UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE BIBLIOTECONOMIA E COMUNICAÇÃO
CURSO DE GESTÃO EM BIBLIOTECAS UNIVERSITÁRIAS

Márcio Rohan Silva

DOCUMENTOS DIGITAIS: um estudo sobre métodos de avaliação de formatos de arquivos digitais e sua adequação à preservação digital.

Porto Alegre
Dezembro de 2009

Márcio Rohan Silva

DOCUMENTOS DIGITAIS: um estudo sobre métodos de avaliação de formatos de arquivos digitais e sua adequação à preservação digital.

Monografia de conclusão do Curso de Especialização em Gestão de Bibliotecas Universitárias da Faculdade de Biblioteconomia e Comunicação da Universidade Federal do Rio Grande do Sul, apresentado como requisito parcial para obtenção do grau de Especialista em Gestão de Bibliotecas Universitárias.

Orientador: Prof. Dr. Rafael Port da Rocha

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Carlos Alexandre Netto
Vice-Reitor: Rui Vicente Oppermann

FACULDADE DE BIBLIOTECONOMIA E COMUNICAÇÃO
Diretor: Ricardo Schneiders da Silva
Vice-Diretora: Regina Helena Van der Laan

DEPARTAMENTO DE CIÊNCIAS DA INFORMAÇÃO
Chefe: Ana Maria Moura
Chefe substituta: Helen Beatriz Frota Rozados

CURSO DE ESPECIALIZAÇÃO EM GESTÃO DE BIBLIOTECAS UNIVERSITÁRIAS
Coordenadora: Ida Regina Chitto Stumpf
Coordenadora substituta: Sônia Caregnato

S586d

Silva, Márcio Rohan.
Documentos digitais: um estudo sobre métodos de avaliação de formatos de arquivos digitais e sua adequação à preservação digital / Márcio Rohan Silva. Porto Alegre: UFRGS, 2009.

81f. : il. + 1 CD-ROM

Orientador: Prof. Dr. Rafael Port Rocha

Monografia (Especialização) – Curso de Especialização em Gestão de Bibliotecas Universitárias. Departamento de Ciências da Informação. Faculdade de Biblioteconomia e Comunicação. Universidade Federal do Rio Grande do Sul. Porto Alegre, RS.

Inclui bibliografia.

1. Documentos digitais. 2. Formatos de arquivos digitais. 3. Métodos de avaliação de formatos de arquivos digitais. I - Título

1. Formatos de arquivos digitais - CDU: 004.6

DEPARTAMENTO DE CIÊNCIAS DA INFORMAÇÃO
Rua Ramiro Barcellos, 2705 – CEP. 90035-007
Telefone: (51) 3308.5067
Campus Saúde
Bairro Santana
Porto Alegre - RS

DOCUMENTOS DIGITAIS: um estudo sobre métodos de avaliação de formatos de arquivos digitais e sua adequação à preservação digital.

Esta monografia foi analisada e julgada adequada para a obtenção do título de Especialista em Gestão de Bibliotecas Universitárias e aprovada em sua forma final pela orientadora e pela Banca Examinadora designada pelo Curso em Especialização em Gestão de Bibliotecas da Faculdade de Biblioteconomia e Comunicação da Universidade Federal do Rio Grande do Sul.

Aprovada em: ____/____/____

Banca Examinadora:

_____

Prof. Dr. Rafael Port da Rocha (orientador)
Doutor em Ciências da Computação pela Universidade Federal do Rio Grande do Sul

_____

Prof. Dra. Sônia Elisa Caregnato
Doutora em Information Studies pela University of Sheffield

_____

Prof. Dra. Helen Beatriz Frota Rozados
Doutora em Comunicação e Informação pela Univesidade Federal do Rio Grande do Sul

AGRADECIMENTOS

Agradeço a todos que colaboraram para elaboração deste trabalho, especialmente à Coordenação do Curso de Especialização em Gestão de Bibiotecas Universitárias, à Direção da Biblioteca Central e ao Professor Orientador Rafael Port da Rocha, que de forma generosa e atuante tornou possível a conclusão deste estudo.

RESUMO

Investiga o problema da preservação de documentos digitais, cujo foco está centrado nos métodos de avaliação de formatos de arquivos digitais. Com base no referencial teórico, desenvolve um modelo de avaliação de formatos de arquivos digitais que mescla critérios de avaliação de três projetos de instituições nacionais: Biblioteca do Congresso Americano, Arquivo Nacional do Reino Unido e Biblioteca Nacional dos Paises Baixos. Aplica o modelo desenvolvido para investigar o *site* do Departamento de Assuntos Estudantis da Universidade Federal do Rio Grande do Sul (UFRGS) e para avaliar a aptidão deste para a preservação digital. Conclui que há um crescimento referente ao desenvolvimento e uso de técnicas para a preservação digital, identifica a existência de métodos para a avaliação de formatos digitais e que estes métodos possuem muitos critérios equivalentes e apresenta um modelo em que critérios são combinados.

Palavras-chave: Formato de arquivo digital. Modelo de avaliação de formato de arquivo digital. Preservação digital por longos períodos

ABSTRACT

This work investigates the problem of the preservation of digital documents, whose focus is centered on methods for the evaluation of digital file formats. Based on the theoretical framework, it develops a model for the evaluation of digital file formats which combines file format evaluation criteria of three projects of national institutions: Library of Congress, National Archives of United Kingdom and National Library of Netherlands. It applies the model to investigate the site of the Department of Student Affairs (SAE, in Portuguese abbreviation) at the Federal University of Rio Grande do Sul (UFRGS, in portuguese abbreviation), and to assess its suitability for digital preservation. It concludes that there is a growing concerning to the development and use of techniques for digital preservation, identifies the existence of methods for the evaluation of digital formats and that these methods have many equivalent criteria, and presents a model that amalgamates criteria.

Keywords: Digital file format. Evaluation model digital file format. Digital preservation. Long-therm preservation

LISTA DE QUADROS

Quadro 1	Vantagens e limitações da preservação digital	17
Quadro 2	Conservação do objeto digital	19
Quadro 3	Representações de bits de um objeto digital	23
Quadro 4	Formatos de arquivos digitais	27
Quadro 5	Fatores de sustentabilidade de preservação digital LC	32
Quadro 6	Critérios para avaliação de formatos de arquivos *NK*	38
Quadro 7	Critérios para avaliação de formatos de arquivos, segundo a KB	42
Quadro 8	Modelo completo para preservação digital	44
Quadro 9	Requisitos das instituições utilizadas como parâmetro na criação do novo modelo de avaliação de formatos digitais	53
Quadro 10	Critérios do novo modelo de avaliação de formatos digitais	54
Quadro 11	Requisitos do modelo	59
Quadro 12	Exemplo de análise do formato PNG, sob a perspectiva do novo modelo	61
Quadro 13	Dados coletados pelo processo de *web archiving*	63
Quadro 14	Pré-condições e limites para o web archiving	64
Quadro 15	Análise dos formatos de arquivos do site da SAE	67

SUMÁRIO

1 INTRODUÇÃO **10**

1.1 *Problema e Justificativa* **13**

1.2 *Objetivos* **14**

2 PRESERVAÇÃO DIGITAL **15**

2.1 *Problemas que envolvem preservação digital* **15**

2.2 *Vantagens e limitações da preservação digital* **15**

2.3 *Estratégias de preservação digital* **17**

3 DOCUMENTOS DIGITAIS **22**

3.1 *Objetos digitais* **23**

3.2 *Metadados* **24**

4 FORMATOS DE ARQUIVOS DIGITAIS **26**

5 AVALIAÇÃO DE FORMATOS DE ARQUIVOS DIGITAIS **29**

5.1 *Projeto Library of Congress (LC)* **30**

5.2 *Metodologia INFORM* **34**

5.3 *Portable Document Format (PDF/A):* *norma ISO 19005-1* **36**

5.4 *National Archives of United Kingdom (UK)*: *guia de preservação digital* **38**

5.5 *National Library of Netherlands (KB)*: *guia de preservação digital* **41**

5.6 *Dissertação de mestrado sobre preservação de documentos digitais de Bodê* **43**

5.7 *Registros de formatos digitais* **44**

6 METODOLOGIA **49**

7 MÉTODO DE AVALIAÇÃO **52**

7.1 *Critérios do modelo de avaliação de formatos digitais* **55**

7.2 *Valoração dos critérios* **60**

8 AVALIAÇÃO DO *SITE* DA SECRETARIA DE ASSISTÊNCIA ESTUDANTIL (SAE) **62**

8.1 *Processo de coleta de dados no site da SAE - UFRGS* **63**

8.2 *Análise da amostra coletada do site da SAE - UFRGS* **66**

9 CONCLUSÕES E CONSIDERAÇÕES FINAIS **75**

REFERÊNCIAS **77**

1 INTRODUÇÃO

É inquestionável que o uso dos documentos digitais vem aumentando consideravelmente. A *School of Information Management and Systems* at *University of Califórnia at Berkeley* (2003), realizou estudo em que foi relatado que uma quantidade superior a 90% das informações produzidas no mundo, no ano de 2002, foram registradas em meio digital. Infere-se que a tendência seja de aumento.

As características e as funcionalidades dos documentos digitais facilitam a vida do usuário moderno sedento de tempo e que busca, incansavelmente, maior velocidade na realização de seus processos cotidianos. Uma das características mais importantes desses documentos consiste na possibilidade de edição de seus conteúdos de forma simples, diferentemente do que ocorre com documentos impressos. Existem outras funcionalidades como a possibilidade de acessibilidade em rede, de agregar metadados para fins preservacionais, entre outras. Assim, nota-se que suas potencialidades são infinitas e que eles vieram para ficar.

Dessa forma, além de dados e documentos, a memória e o patrimônio cultural da humanidade encaminha-se para ser preservada digitalmente. Fato que dá ao assunto uma relevante importância, justificando um estudo mais detalhado de suas especificidades, características e tendências.

O assunto tem grande complexidade pela existência de muitas variantes, possibilidades e interesses comerciais. As tecnologias envolvidas nos processos de criação, acesso e utilização dos documentos eletrônicos vêm sendo rapidamente substituídas por novas tecnologias incompatíveis com as antigas. O risco de perda das informações registradas em arquivos digitais é real, uma vez que esses dados devem ser lidos e acessados através de computadores e *softwares* que tendem a evoluir rapidamente. Essa evolução, por vezes, cria um descompasso que pode tornar inacessíveis tanto a mídia quanto os arquivos

utilizados originalmente no processo de registro do conteúdo informacional dos documentos. Os meios de armazenamento e transporte de dados podem tornar-se instáveis. Questões ligadas à autenticidade e a acessibilidade dos documentos digitais por longos períodos, somadas à volatilidade das mídias usadas para o registro desses dados e à obsolescência tecnológica (tanto de *hardware* como de *software*) são – sem dúvida – os maiores problemas ligados à preservação de documentos digitais. Pouco se sabe a respeito de quanto tempo os dados mantidos em padrões digitais poderão ser utilizados. Há também uma grande dependência dos formatos de arquivos digitais que sofrem constantes modificações, obrigando os gestores de dados a fazer migrações de formatos, na medida em que vão se tornando obsoletos.

Segundo Bodê (2008), diversas propostas de atividades relacionadas à preservação de documentos têm como base a forma e a estrutura na qual estão gravadas nos documentos digitais e constituem, conforme o jargão técnico, os Formatos de Arquivo**.**

Seguindo essa linha, as mudanças tecnológicas ensejam a necessidade da criação de políticas de preservação que vislumbrem a constante necessidade de planejamento, normatização e atualização de todos os processos envolvidos na preservação de documentos digitais.

Diante desse contexto, esta pesquisa está focada no estudo da preservação de objetos digitais, especificamente no que tange aos formatos dos documentos digitais, que codificam em seus conteúdos os textos e as imagens (em movimento ou fixas). Apesar de serem importantes para a preservação da memória digital, os suportes físicos envolvidos nos processos de preservação digital não serão objeto deste estudo.

A monografia está estruturada da seguinte forma: o capítulo 1 apresenta a temática, o problema, a justificativa e os objetivos; os capítulos de 2 a

5, o referencial teórico utilizado para dar sustentação ao estudo; os capítulos 6 e 7 abordam, respectivamente, a metodologia e o método de avaliação de formatos digitais; o capítulo 8 aplica o método proposto na SAE; e finalmente, o capítulo 9 apresenta conclusões e recomendações sobre o estudo.

1.1 Problema e Justificativa

O tema desta pesquisa é a preservação dos documentos digitais disponibilizados nos sites de órgãos administrativos. O foco do estudo está limitado ao conceito de formato de documento digital e à criação de um modelo capaz de verificar se os formatos de arquivos digitais em estudo estão em conformidade com os requisitos de preservação de documentos digitais contemplados na literatura atual. Este estudo foi aplicado na Secretaria de Assuntos Estudantis (SAE) da Universidade Federal do Rio Grande do Sul (UFRGS).

O problema em questão pode ser assim definido: os tipos de formatos de arquivo utilizados nos documentos digitais, disponibilizados nos sites dos órgãos Administrativos da UFRGS, são adequados à preservação por longos períodos?

O avanço tecnológico e as mudanças de suporte informacionais levam os profissionais da informação a questionarem-se sobre qual a melhor forma de preservar os materiais digitais. A UFRGS e seus órgãos administrativos estão inseridos nesse contexto de avanço e mudanças tecnológicas. Os documentos divulgados pela instituição, antes registrados, em sua maioria, no suporte papel, hoje são também disponibilizados no site da Universidade por meio de arquivos digitais.

A importância de se realizar este trabalho nasce da preocupação com o aumento da produção do conhecimento registrado em arquivos digitais e da necessidade de descobrir se estes recebem um tratamento adequado para sua preservação e armazenamento. Dessa forma, é possível contribuir para preservação da memória documental da Universidade.

1.2 Objetivos

O objetivo principal consiste em avaliar os formatos de arquivos digitais e sua adequação à preservação digital, levando-se em consideração os métodos de avaliação, requisitos e registros de formatos de arquivos digitais.

Para tanto, foram definidos os seguintes objetivos específicos:

a) investigar os formatos de arquivos digitais;
b) investigar métodos e critérios de avaliação de formatos de arquivos digitais no que diz respeito a sua adequação à preservação de documentos digitais disponibilizados na *web*;
c) estabelecer critérios para avaliar os formatos de documentos digitais na web, com relação à preservação digital;
d) avaliar, com relação à preservação digital, os formatos dos documentos digitais disponibilizados no *site* da SAE - UFRGS.

2 PRESERVAÇÃO DIGITAL

Arellano (2004) afirma que é cada vez mais imperiosa a necessidade de contar com mecanismos que garantam a preservação dos documentos em formato digital. Os arquivos e as bibliotecas têm de assumir a responsabilidade pela criação de técnicas, políticas institucionais de preservação e normas que vibializem esse trabalho. As bibliotecas são responsáveis pelas coleções para uso permanente, por sua proteção perante as ameaças e eventuais reparos que se façam necessários (WEBB, 2000).

Preservação Digital é, então, segundo FERREIRA (2006, p. 20):

> [...] um conjunto de actividades ou processos responsáveis por garantir o acesso continuado a longo-prazo à informação e restante património cultural existente em formatos digitais. A Preservação Digital consiste na capacidade de garantir que a informação digital permanece acessível e com qualidades de autenticidade suficientes para que possa ser interpretada no futuro recorrendo a uma plataforma tecnológica diferente da utilizada no momento da sua criação.

Há que se ter a devida preocupação com eventuais problemas que o processo de Preservação Digital possa apresentar, o quais são apresentados na seção que segue.

