elemento 5	Elemento 7	Não utilização de recursos de proteção ao acesso.
Elemento 6	Elemento 1	Abertura da especificação/formatos não proprietários
Elemento 7	Elemento 2	Adoção do formato de arquivo
Elemento 2	Sem correspondência	Auto suficiência para execução
Sem correspondência	Elemento 6	Independência de patentes (*royalties*)

Tabela 10 - Correspondência entre tabelas 7 e 8

Podemos observar que o elemento comum "Abertura da especificação e formatos não proprietários" aparece com freqüência como recomendação adequada à preservação, como exemplificado e citado anteriormente em WILLIANSON (2005) e ANDERSON (2005).

No caso da *tabela 9*, de seus onze elementos (fatores de risco para a preservação de formatos), os de número sete, dez e onze não se referem diretamente aos formatos de arquivos. O de número sete se refere a formatos de mídias como o DVD ou o CD e os números dez e onze se referem a fatores da equipe de indivíduos envolvida com o processo de preservação digital. Por outro lado, o elemento de número 1 corresponde ao elemento comum "Independência de patentes (royalties)", os elementos de número dois a cinco são riscos solucionáveis com o uso de formatos de arquivo não proprietários e abertos, portanto elemento comum "Abertura da especificação/formatos não proprietários" na *tabela 10*. O elemento seis refere-se a Metadados exportáveis e, como não há um elemento comum antes definido, trata-se de uma nova característica. O fator de número oito na *tabela 9* corresponde ao elemento comum "Não utilização de recursos de proteção ao acesso". Finalmente, o fator

de risco de número nove equivale ao elemento comum "Auto-suficiência para execução". Cabe lembrar novamente que os fatores de risco da *tabela 9* são apenas uma sugestão inicial proposta pelo autor STANESCU (2005) e não são exaustivos; portanto, esses fatores são considerados aqui como uma referência de relativa importância.

Item	Fatores na Tabela 9	Equivalências com Tabela 10
1	Taxas de licença ou *royalties* podem ser necessárias	Independência de patentes (royalties)
2	Especificação não disponível para inspeção independente	Abertura da especificação/formatos de arquivo não proprietários
3	Versões anteriores da especificação são incompatíveis umas com as outras	
4	Especificação muito complexa, extensa, ambígua ou pouco documentada	
5	Especificação não é largamente aceita, a *de jure* ou a *de facto*	
6	A especificação é única em sua classe e não pode ser mapeada para outra ou Metadados embutidos não podem ser mapeados para outros formatos	Metadados exportáveis (NOVA CARACTERÍSTICA sem equivalência anterior)
7	Especificação não permite cópias idênticas, tornando a operação de *refresh* impossível.	NÃO SE REFERE A FORMATO DE ARQUIVO
8	Especificação utiliza esquemas DRM, envelopes assinados, seções criptografadas ou marcas d´agua	Não utilização de recursos de proteção ao acesso
9	Especificação permite extensões ou características largamente suportadas como *JavaScript* e outras	Auto-suficiência para execução
10	Equipe de pessoas com o conhecimento necessário não está disponível	NÃO SE REFERE A FORMATO DE ARQUIVO
11	Procedimentos de teste e equipe são rapidamente superados por mudanças de especificações.	NÃO SE REFERE A FORMATO DE ARQUIVO

Tabela 11 - Equivalências entre tabela 9 e 10

Dessa maneira, os fatores de risco listados na *tabela 9* equivalem e, por isso, reforçam os seguintes elementos comuns listados na *tabela 10*:

- Independência de patentes (royalties);
- Abertura da especificação/formatos não proprietários;
- Não utilização de recursos de proteção ao acesso;
- Auto suficiência para execução;

É preciso ainda tecer algumas considerações sobre as características já identificadas nas *tabelas 7, 8, 9, 10 e 11*. Primeiro, sobre a nova característica evidente na *tabela 11*:

Metadados exportáveis. A possibilidade de exportação dos metadados presentes e embutidos num **arquivo digital** especificado dentro de determinado formato de arquivo é algo que dependerá principalmente do novo arquivo digital resultante dessa exportação e menos do arquivo original que contém os metadados. Assim, não se trata de uma característica relevante e desejável para nosso **Modelo** que se refere a características inerentes ao formato de arquivo escolhido para a preservação, como a característica *Transparência de conteúdo*, aliás, comum nas *tabelas 7* e *8* e listada na *tabela 10*. Dessa forma, essa "nova característica" será desconsiderada para efeitos de aproveitamento em nosso **Modelo**.

Outra característica que também será desconsiderada é o elemento comum na *tabela 10*: "Adoção do formato de arquivo". Essa característica, conforme já expusemos antes, refere-se ao quanto o ambiente externo de usuários e instituições efetivamente utiliza e aceita determinado formato de arquivo. Apesar de ser um elemento relevante a ser considerado para a preservação digital, estritamente dentro dos objetivos de nossa pesquisa, é algo de difícil mensuração no mundo real e acabaríamos utilizando referências pouco confiáveis para verificação da conformidade dessa característica. Considerando esses argumentos, optamos por desconsiderar a característica.

Outro ajuste também se faz necessário: o elemento comum "Abertura da especificação/formatos não proprietários" deve ser individualizado em dois elementos distintos: **Especificação Não-Proprietária** e **Especificação Aberta**. Ainda é preciso notar que o elemento comum "independência de patentes (royalties)" se torna irrelevante considerando essas duas características; e pode, por isso, ser desconsiderado.

Tomando como ponto de partida os elementos comuns listados na *tabela 1*0, reforçados pelos elementos presentes *tabela 11* E ignorando o elemento novo "metadados exportáveis" (*tabela 11*), o elemento comum "adoção do formato de arquivo" (*tabela 10*) e o elemento comum "independência de patentes (royalties)" (*tabela 10*), odos pelos motivos

expostos anteriormente. E, finalmente, pelo desmembramento em dois elementos distintos do elemento comum “abertura da especificação/formatos não proprietários”. O resultado final de nossa seleção de características relevantes para a preservação será a seguinte lista:

1. Independência de dispositivos externos;
2. Metadados incorporados;
3. Transparência do conteúdo;
4. Não-utilização de recursos de proteção ao acesso;
5. Especificação não-proprietária;
6. Especificação aberta;
7. Auto-suficiência na execução;

6.4 O MODELO DE FORMATO DE ARQUIVO E FORMATOS REAIS

Definidas então essas **sete características** como fundamentais para um modelo próximo do ideal, vamos aprofundar detalhadamente o significado que atribuímos a cada uma delas:

6.4.1 INDEPENDÊNCIA DE DISPOSITIVOS EXTERNOS

Essa característica se refere à capacidade de um arquivo, dentro de uma determinada versão de especificação de formato de arquivo, ser capaz de não depender de hardware ou software específicos. Obviamente sempre será necessária uma plataforma de software X sendo executado numa plataforma Y de hardware. Arquivos digitais não têm utilidade alguma sem o devido suporte dessas duas plataformas. No entanto, há casos em que determinado arquivo somente funcionará ser existir um **determinado e específico** equipamento e *software* associado.

Vamos exemplificar. Suponha a existência de um determinado dispositivo portátil como um *Handheld* ou tocador *MP4* de bolso. Há arquivos de texto ou imagem, por exemplo, feitos especialmente para esses dispositivos e somente funcionam neles, de maneira que

quando não houver mais a disponibilidade dos equipamentos no mercado, da mesma forma não poderemos mais executar esses aplicativos.

Claro que esse fenômeno ocorre mais ou menos com qualquer *software* em qualquer computador. Se considerarmos uma plataforma comum atualmente de *hardware* (processador qualquer do fabricante *Intel* por exemplo) ou *software* (sistema operacional e demais recursos da empresa *Microsoft* por exemplo), há arquivos que somente funcionam nesse conjunto tecnológico atual. Não poderiam ser compatíveis com versões antigas pois antigamente os engenheiros não poderiam imaginar como seriam hoje as especificações de arquivo hoje e nem podem ser compatíveis com a tecnologia que ainda será inventada. Porém, há arquivos que podem ser acessados (executados e ter seu conteúdo acessado) em diferentes aplicativos hoje. Isso pode ocorrer por exemplo através de *browser* para Internet, em diferentes aplicativos e até em diferentes sistemas operacionais. É o caso de arquivos gerados numa especificação qualquer do formato PDF; esse tipo de arquivo pode ser visualizado no mesmo aplicativo de navegação na Internet ou em máquinas diferentes como um modelo *MacIntosh* ou PC de mesa. Num outro extremo estão os arquivos de bases de dados que necessitam de um sistema gerenciador de bases de dados (conjunto específico de *software*) que por sua vez foi projetado para ser executado somente em determinados tipos de *hardware*. Trata-se, portanto, do **quanto** um arquivo é dependente de plataformas tecnológicas de *hardware* e *software*.

Pode-se responder essa característica verificando se a especificação do formato de arquivo foi projetada para funcionar em uma plataforma específica ou em plataformas diferentes e comuns no mercado. Alguns formatos proprietários e com especificação fechada, como têm sido com a maioria dos arquivos da fabricante *Microsoft*, em função do grande alcance de mercado e popularidade, atendem esse requisito. É o caso do formato **DOC** para editores de texto, que pode ser acessado em diferentes tipos de computadores, sistemas

operacionais e até aplicativos. Por outro lado, alguns bancos de dados que se baseiam na linguagem padrão para bases de dados *Structure Query Language* (SQL), em função da possível utilização de uma série de aprimoramentos específicos de cada fabricante, podem ser virtualmente impossíveis de serem migradas ou executadas em outras plataformas que não a original. No caso específico de bancos de dados, caso comum de dependência de plataformas originais, as tecnologias XML tem sido apontadas como uma possível solução *"Num arquivo XML adequadamente projetado, cada documento terá um auto-conteúdo e toda a informação necessária para reconstruir o significado original do negócio contido na informação armazenada no documento"* (WILLIAMS *et al*, 2000, p. 696).

6.4.2 METADADOS INCORPORADOS

Essa parece ser uma característica das mais importantes para a preservação de um arquivo digital e está relacionada à preservação de diferentes aspectos de um documento digital, indo além do simples aspecto tecnológico. O uso de Metadados extensivos, ou seja, a existência da maior quantidade possível de informação relacionada ao documento – como unidade ou pessoa produtora, descrições do conteúdo, relações do conteúdo com outros documentos, direitos autorais e muitas outras – possibilitará viabilizar questões importantes de **gestão de documentos** como aspectos de autenticidade e recuperação da informação (descrição, classificação arquivística, catalogação biblioteconômica e outros). O modelo de referência OAIS citado em (ARMS; FLEISCHHAUER, 2005, p. 4) lista a necessidade de várias categorias de Metadados como a "representação (permite que os dados sejam montados e utilizados como informação)", "referência (para identificar e descrever o conteúdo)", "contexto (por exemplo, para documentar o propósito para a criação do conteúdo)", "fixidez (permitir verificações na integridade dos dados do conteúdo)" e "proveniência (para documentar a cadeia de custódia e qualquer mudança desde que o conteúdo foi originalmente criado)".

No entanto, não cabe à especificação do formato de arquivo a própria existência dessas informações, ou seja, não há como o formato de arquivo exigir que essas informações estejam presentes ou não. Mas pode caber a responsabilidade de, caso essas informações existam e exista interesse dos responsáveis humanos, encapsular esses Metadados juntamente com o conteúdo do documento num mesmo arquivo. A vantagem principal por trás dessa abordagem está em diminuir a necessidade de uma base de dados (que pode ser um outro arquivo) com os Metadados, eliminando assim um possível problema futuro em migrações, por exemplo.

6.4.3 TRANSPARÊNCIA DO CONTEÚDO

Essa característica se aplica particularmente para documentos digitais textuais e se refere à possibilidade de leitura direta do conteúdo textual presente nos arquivos. No caso de documentos não textuais, será importante que o conteúdo textual referente aos Metadados incorporados no arquivo também sejam de fácil leitura humana. Como já expusemos antes, um arquivo digital contém muito mais que o conteúdo propriamente, seja ele texto, imagem, som ou combinações desses. A característica da transparência exige que a parte do arquivo (bitstream da seqüência de bits total) que corresponde ao **texto** do **Documento Textual** possa ser lida diretamente e na ordem original do texto ou, no caso de Metadados, que exista legibilidade entre os campos preenchidos e a função específica de cada campo.

Infelizmente, em termos tecnológicos a implementação desse recurso não é tão simples como pode parecer em princípio. Há vários problemas relacionados ao acesso de texto num arquivo. O primeiro deles se refere à existência de diferentes idiomas no planeta, cada um com seu sistema próprio de registro gráfico:

> Programadores estadunidenses acostumados a trabalhar com 128 caracteres do conjunto de caracteres US ASCII, precisam ter em mente que bem mais que 250 caracteres são necessários para lidar com duas dúzias ou mais de línguas européias baseadas no alfabeto românico. Outros alfabetos – cirílico, grego, hebreu, árabe, devanagari, sânscrito e outros – acrescentam centenas de outros caracteres e os ideogramas chineses, japoneses e coreanos acrescentam dezenas de centenas mais. (KIENTZLE, 1995, p. 19)

Representar os diferentes tipos de textos humanos em termos computacionais de maneira adequada exige a utilização de um sistema de **Tabelas de Códigos**. Assim, para cada código corresponderá um sinal gráfico (caractere, número, indicativos de acentuação e pontuação e outros.). Numa tentativa de padronizar o conceito de **Código** empregado nas referidas tabelas, David Connolly prefere *"é tipicamente um símbolo cujas várias representações são compreendidas da mesma maneira por uma comunidade de pessoas"* (CONNOLLY, 1995). É importante que as **Tabelas de Códigos** sejam amplamente conhecidas, de acesso público irrestrito e aceitas oficialmente por órgãos independentes. Atualmente existem diversas "tabelas" com essa finalidade como a **UNICODE** (ISO 10.646).

A especificação original do formato de arquivo poderia conter uma descrição do significado do texto contido no conteúdo ou nos Metadados; porém, o uso de Tabelas Oficiais de Códigos, como a UNICODE, é um recurso extremamente mais adequado e seguro.

6.4.4 NÃO UTILIZAÇÃO DE RECURSOS DE PROTEÇÃO AO ACESSO

Os recursos de proteção ao acesso podem ser implementados utilizando-se diferentes tecnologias disponíveis atualmente, algumas viáveis para determinadas especificações de formatos de arquivo outras nem tanto. Há muitas razões possíveis para se aplicar mecanismos de proteção ao acesso em arquivos digitais, uma das mais contundentes é a proteção a direitos autorais. Nesse caso específico, os fabricantes interessados em proteger cópias não autorizadas podem lançar mão de recursos como a criptografia, onde somente através da posse de uma senha específica (ou um número de série) o usuário terá acesso ao conteúdo. Em outros momentos, pode-se ter acesso parcial aos recursos de um documento digital, um exemplo nessa direção são arquivos no formato PDF com restrições impostas como a proibição de impressão do conteúdo ou extração de texto do documento.

Freqüentemente, encontramos também o procedimento de encapsular vários arquivos num mesmo arquivo (formatos de arquivo ZIP ou TAR, por exemplo) e aplicar processos de

criptografia no arquivo **encapsulado**, que, além disso, pode ter passado por um processo de **compactação** por *software*.

Os recursos de proteção ao acesso do conteúdo do documento, integralmente ou parcialmente - sendo as técnicas mais comuns os procedimentos de encapsulamento e compactação - não são, em geral, procedimentos irreversíveis ou obrigatórios, em geral. Ou seja, após o uso corrente em ambiente de negócios dos documentos e após a decisão de preservá-los para a posteridade, os recursos de proteção podem e devem ser removidos. Uma especificação de formato de arquivo que necessite **inerente** e **obrigatoriamente** e não permita a **remoção** desses mecanismos no futuro como a **criptografia**, não é adequado para efeitos de preservação digital.

Os responsáveis pelos procedimentos de preservação digital precisam ter pleno controle sobre os objetos digitais sob sua responsabilidade. Pelo menos se considerarmos o cenário atual tecnológico. Fazer preservação digital envolve procedimentos de **cópias** de arquivos para novos suportes tecnológicos (e, até mesmo, procedimentos comuns de *backup*) ou **migração** dos formatos originais para novas opções tecnológicas.

6.4.5 ESPECIFICAÇÃO NÃO-PROPRIETÁRIA

Arquivos gerados por *software* e suas correspondentes especificações técnicas são, em geral, produtos de uma indústria específica que busca lucros. Como corolário da disputa com a concorrência por produtos inovadores que atendam necessidades de mercado e assim alavancando, assim, os investimentos, as empresas comumente estabelecem mecanismos de proteção como **segredos industriais**. Com a honrosa exceção dos produtos *Open Source* a imensa maioria de produtos de *software*, incluindo a especificação dos formatos de arquivo, não têm seus detalhes técnicos divulgados ao público por razões comerciais. É preciso ressaltar que os produtos com código aberto são uma novidade bastante recente no mercado e ainda minoria entre os produtos disponíveis.

Por outro lado, os procedimentos atualmente disponíveis para efetivar a preservação digital, como a **emulação**, **encapsulamento** e a **migração**, precisam ter acesso aos detalhes técnicos dos formatos de arquivo. O resultado natural dessa argumentação é o *status* pouco promissor dos formatos de arquivo protegidos por segredo industrial. Há uma quase unanimidade nas vantagens para a preservação digital no uso de formatos de arquivo **não proprietários**.

Por outro lado é preciso lembrar que um formato de arquivo com especificação proprietária não tem, necessariamente, sua especificação fechada ao acesso público. Formatos de arquivo bastante utilizados atualmente para a preservação digital, como o formato PDF ou o formato TIFF, são proprietários, apesar da estratégia dos fabricantes correspondentes de liberar o acesso à especificação do formato. Essa situação, apesar de menos ruim do que a de formatos proprietários com especificação protegida, como é o caso de quase todos os produtos da fabricante *Microsoft*, não é a ideal. Nada impede que os fabricantes decidam alterar sua estratégia e passar a não mais divulgar sua especificação ou, pelo menos, a especificação de novas versões a serem lançadas ou até resolverem liberar apenas parte das informações necessárias.

Essa característica, em nosso modelo, portanto, precisa ser analisada em combinação com a característica seguinte: especificação aberta.

6.4.6 ESPECIFICAÇÃO ABERTA

A característica "especificação de formato de arquivo aberta" significa que o público em geral pode ter acesso aos detalhes técnicos correspondentes a determinado formato de arquivo. Note-se aqui que não se trata de ter acesso ao código fonte dos aplicativos que geram os arquivos (como o aplicativo *Word* da *Microsoft*). Por trás de processos como a migração de formatos de arquivo está a intenção de (re)montar um arquivo (que esteja numa determinada especificação de formato de arquivo) em uma nova estrutura (especificada pelo novo

formato). O **novo** *software* necessário para executar o **novo** arquivo conterá um **novo** código fonte que não precisa ser necessariamente igual ou até parecido com o código fonte original do primeiro **aplicativo**.

Como já exposto na seção anterior, uma **especificação aberta** pode ser encontrada mesmo em casos onde o formato de arquivo é **proprietário**. Dessa maneira, quando dissemos aqui que desejamos como característica relevante do modelo a existência de especificação aberta, na verdade, estamos nos referindo ao uso de **Normas Oficiais**. Não basta que o público tenha acesso à especificação do formato de arquivo; é preciso ter a segurança de que essa especificação se manterá aberta no futuro. Uma maneira segura de fazer isso é através da criação, adoção ou transformação de uma **especificação de formato de arquivo** em **Norma Técnica**, como por exemplo através da *International Standard Organization* (ISO). Foi assim que, em 2005, após anos de discussão entre várias organizações, foi homologada oficialmente a norma **ISO 19005-1:2005(E)** que corresponde ao que ficou popularmente conhecido como formato de arquivo PDF/A. Além da segurança de acesso à especificação no futuro, a existência de uma norma oficial pública implica também, a exemplo das normas ISO, na existência de uma boa documentação sobre a especificação. O acesso direto a uma especificação não garante necessariamente sua compreensão; é preciso que os registros tenham sido feitos de maneira clara, coerente e com todas as informações necessárias. Sobre isso, num trabalho que tentou fazer uma coletânea de formatos de arquivo utilizados em imagens de todos os tipos, justamente para documentá-los adequadamente para a posteridade, em outras palavras, descrever os detalhes técnicos dos formatos de arquivo, um dos argumentos utilizados pelos autores sobre a importância dessa documentação foi *"Nem todos os formatos são documentados, porém, e alguns documentos são tão esparsos, pobremente escritos ou desatualizados que são essencialmente inúteis"* (MURRAY, VanRYPER, 1994, p. xv).

