Eduardo Alves do Valle Jr.

sob a orientação do

Prof. Arnaldo de Albuquerque Araújo

Sistemas de Informação Multimídia na Preservação de Acervos Permanentes

Dissertação apresentada ao Curso de
Pós-Graduacao em Ciência da Computação
da Universidade Federal de Minas Gerais,
como requisito parcial para obtenção do
grau de Mestre em Ciência da Computação

Universidade Federal de Minas Gerais
Departamento de Ciência da Computação
Instituto de Ciências Exatas
Belo Horizonte / 2003

* * * * * * * * * * * * * * * * *

Instrução ao encadernador:

⇨ Modelo da capa:

O encadernado deve ser obrigatoriamente:
a. capa dura (imitação comum de couro)
b. **capa preta**
c. **letras douradas**

Universidade Federal de Minas Gerais
Departamento de Ciência da Computação
Instituto de Ciências Exatas

Sistemas de Informação Multimídia na Preservação de Acervos Permanentes

Eduardo Alves do Valle Jr.

Belo Horizonte
2003

⇨ **Se** a lombada ficar larga o suficiente para incluir texto, escrever:

Eduardo A. do Valle Jr. / 2003

⇨ A margem esquerda foi propositalmente colocada em 10 mm a mais, para compensar a encadernação

⇨ **Substitua esta folha por uma cópia da folha de aprovação**
Esta página de instrução **não** deve constar no volume encadernado!

* * * * * * * * * * * * * * * * *

Sumário

Lista de Ilustrações

Lista de Equações

Lista de Tabelas

Siglas e Abreviaturas

Sigla	Significado
CD	Disco Compacto, do inglês *Compact Disc*
DAT	Fita de Áudio Digital, do inglês *Digital Audio Tape*
DCMI	*Dublin Core Metadata Initiative*
DDS	Armazenamento de Dados Digitais, do inglês *Digital Data Storage*
DLT	Fita Digital Linear, do inglês *Digital Linear Tape*
DVD	Disco Digital Versátil, do inglês *Digital Versatile Disc*
EAD	Descrição Arquivística Codificada, do inglês *Encoded Arquival Description*
GIF	Arquivo de Informações Gráficas, do inglês *Graphics Information File*
HTML	Linguagem de Marcação de Hipertexto, do inglês *Hypertext Markup Language*
ICR	Reconhecimento Inteligente de Caracteres, do inglês *Intelligent Character Recognition*
ISAAR(CPF)	Normas Internacionais para Registro de Autoridades em Arquivísitica (para Instituições, Pessoas e Famílias), do inglês *International Starndard Archival Authority Record (for Corporate bodies, Persons and Families)*
ISAD(G)	Normas Internacionais para Descrição Arquivística (Geral), do inglês *International Standard Archival Description (General)*
JBOD	Apenas um Monte de Dispositivos, do inglês *Just a Bunch of Drives*
JFIF	Formato de Arquivo para Imagem JPEG, do inglês *Jpeg File Image Format*
MARC	catalogação legível por máquina, do inglês *MAchine Readable Cataloguing*
METS	Padrão para Codificação e Transmissão de Metadados, do inglês *Metadata Encoding and Transmission Standard*
OCR	Reconhecimento Óptico de Caracteres, do inglês *Optical Character Recognition*
PDF	Formato para Documentos Portáveis, do inglês *Portable Document Format*
PNG	Gráficos Portáveis de Rede, do inglês *Portable Network Graphics*
RAID	Arranjo Redundante de Discos Baratos, do inglês *Redundant Array of Inexpensive Disks*

Sigla	Significado
SGML	Linguagem de Marcação Generalizada Padrão, do inglês *Standard Generalized Markup Language*
SLED	Único Disco Grande e Caro, do inglês *Single Large Expensive Disk*
TIFF	Formato de Arquivo Etiquetado de Imagem, do inglês *Tagged Image File Format*
W3C	Consórcio da Teia de Escala Mundial, do inglês *World Wide Web Consortium*
XML	Linguagem de Marcação Extensível, do inglês *eXtensible Markup Language*
Z39.50	protocolo para troca de informações codificadas no formato MARC entre diferentes repositórios

Dedicatória

A Memória, nos mitos dos povos ocidentais, se personaliza em figura feminina: guerreira sutil, que, no conservar, provoca as mais profundas revoluções. Este trabalho é dedicado à *femina* que reconhece o poder do transformar, preservando.

Particularizo minha admiração nestas mulheres notáveis:

- Carla da Costa, colega tão querida, mostrou-me na simplicidade do coditiano o que o professor de administração enrolava a língua, charlatando de eficientificidade (*eficiência* + *eficácia*).
- Edilane Carneiro: mulher-maravilha ou titânia? Talvez apenas possuidora de monumental sabedoria para desconfiar que o feminino de patrão não cabe na proposta simplista de patroa.
- Eliane Amorim, que me mostrou um novo significado da palavra diretor, e o poder de dobrar sem partir.
- Fernanda Canabrava: poeta, poeta, poeta! A sinergia da sua palavra com a minha palavra, o sulco no texto, nas tintas, nos gestos, lavoura fecunda, volto a ter fé!
- Fernanda Vieira, entre certezas e intuições, o tempo todo com muito balanço, a essência da informaticista, modelo para uma geração que esta custando tanto a vir!
- Tônia Lopes, trabalhar à borda do caldeirão do desejo em fervura, tomar o ser recortado em discursos, cozinhá-lo até que dali saia de novo íntegro e juvenescido? Moderna Medéia...
- Vanessa Viegas: a amiga pode surgir mágica, repentina! Nunca pensei que alma irmã se encontrasse assim, num instante ligeiro, felicidade-bolha-de-sabão. Votos de novas instâncias dessa esfera iridescente...
- Valéria Valle, o claque alto dos seus saltos altos ressoa sempre na minha mente: estímulo para saltos sempre mais altos, desejo de uma percussão permanente.

Agradecimentos

O que aqui se lê é o desdobramento do esforço de muitos, que dedicaram seu empenho profissional e pessoal para a realização do trabalho de onde se extraiu este texto.

Gostaria de agradecer ao Arquivo Público Mineiro, seus funcionários e colaboradores, pela seriedade do trabalho desenvolvido, pela coragem e sacrifício diuturno na implementação de uma iniciativa muito acima dos horizontes ordinários da administração pública, e pelo calor da acolhida que me ofereceram.

Quero agradecer às Usinas Siderúrgicas de Minas Gerais — USIMINAS que têm provido substancioso apoio financeiro a este projeto, através da Lei Estadual de Incentivo à Cultura, bem como à Companhia Brasileira de Metalurgia e Mineração — CBMM, que generosamente nos financiou através da Lei Nacional de Incentivo à Cultura. Nessa mesma linha, quero agradecer ao apoio da Microsoft do Brasil, que tem provido o *software* necessário para o sucesso deste trabalho. Quero ainda agradecer à Companhia de Processamento de Dados de Minas Gerais — PRODEMGE, e ao grupo Rede Metropolitana de Alta Velocidade — REMAV-BH2 pelo apoio técnico e colaboração administrativa.

Durante os dois anos de duração deste mestrado, pude contar com uma pequena bolsa do Conselho Nacional de Desenvolvimento Científico e Tecnológico — CNPq, que me permitiu empenhar-me em regime de dedicação exclusiva, sem graves aflições financeiras. Assim, agradeço ao apoio dessa agência. Registro também meu reconhecimento ao Projeto CNPq/PRONEX DCC/SIAM, que tem provido suporte material em mobiliário e equipamentos de informática, bem como outros patrocínios esporádicos, ao Núcleo de Processamento Digital de Imagens, grupo de trabalho ao qual estive vinculado durante todo o desenvolvimento deste trabalho.

Quero deixar registrada minha profunda gratidão e reconhecimento ao meu orientador, Prof. Arnaldo Albuquerque, que me propiciou inúmeras e diversificadas oportunidades de desenvolvimento acadêmico, e a quem devo a oportunidade deste trabalho. Se, com um pouco de dramaticidade, pode-se comparar a graduação em Ciência da Computação ao Inferno e o mestrado ao Purgatório, o Prof. Arnaldo, tem sido, sem dúvida, meu Virgílio.

Aos demais membros da banca, quero agradecer o encorajamento, e as valiosas sugestões que contribuíram para o enriquecimento deste trabalho. Ao Prof. Luis Souza agradeço o constante

esforço em me facilitar o acesso ao universo da conservação/restauração de bens culturais, e os valiosos contatos para a continuidade dos meus estudos nessa área. A Nelson Spangler, agradeço o envolvimento pessoal nesta iniciativa e a disponibilidade em facilitar o diálogo entre o Arquivo e a PRODEMGE. Ao Prof. Marcelo Bax agradeço ter me facilitado o acesso ao universo da Ciência da Informação e sua bibliografia.

Por fim, quero destacar a importância da participação e da cooperação dos meus amigos e da minha família.

Agradeço ao Gervázio (in memoriam), ao Genézio e ao Genesco pelos momentos de relaxamento e diversão, essenciais para que este trabalho pudesse ser levado a termo.

Agradeço à Ágata, ao Bruno, ao Eduardo, à Fernanda, ao Jacques, ao Marcos, ao Matias, ao Paulo, ao Ranério todos os telefonemas atendidos, todas as lamentações consoladas, todos os passeios agradaveis, todos os ouvidos alugados. E a tantos outros, tantas outras coisas...

Agradeço ao meu pai, Eduardo, à minha mãe, Valéria, e às minhas irmãs Isabela e Adriana pelo carinho, presença e pachorra do dia a dia. À Isabela e à Valéria agradeço terem tido a paciência adicional de ajudarem na correção ortografica de um texto tão técnico. E à Valéria agradeço, em especial, a substancial complementação financeira que me permitiu levar este trabalho a termo em dedicação exclusiva — generoso mecenato.

Resumo

A preservação documental é atividade chave para a análise histórica e para a construção da identidade cultural. Entretanto essa tarefa enfrenta a dificuldade de classificar e armazenar enormes massas documentais, e o severo compromisso entre as dimensões *conservação* × *acesso*. A tecnologia digital oferece a possibilidade de romper esse compromisso, fornecendo aos usuários cópias de alta qualidade, ao mesmo tempo em que preserva os artefatos originais da manipulação desnecessária. Além disso, o uso de ferramentas de gestão documental e fluxo de trabalho permite multiplicar a produtividade de arquivistas e técnicos de conservação.

Entretanto, os computadores não oferecem soluções simples para a preservação documental. A informação digital é mais sujeita à adulteração, vandalismo, e acidentes do a analógica; seus suportes são frágeis e perecíveis. Mais grave ainda: devido ao fato de os dados de computador dependerem de complexos sistemas cooperantes de *hardware* e *software*, a rápida obsolescência torna-se uma ameaça constante de perda de acesso aos acervos.

Este trabalho explora os benefícios e desafios trazidos pela aplicação de sistemas de informação multimídia aos acervos de valor permanente. Não focalizamos apenas a disponibilização dos documentos, mas também o auxílio à preservação do acervo digital, a criação de instrumentos de pesquisa eletrônicos e convencionais e o suporte ao profissional da informação em suas tarefas de indexação, classificação e reformatação do acervo.

Identificamos os requisitos e necessidades para sistemas de informação, sistemas de gestão documental, sistemas de controle do fluxo de trabalho e plataformas de representação e recuperação da informação que visem beneficiar o usuário e o profissional do universo arquivístico, explorando algumas soluções já disponíveis, tanto de modelos teóricos, quanto de sistemas já implementados e disponíveis.

Por fim, propomos nosso modelo de sistema de gestão documental para acervos permanentes, e descrevemos sua atual implementação, aplicada especificamente ao acervo de fotografias do Arquivo Público Mineiro, instituição responsável pela guarda da documentação histórica produzida pela administração executiva de Minas Gerais, bem como aquela proveniente de arquivos privados de personalidades que estejam ligados à história político-administrativa do Estado.

Résumé

La conservation des documents est une activité clé pour l'analyse historique et pour construire l'identité culturelle des nations. Cependant cette tâche fait face au défi de classifier et de garder d'énormes quantités de documents, et au compromis sévère entre la conservation et l'accès. La technologie numérique offre la possibilité de fournir des copies de haute qualité aux utilisateurs, tout en préservant les originaux de manipulations inutiles. En plus, l'usage d'outils de gestion de documents et de *workflow* peut multiplier la productivité des archivistes et des techniciens de conservation.

Les ordinateurs, cependant, n'offrent pas de solutions simples pour la préservation documentaire. L'information numérique est plus exposée à l'altération, au vandalisme, et aux accidents que l'information analogique. Les médias numériques sont fragiles et périssables. Pire encore: comme les données dépendent des systèmes complexes de logiciels et d'ordinateurs, le obsolescence rapide constitue une menace constante de perte d'accès aux collections.

Ce travail explore les profits et les défis apportés par l'application de systèmes informatiques multimédia sur les collections permanentes. Nous nous focalisons sur les systèmes qui non seulement rendent les documents disponibles, mais aussi aident à la préservation des collections numériques, créent des aides de recherche conventionnelles et électroniques, et soutiennent le professionnel dans les tâches d'indexation, de classification et de reformatage des collections.

Nous identifions les conditions requises pour les systèmes informatiques, les systèmes de gestion des documents, les systèmes de gestion de *workflow* et pour les structures de représentation et de récupération d'information, visent à aider à l'utilisateur et le professionnel des Archives. Nous explorons des modèles théoriques et des systèmes déjà implantés et disponibles.

Finalement, nous proposons notre propre modèle de système de gestion des documents pour les collections permanentes, et nous décrivons son implémentation actuelle, appliquée à la collection photographique de Arquivo Público Mineiro.

L'Arquivo Público Mineiro c'est l'institution chargée du stockage de la documentation produite par l'administration exécutive de l'Etat de Minas Gerais, et des fonds privés de personnalités en relation à l'histoire politique administratif de l'Etat.

Resumen

La conservación documental es una actividad clave para el análisis histórico y para construir la identidad cultural de las naciones. Sin embargo, esta tarea encara el desafío de clasificar y almacenar inmensas cantidades de documentos, y el severo compromiso entre la conservación y el acceso. La tecnología digital ofrece la posibilidad de romper este compromiso suministrando a los usuarios copias de alta calidad, mientras los originales estan protegidos de unaa manipulación innecesaria. Además, el uso de herramientas de administración de documentos y de *workflow* puede multiplicar la productividad de archiveros y técnicos de conservación.

Las computadoras, sin embargo, no ofrecen soluciones sencillas para la conservación documental. La información digital está más expuesta a la adulteración, al vandalismo, y a los accidentes que la información analógica. Los medios digitales son frágiles y perecederos. Aún peor: debido a la dependencia de los datos informáticos en complejos sistemas cooperarantes de *hardware* y *software*, la rápida obsolescencia llega a ser una amenaza constante de pérdida de acceso a las colecciones.

Este trabajo explora los beneficios y los desafíos generados en la aplicación de sistemas de información multimedias en colecciones permanentes. Enfocamos en sistemas que no sólo hacen los documentos disponibles, sino que permiten la conservación de la colección digital, permiten la creación de instrumentos de pesquisa convencionales y electrónicos, y ayudan al profesional en sus tareas de indexación, clasificación y reformatación de la colección.

Identificamos los requisitos para sistemas de información, sistemas de administración de documento, sistemas de administración de *workflow* y plataformas para la representación y la recuperación de la información que objetivan beneficiar al usuario y al profesional de Archivo. Exploramos ambos modelos teóricos y sistemas ya aplicados y disponibles.

Finalmente, proponemos nuestro propio modelo de sistema de administración de documentos para colecciones permanentes, y describimos su implementación actual, aplicada a la colección fotográfica de Arquivo Público Mineiro, la institución a cargo del guardia de la documentación producida por la administración ejecutiva del Estado de Minas Gerais, así como también de fondos privados de personalidades conectadas a la historia político-administrativa del Estado.