2.1 Problemas que envolvem a preservação digital

O material digital carrega consigo um problema estrutural que coloca em risco a sua longevidade. Embora um documento digital possa ser copiado infinitas vezes sem nenhuma perda de qualidade, este exige a presença de um contexto tecnológico para que possa ser utilizado. Essa dependência tecnológica torna-o vulnerável à rápida obsolescência a que geralmente a tecnologia está sujeita.

Como ressaltam Boeres e Arellano (2005), outro parâmetro preocupante é falta de pessoal preparado para levar adiante os processos de preservação digital. As bibliotecas - em geral – enfrentam o problema de falta de funcionários para as tarefas simples às quais cotidianamente estão envolvidas: empréstimos, guarda de material. Atividades que envolvam qualificações mais complexas tornam-se ainda mais difíceis de serem executadas dada a quantidade insuficiente de pessoal e a falta de qualificação.

Mas um problema bastante preocupante relacionado à preservação digital, quando associada ao objeto deste trabalho - formatos de arquivos digitais – está na possibilidade de o arquivo digital possuir formato proprietário. Nessa situação, os detalhes técnicos podem ou não ser divulgados ao público. A consequência disso é que, quando ocorrer a obsolescência de software haveria a necessidade de se manter toda uma gama de hardware e software capazes de acessarem a codificação não divulgada do arquivo proprietário. Caso se conhecessem os detalhes técnicos do arquivo, quaisquer *softwares* poderiam ser implementados com a capacidade de interpretá-los, garantindo assim a preservação do acesso a seu conteúdo. De nada valem estratégias de preservação como emulação ou migração se não conhecermos as características e detalhes técnicos dos arquivos a serem preservados.

2.2 Vantagens e Limitações da Preservação Digital

A Preservação Digital apresenta vantagens e limitações. Enumeram-se a seguir aspectos positivos e negativos associados à preservação de documentos digitais, conforme mostrado no Q*uadro 1* - vantagens e limitações da preservação digital, elaborado por Proença e Lopes (200?).

Vantagens	Limitações
Diminuição do acesso aos originais	Capacidade incipiente dos sistemas de informação em assegurar a preservação a longo prazo.
Fidelidade ao original	Fragilidade dos suportes de armazenamento digital
Acesso direto e imediato	Complexidade e custos da preservação digital
Possibilidade distribuição em rede	Dependência social da informação digital
Possibilidade de compartilhamento de informação	Limitações decorrentes dos direitos autorais
Fim das barreiras geográficas	Ausência de normas e padrões
Economia de tempo e espaço	Constante preocupação com armazenamento e migração de dados
Rapidez na atualização dos dados	
Possibilidade de manter cópias de segurança	

Quadro 1 – vantagens e limitações da preservação digital
Fonte: Proença e Lopes (200?)

Como se pode notar, através da leitura do *Quadro 1*, as limitações decorrem de situações inerentes ao processos naturais de Preservação Digital. Por exemplo, é muito provável que a capacidade incipiente dos sistemas de informação em assegurar a preservação a longo prazo finde com o andamento dos procedimentos e a conseqüente aquisição de experiência pelos gestores dos sistemas de informação. Esse aspecto de temporalidade contido nas limitações da preservação digital diminui seu grau de importância, quando comparado às vantagens da preservação digital. Portanto, quanto ao grau de importância, as vantagens suplantam as limitações da preservação digital.

2.3 Estratégias de preservação digital

Os problemas da preservação estão sendo sanados a partir de propostas de estratégias de preservação digital, que podem ser agrupadas, segundo Lee (2002), em três grupos: emulação, migração e encapsulamento. Thomaz e Soares

propõem estratégias de preservação de documentos digitais, classificando-as em estratégias estruturais e estratégias operacionais:

> As estratégias estruturais dizem respeito aos investimentos ou esforços iniciais por parte da instituição, preparando seu ambiente para o processo da preservação digital. As estratégias operacionais, por sua vez, representam as atividades ou medidas concretas de preservação digital (THOMAZ; SOARES, 2004, p. 3).

Assim, as políticas de preservação de documentos digitais, além dos procedimentos operacionais, têm um viés gerencial e administrativo, uma vez que traçam diretrizes e normas, a fim de preparar o ambiente para o processo de preservação digital.

Nessa linha, o Conselho Nacional de Arquivos (CONARQ), com o intuito de garantir a autenticidade e a integridade da informação, manifesta a importância da implementação de ações:

> É necessário o estabelecimento de políticas públicas, diretrizes, programas e projetos específicos, legislação, metodologias, normas, padrões e protocolos que minimizem os efeitos da fragilidade e da obsolescência de *hardware*, *software* e formatos e que assegurem, ao longo do tempo, a autenticidade, a integridade, o acesso contínuo e o uso pleno da informação a todos os segmentos da sociedade brasileira (CONSELHO NACIONAL DE ARQUIVOS, 2006).

Ferreira (2006) define a existência de dois tipos de conservação: conservação do objeto digital em seu formato original e conservação do conteúdo intelectual do objeto digital através das diversas formas de migração.

O Q*uadro 2* - conservação do objeto digital ilustra as tipologias definidas por Ferreira (2006), associando o tipo de conservação com sua respectiva estratégia de preservação.

Conservação	Estratégia de preservação
conservação do objeto digital em seu formato original	preservação da tecnologia
	refrescamento
	emulação
conservação do conteúdo intelectual do objeto digital	migração para suportes analógicos
	atualização de versões
	conversão para formatos concorrentes
	normalização
	migração a-pedido
	migração distribuída

Quadro 2 – Conservação do objeto digital
Fonte: criação do autor (2009)

A conservação do objeto digital em seu formato original, apresenta três estratégias de preservação (FERREIRA, 2006):

a) preservação de tecnologia: essa estratégia está voltada à preservação dos equipamentos e programas em seu estado original. Consiste, essencialmente, na conservação e manutenção de todo o hardware e software necessários à correta apresentação dos objetos digitais Os museus de tecnologia exemplificam muito bem os conceitos que envolvem essa estratégia. O mesmo autor critica essa estratégia ao afirmar que "[...] este tipo de estratégia introduz dificuldades ao nível da gestão do espaço físico, manutenção e custo de operação, tornando-as inadequadas para aplicação a longo-prazo" (FERREIRA, 2006, p. 33);

b) refrescamento: na corrida contra o tempo, surge o refrescamento de suporte. Tecnologia utilizada para a transferência de informação de um suporte em vias de se tornar obsoleto para outro suporte físico de armazenamento mais atual;

c) emulação: estratégia voltada à criação de *software* que reproduz as funções de um determinado ambiente, a fim de permitir a reprodução de outros *softwares* criados a partir do ambiente reproduzido. É uma técnica muito conhecida em aplicações de jogos eletrônicos.

Ferreira (2006) enumera oito estratégias de preservação associadas à conservação do conteúdo intelectual do objeto digital através da migração. Esse tipo de conservação consiste em transferir periodicamente um objeto digital de uma tecnologia de hardware e/ou software para outra mais atual. Apresenta-se a definição de cada uma das estratégias, segundo orientações do autor:

a) migração para suportes analógicos: consiste em converter um objeto digital para um suporte não digital, por exemplo, imprimir um texto e armazená-lo em papel;

b) atualização de versões: utilizado essencialmente para *software*, consiste em criar uma versão mais atual do mesmo;

c) conversão para formatos concorrentes: consiste em converter o objeto digital para um outro formato concorrente, por exemplo, converter uma imagem para o formato jpeg;

d) normalização: consiste em reduzir o número de formatos de um repositório de objetos digitais, criando condições favoráveis ao processo de interoperabilidade entre sistemas distintos;

e) migração-a-pedido: consiste em aplicar processos de conversão no objeto digital original, pois os diversos procedimentos de conversão podem gradativamente degradar o formato original do objeto;

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

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

h) pedra de rosetta digital: como no caso da Pedra de Rosetta descoberta no delta do Nilo em 1799, essa estratégia propõe preservar amostras representativas dos objetos digitais, às quais permitirão sua recuperação, ao invés das regras de decodificação do objeto.

Das idéias de todos os autores citados, no que tange às estratégias de preservação digital, há de se ressaltar a importância de identificação e análise do contexto institucional . A consequência da análise bem sucedida do panorama identificado na instituição trará como consequência uma maior probalbilidade de que seja escolhida a melhor estratégia para solucionar a situação em estudo.

A aplicação das técnicas e procedimentos ligados à operacionalização da estratégia deve estar subsidiada pela análise do contexto institucional, considerando principalmente:

a) a vontade política dos gestores;
b) a capacidade de investimento financeiro;
c) a situação dos formatos de arquivos divulgados pela instituição.

Assim, uma vez traçadas e identificadas essas características, a possibilidade de erros no processo de preservação digital diminuirão drasticamente.

3 DOCUMENTOS DIGITAIS

Diversos tipos de suportes foram utilizados para registrar o conteúdo de documentos em distintos períodos históricos da humanidade. Com o passar dos anos e a evolução tecnológica, os documentos passaram a ser registrados em meio digital.

Marcacini explica muito bem o conceito de documento digital, através de uma anologia com o documento tradicional:

> Pode-se dizer que experimentamos hoje um mundo virtual onde, no lugar de átomos, agora temos que nos acostumar com uma realidade de coisas formadas tanto por átomos como por bits. O documento tradicional, em nível microscópico, não é outra coisa senão uma infinidade de átomos que, juntos, formam uma coisa que, captada pelos nossos sentidos, nos transmite uma informação. O documento digital, então, é uma das seqüências de bits que, captada pelos nossos sentidos com o uso de um computador e um software específico, nos transmite uma informação (MARCACIANI, 1999, p. 5)

Assim, o documento digital é criado em um ambiente eletrônico que codifica a informação em dígitos binários e deve privilegiar o pensamento ou fato que se quer perpetuar e não o suporte no qual estes se materializam.

Os documentos digitais possuem as seguintes características, segundo Bodê (2008): legibilidade por máquinas, independência entre suporte físico e sua correspondente mensagem e, finalmente, o fato de serem codificados em linguagem binária digital.

3.1 Objetos digitais

Os objetos digitais são cadeias de bits acessados a partir de uma informação de representação, que faz com que esses bits sejam visíveis e compreensíveis, através da combinação de *hardware*, *software*, suporte e pessoal (ROCHA, 2009).

Os objetos digitais, de acordo com pesquisadores da *University of Leeds* (200?), são sequências de zeros e uns que representam dados codificados. São formados por informações de conteúdo, de estrutura e de apresentação de conteúdo. Os elementos dessa composição estão ligados à forma pela qual esses objetos transmitem informações a seus usuários, à forma com que essas informações estão organizadas no objeto digital, à forma com que seu conteúdo será apresentado .

O *Quadro 3* - representações de *bits* de um objeto digital elaborado por Rocha (2007) ilustra a representação das informações dos objetos digitais, desde a visualização dos dados codificados pelo computador, através de uns e zeros, até a visualização da informação pelo usuário, através de uma representação de alto nível.

Bits	**Tags**	**Textos**
1010110101 1010101010 1011001101 1010101010 1010111010 1101010101	 Objeto Digital 	Objeto Digital
Informações de Representação Bits → Txt	Informações de Representação Txt → html	

Quadro 3 - Representações de bits de um objeto digital
Fonte: Rocha (2007)

Os objetos digitais podem ser natos, quando nascem em ambientes digitais: o arquivo .doc nativo do *MS WORD* é um exemplo; ou, convertidos de um objeto físico. Os livros impressos que foram digitalizados são exemplos de objetos digitais convertidos de um objeto físico.

3.2 Metadados

O objetivo deste item é abordar, de forma sucinta, a importância dos metadados para a preservação do documento digital.

Os metadados viabilizam a execução de procedimentos importantes para a preservação dos documentos digitais: propiciam a interoperabilidade de bibliotecas digitais, facilitam os processos e as estratégias de preservação de documentos digitais, pois podem registrar a história das migrações de um documento no decorrer de sua vida útil, entre outros. Segundo Arellano (2004), os metadados informam as partes importantes do objeto digital e indicam a sua localização.

Além disso, o processo de recuperação da informação armazenada e preservada em repositórios é feito através do uso de metadados. Daí sua importância para preservação de documentos digitais.

Ainda do ponto de vista preservacional, os metadados servem para descrever objetos digitais em detalhe. Thomaz e Soares (2004) enumeram três razões para adotar metadados na descrição de objetos digitais:

a) facilitar a pesquisa e a identificação de suas fontes de informação;
b) gerenciar seu fluxo dentro de processos;
c) representar suas estruturas para possibilitar o acesso.

Concluem os autores que a criação e a manutenção de metadados tendem a ser um componente fundamental para a maioria das estratégias de preservação digital.

4 FORMATOS DE ARQUIVO DIGITAIS

O que se entende pelo termo "formato de arquivo”? Nos primórdios da existência da computação, os formatos de arquivo eram classificados de acordo com o tipo de programa utlizado para gerar o arquivo. Os arquivos continham um único tipo de dado: um texto, uma imagem ou um som. Hoje, fala-se em formato de arquivo de **conteúdo predominante**. “Pode-se falar em aplicativos que geram predominantemente textos, imagens fixas, sons, etc., mas não exclusivamente esses tipos de conteúdos.” (BODÊ, 2008, p. 57). Cita ainda, o referido autor, como exemplo, o formato MP3, utilizado especialmente para o registro de sons, mas que pode conter legendas e textos explicativos.

Uma definição de formato de arquivos deve incluir o conceito da “especificação”, cuja função é guardar os requisitos organizacionais do arquivo. A especificação do formato de arquivo descreve as funções e a disposição dos *bits* que formam o arquivo digital. Bodê ilustra esse conceito:

> 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. (BODÊ, 2008, p. 21).

Um formato de arquivo, então, pode ser definido como uma especificação de dados digitais, que registra seu conteúdo e mostra como está organizado.

Um programa que utiliza os dados de um arquivo deve ser capaz de reconhecer e, eventualmente, ter acesso a esses dados. Por exemplo, um navegador da *web* é capaz de processar e exibir um arquivo no formato *HTML* para que ele apareça como uma página da Web, mas não pode exibir um arquivo em formato *XLS*, concebido para o *Microsoft Excel*. Um formato específico de arquivo é formado pelo nome do arquivo e pela sua extensão, que serve basicamente para permitir que o sistema operacional - instalado no computador -

associe esse arquivo a um programa, que poderá ser disparado ou não pelo usuário, caso queira executar determinada tarefa. Convencionalmente, a extensão é separada do nome do arquivo e contém três ou quatro letras que identificam o seu formato. Um programa pode ou não utilizar a extensão como parâmetro de reconhecimento do formato do arquivo. Para isso, analisará os *bits* no arquivo, identificando, desta maneira o seu formato.

A microinformática, os computadores pessoais, a explosão na criação de *softwares* novos trazem como consequência o surgimento - cada vez maior - de novas especificações de formatos de arquivos. Bodê (2008) chama atenção para existência de uma grande quantidade de especificações técnicas e para uma infinidade de formatos de arquivos diferentes, que evoluíram a partir de versões antigas de aplicativos, hoje, descontinuados.

Existem muitos formatos diferentes de arquivo; a título de exemplificação, o *Quadro 4* - Formatos de arquivos enumera os mais conhecidos.