6.4.7 AUTO-SUFICIÊNCIA NA EXECUÇÃO

Arquivos de computador, como um agrupamento de *bits* organizados de determinada maneira, são executados para desempenhar diversas funções, a maioria delas invisíveis e imperceptíveis diretamente por nós, humanos. Por exemplo, um programa pode estar sendo executado em segundo plano e invisivelmente para monitorar a existência de um vírus. Em nosso campo de interesses nessa dissertação estão apenas os arquivos que, quando executados, permitem a nós, humanos, receber informações **visuais**, **auditivas** ou ambas simultaneamente e interpretá-las cognitivamente: é o que ocorre quando um arquivo contendo uma imagem fotográfica é executado, ou o mesmo com um arquivo contendo o áudio de um discurso de posse ou até um vídeo documentário sobre determinado tema. Ocorre, porém, que um mesmo arquivo pode precisar de outros elementos (que podem ser até mesmo outros arquivos) para serem executados, dependendo do grau de complexidade de determinada especificação de formato de arquivo.

Um dos exemplos mais simples nesse sentido se refere ao uso de *fontes de texto*, normalmente apenas chamadas de **Fontes**. Uma "Fonte, no sentido aqui utilizado, é um elemento que se refere ao aspecto visual de determinado conjunto de caracteres e recebem um nome específico: por exemplo, a fonte utilizada no presente texto chama-se *Times New Roman*, mas poderia ser a `Courier` ou tantas outras centenas e talvez milhares de opções. Tecnologicamente falando, existem arquivos em separado especialmente para cada **Fonte** específica. Quando um texto é produzido, através de um editor de textos, faz uso de determinadas fontes e então salva um arquivo correspondente contendo o texto editado e, opcionalmente, as fontes (além de outros elementos). Esse arquivo não incorporará necessariamente o arquivo com as fontes utilizadas na edição. Na prática, o arquivo gerado pode conter apenas uma referência ao nome da(s) Fonte(s) utilizada(s) e quando for executado exigirá a presença do arquivo com a Fonte correspondente ou outra similar.

Aspectos relacionados à autenticidade de um documento textual podem estar relacionados à utilização do tipo específico original de fonte de texto e o uso de fontes alternativas pode comprometer o quanto se confia no documento original. Apesar de haver casos onde somente o texto em si é o mais importante, não importando se ele é visualizado (lido) em qualquer fonte disponível. De qualquer forma, a situação mais segura para efeitos de preservação digital adequada é a incorporação do arquivo com as fontes originais no arquivo que contém o texto produzido.

É comum também encontrarmos um documento digital predominantemente textual, mas que contém imagens (fotos, desenhos) junto ao texto. Um arquivo digital de uma folha de jornal impresso, enquanto ainda na fase de edição, é um excelente exemplo. Nesse tipo de arquivo facilmente encontramos diversas outras ilustrações junto ao texto. Na *figura 5* exemplificamos uma página de um grande jornal brasileiro, o arquivo foi "baixado" no sítio do jornal. Através de ferramentas adequadas extraímos do arquivo abaixo (pdf versão 1.4 gerado pelo *Acrobat Destiller*) as imagens contidas (oito no formato jpg), nesse mesmo arquivo, identificamos 18 fontes utilizadas, como exemplificado na *figura 5*.


O ESTADO DE S. PAULO

SEXTA-FEIRA

EUA crescem acima do previsto e aliviam mercado

PIB americano teve alta de 3,3% no segundo trimestre; bolsas subiram

É a recessão?

STF pode diminuir reserva indígena

Serra quer proibir fumo em locais fechados

Obama ataca Bush e diz que cortará impostos

PCC pagou a ONG para se infiltrar no Congresso

CADERNO 2

Alcova de Juca Ferreira

Figura 5 - Arquivo digital (pdf) de página de jornal (parte)

Portanto, a característica de possibilitar a inserção de outros arquivos, fontes ou qualquer outro recurso necessário à correta e completa execução idêntica ao original criado é uma característica importante em nosso modelo de formato de arquivo.

6.5 ÚLTIMAS CONSIDERAÇÕES

A *figura 6* ilustra conceitualmente todas as sete características apresentadas anteriormente na forma de um objeto que chamamos de **Modelo** para a preservação digital.

Figura 6 - Modelo Completo para preservação digital

Têm surgido propostas e casos concretos de formatos de arquivo objetivando a preservação digital. Era de se esperar que esses formatos de arquivo contivessem todas as características relacionadas no capítulo anterior. De fato, muitas delas são implementadas, o formato de arquivo PDF/A como norma ISO, por exemplo, incorpora todas as características. É preciso, no entanto, fazer duas críticas aos formatos de arquivo especialmente desenvolvidos para a preservação digital.

Primeiro, uma especificação de formato de arquivo desenvolvido visando à preservação digital não necessariamente conterá todas as características desejadas. Por exemplo, pode não conter a previsão para extração textual dos Metadados descritos no documento ou até mesmo não prever a incorporação de Metadados. Apesar da tendência de que essas características de fato estejam presentes, não há garantias nesse sentido.

Em segundo lugar, ainda não surgiu uma especificação de formato de arquivo que possa ser universalmente utilizada em todos os tipos de arquivos (texto, som, imagens e outros). O formato de arquivo PDF/A, na versão atual, basicamente foi feito pensando na preservação de documentos textuais. Dessa maneira, ainda há uma forte necessidade de se lançar mão de vários formatos de arquivo num mesmo acervo, nem todos contendo necessariamente todas as características desejáveis para a preservação digital.

Por isso é tão importante um procedimento de verificação dos formatos de arquivo efetivamente em uso e suas características ideais. É nesse sentido que surge a importância de um modelo como referência de análise.

Um outro aspecto que é preciso ressaltar ainda mais é o de que algumas características presentes na especificação de formato de arquivo se referem ao **uso potencial**.O uso obrigatório para algumas características transcende a própria especificação do formato. É o caso de permitir encapsulamento de Metadados ou Fontes utilizadas em documentos textuais, o formato permite esses encapsulamentos mas não obriga seu uso. Essa obrigatoriedade, no entanto, pode ser imposta na produção desses arquivos digitais, através da utilização de outras ferramentas tecnológicas.

Com relação à característica das dependências externas, como descrita anteriormente, cabe esclarecer que se refere à necessidade imposta na especificação do formato pelo uso obrigatório de determinado *hardware* ou *software*. Note-se que todo arquivo de computador necessitará de *hardware* e *software* para ser executado; no entanto, é uma situação diferente

quando há uma necessidade **específica de dispositivos**, principalmente se se tratar de dispositivos proprietários e de fácil desatualização tecnológica.

Por último, é necessário registrar que existem várias propostas objetivando a identificação formal e detalhada de um formato de arquivo. Expomos esse tema em mais detalhes na seção que trata do projeto PRONOM do Arquivo Nacional do Reino Unido.

7 COLETA DE DADOS

7.1 MÉTODOS E PROCEDIMENTOS

7.1.1 INTRODUÇÃO

Nesse capítulo, trataremos dos dados coletados, ou seja, estamos tratando aqui da metodologia utilizada no que diz respeito ao universo de pesquisa, às amostras coletadas e à análise correspondente. Expomos, a seguir, o objetivo da coleta de dados, quais dados foram coletados e como foram coletados e analisados.

7.1.2 UNIVERSO DE AMOSTRA DE DADOS

A coleta de dados num determinado universo de pesquisa, no nosso caso órgãos do poder judiciário brasileiro, limita-se entre o desejo ideal de coletar dados de todas as unidades desse universo e a necessidade prática de selecionar uma amostra que corresponda ao universo pesquisado, ou seja, que represente corretamente esse universo.

Nosso universo de pesquisa é composto por órgãos do **Poder Judiciário Brasileiro;** o **Anexo II** contém a lista completa desses órgãos. Trata-se de um universo de 89 órgãos a serem pesquisados. Em função da própria organização legal desses órgãos, podemos organizá-los em grupos similares, o que facilitará a análise do universo. A *tabela 12* resume os grupos dos órgãos.

Nome do Grupo	Unidades
Conselho da Justiça Federal (CJF)	1
Tribunais Superiores	5
Justiça Federal de 1ª e 2ª Instâncias (TRFs)	5
Justiça Estadual/Distrital (TJs)	27
Justiça do Trabalho de 1ª e 2 ª Instâncias	24
Justiça Eleitoral (TREs)	27
Total	**89**

Tabela 12 - Grupos no Universo de Pesquisa

O *gráfico 1* mostra o percentual desses grupos em relação ao universo total. O **Anexo III** contém a organização por Unidade da Federação.

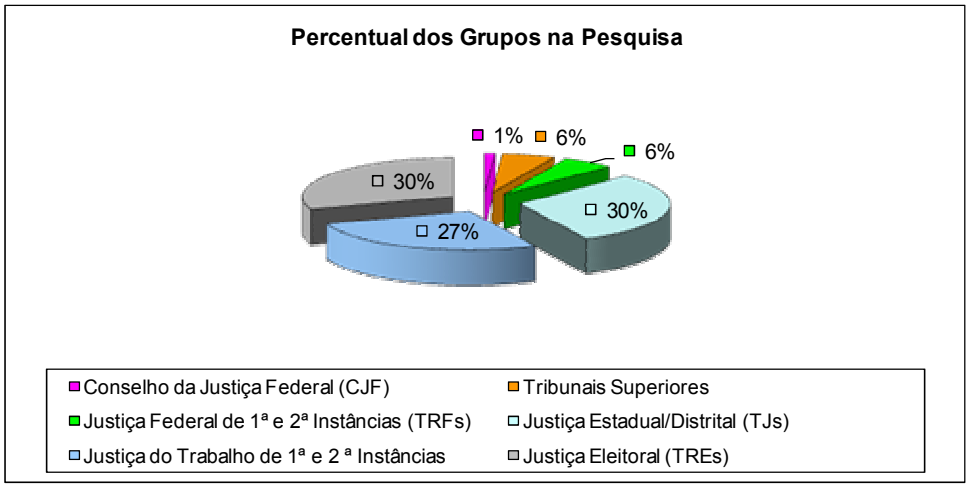


Gráfico 1 - Grupos de pesquisados

7.1.3 WEB ARCHIVING

Nessa pesquisa utilizamos coleta de dados em sítios da Internet. Esse tipo de coleta traz alguns problemas que podem comprometer a qualidade dos itens coletados no que diz respeito, principalmente, à possibilidade de verificação da coleta. Isso ocorre em função da característica dinâmica dos conteúdos na Internet. Nesse contexto, como coletar dados em sítios que possam ser considerados confiáveis do ponto de vista da verificação científica? A resposta que encontramos foi a utilização de técnicas de arquivamento de páginas na Internet (*Web Archiving*).

Segundo Neils Brügger,

> A razão pela qual a pesquisa na Internet se preocupa com o arquivamento na Internet é porque, em algum ponto, a pesquisa que tem a Internet como objeto concreto de estudo precisa estabilizar e manter esse objeto para preservá-lo, ou para uso imediato de análise e/ou para documentação posterior e, dessa forma, como uma base para criticar e discutir a análise efetuada. (BRÜGGER, 2007, p. 9)

Em termos práticos, o problema todo consiste em fazer uma cópia de todo o conteúdo de um sítio disponível em um determinado endereço da Internet. Essa cópia deverá conter todos os arquivos da página e é salva em disco local.

A visualização de uma página *web* não é apenas o resultado não de um arquivo, mas de um conjunto de vários tipos diferentes de arquivos, como imagens, texto, código em linguagem *html* e/ou outras. Se executado corretamente, o processo de *web archiving* num determinado sítio permitirá, sempre que necessário, visualizar o sítio *off-line*, ou seja, localmente no computador, sem o acesso à Internet. É possível também salvar individualmente em disco os arquivos disponíveis e relacionados através dos *hiper-links* nas páginas do sítio.

Esse processo de coleta de arquivos possui um limite. Esse limite se refere, principalmente, ao espaço em disco necessário para armazenar os arquivos. Uma página qualquer na Internet segue o princípio do uso de *links*, ou seja, clicar num determinado objeto (palavra, figura ou outros elementos) remete a uma outra página, arquivo ou aplicativo de busca, por exemplo. Assim, capturar uma página na Internet consiste em copiar todos os arquivos cujos *links* remetem a esses mesmos arquivos. Porém, esse princípio pode ser teoricamente infinito já que uma página pode remeter a outras páginas fora do contexto da original. Efetuar o *web archive* de uma página, portanto, exige definir em quantos níveis de *hiper-links* os arquivos serão baixados e mesmo se serão copiados também arquivos em outros contextos de páginas e endereços. É possível também definir quais tipos de arquivos serão copiados, por exemplo, não copiando arquivos do tipo aplicativo (programas que executam algum tipo de código).

7.1.4 COLETA DE DADOS *ON-LINE*

Efetuamos coleta de dados *on-line*, especificamente disponível em sítios da Internet correspondentes aos órgãos do judiciário listados no Anexo II ao final dessa dissertação, o anexo IV contém os endereços na Internet. Para que essa amostra possa ser considerada cientificamente confiável, além de verificável, optamos por um método de **coleta de dados automatizado**. Esse método possibilita uma coleta uniforme em todos os sítios e o

armazenamento de arquivos para verificação. Como os procedimentos de coleta são uniformes (tempo de *download*, tipo e tamanho dos arquivos baixados, além de outras diretivas), é possível uma comparação das diferenças entre os sítios da Internet objeto da coleta ou assegurar que a coleta individual possui o mesmo peso na composição da amostra. Além disso, o armazenamento dos arquivos possibilita um "congelamento" do sítio na data da coleta, o que mitiga a característica da dinamização da Internet.

Nosso processo de *web archiving*, como descrito na seção anterior, adotou as seguintes pré-definições de limites para o processo de cópia do sítio:

Nome do parâmetro	Configurado para	Descrição parâmetro
Alcance dos *links*	3 níveis	Especifica até quantos *links* os arquivos serão copiados.
Tipo de extensões	Extensões típicas utilizadas em arquivos textuais, imagens e sons.	Especifica qual o tipo de arquivo que será copiado.
Tamanho dos arquivos	Mínimo de 20Kb	Especifica o tamanho mínimo dos arquivos que serão copiados.
Servidores de páginas	Somente o servidor original do sítio.	Especifica até quantos servidores de páginas web serão acessados e copiados.

Tabela 13 - Parâmetros para *web archiving*

A opção **"alcance dos links"** refere-se a quantos arquivos serão copiados a partir do endereço inicial (nível 0) do sítio. Exemplificando: a página inicial é o nível 0, que pode conter *links* para um nível 1, nesse nível pode haver *links* para um nível 2. Nesse último nível, caso haja *links* para outras páginas ou arquivos, serão ignorados.

A opção **"tipo de extensões"** refere-se aos formatos de arquivo que serão copiados no processo de *web archiving*. Uma maneira de identificar um determinado formato de arquivo é através de sua extensão, três caracteres em geral. Por exemplo, um arquivo estruturado no **formato de arquivo** *Portable Document Format* (PDF) terá um nome e uma extensão .pdf (NomeQualquerDoArquivo.pdf). A limitação no número de extensões que serão copiadas é importante pois estamos interessados apenas em certos tipos de arquivos: imagem, texto e som, os quais, em princípio, devem corresponder a documentos do órgão mantenedor do sítio.

Dessa forma, estamos impedindo a cópia de arquivos com extensões como a *.exe* ou *.jsp* que correspondem a pequenos aplicativos executáveis e não a documentos propriamente falando.

A opção **"tamanho dos arquivos"** é importante pois, mesmo no grupo dos tipos de arquivos que procuramos, podemos encontrar exemplares sem utilidade para nossa pesquisa, notadamente nos tipos de arquivos para imagens; por exemplo, o tipo *.gif* é bastante utilizado para arquivos que têm a utilidade de funcionar como botões nas páginas dos sítios ou apenas como elementos decorativos. Apesar do sítio como um todo poder ser **considerado um documento** e daí todo e qualquer elemento possui sua importância nesse contexto, não é nosso objetivo nesse trabalho utilizar esse tipo de amostra, que está **fora de nosso escopo**.

Por fim, especificamos também a opção **"servidores de páginas"**, que se refere aos servidores de Internet que serão objeto da cópia de arquivos. Um sítio qualquer estará sempre alocado dentro de um certo servidor para acesso dos usuários da Internet. Ocorre que um servidor pode remeter a acessos em um outro servidor, por exemplo, um servidor de um determinado órgão da justiça pode remeter aos acessos no servidor do Diário Oficial da União, o que fugiria à delimitação de órgãos pesquisados em nossa amostra.

Em primeiro lugar, todas essas restrições objetivam assegurar a coleta de documentos realmente relevantes à nossa análise e, em segundo lugar, limitar o espaço físico necessário para armazenar os arquivos. Sem essas limitações, teoricamente, um processo de *web archiving* poderia ocupar um espaço infinito em disco.

7.1.4.1 COLETA EM *WEB ARCHIVING*

Existem diversos aplicativos disponíveis para efetivar o processo de *Web Archiving* de sítios na Internet, no nosso caso, mais especificamente, endereços no protocolo http://. Para a escolha de um aplicativo para *Web Archiving*. Consultamos o trabalho de David Kellog (KELLOG, 2005) sobre *software* livre, o que em princípio facilitaria o trabalho, pelo menos com relação aos custos. No entanto, percebemos que se tratava de aplicativos excessivamente

sofisticados para nossos objetivos; naquele trabalho se afirmava que "Uma replicação próxima da perfeição é necessária para dar aos futuros usuários a sensação e visual reais do sítio original" (KELLOG, 2005, p. 7). Nossos objetivos não envolvem a necessidade de cópia e replicação de sítios mas tão somente a captura de arquivos utilizados como documentos no sítio, como fotografias no formato JPG ou relatórios no formato PDF.

Um outro trabalho disponibilizado pelo Centro para Pesquisa sobre Internet (*The Centre for Internet Research*) de autoria de Thomasen Bo Hovgaard, relacionava outros aplicativos os quais submetemos a testes de instalação e uso. Como resultado desses testes, surgiu o aplicativo *Web Reaper* (http://www.webreaper.net) como uma alternativa viável. Na verdade, o próprio documento de Bo Hovgaard não aprova o aplicativo *Web Reaper* como melhor alternativa para cópia de sítios *web "o processo de archiving levado a cabo por aqueles programas [outros avaliados além do Web Reaper] têm mais defeitos do que aqueles outros dois programas antes mencionados [os escolhidos nas avaliações]"* (BO HOVGAARD, 2004, p. 9, inserções nossas). Pelo menos, não ideal para cópias perfeitas dentro dos objetivos de arquivar sítios disponíveis na Internet. Porém, o aplicativo em questão se mostrou simples de instalar e utilizar, além de não exigir custos financeiros. De qualquer forma, através dele nos foi perfeitamente possível executar o *download* de todos os arquivos que necessitamos para compor nossas amostras de pesquisa. A *figura 7* mostra a tela principal do aplicativo.