Abstract

The documental preservation is a key activity for historical analysis and for building the cultural identity of nations. However this task faces the challenge of classifying and storing huge amounts of documents, and the severe compromise between conservation and access. Digital technology offers the possibility of breaking this compromise, by supplying to the users high quality copies, while preserving the originals from unnecessary manipulation. In addition, the use of document and workflow management tools can multiply the productivity of archivists and conservation technicians.

Computers, however, do not offer simple solutions for documental preservation. Digital information is more exposed to the adulteration, vandalism, and accidents than analog information. Digital media are fragile and perishable. Even worse: because of the dependency of computer data on complex cooperating systems of hardware and software, the quick obsolescency becomes itself a constant threat of loss of access to the collections.

This work explores the benefits and challenges brought by the application of multimedia information systems on permanent collections. We focus on systems that not only make the documents available, but also aid the preservation of the digital collection, create conventional and electronic finding aids, and support the professional in his/her tasks of indexing, classification and reformatting the collection.

We identify the requirements for information systems, document management systems, workflow management systems and frameworks for representation and retrieval of information that aim to benefit the user and the professional of Archives. We explore both theoretical models and already implemented, available systems.

Finally, we propose our own model of document management system for permanent collections, and we describe its current implementation, applied to the photographic collection of Arquivo Público Mineiro, the institution in charge of the guard of the documentation produced by the executive administration of the State of Minas Gerais, as well as private fonds of personalities connected to the political-administrative history of the State.

“Une Montagne en mal d’enfant
Jetait une clameur si haute,
Que chacun au bruit accourant
Crut qu’elle accoucherait, sans faute,
D’une Cité plus grosse que Paris:
Elle accoucha d’une Souris.”

Jean de la Fontaine

1 Introdução

1.1 Preservando a memória ancestral e a contemporânea

A preservação da memória documental é uma atividade chave para a análise da história do ser humano e para a construção da identidade cultural dos povos. Entretanto, essa tarefa não se faz sem percalços: além da dificuldade em classificar e armazenar enormes massas documentais de forma sistemática, a fragilidade dos artefatos impõe um severo compromisso entre conservação e acesso.

Para garantir sua preservação, itens valiosos são guardados em arquivos seguros, disponíveis apenas para uns poucos pesquisadores. Isso, sem dúvida, é frustrante, pois quando os documentos são protegidos, mas não estão ao alcance do público, a tarefa de manter a memória viva não está sendo cumprida adequadamente. Por outro lado, a manipulação direta e constante dos originais provoca, inevitavelmente, sua degradação.

A tecnologia digital surge como uma possibilidade de romper esse compromisso, permitindo dar amplo acesso a cópias digitais de alta qualidade de determinados documentos, resguardando ao mesmo tempo os originais da manipulação desnecessária.

Mais ainda, sistemas computadorizados têm um potencial enorme de facilitar a própria tarefa de organização e descrição dos acervos. Através de ferramentas de gestão documental e fluxo de trabalho, é possível multiplicar a capacidade de trabalho de historiadores, arquivistas e técnicos de conservação. Sistemas bem engendrados permitem a esses profissionais automatizar e gerir de forma racional suas complexas atividades.

Do lado dos consulentes, o uso da tecnologia digital permite operações de busca muito mais rápidas e sofisticadas do que é possível com os instrumentos de pesquisa em papel, como inventários e índices. As redes de computadores, em particular a *Internet*, com sua *World Wide Web*, adicionam a possibilidade da consulta remota, expandindo dramaticamente o universo de pesquisadores com acesso ao acervo.

Entretanto, a tecnologia dos computadores não oferece soluções simples para o problema da preservação documental. A informação digital é muito mais sujeita à adulteração, vandalismo, e perdas acidentais do que sua contraparte analógica; e seus suportes são frágeis e perecíveis. Mais grave do que isso: devido ao fato dos dados de computador dependerem, para serem visualizados, de complexos sistemas cooperantes de *hardware* e *software*, a rápida obsolescência dos componentes torna-se uma ameaça constante de perda de acesso aos acervos. Considerando isoladamente a questão da conservação, os dados digitais oferecem dificuldades muito mais severas do que os convencionais. De fato, a maioria dos arquivistas prefere, sempre que possível, associar técnicas analógicas e digitais para garantir ao mesmo tempo a segurança de preservação das primeiras, e a facilidade de manipulação e acesso das segundas, nos chamados *sistemas híbridos de preservação*.

Infelizmente, a preservação dos dados digitais se tornou uma necessidade, e não um luxo, uma vez que um volume cada vez maior de documentos é criado em computadores. Para muitos desses documentos, como programas aplicativos, jogos de *video-game*, documentos em hipermídia, é impossível encontrar uma representação analógica satisfatória: eles precisam ser preservados em seus formatos originais. Mesmo documentos mais convencionais, como planilhas eletrônicas e textos formatados, embora passíveis de serem impressos em papel ou microfilme, perdem muitas de suas características essenciais ao serem arrancados do universo digital.

A não ser que esforços sejam rapidamente coordenados para endereçar a preservação dos artefatos digitais, as gerações futuras estarão inexoravelmente deserdadas de um grande percentual da memória documental contemporânea.

1.2 Objetivos

Nosso objetivo foi explorar os benefícios e desafios trazidos pela aplicação de sistemas de informação multimídia às instituições que mantêm custódia de acervos de valor permanente. Nosso foco não é apenas a disponibilização de documentos nos mais diversos formatos (textos, manuscritos, mapas, fotografias, trechos de áudio, filmes…), mas também o

auxílio à preservação do acervo digital, a criação de instrumentos de pesquisa eletrônicos e convencionais e o suporte ao profissional da informação em suas tarefas de indexação, classificação e reformatação do acervo.

O que fazemos aqui é identificar os requisitos e necessidades para sistemas de informação, sistemas de gestão documental, sistemas de controle do fluxo de trabalho e plataformas de representação e recuperação da informação que visem beneficiar o usuário e o profissional do universo arquivístico.

Muitos estudos têm sido conduzidos a respeito de cada um desses aspectos, em suas feições isoladas, mas estava por fazer uma análise mais aprofundada de como essas diferentes tecnologias e metodologias podem se articular e serem utilizadas conjuntamente para favorecer a preservação dos acervos e o acesso a eles.

Exploramos algumas soluções já disponíveis, tanto de modelos teóricos, quanto de sistemas já implementados e disponíveis para o tratamento e disponibilização de acervos digitais. Por fim, propomos nosso modelo de sistema de gestão documental para acervos permanentes, e descrevemos sua atual implementação, aplicada ao acervo de fotografias do Arquivo Público Mineiro.

O Arquivo Público Mineiro é a instituição responsável pela guarda da documentação permanente produzida pela administração executiva estadual, bem como aquela proveniente de arquivos privados de personalidades que estejam ligados à história político-administrativa do Estado. Encontram-se sob sua custódia fundos de documentos da administração da província desde o Brasil colônia até o século XX. Seu acervo fotográfico, composto por dezenas de milhares de itens, está sendo disponibilizado na *World Wide Web* através do sistema que implementamos.

1.3 Acervos correntes × Acervos permanentes

Em arquivística identificam-se os conceitos de arquivos correntes e arquivos permanentes, correspondendo às duas fases bem distintas da vida dos documentos [SCHELLENBERG 73] [DUCHEIN 86].

A fase corrente corresponde à etapa de criação, modificação e tramitação do documento — i.e., aquela em que ele se encontra em produção. Nela, o documento é peça participante dos processos desempenhados por uma instituição, registrando informações relevantes produzidas no desenvolvimento de suas tarefas. Nos arquivos correntes, os documentos estão sujeitos às modificações, e também ao descarte, uma vez que tenham cumprido seu papel.

O objetivo principal de um sistema de arquivo corrente é fazer com que os documentos sirvam às finalidades para as quais foram criados, da forma mais eficiente e econômica possível. Os documentos, ao final de suas vidas úteis, recebem uma destinação: que pode ser a destruição imediata, a transferência para um depósito de armazenamento temporário, ou a passagem para um arquivo de preservação em caráter permanente. A decisão do destino final dos documentos segue normalmente uma tabela de temporalidade, ou agenda de destinação[1] [SCHELLENBERG 73].

A fase permanente corresponde à etapa de recolhimento dos documentos cuja temporalidade determinou que eles não deveriam ser descartados, por serem importantes registros, seja para fins legais e comprobatórios, seja para fins culturais. Num arquivo permanente, os documentos estão disponíveis apenas para consulta, não estando mais sujeitos à modificação ou descarte.

Sistemas de gestão documental para arquivos correntes vêm sendo estudados e aplicados desde a década de 1970. Entretanto, a aplicação desses sistemas nos arquivos permanentes é relativamente recente e requer um esforço de adequação para atender aos requisitos específicos dessa atividade:

- Longevidade digital, que tratamos nos Capítulos 3 e 4.4.
- Acesso e recuperação de informação em grandes acervos, que tratamos no Capítulo 4.
- Uso extensivo de metadados que tratamos no Capítulo 6.
- Gestão do fluxo de trabalho para as atividades do acervo, que tratamos no Capítulo 6.

[1] Do inglês: *disposal schedule*

1.4 Estrutura da dissertação

Este trabalho foi estruturado da seguinte forma:

- No Capítulo 2 fazemos uma contextualização da dissertação, analisando diversos trabalhos relacionados ao nosso: experiências de aplicação de sistemas de informação em acervos de arquivos e bibliotecas, bem como alguns sistemas de informação para gestão de acervos digitais atualmente disponíveis.
- No Capítulo 3 delineamos as duas principais situações em que a tecnologia digital interage com a preservação de acervos — a preservação de documentos originalmente digitais e a digitalização de documentos analógicos para fins de preservação. Nesse capítulo condensamos as principais recomendações encontradas na literatura, e apresentamos as diretrizes que foram utilizadas em nosso trabalho no Arquivo Público Mineiro, segundo o resultado de nossos experimentos.
- O Capítulo 4 analisa como a questão do acesso permeia todos os esforços de preservação no universo digital, e discute diferentes mecanismos que permitem garantir esse acesso.
- O Capítulo 4.4 discute as questões de preservação do acervo digital, relacionadas tanto à confiabilidade dos suportes, quanto à necessidade de se combater os efeitos da obsolescência dos mecanismos de visualização dos dados.
- O Capítulo 6 estabelece os conceitos de sistemas de gestão documental e de fluxo de trabalho para acervos permanentes. Nesse capítulo descrevemos a análise e a modelagem do sistema que concebemos, e apresentamos alguns aspectos de sua implementação.
- No Capítulo 7 apresentamos as conclusões deste trabalho.
- Finalizamos o texto com uma lista completa das referências bibliográficas.

2 Contexto e Trabalhos Relacionados

Neste capítulo procuramos contextualizar nosso trabalho, relacionando-o com os que lhe são afins em um de dois aspectos: ou aqueles que aplicam sistemas de informação multimídia em acervos históricos, arquivísticos e de bibliotecas, ou aqueles que propõem modelos e implementações desses sistemas.

Na apresentação dos sistemas de informação já disponíveis, não fazemos uma exposição exaustiva — preferimos escolher três exemplos que consideramos arquetípicos: o *AdLib Toolkit*, um conjunto de sistemas comerciais desenvolvido para instituições de guarda pela *AdLib Information Systems* [ADLIB 03]; o *Greenstone*, uma plataforma gratuita e de código aberto desenvolvida pelo Projeto Neozelandês de Bibliotecas Digitais para distribuição de conteúdo digital em sistemas eletrônicos [GREENSTONE 03]; e o *GNU EPrints*, um sistema também gratuito e de código aberto desenvolvido pelo Departamento de Eletrônica e Ciência da Computação da Universidade de Southampton para criação de arquivos distribuídos de materiais eletrônicos, visando principalmente o auto-arquivamento de trabalhos científicos e acadêmicos [EPRINTS 03].

2.1 Precursores

Araújo descreve um sistema para codificação e armazenamento de imagens históricas que inaugura a linha de pesquisa na qual se apoia o corrente trabalho [ARAÚJO 92] [ARAÚJO 93].

Andrade estuda os sistemas de informação multimídia e implementa um protótipo de um sistema para dar acesso às informações de um dos fundos do Arquivo Público Mineiro. Sua implementação propõe o uso de sistemas de bancos de dados orientados por objeto, e a

linguagem *Java*, da *Sun Microsystems*. A arquitetura de documentos do acervo do Arquivo é modelada através de um diagrama de classes. O trabalho também se ocupa de modelar os mecanismos de acesso e navegação ao sistema [ANDRADE 98a] [ANDRADE 98b].

O sucesso desse protótipo foi um importante estímulo para o desenvolvimento do presente trabalho — procuramos estender a visão orientada por objetos aplicada à modelagem do arranjo arquivístico, à gestão de documentos e do fluxo de trabalho. Se o trabalho anterior explorou as potencialidades dos sistemas de informação na dinamização da consulta ao acervo, o trabalho presente tencionou endereçar às questões concernentes à preservação de longo prazo da informação digital, e sua correlação com a organização do acervo e o acesso.

2.2 Instituições e consórcios aplicando sistemas digitais a acervos

As facilidades trazidas pela tecnologia digital têm estimulado instituições das mais diversas naturezas — arquivos, museus, bibliotecas e consórcios de distribuição da informação — a disponibilizarem seus acervos em formato digital, quer se trate de esforços de reformatação de acervos convencionais, quer se trate da organização de documentos criados desde o início em formato digital.

A quase totalidade das iniciativas de implantação de sistemas de informação em acervos o faz pelas vantagens trazidas ao acesso, e não por questões de preservação. De fato, as dificuldades de preservação da informação digital, com algumas exceções, não são endereçadas, ou são encaradas com perigoso otimismo:

> "Também as pessoas, por analogia com documentos de hoje que se tornaram ilegíveis em processadores de texto ou periféricos obsoletos, se preocupam se o código digital, mesmo se preservado será passível de ser acessado e inteligível.
>
> "A resposta é novamente a probabilidade. A razão pela qual o texto impresso tem sido fielmente preservado através das gerações (quando ele tem sido) é que o interesse coletivo dos literatos é revertido na garantia dessa preservação. Essa mesma continuidade de interesses coletivos existirá para o código digital, pela mesma razão, porém **o código digital será muito mais fácil de migrar para cada nova tecnologia sucessiva do que o texto impresso** jamais o foi em cada construção ou regime sucessivo." [EPRINTS 02b] (grifo nosso, original em inglês)

Por várias razões, que exploramos detalhadamente nas Seções 3.1 e 5.2, discordamos fortemente desta visão otimista acerca da preservação da informação digital, e conosco se alinham a maioria dos estudos aprofundados a respeito do tema [BESSER 00] [CONWAY 97] [LAWRENCE 00] [ROTHENBERG 98] [TFADI 96] [UNESCO 00].

Uma dificuldade na análise dos estudos de caso da implantação de sistemas de informação é a evolução extremamente rápida da tecnologia. Suposições razoáveis a respeito do estado-da-arte da tecnologia, e observações sobre o custo, desempenho, capacidade operacional e velocidade dos equipamentos perdem a validade no universo de quatro ou cinco anos. Assim, é inútil mimetizar as decisões técnicas de um projeto de acervos digitais com pouco mais do que alguns meses de antecedência.

Entretanto, uma análise dos pressupostos racionais que levaram a essas decisões, do embasamento teórico que gerou as medidas práticas, das preocupações relacionadas à natureza da informação, e à natureza do relacionamento desta com seus consulentes, fornece informações valiosas para o desenvolvimento de novos projetos de digitalização. Em outras palavras, embora as premissas e as conclusões das decisões técnicas de um projeto envelheçam rapidamente, a argumentação e o raciocínio usados para levar as premissas até as conclusões permanecem importantes para projetos futuros.