Conteúdo predominante	**Tipo de arquivo**
TEXTO	DOC, RTF, OPEN OFFICE, ODF
PÁGINAS DA WEB	HTM, HTML, XML
IMAGENS FIXAS	GIF, JPG, TIFF
IMAGENS EM MOVIMENTO	MEU, KAR, MP3, MP4
IMAGENS EM 3D	CAD, BIFF, X4D
PROGRAMAS	.EXE, COM
MULTIMÍDIA	.MP3 , KAR, MP4

Quadro 4 – Formatos de arquivos digitais
Fonte: criação do autor (2009)

As modificações e os aperfeiçoamentos que ocorrem com os programas incidem nos formatos de arquivos a eles associados. Assim, um determinado formato de arquivo poderá ter inúmeras versões. O formato PDF, por exemplo,

possui versões distintas: PDF 1.4, VERSÃO 1.0 1-4; PDF 1.5, VERSÃO 1.5; PDF / A 1 baseado na versão 1.4 ; PDF / A 1b ; PDF / X .

5 AVALIAÇÃO DE FORMATOS DE ARQUIVOS DIGITAIS

Este capítulo visa à identificação, na literatura, de métodos e critérios de avaliação de formatos de documentos digitais para fins de preservação digital, com objetivo de preservá-los por longos períodos. Os dados encontrados subsidiarão a escolha das características que contemplarão a criação do modelo de formato para preservação dos documentos digitais objeto desse estudo.

A avaliação de formatos de arquivos digitais consiste basicamente em analisar as características do formato de arquivo de um documento de acordo com requisitos estabelecidos em um determinado modelo, projeto ou método de avaliação. A partir daí, identifica-se quais requisitos estão sendo contemplados, atribui-se pesos distintos em função da conformidade ou não aos requisitos e verifica-se a aptidão do formato em estudo para fins de preservação digital.

A *Library of Congress* (2009) aponta as finalidades do processo de avaliação de formatos de arquivos, dentre os quais destacam-se os seguintes:

a) apoiar o planejamento estratégico sobre formatos de conteúdo digital, a fim de garantir a preservação a longo prazo ;
b) identificar e descrever os formatos que são promissores para a sustentabilidade a longo prazo, e desenvolver estratégias para sustentá-los;
c) identificar e descrever os formatos que não são promissores para a sustentabilidade a longo prazo e desenvolver estratégias para manter o conteúdo que eles contêm.

Nessa linha, Stanescu (2004) faz considerações com relação aos requisitos que a avaliação do formato digital deve prever: avaliação do risco, identificação das propriedades do modelo em questão e as características do formato.

Através de revisão de literatura foi possível identificar um amplo rol de modelos aptos a darem embasamento a esse estudo; dentre todos, destacam-se os seguintes:

a) projeto *Library of Congress*: fatores de sustentabilidade de formatos digitais[1];

b) projeto UPF: formato universal para preservação[2];

c) metodologia *Investigation of Formats based on Risk Management* (INFORM): seleção de arquivos digitais para registrar informações por longos períodos[3];

d) Norma ISO 19005-1 – *Portable Document Format* (PDF/A)[4].

e) *National Archives of United Kingdom*: guia de preservação digital[5];

f) *National Library of Netherlands*: guia de preservação digital[6];

g) Preservação de documentos digitais: o papel dos formatos de arquivo dissertação de Mestrado (BODÊ, 2008)[7].

A seguir, detalha-se cada um destes projetos voltados à avaliação de formatos de arquivos digitais identificados na revisão da literatura.

5.1 Projeto *Library of Congress* (LC)

O projeto da *Library of Congress* (LC) dedica-se à análise dos aspectos técnicos de formatos digitais. Preocupa-se com os formatos associados às mídias independentes de conteúdo digital, ou seja, documentos, cujo conteúdo, que nor-

---

[1] http://www.digitalpreservation.gov/formats/
[2] http://info.wgbh.org/upf/
[3] http://www.dlib.org/dlib/november04/stanescu/11stanescu.html
[4] http://www.iso.org/iso
[5] http://www.nationalarchives.gov.uk/
[6] http://www.kb.nl/index-en.html
[7] http://www.cipedya.com/doc/176563

malmente é gerido como arquivo, não depende de um determinado suporte físico.

Dessa forma, não fazem parte do escopo desse projeto os formatos de mídia associado-dependente de conteúdo digital, isto é, os formatos que são dependentes e indissociavelmente ligados à mídia física, por exemplo, DVDs, CDs de áudio, fitas de vídeo, entre outros.

Esse projeto serve de apoio para tomada de decisões no que tange à preservação dos formatos digitais. Para isso, dentre outras questões, o projeto visa responder:

> Que formatos digitais devem ser plenamente apoiados por sistemas, ferramentas automatizadas, ou fluxo de trabalho associados com o ciclo de vida de processos de conteúdo digital em debate na Biblioteca? Em outras palavras, serve de apoio para a recepção e validação de conteúdos digitais , a seleção de conteúdos digitais, preparação de conteúdo digital para a guarda responsável de longo prazo, e estabelecimento de estratégias de conservação. (*LIBRARY OF CONGRESS*, 2007, tradução nossa)

Os objetivos do projeto da *Library of Congress*, segundo Arms e Fleischhauer (2003), são os seguintes:

a) desenvolver uma estrutura de planejamento e estratégias sobre formatos de conteúdo digital, a fim de garantir a preservação a longo prazo de conteúdos digitais;

b) construir um inventário de informações sobre os formatos atuais e emergentes, incluindo a identificação de ferramentas e documentação detalhada que são necessárias para garantir que a Biblioteca do Congresso pode gerenciar o conteúdo criado ou recebido nestes formatos;

c) identificar e descrever os formatos que são promissores para a sustentabilidade a longo prazo e desenvolver estratégias para sustentá-los, incluindo

recomendações para utilização de ferramentas e a documentação necessária para a sua gestão.

A *Library of Congress* (2007) definiu sete fatores que se aplicam na avaliação de formatos digitais para preservação de documentos digitais. Esses fatores influenciam a viabilidade e o custo de preservação da informação. Eles são importantes quaisquer que sejam as estratégias adotadas como base para futuras ações de preservação: migração para novos formatos, emulação de software, ou de uma abordagem híbrida. Os fatores definidos pela LC são apresentados no *Quadro 5* - fatores de sustentabilidade de preservação digital segundo a *Library of Congress.*

Fatores de sustentabilidade	**Significado**
Disclosure	Divulgação da especificação
Adoption	Adoção
Transparency	Transparência
Self-documentation	Autodocumentação
External dependencies	Dependências externas
Impact of patents	Impacto das patentes
Technical protection mechanisms	Mecanismos técnicos de proteção

Quadro 5 - Fatores de sustentabilidade de preservação digital LC
Fonte: Library of Congress (2007)

A seguir são detalhados os requisitos definidos pela LC, que se aplicam na avaliação de formatos digitais orientados à preservação digital:

a) divulgação da especificação: diz respeito à existência e ao grau de acessibilidade às especificações técnicas do arquivo. Serve para viabilizar o entendimento de como a informação é representada (codificada a partir de *bits* e/ou *bytes*) no arquivo digital;

b) adoção: a adoção relaciona-se à aprovação do formato; refere-se ao grau em que o formato já é utilizado pelos principais criadores, divulgadores, ou usuários de recursos de informação. Se um formato é amplamente adotado, é menos provável que se torne obsoleto rapidamente e as ferramentas de migração e emulação estariam mais propensas a sair da indústria, sem investimentos específicos para acompanhar a obsolescência, acarretando gastos menores;

c) transparência: refere-se ao grau em que a representação digital está aberta à análise. A transparência é reforçada se os conteúdos textuais (incluindo metadados embutidos em arquivos de conteúdo não-texto) estiverem codificados na norma de codificações de caracteres (por exemplo, no UNICODE UTF-8) e armazenados em ordem natural de leitura. Muitos formatos digitais usados para a divulgação de conteúdo utilizam a encriptação ou a compressão. Cabe lembrar que a criptografia é incompatível com a transparência; compressão inibe transparência;

d) autodocumentação: objetos digitais auto-documentáveis são menos vulneráveis a catástrofes do que os objetos de dados que são armazenados separadamente dos seus respectivos metadados. Um objeto digital que contém metadados descritivos e incorpora metadados técnicos e administrativos relativos à sua criação e fases iniciais do seu ciclo de vida torna mais fácil o gerenciamento, controle de sua integridade e usabilidade e a transferência de confiabilidade de um sistema de arquivos para o seu sistema sucessor;

e) dependências externas: refere-se ao grau em que um determinado formato depende especificamente de um hardware, sistema operacional, ou software para sua utilização. Algumas formas de conteúdos digitais interativos, embora não estejam vinculadas a determinados suportes físicos, são projetadas para uso com hardware específico, como um microfone ou um *joystick*;

f) impacto das patentes: patentes relacionadas com um formato digital podem inibir a capacidade das instituições para sustentar conteúdo nesse formato. A existência de patentes pode atrasar o desenvolvimento de códigos abertos. Codificadores e decodificadores para transcodificação de conteúdos de formatos obsolescentes podem aumentar a taxa de licença. Cabe ressaltar que a existência de patentes não é um problema potencial, mas que se deve levar em conta as condições que os titulares de patentes definirão para seu uso. É sobre isso que sobrecaem os custos;

g) mecanismos técnicos de proteção: o conteúdo de um repositório não deve ser protegido por mecanismos técnicos, tais como criptografia, implementados de maneira a impedir que seu conteúdo possa ser preservado e acessível às futuras gerações.

5.2 Metodologia INFORM

Fundada em 1967, a *Online Computer Library Center* (OCLC) é uma organização, sem fins lucrativos, criada para prestar serviços bibliotecários computadorizados e de pesquisa. É dedicada a facilitar o acesso à informação mundial. Mais de 71.000 bibliotecas, em 112 países e territórios em todo o mundo, usam os serviços da OCLC para localizar, obter, catalogar, emprestar e preservar materiais de biblioteca. Pesquisadores, estudantes, professores, bibliotecários, arquivistas e outras intituições utilizam os serviços da OCLC para obter informações bibliográficas, resumos e textos completos quando e onde precisarem.

INFORM (Investigação de Formatos Baseados em Gerenciamento de Riscos) é uma metodologia desenvolvida na OCLC para investigar e medir os fatores de risco dos formatos digitais e fornecer orientações para os planos de ação de preservação. Em outras palavras, tenta informar e descobrir ameaças

específicas para a preservação digital e medir o seu impacto sobre as decisões de preservação.

A metodologia INFORM, na seção de formatos digitais, analisa os riscos presentes em formatos de arquivos digitais. Esta análise serve de parâmetro para tomada de decisões, principalmente no que concerne à migração de arquivos digitais.

> A metodologia INFORM analisa as ameaças à preservação e durabilidade, define ferramentas e processos de avaliação necessários para se chegar a uma classificação de risco para qualquer objeto digital, formato e plano de preservação. (STANESCU, 2005, p. 75, tradução nossa).

Stanescu (2005) aborda a metodologia INFORM focando seu objetivo precípuo que é o de definir ferramentas, processos e métrica necessária para selecionar formatos mais aptos a suportar a passagem do tempo.

A metodologia INFORM sugere ainda uma lista contendo 11 (onze) fatores de risco capazes de comprometer a preservação digital:

a) taxas de licença ou *royalties* podem ser necessárias;
b) especificação não disponível para inspeção independente;
c) versões anteriores da especificação são incompatíveis com a atual;
d) especificação muito complexa, extensa, ambígua ou pouco documentada;
e) especificação não é amplamente aceita;
f) especificação é única em sua classe e não pode ser mapeada para outra ou os metadados embutidos não podem ser mapeados para outros formatos;
g) especificação não permite cópias idênticas, tornando a operação de *refresh* impossível;

h) especificação utiliza esquemas DRM[1] (*Digital Rights Management*), envelopes assinados, seções criptografadas ou marcas d´agua;
i) especificação permite extensões ou características largamente suportadas como *JavaScript* e outras;
j) indisponibilidade de equipe com os conhecimentos necessários;
l) procedimentos de teste e equipe são rapidamente superados por mudanças de especificações.

Destaca-se a necessidade de um elevado nível de especialização para pôr em prática os preceitos técnicos da Metodologia INFORM.

> As principais desvantagens deste sistema é que, para a atribuição de um fator de risco para uma das seis áreas de risco, um elevado nível de especialização é exigido para cada ambiente individual. Assim, o processo deve possuir funcionários altamente qualificados e não pode ser padronizados com facilidade. (RAUCH, 2007, p. 102, tradução nossa).

O processo, além de não poder ser padronizado, deverá contar com pessoal altamente qualificado. Essa constatação limita a implantação dessa metodologia de avaliação em instituições cujo *staff* não tenha a qualificação exigida. Para sua implantação há necessidade de avaliar, num primeiro momento, a qualificação do grupo de trabalho ligado à instituição.

5.3 *Portable Document Format* (PDF/A)

O projeto PDF/A, conforme a ISO (2006), teve início em outubro de 2002.

Grupos de arquivistas, gerentes de bancos de dados, bibliotecários e usuários, através de suas discussões, deram origem ao processo de conservação

---

[1] termo genérico que se refere ao acesso a tecnologias de controle que podem ser utilizados pelos fabricantes de hardware, os editores, os detentores de direitos autorais e os indivíduos para tentar impor limitações ao uso de conteúdos digitais e aparelhos (WIKIPEDIA, 2009)

do documento digital por longos períodos – projeto PDF/A. Mais adiante, em 2003, passa a ser desenvolvido como norma ISO *international*. O documento ISO 19005-1 foi publicado em 2005 e normatiza as características, os dados e os aspectos relativos ao padrão PDF/A.

A norma baseia-se na utilização de arquivos PDF especificamente orientados para preservação a longo prazo. O PDF/A é ideal para documentos cujo conteúdo e aparência devam permanecer estáveis durante longos períodos de tempo. A intenção foi a de criar um formato para representar documentos originalmente criados em PDF, convertidos a partir de outros formatos digitais ou digitalizado em papel ou microfilme. Assim, o formato PDF/A é uma variação do formato PDF, incrementado de algumas funcionalidades voltadas à preservação. Suas características, conforme Sullivan (2006), são as seguintes:

a) independência de dispositivo;
b) autoconteúdo;
c) autodescrição;
d) transparência;
e) acessibilidade;
f) abertura da especificação;
g) adoção.

Como se vê, o fomato PDF/A possui várias características que vão ao encontro do que se pretende definir como modelo de arquivo ideal para preservação. Tais características estão muito próximas daquelas estabelecidas pelo projeto da *Library of Congress,* como as ideais para preservação por longos períodos.

5.4 *National Archives of United Kingdom* (UK): guia de preservação digital

O *National Archives of United Kingdom* (2009) elaborou um documento para orientar os gerentes e administradores de acervos de documentos digitais no que tange ao planejamento e seleção de critérios de avaliação de formatos de arquivos voltados à preservação digital. Essas orientações estabelecem uma série de critérios, cuja finalidade é ajudar os criadores de dados e arquivistas nas tomadas de decisões sobre questões ligadas aos formatos de arquivos.

Os requisitos definidos pelo *National Archives of United Kingdom* são apresentados no *Quadro 6* – Critérios para avaliação de formatos de arquivos.