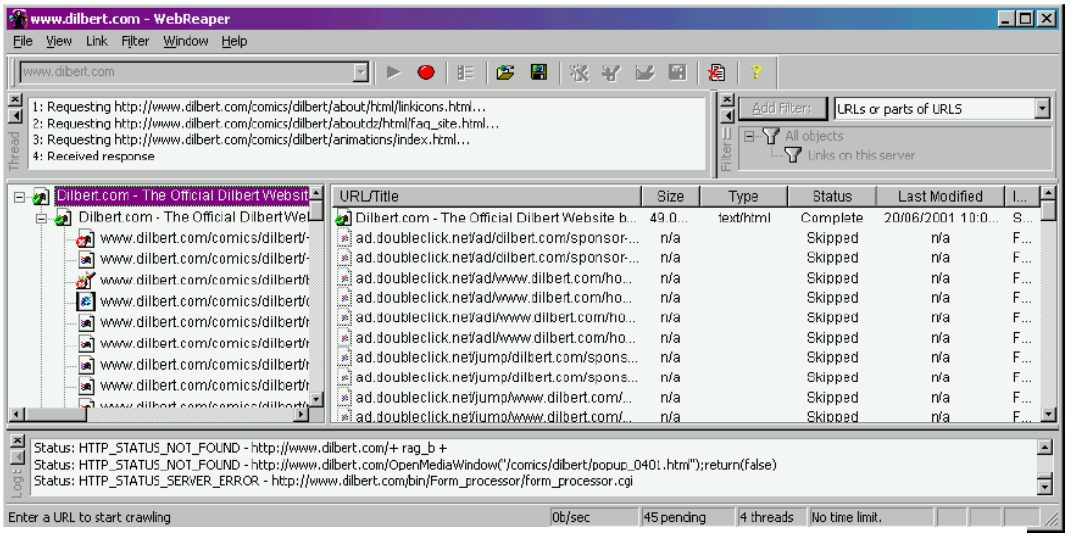


Figura 7 - Tela inicial *WebReaper*

A *figura 8* mostra um exemplo de dados coletados através do aplicativo acima. A estrutura de pastas que vemos na parte esquerda da figura (buscasite, certidaoquitacao, consultaCnpj, etc.) refere-se à estrutura do sítio que foi acessado no processo de *archiving*.

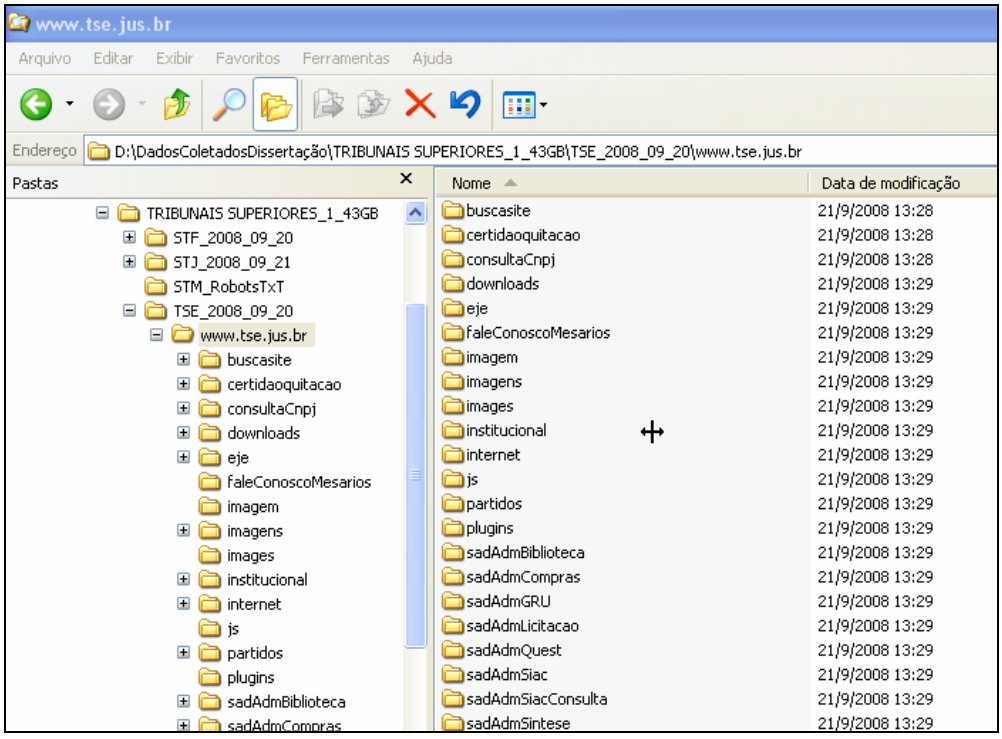


Figura 8 - Exemplo de *archiving* para um sítio da Internet (http://www.tse.jus.br)

Na *figura 8*, no lado direito, podemos ver quatro pastas principais, cada uma com uma sigla de três letras (STF, STJ, STM e TSE) e datas (as datas em que efetivamos o processo de *web archiving*). No entanto, ao lado da sigla STM encontramos a mensagem Robots.txt. Inserimos esse nome para indicar uma restrição encontrada. Nesse caso, os arquivos referentes ao órgão STM (Superior Tribunal Militar) não foram copiados em *download* em função de uma diretiva do órgão que impede esse processo. Por diferentes motivos, qualquer detentor de um sítio na Internet pode sinalizar que não deseja que os arquivos em seu sítio sejam copiados. É claro que o processo de entrar num sítio qualquer da Internet implica em copiar arquivos pelo *browser* (visualizador *Internet Explorer* por exemplo) para visualização em nossos computadores, mesmo quando não estamos fazendo um *download* propriamente. Essa diretiva se aplica para processos de cópias automáticas através de programas especialmente desenvolvidos para esse fim, como é o caso do software ***Web Reaper***. Aliás uma outra vantagem desse aplicativo é que ele não permite que essa diretiva seja ignorada, ou seja, se um sítio possui a diretiva *Robots.txt*, ele a obedecerá obrigatoriamente. Vale lembrar também que alguns sítios impõem essa diretiva restritiva às cópias automáticas apenas em partes de seus sítios.

Trata-se de uma questão entre ética, mais especificamente seguir o desejo do detentor do sítio, e o direito de arquivar certos sítios na Internet, como expõe o professor Neils Brügger:

> Com relação a condições específicas de *archiving*, uma questão a ser considerada é sobre se devemos seguir a orientação do sítio de que expressamente não deseja ser arquivado (expressa no arquivo robots.txt, que certos *softwares* de *archiving* podem ser configurados para seguir ou não). (BRÜGGER, 2005, p. 13).

Em nosso trabalho optamos por seguir rigorosamente o desejo dos órgãos que impuseram a diretiva de coleta de informações *Robots.txt*. Essa diretiva, no entanto, não comprometeu a qualidade de nossa coleta, pois poucos sítios impuseram essa diretiva para todo um sítio, sendo apenas 1 (um) no grupo de Tribunais Superiores e no grupo da Justiça

Estadual, 2 (dois) no grupo da Justiça Federal e 3 (três) no grupo da Justiça do Trabalho. 7 (sete) órgãos, portanto, o que equivale a menos de 8% da quantidade total de órgãos (89).

7.1.5 IDENTIFICAÇÃO DOS FORMATOS DE ARQUIVO

A partir dos parâmetros definidos na seção anterior, para cada órgão do poder judiciário dentro de nosso universo de pesquisa, o aplicativo utilizado para o procedimento de *web archiving* produz um conjunto de arquivos em disco. A quantidade de arquivos, o tamanho desses arquivos em *kilobytes* no disco, o formato de arquivo e a versão do formato de arquivo são elementos variáveis que dependem de uma série de fatores para cada sítio pesquisado.

Como nosso principal objetivo se relaciona à análise dos formatos de arquivo efetivamente em uso nos órgãos do universo de pesquisa, faz-se necessário um processo de identificação das características de cada arquivo. Há várias maneiras de se fazer a identificação das características de um arquivo. No entanto, estamos lidando com grandes quantidades de arquivos, da ordem de centenas e até milhares. Em função disso, optamos pela utilização de uma ferramenta de *software* que automatiza o processo de identificação de arquivos: o aplicativo ***DROID***.

7.1.6 O PROJETO *PRONOM* E O APLICATIVO *DROID*

O Arquivo Nacional do Reino Unido, dentro de seu programa de preservação digital, desde o ano de 2002 mantém o projeto PRONOM. "Sua gênese reside na necessidade de ter acesso imediato a informações técnicas confiáveis a respeito da natureza dos documentos eletrônicos agora sendo armazenados em nosso Arquivo Digital"[42]. O projeto ***PRONOM*** existe em função do reconhecimento de que para se efetivar ações de **preservação digital** há

[42] *Background* sobre o projeto *PRONOM*, disponível em < http://www.nationalarchives.gov.uk/aboutapps/PRONOM/default.htm >. Acesso em 02 de julho de 2008.

a necessidade de subsídios técnicos sobre documentos digitais, "Informações técnicas sobre a estrutura daqueles formatos de arquivo e os produtos de software correspondentes são portanto um pré-requisito para qualquer ação de preservação digital"[43].

A quarta edição do projeto ***PRONOM***[44] disponibilizou uma ferramenta desenvolvida em *software* livre que pode ser utilizada para identificar informações técnicas sobre arquivos digitais. Mais que isso, a ferramenta permite que essa tarefa seja executada com bastante rapidez em lotes de arquivos. Nos primeiros testes que efetivamos, a identificação de algumas centenas de arquivos demorou aproximadamente três minutos. Essa ferramenta foi batizada de DROID (**D**igital **R**ecord **O**bject **I**dentification) e está disponibilizada gratuitamente para a comunidade interessada em atividades de preservação digital, através de *download* em sítio específico[45] do Arquivo do Reino Unido. A *figura 9* mostra uma tela inicial do aplicativo DROID como exemplo.

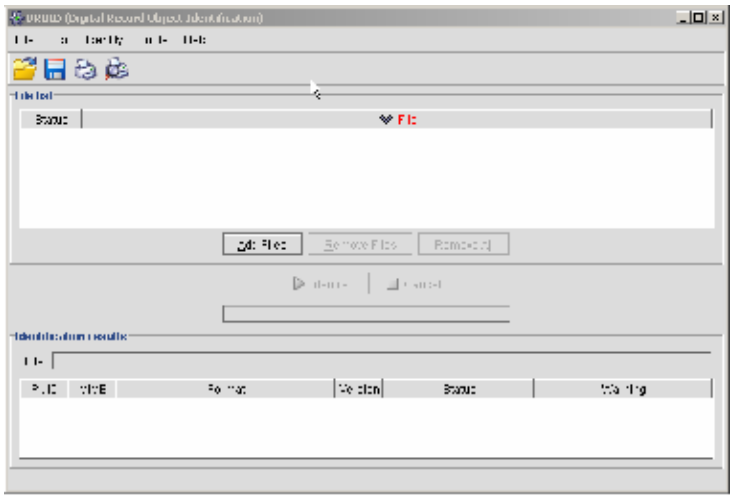


Figura 9 - Tela do aplicativo *DROID*

[43] Idem.

[44] Completada em outubro de 2005, conforme informações disponíveis na página do projeto.

[45] http://www.nationalarchives.gov.uk/aboutapps/PRONOM/tools.htm

A *figura 10* mostra os detalhes identificados (*Identification results*) para um arquivo específico que já passou pelo processo de identificação.

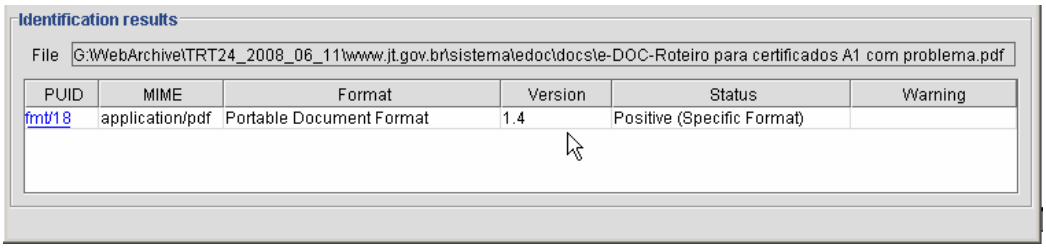


Figura 10 - Detalhe no aplicativo DROID com características identificadas

Na parte superior da *figura 10*, na linha **File**, o aplicativo mostra o local onde o arquivo está gravado, incluindo o caminho completo. No caso do exemplo acima, o arquivo chama-se 'e-DOC-Roteiro para certificados A1 com problema.pdf'. Abaixo dessa linha existem seis colunas com detalhes sobre esse arquivo específico.

A primeira coluna, ***PUID***, significa ***P**ronom **U**nique **Id**entifier* (identificador único no PRONOM). O projeto ***PRONOM*** tem coletado diversas informações sobre formatos de arquivo, como o *software* associado a determinado formato. Principalmente com o intuito de padronizar a identificação desses formatos foi criado o ***PUID***. No exemplo acima, o termo **fmt/18** se consultado nas bases do projeto ***PRONOM*** (há um *link* no próprio aplicativo ***DROID***), mostraria as seguintes informações (parte do total de informações relacionadas a este ***PUID***):

Name	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 text, images, video and audio.
Orientation	Binary
Byte order	Big-endian (Motorola)

Figura 11 – Parte das informações disponibilizadas sobre o formato fmt/18

Na *figura 11*, a descrição (*Description*) do formato fmt/18 contém informações bastantes relevantes sobre o formato.

A segunda coluna da *figura 10* contém informações ***MIME*** (***M****ultipurpose* ***Internet M****ail* ***E****xtensions*). As informações ***MIME*** são uma tentativa para padronizar os arquivos que trafegam na Internet, esses tipos são mantidos pela organização ***IANA*** (*Internet Assigned Numbers Authority*[46]).

Ao lado dessa coluna temos as colunas ***Format*** (nome do formato) e ***Version*** (versão desse formato). É importante aqui frisar que se deve dar atenção para a versão de cada formato. Dependendo dessa versão e formato haverá alterações significativas na estrutura desses arquivos.

A coluna ***Status*** indica se a operação de identificação do formato de arquivo foi bem sucedida ou não. De acordo com os manuais do aplicativo ***DROID***, as possibilidades da coluna Status são:

Positive (Specific): ocorre se o arquivo confere com uma assinatura binária que identifica um formato de arquivo único.

Positive (Generic): ocorre se o arquivo confere com uma assinatura binária que identifica vários formatos de arquivo.

Tentativ: ocorre se o arquivo tem uma extensão de arquivo usado pelo formato de arquivo e não há uma assinatura binária disponível para esse formato.

---

[46] A *Internet Assigned Numbers Authority* (IANA) é responsável pela coordenação global dos nomes DNS, endereços de IP e outros protocolos na Internet. Seu sítio é http://www.iana.org/.

Uma quarta opção ocorre quando o formato de arquivo simplesmente não pode ser identificado pelo aplicativo, nesse caso, ele reportará a mensagem: *"The format could not be identified"*

Finalmente, a última coluna ***Warning*** pode conter avisos relevantes, como por exemplo, o arquivo possuir uma determinada extensão, como *.jpg*, mas ser identificado como sendo do tipo *.gif*, nesse caso, a mensagem seria: *"Possible file extension mismatch"*.

7.1.6.1 IDENTIFICAÇÃO DOS FORMATOS DE ARQUIVO

Para cada sítio pesquisado nos órgãos do Universo de Pesquisa, utilizamos o aplicativo *DROID* para efetuar a identificação de cada um dos arquivos presentes. Em seguida, através do próprio aplicativo geramos uma lista num formato reconhecido pelo aplicativo ***Excel*** da *Microsoft*, gerando assim planilhas para cada órgão pesquisado. No **Anexo VI** encontramos uma parte de uma planilha para um órgão que já passou pelo processo de *archiving* e análise pelo aplicativo ***DROID***. A primeira coluna contém a identificação ***PUID*** (fmt/17 por exemplo) do arquivo, em seguida o nome ***MIME*** (*application/pdf* por exemplo) e na seqüência o nome e versão do formato, a última coluna refere-se ao ***Status*** da análise.

Do total de arquivos que passaram pelo processo de *archiving*, aplicamos uma filtragem inicial. Primeiro, foram excluídos todos aqueles que não tiveram um Status positivo de análise (*Positive Specific Format*), como no exemplo abaixo:

fmt/17	application/pdf	Portable Document Format	1.3	Positive (Specific Format)

Arquivos que não tiveram o mesmo *status* do arquivo acima, ou seja, tiveram apenas o *status Tentative*, foram excluídos pois não oferecem uma informação segura de seu tipo, a não se em casos muito especiais onde ficou claro se tratar de um formato **importante na amostra**. Em segundo lugar, excluímos também todos os arquivos que claramente compõem a própria página do sítio pesquisado, ou seja, trata-se de elementos do **sítio** como um

documento em si e não documentos que podem ser acessados através de *links* no sítio. Como já definimos antes, a análise de sítios como documentos está fora de nosso escopo. Assim, arquivos como esses dois especificados abaixo codificados em *html* e *xml* são considerados como fazendo parte do sítio em si.

text/html	Hypertext Markup Language	null	Positive (Specific Format)
txt/xml	Extensible Markup Language	1.0	Positive (Specific Format)

Tabela 14 - Arquivos excluídos da amostra de dados

Após essa filtragem inicial, efetuamos uma compilação dos formatos de arquivo presentes no sítio do órgão pesquisado, a *tabela 15* exemplifica um órgão pesquisado:

1	fmt/17	application/pdf	Portable Document Format	1.3
15	fmt/18	application/pdf	Portable Document Format	1.4
7	fmt/3	image/gif	Graphics Interchange Format	1987a
175	fmt/4	image/gif	Graphics Interchange Format	1989a
115	fmt/43	image/jpeg	JPEG File Interchange Format	1.01
32	fmt/44	image/jpeg	JPEG File Interchange Format	1.02
10	fmt/11	image/png	Portable Network Graphics	1.0

Tabela 15- Dados Compilados por Órgão

No exemplo do órgão utilizado nessa coleta, após as filtragens dos formatos positivamente identificados e a exclusão de arquivos claramente utilizados para a construção do sítio do órgão, o resultado foi 355 (trezentos e cinqüenta e cinco) arquivos (soma da primeira coluna), sendo que encontramos 7 (sete) diferentes formatos. Observe-se aqui que esse número de formatos refere-se ao tipo ***PUID*** que considera formatos de versões diferentes como sendo diferentes formatos de arquivo.

8 ANÁLISE DOS DADOS COLETADOS

Nesse capítulo, procederemos a uma análise detalhada dos dados coletados na amostra do universo. Os procedimentos de ***WebArchiving***, **Identificação dos Formatos de Arquivo**, **Filtragem** e **Compilação dos Dados** foram executados para cada um dos 89 (oitenta e nove) órgãos pesquisados. Na seqüência, iremos expor os dados coletados e analisados para então comparar os formatos identificados em relação ao **Modelo de Formato de Arquivo** que definimos anteriormente nesse trabalho.

8.1 DADOS COLETADOS NO PROCESSO DE *WEB ARCHIVING*

O anexo IV dessa dissertação contém a relação completa dos órgãos pesquisados e os respectivos endereços na *web*. O processo de *Web Archiving* baseou-se nessa relação, no entanto, não foi possível efetuar o *download* de arquivos em todas as URL´s (*Uniform Resource Locator*, endereços web do tipo http://.....); primeiro, em função da diretiva *Robot.txt* já exposta no item 8.1.4.1 anterior; segundo, em função de problemas técnicos não identificados nos órgão pesquisados. Em um caso específico, identificamos que o órgão havia passado por um incêndio que também envolveu o setor de informática (Tribunal Regional do Trabalho da 11ª, Amazonas). Na maioria dos casos com problemas técnicos, o processo de *Web Archiving* efetuou *download* de somente dois ou três arquivos e decidimos então excluir esses casos da amostra. Por coincidência ou não, a maioria dos problemas ocorreu no grupo dos Tribunais Eleitorais Regionais.[47], 8 (oito) casos do total de 14 (quatorze), sendo 1(um) na Justiça Estadual e 3 (três) na Justiça do Trabalho. Além do quase 8% (oito por cento) em função da diretiva *Robots.txt* mais 12 (doze) órgãos, 13,5% (treze e meio por cento), da amostra total foi comprometida.