2.2.1 Arquivos públicos e privados

Arquivos e consórcios de arquivos têm se utilizado de sistemas de informação para dar acesso aos seus instrumentos de pesquisa e a versões digitalizadas de seus acervos.

Documentos de arquivos têm peculiaridades especiais, no tocante à sua custódia, organização e tratamento informacional, que se refletem na forma como suas versões digitalizadas são tratadas. Em particular, a organização de documentos arquivísticos é fortemente baseada no *princípio de respeito aos fundos*, que estabelece que as coleções devem ser organizadas de acordo com sua proveniência, e nunca desmembradas e misturadas com outras coleções [DUCHEIN 86].

Para permitir que os documentos relevantes sejam encontrados nas coleções custodiadas, são criados *instrumentos de pesquisa*. Esforços internacionais foram feitos para normalizar a descrição arquivística e a escrita dos instrumentos de pesquisa, gerando os

padrões ISAD(G)[1] [CAND 94] e ISAAR(CPF)[2] [CAND 95]. O primeiro prescreve os princípios de descrição dos fundos de arquivos, enquanto o segundo normaliza o registro de autoridades.

Para permitir o compartilhamento eletrônico dessas descrições, uma convenção ainda mais detalhada foi criada: o EAD[3], que, basicamente, define como uma descrição ISAD(G) pode ser codificada em um documento XML[4] padronizado [TSAA 00a] [TSAA 00b].

2.2.1.1 *The Making of America* e *The Making of America II*

Uma cooperação entre a Universidade de Michigan e a Universidade de Cornell, resultou no projeto *Making of America* (MOA), uma coleção que inclui mais de 1,5 milhões de páginas de texto digitalizadas, disponíveis na *World Wide Web*. Os documentos, relacionados à história norte-americana do século XIX, foram digitalizados, passaram por um processo de reconhecimento de texto utilizando a tecnologia OCR[5], e em seguida foram codificados em SGML[6], segundo as normas do *Text Encoding Initiative* [TEI 02]. O texto codificado foi utilizado para prover facilidades de busca em texto livre, mas uma vez localizado o recurso desejado, é oferecida ao usuário a página digitalizada, que preserva a aparência original do documento [SHAW 97].

O experimento do MOA levou ao projeto *The Making of America II* (MOA2), sob coordenação da Universidade da Califórnia em Berkeley, com a participação das Universidades de Cornell, Stanford, Penn State e da Biblioteca Pública de Nova Iorque. O projeto focalizou-se em coleções acerca do tema de transportes durante a época da Corrida do Ouro. O objetivo foi a criação de versões digitalizadas das fontes primárias, de forma a permitir o acesso direto a essas cópias, independentemente da instituição de custódia. O projeto baseou-se num complexo conjunto de metadados, tanto descritivos (aqueles relacionados à identificação do conteúdo dos artefatos), quanto estruturais (aqueles

[1] Normas Internacionais para Descrição Arquivística (Geral), do inglês *International Standard Archival Description (General)*

[2] Normas Internacionais para Registro de Autoridades em Arquivísitica (para Instituições, Pessoas e Famílias), do inglês *International Starndard Archival Authority Record (for Corporate bodies, Persons and Families)*

[3] Descrição Arquivística Codificada, do inglês *Encoded Arquival Description*

[4] Linguagem de Marcação Extensível, do inglês, *eXtensible Markup Language*

[5] Reconhecimento Óptico de Caracteres, do inglês, *Optical Character Recognition*

[6] Linguagem de Marcação Generalizada Padrão, do inglês, *Standard Generalized Markup Language*

relacionados ao formato, comportamento e interligação dos documentos), e administrativos (aqueles relacionados ao fluxo de trabalho, proveniência, direitos de acesso, etc…) [CLIR 98].

Ao incluir aspectos comportamentais nos metadados, o projeto procura capturar aspectos dos objetos digitais freqüentemente ignorados nas iniciativas de reformatação de acervos. De fato, o MOA2 propõem um Modelo de Serviços para Bibliotecas Digitais fundado nos conceitos de orientação por objeto herdados da Ciência da Computação [HURLEY 99]. O modelo do MOA2 é baseado em três camadas:

a) **Objetos da biblioteca digital:** contendo instrumentos de pesquisa e os objetos digitalizados (incluindo conteúdo, metadados, comportamento e métodos).

b) **Ferramentas:** visualizadores para os objetos digitalizados, catálogos *on-line*, bancos de dados para administração dos instrumentos de pesquisa.

c) **Serviços:** recuperação, visualização, e navegação dos materiais arquivados.

O projeto MOA2 prevê diversas iniciativas interessantes. Uma delas é a possibilidade de diferentes interfaces de serviços para diferentes tipos de usuários. Assim, pesquisadores acadêmicos podem utilizar serviços sofisticados e precisos, baseados em instrumentos de pesquisa eletrônicos, enquanto estudantes pré-universitários podem usar facilidades menos exatas, porém mais simples de entender e usar. Outra é o uso do EAD para representação dos instrumentos de pesquisa eletrônicos. Por fim, um conjunto de medidas visa especificamente a longevidade digital, incluindo metadados previstos para essa finalidade.

2.2.1.2 The Minnesota Historical Society

A Sociedade Histórica de Minnesota (MHS) é uma instituição privada, cultural, sem fins lucrativos, estabelecida em 1849 para preservar e divulgar a história do Estado de Minnesota — E.U.A. Seu acervo tem um caráter misto de biblioteca, coleções museológicas e fundos arquivísticos. O MHS disponibiliza versões eletrônicas de vários artefatos sob sua custódia, bem como instrumentos de pesquisa, codificados utilizando o *EAD* [MHS 03].

A MHS tem uma iniciativa de proteção de registros eletrônicos. O Arquivo do Estado (ligado à MHS) publica um volume orientando os funcionários do governo a lidarem com registros eletrônicos, visando sua preservação, incluindo imagens digitais, e correio eletrônico [MSA 03]. Há uma preocupação de motivar os funcionários públicos a seguirem

essa iniciativa, orientando-os em detalhes, e alertando-os para a importância das medidas, inclusive em seus aspectos legais, como se depreende deste trecho, extraído da página do Arquivo:

> "Você rotineiramente cria, usa e gerencia informação eletrônica em seu trabalho diário, já que você usa computadores para enviar *e-mail*, criar planilhas eletrônicas, publicar páginas da *web*, gerenciar bancos de dados e criar outros materiais eletrônicos. Porque você trabalha para uma agência governamental, as leis de Minnesota e da federação o obrigam a tratar essas informações como registros oficiais do governo.
>
> "Você provavelmente já tem uma estratégia para gerenciar seus registros em papel. Clique aqui para aprender como desenvolver uma estratégia para gerenciar registros eletrônicos." [MSA 02a] (original em inglês)

Há também uma preocupação acerca da normalização do uso de metadados para permitir a identificação, preservação e o acesso aos registros governamentais, independentemente de seus formatos nativos [MSA 02b].

2.2.1.3 Projeto Nabuco

O Projeto Joaquim Nabuco teve por objetivo o "desenvolvimento de um ambiente para aquisição, filtragem, compressão, consulta e impressão de imagens de documentos de valor não só textual, mas também iconográfico" [PJN 00]. O projeto envolveu uma cooperação entre a Fundação Joaquim Nabuco e o Departamento de Informática da Universidade Federal do Paraná, para disponibilizar os documentos de um único fundo privado, com o legado documental do importante político recifense.

O Projeto Nabuco permite o acesso a versões digitalizadas de fotografias, cartas, e escritos de Nabuco. Uma interface de busca por palavras-chave é oferecida.

Há uma grande preocupação com o aspecto da compressão de imagens, e reconhecimento de texto e outros tipos de processamento digital de imagens. Um dos estudos do projeto propõem um método para comparação de diferentes produtos de OCR, aqui endereçado por nós na Seção 3.2.3. Questões importantes, contudo, como a longevidade do acervo digital, ou o acesso aos metadados não recebem atenção [ROSA 94].

2.2.1.4 Arquivo Público do Paraná

O Arquivo Público do Paraná tem sido um dos pioneiros na conversão dos seus instrumentos de pesquisa para o formato EAD, e na divulgação de parte do seu acervo em formato eletrônico. Atualmente, o Arquivo, disponibiliza através do seu *website* o seu guia de

fundos, instrumentos de pesquisas detalhados para parte da coleção, e alguns documentos digitalizados [APP 03].

O *website* não traz considerações de longevidade digital.

2.2.2 Bibliotecas e diretórios de informação

Bibliotecas têm trabalhado com catálogos eletrônicos desde o final da década de 1960, com a criação do MARC — catalogação legível por máquina[1]. Com a criação do protocolo Z39.50, no final da década de 1980, as bibliotecas passaram a poder compartilhar eletronicamente seus catálogos codificados segundo o MARC. Esse compartilhamento levou instituições cooperantes a criarem os chamados diretórios ou portais de recursos — grandes catálogos onde o usuário podia localizar recursos do seu interesse em múltiplos repositórios, simultaneamente.

Com a evolução da tecnologia muitas bibliotecas e portais passaram a oferecer, além dos catálogos, versões digitalizadas de elementos do seu acervo, especialmente versões esgotadas ou raras de livros e documentos. Com o aumento do volume de publicações em formato originalmente eletrônico, cresceu também o interesse por diretórios que disponibilizassem esse tipo de artefato.

2.2.2.1 Biblioteca do Congresso Norte-americana

Nos Estados Unidos, a prestigiosa Biblioteca do Congresso é ao mesmo tempo um grande fomentador e usuário de sistemas de informação aplicados ao tratamento de acervos [TLOC 03a].

A Biblioteca oferece diferentes facilidades de pesquisa para diferentes tipos de usuários — acadêmicos, crianças em idade escolar, colegiais e interessados em cultura norte-americana e universal. Um portal no protocolo Z39.50 permite que outras instituições façam consulta ao catálogo eletrônico, que abrange a totalidade do acervo. Materiais das mais diversas naturezas: livros, registros sonoros, *videotapes* podem ser localizados utilizando essas interfaces.

De sua gigantesca coleção a Biblioteca oferece apenas uma pequena fração em formato digital. Ainda assim, são mais de 7,5 milhões de itens digitais, amostrados de mais de

[1] Do inglês: *MAchine Readable Cataloguing*

100 coleções históricas. A Biblioteca conta ainda com um programa para a distribuição de materiais acerca de lei e legislação em formato eletrônico, um programa de reformatação para meio digital e distribuição ao público de materiais sob riscos de conservação e um programa para preservação de materiais originalmente eletrônicos [TLOC 03b].

A Biblioteca é ainda fonte de referência para vários padrões importantes em bibliotecas digitais, como o MARC, o Z39.50, o EAD e o METS[1], todos concernentes à coleta e compartilhamento de metadados.

2.2.2.2 Biblioteca Nacional Brasileira

A Fundação Biblioteca Nacional oferece uma interface de busca ao seu catálogo, disponibilizando ao usuário diferentes formas de consulta. Os catálogos eletrônicos da Biblioteca estão em formato MARC [BN 03].

A biblioteca oferece parte do seu acervo em formato digital. Ao efetuar a busca, o usuário pode especificar se está interessado em todo o acervo, ou apenas nos documentos que já possuam versões digitalizadas. Exposições temáticas virtuais permitem um acesso contextualizado a subconjuntos de objetos digitais. Os metadados dos objetos digitais também obedecem ao formato MARC.

O *website* não traz considerações de longevidade digital.

2.2.2.3 *Open Archives Initiative* e *EPrints*

Repositórios de trabalhos acadêmicos e artigos científicos, por causa da penetração do uso de computadores entre seus produtores e leitores, têm sido constantemente considerados projetos pilotos ideais para a implantação de bibliotecas digitais.

O *Open Archives Initiative* (OAI) adotou, juntamente com o *EPrints*, uma iniciativa de desenvolver e promover padrões para facilitar a difusão de conteúdo. O OAI criou um padrão para a implementação de bibliotecas digitais distribuídas, formadas por componentes independentes e cooperantes [OAI 03].

O OAI se baseia em dois tipos de participantes: os provedores de dados e os provedores de serviços. Os primeiros disponibilizam seus dados e metadados, utilizando

[1] Padrão para Codificação e Transmissão de Metadados, do inglês *Metadata Encoding and Transmission Standard*

protocolos e formatos padronizados. Os segundos fornecem facilidades para o uso desses dados, normalmente ferramentas de recuperação e visualização.

O *EPrints*, um dos parceiros do OAI tem promovido o conceito de auto-arquivamento, estimulando acadêmicos e cientistas a armazenarem, em arquivos abertos, disponíveis para o público, documentos relativos às diferentes etapas de sua produção [EPRINTS 02a].

Tanto o OAI quanto o *EPrints* concentram-se na padronização dos metadados de descrição. Metadados estruturais e a questão da longevidade digital são apenas tangencialmente endereçados.

Atualmente, o OAI passa por uma transição. Embora o projeto continue fortemente centrado na disseminação de materiais acadêmicos e científicos, há o desejo de fomentar a aplicação de sua tecnologia em outros tipos de conteúdo digital [OAI 03].

2.3 Sistemas de informação para acervos digitais

Já existem vários sistemas de gestão de documentos multimídia disponíveis. Embora a grande maioria desses sistemas seja voltada para documentos em fase corrente, alguns já se propõem (ou ao menos se prestam) ao tratamento de documentos permanentes, notadamente às aplicações de bibliotecas digitais.

2.3.1 AdLib Toolkit

A *AdLib Information Systems* desenvolve e distribui o produto chamado *AdLib Toolkit*, uma plataforma para o auxílio a bibliotecas digitais e convencionais. O produto conta com diferentes versões, adaptadas especificamente para bibliotecas, museus ou arquivos.

As ferramentas do *AdLib Toolkit* permitem a gestão integrada não apenas do acervo, mas também de outros processos ligados à instituição como o fluxo de consulentes, empréstimo e devolução de documentos físicos, empréstimo entre instituições, gestão de permissões de acesso e direitos autorais, etc… O *AdLib* suporta os principais protocolos

envolvidos no universo da catalogação eletrônica: o Z39.50 e o MARC para bibliotecas e o ISAG(G), o ISAAR(CPF) e o EAD para arquivos. A indexação permite o uso de tesauros e arquivos de autoridades.

Embora permita o armazenamento de imagens digitais, e portanto a disponibilização do acervo em formato digital, o forte do sistema está na manutenção de catálogos eletrônicos. Não há suporte específico para fluxo de trabalho de reformatação de documentos. O *AdLib* foi concebido principalmente para instituições que possuem acervos convencionais e desejam disponibilizar em forma digital seus catálogos, e possivelmente, uma parte de seus artefatos [ADLIB 03].

O *AdLib Toolkit* está disponível para a plataforma *Microsoft Windows*.

2.3.2 Greenstone

O Projeto Neozelandês de Bibliotecas Digitais, da Universidade de Waikato, desenvolve e distribui, em cooperação com a *UNESCO* e com a *Human Info NGO* uma plataforma gratuita e de código aberto para a geração de bibliotecas digitais denominada *Greenstone* [GREENSTONE 03]. O *Greenstone* permite a busca baseada em texto-livre, e a navegação baseada em metadados. Nesse produto, as informações são armazenadas em seus formatos nativos, que podem incluir textos em formatos HTML, PDF, *PostScript*, *Microsoft Word* e muitos outros. Novos formatos podem ser reconhecidos pelo sistema através da adição de *plug-ins*, que traduzem o conteúdo de formatos proprietários para o formato nativo XML do *Greenstone*. Uma vez incorporados a uma coleção, os documentos passam a gozar das facilidades de pesquisa e navegação providas pelo sistema [WITTEN 01].