Fatores de sustentabilidade	Significado
Ubiquity	Popularidade
Support	Suporte
Disclosure	Divulgação
Documentation quality	Qualidade da documentação
Stability	Estabilidade
Intellectual property rights	Direitos de propriedade intectual
Metadata support	Metadados de apoio
Complexity	Complexidade
Interoperabiblity	Interoperabilidade
Viability	Viabilidade
Re-usability	Reutilização

Quadro 6 – Critérios para avaliação de formatos de arquivos *NK*.
Fonte: Adaptado de National Archives of United Kingdom (2009)

Os critérios estabelecidos pelo *National Archives of United Kingdom* são os seguintes:

a) popularidade: verifica se o formato está estabelecido na comunidade a que se destina. É importante, pois, caso seja verdadeira essa premissa, os fornecedores de software e os usuários apoiarão de forma mais ampla e duradoura, fazendo com que o formato tenha um ciclo de vida mais prolongado. Formatos populares são, portanto, preferíveis;

b) suporte: a disponibilidade de uma ampla gama de ferramentas e suporte de software é muito importante, pois eliminam a dependência de um único fornecedor e por isso é preferível;

c) divulgação: os responsáveis pela gestão e preservação a longo prazo dos registros exigem acesso a informações técnicas detalhadas sobre os formatos de arquivo utilizados. Formatos que têm especificações técnicas disponíveis no domínio público são os mais recomendados. É o caso dos padrões abertos, tais como JPEG. Os desenvolvedores de formatos proprietários também podem publicar suas especificações livremente, como, o PDF, ou comercialmente, como, Adobe Photoshop. O uso de formatos abertos é altamente recomendado, sempre que possível;

d) qualidade da documentação: a disponibilidade de documentação do formato não é, por si só, suficiente. A documentação também deve ser completa, precisa e compreensível;

e) estabilidade: a especificação de formato deve ser estável e não sujeita a constantes ou grandes mudanças ao longo do tempo. Novas versões do modelo também devem ser compatíveis com os programas utilizados anteriormente;

f) facilidade de identificação e validação:a capacidade de identificar com precisão o formato de um arquivo de dados e confirmar que é um exemplo válido é vital para o uso continuado. Quando bem concebidos, os formatos devem facilitar sua identificação através do uso de "números mágicos";

g) direitos de propriedade intelectual: formatos podem utilizar tecnologias cobertas por patentes ou outros direitos e restrições de propriedade, como algoritmos de compressão de imagem. Isso pode limitar a utilização futura de objetos nesse formato. Em particular, "patentes submarinas", que são patentes registradas por uma empresa ou grupo que não as desenvolve, mas que possui seus direitos.

Ocorre que o objeto patenteado acaba sendo utilizado ou desenvolvido por outro grupo o qual não tem idéia da existência do registro. Quando o desenvolvimento está concluído o produto toma corpo no mercado, surge a empresa detentora dos direitos e cobra a conta. Por isso essas patentes são chamadas patentes submarinas - quando menos se espera, elas emergem. Formatos que são livres de patentes são recomendados;

h) metadados de apoio: alguns formatos de arquivo prevêem a inclusão de metadados. Esses metadados podem ser gerados automaticamente quando são criados, quando inseridos pelo usuário ou numa combinação de ambos. São importantes sobretudo para para a preservação a longo prazo, pois fornecerão as informações sobre a procedência e características técnicas do dados. Por exemplo, um arquivo TIFF pode incluir campos de metadados para registrar detalhes como a marca e o modelo do scanner, o software e sistema operacional utilizados, o nome do criador e uma descrição da imagem. Da mesma forma, Documentos do Microsoft Word podem incluir uma série de metadados para apoio, fluxo de documentos e controle de versão, dentro das propriedades do documento. Os formatos que oferecem apoio metadados são preferíveis;

i) complexidade: formatos devem ser selecionados para serem usados sobre uma base capaz de suportar toda a gama de características e funcionalidades necessárias para que seu propósito seja atingido. É igualmente importante, no entanto, evitar a escolha de formatos excessivamente complexos. Quanto mais complexo o formato maior será o custo para geri-lo e preservá-lo;

j) interoperabilidade: a capacidade de troca de registos eletrônicos com outros usuários e sistemas de TI é também um importante fator a ser considerado. Formatos que são suportados por uma ampla gama de softwares ou que sejam independentes de plataforma são os mais desejáveis. Isso também tende a apoiar a sustentabilidade a longo prazo de dados, facilitando a migração de um ambiente técnológico para outro;

l) viabilidade: alguns formatos devem proporcionar facilidades de detecção de erros para permitir a detecção de arquivos quebrados ou corrompidos. Muitos formatos incluem um CRC (*Cyclic Redundancy Check*). Existem outras técnicas que também podem ser utilizadas. Por exemplo, o formato PNG incorpora sequências para verificar três tipos específicos de erros que poderiam ser introduzidas. Formatos que fornecem equipamentos, tais como esses, são mais robustos e, portanto, preferíveis;

m) reutilização: certos tipos de dados devem conservar a capacidade de serem tratados, pois sua reutilização dependerá dessa funcionalidade. Por exemplo, uma planilha criada em formato XLS gravada e convertida em formato PDF, efetivamente, perde grande parte da sua capacidade de ser processada. Assim, a exigência de manter uma versão do registro capaz de ser processado também deve ser considerada.

5.5 *National Library of Netherlands*: guia de preservação digital

Uma das tarefas mais importantes da *Koninklijke Bibliotheek - National Library of Netherlands (KB)* - é o cuidado com a preservação do patrimônio cultural holandês. Essa preocupação trouxe como conseqüência a criação do *e-Depot*, ambiente digital que garante acesso a objetos digitais por longos períodos (NATIONAL LIBRARY OF NETHERLANDS, 2009a).

Em 2003, a *National Library of Netherlands* (2009b) criou duas unidades organizacionais: o e-Depot, que é responsável pela gestão operacional do repositório dos arquivos digitais que constituem parte do acervo da KB, e o Departamento de Preservação Digital, responsável pela realização e desenvolvimento de pesquisas na área da preservação digital. Estas tarefas de investigação incluem a vigilância tecnológica e a colaboração e o estudo de todos os aspectos de preservação em relação ao *e-Depot*, tais como metadados,

formatos de arquivo, controle e desenvolvimento de estratégias para o acesso permanenente.

A *National Library of Netherlands*, conforme artigo criado por Rog e Wijk (2007), desenvolveu um método de avaliação de riscos para formatos de arquivos. Esse método pode ser usado para definir estratégias de preservação digital para os formatos de arquivo. A escolha de um formato de arquivo no momento da criação influencia o acesso desse objeto a longo prazo. O método de avaliação da KB contém sete critérios de sustentabilidade para os formatos de arquivo, que são avaliados e recebem pesos e notas em função do seu grau de importância de acordo com metodologia própria.

Os requisitos definidos pela KB são listados no *Quadro 7* – critérios para avaliação de formatos de arquivos, segundo a *National Library of Netherlands* - KB.

Fatores de sustentabilidade	**Significado**
Openness	Abertura
Adoption	Adoção
Complexity	Complexidade
Technical Protection Mechanism	Mecanismos de proteção técnica
Self-documentation	Auto-documentação
Robustness	Robustez
Dependencies	Dependencias

Quadro 7 – Critérios para avaliação de formatos de arquivos, segundo a KB
Fonte: Adapatado de National Library of Netherlands (2009a)

Os critérios estabelecidos pela KB são apresentados, a seguir, com suas especificidades:

a) abertura: indica a possibilidade de acumular-se conhecimento sobre a estrutura do formato de arquivo. O conhecimento sobre um formato de arquivo irá aumentar a chance de sucesso do planejamento de preservação digital;

b) adoção: indica a popularidade e a onipresença de um formato de arquivo;

c) complexidade: indica o quão complicado um formato de arquivo pode ser de decifrar;

d) mecanismos de proteção técnica: indica a possibilidade de um formato de arquivo de restringir o acesso a seu conteúdo. Limitações ao acesso do conteúdo de arquivos digitais podem ser problemáticas, quando houver necessidade de colocar em prática estratégias como a de migração;

e) autodocumentação: indica as possibilidades de o formato facilitar o processo de encapsulamento de metadados. Essa funcionalidade é de extrema importância, pois será possível preservar o arquivo sem a necessidade de utilizar outras fontes de informações;

f) robustez: indica se o formato de arquivo corre o risco de corromper-se ou se possui proteção contra *cracks*. Abrange também informações relativas às suas mudanças. Mudanças frequentes podem ameaçar a continuidade da acessibilidade do formato para longos períodos. Grandes diferenças entre as gerações de um formato de arquivo poderiam pôr em risco a sua continuidade;

g) dependências: indica a dependência de recursos específicos, representando um risco para preservação digital pelo simples fato de que perdido o recurso, perdido o arquivo.

5.6 Dissertação de mestrado sobre a preservação de documentos digitais de Ernesto Carlos Bodê

Os requisitos mencionados nesta seção fazem parte de uma Dissertação de Mestrado de autoria de Ernesto Bodê, de 2008, submetida ao Programa de Pós-Graduação em Ciência da Informação do Departamento de Ciência da Informação e Documentação da Universidade de Brasília. O estudo discorre sobre

a preservação de documentos digitais, focando a relação dos formatos de arquivo com a preservação por longos períodos, limitando o universo da pesquisa ao Poder Judiciário brasileiro. Para isso, o pesquisador valeu-se da criação de um modelo de avaliação de formatos de arquivo adequado para a preservação de documentos digitais por longos períodos. Esse modelo foi criado a partir da análise comparativa dos requisitos da Norma ISO 19005-1, de 2005 (PDF/A), Projeto da *Library of Congress* (fatores de sustentabilidade) e metodologia *INFORM*. Tal modelo consta do Q*uadro 8* e denomina-se, segundo o próprio autor, Modelo Completo para Preservação Digital:

Requisitos
independente de dispositivos
metadados incorporados
transparência
desativação de proteções
especificação não-proprietária
especificação normatizada
auto-sufiência

Quadro 8 – Modelo completo para preservação digital
Fonte: Bodê (2008)

5.7 Registros de formatos digitais

Registros de formatos digitais são bases de dados que disponibilizam informações técnicas *on line* e se prestam para subsidiar a avaliação de formatos digitais. Entre os registros *on-line* de informações técnicas de formatos digitais,

pode-se destacar o projeto PRONOM, o *Global Digital Format Registry* (GDFR) e os Fatores de Sustentabilidade da *Library of Congress.*

O projeto PRONOM foi criado pelo *National Archives of United Kingdom*, lançado em março de 2002. É um recurso que disponibiliza informações sobre formatos de arquivos, produtos de *softwares* e outros componentes técnicos necessários para apoiar o acesso a registos eletrônicos e objetos digitais por longos períodos. A versão mais recente representa um avanço significativo. Entre as funcionalidades desenvolvidas, destacam-se: a reformulação do modelo de dados (que permite a captura de informações técnicas detalhadas sobre formatos de arquivo, cujo objetivo é servir de apoio à futura interoperabilidade com outros sistemas de registro previstos) e o lançamento do *software* DROID *(Digital Record Object Identification)* para a identificação automática de formatos de arquivos digitais. Trata-se de um *software* livre, que lê arquivos digitais e analisa sua estrutura, identificando a que formato pertence, tomando como base arquivos de assinatura XML, gerados a partir de informações disponibilizadas no Registro Técnico PRONOM (NATIONAL ARCHIVES OF UNITED KINGDOM, 2009).

A *figura 1*, a título exemplificativo, mostra a tela de abertura do projeto PRONOM.

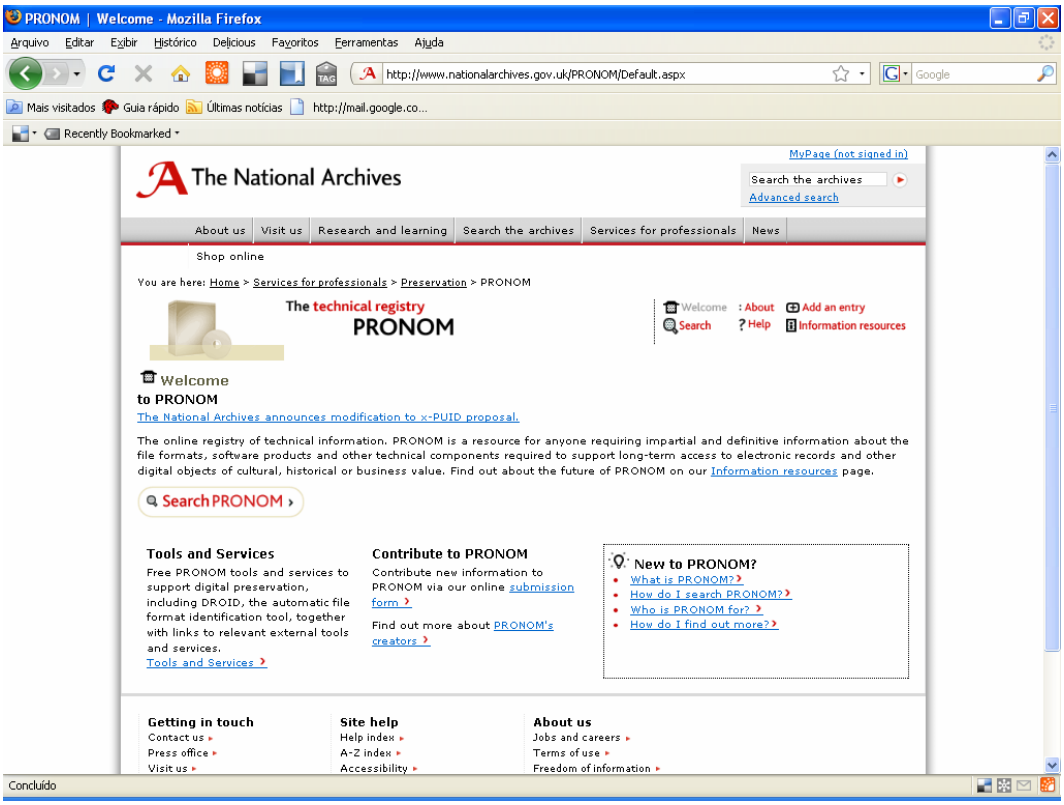

Figura 1 – Projeto PRONOM
Fonte: Pronom (2009)

O *Global Digital Format Registry* (GDFR) foi criado pela *Harvard University Library* e seu objetivo é dar apoio aos processos que envolvam preservação digital. É um registro de formatos digitais que está sendo substituído pelo Registro Único de Formatos Digitais (UDFR). O UDFR é uma proposta de criação de um registro único envolvendo a *Global Registry Digital Formats* (GDFR) e o PRONOM. O PRONOM tem, como ponto forte, uma significativa base tecnológica e construiu um poderoso banco de dados de informações sobre os vários formatos digitais. No entanto, é mantido por uma única organização, tornando-o vulnerável a mudanças políticas e diretivas que ocorram na referida instituição. Já o GDFR, desenvolvido pela Biblioteca da Universidade de Harvard, mostra um modelo para um registro com base na gerência partilhada, hospedagem e contribuição de dados cooperativa. No entanto, GDFR é tecnicamente menos desenvolvida. A idéia é agregar os pontos fortes dos dois registros, criando, dessa forma, uma

fonte de informação de formatos digitais ainda mais robusta e eficiente (GLOBAL DIGITAL FORMAT REGISTRY, 2009).

A *figura 2* mostra a tela incial do *Global Digital Format Registry* (GDFR).

Figura 2 – Tela inicial da Global Digital Format Registry
Fonte: *Global Digital Format Registry* (2009).

A *Library of Congress* disponibiliza, em seu *site,* informações sobre formatos de conteúdo digital. Esse serviço teve início em 2004 e está sendo constantemente ampliado e atualizado. As descrições dos conteúdos dos formatos digitais fornecem informações específicas sobre suas características. Cabe ressaltar que a inclusão de uma descrição para um formato não implica que este seja preferido pela Biblioteca do Congresso. Fato que garante independência às informações encontradas neste registro (LIBRARY OF CONGRESS, 2007).

. A *figura 3* apresenta a tela inicial, que mostra a descrição dos formatos de arquivos digitais classificados por categorias no registro da *Library of Congress*.