[47] Na época desta pesquisa (segundo semestre de 2008), passamos por um período de eleições estaduais em quase todo o Brasil, com exceção do Distrito Federal.

A *tabela 16* resume os dados coletados no processo de *Web Archiving* nos 89 (oitenta e nove) órgãos pesquisados.

Nome do Grupo	Dados do Grupo			
		Tamanho	Arquivos	Pastas
Conselho da Justiça Federal	Totais	3227648	86	10
	Médias:	0	0	0
	Desvios Padrão:	0	0	0
	Número de órgãos pesquisados:	1	0	0
	Número efetivamente pesquisado:	1	0	0
Tribunais Superiores	Totais	1536167936	23559	2592
	Médias:	384041984	5889,75	648
	Desvios Padrão:	276394524	3005,468169	654,3388521
	Número de órgãos pesquisados:	5	0	0
	Número efetivamente pesquisado:	4	0	0
Justiça Federal	Totais	435843072	1831	620
	Médias:	145281024	610,3333333	206,6666667
	Desvios Padrão:	142233526,3	514,4952219	255,7974459
	Número de órgãos pesquisados:	5	0	0
	Número efetivamente pesquisado:	3	0	0
Justiça Estadual/Distrital	Totais	4013686784	39972	4277
	Médias:	160547471,4	1598,88	171,08
	Desvios Padrão:	169305118,1	1676,004756	184,4959168
	Número de órgãos pesquisados:	27	0	0
	Número efetivamente pesquisado:	25	0	0
Justiça do Trabalho	Totais:	3703074816	50094	2425
	Médias:	205726378,7	2783	134,7222222
	Desvios Padrão:	342394955,3	3685,245319	134,9757301
	Número de órgãos no grupo:	24	0	0
	Número efetivamente pesquisado:	18	0	0
Justiça Eleitoral	Totais	4566188032	71077	10599
	Médias:	240325685,9	3740,894737	557,8421053
	Desvios Padrão:	267237745,6	4226,517372	1047,878559
	Número de órgãos pesquisados:	27	0	0
	Número efetivamente pesquisado:	19	0	0
Número de Grupos:			**6**	
Número Órgãos na Pesquisa:			**89**	
Número Órgãos com DownLoad:			**70**	**78,65%**
Tamanho DownLoad (Mbytes)			**13.924.012**	
Quantidade Arquivos no DownLoad:			**186.619**	

Tabela 16 - Quadro geral *Web Archiving*

A *tabela 16* mostra que fizemos o *download* total de 186.619 (cento e oitenta e seis mil e seiscentos e dezenove) arquivos em aproximadamente 12 Gbytes (13.924.012 bytes). Esse ainda não é o número total de arquivos da amostra de formatos de arquivo identificados que compõe a Amostra Final pois inclui também arquivos que compõem a estrutura do sítio e serão eliminados no processo de filtragem manual. Esses números foram obtidos em 78,65% (setenta e oito por cento e sessenta e cinco centésimos) do total de 89 órgãos pesquisados.

8.2 FORMATOS DE ARQUIVOS IDENTIFICADOS NA AMOSTRA

A *tabela 17* resume os arquivos que foram analisados com relação à identificação do formato de arquivo específico utilizado. Essa análise foi feita após a operação de filtragem do total de arquivos baixados, como já explicamos anteriormente em 8.1.6.1.

Nome do Grupo	Dados do Grupo	
Conselho da Justiça Federal	Quant. Arquivos após Filtragem	218
	Média Formatos Únicos por Órgão	7,00
	Média Notas de Formatos no Grupo	67,35
Tribunais Superiores	Quant. Arquivos após Filtragem	9476
	Média Formatos Únicos por Órgão	12,75
	Média Notas de Formatos no Grupo	66,32
Justiça Federal	Quant. Arquivos após Filtragem	924
	Média Formatos Únicos por Órgão	8,00
	Média Notas de Formatos no Grupo	64,88
Justiça Estadual/Distrital	Quant. Arquivos após Filtragem	3430
	Média Formatos Únicos por Órgão	12,60
	Média Notas de Formatos no Grupo	66,76
Justiça do Trabalho	Quant. Arquivos após Filtragem	11683
	Média Formatos Únicos por Órgão	10,00
	Média Notas de Formatos no Grupo	66,37
Justiça Eleitoral	Quant. Arquivos após Filtragem	24095
	Média Formatos Únicos por Órgão	12,63
	Média Notas de Formatos no Grupo	63,28
Número de Grupos:		6
Total Geral Arquivos após Filtragem		49.826
Média Nota Formatos da Pesquisa		65,83

Tabela 17 - Resumo Identificação Formatos de Arquivo

Pela *tabela 17* verificamos que a análise de formatos de arquivo disponíveis nos sítios do órgãos efetivamente pesquisados ocorreu em 49.826 (quarenta e nove mil oitocentos e vinte e seis) arquivos. O restante dos arquivos que foram baixados dos sítios - 186.619 total de arquivos menos 49.826 analisados –, num total de 136.793 (cento e trinta e seis mil e setecentos e noventa e três) arquivos, foram considerados como não relevantes para a análise. Na verdade, esses arquivos foram efetivamente analisados com relação a seu formato de arquivo, porém o formato de arquivo identificado foi associado a **código de programação** ou **componentes** em geral do sítio (elementos de um sítio na Internet) e, por isso, foram excluídos da análise mais detalhada com relação às notas dadas aos arquivos, como veremos na seção seguinte.

8.3 AVALIAÇÃO DOS FORMATOS DE ARQUIVO DA AMOSTRA

Essa talvez seja a seção mais importante nesse capítulo dedicado à análise dos dados coletados em nossa pesquisa. Os procedimentos antes descritos, ou seja, **coleta de dados** através de uma ferramenta de ***web archiving***, **identificação** dos formatos de arquivo presentes em cada órgão que foi possível colher dados – 70 de um total de 89 no universo de pesquisa -, **filtragem** dos arquivos que realmente nos interessam para compor a análise, todos eles objetivaram criar a amostra final de **49.826 arquivos** (conforme *tabela 16*) para que fosse possível um **diagnóstico qualitativo** dos formatos de arquivo efetivamente em uso.

Para que tal diagnóstico fosse possível elaboramos uma planilha onde aplicamos uma nota para cada formato de arquivo disponível na Amostra Final. Essa nota foi obtida tomando como referência o **Modelo** de formato de arquivo que definimos na seção 6.5. Mais especificamente, o que fizemos foi responder, para cada uma das 7 (sete) características presentes no **Modelo,** se o formato de arquivo utilizado na **Amostra Final** atende ou não os requisitos ou até que ponto os atende. Caso um determinado formato de arquivo atendesse às sete características estaria então 100% (cem por cento) em conformidade com o modelo, da mesma forma, o não atendimento de qualquer característica implicaria em um nota 0% (zero por cento). Entre os limites de 0% e 100% surgiram as notas obtidas.

Para ilustrar o procedimento acima, vejamos um exemplo. O formato de arquivo PDF em sua versão 1.4 (um ponto quatro) foi um dos formatos de arquivo que encontramos em nossa amostra final: para esse formato de arquivo, então, aplicamos a planilha da *Tabela 18*, a nota que aparece no campo **Nota do Formato** é calculado automaticamente pela planilha.

PUID do formato:	fmt/18
Característica	**Resposta**
INDEPENDE DE DISPOSITIVOS	sim
METADADOS INCORPORADOS	sim
TRANSPARÊNCIA	sim
DESATIVAÇÃO DE PROTEÇÕES	sim
ESPECIFICAÇÃO NÃO PROPRIETÁRIA	não
ESPECIFICAÇÃO NORMATIZADA	não
AUTO SUFICIÊNCIA	sim
Nota do Formato:	**71,43**

Tabela 18 - Análise do formato de arquivo PDF versão 1.4

No caso do formato de arquivo PDF (versão 1.4) sabemos que ele atende a todas as características em relação ao **Modelo** que definimos, exceto ser uma **Especificação Proprietária** e não **Normatizada.** Portanto, das 7 (sete) características, ele atende 5 (cinco) delas o que incorre em 71.43% (setenta e um e quarenta e três centésimos por cento) do ideal que seria 100% de conformidade.

Vendo de outra maneira, um determinado formato de arquivo poderia receber até 8 (oito) notas diferentes: caso não atenda a qualquer requisito, 0 (**0%**); um requisito (**14,29%**), dois (**28,57%**); três (**42,86%**); quatro (**57,14%**); cinco (**71,42%**); seis (**85,71%**) ou sete (**100%**) equivalente a uma conformidade total.

Vejamos mais um exemplo da análise para ilustrar melhor. O formato de arquivo RTF em sua versão 1.2 foi analisado através da seguinte planilha:

PUID do formato:	fmt/47
Característica	**Resposta**
INDEPENDE DE DISPOSITIVOS	sim
METADADOS INCORPORADOS	não
TRANSPARÊNCIA	sim
DESATIVAÇÃO DE PROTEÇÕES	sim
ESPECIFICAÇÃO NÃO PROPRIETÁRIA	não
ESPECIFICAÇÃO NORMATIZADA	não
AUTO SUFICIÊNCIA	sim
Nota do Formato:	**57,14**

Tabela 19 - Análise do formato RTF versão 1.2

Notem que nesse formato de arquivo a característica **Metadados Incorporados** não é obedecida , além de também não obedecer às mesmas duas características do formato de arquivo PDF (1.4); dessa forma, o formato RTF atende a **4** (quatro) requisitos o que equivale à sua nota **57,14%**.

O processo de análise dos formatos de arquivo na **Amostra Final** encontrou 46 (quarenta e seis) diferentes tipos de formatos de arquivo (***PUIDs***). Na verdade, é importante lembrar que esse número de diferentes tipos considera que um mesmo formato de arquivo em versões diferentes deve ser tratado como um formato diferente. Assim, o formato de arquivo PDF foi encontrado em quase todas as suas versões até o momento o que implicou em seis diferentes tipos somente para esse formato.

8.3.1 FONTES PARA AVALIAR FORMATOS DE ARQUIVO

Com base em quais informações e fontes de consulta nós respondemos a cada uma das 7 (sete) indagações sobre características do **Modelo** confrontadas com cada **Formato de Arquivo** na **Amostra Final,** como nos exemplos da seção anterior para PDF e RTF?

Responder essa questão foi possível através da consulta a documentos disponíveis sobre cada um desses formatos de arquivo. Tais documentos são comentários de especialistas, reportagens publicadas em sítios especializados ou *blogs*, mas principalmente especificações

sobre formatos de arquivo. De grande importância no processo de análise foi o repositório de informações sobre formatos de arquivo do Reino Unido: o ***PRONOM***, já discutido anteriormente na seção 8.1.6. Naquela seção discutimos mais especificamente o processo de identificação de formatos de arquivo através do uso do aplicativo ***DROID***. Na verdade, o projeto do **Arquivo Nacional** do Reino Unido vai muito além da disponibilização e manutenção daquele aplicativo. Há um repositório de informações com mecanismos de busca e visualização de relatórios detalhados com informações técnicas e gerais sobre formatos de arquivo. A *figura 12* mostra a página inicial na Internet[48] do repositório *PRONOM*.

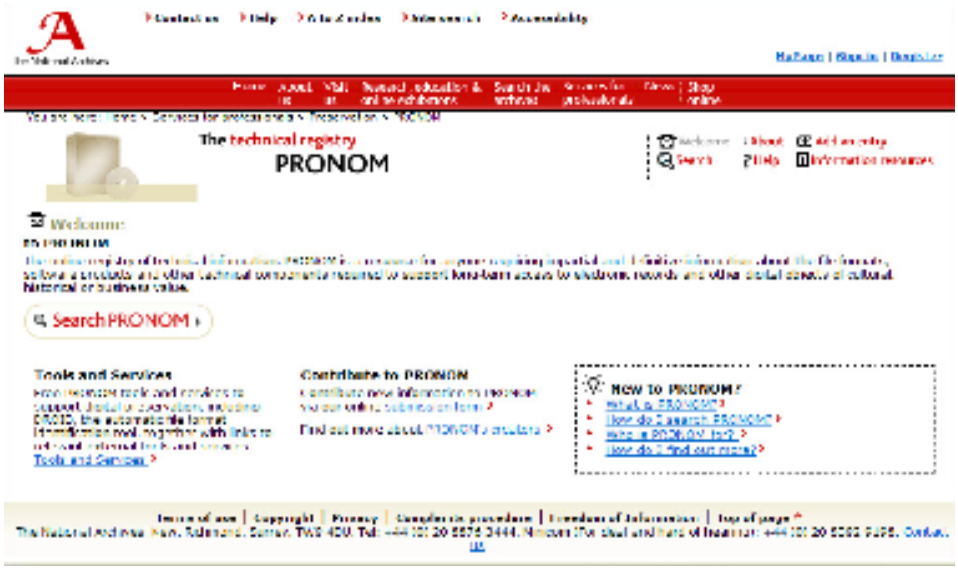


Figura 12 - Página inicial PRONOM

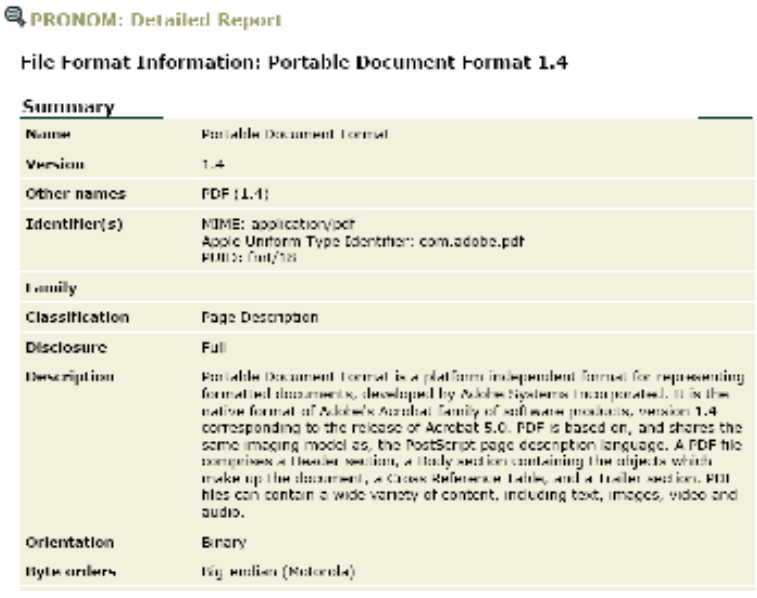
Aproveitando-se do fato de que o aplicativo ***DROID*** gera, no processo de identificação de formatos de arquivo, o código ***PUID*** (*Pronom Unic Identifier*). Utilizamos esse código para cada formato de arquivo na **Amostra Final,** utilizando-o nas buscas pelos

[48] (http://www.nationalarchives.gov.uk/PRONOM)

relatórios técnicos detalhados, como no exemplo da *figura 13*, onde efetuamos uma busca no **PUID** fmt/18 (PDF versão 1.4):

Figura 13 - Busca de relatório formato fmt/18

Isto resultou num relatório de 3 páginas. Uma parte do *Summary* desse relatório está na *figura 14*:



PRONOM: Detailed Report

File Format Information: Portable Document Format 1.4

Summary

Name	Portable Document Format
Version	1.4
Other names	PDF (1.4)
Identifier(s)	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 text, images, video and audio.
Orientation	Binary
Byte orders	Big endian (Motorola)

Figura 14 – Parte do relatório *PUID* fmt/18

Finalmente, uma outra vantagem no uso do repositório de informações **PRONOM** como fonte de análise dos formatos na **Amostra Final** é sua idoneidade. O Arquivo Nacional do Reino Unido é uma instituição governamental, em princípio neutra com relação a posições comerciais, além de ser uma instituição tradicional ligada ao tratamento de diferentes aspectos dos documentos.

9 CONCLUSÕES SOBRE DADOS COLETADOS

9.1 DADOS COLETADOS

As primeiras conclusões sobre os dados coletados devem ser com relação ao êxito ou não de nossos objetivos iniciais. Em outras palavras, conseguimos coletar os dados de que necessitávamos? Esses dados são adequados quantitativa e qualitativamente? É possível responder ao problema proposto inicialmente?

A resposta para a primeira pergunta é evidentemente positiva. Ao todo, coletamos mais de 12Gigabytes de dados, em órgãos espalhados por todo o território nacional. O método de coleta *web-archiving* se mostrou extremamente eficiente para essa coleta. Com relação à adequação dos dados coletados, consideramos que a resposta também é afirmativa. Não apenas colhemos uma quantidade imensa bruta de arquivos (**186.619)**, mas houve também uma dispersão equilibrada da coleta. Com exceção do **Grupo Justiça Federal** que teve 60% dos órgãos com dados coletados e analisados[49], todos os outros grupos de órgãos pesquisados tiveram pelo menos 70% das unidades coletadas e analisadas (vide *tabela 15*). O número de arquivos que compõem a **Amostra Final** (**49.826)**, nos parece expressivo, se considerarmos que nosso universo original era de 89 órgãos apenas. Assim, em média, colhemos quase 560 (49.826 / 89) amostras em cada órgão.

Um outro aspecto da coleta de dados que nos parece importante salientar diz respeito ao problema de *bias* em procedimentos de coleta. O termo refere-se a possíveis interferências tanto do **pesquisador** como dos **pesquisados** na formulação dos questionamentos e obtenção das respostas, interferências essas em função de possíveis fatores humanos como o receio de passar informações que podem ser comprometedoras e vários outros fatores emocionais. Como nossa coleta se deu através de **mecanismos automatizados**, sem que os pesquisados

[49] Porém é importante salientar que esse grupo possui apenas 5 (cinco) unidades e três foram coletadas com sucesso

nem mesmo tivessem consciência de que estavam sendo pesquisados, acreditamos que o fator *bias* foi bastante mitigado. Além disso, a principal fonte (**Arquivo Nacional do Reino Unido**) para **avaliação dos formatos de arquivo**, como já expusemos antes, é uma instituição neutra do ponto de vista dos interesses industriais e comerciais.

9.2 LIMITES DA COLETA DE DADOS

Apesar de estarmos plenamente satisfeitos com a coleta de dados que efetivamos no que diz respeito tanto aos aspectos quantitativos quanto aos qualitativos, acreditamos ser importante fazer notar algumas observações sobre os limites de nossa coleta e conseqüente análise dos dados.

Antes de tudo, como já foi exposto antes através das tabelas, resumo do processo de coleta, nem todos os 89 órgãos tiveram dados coletados: o número de 70 órgãos nos foi imposto em função de políticas expressas através do arquivo *Robots.txt* e problemas técnicos encontrados. Mesmo dentro do universo de 70 órgãos pesquisados, não coletamos todos os arquivos de cada órgão pois fomos obrigados a impor limites: como exposto antes, em relação ao **tamanho dos arquivos coletados** e a conseqüente **duração de coleta**, necessidade de (muito) **espaço** em disco[50], **tempo geral** para coleta e **velocidade** de acesso (banda) na Internet[51].

Por último, mas não menos importante, é preciso lembrar que a Internet possui várias camadas, sendo algumas às vezes referidas como ***web* profunda**, outra maneira de dizer que certos arquivos estão disponíveis somente após a execução de procedimentos especiais como processadores de bancos de dados ou uso de senhas para acesso. É forçoso lembrar que em

[50] Mesmo com as limitações impostas colhemos mais de 12Gb de informações.

[51] A coleta foi feita utilizando-se um acesso doméstico para a Internet.

função dessa estrutura, provavelmente deixamos de coletar um número indeterminado mas talvez relevante de arquivos para análise.