As primeiras versões do *Greenstone* confiavam na busca em texto-livre como a única forma de recuperação da informação, não oferecendo facilidades mais sofisticadas. Contudo, as versões atuais permitem que a busca seja feita considerando partes dos documentos (apenas o título, apenas o corpo documento, apenas em certos metadados, etc…) permitindo uma recuperação mais precisa das informações. O sistema atual também oferece suporte à navegação, através de estruturas criadas automaticamente pela extração dos metadados [WITTEN 01].

O *Greenstone* não oferece suporte ao controle do fluxo de trabalho, embora seus criadores tenham escrito um documento sugerindo um fluxo de trabalho para reformatação de acervos visando o uso da plataforma [LOOTS 02].

O *Greenstone* está disponível tanto para as plataformas *Microsoft Windows* quanto para *Unix*/*Linux*.

2.3.3 GNU EPrints

O Departamento de Eletrônica e Ciência da Computação da Universidade de Southampton desenvolveu um sistema para arquivamento de documentos eletrônicos gratuito e de código aberto denominado *GNU EPrints*. O projeto é relacionado ao *EPrints* e ao *Open Archives Initiative*.

Embora o *GNU EPrints* tenha sido inicialmente concebido para o armazenamento de documentos acadêmicos e científicos, ele pode ser configurado para o arquivamento de documentos de qualquer natureza, em qualquer formato: uma arquitetura de documentos simplificada permite definir os metadados relevantes para cada tipo de documento. O sistema suporta uma arquitetura extensível e configurável de assuntos, que pode ser usada para a navegação através do acervo e nas pesquisas. As buscas no *GNU EPrints* são baseadas no conteúdo dos metadados [EPRINTS 03].

O grande diferencial do *GNU EPrints* está em sua capacidade de agregar o arquivo à *Open Archives Initiative*, permitindo o acesso da comunidade global às informações armazenadas [OAI 03].

O *GNU EPrints* está disponível para sistemas baseados em *Unix*/*Linux*.

3 Tecnologia Digital e a Preservação de Acervos

Existem duas grandes aplicações para sistemas de informação na preservação de acervos. A primeira é a preservação de artefatos de origem digital, que apresenta grandes desafios, e é de crescente importância, por causa do volume crescente de produção técnica, intelectual e artística que vem sendo criado com o uso de computadores.

A segunda é o uso da digitalização como uma técnica de reformatação, aplicada à preservação do conteúdo intelectual de artefatos originalmente analógicos. Quase sempre, essas iniciativas são motivadas pelas facilidades de acesso — e não de preservação — dos dados digitais: sua perfeita replicabilidade, a facilidade de recuperação e transmissão, a possibilidade de gerar cópias de alta qualidade. Outras vezes, contudo, a digitalização é utilizada como meio, ou pelo menos como coadjuvante, na preservação do conteúdo de documentos particularmente difíceis de serem tratados por meios convencionais, como fotografias coloridas e registros de áudio.

Neste capítulo discutimos detalhadamente a natureza da informação digital, enquanto registro documental, e os motivos de sua preservação ser tão problemática. Os meios para promover essa preservação — o refrescamento dos suportes, a emulação de plataformas e a migração entre formatos de arquivos — são discutidos na Seção 5.2.

Quanto à reformatação, endereçamos a digitalização de imagens, tanto reunindo as principais recomendações da literatura, quanto apresentando nossas próprias diretrizes, e os experimentos que as motivaram.

3.1 Documentos de origem digital

O computador digital é uma invenção recente, datando da década de 1940, e seu uso na administração governamental e comercial é ainda mais atual - começa na década de 1960 e se populariza na década de 1980. Contudo, a penetração da tecnologia digital se deu com tal ímpeto que, desde sua introdução, vêm-se assistindo a um crescimento exponencial de praticamente todos os indicadores a ela relacionados — número de computadores, número de usuários, número de computadores ligados em rede, e volume de documentos produzidos em forma digital [CONLAN 88].

De acordo com dados do *United States Census Bureau*, 51,0% dos lares norte-americanos tinham um computador e 41,5% deles estavam ligados à *Internet* em 2000. Dados de 1997 apontam que 49,8% dos empregados estadunidenses utilizavam computadores em suas tarefas de trabalho — um percentual que provavelmente já aumentou [USCB 00]. A maior parte dos trabalhos de impressão e escrita de textos é feita com computadores, a gravação profissional de som é quase toda digital, câmeras fotográficas digitais estão substituindo as baseadas em filmes e mesmo a gravação de vídeo caminha inexoravelmente no mesmo sentido.

O fato de a tecnologia digital ser ao mesmo tempo tão recente, tão ubíqua e estar evoluindo com tanta rapidez traz severas conseqüências para a preservação dos documentos que são criados sob sua égide.

Um incidente, já famoso, envolveu os dados do censo norte-americano de 1960, que precisaram ser recuperados em 1976 — o que foi quase impossível, já que tais dados residiam num tipo de fita que só podia ser lido por um dispositivo então há muito obsoleto. O *Census Bureau* enfrentou grandes dificuldades para recuperar esses dados, e embora o tenha conseguido em grande extensão, o evento serviu de alarma para os riscos de preservação da memória documental contemporânea [TFADI 96].

Menos bem sucedido foi o Arquivo do Estado de Nova Iorque, que em meados da década de 1980, tentou resgatar um grande volume de informações gerado ao final da década de 1960 acerca do uso das terras do Estado. Embora o Arquivo tenha obtido uma cópia das fitas com os dados, os programas necessários para interpretá-los não haviam sido preservados, e as dificuldades em recuperar as informações das fitas, sem eles, se mostraram insuperáveis.

Hoje, os únicos testemunhos sobreviventes do trabalho são listagens de dados impressas, que estavam sob a custódia dos Arquivos da Universidade de Cornell [TFADI 96].

Uma força-tarefa criada pela *Comission on Preservation and Access* e o *Research Libraries Group* identificou que a preservação dos dados digitais está sujeita aos seguintes fatores: [TFADI 96]

a) **Conteúdo:** no mundo digital, a preservação do conteúdo é complexa e aparece em diferentes níveis de abstração. Sem dúvida é necessário preservar a cadeia de zeros e uns que compõe toda e qualquer informação digital, mas não é suficiente: é preciso também preservar os mecanismos que tornam essa seqüência inteligível para seres humanos.

b) **Imutabilidade:** é preciso preservar com segurança a autenticidade dos documentos, i.e., garantir que não haja possibilidade de adulteração ou supressão dos dados. Caso contrário, o valor comprobatório do registro digital fica nulo, e mesmo seu valor cultural é grandemente diminuído.

c) **Referência:** é preciso criar mecanismos consistentes de referência aos dados, de forma que eles possam ser associados a identificadores, e recuperados através desses identificadores.

d) **Proveniência:** toda a arquivística moderna é centrada no princípio da proveniência, que estabelece que parte da integridade de um documento reside em conhecer sua origem. Para preservar a integridade da informação digital, deve ser possível registrar a sua origem e a sua cadeia de custódia.

e) **Contexto:** para que a informação permaneça íntegra é preciso preservar o seu contexto. No caso da informação digital, há um contexto técnico — o *hardware*, *software*: a tecnologia de que eles dependem; mas também um contexto de interdependência informacional, uma vez que no mundo digital, freqüentemente os itens documentais fazem referências uns aos outros. Há ainda que se considerar o contexto comportamental daquelas informações, a forma como o usuário se relacionava com elas, que é dependente da tecnologia disponível no momento em que as informações foram criadas.

Os dados digitais são afetados pela fragilidade de seus suportes, um problema que endereçamos especificamente na Seção 5.1, mas muito mais severos são os problemas provocados pela obsolescência da tecnologia que permite manipulá-los. As políticas das instituições e da indústria de informática, bem como certos comportamentos dos usuários também não facilitam a tarefa de preservação digital. Besser caracteriza esses problemas da seguinte forma: [BESSER 00]

a) **Visualização:** informação digital criada no passado requer a manutenção de uma infraestrutura e de uma base de conhecimento para ser visualizada. Não basta preservar os dados digitais gerados por um processador de texto, por exemplo, sem preservar o *software* necessário para visualizá-lo, e o *hardware* necessário para executar o *software*. Ao menos é preciso reter o conhecimento acerca da codificação do arquivo, para que se possa interpretar seu conteúdo.

b) **Embaralhamento:** práticas comuns para resolver problemas de curto prazo no uso da informação digital acabam resultando em problemas de preservação. Dois exemplos são a compressão de dados e a criptografia. A compressão de dados adiciona uma camada de complexidade à interpretação dos dados digitais, impedindo que esses dados sejam interpretados a não ser que o método de descompressão seja conhecido, ou o *software* de descompressão esteja disponível. A criptografia apresenta ainda mais problemas, uma vez que mesmo que os métodos sejam conhecidos, pode ser impossível resgatar as mensagens sem as chaves, e é fácil que essas sejam acidentalmente perdidas, no decorrer da vida dos dados.

c) **Inter-relação:** no mundo digital as informações estão cada vez mais inter-relacionadas, através de recurso como a incorporação e hiperligação. Páginas da *web*, por exemplo, mesmo que sejam exibidas de forma unificada, são compostas por diversos arquivos, que às vezes residem em computadores muito distantes entre si. Está se tornando progressivamente mais difícil delimitar os itens de informação e mesmo identificar seus contextos.

d) **Custódia:** embora as organizações tradicionalmente tenham desenvolvido a preocupação de preservar e manter vários tipos de material analógico, esse cuidado não se estendeu ainda aos dados digitais. Por isso, a maior parte do

material produzido digitalmente não é atribuída a responsáveis por sua custódia e provavelmente não estará disponível para as futuras gerações.

e) **Tradução:** quando o conteúdo é traduzido para novos formatos, freqüentemente a mudança de forma provoca uma certa mudança no conteúdo. Sucessivas migrações de formato em arquivos digitais provocam constantes oportunidades para erros e imprecisões de tradução. Esforços de emulação nem sempre conseguem capturar todos os aspectos do ambiente simulado. Para que a operação de salvaguardar um arquivo seja fiel, é preciso não apenas armazenar seu conteúdo, mas permitir que seu comportamento junto ao usuário ou consulente, e sua forma de interação com este sejam preservados.

Dentre outras coisas, os documentos digitais não sobrevivem sem uma estratégia constante de proteção aos seus mecanismos de armazenamento e visualização, uma vez que esses estão sujeitos a se tornarem indisponíveis devido à rápida obsolescência. Os recursos disponíveis para efetuar essa proteção — refrescamento, emulação e migração — são discutidos na Seção 5.2.

3.2 Documentos de origem analógica

Dada a enorme dificuldade em obter longevidade digital, parece contraproducente sequer pensar em utilizar a digitalização para reformatação de preservação [WEBER 97].

Entretanto certos benefícios tornam os sistemas digitais tão atraentes, que compensam muitas vezes o risco de sua utilização como meios, ou ao menos como coadjuvantes, na preservação de acervos.

O primeiro, e talvez mais importante desses benefícios, é a perfeita replicabilidade dos dados digitais: uma cópia de um artefato digital é um clone, indistinguível do seu original, ao contrário do que acontece com os dados analógicos, em que as cópias são sempre imperfeitas, sujeitas a ruídos e atenuações.

O segundo fator é a facilidade de manipulação e análise dos dados digitais, por causa de sua natureza numérica. Imagens digitais, por exemplo, permitem uma vasta gama de operações de realce, restauração, filtragens, ampliações e correções que só estão disponíveis para fotografias analógicas através de morosas e delicadas intervenções de laboratório.

Não se pode negar que, no mínimo, a tecnologia digital pode propiciar um excelente acesso ao conteúdo intelectual dos artefatos digitalizados, resguardando os originais da manipulação desnecessária, e retardando sua deterioração.

Em muitos casos, porém, seu papel vai muito além de simplesmente romper o compromisso entre preservação e acesso. Freqüentemente, os resultados obtidos na análise dos dados digitais se tornam cruciais para o processo de conservação e restauração dos artefatos originais [LAHANIER 02]. Em outros casos, a natureza dos artefatos a serem preservados é tão complexa, que a tecnologia digital pode oferecer uma forma mais conveniente de representar seu conteúdo intelectual — como na preservação de sítios arqueológicos.

Quando sistemas de informação são utilizados em esforços de reformatação de acervos, além de todas as preocupações relativas à preservação dos dados digitais — que abordamos na Seção 3.1 — acresce-se a preocupação de que o próprio processo de digitalização seja feito de forma a capturar o mais fielmente possível os aspectos de interesse dos artefatos.

Em nosso trabalho iremos nos limitar a explorar a reformatação de imagens bidimensionais, a tarefa de reformatação mais amplamente utilizada para a tecnologia digital, e que se aplica a uma vasta gama de acervos documentais, desde livros tipografados, até documentos manuscritos, recortes, fotografias, slides e microfilmes.

Um grande número de estudos tem sido conduzido a respeito da digitalização de imagens, seus formatos de armazenamento e sua conversão para texto — com fins de preservação de acervos.

O Conselho em Bibliotecas e Recursos de Informação editou um conjunto de guias para a qualidade em projetos de imagem digital. Esses guias provêem diretrizes para todo o processo de captura do acervo, desde o planejamento da digitalização, passando pela seleção do *scanner*, até a mensuração da qualidade e o formato de armazenamento para as matrizes digitais [CLIR 00].

3.2.1 A natureza das imagens digitais

A forma mais convencional de representar e armazenar imagens em sistemas digitais é através das chamadas *imagens de varredura*[1]. Essas imagens são formadas por uma grade bidimensional de amostras de valores de tonalidade, para imagens coloridas, ou luminosidade, para imagens monocromáticas. Cada uma dessas amostras é chamada de *pixel* ou *pel*, uma abreviação para *elemento da figura*[2]. (Figura 3.1)

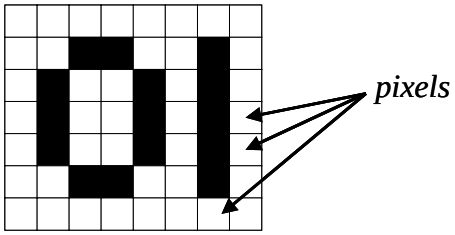


Figura 3.1: Uma imagem de varredura é uma grade de amostras de tonalidades; as amostras são chamadas *pixels*

A correspondência entre as dimensões de cada amostra da imagem digital e as dimensões da imagem representada no mundo real fornece a *resolução* da imagem. A resolução pode ser expressa como a dimensão das amostras (e.g., *pixels* de 250 µm) ou mais comumente pelo número de amostras por unidade linear, como 300 *pixels por polegada* ou 300 dpi[3]. No universo das imagens analógicas, especialmente na fotografia a micrografia, a resolução é freqüentemente expressa em *pares de linha por milímetro*. Como cada par de linhas equivale a dois *pixels* e uma polegada tem 25,4 mm, a conversão entre as duas unidades é simples (Equação 3.1).

$$1\,{}^{lp}\!/_{mm} = 50{,}8\ dpi$$

Equação 3.1: Conversão das medidas de resolução em lp/mm para dpi

Para imagens monocromáticas, cada amostra representa um valor numérico de luminosidade. Este valor representa uma escala numérica que vai desde o preto até o branco, passando pelos valores intermediários. A capacidade de representação tonal da imagem é determinada pelo número de *bits* das amostras, às vezes chamado de *profundidade de bit*. No caso limítrofe, em que cada amostra é composta por apenas um *bit*, só é possível representar

[1] Do inglês: *raster images*
[2] Do inglês: *PICTure ELement*
[3] Do inglês: *Dots Per Inch*

duas luminosidades: a presença e a ausência da luz, formando as chamadas imagens bitonais. À medida que se aumenta o número de *bits* por amostra, cresce a capacidade de representar luminosidades intermediárias; valores típicos são 4, 8, 10, 12 e 16 *bits* por amostra, que permitem representar, respectivamente 16, 256, 1024, 4096 e 65.536 níveis de luminosidade diferentes.