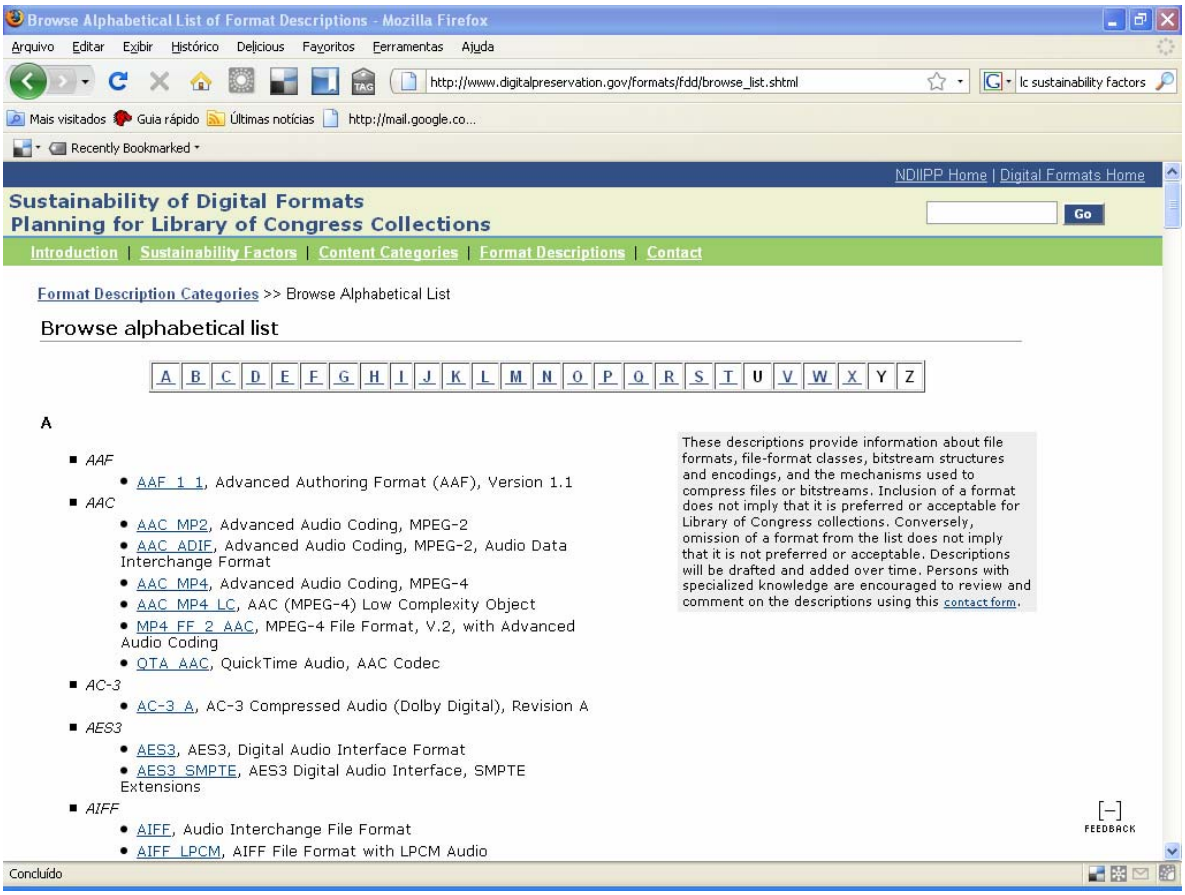

Figura 3 – Fatores de sustentabilidade de formatos digitais da Library of Congress
Fonte: *Library of Congress* (2009).

6. METODOLOGIA

A abordagem utilizada neste estudo é de cunho quantitativo, do tipo exploratório. Para Gil (2002), a pesquisa exploratória é aquela que tem como principal finalidade o aperfeiçoamento de idéias e a descoberta de intuições. Este tipo de pesquisa possui o seu planejamento bastante flexível; já, Minayo (1999) afirma que a abordagem quantitativa é definida por uma população e busca um critério numérico que possibilite gerar e generalizar conceitos teóricos que se pretende testar. Ela transforma em números, opiniões e informações, por meio de recursos e técnicas estatísticas, classificando e analisando-as.

Para contemplar os objetivos propostos, este trabalho desenvolveu um modelo de avaliação de formatos de arquivos digitais, destinado a auxiliar a tomada de decisão no que diz respeito à escolha de formatos de documentos eletrônicos, que facilitem a preservação de documentos digitais por longos períodos. Posteriormente a isso, o modelo desenvolvido foi aplicado para investigar os documentos do site da Secretaria de Assuntos Estudantis (SAE) da UFRGS, a fim de verificar se os seus formatos favorecem a preservação digital por longos períodos.

Para o desenvolvimento do modelo de avaliação, identificou-se, num primeiro momento, na literatura pertinente, três modelos desenvolvidos por instituições renomadas na área da Ciência da Informação: o modelo de avaliação do *National Archives of United Kingdom (UK)*, apresentado na seção 5.4; o modelo de avaliação da *Library of Congress (LC)*, apresentado na seção 5.1 e o modelo de avaliação da *National Library of Netherlands* (KB), apresentado na seção 5.5.

A fim de subsidiar a criação do modelo de avaliação deste projeto, foram comparados os modelos UK, LC e KB, com intuito de identificar-se equivalências e diferenças entre os requisitos que os compunham. O novo modelo foi então

criado a partir da seleção dos requisitos relevantes e presentes em todos os modelos analisados.

No modelo desenvolvido, parâmetros de avaliação foram estabelecidos para os requisitos, a fim de identificar o grau de conformidade de um determinado formato com o requisito. Estabeleceu-se que um formato de arquivo poderá contemplar um requisito da seguinte forma: totalmente conforme, parcialmente conforme e totalmente desconforme.

A fim de buscar uma medida quantitativa, esses parâmetros receberam pesos distintos. Quando um formato estiver totalmente conforme com as características de um requisito, recebe o peso 10; quando estiver parcialmente conforme, recebe o peso 5 e quando estiver desconforme recebe peso 0. Dessa forma, a cada um desses parâmetros, foram atribuídos os valores 10 (dez), 5 (cinco) e 0 (zero), respectivamente.

A aplicação desse método para avaliar os documentos do site da na Secretaria de Assuntos Estudantis (SAE) da UFRGS foi composta por três etapas:

a) definição de uma amostra, cujo universo foram os arquivos digitais disponibilizados no site do SAE – UFRGS;

b) identificação dos principais formatos existentes nessa amostra;

c) avaliação desses formatos de arquivos conforme o novo modelo de avaliação.

Estabeleceu-se, para definição da amostra, que somente os formatos de arquivos de conteúdo predominantemente textuais, sonoros e de imagem fossem selecionados. A identificação dos principais formatos existentes na amostra, bem

como sua quantidade, foi fruto da aplicação do *software* DROID[1], disponibilizado através do projeto PRONOM. Antes disso, porém, houve a necessidade da utilização do *software WebReapper* [2], cuja finalidade foi extrair todos os arquivos do *site* objeto da coleta de dados.

Para analisar os documentos do *site* da SAE foi utilizada a técnica de *web archiving*, que consiste em fazer uma cópia de todo o seu conteúdo. Esse método foi escolhido para não comprometer a qualidade dos dados coletados, uma vez que toda a análise é feita a partir de um único estado do *site*, representado por essa cópia. Sabe-se que os *sites* de *Internet* têm funcionamento dinâmico. As informações neles postadas podem ser inseridas e retiradas rapidamente, quando menos se espera e sem quaisquer avisos.

Uma vez obtidos os arquivos que espelhavam fidedignamente o *site* da SAE, utilizou-se o *software* DROID para identificar automaticamente os formatos desses arquivos e suas quantidades. Esses procedimentos objetivaram identificar e verificar quantitativamente os formatos utilizados, a fim de avaliá-los com relação à preservação digital.

A análise dos formatos dos arquivos digitais utilizou o modelo de avaliação de formatos de arquivos digitais desenvolvido por este trabalho como parâmetro de comparação, atribuindo-lhes pesos distintos, conforme seu grau de contemplação, tal como foi explicado anteriormente. Essa avaliação calcou-se em informações sobre formatos disponibilizadas nos registros *on-line* do projeto PRONOM, *Global Digital Format Registry* (GDFR), Fatores de Sustentabilidade da LC e documentação dos formatos de arquivos.

---

[1] http://droid.sourceforge.net/
[2] http://www.webreaper.net/

7 MÉTODO DE AVALIAÇÃO

O método de avaliação desenvolvido foi baseado nos critérios de avaliação criados pelos projetos da *Library of Congress, National Archives of United Kingdom* e *National Library of Netherlands*. Consiste na combinação dos principais requisitos contemplados por esses três modelos de avaliação de formatos digitais.

A escolha desses métodos deve-se ao fato de terem sido elaborados por instituições governamentais consagradas no apoio ao desenvolvimento de arquivos e bibliotecas digitais. Essas instituições são referência na área da Ciência da Informação e exercem um papel de vanguarda nos processos que envolvem preservação digital. Além disso são organizações imparciais e transparentes, cujos projetos não estão ligados a nenhum produtor comercial de formatos de arquivos. Tanto a *Library of Congress*[1] quanto o *National Archives of United Kingdom*[2] disponibilizam em seus *sites* um extenso e pertinente rol de informações, dados, ferramentas e documentos sobre preservação digital. Esses materiais auxiliam e dão subsídio a todo processo de preservação digital, servindo como fonte de estudo, pesquisa e desenvolvimento da área. Já a *National Library of Netherlands*[3] iniciou atividades na área de preservação digital no ano de 1992, o que demonstra sua grande experiência e preocupação no avanço investigativo nessa área do conhecimento.

Em função dessa gama de motivos, os métodos citados foram escolhidos para dar suporte e embasar a elaboração de um modelo próprio de avaliação de arquivos digitais, que foi utilizado - neste estudo - para avaliar se os formatos de arquivos disponibilizados no site da Secretaria de Assistência Estudantil da UFRGS estão aptos à preservação digital.

---

[1] http://www.digitalpreservation.gov/formats/
[2] http://www.nationalarchives.gov.uk/
[3] http://www.kb.nl/index-en.html

A determinação dos requisitos a serem utilizados pelo método desenvolvido por este trabalho partiu da análise comparativa dos requisitos presentes nos modelos *National Archives of United Kingdom (UK)*, *Library of Congress (LC)* e nos requisitos da *National Library of Netherlands (KB)*. O *quadro 9* revela esta comparação. Seu arranjo tem como objetivo facilitar a visualização dos elementos que possuem características semelhantes. Além de identificar os requisitos dos modelos analisados, o *quadro 9* mostra os requisitos, cujos conceitos possuam equivalência. Assim, elementos de conteúdos semelhantes aparecem na mesma linha.

NATIONAL ARCHIVES OF UNITED KINGDOM		***NATIONAL LIBRARY OF NETHERLANDS***		***LIBRARY OF CONGRESS***	
Requisito	**Significado**	**Requisito**	**Significado**	**Requisito**	**Significado**
Ubiquity	Popularidade	Openness	Abertura	Adoption	Adoção
Support	Suporte	Adoption	Adoção	Disclosure	Divulgação da especificação
Disclosure	Divulgação	Complexity	complexidade	Impact of patents	Impacto das patentes
Documentation quality	Qualidade da documentação	Technical protection mecanism	Mecanismos técnicos de proteção	Transparency	Transparência
Stability	Estabilidade	Self-documentation	Autodocumentação	Self-documentation	Autodocumentação
Intellectual property rights	Direitos de propriedade intectual	Robustness	Robustez	External dependencies	Dependências externas
Metadata support	Metadados de apoio	Dependecies	Dependências	Technical protection mechanisms	Mecanismos técnicos de proteção
Complexity	Complexidade				
Interoperabiblit y	Interoperabilidad e				
Viability	Viabilidade				
Re-usability	Reutilização				

Quadro 9 – Requisitos das instituições utilizadas como parâmetro na criação do novo modelo de avaliação de formatos digitais
Fonte: dados de pesquisa

Observando o *quadro 9*, percebe-se que a maioria dos requisitos se repete nos três métodos, embora com rótulos e nomes diferentes. Exemplo disso são os critérios ***Support*** (suporte), ***Ubiquity*** (popularidade) e ***Adoption*** (adoção).

Nota-se, também, que os critérios arrolados pela *Library of Congress* e pela *National Library of Netherlands* ocorrem em um número bem menor do que os critérios definidos pelo Arquivo Nacional do Reino Unido. Há, portanto, uma disparidade numérica nos requisitos que formam os dois modelos. Com isso, resolveu-se criar um modelo de avaliação que contivesse a intersecção dos requisitos utilizados pelas três instituições, isto é, que contivesse os critérios presentes nas três instituições, a saber: *Library of Congress*, *National Archives of United Kingdon* e *Library of Netherlands*. Optou-se pela intersecção dos requisitos pela necessidade de se criar um modelo conciso e, ao mesmo tempo, o mais completo possível, capaz de abranger e avaliar todos os formatos de arquivos passíveis de serem encontrados no *site* objeto da coleta de dados.

Em decorrência da análise dos requisitos nas três Instituições, foram identificadas diversas similaridades e, com isso, foi possível elaborar uma lista contendo critérios de conteúdo equivalente. Fato que reforça a importância dos requisitos escolhidos. A partir daí, criou-se a lista UK ∩ LC ∩ KB, cujos elementos são arrolados no *quadro 10* – critérios do novo modelo de avaliação de formatos digitais.

A aplicação da intersecção nos requisitos das instituições analisadas mostrou os seguintes resultados:

Modelo UK ∩ LC ∩ KB	
Requisitos	**Significado**
Openness	Abertura
Support	Suporte
Technical protection mecanism	Mecanismos técnicos de proteção
Self-documentation	Autodocumentação
Robustness	Robustez
Dependecies	Dependências

Quadro 10 – Critérios do novo modelo de avaliação de formatos digitais
Fonte: dados de pesquisa

Assim, os elementos listados no *quadro 9* são, respectivamente, os elementos equivalentes aos dos *quadros 5, 6* e *7*. Por essa razão, esse modelo assume as características contempladas nos três modelos analisados.

7.1 Critérios do método de avaliação de formatos digitais

As seções seguintes apresentam os critérios utilizados no desenvolvimento do método proposto e suas relações de equivalência e similaridades com outros métodos identificados na revisão de literatura.

7.1.1 Abertura

O indicador abertura foi definido a partir de uma expressão de equivalência, que envolveu alguns requisitos dos modelos da LC, KB e UK, e é apresentado a seguir :

OPENNESS$_{(BK)}$ ≈ DISCLOSURE$_{(UK)}$ ≈ INTELECTUAL PROPERTIES RIGHTS$_{(UK)}$ ≈ IMPACT PATENTS$_{(LC)}$ ⇒ **ABERTURA (OPENNESS)**

O indicador abertura ***(openness)***, da National Library of Netherlands (BK), está associado com a possibilidade de se ter acesso à estrutura do formato de arquivo e o indicador divulgação ***(disclosure)***, da *National Archives of United Kingdon* (UK), da mesma forma, ocupa-se da divulgação das informações técnicas detalhadas do formato de arquivo. Dar acesso aos formatos ou divulgá-los são ações semelhantes, distinguindo-se principalmente pela intensidade com que ocorrem. A primeira disponibiliza acesso às informações técnicas, dando a idéia de uma ação com maior passividade; a segunda divulga-as, dando a idéia de pró-atividade e maior complexidade. Porém ambas têm um objetivo em comum: fazer com que as informações técnicas e estrutura dos formatos de arquivos estejam disponíveis ao domínio público. Os requisitos direitos de propriedades intelectuais (***intelecutal properties rights)*** e impactos das patentes (***impact patents)***

também estão ligados à questão do acesso aos formatos, agindo como possíveis limitadores.