10 CONCLUSÕES GERAIS

Nesse último capítulo, procuramos expor os resultados que obtivemos nos diferentes procedimentos que adotamos para solucionar o problema dessa dissertação, conforme descritos nos objetivos do **Capítulo 1**, sendo que o ponto principal diz respeito a concluir sobre a qualidade dos **formatos de arquivos** utilizados nos documentos digitais da administração pública. Antes de abordar esse ponto, porém, teceremos outros comentários que consideramos igualmente relevantes.

10.1 SOBRE O MODELO DE FORMATOS DE ARQUIVO

O **Modelo** de formato de arquivo que utilizamos como referência para efetuar a análise qualitativa das especificações de formatos de arquivo efetivamente utilizadas é uma ferramenta criada no contexto dessa dissertação para utilização dentro de nossos objetivos de pesquisa. No entanto, após sua elaboração e aplicação real na amostra final que recolhemos nos órgãos do **Universo da Pesquisa**, acreditamos que pode se tratar de uma poderosa ferramenta de análise que extrapola em muito os limites dessa dissertação.

Programas de Gestão Documental ou administração de **Acervos**, como **bibliotecas**, **arquivos** e outras **unidades de informação**, de **Documentos Digitais** podem valer-se dessa ferramenta como um referencial para estabelecer quais formatos de arquivo devem ser recebidos no **Acervo** ou avaliar os riscos na utilização dos formatos de arquivo que já foram incorporados ao acervo e dessa forma planejar providências como, por exemplo, a migração dos formatos atuais para novos formatos mais adequados.

Normalmente, os responsáveis por **Acervos Digitais** ficam a reboque do desenvolvimento tecnológico no que cabe à adoção de formatos de arquivos para seus documentos. Mas, com o uso do **Modelo** proposto pode-se, pelo menos, estabelecer critérios para a escolha dos formatos mais adequados.

Por outro lado, da mesma forma que percebemos o potencial de aplicação do **Modelo** em outros contextos. Através da aplicação prática do mesmo, percebemos limites em sua aplicação. Ocorre que para determinadas especificações de formatos de arquivos certas características não puderam ser checadas com 100% de certeza em relação à especificação, em parte talvez pela falta de disponibilidade de informações técnicas, mas principalmente pela característica no modelo não se encaixar numa simples resposta **sim** ou **não**. Ao que parece, para certos formatos, a resposta está entre o sim e o não. Esse fato sugere a necessidade de futuras melhorias no **Modelo**, possivelmente exigindo sub-características para cada uma das sete características já definidas.

É também importante prever que novas características podem surgir, além das atuais sete e finalmente há a questão dos tipos de documentos. Nosso Modelo atualmente se aplica genericamente para qualquer tipo de documento, seja ele som, imagem, texto ou outros. Talvez fosse mais vantajosa a utilização de um **Modelo** para cada tipo de documento.

São questões que parecem exigir mais reflexão e pesquisa para possíveis aprimoramentos; de qualquer forma, dentro dos limites propostos, o **Modelo** foi adequado.

10.2 OS FORMATOS SÃO ADEQUADOS PARA A PRESERVAÇÃO?

Os formatos de arquivo atualmente um uso pela **Administração Pública Brasileira** são adequados para a **Preservação Digital** de maneira que as gerações futuras poderão ter acesso ao legado atual produzido nesse segmento de nossa sociedade brasileira?

Essa foi a grande questão que nos propusemos a responder através do presente trabalho e sua resposta não será um simples sim ou não. A complexidade do problema e o modo como deve ser analisado exige várias considerações.

Se considerarmos a **Nota Final** que obtivemos para os formatos analisados: **65,83%** em relação ao **Modelo** (ver *tabela 16*), trata-se de uma nota quase mediana, um pouco acima da média. Porém, é preciso lembrar que essa nota é uma média aritmética de todos os

formatos encontrados em todas as unidades pesquisadas e precisa ser encarada como o que é: uma **média**. Essa nota oscilou pouco entre as notas finais de cada grupo como se pode perceber na mesma *tabela 16*. A média aritmética dos **46** formatos de arquivo identificados na amostra final foi de **58,10%** (no Anexo V há uma tabela com todos os formatos analisados e suas respectivas notas). Mais uma vez um número não abaixo da média, mas ainda menos acima dessa que o anterior. Com base nessas notas, é forçoso admitir que o **cenário geral** em relação ao uso de formatos de arquivo e sua preservação é apenas **regular**, numa escala entre **péssimo** (0 a 20%), **ruim** (entre 20 e 40%), **regular** (entre 40 e 60%), **bom** (entre 60 e 80%) e **ótimo** (acima de 80%).

É preciso sempre lembrar que essa avaliação "**regular**" refere-se a um cenário geral, não é uma nota específica para os documentos digitais tomados individualmente. Nesse caso, se tomarmos um documento digital individualmente essa avaliação como regular torna-se muito pior. Podemos afirmar isso pois para quase todos os tipos básicos de documentos digitais - como o texto, a imagem fixa e o som – há, atualmente, no mercado tecnológico disponibilidade de especificações de formatos de arquivo com excelentes avaliações como a especificação de formato PDF/A (norma ISO 19005-1) para texto. Essa especificação atende todos os requisitos de nosso **Modelo** de referência. Então, porque, essas especificações especiais não são utilizadas?

Com base nas análises e conclusões dos parágrafos anteriores que sugerem fortemente a não adoção de especificações de formatos de arquivo adequados para a preservação podemos questionar o porquê dessa situação. Sabemos que existem especificações disponíveis e mais adequadas aos objetivos da preservação digital; por que, então, não são adotadas? Acreditamos que uma das respostas está na ausência de **políticas** (pelo menos que incluam preocupações com a preservação digital) nos órgãos que orientem a produção, recebimento e disponibilização de documentos digitais.

Antes de utilizar o processo de *web archiving* automatizado para coleta dos arquivos em nossa pesquisa, pretendíamos efetuar a análise sítio por sítio, acessando cada um dos endereços em nossa relação e fazendo uma análise individualizada pessoalmente. Abandonamos essa abordagem após a decisão de utilizar um processo automatizado com todas as suas vantagens correspondentes. Mas, na sondagem que fizemos, entre outras perguntas, procuramos verificar se havia, disponível no sítio do órgão, algum tipo de manual ou procedimentos documentais e, se sim, se havia alguma orientação para uso de formatos de arquivo e se esse uso visava à preservação digital. Aproveitamos o resultado desse material para todos os **Tribunais de Justiça** em cada um dos 26 estados e Distrito Federal e também para cada um dos 24 **Tribunais Regionais do Trabalho**. O resultado está disponível no Anexo VII. As respostas foram obtidas em todos os órgãos exceto um e é preocupante notar que **não** encontramos menção à preservação digital em qualquer órgão. Esse levantamento sugere a falta de políticas de preservação digital anteriormente citada. Apesar de ser importante frisar que se trata de um levantamento do que está disponível nos sítios e talvez exista algum tipo de política interna aos órgãos.

Concluímos, portanto, que os dados recolhidos e analisados sugerem fortemente a falta de uma política definida e implementada de política de preservação digital, pelo menos no que cabe aos formatos de arquivo. Essa política deveria, antes de mais nada, definir quais os formatos de arquivo que poderiam ser utilizados em acervos digitais e optar pelos formatos de arquivo mais adequados. Uma prospecção de formatos de arquivo utilizados num contexto onde essa política estivesse em funcionamento apontaria um mapa de formatos bem diferente do que encontramos de fato.

REFERÊNCIAS

ANDERSON, Cokie. Digital preservation: will your files stand the test of time ? **Library High Tech News**, v. 6, pp. 9-10. Emerald Publishing, 2005.

ARMS, Caroline; FLEISCHHAUER, Carl. **Digital Formats**: Factors for sustainability, functionality, and quality. Washington: Office of Strategic Initiatives. Library of Congress, 2005. Disponível em: <http://www.digitalpreservation.gov/formats>. Acesso em: 15 ago. 2008.

ASCHENBRENNER, Andreas. The bits and bites of data formats: stainless design for digital endurance. New York: **RLG Diginews**, v. 8, n. 1. Disponível em: <http://www.rlg.org/>. Acesso em: 20 fevereiro 2006.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS (ABNT). **NBR 15472**: Sistemas espaciais de dados e informações - Modelo de referência para um sistema aberto de arquivamento de informação (SAAI). 2007.

AXT, Gunter. Justiça e memória: a experiência do memorial do judiciário do estado do Rio Grande do Sul. **Justiça & Memória**, Porto Alegre, v. 2, n. 4, p. 215-238, 2002.

BO HOVGAARD, Thomasen. **Tests of software and strategies for micro-archiving websites**. Denmark: Centre for Internet Research, University of Aarhus, 2004. Disponível em: <http://www.cfi.au.dk/eng/pub/webarc>. Acesso em: 26 set. 2008.

BODÊ, Ernesto C. Assinaturas digitais e arquivologia. **Arquivística.net**. v. 2, n. 1, 2006. Disponível em: <http://www.arquivistica.net/ojs/viewarticle.php?id=51>. Acesso em: 10 jun. 2007.

_____. **Preservação de coleções de documentos digitais**. In: SEMINÁRIO INTERNACIONAL DE BIBLIOTECAS DIGITAIS, 2007, São Paulo. Disponível em < <http://www.cipedya.com/doc/175640>.

_____. **Formatos de arquivo e a preservação de documentos digitais**. Comunicação livre apresentada no XIV Congresso Brasileiro de Arquivologia, Rio de Janeiro, 2006.

BRASIL, **Resolução 45** do Conselho Nacional de Justiça (CNJ). Dispõe sobre a padronização dos endereços eletrônicos dos órgãos do Poder Judiciário.

_____ . Tribunal de Contas da União (TCU). **Relatório e pareceres sobre as contas do governo da República**: exercício de 2006. Brasília: TCU, 2007.

BROWN, Adrian. **Selecting file format media for Long Term Preservation**. UK: The National Archives, 2003. Disponível em: <http://www.nationalarchives.gov.uk/documents/selecting_file_formats.pdf>. Acesso em: 10 jun. 2007.

BRÜGGER, Niels. **Archiving websites**: general considerations and strategies. Denmark: Centre for Internet Research, University of Aarhus, 2005. Disponível em: <http://www.cfi.au.dk/eng/pub/webarc>. Acesso em: 26 set. 2008.

BYERS, Fred R. **Care and handling of CDs and DVDs**. Washington: Council on Library and Information Resources, 2003.

CAMPOS, Luiz F. de Barros. **Metadados Digitais**: revisão bibliográfica da evolução e tendências por meio de categorias funcionais. In: Revista Eletrônica de Biblioteconomia e

Ciência da Informação, v. 12, n. 23, 2007. Disponível em: <http://www.periodicos.ufsc.br/index.php/eb/article/viewfile/318/390>.Acesso em: 17 jul. 2008.

CHEN, Ching-Chih; KIERNA, Kevin (ed.s). **DELOS-NSF Working Group on Digital Imagery for Significant Cultural and Historical Materials**. . Dec. 2002. Disponível em: <http://dli2.nsf.gov/internationalprojects/working_group_reports/digital_imagery.html>. Acesso em: 15 abr. 2008.

CONNONLLY, David W. **Character set considered harmful**. Documento (minuta) publicado na Internet em 1995. Disponível em:

<http://www.w3.org/MarkUp/html-spec/charset-harmful.html>. Acesso em: 20 ago. 2008.

CONWAY, P. **Preservação no universo digital**. 2 ed. Rio de Janeiro: Projeto Conservação Preventiva em Bibliotecas e Arquivos: Arquivo Nacional, 2001.

_____. Overview: rationale for digitization and preservation. In: SITTS, Maxine K. **Handbook for digital projects**: a management tool for preservation and access. Massachuttes: Northeast Document Conservation Center, 2000. Disponível em: <http://www.nedcc.org/oldnedccsite/digital/ii.htm>. Acesso em: 15 abr. 2008.

DBTA (**Dicionário Brasileiro de Terminologia Arquivística**). Rio de Janeiro: Arquivo Nacional, 2005.

DELLAVALLE, Robert P. Et al.. Going, going, gone: lost internet references. **Science**, vol. 302, n. 31, oct./2003. Disponível em: <http://www.sciencemag.org>. Acesso em: 05 ago. 2008.

DELOS-NSF Working Group on Digital Imagery for Significant Cultural and Historical Materials. edited by Ching-chih Chen and Kevin Kiernan. December 2002. Disponível em: <http://dli2.nsf.gov/internationalprojects/working_group_reports/digital_imagery.html>. Acesso: 15 abr. 2008.

DPC - DIGITAL PRESERVATON COALITION. **The handbook**. Disponível em: <http://www.dpconline.org/graphics/handbook>. Acesso: 05 abr. 2006.

DOCTORS, Márcio. **A cultura do papel**. Rio de Janeiro: Casa da Palavra, 1999.

DURANTI, L., EASTWOOD, Terry, Macneil, Heather. **Preservation of the integrity of electronic records**. The Netherlands: Kluwer Academic Publishers, 2002.

FERREIRA, Aurélio B. de Holanda. **Novo dicionário da língua portuguesa**. 2ª ed. Rio de Janeiro: Nova Fronteira, 1986.

FERREIRA, Miguel. Introdução à preservação digital: conceitos, estratégias e actuais consensos. Guimarães: Escola de Engenharia da Universidade do Minho, 2006. Disponível em: <https://repositorium.sdum.uminho.pt/bitstream/1822/5820/1/livro.pdf>. Acesso em: 15 nov. 2008.

FISCHER, S. R. **A history of writing**. London, Reino Unido: Reaktion Books, 2003.

HEREDIA HERRERA, Antonia. **Archivística general:** teoria y práctica. Sevilla: Diputación Provincial, 1991.

HOFMAN, Hans. **Can bits and bytes be authentic?**: **preserving the authenticiy of digital objects**. IFLA conference in Glasgow (revised paper), 2002. Disponível em: <http://www.digicult.info>. Acesso em 30 de julho de 2008.

HUNTER, Dard. **Papermaking**: the history and technique of na ancient craft. New York: Dover, 1978.

IKEMATU, Ricardo Shoiti. Gestão de metadados: sua evolução na tecnologia da informação. In: **DataGramaZero** v. 2, n. 6 dez. 20011. Disponível em: <http://www.dgz.org.br/dez01/Art_02.htm>. Acesso em: 27 jul. 2008.

INNARELLI, Humberto C. **Preservação de documentos digitais :** confiabilidade de mídias de CD-ROM e CD-R . Campinas, 2006. Dissertação (Mestrado em Engenharia Mecânica) Universidade Estadual de Campinas.

IM (Information Management Journal). The digital explosion. Lenexa: ARMA, 2007.

KELLOG, David. **Evaluation of open source spidering technology**. Reports from Aquifer Meeting, 2004. Disponível em: <http://www.diglib.org/aquifer/oct2504/>. Acesso em: 26 set. 2008.

KIDDER, Tracy. **A alma da nova máquina**. São Paulo: Melhoramentos, 1981.

KIENTZLE, Tim. **Internet file formats**. Arizona: Coriolis Group, 1995.

LAURENT, Gilles. **Guarda e manuseio de materiais de registro sonoro**. Projeto Conservação Preventiva em Bibliotecas e Arquivos: Arquivo Nacional, 2001.

LAVOIE, B.; DEMPSEY, L. Thirteen ways of looking at... digital preservation. In: **D-Lib Magazine**, v. 10, n. 7/8. Disponível em: <http://www.dlib.org/dlib/july04/lavoie/07lavoie.html>. Acesso em: 30 jul. 2008.

LeFURGY, William G. PDF/A: **Developing a file format for long-term preservat**ion. RLG News, New York, v. 7, n. 6, 2003. Disponível em: <http://www.rlg.org>. Acesso em: 10 nov. 2005.

MacCARN, Dave. **Toward a universal data format for the preservation of media**. SMPTE Journal, 1997. Disponível na web em: <http://info.wgbh.org/upf/papers/smpte_upf_paper.html>. Acesso em: 15 ago. 2008.

MacNEIL, Heather. **Trusting records**: legal, historical, and diplomatic perspectives. The Netherlands: Kluwer Academic Publishers, 2000.

MANINI, Miriam Paula; MARQUES, Otacílio Guedes. Informação histórica: recuperação e divulgação da memória do poder judiciário brasileiro. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 8, 2007, Salvador. **GT2**. Disponível em: <http://www.enancib.ppgci.ufba.br/artigos/GT2--149.pdf>. Acesso em: 18 nov. 2008.

MONTE, A. C. LOPES, Luis F. D. **A qualidade dos suportes no armazenamento de informações**. Florianópolis: VisualBooks, 2004.

MUÑOZ VIÑAS, Salvador. **Contemporary theory of conservation**. Reino Unido: Elsevier, 2005.

MURRAY, James; VanRYPER, William. **Encyclopedia of graphics file form**ats. California: O'Reilly & Associates, 1994.

NAPOLITANO, Marcos. Fontes audiovisuais: a história depois do papel. In: PINSKY, C. B. (org). **Fontes históricas**. São Paulo: Contexto, 2006.

OCLC/RLG. **Preservation metadata for digital objects**: a review of the state of the art. OCLC/RLG: 2001.

RONDINELLI, Rosely Curi. **Gerenciamento arquivístico de documentos eletrônicos**: uma abordagem teórica da diplomática arquivística contemporânea. Rio de Janeiro: FGV, 2002.

ROTHENBERG, Jeff. **Avoiding technological quicksand**: finding a viable technical foundation for digital preservation. Washington: Council on Library and Information Resources, 1999. ISBN 1-887334-63-7.

SANTOS, Vanderlei B. dos. **Gestão de documentos eletrônicos: uma visão arquivística**. 2ª edição. Brasília: ABARQ, 2005.

SHEPARD, Thom; MacCARN, Dave. **The universal preservation format**: a recommended practice for archiving media and electronic records. Boston, 1998. Disponível em: <http://info.wgbh.org/upf/>. Acesso em: 22 mar. 2008.

SMIT, Johanna; GONÇALVES, Cássia Denise. **Como organizar arquivos fotográficos: projeto como fazer**. São Paulo: AASP, 2005. Apostila do curso

STANESCU, Andreas. Assessing the durability of formats in a digital preservation environment. **OCLC Systems & Services**, v. 21, n. 1, pp. 61-81. Emerald Publishing, 2005. Disponível em: <http://www.emeraldinsight.com/1065-075X.htm>. Acesso em: 20 ago.2008.

SULLIVAN, Susan J. An archival/records management perspective on PDF/A. In: **Records Management Journal**, v. 16, n. 1, pp. 51-56. Emerald Group Publishing Limited, 2006. Disponível em: <http://www.emeraldinsight.com/0956-5698.htm>. Acesso em: 15 ago. 2008.

THOMAZ, K. P. **A preservação de documentos eletrônicos de caráter arquivístico**: novos desafios, velhos problemas. Belo Horizonte, 2004. Tese (Doutorado em Ciência da Informação) - Programa de Pós-graduação da Escola de Ciência da Informação da UFMG.

_____. Gestão e preservação de documentos eletrônicos de arquivo: revisão de literatura – parte 2. In: **Arquivistica.net**. v.2, n.1, p.114-131, jan./.jun. 2006. Disponível em: <http:// www.arquivistica.net>. Acesso em: 09 jan. 2007.

_____; SOARES, A. José. A preservação digital e o modelo de referência open archival information system (OAIS). **Datagramazero**, Rio de Janeiro, v. 5, n. 1 fev. 2004.

_____; SANTOS, Vilma M. Metadados para o gerenciamento eletrônico de

arquivos – GED/A. **DatagramaZero**, Rio de Janeiro, v. 4, n. 4, ago. 2003

Understanding CD-R and CD-RW. California: Optical Storage Technology Association, 2003.

UNIVERSITY OF LEEDS. **Survey and assessment of sources of information on file formats and software documentation**. The representation and rendering project. Reino Unido, [s.d]. 48 p. Disponível em: <http://www.leeds.ac.uk/reprend>. Acesso em: 22 mar. 2008.

VAN BOGART, John W.C. **Armazenamento e manuseio de fitas magnéticas**: um guia para bibliotecas e arquivos. Rio de Janeiro: Projeto Conservação Preventiva em Bibliotecas e Arquivos: Arquivo Nacional, 2001.