Para imagens coloridas é preciso que cada amostra contenha a representação não apenas da luminosidade, mas da tonalidade. A informação de cor está ligada ao espectro da luz, i.e., à distribuição dos diferentes comprimentos de onda combinados num raio luminoso. Se fosse necessário que cada uma das amostras descrevesse exatamente o espectro da luz coletado por ela, a representação de cor se tornaria inexeqüível nos sistemas digitais, mas felizmente, o sistema visual humano simplifica bastante a percepção da cor, dividindo-a em três faixas de comprimento de onda: mais curtas, correspondendo ao azul, mais longas, correspondendo ao vermelho e intermediárias, correspondendo ao verde.

Assim para efeitos de percepção humana, a complexa informação espectral só precisa ser descrita através de três valores referentes a estas *cores primárias*: verde, vermelho e azul. Isso faz com que uma imagem colorida seja equivalente, grosso modo, a três imagens monocromáticas superpostas. De fato, para uma dada profundidade de *bit*, uma imagem colorida será três vezes maior que a equivalente monocromática.

3.2.2 Considerações de uso das imagens

A primeira consideração a fazer na determinação dos parâmetros de qualidade das imagens digitais é o seu propósito. Essa questão afeta todas as decisões subseqüentes, desde a resolução até o formato de armazenamento.

Em ordem crescente de necessidade de fidelidade ao original, as imagens digitais podem ter o propósito de:

a) **Índice**: são imagens de pequeno tamanho e baixa qualidade, utilizadas como miniaturas em resultados de pesquisa ou em listagens impressas. Seu propósito é apenas a rápida identificação do conteúdo geral do documento, para propósitos de busca e identificação.

b) **Acesso**: são imagens de média qualidade, que permitam uma visualização razoavelmente detalhada do conteúdo do documento. Seu propósito é reduzir o acesso ao documento original, mas não substituí-lo.

c) **Reprodução**: são imagens de alta qualidade, que permitam a duplicação do documento, em impressora ou outro equipamento de imagem. Seu propósito é capturar a aparência do documento original em grau suficiente para permitir uma reprodução satisfatória, tão próxima do original quanto possível.

d) **Preservação**: são imagens da mais alta qualidade possível, que procuram capturar todos os detalhes de interesse do documento, não apenas para as aplicações atuais, mas para possíveis aplicações futuras.

Uma segunda dimensão que precisa ser analisada é o tipo de análise que se espera fazer com a imagem digital. Se a imagem for preparada apenas para visualização por olhos humanos, os padrões de qualidade podem ser mais relaxados, inclusive com a aplicação de métodos de compressão mais agressivos. Entretanto, se a imagem for utilizada para mensurações e análises quantitativas a serem geradas por computador para aplicações de reconhecimento de caracteres, ou em aplicações de visão computacional, os requisitos de qualidade podem ser mais estritos.

Ao se tomar as decisões de qualidade das imagens, é recomendado ser o mais conservador possível no sentido da alta qualidade. Isto porque se deve levar em consideração não apenas os usos presentes, mas os possíveis usos futuros das imagens. Assim, uma imagem de acesso de baixa resolução pode se tornar insatisfatória com o aumento do nível de exigência dos usuários à medida que eles têm acesso a novos sistemas de alta qualidade. Uma imagem de reprodução concebida para as atuais impressoras coloridas de 300 dpi terá resultados sub-ótimos em um sistema de imagem de 600 dpi.

O custo mais expressivo no processo de reformatação de uma coleção consiste no esforço de remoção e preparação dos materiais para serem reformatados e sua posterior devolução aos seus depósitos permanentes. De fato, acessar cada página ou item a ser convertido é o maior dispêndio desses esforços. Assim, uma certa extrapolação da qualidade de escaneamento evita incorrer no ônus inaceitável de passar por sucessivos processos de digitalização, à medida que surgem novas aplicações para a imagem digital.

Quando for necessário, por razões de acesso, disponibilizar para os usuários arquivos de menor tamanho, é possível gerar esses arquivos a partir dos registros de alta qualidade, utilizando-se programas para a redução de resolução e compressão de dados. O contrário, entretanto, não é possível — nenhum procedimento algorítmico é capaz de aumentar a resolução real de uma imagem, e os métodos de compressão de imagens mais eficazes são irreversíveis.

Assim, um esquema freqüentemente adotado na reformatação de acervos é o escaneamento na mais alta qualidade possível, visando a geração de um máster, ou matriz digital de preservação. Desses arquivos, freqüentemente grandes demais para serem acessados diretamente pelo usuário ou mesmo para residirem no sistema da informação *on-line*, são derivadas, automaticamente, as cópias para as finalidades de índice, acesso e reprodução. As matrizes são então transferidas para mídias terciárias (fitas magnéticas ou discos ópticos), onde ficam armazenadas até que seja necessário utilizá-las novamente, por exemplo, para gerar novas imagens de acesso com maior qualidade.

3.2.3 Conversão de imagens para texto — *OCR* e *ICR*

Freqüentemente, após a captura da imagem de uma página textual, o usuário se interessa por convertê-la em texto codificado em caracteres. As vantagens do segundo formato são inúmeras: o texto pode ser copiado e colado em editores, e podem ser realizadas buscas por determinadas palavras ou expressões.

Existem duas tecnologias de conversão de texto em imagens para texto codificado em caracteres. A primeira, chamada de OCR — reconhecimento óptico de caracteres[1], é normalmente utilizada para texto tipografado ou impresso em alta qualidade, em que o tipo dos caracteres é bastante legível e regular. O OCR utiliza métodos mais expressos e convencionais de reconhecimento da forma das letras. Uma segunda técnica chamada de ICR — reconhecimento inteligente de caracteres[2], é utilizada para textos mais problemáticos como impressos matriciais, tipografias antigas, dactilografia e até mesmo manuscritos.

Quando essas tecnologias são utilizadas para fins de apresentação do texto convertido, taxas de reconhecimento muito elevadas são requeridas. De fato, estudos apontam

[1] Do inglês: *Optical Character Recognition*
[2] Do inglês: *Intelligent Character Recognition*

que quando o número de erros é superior a 4 ou 5 por 1000 caracteres, uma taxa tão elevada quanto 99,5%, é mais econômico processar o texto manualmente do que reconhecê-lo e depois corrigir os erros.

Mello e Lins estabelecem uma comparação do desempenho de ferramentas comerciais de OCR aplicados a documentos históricos digitalizados em diferentes resoluções e tons de cinza [MELLO 99].

3.2.4 Formatos de armazenamento e compressão

Uma vez adquirida a imagem digital, é preciso escolher o formato de arquivo usado para armazená-la.

Para fins de preservação, essa decisão é muito importante. Deve-se utilizar um formato amplamente compatível, bem documentado, de preferência definido por um consórcio aberto (em contraposição aos formatos proprietários de empresas), que preserve fielmente as informações coletadas na aquisição, e ainda permita o acréscimo de alguns metadados à imagem digital — no mínimo os parâmetros utilizados na sua digitalização, e uma identificação do documento original. Na prática, entretanto, nenhum formato atualmente disponível atende plenamente a esses objetivos.

Um formato que se aproxima dessa proposta é o TIFF, formato de arquivo etiquetado de imagem[1]. O TIFF permite armazenar a imagem digital preservando toda a integridade dos dados. A documentação do formato é propriedade da *Adobe Systems*, mas o formato em si é amplamente documentado e a *Adobe* divulga livremente a especificação técnica [ADOBE 92]. O formato conta também com uma vasta comunidade de usuários, que mantêm e trocam informações a seu respeito [RITTER 97]. O TIFF possui o conceito de *etiquetas*, que são metadados que podem ser acrescentados ao cabeçalho do arquivo, com informações de identificação da imagem, e detalhadas descrições de seu formato digital.

Entretanto, por ser tão flexível, o TIFF tornou-se um formato complexo. Isso faz com que ele apresente problemas de compatibilidade, uma vez que nem todos os aplicativos dão suporte a todos os detalhes de sua especificação. É comum dois aplicativos não conseguirem compartilhar seus arquivos, mesmo quando ambos usam o TIFF. Tentativas de

[1] Do inglês: *Tagged Image File Format*

normalização, como o TIFF/IT, tiveram o efeito contrário de ampliar as variações encontráveis para o formato.

O TIFF permanece, entretanto, como um dos principais formatos de escolha para preservação digital. Sugere-se utilizar com parcimônia os recursos opcionais mais avançados do formato, de forma a evitar problemas de compatibilidade.

Uma alternativa pode ser o formato PNG — gráficos portáveis de rede[1]. O PNG foi criado como um sucessor do GIF, um formato limitado e que, por usar um método de compressão patenteado, começou a ter seu uso desencorajado na *Internet* [ROELOFS 03]. As especificações do PNG foram publicadas pelo W3C e admitem menos variações de implementação que o TIFF, tornando o formato menos sujeito a incompatibilidades [LILLEY 03]. O PNG oferece a possibilidade de anotação de metadados textuais, bem como características interessantes como correção de gama embutida e suporte aos principais espaços de cor dependentes e independentes de dispositivo.

O formato PNG é um formato de compressão sem perdas. Isso significa que um algoritmo reversível é utilizado para reduzir o tamanho da imagem armazenada, preservando-a sem nenhuma alteração visível ou invisível. Entretanto, mesmo a compactação sem perdas pode afetar a longevidade digital, dificultando eventuais tarefas de resgate de arquivos obsoletos.

O formato JFIF — formato de arquivo para imagem JPEG[2], utiliza compressão com perdas para reduzir dramaticamente o tamanho das informações. Aproveitando-se de características do sistema visual humano, a compressão JPEG descarta parte da informação da imagem de forma a tornar o arquivo propício à compactação e em seguida aplica métodos que lhe permitem reduzir o tamanho do arquivo de 10 a 50 vezes, e às vezes mais [JPEG 03].

O grau de compressão determina a qualidade resultante do arquivo. Para reduções de até 10 vezes, as variações costumam ser imperceptíveis. O fato, entretanto, de o JPEG ser um formato com perdas faz com que sua compressão não seja reversível: uma vez que um arquivo foi gravado em formato JPEG, a imagem original jamais pode ser reconstruída, exatamente, a partir dele. Além disso, a cada vez que o arquivo é salvo, novas perdas

[1] Do inglês: *Portable Network Graphics*
[2] Do inglês: *Jpeg File Image Format*

acontecem, fazendo com que o JPEG seja um formato inadequado para arquivos que precisem sofrer vários ciclos de retoque ou alteração.

Embora o JFIF seja uma escolha pobre para as matrizes digitais, ele deve ser considerado para as imagens de acesso ou reprodução, por causa da sua grande capacidade de compressão, mantendo a aparência da imagem.

3.2.5 Sistemas híbridos de microfilmagem e digitalização

O fato de os dados digitais estarem sujeitos a tantos problemas de longevidade faz com que a digitalização não seja uma forma plenamente confiável de reformatação para conservação. Entretanto, a associação da microfilmagem com o uso da tecnologia digital tem se mostrado extremamente frutífera para resolver os problemas de preservação e acesso aos acervos documentais [WILLIS 97].

O microfilme é um meio seguro de preservação. Se exposto, processado e armazenado em condições adequadas, sua longevidade esperada chega a 500 anos. Além disso, os filmes e câmeras usados para acervos permanentes permitem atingir resoluções tão elevadas quanto 150 ou 200 lp/mm (7.500 a 10.000 dpi), valores inimagináveis para os atuais sistemas digitais. Entretanto, o filme só pode ser lido de forma seqüencial, em equipamentos de leitura caros e desconfortáveis.

A associação das duas tecnologias permite endereçar ao mesmo tempo preservação e acesso, gerando tanto a imagem de preservação em filme, quanto a de acesso, em meio digital. Isso pode ser feito de três formas: [BECK 97]

a) **Digitalizar o material e imprimir a imagem digital em microfilme**: esta abordagem permite que as imagens sejam tratadas adequadamente pelo sistema digital, antes da geração do filme, permitindo que páginas que contenham ao mesmo tempo regiões de alto e baixo contraste sejam produzidas de maneira legível no filme resultante. Além disso, a imagem digital pode ser colorida, enquanto o filme de preservação descarta a informação de cor. Entretanto as imagens ficam limitadas tanto pelo sistema de aquisição (*scanner*) quanto pela impressora de microfilmes. Atualmente, o estado-da-arte dessa tecnologia permite atingir apenas cerca de 40 ou 50 lp/mm, o que é considerado pouco para fins de preservação.

b) Digitalizar e microfilmar os materiais simultaneamente: esta abordagem permite que as imagens sejam geradas em paralelo, utilizando câmeras especiais que fazem ao mesmo tempo o registro em microfilme e em formato digital. Isso apresenta algumas vantagens: a geração de um microfilme de altíssima resolução, acompanhado de uma imagem digital que preserve a informação de cor, por exemplo. Esta tecnologia, entretanto, não está plenamente desenvolvida para propósitos de acervos permanentes, e os equipamentos que hoje estão disponíveis têm qualidade suficiente apenas para a microfilmagem comercial.

c) Microfilmar o material e digitalizar o microfilme: esta é quase sempre a opção de escolha, pois permite gerar um filme de preservação de alta resolução, que marcado com sinais especiais — códigos de barras ou *blips*, pode depois ser digitalizado de forma semi-automatizada. O problema dessa abordagem é que toda a informação de cor é descartada, e mesmo a gama tonal da imagem fica dramaticamente reduzida, quando se utilizam os filmes de alto contraste, mais comuns nas tarefas de microfilmagem. Essa questão pode ser minimizada com o uso de filmes de preservação de tons contínuos[1], que preservam melhor a gama tonal.

Um esforço de sistema híbrido de preservação não consiste simplesmente em microfilmar o material, de forma convencional, e depois submeter os filmes para a digitalização. Na verdade o processo envolve mudanças desde ao fluxo de trabalho até às decisões técnicas de processamento do filme [WATERS 97].

A microfilmagem feita como processo isolado procura ajustar contraste e exposição dos filmes de forma a facilitar sua leitura direta. Além disso, nem sempre são utilizados recursos de marcação, como *blips*, pois se assume que o filme será lido por um usuário humano, que avançará os quadros manualmente.

A microfilmagem para posterior digitalização, entretanto, deve se concentrar em capturar o documento o mais fielmente possível, sem procurar melhorar sua legibilidade — isso será papel da imagem digital. O uso de técnicas de marcação se torna quase obrigatório, para permitir a digitalização semi-automática dos filmes, em lotes — sem a marcação o

[1] Conhecidos como CFT — *Continuous Tone Filming*

usuário precisa enquadrar manualmente os fotogramas no *scanner* de microfilmes, um processo moroso e sujeito a erros. O aspecto mais importante, talvez, seja a preocupação de indexar o material *durante a microfilmagem*, registrando em um banco de dados a correspondência entre as páginas de documentos e os fotogramas do filme. Esse passo irá permitir que, durante a digitalização os fotogramas sejam associados, automaticamente aos documentos que se lhes correspondem.

3.2.6 Decidindo os parâmetros de qualidade

Nesta seção apresentamos uma série de experimentos, bem como considerações teóricas, para guiar a determinação dos parâmetros de qualidade na digitalização de imagens.

Começamos por criar uma página de texto que pudesse ser utilizada na avaliação quantitativa e qualitativa dos diferentes parâmetros de aquisição da imagem digital. Em seguida, geramos diversas versões desta página com diferentes graus de qualidade, correspondendo a várias condições típicas de documentação encontráveis em acervos textuais. Acrescentamos a este conjunto artificialmente criado, as imagens de alguns manuscritos em diferentes condições de conservação, selecionados como representativos do acervo do Arquivo Público Mineiro. Adicionamos ainda imagens de um livro tipografado, com extensos comentários escritos à mão. Por fim, incluímos a imagem de um cartão de teste de resolução. O conjunto total de imagens é descrito na **Tabela 3.1**.