A idéia da divulgação transcende a finalidade precípua do requisito: estar acessível tão somente; a das patentes limita-se a frear o interesse de acesso . Escolheu-se, assim, o requisito abertura (***openness)*** por transmitir a idéia de liberdade de acesso e disponibilidade permanente aos interessados. Tal requisito, engloba, porém, as características dos demais elementos no momento em que se estiver avaliando o formato.

7.1.2 Suporte

O indicador suporte foi definido a partir de uma expressão de equivalência, que envolveu alguns requisitos dos modelos da LC, KB e UK, e é apresentado a seguir:

$$\text{ADOPTION}_{(KB,\ LC)} \approx \text{SUPPORT}_{(UK)} \approx \text{UBIQUITY}_{(UK)} \approx \text{INTEROPERABILITY}_{(UK)} \Rightarrow \textbf{SUPORTE(SUPPORT)}$$

O indicador popularidade (***ubiquity)*** utilizado no modelo da UK tem função de verificar se o formato está estabelecido na comunidade a que se destina. Na mesma, linha o indicador adoção (***adoption)*** utilizado no modelo da LC, refere-se ao grau com que o formato é usado pelos principais criadores, divulgadores e usuários. Ambos critérios medem a popularidade do formato, pois partem da premissa de que, quanto mais popular maior o interesse dos fornecedores de software e divulgadores pelo formato e menor a probabilidade de tornar-se obsoleto.

Note-se que o indicador suporte (***support)*** utilizado no modelo da UK também segue nessa linha, mas em sentido inverso; chama, pois, a atenção para

a necessidade de uma ampla gama de ferramentas e fornecedores dedicados a darem suporte ao formato.

Após essa análise, escolheu-se, para formar o rol do nosso modelo, o requisito suporte (***support)***, pois se trata de um conceito mensurável, ao contrário do dos critérios *adoption, ubiquity* e *interoperability* que têm um viés subjetivo.

7.1.3 Mecanismos técnicos de proteção

O indicador mecanismos técnicos de proteção **(technical protection mechanisms)** foi definido a partir de uma expressão de equivalência, que envolveu alguns requisitos dos modelos da LC e KB, e é apresentado a seguir:

TECHNICAL PROTECTION MECHANISMS $_{(LC, KB)}$ ≈ TRANSPARENCY$_{(LC)}$ ≈ COMPLEXITY ⇒ **MECANISMOS TÉCNICOS DE PROTEÇÃO (TECHNICAL PROTECTION MECHANISM)**

Os requisitos mecanismos técnicos de proteção (***technical protection mechanisms),*** transparência ***(transparancy)*** e complexidade (***complexity)*** indicam a possibilidade de o formato de arquivo restringir o acesso a seu conteúdo. Características como proteção por senha, assinatura digital e proteção contra cópias são limitadores no processo de preservação.

A transparência (***transparency***) faz menção ao grau com que o formato está aberto à análise. A complexidade **(*complexity*)**, o quão complicado um formato de arquivo pode ser de decifrar. A preocupação está centrada na possibilidade de o formato ser passível de encriptação ou compressão. Tais funcionalidades são incompatíveis com a preservação digital e inibem a análise do conteúdo dos arquivos. Do mesmo modo, os mecanismos técnicos de proteção, tais como criptografia, dificultam o acesso ao conteúdo do arquivo. Esses três requisitos abordam a mesma preocupação e têm a mesma finalidade: tornar o

conteúdo dos formatos de arquivo abertos a análises e acessíveis a futuras gerações. Por esses motivos, resolveu-se enquadrá-los em um único requisito. Optou-se pelo nome mecanismos técnicos de proteção ***(technical protection mechanism)*** por retratar de forma clara a finalidade a que se propõe.

7.1.4 Autodocumentação

O indicador autodocumentação **(self-documentation)** foi definido a partir de uma expressão de equivalência, que envolveu alguns requisitos dos modelos da LC, KB e UK, e é apresentado a seguir:

SELF-DOCUMENTATION $_{(LC,\ KB)}$ ≈ METADATA SUPPORT$_{(UK)}$ ⇒ **AUTO-DOCUMENTAÇÃO (SELF-DOCUMENTATION)**

Os objetos digitais autodocumentáveis podem conter metadados descritivos, de preservação, técnicos e administrativos. Quando isso ocorre, a possibilidade de encapsulamento do formato é facilitada. Assim, esse requisito é muito importante por propiciar suporte à preservação sem a necessidade da utilização de outras fontes de informação, tais como documentos de criação, entre outros. Em ambos requisitos é abordada a necessidade da presença de metadados, que se justificam pelos motivos expostos. Assim, se optou pela escolha do requisito autodocumentação (***self-documentation)*** por representar melhor a idéia de formato encapsulável.

7.1.5 Robustez

O indicador robustez **(robustness)** foi definido a partir de uma expressão de equivalência, que envolveu alguns requisitos dos modelos da KB e UK, e é apresentado a seguir:

ROBUSTNESS$_{(KB)}$ ≈ VIABILITY$_{(UK)}$≈STABILITY$_{(UK)}$ ⇒ **ROBUSTEZ (ROBUSTNESS)**

O requisito estabilidade (***stability)*** define que o formato não deve se sujeitar a grandes mudanças ao longo do tempo e que as novas versões devem ser compatíveis com os softwares e ferramentas utilizadas até então. Já o requisito viabilidade **(*viability)*** foca atenção na necessidade de o formato possuir a funcionalidade CRC (Cyclic Redundance Check), que é um dispositivo de detecção de corrupção de arquivos. Característica que o torna mais robusto. O requisito robustez **(*robustness)*** engloba essas duas preocupações: os formatos devem possuir funcionalidades que os protejam contra *cracks* e ter em vista a preocupação de não mudar freqüentemente, pois essas mudanças podem ameaçar a continuidade da acessibilidade por longos períodos. Grandes diferenças entre as gerações de formatos poderiam pôr em risco sua continuidade.

Por ser mais completo e abordar as duas características simultaneamente, escolheu-se o requisito robustness.

7.1.6 Dependências

O indicador dependências **(dependencies)** foi definido a partir de uma expressão de equivalência, que envolveu alguns requisitos dos modelos da LC e KB, e é apresentado a seguir:

DEPENDENCIES$_{(KB)}$ ≈ EXTERNAL DEPENDENCIES$_{(LC)}$ ⇒ **DEPENDÊNCIAS (DEPENDENCIES)**

Indica a dependência de recursos específicos, tais como *hardware* específico, *software* específico, *codecs*, entre outros. A dependência de um recurso específico representa um risco para preservação digital por longos períodos, pois a consequência do extravio de quaisquer um desses recursos implicará a perda do conteúdo do formato.

7.2 Valoração dos critérios

Uma vez definidos os requisitos do modelo, foi estabelecido um método que avaliasse quantitativamente o grau com que um determinado formato atende a um requisito. Isso permite a construção de um *ranking* de formatos, cujo intuito é o de identificar os formatos mais bem colocados. Com isso, é possível definir que os mais bem colocados têm um grau de aptidão maior para a preservação digital do que os formatos que obtiveram uma pontuação menor.

Os valores definidos para identificar o grau de atendimento de um requisito variam numa escala de 0 (zero) a 10 (dez), podendo receber os seguintes conceitos:

a) totalmente desconforme → 0 (zero)

b) parcialmente conforme → 5 (cinco)

c) totalmente conforme → 10 (dez)

O *quadro 12* exemplifica a análise de um arquivo do formato *Portable Network Graphics* (PNG), que é um formato utilizado para representar imagens, de acordo com os critérios utilizados pelo modelo, constantes do *quadro 11*:

Sequência	Requisito
1	Abertura
2	Suporte
3	Mecanismos técnicos de proteção
4	Autodocumentação
5	Robustez
6	Dependências

Quadro 11 – Requisitos do modelo
Fonte: dados de pesquisa

A sequência numérica apresentada no quadro 10 está associada aos requisitos do modelo. O requisito abertura recebe a sequência de número 1, o requisito suporte recebe a sequência de número 2, e assim sucessivamente. A sequência definida no quadro 10 foi inserida no quadro 11. Assim, a coluna 1 (um) do quadro 11 representa o requisito abertura; a coluna 2 (dois), o requisito suporte; a coluna 3 (três), o requisito mecanismos técnicos de proteção; a coluna 4 (quatro), o requisito auto-documentação; a coluna 5 (cinco), o requisito robustez, e a coluna 6 (seis) representa o requisito dependências.

CRITÉRIOS PARA SELEÇÃO DE FORMATOS DE ARQUIVOS- SAE MODELO						
PUID	**1**	**2**	**3**	**4**	**5**	**6**
fmt/11	*10*	*10*	*0*	*5*	*0*	*10*

Quadro 12 – Exemplo de análise do formato PNG, sob a perspectiva do novo modelo.
Fonte: dados de pesquisa

O formato fmt/11 (PNG v. 1.0) contempla totalmente os requisitos abertura, pois se trata de um *open archive*; suporte, pois, segundo a LC, é amplamente utilizado e trata-se do formato preferido pela *Florida Center for Library Automation* (FCLA) e *Library and Archives of Canada* para registrar imagens estáticas; dependências, pois não depende de nenhum recurso externo.

Preenche parcialmente o requisito mecanismos técnicos de proteção, pois permite que seu conteúdo seja comprimido, fato que – como se viu na seção 5.1, alínea “c”, não é recomendável para formatos de arquivos destinados à preservação digital; e, por fim, não contempla, segundo a LC, os requisitos autodocumentação, tampouco robustez, conforme dados extraídos do registro PRONOM.

8 AVALIAÇÃO DO *SITE* DA SECRETARIA DE ASSISTÊNCIA ESTUDANTIL (SAE) - UFRGS

O universo de pesquisa nesta fase do estudo limita-se à Secretaria de Assuntos Estudantis (SAE) da UFRGS. A SAE foi criada em junho de 2001, com o objetivo de tratar dos aspectos relativos à política de atendimento à comunidade discente da UFRGS (UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL, 2009).

As atividades precípuas da SAE estão voltadas ao desenvolvimento de programas e projetos, com o intuito de integrar a comunidade discente à vida universitária. Destacam-se entre as principais atividades desenvolvidas pela SAE:

a) a coordenação dos estágios não obrigatórios;

b) o gerenciamento dos estágios obrigatórios;

c) a concessão de benefícios aos estudantes carentes de recursos sócio-econômicos, tais como Bolsa Permanência, Bolsa Treinamento, Moradia Estudantil e Auxílio Alimentação;

d) a administração dos restaurantes universitários, das casas de estudantes e da colônia de férias;

e) o apoio financeiro e logístico a Diretórios e Centros Acadêmicos para a realização de projetos sociais e eventos desenvolvidos pelos estudantes.

Esses processos envolvem a divulgação de normas, portarias, códigos disciplinares, editais, informativos sobre RU´S, colônia de férias, casas de estudantes, diretrizes sobre estágios e documentos vinculados à vida acadêmica discente,.que podem ser disponibilizados no *site* da secretaria.

8.1 Processo de coleta de dados no *site* da SAE – UFRGS

O processo de coleta de dados no *site* da SAE foi descrito, conforme item 6, p. 53.

A coleta de dados no *site* da SAE propiciou que fossem copiados 1255 arquivos, os quais representam 40,9 *megabytes* de espaço ocupado em disco rígido.

Apresenta-se, no *Quadro 13* - Dados coletados pelo processo de *web archiving*, a quantidade de arquivos e o tamanho do universo da pesquisa representado em *Megabytes*.

SECRETARIA	Nº de ARQUIVOS	TAMANHO (MB)
Secretaria de Assuntos Estudantis – SAE	1255	40,9

Quadro 13– Dados coletados pelo processo de *web archiving*
Fonte: dados de pesquisa

A diretiva *robot.txt,* presente em alguns diretórios do *site* do SAE, impossibilitou que a totalidade dos arquivos fosse coletada. A referida diretiva aplica-se para inibir a realização de *donwloads* automáticos de documentos, através de programas desenvolvidos especialmente para esse fim.

Através dessa diretiva, o detentor dos direitos autorais de um *site* pode definir quais as partes do *site* receberão permissão para *download*. Ao utilizá-la, os responsáveis pelo site sinalizam o que não tem permissão para *download*, manifestando claramente sua vontade com relação à cópia de determinados elementos, variáveis ou dados.

Para realizar o *web archiving*, neste estudo, escolheu-se o aplicativo *WebReaper*, que possui funcionalidades extremamente importantes. Tal programa permite definir quantos níveis de *links* serão copiados e dessa forma excluir o risco de copiar dados oriundos de outros *sites*; permite definir quais os

tipos de arquivos serão copiados e com isso evitar que arquivos executáveis, de bases de dados, sejam coletados; permite definir o tamanho máximo ou mínimo dos arquivos copiados, evitando-se assim a cópia de logos, botões ou objetos de sinalização do *site*. O *Quadro 14* - Pré-condições e limites para o *web archiving* arrola as definições e os limites para o processo de busca pelo método W*eb Archiving*, utilizados neste estudo.

Parâmetro	Configuração	Descrição
Alcance dos *links*	3 níveis	Nº de níveis de *links* a serem copiados
Tipos de extensões	Textos, imagens, sons	Tipo de formato de arquivo a ser copiado
Tamanhos arquivos	Mínimo 30 kb	Tamanho mínimo dos arquivos a serem copiados
Servidores de páginas	Servidor original do *site*	Nº de servidores *web* acessados e copiados

Quadro 14 – Pré-condições e limites para o web archiving
Fonte: Adaptado Bodê (2008).

O parâmetro alcance dos *links* refere-se a quantos níveis de *links* o programa irá penetrar no site para proceder à colheita de dados. Esses arquivos serão copiados a partir da raiz do *site*. A página inicial é a de nível 0 (zero) e pode conter vários *links* formando diferentes níveis. A configuração escolhida limita a coleta até o nível 3, impedindo, assim, que o *WebReaper* recupere dados de outros *sites*.

O parâmetro tipo de extensões refere-se aos formatos de arquivo que serão copiados no processo de *web archiving*. A extensão pode ser um parâmetro capaz de identificar um determinado formato. O formato de arquivo *Joint Photographic Experts Group*, por exemplo, recebe a extensão .JPEG . A limitação dos tipos de extensões coletadas faz-se necessária, pois o interesse do estudo está focado em formatos de arquivos que possam representar documentos digitais. Formatos de arquivos executáveis, de bases de dados, entre outros, foram descartados, porque não representam documentos digitais.

Há que se ressaltar a possibilidade de que aplicativos diferentes salvem seus documentos com mesmas terminações. Por isso, a terminação não pode ser utilizada como único e decisivo critério para identificar o formato de um documento. Isso é resolvido através de aplicativos que identificam o formato de um documento analisando suas estruturas, tais como o DROID.

A especificação do “tamanho dos arquivos” é importante, pois é possível encontrar arquivos, cuja função nos *sites* seja simplesmente decorativa ou a representação de botões. Arquivos do tipo *Joint Photographic Experts Group* (JPEG) e, principalmente, do tipo *Graphics Interchange Format* (GIF) são os mais utilizados para essas funções.

A restrição a um único “servidor de páginas” justifica-se pela possibilidade de determinado servidor A remeter – através de um *link* – o acesso a um servidor B, cujo conteúdo ou *sites* alocados não represente o *site* objeto de coleta de dado. Com essa medida, tem-se certeza de que o processo de extração de arquivos deu-se somente no *site* objeto da coleta de dados.