WATERS, Donald. Transforming libraries through digital preservation. In: **Going Digital**: strategies for access, preservation, and conversion of collections to a digital format. New York: The Haworth Press, 1998.

WILLIAMS, K. et al.. **Professional XML databases**. Reino Unido: Wrox Press, 2000.

WILLIANSON, Andrew. Strategies for managing digital content formats. **Library Review**, v. 54, n. 9, pp. 508-513. Emerald Publishing, 2005. Disponível em: <http://www.emeraldinsight.com/0024-2535.htm>. Acesso em: 20 ago. 2008.

ANEXO I – EXEMPLO FORMATO DE ARQUIVO: WRI

.WRI Write File Format

This topic describes the binary file format used by Microsoft Write. A Write binary file contains information about file content, text and pictures (including object-linking-and-embedding, or OLE, objects), and formatting.

(Some stuff seems to be missing, so I've added it. Comments to sean@mess.org please.)

Write-File Header

The Write-file header describes the content of the file. It contains data, pointers to subdivisions of the formatting section, and information about the length of the file. The file header has the following form:

Word	Name	Description
0	wIdent	Must be 0137061 octal (or 0137062 octal if the file contains OLE objects)
1	dty	Must be zero
2	wTool	Must be 0125400 octal
3		Reserved; must be zero
4		Reserved; must be zero
5		Reserved; must be zero
6		Reserved; must be zero
7-8	fcMac	Number of bytes of actual text plus 128, the bytes in one sector (low-order word first)
9	pnPara	Page number for start of paragraph information
10	pnFntb	Page number of footnote table (FNTB) or pnSep, if none
11	pnSep	Page number of section property (SEP) or pnSetb, if none
12	pnSetb	Page number of section table (SETB) or pnPgtb, if none
13	pnPgtb	Page number of page table (PGTB) or pnFfntb, if none
14	pnFfntb	Page number of font face-name table (FFNTB) or pnMac, if none
15-47	szSsht	Reserved for Microsoft Word compatibility
48	pnMac	Count of pages in whole file (last page number plus 1)

In the preceding list, a "page number" means an offset in 128-byte blocks from the start of the file. For example, if pnPara equals 10, the paragraph information is at offset 10*128 = 1280 in the file.

The starting page number of character information (pnChar) is not stored but is computable, as follows:

pnChar = (fcMac + 127) / 128

Examining the value of word 48 of the header is a good way to distinguish Write files from Microsoft Word files. If pnMac equals zero, the file originated in Word. Any other value identifies a Write file.

Text and Pictures

After the header comes information about text and pictures. This information constitutes a separate section of the file.

Text

The text of the Write file starts at word 64 (page 1). Write uses the Windows character set (except for the pictures in the file) as well as the following special characters:

o ASCII character codes 13, 10 (carriage return, linefeed) for paragraph

ends. No other occurrences of these two characters are allowed.

o ASCII character code 12 for explicit page breaks.

o ASCII character code 9 (normal) for tab characters.

Other line-break or wordwrap information is not stored.

Pictures

Pictures (including OLE objects) are stored as a sequence of bytes in the text stream. These bytes can be identified as picture information by examining their paragraph formatting. One picture is exactly one paragraph.

Paragraphs that are pictures have a special bit set in their paragraph property (PAP) structure. For more information on the PAP structure, see Section 8.3, "Formatting."

(note: Write that comes with Windows 3.0 uses the picture stuff below, and does not support OLE; Write that comes with Windows 3.1 always uses OLE, but can read the picture stuff below.

Proof of this is that if you paste a picture into Write 3.1 (and thus it is OLE) you get an extra option in Save As; you get the possiblity to save it for Write 3.0. If you choose this it will say that all OLE objects will be removed in the file.

Also I have been unable to paste pictures with colour into Write 3.0, it always seems to convert it to monochrome; as a result of that, bmPlanes and bmBitsPixel are always 1.)

Each picture consists of a descriptive header followed by the data that makes up the picture. The header for OLE objects is different from the one used for pictures. The picture header has the following form:

Byte	Name	Description
0-7	mfp	Windows METAFILEPICT structure (hMF member undefined)
8-9	dxaOffset	Offset of picture from left margin, in twips (1/1440 inch)
10-11	dxaSize	Horizontal size, in twips
12-13	dyaSize	Vertical size, in twips
14-15	cbOldSize	Number of following bytes (actual metafile or bitmap bits); set to zero
16-29	bm	Additional information for bitmaps only
30-31	cbHeader	Number of bytes in this header
32-35	cbSize	Number of following bytes (actual metafile or bitmap bits), replacing cbOldSize for new files
36-37	mx	Scaling factor (x)
38-39	my	Scaling factor (y)
40-?	cbHeader	Picture contents, through cbHeader+cbSize-1

The mm member (bytes 0-1) of the METAFILEPICT structure specifies the mapping mode used to draw the picture. The last set of bytes will be bitmap bits if the value of the mm member is 0xE3. This is a special value used only in Write. Otherwise, the bytes will be metafile contents.

If the picture has never been rescaled with the Size Picture command in Write, the scaling factors in each direction will be 1000 (decimal). If the picture has been resized, the scaling factor will be the percentage of the original size that the picture is now, relative to 1000 (100 per cent).

For information about the METAFILEPICT structure and bitmaps, see the Microsoft Windows Guide to Programming and the Microsoft Windows Programmer's Reference, Volumes 1 and 3.

(added note:)

The METAFILEPICT structure looks like:

Word	Name	Description
0	mm	0xe3 for bitmap, metafile otherwise
1	xExt	Horizontal size, Word uses this in stead of dxaSize

2	yExt	Vertical size, Word uses this in stead of dyaSize
3	hMF	Handle to metafile, not used in Write.

If the contents is a bitmap, the bm member is a BITMAP structure, which looks like:

Byte	**Name**	**Description**
0-1	bmType	"BM" for bitmaps, not used in Write
2-3	bmWidth	Width in pixels
4-5	bmHeight	Height in pixels
6-7	bmWidthBytes	Width in bytes, rounded up on two-byte boundary
8	bmPlanes	Number of bit planes
9	bmBitsPixel	Number of bit per pixel
10-13	bmBits	A void FAR* pointer to the data, not used in Write

If the mm member has value 0x88, the file is a metafile (.wmf file). The bm member is empty, but the other members have values like normal. Colour wmf files exist.

(end of added note)

The descriptive header for OLE objects is similar to the one used for pictures. The OLE object header has the following form:

Byte	**Name**	**Description**
0-1	mm	Must be 0xE4
2-5		Not used
6-7	objectType	Type: 1=static, 2=embedded, 3=link
8-9	dxaOffset	Offset of picture from left margin, in twips (1/1440 inch)
10-11	dxaSize	Horizontal size, in twips
12-13	dyaSize	Vertical size, in twips
14-15		Not used
16-19	dwDataSize	Number of bytes in the object data that follows the header
20-23		Not used
24-27	dwObjNum	Hexadecimal number that, when converted to an 8-digit string, represents the object's unique name
28-29		Not used
30-31	cbHeader	Number of bytes in this header
32-35		Not used
36-37	mx	Scaling factor (x)
38-39	my	Scaling factor (y)
40-?	cbHeader	Object contents, through cbHeader+dwDataSize-1

The scaling factors for OLE objects work the same way as they do with pictures.

(added note:)

I couldn't find any information on the OLE objects. There is a libole2, which only works for OLE2 as far as I can see. OLE2 is an entire file-system, while OLE1 (as used here) is only one object.

The following is entirely reverse-enigineered, and therefore might not be correct.

The OLE object always starts with a DWORD with value 0x501, followed by another DWORD is the objectType as above, only with reverse values:

3 = static, 2 = embedded, 1 = link.

Next comes a DWORD which gives the length of the typename, which is immediately followed by that typename. It is a zero-terminated ascii string, and the length includes the 0 at the end.

Static OLE Object

Note that a static OLE object isn't really an OLE object; it is simply a picture which is rendered by Write itself. See:

http://support.microsoft.com/support/kb/articles/Q88/1/16.ASP

If the objectType is static, the typename has one of the following values:

DIB

METAFILEPICT

BITMAP

As usual, the data following that is not the stuff you would expect. The headers are garbled.

DIB

A dib (Device Independant Bitmap, a bmp file) usually has the following structure:

BITMAPFILEHEADER bmfh;

BITMAPINFOHEADER bmih;

RGBQUAD aColors[];

BYTE aBitmapBits[];

In the DIB which is stored in Write, the BITMAPFILEHEADER is missing.

After the string "DIB" (and the 0 terminator), comes the following bytes:

0xb2 0x18 0x00 0x00 0x29 0xec 0xff 0xff, followed by a DWORD which is the size of the dib _without_ the BITMAPFILEHEADER. After that the

BITMAPINFOHEADER follows. You must fill the members of the BITMAPFILEHEADER yourself; you can use the ColorsUsed to calculate the OffsetBits member.

(However, I have one instance of a Write file where this member is 0, although it is a 4 bit image. Maybe BitCount is a better member to use.)

BITMAP

This is the Device Dependant Bitmap (DDB), which is an insane format IMHO as the palette information is not stored. If the image is monochrome, he colours are of course black and white; if it is 4-bits, use the indows colours; if it is 8-bit, the first 8 and last 8 colours in the alette are Windows colours, but the other colours depend on what colour he palette has at that moment.

The data is stored in the BITMAP structure just as above (for Write 3.0 mages). After the "BITMAP" string (with the 0 terminator) comes the ollowing bytes:

0xb4 0x18 0x00 0x00 0x28 0xec 0xff 0xff

Followed by the size in in DWORD; next comes with BITMAP structure with he bmType and bmBits members undefined, followed by the uncompressed its.

METAFILEPICT

This is a Windows metafile (wmf). For reasons unknown Write (or Windows?) onverts some images to metafiles. I have no idea how this is stored.

It seems to be followed by these bytes:

0x4f 0x03 0x00 0x00 0xb1 0xfc 0xff 0xff

Then the size of the metafile in a DWORD; next comes the METAFILEPICT tructure (defined above) again with hMF and mm members undefined. After hat the metafile bits follow, but without a header.

Embedded OLE Object

The typename is the name of the executable, with the exe extension. For Paintbrush it is "Pbrush" for example.

The typename is followed by the filename. First there is a DWORD with he length (including the 0 at the end of the string), and the string tself. If the length is 0, there is no string (so not even a 0 for an empty string).

After that comes a parameter, for example the size of a picture in a string: "0 0 320 240". I don't know what use this has but it's there.

Just like with the filename, first there is a DWORD with the length of the string, and then the string itself (if the length is non-zero).

Last comes a DWORD with the offset to the next part of the OLE Object, followed by the data of the file itelf. That length is enough information on the length of the file, but it seems to be padded with crap; I have no idea how to acquire the length of the file without looking at the file itself (note that this depends on the type of file).

The data itself is really the file. For example for Paintbrush this would simply be a .bmp file, so it would start with "BM". Also note that some files cannot be read; if you use Paint Shop Pro for embedded objects, the file cannot be read into Paint Shop Pro when you extract it manually (so all of this is application specific).

After the file (add the offset to the byte after the DWORD where the offset is stored) comes the next part. Again this works like the whole OLE stream all over again, but with a difference: if the objectType is 0, there is nothing any more. If it is 5, it probably means "alternative display," like the Sound Recorder icon if the file was a .wav file.

Link OLE Object

This type is supposed to the type where the actual data is somewhere else; the filename points to the data of the file. It works very much like the embedded OLE Object type.

Suppose you have a Paintbrush OLE Object, type link. The filename is

"C:\WINDOWS\WINLOGO.BMP". The first part is stored as with embedded stuff, but after the parameter (which would be "0 0 320 240" in this case), there are 12 bytes padding and then the next OLE object. This could very well be the actual picture again as a embedded OLE object. However if a link is stored as a link OLE Object, the next OLE object will be the Sound Recorder icon.

Formatting

Write files contain both character and paragraph formatting information. There can be no gaps in either; each must begin with the first text character (byte 128) and continue through the last. The format descriptors (FODs) for the first and last paragraph must, therefore, have the value of fcLim equal to the value of fcMac, as defined in the header section.

(note: Write 3.0 sometimes saves a fcLim > fcMac, you have to check for this!)

There is a difference between paragraph and character FODs. A character FOD may describe any number of consecutive characters with the same formatting. However, there must be exactly one paragraph FOD for each text paragraph. In either case, it is advisable to have multiple FODs point to the same formatting properties (FPROPs) on a given page because it saves space in the file. No FOD may point off its page.

Characters and Paragraphs

Both the character and paragraph sections are structured as a set of pages. Each page contains an array of FODs and a group of FPROPs, both of which are described later in this section. Following is the format of a page:

Byte	Name	Description
0-3	fcFirst	Byte number of first character covered by this page of formatting information; equals 128 for first character in the text (low-order byte first)
4-n	rgfod	Array of FODs
n+1-126	grpfprop	Group of FPROPs
127	cfod	Number of FODs on this page

An FOD is fixed in size. It contains the byte offset to the corresponding

FPROP. Following is the structure of an FOD:

Word	Name	Description

0-1	fcLim	Byte number after last character covered by this FOD
2	bfprop	Byte offset from beginning of FOD array to corresponding FPROP for these characters or this paragraph

(note: sometimes bfprop is 0xffff; it seems that that means that the CHP or PAP has the default values.)

An FPROP is variable in size. It contains the prefix for a character property (CHP) or paragraph property (PAP), both of which are described later in this section. Following is the structure of an FPROP:

Byte	Name	Description
0	cch	Number of bytes in this FPROP
1-n	rgchProp	Prefix for a CHP (for characters) or a PAP (for paragraphs) sufficient to include all bits that differ from the default CHP or PAP

Following is the format of a CHP:

Byte	Bit	Name	Description
0			Reserved; ignored by Write
1	0	fBold	Bold characters
	1	fItalic	Italic characters
	2-7	ftc	Font code (low bits); index into the FFNTB
2		hps	Size of font, in half points (standard is 24)
3	0	fUline	Underlined characters
	1	fStrike	Reserved; ignored by Write
	2	fDline	Reserved; ignored by Write
	3	fOverset	Reserved; ignored by Write
	4-5	csm	Reserved; ignored by Write
	6	fSpecial	Set for "(page)" only
	7		Reserved; ignored by Write
4	0-2	ftcXtra	Font code (high-order bits, concatenated with ftc)
	3	fOutline	Reserved; ignored by Write
	4	fShadow	Reserved; ignored by Write
	5-7		Reserved; ignored by Write
5		hpsPos	Position: 0=normal, 1-127=superscript, 128-255=subscript

If the user doesn't select any special character properties, the CHP is filled with the following default values:

Byte	Value
0	1
2	24
3-5	0

Each character FPROP must, therefore, have a count of characters (cch) greater than or equal to 1.

Each PAP can contain up to 14 tab descriptors (TBDs), which are described later in this section. Following is the structure of a PAP:

Byte	Bit	Name	Description

0			Reserved; must be zero
1	0-1	jc	Justification: 0=left, 1=center, 2=right, 3=both
	2-7		Reserved; must be zero
2			Reserved; must be zero
3			Reserved; must be zero
4-5		dxaRight	Right indent, in 20ths of a point
6-7		dxaLeft	Left indent, in 20ths of a point
8-9		dxaLeft1	First-line left indent (relative to dxaLeft)
10-11		dyaLine	Interline spacing (standard is 240)
12-13		dyaBefore	Reserved; ignored by Write (standard is zero)
14-15		dyaAfter	Reserved; ignored by Write (standard is zero)
16	0	rhcPage	0=header, 1=footer
	1-2		Reserved; 0=normal paragraph, nonzero=header or footer paragraph
	3	rhcFirst	Start of printing: 1=print on first page, 0=do not print on first page
	4	fGraphics	Paragraph type: 1=picture, 0=text
	5-7		Reserved; must be zero
17-21			Reserved; must be zero
22-78			Tab descriptors (up to 14)

Following is the format of a TBD:

Byte	Bit	Name	Description
0-1		dxa	Indent from left margin of tab stop, in 20ths of a point
2	0-2	jcTab	Tab type: 0=normal tabs, 3=decimal tabs
	3-5	tlc	Reserved; ignored by Write
	6-7		Reserved; must be zero
3		chAlign	Reserved; ignored by Write

If the user doesn't select any special paragraph properties, the PAP is filled with the following default values:

Byte	Value
0	61
2	30
10-11	240 (word)
12-78	0

Each paragraph FPROP must have a count of characters (cch) greater than or equal to 1.

Footnotes

Write documents do not have footnote tables (FNTBs), so pnFntb is always equal to pnSep. In fact, all their header and footer paragraphs appear at the beginning of the document before any normal paragraphs. When reading files created by Word, Write recognizes only those headers and footers that appear at the beginning of the document; it treats all others as normal text.

Sections

A Write document has only one section. If the section properties of a Write document differ from the defaults, the document contains a section property (SEP) section and a section table (SETB) section. If not, then neither section is present and pnSep and pnSetb are both equal to pnPgtb.

Following is the format of an SEP:

Byte	Name	Description
0	cch	Count of bytes used, excluding this byte (all properties at byte positions greater than cch are set to their default values)
1-2		Reserved; must be zero
3-4	yaMac	Page length, in 20ths of a point (default is 11*1440=15840)
5-6	xaMac	Page width, in 20ths of a point (default is 8.5*1440=12240)
7-8		Reserved; must be 0xFFFF
9-10	yaTop	Top margin, in 20ths of a point (default is 1440)
11-12	dyaText	Height of text, in 20ths of a point (default is 9*1440=12960)
13-14	xaLeft	Left margin, in 20ths of a point (default is 1.25*1440=1800)
15-16	dxaText	Width of text area, in 20ths of a point (default is 6*1440=8640)

(add note: this table is incomplete)

Byte	Name	Description
1-2		Start page numbers at # if not 0xFFFF
19-20	yaHeader	Distance from top to header (default is 0.75*1440=1080)
21-22	yaFooter	Distance from top to footer (default is yaMac-0.75*1440=15760)

(end of added note)

The page length (yaMac) is equal to yaTop+dyaText. The page width (xaMac) is equal to xaLeft+dxaText+(right margin, not stored).

If all the above properties are set to their defaults, no SEP or SETB is needed. Otherwise, the count of characters (cch) is greater than or equal to 1 and less than or equal to 16.

The SETB section contains an array of section descriptors (SEDs), described later in this section. Following is the structure of an SETB:

Word	Name	Description
0	csed	Number of sections (always 2 for Write documents)
1	csedMax	Undefined
2-n	rgsed	Array of SEDs plus zero-padding to fill the sector

Following is the structure of an SED:

Word	Name	Description
0-1	cp	Byte address of first character following section

2	fn	Undefined
3-4	fcSep	Byte address of associated SEP

A Write document always has exactly two SED entries. The cp value of the first entry indicates that it affects all the characters in the document. The fcSep value of the first entry points to the one SEP in the file. The second SED entry is a dummy with fcSep set to 0xFFFFFFFF.

The PGTB section (optional) is on the page immediately after the SEP section.

(added note: AFAICS these are not used in Write.)

Note: The term "page" used in the rest of this section refers to printed pages of a Write document, not 128-byte "pages" of a disk file.

The page table (PGTB) contains an array of page descriptors (PGDs), which are described later in this section. Following is the structure of a PGTB:

Word	**Name**	**Description**
0	cpgd	Number of PGDs (1 or more)
1	cpgdMac	Undefined
2-n	rgpgd	Array of PGDs plus zero padding to fill the sector

Following is the structure of a PGD:

Word	**Name**	**Description**
0	pgn	Page number in printed Word documents
1-2	cpMin	Byte address of first character on printed page Font Table

The font face-name table (FFNTB) contains the number of font face names (FFNs) and a list of FFNs. Following is the structure of an FFNTB:

Byte	**Name**	**Description**
0-1	cffn	Number of FFNs
2-n	grpffn	List of FFNs

Following is the structure of an FFN:

Byte	**Name**	**Description**
0-1	cbFfn	Number of bytes following in this FFN (not including these 2 bytes)
2	ffid	Font family identifier (see below)
3-(cbffn+2)	szFfn	Font name (variable length; null-terminated)

A cbFfn value of 0xFFFF means that the next FFN entry will be found at the start of the next 128-byte page. A cbFfn value of zero means that there are no more FFN entries in the table.