Tabela 3.1: Documentos utilizados nos experimentos de digitalização

Doc.	Descrição	Características
0	Cartão de resolução *A-2289 CAT 800 7403 Rev.2/00* da *Kodak*	Alta resolução, alto contraste
1	Página concebida para o experimento, impressa a 600 dpi em impresora a *laser* sobre papel branco	Alta resolução, alto contraste, contraste regular
2	Página concebida para o experimento, impressa a 600 dpi em impresora a *laser* sobre papel timbrado (com marca d'água)	Alta resolução, alto contraste, variações de contraste
3	Página concebida para o experimento, impressa a 300 dpi em impresora a jato de tinta sobre papel branco	Média resolução, alto contraste, contraste regular
4	Página concebida para o experimento, impressa a 300 dpi em impresora a jato de tinta sobre papel branco. Depois de impressa, a página foi imersa em álcool por alguns minutos.	Média resolução, alto contraste, contraste regular, leve borramento
5	Tiramos uma cópia xerográfica do Documento 1, desta cópia tiramos uma outra, resultando no Documento 5.	Baixa resolução, alto contraste, contraste irregular, letras erodidas
6	Imergimos uma versão do Documento 5 em uma solução de corante, para manchá-lo irregularmente.	Baixa resolução, médio contraste, contraste irregular, letras erodidas
7	Microfilme de manuscrito antigo (Século XVIII) em bom estado de conservação	Manuscrito, alto contraste, contraste regular
8	Microfilme de manuscrito antigo (Século XVIII) com letras erodidas	Manuscrito, médio contraste, contraste irregular, letras erodidas
9	Microfilme de manuscrito antigo (Século XVIII) esmaecido, papel manchado	Manuscrito, baixo contraste, contraste irregular, letras erodidas
10	Microfilme de manuscrito antigo (Século XVIII) com migração de tinta do verso	Manuscrito, auto contraste, contraste irregular, migração de tinta
11	Livro tipografado sob papel jornal, com adição de notas de margem manuscritas e marcações	Tipografia de baixa qualidade com notas manuscritas

3.2.6.1 Resolução

Freqüentemente, a primeira decisão a tomar em um projeto digitalização de acervos é a resolução adequada para a captura do material. Quanto maior a resolução, maior a qualidade da imagem digital, contudo, maior também o tamanho dos arquivos. De fato, devido à natureza bidimensional das imagens, quando a resolução dobra, o tamanho do arquivo resultante quadruplica. Assim, dados os custos de armazenamento, procura-se utilizar a menor resolução que atenda com segurança aos propósitos da imagem digital.

Para documentos iconográficos (fotografias, mapas, etc…), em que o valor visual do documento é essencial para a sua integridade, é desejável que as matrizes digitais capturem o máximo de informação, se possível, todos os detalhes do documento original. Para isso, é preciso utilizar uma resolução de escaneamento igual ou superior à do material original.

A resolução de um negativo ou *slide* fotográfico é determinada por fatores complexos, que incluem o tipo do filme, as qualidades ópticas e mecânicas da câmera, e os cuidados e decisões no processamento do filme. Os melhores filmes fotográficos coloridos atualmente disponíveis atingem uma resolução de 100 lp/mm, filmes preto-e-branco podem

atingir enormes valores de resolução, até 600 lp/mm em condições especiais. Entretanto, considerando o conjunto do sistema de captura fotográfica, com seus múltiplos componentes ópticos e mecânicos, esses níveis de qualidade podem ser considerados utopias teóricas. Valores mais típicos situam-se em torno de 20 a 80 lp/mm (1000 a 4000 dpi) [DIGITAL 02] [PUTS 01]. Na prática, isso implica em tamanhos de arquivos bastante elevados, como mostra a **Tabela 3.2**, embora muitos *scanners* de filmes disponíveis atualmente ofereçam resoluções dessa ordem.

Tabela 3.2: Tamanho da matriz digital para preservação de negativos e *slides*

Resolução do Original	Tamanho da Matriz Digital* (em MB) para cada Formato de Filme							
	Peq. Formato	Médio Formato				Formatos de Estúdio		
	35 mm	6 x 4,5 cm	6 x 6 cm	6 x 7 cm	6 x 9 cm	5 x 4"	5 x 7"	10 x 8"
20 lp/mm	4,0	12	16	19	25	59	103	236
60 lp/mm	36	111	148	173	222	532	930	2.126
80 lp/mm	63	198	264	308	396	945	1.654	3.780

* O tamanho se refere a um arquivo não comprimido, em cores, com 8 *bits* por cor.

Quando a digitalização é feita através da cópia fotográfica em papel, deve-se considerar que resoluções menores são necessárias. De fato, a resolução de uma cópia fotográfica será sempre inversamente proporcional à sua ampliação, e às vezes menor ainda devido à óptica do ampliador. Considere uma fotografia no formato mais popular atualmente, 10×15 cm, gerada a partir de um negativo de 35 mm (24×36 mm), cuja resolução seja de 50 lp/mm. A resolução da cópia, considerando um ampliador perfeito, será dada pela Equação 3.2.

$$\frac{\text{Resolução Original}}{\text{Ampliação}} = \frac{\text{Resolução Original}}{\text{Dimensão da Cópia} \,/\, \text{Dimensão do Original}}$$

$$\frac{50\ {}^{lp}/_{mm}}{10\ cm \,/\, 24\ mm} = \frac{50\ {}^{lp}/_{mm}}{100\ mm \,/\, 24\ mm} = \frac{50\ {}^{lp}/_{mm}}{4{,}2} = 12\ {}^{lp}/_{mm} = 610\ dpi$$

Equação 3.2: Calculando a resolução necessária para a digitalização de preservação de uma cópia fotográfica

Fizemos um experimento de digitalizar uma fotografia de 10 × 15 cm, ampliada de um negativo colorido, de 35 mm, com sensibilidade ISO 400. A resolução adequada para digitalizar esse filme, que tem resolução entre 20 e 40 lp/mm, extraindo toda a informação útil, estaria entre 1000 e 2000 dpi. Com a digitalização feita a partir da ampliação, a resolução

adequada ficaria em torno de 240 e 480 dpi. De fato, como se pode observar na Figura 3.2, o escaneamento a 600 dpi já começa a revelar o grão do filme, e em 1200 dpi a granularidade é bastante acentuada.

(a)

(b)

(c)

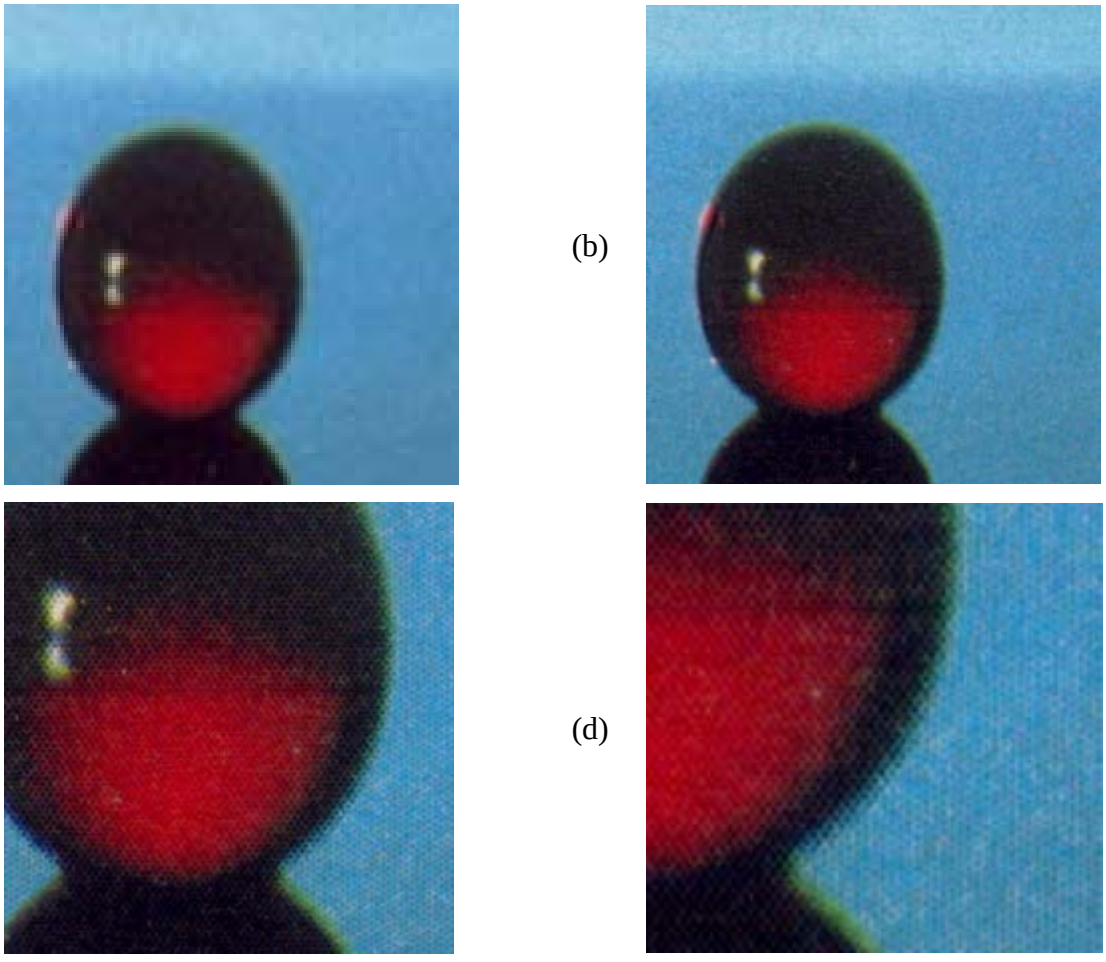
Figura 3.2: Exame de imagens digitais geradas a partir do escaneamento de uma ampliação de 15x10 cm de um negativo de 35 mm. Escaneamentos em: *a)* 300 dpi, *b* 600 dpi, *c)* 1200 dpi

Imagens produzidas em tipográficas ou *offset*, como pôsters, gravuras de livro, utilizam *meios-tons* na reprodução da gama tonal. A imagem não é formada por tons contínuos, como numa fotografia, mas decomposta em minúsculos pontos de cor pura, cujas dimensões ou densidade irão promover a sensação de cores intermediárias, quando o

documento é visto à distância normal de leitura. A resolução da imagem equivale à da retícula utilizada para gerar o padrão de meios-tons, que é medida em linhas por polegada ou lpi[1]. Jornais utilizam retículas de cerca de 85 lpi, um valor tão baixo, que os pequenos pontos da imagem são visíveis até mesmo a olho nu. Imagens de alta qualidade, utilizadas em publicações, podem atingir até 200 lpi.

A regra simples para fazer o escaneamento de imagens impressas em meios-tons é usar o dobro da resolução da retícula. Assim, por exemplo, para um original gerado em retícula de 80 lpi, deve usar um escaneamento de 160 dpi.

(a) (b)



(c) (d)

Figura 3.3: Exame de imagens digitais geradas a partir do escaneamento de uma imagem de *offset* impressa em tela de 150 lpp. Escaneamentos em: *a)* 150 dpi, *b)* 300 dpi, *c)* 600 dpi; *d)* 1200 dpi

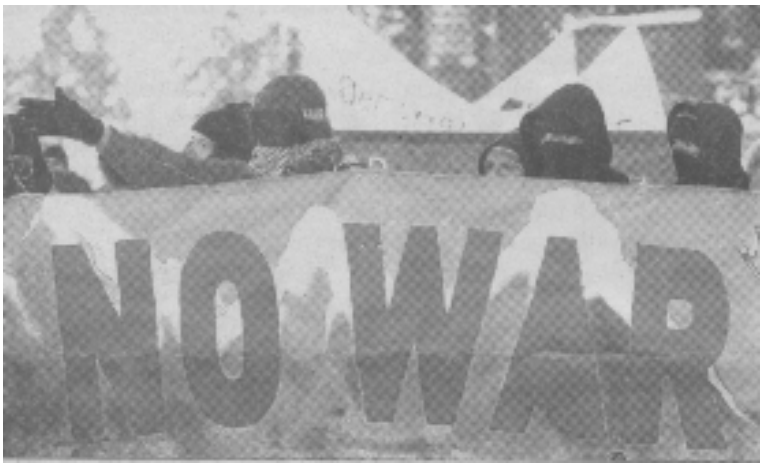
Fizemos um experimento de digitalizar uma impressão colorida de *offset*, gerada em tela de fotocomposição de 150 1pi. A resolução adequada para digitalizar este original seria de 300 dpi. De fato, como se observa na Figura 3.3, há uma melhoria de nitidez da

[1] Do inglês: *Lines Per Inch*

imagem no aumento de 150 para 300 dpi, mas aumentos de resolução subseqüentes simplesmente ampliam a granularidade da imagem.

Na digitalização de imagens de *offset*, um problema que às vezes surge é a formação de padrões de interferência entre o padrão da retícula de meios-tons e a grade de amostragem de digitalização. Esse efeito é chamado *Moiré* e produz uma texturização bastante desagradável na imagem resultante. O problema pode ser amenizado no pós-processamento da imagem, mas não removido completamente. Às vezes uma ligeira rotação do original ou alteração na resolução de escaneamento resolve o problema [VAKULENKO 00], mas a única forma garantida de eliminá-lo é fazer o escaneamento em uma resolução pelo menos duas vezes superior à da retícula. (Figura 3.4)

(a)



(c)

Figura 3.4: O escaneamento de imagens impressas em meios-tons pode gerar o aparecimento dos padrões de *Moiré*. Um aumento de resolução irá eliminar o problema. Escaneamentos em: *a)* 75 dpi, *b)* 150 dpi

As imagens de acesso e reprodução, quer sejam diretamente digitalizadas, quer sejam produzidas pela compressão das matrizes digitais, têm requisitos de resolução menos estritos do que essas últimas.

Ao se determinar a resolução utilizada nessas imagens, considera-se mais a capacidade do sistema visual humano do que a natureza dos artefatos. A resolução da visão humana é medida em unidades angulares, não em unidades lineares, revelando o fato óbvio de que a capacidade de distinguir um par de linhas depende da distância desse par de linhas. A resolução de visão não excede um minuto de grau, o que equivale, para um objeto a 25 cm (a distância normal de leitura), a cerca de 8 lp/mm ou 400 dpi.

Contudo, se o usuário deseja ter a possibilidade de ampliar a imagem, sem perdas de resolução (como se ele estivesse utilizando uma lupa para observar o documento original) a digitalização deve levar isso em consideração. A fórmula da resolução de aquisição é dada pela Equação 3.3. Assim, supondo um original de 10×15 cm, que se deseja permitir ao usuário imprimir nas dimensões de 15×22 cm em uma impressora de 300 dpi, deve-se escaneá-lo a 450 dpi.

$$\text{Resolução Final Desejada} \times \text{Ampliação}$$

$$300\ dpi \times \frac{15\ cm}{10\ cm} = 450\ dpi$$

Equação 3.3: Determinando a resolução das imagens de acesso

Ao determinar a resolução de imagens de acesso e reprodução, considere os seguintes fatores:

- Se o objetivo for a visualização da imagem em um monitor de vídeo, a resolução desses dispositivos não ultrapassa 150 dpi (valores mais típicos estando em torno de 90 dpi). Assim, uma imagem de 450 dpi ultrapassa em três vezes a resolução necessária (desconsiderando a ampliação).
- Se não estiver salvando as matrizes de digitalização, extrapole, dentro das possibilidades do projeto, a qualidade das imagens de acesso e reprodução. Isso irá garantir que elas não se tornem o fator limitante de qualidade, à medida que a tecnologia de monitores e impressoras avança.
- Não faz sentido utilizar uma resolução de escaneamento superior à de preservação, mesmo que se deseje possibilitar grandes ampliações, em uma impressora de altíssima resolução. O resultado final sempre será limitado pela resolução do original.