Essas medidas restritivas e limitadoras objetivaram assegurar que os dados da coleta sejam relevantes e garantir que a análise seja fidedigna à realidade encontrada na Secretaria de Assuntos Estudantis da UFRGS.

Os processos de coleta de dados, através do método de *Web Archiving* e a identificação automática dos formatos de arquivos, através do *software* DROID, são similares aos usados por Bodê (2008). Entretanto, o referido autor, utiliza esse processo para identificar quais os formatos foram usados pelo *site* objeto de sua pesquisa, enquanto neste trabalho esses processos são usados tanto para identificá-los, quanto para quantificá-los. E, dessa forma, obter subsídios que propiciem uma análise mais detalhada dos formatos de arquivos disponibilizados no *site* da SAE.

8.2 Análise da amostra coletada do *site* da SAE – UFRGS

Nesta seção, o processo de análise de dados será especificado e detalhado. O processo de coleta de dados e identificação dos formatos de arquivos digitais foi implementada através da aplicação do método de *web archiving* e da utilização do aplicativo DROID, respectivamente.

A seguir tem-se a exposição dos dados coletados e a comparação com os modelos de avaliação de critérios usados para seleção de formatos de arquivos com vistas à preservação digital.

8.2.1 Formatos identificados no *site* da SAE: a amostra

A análise dos formatos de arquivos disponíveis no *site* da Secretaria de Assuntos Estudantis da UFRGS ocorreu em **1255** arquivos. Após filtragem, identificou-se um total de **496** arquivos, que foram considerados irrelevantes para a análise, uma vez que se enquadram no grupo de elementos desprezados para preservação digital, pois são formatos de arquivos identificados ou associados a códigos de programação, bancos de dados, entre outros.

A diferença entre os arquivos disponíveis no *site* do SAE-UFRGS e os arquivos associados a bases de dados ou executáveis estabeleu uma amostra final de **759** arquivos.

8.2.2 Diagnóstico da amostra do *site* da SAE

Partindo-se do estabelecimento da amostra foi possível elaborar um diagnóstico dos arquivos que compõem o *site* da SAE.

Para tanto, cotejou-se a amostra com nosso modelo de avaliação de formatos dos arquivos voltados à preservação digital. Em seguida, utilizando-se

dos procedimentos expostos na metodologia, dentre eles o de atribuição de critérios e valores para os formatos de arquivos analisados, estabeleceu-se o diagnóstico. O Q*uadro 15* demonstra essa análise e apresenta dados quantitativos dos formatos de arquivos encontrados no *site* da SAE. Quanto à estrutura, cada formato é apresentado de acordo com sua versão, ocorrência na SAE, valor obtido para cada critério do método de avaliação e índice, que representa a média aritmética dos valores de cada critério. As colunas de 1 a 6 correspondem, respectivamente, aos critérios: abertura, coluna 1; suporte, coluna 2; mecanismos técnicos de proteção, coluna 3; autodocumentação, coluna 4; robustez, coluna 5 e dependências, coluna 6.

AVALIAÇÃO DOS FORMATOS DE ARQUIVOS – SAE											
NOVO MODELO											
Puid	**Mime(finalidade)**	**versão**	**ocorrência**	**1**	**2**	**3**	**4**	**5**	**6**	**índice**	**%**
Fmt/3	Gif(imagem)	1987a	08	10	10	5	0	0	10	5,833	58,00
Fmt/4	Gif(imagem)	1989a	177	10	10	5	0	0	10	5,833	58,00
Fmt/11	Png (imagem)	1.0	43	10	10	0	5	0	10	5,833	58,33
Fmt/17	Pdf(texto)	1.3	02	10	10	5	5	5	10	7,500	75,00
Fmt/18	Pdf(texto)	1.4	07	10	10	5	5	5	10	7,500	75,00
Fmt/40	Msword(texto)	97-2003	08	0	10	5	5	5	5	5,000	50,00
Fmt/42	Msword(texto)	97-XP	09	0	10	10	0	5	5	5,000	50,00
Fmt/43	Jpeg(imagem)	1.01	116	10	10	10	0	5	10	7,500	75,00
Fmt/44	Jpeg(imagem)	1.02	109	10	10	10	0	5	10	7,500	75,00
Fmt/96	Html(text-markup)	-	152	10	10	0	5	5	10	6,600	66,00
Fmt/99	Html(text-markup)	4.0	09	10	10	0	5	5	10	6,600	66,00
Fmt/100	Html(text-markup)	4.01	45	10	10	0	5	5	10	6,600	66,00
Fmt/102	Xhtml(text-markup)	1.0	21	10	10	0	10	0	10	6,600	66,00
Fmt/116	Bmp(imagem)	3.0	13	5	5	10	0	0	10	5,000	50,00
Fmt/391	Jpeg(imagem)	2.2	40	10	10	0	0	10	10	6,600	66,00
∑	**15**		**759**	**8,27**	**9,60**	**4,3**	**3,20**	**3,70**	**9,30**	**6,2**	

Quadro 15 – Análise dos formatos de arquivos do site da SAE
Fonte: dados de pesquisa

Através da análise do Q*uadro 15* – Análise dos formatos de arquivos do *site* da SAE, podemos inferir que: dos 15 (quinze) formatos de arquivos identificados no *site* da SAE, 7 (sete) são arquivos preponderantemente de imagens e 8 (oito) são arquivos preponderantemente de texto. Não houve identificação de arquivos que registrem preponderantemente sons. Assim, são utilizados vários formatos diferentes que têm como finalidade registrar texto ou imagens.

Os formatos que obtiveram os melhores e os piores índices, considerando a totalidade da amostra, sem quaisquer distinções quanto às suas finalidades, foram, respectivamente, **fmt/43** (*.jpeg*), versão 1.01, que alcançou um índice de 7,5 e **fmt/116** (*.bmp*), versão 3.0, **fmt/40** (*.doc*), versão 97-2003 e **fmt/42** (*.doc*), versão 97-XP, que alcançaram o índice de 5,0 cada.

A *tabela 1* – verificação dos requisitos do modelo em relação aos formatos de arquivos analisados é um recorte do *quadro 15.* Foi elaborada para analisar o comportamento médio dos formatos utilizados pela SAE para cada requisito do modelo, identificando quais requisitos foram menos atendidos e, conseqüentemente, quais os riscos para a preservação digital o *site* da SAE está correndo.

Puid	Mime(finalidade)	versão	1	2	3	4	5	6
Fmt/3	Gif(imagem)	1987a	10	10	5	0	0	10
Fmt/4	Gif(imagem)	1989a	10	10	5	0	0	10
Fmt/11	Png (imagem)	1.0	10	10	0	5	0	10
Fmt/17	Pdf(texto)	1.3	10	10	5	5	5	10
Fmt/18	Pdf(texto)	1.4	10	10	5	5	5	10
Fmt/40	Msword(texto)	97-2003	0	10	5	5	5	5
Fmt/42	Msword(texto)	97-XP	0	10	10	0	5	5
Fmt/43	Jpeg(imagem)	1.01	10	10	10	0	5	10
Fmt/44	Jpeg(imagem)	1.02	10	10	10	0	5	10
Fmt/96	Html(text-markup)	-	10	10	0	5	5	10
Fmt/99	Html(text-markup)	4.0	10	10	0	5	5	10
Fmt/100	Html(text-markup)	4.01	10	10	0	5	5	10
Fmt/102	Xhtml(text-markup)	1.0	10	10	0	10	0	10
Fmt/116	Bmp(imagem)	3.0	5	5	10	0	0	10
Fmt/391	Jpeg(imagem)	2.2	10	10	0	0	10	10
∑	**15**		**8,27**	**9,60**	**4,3**	**3,20**	**3,70**	**9,30**

Tabela 1 - Verificação dos requisitos do modelo em relação aos formatos de arquivos analisados
Fonte: dados de pesquisa

Dos 6 (seis) requisitos que constam do modelo, pode-se apontar o critério **suporte** (coluna número 2), com pontuação de 9,60, como aquele que foi mais bem atendido pelos formatos de arquivos que compõe o *site* da SAE. Os requisitos **abertura** (coluna 1) e **dependências** (coluna 6) também obtiveram boa pontuação. Da análise desses três requisitos, e, conforme seus respectivos preceitos, verifica-se que os documentos da SAE possuem baixa probabilidade de se tornarem obsoletos, possuem liberdade de acesso e disponibilidade permanente e apresentam significativa independência de dispositivos específicos.

São, portanto, requisitos que contribuem para preservação dos documentos digitais por longos períodos.

Estendendo essa análise aos demais requisitos do modelo, tem-se que o requisito **autodocumentação** (coluna 4), cuja pontuação é de 3,20, foi o que mostrou ser menos atendido pelos formatos de arquivos utilizados pela SAE. Chama-se a atenção para os requisitos **robustez** (coluna 5), cuja pontuação atinge 3,70 pontos, e **mecanismos técnicos de proteção** (coluna 3), cuja pontuação soma 4,70 pontos, que sinalizam a possibilidade da presença de riscos no que tange à preservação digital. A baixa pontuação alcançada por esses requisitos mostra uma tendência desfavorável e revela que os documentos da SAE apresentam características que dificultam sua preservação por longos períodos. Por exemplo, a pontuação baixa alcançada pelo requisito **autodocumentação** leva a crer que os documentos disponibilizados no *site* da SAE necessitam de outros documentos que registrem seus formatos de arquivos. E mais, permite deduzir que alguns formatos de arquivos constantes do *site* da SAE sejam versões antigas que não possibilitam a inserção de metadados. Exemplo disso é o formato de imagem **fmt/3** (.gif), cuja versão data do ano de 1987. A tecnologia que possibilita a inserção de metadados está contemplada em formatos de arquivos mais recentes, motivo que explica a precariedade desse recurso em formatos de arquivos antigos e por consequência justifica sua baixa pontuação. Da mesma forma, a baixa pontuação obtida pelo requisito **robustez** significa que parte significativa dos formatos adotados muda frequentemente; já a baixa pontuação somada pelo requisito **mecanismos técnicos de proteção** mostra que alguns formatos permitem a compressão e/ou a criptografia. Esses fatores seguramente acarretam dificuldades à preservação digital.

A *tabela 2* – índices dos formatos de arquivos em função das suas finalidades é um recorte do *quadro 15* e foi elaborada para evidenciar e analisar os índices obtidos pelos formatos de arquivos que constam do *site* da SAE com suas respectivas finalidades: texto ou imagem.

Puid	Mime	versão	índice	finalidade
Fmt/3	Gif(imagem)	1987a	5,833	Imagem
Fmt/4	Gif(imagem)	1989a	5,833	
Fmt/11	Png (imagem)	1.0	5,833	
Fmt/43	Jpeg(imagem)	1.01	7,500	
Fmt/44	Jpeg(imagem)	1.02	7,500	
Fmt/116	Bmp(imagem)	3.0	5,000	
Fmt/391	Jpeg(imagem)	2.2	6,600	

Puid	Mime	versão	Índice	finalidade
Fmt/17	Pdf(texto)	1.3	7,500	Texto
Fmt/18	Pdf(texto)	1.4	7,500	
Fmt/40	Msword(texto)	97-2003	5,000	
Fmt/42	Msword(texto)	97-XP	5,000	
Fmt/96	Html(text-markup)	-	6,600	
Fmt/99	Html(text-markup)	4.0	6,600	
Fmt/100	Html(text-markup)	4.01	6,600	
Fmt/102	Xhtml(text-markup)	1.0	6,600	

Tabela 2 – Índices dos formatos de arquivos em função das suas finalidades
Fonte: dados de pesquisa

Considerando-se as suas finalidades, o formato de imagem que recebeu melhor índice continua sendo o ***fmt/43*** (.jpeg), cuja soma alcança 7.5, assim como o formato de imagem que recebeu o pior índice foi o ***fmt/116*** (.bmp), cuja soma atinge 5.0. Já os formatos de texto que receberam os melhores e piores índices, respectivamente, foram os ***fmt/17*** (.pdf) e ***fmt/18*** (.pdf)***,*** com 7.5, e os formatos ***fmt/40*** (.doc)***, fmt/42*** (.doc)***,*** com índice de 5.0, cada.

A partir desses dados, tem-se que o intervalo entre os melhores e piores índices identificados no *site* da SAE, considerando-se suas finalidades, apresenta situações idênticas. Por exemplo, para os formatos de função textual, a diferença entre o melhor índice (7.5) e o pior índice (5.0) é a mesma diferença encontrada entre o melhor índice (7.5) e o pior índice (5.0) dos formatos de imagem. Fato que permite concluir que os formatos de arquivos com finalidades textuais, disponibilizados no *site* da SAE, do ponto de vista preservacional, assemelham-se, entre si, tanto quanto os formatos de arquivos de imagens.

Há, portanto, um risco equivalente nos formatos de arquivos de textos e de imagens, quando analisam-se seus índices em função de suas finalidades.

A *tabela 3* – Análise da redundância dos formatos de arquivos em função das suas finalidades é um recorte do *quadro 15* e foi elaborada para identificar e apresentar de forma mais clara o uso de diferentes formatos para as mesmas funções.

Puid	Mime	versão	índice	%	finalidade
Fmt/3	Gif(imagem)	1987a	5,833	1,05	Imagem
Fmt/4	Gif(imagem)	1989a	5,833	23,32	
Fmt/11	Png (imagem)	1.0	5,833	5,66	
Fmt/43	Jpeg(imagem)	1.01	7,500	15,28	
Fmt/44	Jpeg(imagem)	1.02	7,500	14,36	
Fmt/116	Bmp(imagem)	3.0	5,000	1,71	
Fmt/391	Jpeg(imagem)	2.2	6,600	5,27	
Σ				**66,65**	
Puid	**Mime**	**versão**	**Índice**	**%**	**finalidade**
Fmt/17	Pdf(texto)	1.3	7,500	0,26	Texto
Fmt/18	Pdf(texto)	1.4	7,500	0,92	
Fmt/40	Msword(texto)	97-2003	5,000	1,05	
Fmt/42	Msword(texto)	97-XP	5,000	1,18	
Fmt/96	Html(text-markup)	-	6,600	20,02	
Fmt/99	Html(text-markup)	4.0	6,600	1,18	
Fmt/100	Html(text-markup)	4.01	6,600	5,92	
Fmt/102	Xhtml(text-markup)	1.0	6,600	2,76	
Σ				**33,29**	

Tabela 3 - Redundância dos formatos de arquivos de acordo com suas finalidades
Fonte: dados de pesquisa

Como se observa na *tabela 3*, 66,65% dos formatos de arquivos usados no *site* da SAE são de imagem e 33,29% são formatos que representam textos.

Dos arquivos utilizados para imagens, três são versões de *.jpeg*, dois são versões de *.gif*, um é versão de *.png* e um é versão de *.bmp*. Analisando os formatos agrupados por versões, 24,37% são versões de *.gif*; 45,45% são versões de *.jpeg*; 1,71% é *versão de .bmp* e 5,66% é versão do tipo *.png.* Portanto, existem **quatro** diferentes formatos de arquivos desempenhando o **mesmo papel**: registrar imagens .jpeg, .gif e .bmp. O formato que obteve melhor índice (.jpeg) é o mais usado, com 45,45%.