Possible values for ffid are FF_DONTCARE, FF_ROMAN, FF_SWISS, FF_MODERN, FF_SCRIPT, and FF_DECORATIVE. These constants are defined in WINDOWS.H.Additional values may be added to the list in future versions of Windows.

(added note) These are the definitions taken from WINDOWS.H:

```
#define FF_DONTCARE 0x00  /* Don't care or don't know. */
#define FF_ROMAN    0x10  /* Variable stroke width, serifed. */
#define FF_SWISS    0x20  /* Variable stroke width, sans-serifed. */
#define FF_MODERN   0x30  /* Constant stroke width, serifed or sans-serifed. */
#define FF_SCRIPT   0x40  /* Cursive, etc. */
#define FF_DECORATIVE       0x50  /* Old English, etc. */
```

ANEXO II – ÓRGÃOS PESQUISADOS NO UNIVERSO

Órgão a ser pesquisado	Cidade/UF	Outros dados	Fone
Grupo CJF			
Conselho da Justiça Federal (CJF)	Brasília/DF	SAFS, Quadra 6, Lote 1, Trecho III CEP 70095-900	(61) 3319-8000
Tribunais Superiores			
Supremo Tribunal Federal (STF)	Brasília/DF	Praça dos Três Poderes, CEP 70175-900	(61) 3217-3000
Tribunal Superior do Trabalho (TST)	Brasília/DF	SAFS - Qd 8 Lote 1 CEP 70070-600	(61) 3314-4808
Tribunal Superior Eleitoral (TSE)	Brasília/DF	Praça dos Tribunais Superiores Bloco C CEP 70096-900	(61) 3316-3000
Superior Tribunal Militar (STM)	Brasília/DF	Praça dos Tribunais Superiores SAS CEP 70098-900	(61) 3313-9292
Superior Tribunal de Justiça (STJ)	Brasília/DF	SAFS, Quadra 6, Lote 1, Trecho III CEP 70095-900	
Justiça Federal de 1ª e 2ª Instâncias (TRFs)			
Tribunal Regional Federal da 1ª Região	Brasília/DF	SAU/SUL - Quadra 2 – Blocos A (Sede I) e K (Sede II) Praça dos Tribunais Superiores CEP 70070-900	(61) 3314-5225
Tribunal Regional Federal da 2ª Região	Rio de Janeiro/RJ	Rua Acre, 80 - Centro - 20.081-000	(21) 2276-8000
Tribunal Regional Federal da 3ª Região	São Paulo/SP	Av. Paulista, 1842 - Torre Sul - Cep:01310-936	
Tribunal Regional Federal da 4ª Região	Porto Alegre/RS	Rua Otávio Francisco Caruso da Rocha, 300 - Bairro Praia de Belas - CEP 90010-395	(51) 3213 3000
Tribunal Regional Federal da 5ª Região	Recife/PE	Av. Martin Luther King, S/N - Edifício Ministro Djaci Falcão - Cais do Apolo -	(81) 3425.9000

		CEP: 50030-908	
Justiça Estadual/Distrital (TJs)			
Tribunal de Justiça do Estado do Acre	Rio Branco/AC	Rua Floriano Peixoto, 460 – Centro	(68) 3211-5300
Tribunal de Justiça do Estado de Alagoas	Maceió/AL	Praça Marechal Deodoro, 319, Centro	(82) 3216-0100
Tribunal de Justiça do Estado do Amapá	Macapá/AP	Av. General Rondon, 1295 Centro 68906-390	(06) 3312-3301
Tribunal de Justiça do Estado do Amazonas	Manaus/AM	Av. André Aráujo s/n - CEP:69097-788	(92) 2129-6666
Tribunal de Justiça do Estado da Bahia	Salvador/BA	5ª Av. do CAB, nº 560, CEP 41746-900	(71) 3372-5686
Tribunal de Justiça do Estado do Ceará	Fortaleza/CE	Av. Gal. Afonso A. Lima, s/n Cambeba CEP 60.830-120	(85) 3216-2500
Tribunal de Justiça do Distrito Federal e Territórios	Brasília/DF	Palácio da Justiça Praça Municipal, lote 01 CEP 70094-900	(61) 3343-7000
Tribunal de Justiça do Estado do Espírito Santo	Vitória/ES	Rua Desembargador Homero Mafra, 60 Enseada do Suá - CEP 29050-275	(27) 3334-2000
Tribunal de Justiça do Estado do Mato Grosso	Cuiabá/MT	Centro Político Administrativo - CEP 78050-970 Caixa Postal - 1071	(65) 3617-3000
Tribunal de Justiça do Estado do Mato Grosso do Sul	Campo Grande/MS	Av. Mato Grosso - Bloco 13 - Parque dos Poderes - CEP 79031-902	(67) 3314-1300
Tribunal de Justiça do Estado de Minas Gerais	Belo Horizonte/MG	Rua Goiás, 229 - Centro - 30190-030	(31) 3237-6100
Tribunal de Justiça do Estado do Maranhão	São Luís/MA	Praça D. Pedro II s/n - Centro - Cep: 65.010-905	0800-707-1581

Tribunal de Justiça do Estado de Goiás	Goiânia/GO	Av. Assis Chateaubriand nº 195 St. Oeste CEP:74130-012	(62) 3216-2000
Tribunal de Justiça do Estado da Paraíba	João Pessoa/PB	Praça João Pessoa, s/n - CEP 58013-902	(83) 3216-1400
Tribunal de Justiça do Estado do Paraná	Curitiba/PR	Pç. Nossa Senhora da Salete - Centro Cívico - 80.530-912	(41) 3200-2000
Tribunal de Justiça do Estado da Pará	Belém/PA	Av. Almirante Barroso nº 3089 - Bairro: Souza - CEP:66613-710	(91) 3205-3000
Tribunal de Justiça do Estado de Pernambuco	Recife/PE	PRAÇA DA REPÚBLICA S/N - SANTO ANTÔNIO CEP: 50010-040	(81) 3419-3311
Tribunal de Justiça do Estado do Piauí	Teresina/PI	Pça. Des. Edgard Nogueira s/n, Centro Cívico	(86) 3216-7400
Tribunal de Justiça do Estado do Rio Grande do Sul	Porto Alegre/RS	Praça Marechal Deodoro, 55 – Centro	(51) 3210-7000
Tribunal de Justiça do Estado do Rio Grande do Norte	Natal/RN	Praça 7 de Setembro, S/N, Natal/RN, 59025-000	(84) 3216-6800
Tribunal de Justiça do Estado do Rio de Janeiro	Rio de Janeiro/RJ	Av. Erasmo Braga, 115 - Centro / CEP: 20020-903 - Rua Dom Manuel, 29, Centro / CEP: 20010-090	(21) 3133-2000
Tribunal de Justiça do Estado de Rondônia	Porto Velho/RO	Rua Rogério Weber, 1872 - Centro CEP 78916-050	(69) 3217-1152
Tribunal de Justiça do Estado de Roraima	Boa Vista/RR	Praça do Centro Cívico, s/n – Centro. CEP: 69.301-380	
Tribunal de Justiça do Estado de Santa Catarina	Florianópolis/SC	Rua Álvaro Millen da Silveira, n. 208	(48) 3221-1000
Tribunal de Justiça do Estado de São Paulo	São Paulo/SP	Praça da Sé, s/nº CEP 01018-001	(11) 3242-9366

Tribunal de Justiça do Estado de Sergipe	Aracajú/SE	Praça Fausto Cardoso, 112 - Centro. CEP:49010-080	(79) 3226-3100
Tribunal de Justiça do Estado do Tocantins	Palmas/TO	Praça do Girassóis, s/n CEP 77015-007	(63) 3218-4300
Justiça do Trabalho de 1ª e 2 ª Instâncias			
Tribunal Regional do Trabalho da 1ª Região (Rio de Janeiro)	Rio de Janeiro/RJ	Av. Presidente Antônio Carlos, 251- Castelo - CEP: 20.020-010	(21)3907-6150
Tribunal Regional do Trabalho da 2ª Região (São Paulo)	São Paulo/SP	Rua da Consolação, 1272 Consolação CEP 01302-906	(11) 3150-2000
Tribunal Regional do Trabalho da 3ª Região (Minas Gerais)	Belo Horizonte/MG	Av. Getúlio Vargas, 225 Bairro Funcionários CEP 30112-900	(31) 3228-7000
Tribunal Regional do Trabalho da 4ª Região (Rio Grande do Sul)	Porto Alegre/RS	Av. Praia de Belas, 1100 CEP 90110-903	(51) 3255-2000
Tribunal Regional do Trabalho da 5ª Região (Bahia)	Salvador/BA		
Tribunal Regional do Trabalho da 6ª Região (Pernambuco)	Recife/PE	Cais do Apolo, 739 Bairro do Recife CEP 50030-902	(81) 2129-2000
Tribunal Regional do Trabalho da 7ª Região (Ceará)	Fortaleza/CE	Av. Santos Dumont, 3384 Aldeota CEP 60150-162	(85) 3388-9400
Tribunal Regional do Trabalho da 8ª Região (Pará)	Belém/PA	Tv. D. Pedro I, 746 Umarizal CEP 66050-100	(91) 4008-7000
Tribunal Regional do Trabalho da 9ª Região (Paraná)	Curitiba/PR	Rua Vicente Machado, 147 Centro CEP 80420-905	(41) 3310-7000
Tribunal Regional do Trabalho da 10ª Região - Distrito Federal	Brasília/DF	SAS Quadra 01 Bloco D Praça dos Tribunais Superiores CEP 70097-900	(61) 3348-1100
Tribunal Regional do Trabalho da 11ª Região (Amazonas)	Manaus/AM	Rua Visconde de Porto Alegre, 1265 Praça 14 de Janeiro	(92) 3621-7200

		CEP 69.020-130	
Tribunal Regional do Trabalho da 12ª Região (Santa Catarina)	Florianópolis/SC	Rua Esteves Júnior, 395 Centro CEP 88015-905	(48) 3216-4000
Tribunal Regional do Trabalho da 13ª Região (Paraíba)	João Pessoa/PB	Av. Coralio Soares de Oliveira, s/n Centro CEP 58013-260	(83) 3533-6533
Tribunal Regional do Trabalho da 14ª Região (Rondônia)	Porto Velho/RO	Rua Almirante Barroso, 600 Centro CEP 78916-020	(68) 3211-6300
Tribunal Regional do Trabalho da 15ª Região (Campinas)	Campinas/SP	Rua Barão de Jaguara, 901 Centro CEP 13015-927	(19) 3236-2100
Tribunal Regional do Trabalho da 16ª Região (Maranhão)	São Luís/MA	Av. Senador Vitorino Freire, 2001 Areinha CEP 65030-015	(98) 3218-9300
Tribunal Regional do Trabalho da 17ª Região (Espírito Santo)	Vitória/ES	Rua Pietrangelo de Biase, 33 Centro CEP 29010-190	(27) 3321-2400
Tribunal Regional do Trabalho da 18ª Região (Goiás)	Goiânia/GO	Rua T-2 nº 1403 S. Bueno CEP 74215-901	(62) 3901-3300
Tribunal Regional do Trabalho da 19ª Região (Alagoas)	Maceió/AL	Avenida da Paz, 2076 Centro CEP 57020-440	(82) 2121-8299
Tribunal Regional do Trabalho da 20ª Região (Sergipe)	Aracajú/SE	Av. Dr. Carlos Rodrigues da Cruz, s/n Centro Adm. Gov. Augusto Franco - Bairro Capucho - CEP 49080-190	(79) 2105-8888
Tribunal Regional do Trabalho da 21ª Região (Rio Grande do Norte)	Natal/RN	Av. Capitão Mor-Gouveia, 1738 Lagoa Nova CEP 59063-400	(84) 4006-3000
Tribunal Regional do Trabalho da 22ª Região (Piauí)	Teresina/PI	Rua 24 de Janeiro, 181 / Norte CEP 64000-921	(86) 2106-9500
Tribunal Regional do Trabalho da 23ª Região (Mato Grosso)	Cuiabá/MT	Av. Historiador Rubens de Mendonça, 3355 Centro Político e	(65) 3648-4100

		Administrativo CEP 78050-955	
Tribunal Regional do Trabalho da 24ª Região (Mato Grosso do Sul)	Campo Grande/MS	Rua Jornalista Belizário Lima, 418 CEP 79004-912	(67) 3316-1771
Justiça Eleitoral (TREs)			
Tribunal Regional Eleitoral do Acre	Rio Branco/AC	Centro Administrativo do Governo Estadual, BR-364 Distrito Industrial CEP 69914-220	(68) 3212-4400
Tribunal Regional Eleitoral de Alagoas	Maceió/AL	Praça Visconde de Sinimbu s/n Centro CEP 57020-720	(82) 2122-7700
Tribunal Regional Eleitoral do Amapá	Macapá/AP		
Tribunal Regional Eleitoral do Amazonas	Manaus/AM	Av. André Araújo s/nº Aleixo	(092) 611-3638
Tribunal Regional Eleitoral da Bahia	Salvador/BA	1ª Avenida do CAB, 150 CEP 41745-901	(71) 3373-7220
Tribunal Regional Eleitoral do Ceará	Fortaleza/CE	Rua Jaime Benévolo, 21 Centro CEP 60050-080	(85) 3388-3500
Tribunal Regional Eleitoral do Distrito Federal	Brasília/DF	Praça Municipal Qd. 02 Lote 06 CEP 70094-901	(61) 3441-1027
Tribunal Regional Eleitoral do Espírito Santo	Vitória/ES	Av. João Batista Parra, 575 Praia do Suá CEP 29052-120	(27) 2121-8500
Tribunal Regional Eleitoral de Goiás	Goiânia/GO	Praça Cívica, 300 CEP 74003-010	(62) 3521-2114
Tribunal Regional Eleitoral do Maranhão	São Luís/MA	Av. Sem. Vitorino Freire, Areinha CEP 65010-917	0800-98-5000
Tribunal Regional Eleitoral do Mato Grosso	Cuiabá/MT		3648-8018
Tribunal Regional Eleitoral do Mato Grosso do Sul	Campo Grande/MS	Rua Desembargados Leão Neto do Carmo, 23 Parque dos Poderes CEP 79037-100	(67) 3326-4002

Tribunal Regional Eleitoral do Minas Gerais	Belo Horizonte/MG	Av. Prudente de Morais, 100 Cidade Jardim CEP 30380-000	(31) 3298-1100
Tribunal Regional Eleitoral do Pará	Belém/PA	Rua João Diogo, 288 Campina CEP 66015-902	
Tribunal Regional Eleitoral da Paraíba	João Pessoa/PB	Av. Princesa Isabel, 201 Centro CEP 58013-250	(83) 3214-1200
Tribunal Regional Eleitoral do Paraná	Curitiba/PR	Rua João Parolin, 224 Prado Velho CEP 80220-902	(41) 3330-8500
Tribunal Regional Eleitoral de Pernambuco	Recife/PE	Av. Agamenon Magalhães, 1160 Graças CEP 52010-904	(81) 4009-9200
Tribunal Regional Eleitoral do Piauí	Teresina/PI	Praça Desembargador Edgar Nogueira, s/n Centro Cívico CEP 64000-830	(86) 2107-9700
Tribunal Regional Eleitoral do Rio de Janeiro	Rio de Janeiro/RJ		
Tribunal Regional Eleitoral do Rio Grande do Norte	Natal/RN	Pça. André Albuquerque, 534 Centro CEP 59025-580	(84) 4006-5600
Tribunal Regional Eleitoral do Rio Grande do Sul	Porto Alegre/RS	Rua Duque de Caxias, 350 Centro CEP 90010-280	(51) 3216-9444
Tribunal Regional Eleitoral de Rondônia	Porto Velho/RO	Av. Presidente Dutra, 1889 Areal CEP 78916-100	(69) 3211-2000
Tribunal Regional Eleitoral de Roraima	Boa Vista/RR	Av. Juscelino Kubitschek, 589 São Pedro CEP 69306-685	(95) 2121-7000
Tribunal Regional Eleitoral de Santa Catarina	Florianópolis/SC	Rua Esteves Júnior, 68 Centro CEP 88015-130	(48) 3251-3700
Tribunal Regional Eleitoral de São Paulo	São Paulo/SP	Rua Francisca Miquelina, 123 Bela Vista CEP 01316-900	(11) 6858-2000

Tribunal Regional Eleitoral de Sergipe	Aracajú/SE	Lote 7, Variante 2 CENAF CEP 49081-000	(79) 2106-8600
Tribunal Regional Eleitoral de Tocantins	Palmas/TO	Av. Teotônio Segurado, Conjunto 01 Lotes 1 e 2 Plano Diretor Norte	(63) 218-6401

ANEXO III – ÓRGÃOS POR UNIDADE FEDERATIVA (UF)

Capital/UF	Cidade	Região	Unidades	%
Acre/AC	Rio Branco	Norte	2	2%
Alagoas/AL	Maceió	Nordeste	3	3%
Amapá/AP	Macapá	Norte	2	2%
Amazonas/AM	Manaus	Norte	3	3%
Bahia/BA	Salvador	Nordeste	3	3%
Brasília/DF	Brasília	CO	10	11%
Ceará/CE	Fortaleza	Nordeste	3	3%
Espírito Santo/ES	Vitória	Sudeste	3	3%
Goiás/GO	Goiânia	CO	3	3%
Maranhão/MA	São Luís	Nordeste	3	3%
Mato Grosso do Sul/MS	Campo Grande	CO	3	3%
Mato Grosso/MT	Cuiabá	CO	3	3%
Minas Gerais/MG	Belo Horizonte	Sudeste	3	3%
Pará/PA	Belém	Norte	3	3%
Paraíba/PR	João Pessoa	Nordeste	3	3%
Paraná/PR	Curitiba	Sul	3	3%
Pernambuco/PE	Recife	Nordeste	4	5%
Piauí/PI	Teresina	Nordeste	3	3%
Rio de Janeiro/RJ	Rio de Janeiro	Sudeste	4	5%
Rio Grande do Norte/RN	Natal	Nordeste	3	3%
Rio Grande do Sul/RS	Porto Alegre	Sul	4	5%
Rondônia/RO	Porto Velho	Norte	3	3%
Roraima/RR	Boa Vista	Norte	2	2%
Santa Catarina/SC	Florianópolis	Sul	3	3%
São Paulo/SP	São Paulo e Campinas	Sudeste	5	6%
Sergipe/SE	Aracajú	Nordeste	3	3%

Tocantins/TO	Palmas	Norte	2	2%
Total			89	100%

Região	Unidades	Percentual
Norte	17	19,10
Nordeste	28	31,46
CO	19	21,35
Sul	10	11,24
Sudeste	15	16,85
Total	89	100,00

ANEXO IV – RELAÇÃO ÓRGÃOS PESQUISADOS E ENDEREÇOS WEB

Conselho da Justiça Federal (CJF)	http://w w w .jf.gov.br
Tribunais Superiores	
Supremo Tribunal Federal (STF)	http://w w w .stf.gov.br
Tribunal Superior do Trabalho (TST)	http://w w w .tst.gov.br
Tribunal Superior Eleitoral (TSE)	http://w w w .tse.gov.br
Superior Tribunal Militar (STM)	http://w w w .stm.gov.br
Superior Tribunal de Justiça (STJ)	http://w w w .stj.gov.br
Justiça Federal de 1ª e 2ª Instâncias (TRFs)	
Tribunal Regional Federal da 1ª Região	http://w w w .trf1.gov.br/
Tribunal Regional Federal da 2ª Região	http://w w w .trf2.gov.br
Tribunal Regional Federal da 3ª Região	http://w w w .trf3.gov.br/
Tribunal Regional Federal da 4ª Região	http://w w w .trf4.gov.br/
Tribunal Regional Federal da 5ª Região	http://w w w .trf5.gov.br/