Para documentos de natureza textual, a perfeita reprodução da aparência visual não é tão importante quanto a legibilidade. Kenney e Chapmam adaptam um método para medição da qualidade em microfilmes —o Índice de Qualidade (IQ) — para o estabelecimento de requisitos de resolução em imagens de texto [KENNEY 97]. Segundo eles:

> "As normas técnicas mais exigentes foram desenvolvidas para a indústria de micrográficos e se baseiam no método do Índice de Qualidade. De fato, os procedimentos de controle de qualidade para inspeção de microfilme e o método do IQ para descrever a legibilidade de texto são bem adequados — com certas modificações — para predizer e avaliar o desempenho de sistemas digitais de reprodução de imagens. (…) Contanto que uma câmera ou um *scanner* estejam operando em seus níveis ótimos, o IQ pode ser utilizado para predizer os níveis de qualidade marginal (3,6), médio (5,0) ou alto (8,0) — que serão consistentemente alcançados na cópia de uso." [KENNEY 97]

O método do Índice de Qualidade é muito simples e relaciona a altura do menor caractere de interesse no texto (normalmente a altura da letra "e" minúscula) e a resolução do documento. Ele pressupõe que se o menor caractere do texto for representado por número razoável de pontos, o texto como um todo será legível. No IQ tradicional, para sistemas micrográficos, a altura é medida em milímetros, e a resolução em pares de linha por milímetro. Para se adaptar o sistema para a tecnologia digital é preciso converter a fórmula para a resolução em pontos por polegada. A Equação 3.4 mostra as fórmulas.

$$\text{(a)} \quad IQ = h \times p \qquad \text{(b)} \quad IQ = h \times (r \div 50{,}8)$$

Equação 3.4: A fórmula do índice de qualidade para resolução (a) em lp/mm e (b) em pontos por polegada (dpi). *h* é a altura em milímetros da letra "e" minúscula de menor tipo presente no texto; *p* é a resolução em lp/mm; r é a resolução em dpi.

Sumarizamos, na **Tabela 3.3**, algumas resoluções necessárias para atingir o IQ desejado, para documentos com tipos de diferentes tamanhos.

Tabela 3.3: Relação entre resolução e Índice de Qualidade em documentos textuais

Altura do Menor Tipo h (em mm)	Resolução (em dpi) Necessária Para um IQ		
	Marginal (3,6)	Médio (5,0)	Alto (8,0)
0,5	366	508	813
1,0	183	254	406
1,5	122	169	271
2,0	91	127	203
2,5	73	102	163

A resolução utilizada no cálculo do IQ não é simplesmente a resolução nominal de escaneamento das imagens, e sim a resolução efetivamente obtida, ou a *resolução de saída*. Esta resolução normalmente é menor que a resolução nominal por causa dos chamados *erros*

de registro, que fazem com que os pontos da imagem nem sempre correspondam exatamente às amostras coletadas pelo elemento do *scanner*. Problemas de foco, trepidação mecânica e outros também podem resultar numa resolução de saída inferior à nominal. A resolução de saída pode ser medida através do uso de um cartão de resolução, o que garante a fidelidade dos resultados. (Figura 3.5).

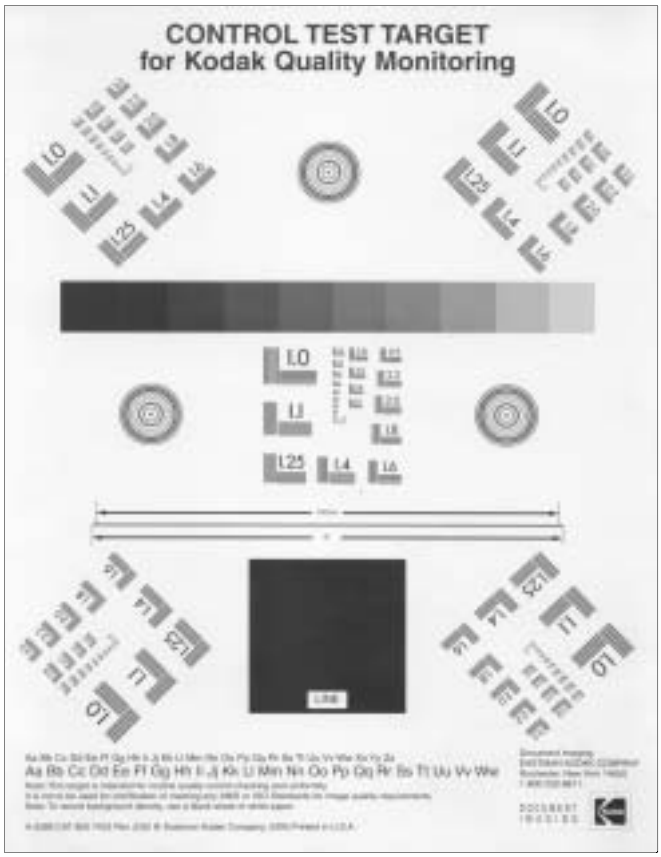


Figura 3.5: Cartão de resolução utilizado nos nossos experimentos. Existem várias versões desses cartões para medida de qualidade de imagem

Na falta deste, Kenney e Chapmam recomendam em escaneamentos multitonais considerar a resolução de saída igual à resolução nominal, e em escaneamentos bitonais considerar a resolução de saída como dois terços da resolução nominal [KENNEY 97].

Nosso experimento indicou, entretanto, que esses números nada têm de absolutos, sendo sempre melhor usar o cartão de resolução (**Tabela 3.1**). Observa-se, ao menos em nosso *scanner*, um *Hewlett-Packard Scanjet 6300c*, que os valores de resolução de saída das imagens bitonais não aumentam na mesma proporção da resolução nominal, requerendo correções cada vez maiores. O comportamento da resolução de saída dependente muito do

equipamento utilizado para a aquisição, uma razão adicional para usar o cartão ao invés dos fatores de ajuste.

Tabela 3.4: Comparação da resolução nominal *versus* a resolução de saída

Modo	Resolução Nominal (dpi)	Resolução no Cartão (lp/mm)	Resolução de Saída* (dpi)	Variação Resolução Nominal / Resolução de Saída
Bitonal	300	4,0	203	68%
Multitonal 8 *bits*	300	5,6	284	95%
Bitonal	450	5,0	254	56%
Bitonal	600	5,6	284	47%
Multitonal 8 *bits*	600	11	559	93%
Bitonal	900	7,1	361	40%

(*) Mensurada com o cartão de resolução. Equipamento: *Scanner HP Scanjet 6300c*.

3.2.6.2 Imagens monocromáticas: bitonais x multitonais

Em grande parte dos acervos textuais, a informação de cor é considerada irrelevante, limitando a escolha ao escaneamento bitonal ou o multitonal (tons de cinza). Isso é particularmente aplicável à digitalização de microformas (microfilmes, microfichas), em que a informação de cor foi perdida de qualquer maneira. Essa escolha, entretanto é muito importante, e afeta profundamente os resultados do trabalho de digitalização.

Uma imagem bitonal, contrastada com uma escala de cinza de 256 cores (suficiente para capturar toda a gama de cinzas perceptível pelo olho humano), é oito vezes menor. Por causa disso, e do alto contraste dos documentos textuais: tinta escura sob fundo claro, muito freqüentemente se advoga o uso da imagem bitonal na digitalização de textos.

Entretanto, a imagem multitonal preserva melhor a informação do documento, e em muitos casos, o torna mais legível. Textos com variações de contraste, borrados, esmaecidos ou irregularmente manchados são mais bem capturados em imagens com tons de cinza. Mesmo documentos relativamente bem preservados, se impressos sobre marcas d'água ou com fundos coloridos, freqüentemente têm sua digitalização mais bem sucedida em modo multitonal. Uma vantagem adicional, para o usuário, é a possibilidade de ajustar, interativamente, brilho e contraste da imagem ao visualizá-la, não só permitindo atender às preferências individuais, mas também facilitando a leitura de páginas com variações de contraste no original.

Nosso parecer recomenda o uso de imagens bitonais apenas para documentos especialmente regulares: tipografados ou impressos, com boa legibilidade, sem grande degradação e em que o contraste entre o texto e o fundo seja elevado e constante. Nos outros

casos, especialmente em textos datilografados, manuscritos, com diferentes níveis de contraste, ou que apresentem deterioração, recomendamos fortemente a digitalização multitonal.

Para contrastar os desempenhos da digitalização bitonal e multitonal, digitalizamos os documentos descritos na **Tabela 3.1** e comparamos os resultados obtidos, usando critérios subjetivos de legibilidade.

No experimento com documentos impressos (Documento 1 a Documento 6), digitalizamos diretamente os originais em nosso *scanner* de mesa, um *HP Scanjet 6300c*, usando os modos bitonal e de 256 tons de cinza fornecidos pelo *software* de interface do *scanner* (**Tabela 3.5**). Não notamos diferenças significativas de legibilidade: as imagens bitonais foram tão aceitáveis quanto as multitonais, mesmo no Documento 4, que apresentava um pequeno nível de borramento, e no Documento 6, que tinha um fundo manchado.

Os documentos históricos manuscritos (Documento 7 a Documento 10) foram microfilmados, e fizemos a digitalização em um *scanner* de microfilmes, o *Kodak 2400 DSV* (**Tabela 3.6**). As diferenças encontradas foram marcantes, especialmente nas imagens mais degradadas do acervo, onde a imagem bitonal se tornou, por vezes ilegível.

No experimento seguinte, simulamos uma situação que freqüentemente ocorre em fundos de personalidades: a reformatação de um livro com notas de margem, sublinhas e outras marcações do seu proprietário. Essas anotações são, muitas vezes, testemunhos documentais tão importantes quanto o texto original do documento. A digitalização do Documento 11, em um *Scanner HP 4400c*, revelou a dificuldade que a digitalização bitonal tem em lidar com essa situação. (**Tabela 3.7**)

Tabela 3.5: Comparação visual da digitalização bitonal × multitonal para páginas impressas

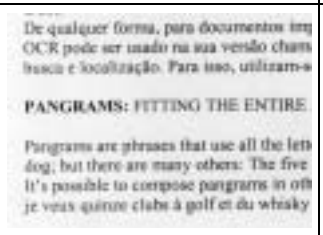
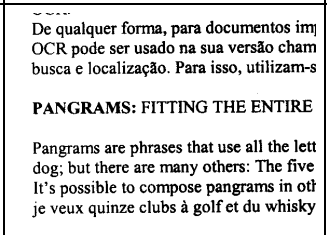
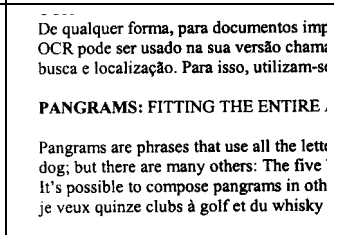
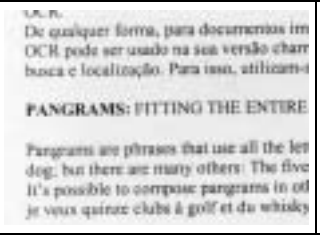
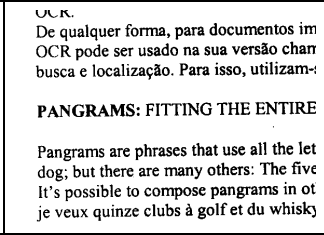
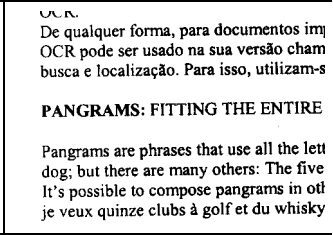
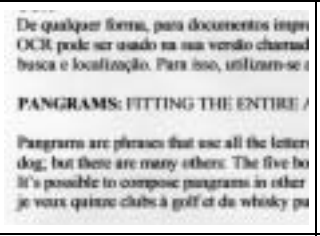
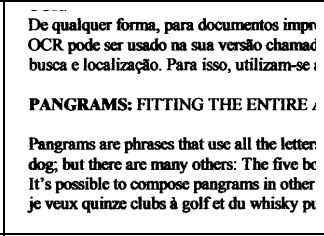
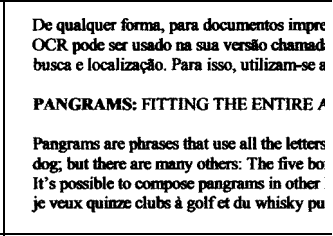
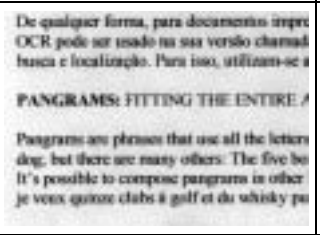
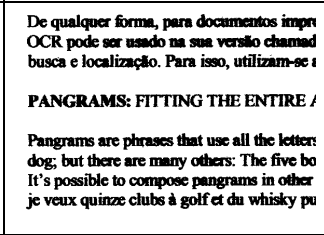
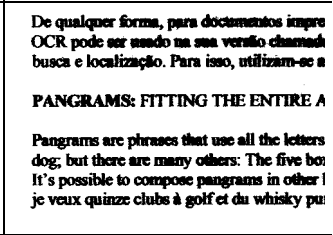
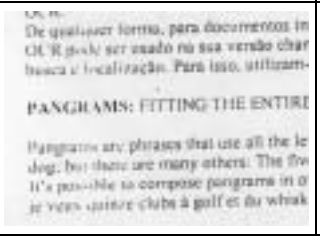
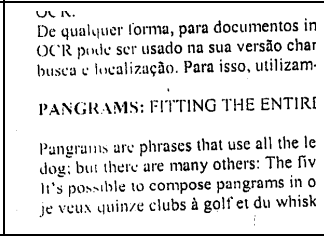
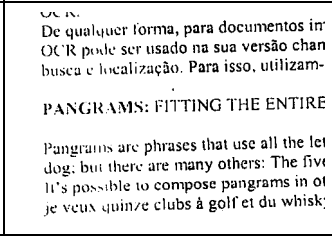
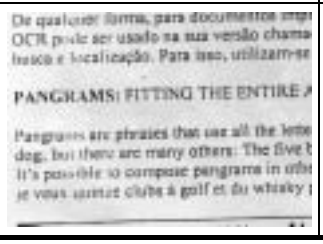
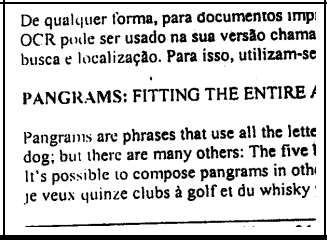
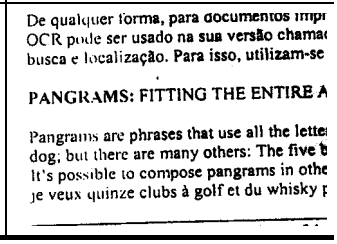
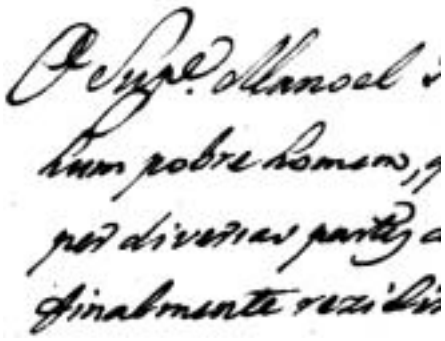
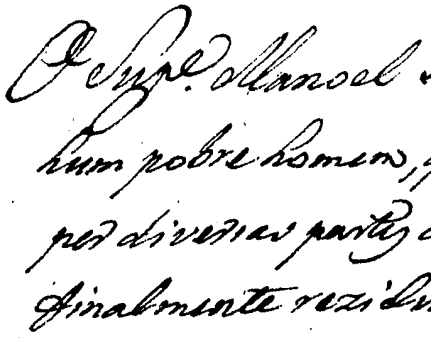
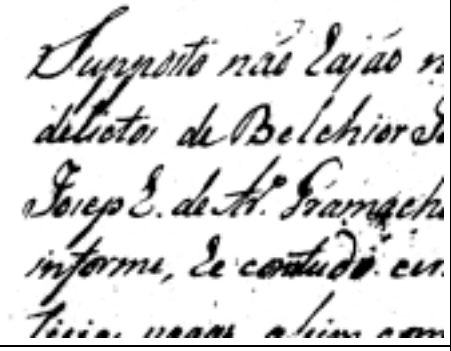
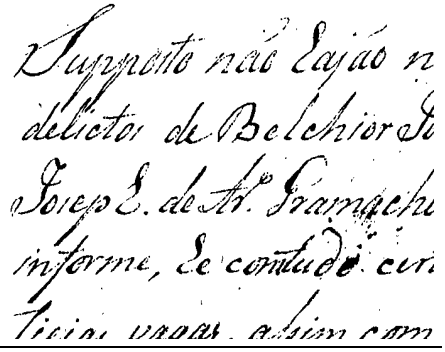
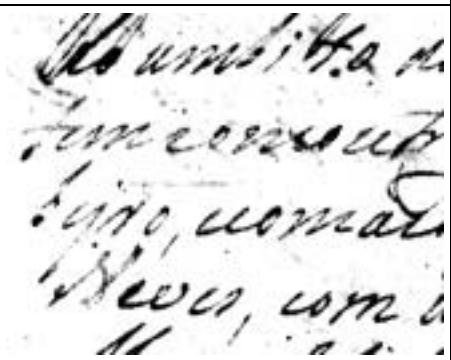
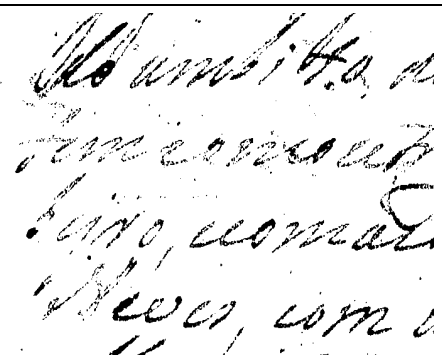
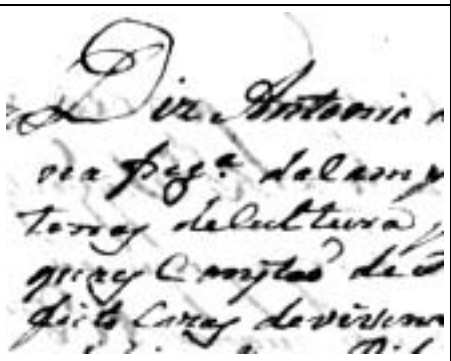
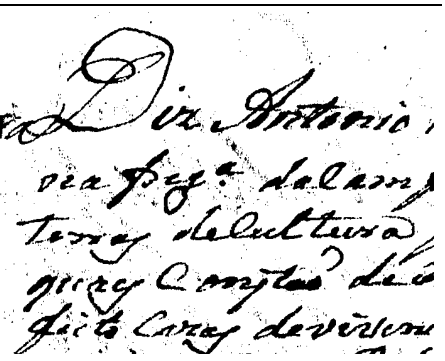
Modo da imagem:	Multitonal	Bitonal	Bitonal
Resolução nominal:	300 dpi	300 dpi	600 dpi
Resolução de saída:	284 dpi	203 dpi	284 dpi
Tamanho do arquivo:	8,24 MB	1,03 MB	4,12 MB
Amostra do Documento 1			
Amostra do Documento 2			
Amostra do Documento 3			
Amostra do Documento 4			
Amostra do Documento 5			
Amostra do Documento 6			