Dos arquivos utilizados para texto, dois são versões de *.pdf* , dois são versões de *.doc* e quatro são versões de *.html*. Desses, 1,18% são versões de *.pdf*, 2,23% são versões de *.doc* e 29,88% são do tipo *.html*. Portanto, existem **três** diferentes formatos de arquivos desempenhando a **mesma função**:registrar textos .pdf e .html. O formato que obteve melhor índice (.pdf) é o menos usado, com 1,18%. Já o formato mais usado (.html), com 29,88%, apresenta índice mediano

Diante dessas constatações e considerando os formatos de arquivos disponibilizados pelo *site* da SAE, é possível concluir: os formatos de arquivos com finalidade de registrar imagens estão numa situação mais delicada do que os formatos que registram textos. Por esse motivo, poderia ser priorizada a escolha desses formatos para iniciar um processo de migração e atualização para formatos mais adequados à preservação digital.

Para melhorar ainda mais a situação dos formatos de arquivos com finalidade textual, utilizando-se os mesmos formatos encontrados no *site* da SAE, sugere-se que estes sejam migrados para o formato .pdf., pois apresentam um baixo percentual de utilização e têm os melhores índices.

É possível afirmar que existe uma **redundância de formatos de arquivos** no *site* da SAE. O ideal seria que um único formato de arquivo fosse escolhido para desempenhar precipuamente cada função estabelecida por seu gestor. E mais, seria desejável que o formato escolhido fosse, entre todos, o mais apropriado. Para isso, bastaria aplicar o modelo a todos os formatos candidatos e o que obtivesse melhor índice fatalmente seria a melhor opção.

Dos formatos de arquivos analisados, o único que contemplou plenamente o requisito **autodocumentação** foi o fmt/102 (*.xhtml*). Os formatos que obtiveram pontuação baixa nesse requisito não contemplam de forma adequada a utilização de metadados como funcionalidade latente. Há que se ratificar que alguns

processos de preservação digital, como o encapsulamento, *seção 3.3, alínea g*, prescindem da presença de metadados para que possam ser implementados.

Por fim, cabe enfatizar que a questão da redundância dos formatos de arquivos encontrados no *site* da SAE parece ser um dos problemas mais preocupantes sob o ponto de vista da preservação digital. A redundância faz com que, no futuro, alguns tipos de formatos de arquivos tenham maior probabilidade de serem preservados do que outros. No futuro, arquivos de conteúdos considerados mais importantes podem não ser recuperados, assim como os arquivos considerados menos importantes podem ser recuperados. Há que se ressaltar que no *site* da SAE não há critérios, no que tange à preservação digital, para a escolha dos melhores formatos de arquivos digitais. Hoje o que define a futura recuperabilidade do arquivo digital disponibilizado no *site* da SAE é a sua própria sorte: ter nascido ou não em um formato de arquivo recomendado.

Para minimizar esse problema, pode-se fazer uma padronização dos formatos de arquivos identificados e migrá-los para os que se mostraram mais aptos à preservação digital. Num primeiro momento, essa é a medida cautelar necessária para minimizar eventuais problemas de preservação. Tal procedimento implica redução do número de formatos de arquivos com as mesmas finalidades: de conteúdo textual ou de imagem. Assim, do universo de formatos de arquivos utilizados no *site* da SAE, poderiam ser escolhidos, para registrar imagens, o formato **fmt/44** (.jpeg), versão 1.02, cuja pontuação, de acordo com o modelo aplicado, teve um índice de **7,5** e, para registrar-se textos, o formato **fmt/18** (.pdf), versão 1.4, cuja pontuação teve um índice de **7,5,** ou até mesmo se fazer um estudo para identificar e definir um *rol* de formatos de arquivos que melhor se adaptem às necessidades do *site* da SAE.

Esse procedimento de avaliação isolado não resolve o problema, sendo apenas um instrumento para orientar a formulação de políticas e estratégias de

preservação digital. A questão passa necessariamente pela definição de diretrizes e pela elaboração de uma política de preservação de documentos digitais.

9 CONCLUSÕES e CONSIDERAÇÕES FINAIS

Ao investigar os formatos de arquivos digitais, observou-se que há uma crecente preocupação no que diz respeito à busca de instrumentos de avaliação de formatos digitais para fins de preservação digital. Isso é reforçado pelo fato de que instituições de cunho governamental como o *National Archives of United Kingdom, a National Library of Netherlands e a Library of Congress* estão se preocupando com a questão da preservação de documentos digitais, através da elaboração de inúmeros e relevantes estudos, projetos e métodos voltados à preservação digital. Dentre essas contribuições, destaca-se a criação dos métodos de avaliação de formatos de arquivos digitais e a criação dos registros de formatos digitais: ferramentas de fundamental importância para o estudo e aprimoramento dos processos ligados à preservação de documentos digitais.

Ao investigar métodos e critérios de avaliação de formatos digitais identificou-se o surgimento, a importância e a relevância dos modelos de avaliação e dos fatores de sustentabilidade de formatos de arquivos digitais, dentre os quais destacam-se métodos como o do UK, da LC e da BK, entre outros. Esses métodos possuem correspondências e similaridades, que ficam claras quando expostas a um estudo mais detido de seus critérios. Em função disso, ao estabelecer critérios para avaliar os formatos de documentos digitais na *web* com relação à preservação digital este trabalho desenvolveu um método de avaliação baseado na intersecção de conceitos e similaridades apresentados nos modelos mais representativos direcionados à avaliação de formatos de arquivos digitais.

Com isso, foi possível estabelecer um novo modelo para avaliar os formatos de documentos digitais, a partir da análise de critérios de métodos de avaliação de arquivos digitais e a posterior criação do método de avaliação de formatos de arquivos digitais aplicados na SAE. Para tanto, fez-se uso da possibilidade de identificar semelhanças, estabelecer conexões e intersecções entre os critérios que compunham os modelos estudados.

Como trabalho futuro, poderia-se investigar a viabilidade de dividir os critérios em subcategorias de avaliação, pois, em algumas ocorrências, somente parte dos requisitos é contemplada pelos formatos de arquivos. Com esse particionamento, possivelmente haverá maior fidedignidade e exatidão no processo de avaliação de formatos de arquivos digitais.

O modelo desenvolvido é um instrumento extremamento útil para as instituições, principalmente para servir como instrumento de alerta sobre a importância e a necessidade de ações de preservação digital. Através do diagnóstico da situação encontrada na instituição objeto de análise, o método desenvolvido oferece subsídios para a implantação de políticas e diretrizes ligadas à preservação digital.

Cabe enfatizar a importância da criação de diretrizes e políticas que, baseadas em processos de avaliação e no contexto das instituições, sejam capazes de definir quais formatos de arquivos digitais poderiam ser utilizados no processo de preservação digital, optando-se pelos mais adequados. O método desenvolvido serve como instrumento facilitador para o estabelecimento destas políticas, servindo tanto para alertar a situação como para subsidiar a tomada de decisões quanto aos formatos a serem adotados. No estabelecimento de políticas, além da utilização de métodos de avaliação de formatos, o contexto político e econômico da instituição objeto de estudo têm de ser considerados, analisados e interpretados conjuntamente com a proposta investigativa.

Com relação à aplicação do método desenvolvido no *site* da SAE, identificou-se que este *site* apresenta como principal problema a redundância dos formatos utilizados, no que diz respeito a suas finalidades, servindo como um instrumento de alerta e de subsídio para a tomada de ações no que diz respeito à preservação digital.

REFERÊNCIAS

ARELLANO, M.A. Preservação de documentos digitais. **Ciência da Informação**, Brasília, DF, v. 33, n.2, p. 15-27, maio/ago. 2004. Disponível em < http://www.scielo.br/pdf/ci/v33n2/a02v33n2.pdf>. Acesso em: 29 jul. 2009.

ARMS, C; FLEISCHHAUER, C. **Digital Formats for Library of Congress Collections**: factors to consider when choosing digital formats. Washington: Office of Strategic Initiatives. Library of Congress, 2003. Disponível em: <http://www.digitalpreservation.gov/formats>. Acesso em: 14 set. 2009.

BODÊ, E.C. **Preservação de documentos digitais**: o papel dos formatos de arquivo. Departamento de Ciência da Informação e Documentação da Universidade de Brasília. Programa de Pós-Graduação em Ciência da Informação (Dissertação de mestrado). Brasília: Unb, 2008.153f.

BOERES, S.A.A.; ARELLANO, M. M. **Políticas e estratégias de preservação de documentos digitais**. Salvador: UFBA, 2005.O. Disponível em: < http://www.cinform.ufba.br/vi_anais/docs/SoniaMiguelPreservacaoDigital.pdf>. Acesso em: 01 ago. 2009.

CAMPOS, L.F.B. Metadados Digitais: revisão bibliográfica da evolução e tendências por meio de categorias funcionais. **Revista Eletrônica de Biblioteconomia e Ciência da Informação**, Florianópolis, v. 12, n. 23, 2007. Disponível em: <http://www.periodicos.ufsc.br/index.php/eb/article/viewfile/318/390>. Acesso em: 07 ago. 2009.

CONSELHO NACIONAL DE ARQUIVOS. **Carta para a Preservação do Patrimônio Arquivístico Digital:** preservar para garantir o acesso. . Rio de Janeiro: CONARQ , 2006. Disponível em: <http://www.conarq.arquivonacional.gov.br/Media/publicacoes/cartapreservpatrimarqdigitalconarq2004.pdf>. Acesso em: 20 ago. 2009.

FERREIRA, M. **Introdução à preservação digital**: conceitos, estratégias e actuais consensos. Guimarães, Portugal: Escola de Engenharia da Universidade do Minho, 2006. Disponível em: <https://repositorium.sdum.uminho.pt/bitstream/1822/5820/1/livro.pdf>. Acesso em: 07 ago. 2009.

GIL, A.C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas; 2002.

GLOBAL DIGITAL FORMAT REGISTRY. **About Global Digital Format Registry**: scope and mission. 2009. Disponível em: <http://www.gdfr.info/> . Acesso em: 11 nov. 2009.

ISO. **PDF/A-1:** ISO 19005-1:2005. Maryland: AIIM , 2006. Disponível em:< http://www.aiim.org/documents/standards/PDF-A/19005-1_FAQ.pdf>. Acesso em: 12 out. 2009. (Frequently asked questions – FAQ).

LEE, K-Ho et al. The state of the art and practice in digital preservation. **Journal of Research of National Standards and Technology**, New York, v. 107, n. 1, p. 93-106, jan./fev., 2002.

LIBRARY OF CONGRESS. **Sustainability of digital formats** : planning for Library of Congress collections. 2007. Disponível em: <http://www.digitalpreservation.gov/formats/sustain/sustain.shtml>. Acesso em: 10 ago. 2009.

MARCACINI, A.T.R. **O documento eletrônico como meio de prova**. 1999. Disponível em:<http://augustomarcacini.cjb.net/index.php/DireitoInformatica/DocumentoEletronico>. Acesso em: 06 ago. 2009.

MINAYO M.C.S. **Pesquisa Social**: métodos e técnicas. São Paulo: Atlas; 1999.

NATIONAL ARCHIVES OF UNITED KINGDOM. **PRONOM**: background. 2009. Disponível em:< http://www.nationalarchives.gov.uk/aboutapps/pronom/default.htm>. Acesso em: 10 set. 2009.

NATIONAL LIBRARY OF NETHERLANDS. **e-Depot and Digital Preservation**. 2009a. Disponível em:< http://www.kb.nl/hrd/dd/index-en.html>. Acesso em: 22 set. 2009.

NATIONAL LIBRARY OF NETHERLANDS. **Digital preservation research.** 2009b. Disponível em:< http://www.kb.nl/hrd/dd/dd-en.html>. Acesso em: 23 set. 2009.

PDF/A Competence Center. **PDF/A**: an iso standard. 2009. Disponível em:< http://www.pdfa.org/doku.php?id=pdfa:en>. Acesso em: 20 set. 2009.

PROENÇA, A.M.P; LOPES, S.G. **Digital preservation** . Covilhã, Portugal: Universidade da Beira Interior, 200?. Disponível em:< http://www.di.ubi.pt/~api/digital_preservation.pdf>. Acesso em: 11 ago. 2009.

RAUCH, C. et al. **File-Formats for Preservation**: evaluating the Long-Term Stability of File-Formats. In: PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON ELECTRONIC PUBLISHING, 11.Vienna, Austria 13-15 June 2007. ElPub Digital Library. 2007. Disponível em:<http://elpub.scix.net/data/works/att/122_elpub2007.content.pdf>. Acesso em: 10 set. 2009.

ROCHA, R.P. **Preservação digital.** Porto alegre: UFRGS, 2009. Disponível em:< http://www6.ufrgs.br/snote/wiki/doc.php?wikipage=/ensino/notasaula/preservacao_digital>. Acesso em: 06 ago. 2009. (Nota de aula da disciplina BIB03200 - Gerenciamento Arquivístico de Documentos Eletrônicos Digitais ,do Curso de Arquivologia da Universidade Federal do Rio Grande do Sul).

ROG, J.; WIJK, C. **Evaluating file formats for long term preservation.** Netherlands, Ipress, 2007. Disponível em:< http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_file_format_evaluation_method_27022008.pdf>. Acesso em: 23 set. 2009. (National Library of the Netherlands)

ROTHENBERG, J. Metadata to support data quality and longevity. In: THE IEEE METADATA CONFERENCE, 1., 1996. Silver Spring, Maryland, US. **Proceedings …** Silver Spring, Maryland, US : IEEE, 1996. Disponível em: <http://www.computer.org/conferences/meta96/rothenberg_ Paper/ieee.data-quality.html>. Acesso em: 28 jul. 2009.

SULLIVAN, S.J. An archival/records management perspective on PDF/A. **Records Management Journal**, United Kingdom, v. 16, n. 1, p. 51-56, 2006. Disponível em: <http://www.emeraldinsight.com/0956-5698.htm>. Acesso em: 25 jul. 2009.

STANESCU, A. Assessing the durability of formats in a digital preservation environment. **OCLC Systems & Services**, United Kingdom, v. 21, n. 1, p. 61-81, 2005. Disponível em: <http://www.emeraldinsight.com/1065-075X.htm>. Acesso em: 10 ago. 2009.

_____. Assessing the durability of formats in digital preservation environment: the INFORM methodology. **D-Lib Magazine**. v.10, n.11, nov. 2004. Disponível em:< http://www.dlib.org/dlib/november04/stanescu/11stanescu.html>. Acesso em: 10 set. 2009.

THOMAZ, K. P.; SOARES, A. J. A preservação digital e o modelo de referência Open Archival Information System (OAIS). **DataGramaZero**, Rio de Janeiro, v. 5, n.1, 2004. Disponível em:< http://dgz.org.br/fev04/F_I_aut.htm>. Acesso em: 07 ago. 2009.

UNIVERSITY OF CALIFORNIA AT BERKELEY, **How Much Information ?**. 2003. Disponível em:<http://www.sims.berkeley.edu/research/projects/how-much-info-2003/index.htm>. Acesso em: 16 ago. 2009.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL. **Secretaria de Assuntos Estudantis – SAE**. Porto Alegre: UFRGS, 2009. Disponível em:< http://www.ufrgs.br/sae/>. Acesso em: 25 set. 2009.

UNIVERSITY OF LEEDS. **Survey and assessment of sources of information on file formats and software documentation**. Reino Unido, 200?. 48 p. Disponível em: <http://www.leeds.ac.uk/reprend>. Acesso em: 22 jul. 2009.

WEBB, C. **The role of preservation and the library of the future**. Canberra: National Library of Australia, 2000. Disponível em: <http://www.nla.gov.au/nla/staffpaper/cwebb9.html>. Acesso em: 28 jun. 2009.

WIKIPEDIA. **Digital rights management.** 2009. Disponível em:< http://en.wikipedia.org/wiki/Digital_rights_management>. Acesso em: 23 set. 2009.