Justiça Estadual/Distrital (TJs)	
Tribunal de Justiça do Estado do Acre	http://www.tj.ac.gov.br
Tribunal de Justiça do Estado de Alagoas	http://www.tj.al.gov.br
Tribunal de Justiça do Estado do Amapá	http://www.tjap.gov.br
Tribunal de Justiça do Estado do Amazonas	http://www.tj.am.gov.br
Tribunal de Justiça do Estado da Bahia	http://www.tj.ba.gov.br
Tribunal de Justiça do Estado do Ceará	http://www.tj.ce.gov.br
Tribunal de Justiça do Distrito Federal e Territórios	http://www.tjdft.gov.br
Tribunal de Justiça do Estado do Espírito Santo	http://www.tj.es.gov.br
Tribunal de Justiça do Estado do Mato Grosso	http://www.tj.mt.gov.br
Tribunal de Justiça do Estado do Mato Grosso do Sul	http://www.tj.ms.gov.br
Tribunal de Justiça do Estado de Minas Gerais	http://www.tjmg.gov.br
Tribunal de Justiça do Estado do Maranhão	http://www.tj.ma.gov.br
Tribunal de Justiça do Estado de Goiás	http://www.tj.go.gov.br
Tribunal de Justiça do Estado da Paraíba	http://www.tj.pb.gov.br
Tribunal de Justiça do Estado do Paraná	http://www.tj.pr.gov.br
Tribunal de Justiça do Estado da Pará	http://www.tj.pa.gov.br
Tribunal de Justiça do Estado de Pernambuco	http://www.tjpe.gov.br
Tribunal de Justiça do Estado do Piauí	http://www.tj.pi.gov.br
Tribunal de Justiça do Estado do Rio Grande do Sul	http://www.tj.rs.gov.br
Tribunal de Justiça do Estado do Rio Grande do Norte	http://www.tjrn.gov.br
Tribunal de Justiça do Estado do Rio de Janeiro	http://www.tj.rj.gov.br
Tribunal de Justiça do Estado de Rondônia	http://www.tj.ro.gov.br
Tribunal de Justiça do Estado de Roraima	http://www.tj.rr.gov.br
Tribunal de Justiça do Estado de Santa Catarina	http://www.tj.sc.gov.br
Tribunal de Justiça do Estado de São Paulo	http://www.tj.sp.gov.br
Tribunal de Justiça do Estado de Sergipe	http://www.tj.se.gov.br
Tribunal de Justiça do Estado do Tocantins	http://www.tj.to.gov.br

Justiça do Trabalho de 1ª e 2 ª Instâncias	
Tribunal Regional do Trabalho da 1ª Região (Rio de Janeiro)	http://www.trt1.gov.br/
Tribunal Regional do Trabalho da 2ª Região (São Paulo)	http://www.trt2.gov.br/
Tribunal Regional do Trabalho da 3ª Região (Minas Gerais)	http://www.trt3.gov.br/
Tribunal Regional do Trabalho da 4ª Região (Rio Grande do Sul)	http://www.trt4.gov.br/
Tribunal Regional do Trabalho da 5ª Região (Bahia)	http://www.trt5.gov.br/
Tribunal Regional do Trabalho da 6ª Região (Pernambuco)	http://www.trt6.gov.br/
Tribunal Regional do Trabalho da 7ª Região (Ceará)	http://www.trt7.gov.br/
Tribunal Regional do Trabalho da 8ª Região (Pará)	http://www.trt8.gov.br/
Tribunal Regional do Trabalho da 9ª Região (Paraná)	http://www.trt9.gov.br/
Tribunal Regional do Trabalho da 10ª Região - Distrito Federal	http://www.trt10.gov.br/
Tribunal Regional do Trabalho da 11ª Região (Amazonas)	http://www.tr11.gov.br/
Tribunal Regional do Trabalho da 12ª Região (Santa Catarina)	http://www.trt12.gov.br/
Tribunal Regional do Trabalho da 13ª Região (Paraíba)	http://www.trt13.gov.br/
Tribunal Regional do Trabalho da 14ª Região (Rondônia)	http://www.trt14.gov.br/
Tribunal Regional do Trabalho da 15ª Região (Campinas)	http://www.trt15.gov.br/
Tribunal Regional do Trabalho da 16ª Região (Maranhão)	http://www.trt16.gov.br/
Tribunal Regional do Trabalho da 17ª Região (Espírito Santo)	http://www.trt17.gov.br/
Tribunal Regional do Trabalho da 18ª Região (Goiás)	http://www.trt18.gov.br/
Tribunal Regional do Trabalho da 19ª Região (Alagoas)	http://www.trt19.gov.br/
Tribunal Regional do Trabalho da 20ª Região (Sergipe)	http://www.trt20.gov.br/
Tribunal Regional do Trabalho da 21ª Região (Rio Grande do Norte)	http://www.trt21.gov.br/
Tribunal Regional do Trabalho da 22ª Região (Piauí)	http://www.trt22.gov.br/
Tribunal Regional do Trabalho da 23ª Região (Mato Grosso)	http://www.trt23.gov.br/
Tribunal Regional do Trabalho da 24ª Região (Mato Grosso do Sul)	http://www.trt24.gov.br/

Justiça Eleitoral (TREs)	
Tribunal Regional Eleitoral do Acre	http://www.tre-ac.gov.br/
Tribunal Regional Eleitoral de Alagoas	http://www.tre-al.gov.br/
Tribunal Regional Eleitoral do Amapá	http://www.tre-ap.gov.br/
Tribunal Regional Eleitoral do Amazonas	http://www.tre-am.gov.br/
Tribunal Regional Eleitoral da Bahia	http://www.tre-ba.gov.br/
Tribunal Regional Eleitoral do Ceará	http://www.tre-ce.gov.br/
Tribunal Regional Eleitoral do Distrito Federal	http://www.tre-df.gov.br/
Tribunal Regional Eleitoral do Espírito Santo	http://www.tre-es.gov.br/
Tribunal Regional Eleitoral de Goiás	http://www.tre-go.gov.br/
Tribunal Regional Eleitoral do Maranhão	http://www.tre-ma.gov.br/
Tribunal Regional Eleitoral do Mato Grosso	http://www.tre-mt.gov.br/
Tribunal Regional Eleitoral do Mato Grosso do Sul	http://www.tre-ms.gov.br/
Tribunal Regional Eleitoral do Minas Gerais	http://www.tre-mg.gov.br/
Tribunal Regional Eleitoral do Pará	http://www.tre-pa.gov.br/
Tribunal Regional Eleitoral da Paraíba	http://www.tre-pb.gov.br/
Tribunal Regional Eleitoral do Paraná	http://www.tre-pr.gov.br/
Tribunal Regional Eleitoral de Pernambuco	http://www.tre-pe.gov.br/
Tribunal Regional Eleitoral do Piauí	http://www.tre-pi.gov.br/
Tribunal Regional Eleitoral do Rio de Janeiro	http://www.tre-rj.gov.br/
Tribunal Regional Eleitoral do Rio Grande do Norte	http://www.tre-rn.gov.br/
Tribunal Regional Eleitoral do Rio Grande do Sul	http://www.tre-rs.gov.br/
Tribunal Regional Eleitoral de Rondônia	http://www.tre-ro.gov.br/
Tribunal Regional Eleitoral de Roraima	http://www.tre-rr.gov.br/
Tribunal Regional Eleitoral de Santa Catarina	http://www.tre-sc.gov.br/
Tribunal Regional Eleitoral de São Paulo	http://www.tre-sp.gov.br/
Tribunal Regional Eleitoral de Sergipe	http://www.tre-se.gov.br/
Tribunal Regional Eleitoral de Tocantins	http://www.tre-to.gov.br/

ANEXO V – RESUMO FORMATOS ANALISADOS

PUIDs	Notas
fmt/3	57,14
fmt/4	57,14
fmt/7	42,86
fmt/11	57,14
fmt/12	57,14
fmt/13	57,14
fmt/14	71,43
fmt/15	71,43
fmt/16	71,43
fmt/17	71,43
fmt/18	71,43
fmt/19	71,43
fmt/20	71,43
fmt/34	57,14
fmt/36	57,14
fmt/38	57,14
fmt/39	57,14
fmt/40	57,14
fmt/41	85,71
fmt/42	85,71
fmt/43	85,71
fmt/44	85,71
fmt/45	57,14
fmt/46	57,14
fmt/47	57,14
fmt/48	57,14
fmt/49	57,14
fmt/50	57,14
fmt/51	57,14
fmt/52	57,14
fmt/53	57,14
fmt/57	57,14
fmt/59	57,14
fmt/61	57,14
fmt/62	57,14
fmt/116	42,86
fmt/125	57,14
fmt/126	57,14
fmt/132	42,86
fmt/133	42,86
fmt/134	28,57
x-fmt/391	57,14
x-fmt/230	28,57
x-fmt/263	28,57
x-fmt/264	28,57

Quant. Formatos Analisados	**46**
Média dos Formatos	**58,10**
Maior Nota	85,71
Menor Nota	28,57

ANEXO VI – PLANILHA IDENTIFICAÇÃO DE FORMATOS

fmt/17	application/pdf	Portable Document Format	1.3
fmt/18	application/pdf	Portable Document Format	1.4
fmt/18	application/pdf	Portable Document Format	1.4
fmt/18	application/pdf	Portable Document Format	1.4
fmt/18	application/pdf	Portable Document Format	1.4
fmt/18	application/pdf	Portable Document Format	1.4
fmt/18	application/pdf	Portable Document Format	1.4
fmt/18	application/pdf	Portable Document Format	1.4
fmt/18	application/pdf	Portable Document Format	1.4
fmt/18	application/pdf	Portable Document Format	1.4
fmt/18	application/pdf	Portable Document Format	1.4
fmt/18	application/pdf	Portable Document Format	1.4
fmt/18	application/pdf	Portable Document Format	1.4
fmt/18	application/pdf	Portable Document Format	1.4
fmt/18	application/pdf	Portable Document Format	1.4
fmt/18	application/pdf	Portable Document Format	1.4
fmt/3	image/gif	Graphics Interchange Format	1987a
fmt/3	image/gif	Graphics Interchange Format	1987a
fmt/3	image/gif	Graphics Interchange Format	1987a
fmt/3	image/gif	Graphics Interchange Format	1987a
fmt/3	image/gif	Graphics Interchange Format	1987a
fmt/3	image/gif	Graphics Interchange Format	1987a

fmt/3	image/gif	Graphics Interchange Format	1987a
fmt/4	image/gif	Graphics Interchange Format	1989a
fmt/4	image/gif	Graphics Interchange Format	1989a
fmt/4	image/gif	Graphics Interchange Format	1989a
fmt/4	image/gif	Graphics Interchange Format	1989a
fmt/4	image/gif	Graphics Interchange Format	1989a
fmt/4	image/gif	Graphics Interchange Format	1989a
fmt/4	image/gif	Graphics Interchange Format	1989a
fmt/4	image/gif	Graphics Interchange Format	1989a
fmt/4	image/gif	Graphics Interchange Format	1989a
fmt/4	image/gif	Graphics Interchange Format	1989a
fmt/4	image/gif	Graphics Interchange Format	1989a
fmt/4	image/gif	Graphics Interchange Format	1989a
fmt/4	image/gif	Graphics Interchange Format	1989a

Tabela 20 - Planilha Coleta em Órgão após filtragem dos formatos de arquivo

ANEXO VII – LEVANTENTO ÓRGÃOS COM POLÍTICA FORMATOS

Justiça Estadual/Distrital (TJs)				
Tribunal de Justiça do Estado do Acre	Não	NA	Não	
Tribunal de Justiça do Estado de Alagoas	Não	NA	Não	
Tribunal de Justiça do Estado do Amapá	Não	NA	Não	
Tribunal de Justiça do Estado do Amazonas	**Sim**	Não	Não	Manual biblioteca 24hs
Tribunal de Justiça do Estado da Bahia	Não	NA	Não	
Tribunal de Justiça do Estado do Ceará	Não	NA	Não	
Tribunal de Justiça do Distrito Federal e Territórios	**Sim**	Não	Não	Política de Gestão Documental do Órgão
Tribunal de Justiça do Estado do Espírito Santo	Não	NA	Não	
Tribunal de Justiça do Estado do Mato Grosso	Não	NA	Não	
Tribunal de Justiça do Estado do Mato Grosso do Sul	Não	NA	Não	
Tribunal de Justiça do Estado de Minas Gerais	Não	NA	Não	
Tribunal de Justiça do Estado do Maranhão	Não	NA	Não	
Tribunal de Justiça do Estado de Goiás	Não	NA	Não	
Tribunal de Justiça do Estado da Paraíba	Não	NA	Não	
Tribunal de Justiça do Estado do Paraná	Não	NA	Não	
Tribunal de Justiça do Estado da Pará	Não	NA	Não	
Tribunal de Justiça do Estado de Pernambuco	Não	NA	Não	
Tribunal de Justiça do Estado do Piauí	Não	NA	Não	
Tribunal de Justiça do Estado do Rio Grande do Sul	Não	NA	Não	
Tribunal de Justiça do Estado do Rio Grande do Norte	Não	NA	Não	
Tribunal de Justiça do Estado do Rio de Janeiro	**Sim**	Não	Não	Tabela de temporalidade com orientação para digitalização
Tribunal de Justiça do Estado de Rondônia	Não	NA	Não	
Tribunal de Justiça do Estado de Roraima	Não	NA	Não	
Tribunal de Justiça do Estado de Santa Catarina	Não	NA	Não	
Tribunal de Justiça do Estado de São Paulo	Não	NA	Não	
Tribunal de Justiça do Estado de Sergipe	Não	NA	Não	
Tribunal de Justiça do Estado do Tocantins	Não	NA	Não	
Justiça do Trabalho de 1ª e 2 ª Instâncias				
Tribunal Regional do Trabalho da 1ª Região (Rio de Janeiro)	Não	NA	Não	
Tribunal Regional do Trabalho da 2ª Região (São Paulo)	Não	NA	Não	
Tribunal Regional do Trabalho da 3ª Região (Minas Gerais)	Não	NA	Não	
Tribunal Regional do Trabalho da 4ª Região (Rio Grande do Sul)	Não	NA	Não	
Tribunal Regional do Trabalho da 5ª Região (Bahia)	Não	NA	Não	
Tribunal Regional do Trabalho da 6ª Região (Pernambuco)	Não	NA	Não	
Tribunal Regional do Trabalho da 7ª Região (Ceará)	Não	NA	Não	
Tribunal Regional do Trabalho da 8ª Região (Pará)	Não	NA	Não	
Tribunal Regional do Trabalho da 9ª Região (Paraná)	Não	NA	Não	
Tribunal Regional do Trabalho da 10ª Região - Distrito Federal	Não	NA	Não	
Tribunal Regional do Trabalho da 11ª Região (Amazonas)	Não	NA	Não	
Tribunal Regional do Trabalho da 12ª Região (Santa Catarina)	Sim	Sim	Não	Envio de petições pede o formato pdf
Tribunal Regional do Trabalho da 13ª Região (Paraíba)	Não	NA	Não	
Tribunal Regional do Trabalho da 14ª Região (Rondônia)	Não	NA	Não	
Tribunal Regional do Trabalho da 15ª Região (Campinas)	Não	NA	Não	
Tribunal Regional do Trabalho da 16ª Região (Maranhão)	Não	NA	Não	
Tribunal Regional do Trabalho da 17ª Região (Espírito Santo)	Não	NA	Não	
Tribunal Regional do Trabalho da 18ª Região (Goiás)	Não	NA	Não	
Tribunal Regional do Trabalho da 19ª Região (Alagoas)	Não	NA	Não	
Tribunal Regional do Trabalho da 20ª Região (Sergipe)	Não	NA	Não	
Tribunal Regional do Trabalho da 21ª Região (Rio Grande do Norte)	Não	NA	Não	
Tribunal Regional do Trabalho da 22ª Região (Piauí)				Não foi possível verificar !
Tribunal Regional do Trabalho da 23ª Região (Mato Grosso)	Não	NA	Não	
Tribunal Regional do Trabalho da 24ª Região (Mato Grosso do Sul)	Não	NA	Não	

ANEXO VIII – TABELA COMPARATIVA METADADOS

	OAIS	CEDARS	NLA	NEDLIB
PRESERVATION DESCRIPTION INFORMATION	Reference Information	Resource Description	Persistent identifier	Creator
		Existing Metadata	Date of Creation	Title
		- existing records		Date of Creation
				Publisher
				Assigned Identifier
				- Value
				- Construction method
				- Responsible agency
				URL
				- Value
				- Date of validation
	Context Information	Related Information Objects	Relationships	<NA>
	Provenance Information	History of Origin	Preservation Action Permission	Change History
		- Reason for Creation	Quirks	- Main metadata concerned
		- Custody History	Archiving Decision (work)	- date
		- Change history before archiving	Decision Reason (work)	- old value
		- Original technical environment	Institution Responsible for Archiving Decision (work)	- new value
		- prerequisites	Archiving Decision (manifestation)	- tool
		- procedures	Decision Reason (manifestation)	- name
		- documentation	Institution Responsible for Archiving Decision (manifestation)	- version
		- Reason for preservation	Intention Type	- reverse
		Management history	Institution with Preservation Responsibility	- Other metadata concerned
		- Ingest process history	Process	- old value
		- Administration history	- Description of Process	- new value
		- action history	- Name of agency responsible for process	
		- policy history	- Critical hardware used	
		Rights Management	- Critical software used	
		- Negotiation history	- How process was carried out	
		- Rights information	- Guidelines used to implement process	
		- copyright statement	- Date and time	
		- name of publisher	- Result	
		- date of publication	- Process rationale	
		- place of publication	- Changes	
		- rights warning	- Other	
		- contracts or rights holder	Record Creator	
		- actors	Other	
		- actions		
		- permitted by statute		
		- legislation pointers		
		- permitted by license		
		- license text pointer		

CONTENT INFORMATION	**Fixity Information**	Authentication Indicator	Validation	Checksum
				- Value
				- Algorithm
				Digital Signature
	Representation Information [& data object]	Underlying Abstract Form Description	Structural Type	Specific Hardware Requirements
		Transformer Objects	Technical Structure of Complex Objects	- Specific microprocessor req.
		- Platform	File Description	- Specific multimedia req.
		- Parameters	- Image	- Specific peripheral req.
		- Render/analyze engines	- image format and version	Operating System
		- Output formats	- image resolution	- Name
		- Input format	- image dimensions	- Version
		Render/analyze/convert Objects	- image color	Interpreter & Compiler
		- Platform	- image tonal resolution	- Name
		- Parameters	- image color space	- Version
		- Render/analyze engines	- image CLUT	- Instruction
		- Output formats	- image orientation	Object Format
		- Input format	- compression	- Name
				- Version
		Render/Analyze Objects	- Audio	Application
		- Platform	- audio format and version	- Name
		- Parameters	- audio resolution	- Version
		- Render/analyze engines	- duration	
		- Output formats	- bit rate	
		- Input format	- compression	
			- encapsulation	
			- track # and type	
			- Video	
			- video file format & version	
			- frame dimensions	
			- duration	
			- frame rate	
			- compression	
			- video encoding structure	
			- video sound	
			- Text	
			- text format & version	
			- compression	
			- text character set	
			- text associated DTD	
			- text structural divisions	
			- Database	
			- database format & version	
			- compression	
			- datatype & representation category	
			- representation form & layout	
			- maximum size of data element values	
			- Executables	
			- code type and version	

		Known System Requirements	
		Installation Requirements	
		Storage Information	
		Access Inhibitors	
		Finding/Searching Aids & Access Facilitators	