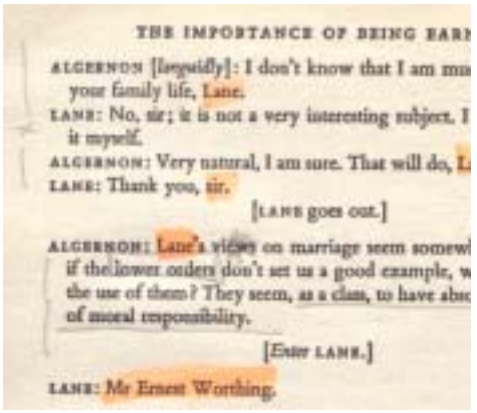
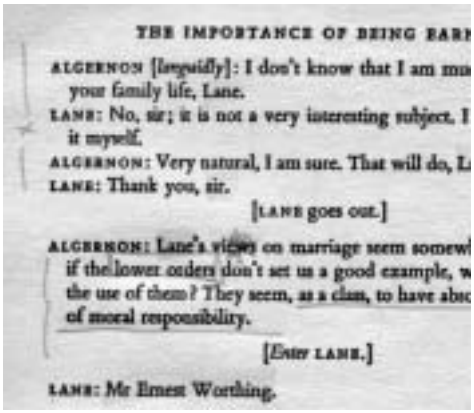
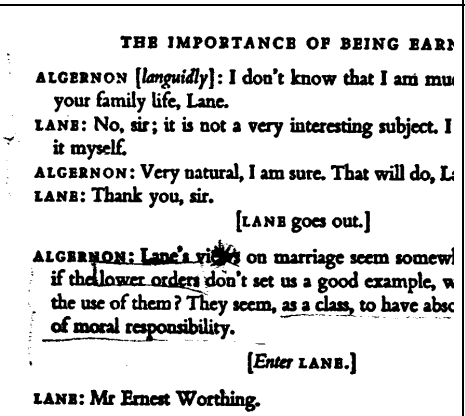
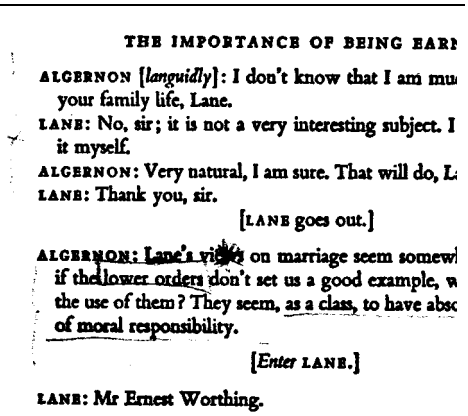
Tabela 3.6: Comparação visual da digitalização bitonal × multitonal para manuscritos

Modo da imagem:	Multitonal*	Bitonal**
Resolução nominal:	300 dpi	400 dpi
Resolução de saída:	203 dpi	183 dpi
Tamanho do arquivo:	8,24 MB	4,12 MB
Amostra do Documento 7		
Amostra do Documento 8		
Amostra do Documento 9		
Amostra do Documento 10		

* As imagens passaram por uma correção de contraste

** As imagens foram bitonalizadas no próprio *scanner* com um filtro combinado de aplicação de limiar e meios-tons, que proporciona uma maior legibilidade

Tabela 3.7: Comparação visual da digitalização bitonal × multitonal para página de livro tipografado com anotações manuscritas

Modo da imagem:	Cor	Multitonal
Resolução nominal:	300 dpi	300 dpi
Resolução de saída:	N/D	N/D
Tamanho do arquivo:	7,39 MB	2,46 MB
Amostra do Documento 11		
Modo da imagem:	Bitonal	Bitonal
Resolução nominal:	300 dpi	900 dpi
Resolução de saída:	N/D	N/D
Tamanho do arquivo:	315 KB	2,77 MB
Amostra do Documento 11		

Quando, por razões de tamanho de arquivo, for considerado essencial utilizar imagens bitonais, um procedimento que resulta em melhores resultados é fazer o escaneamento em tons de cinza, e depois converter a imagem utilizando um algoritmo selecionado, ao invés de confiar na bitonalização feita pelo próprio *scanner*, que normalmente é mais grosseira. A imagem bitonal gerada por esse processo pode ser colocada em uso normalmente, e a imagem em tons de cinza usada de intermediário pode ser descartada ou armazenada como matriz digital.

O método mais simples utilizado para bitonalização é a *aplicação de limiar*, e consiste em escolher uma tonalidade de cinza como limite entre os tons mais claros, que serão convertidos para branco, e os tons mais escuros, que serão convertidos em preto. A maioria dos *scanners*, ao fazer a digitalização em modo bitonal, o faz pela aplicação de limiar.

Para algumas imagens, entretanto, pode ser vantajoso utilizar os chamados métodos de meios-tons ou *pontilhamento*[1]. Esses métodos distribuem uma combinação de pontos pretos e brancos na imagem, procurando simular o efeito de escala de cinza original. Diferentes métodos de bitonalização podem ser vistos na Figura 3.6. As imagens em meios-tons podem inclusive ser visualizadas em tons de cinza, quando o usuário estiver utilizando monitores de baixa resolução, proporcionando excelente legibilidade. Elas não devem, porém, serem usadas em aplicações de reconhecimento de texto.

Se a decisão for pela digitalização multitonal, resta definir a profundidade de *bit*. Para aplicações convencionais, 8 *bits* costumam ser suficientes — o olho humano não consegue distinguir uma variação contínua de 256 gradações de cinza. Valores mais elevados são utilizados em aplicações especiais de visão computacional, quando as imagens precisam passar por muitas etapas de pós-processamento antes de serem entregues ao usuário final — e nesse caso uma maior profundidade de *bits* garante que os pequenos erros de arredondamento induzidos nos processamentos intermediários não irão prejudicar a aparência final da imagem. Para essas aplicações 12 *bits* apresentam uma margem segura. Em alguns casos, como na digitalização de microfilmes de alto contraste, a gama tonal é muito limitada, e 4 *bits* (16 tons de cinza) são suficientes.

[1] Este método é chamado, em língua inglesa, de *dithering*

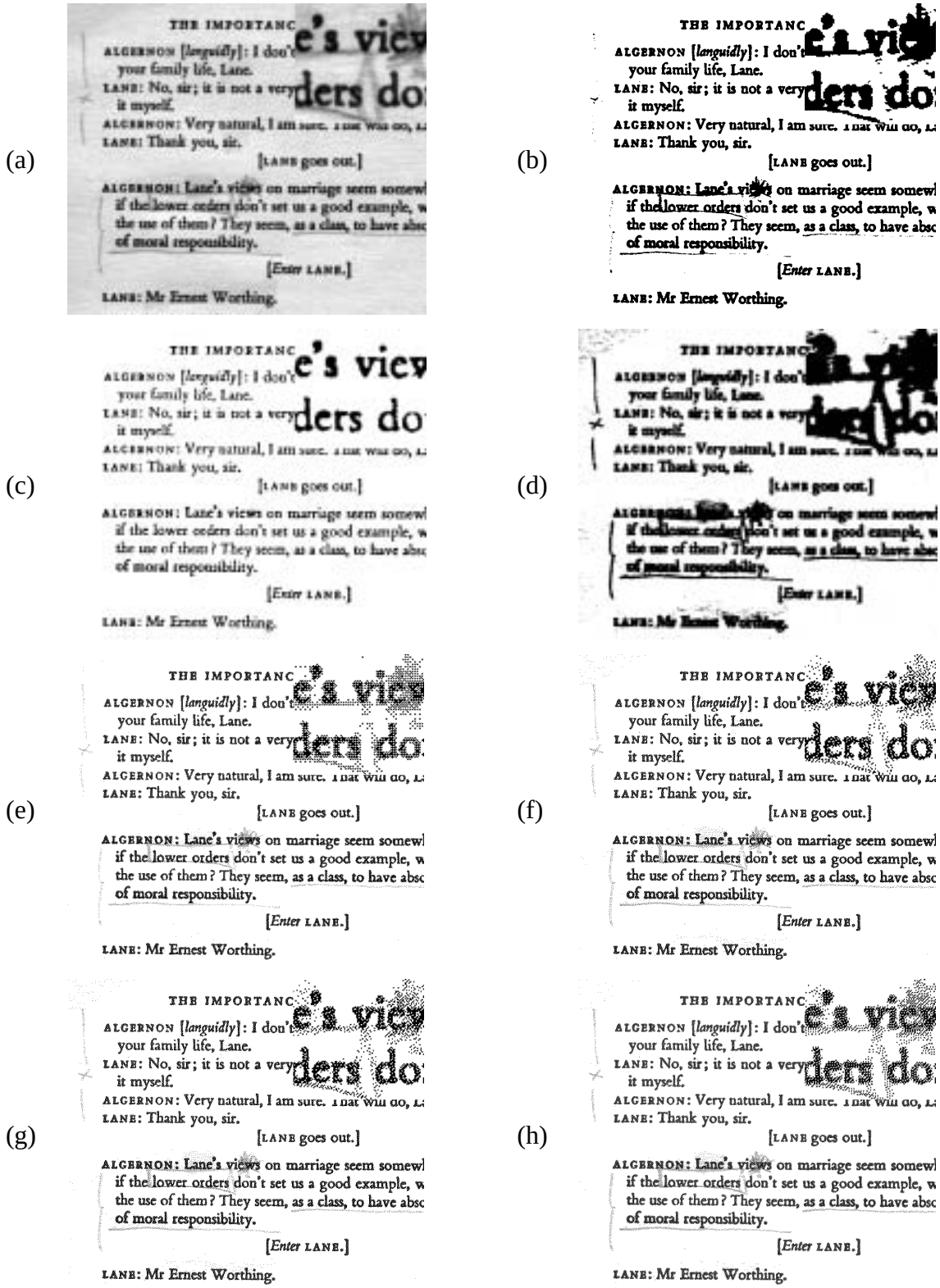


Figura 3.6: Resultados de diferentes métodos de bitonalização de imagens. Imagem multitonal *(a)* e bitonal geradas pelo *scanner (b)*. Imagens derivadas de (a) por aplicação de limiar, escolhido para preservar o texto tipografado (*c)* e as anotações manuscritas *(d)*. Imagens derivadas de (a) por pontilhamento pelo método de grade ordenada *(e)*, Jarvis *(f)*, Stucki *(g)* e Floyd-Steinberg *(h)*. O canto superior direito de cada imagem mostra a ampliação de um detalhe.

3.2.6.3 Imagens coloridas

A preservação da informação de cor, quando é necessário alto grau de fidelidade, apresenta um desafio adicional para os atuais sistemas de imagem. Isso, porque, embora toda imagem colorida seja formada por uma composição das cores primárias, existe uma grande variação na escolha exata da composição do espectro das cores que irão representar as primárias.

Isso significa que a forma como os diferentes *scanners* fazem a leitura dos valores de *verde*, *vermelho* e *azul*, não corresponde exatamente ao modo como os diferentes monitores definem e exibem essas cores. Pior do que isso, a maioria dos sistemas de impressão não se baseia nessas primárias aditivas, mas nas primárias subtrativas *amarelo*, *magenta* e *ciano*. Existe uma consideração adicional, que é o fato de que quase todos sistemas de imagem — monitores, impressoras, *scanners* e até mesmo o olho humano — não se comportam de maneira linear quanto à percepção ou produção das cores, mas de forma exponencial.

Na prática, isso provoca as seguintes dificuldades:

a) O números que representam as cores precisam ser traduzidos entre sistemas que usam cores primárias diferentes.

b) As escalas exponenciais de componentes diferentes precisam ser ajustadas de forma a fazer uma correspondência entre os valores de cor.

c) Algumas vezes um dispositivo não irá conseguir, de forma alguma, representar determinada cor, indicada por outro dispositivo, e uma cor aproximada precisa ser encontrada.

Para endereçar essas questões, vários sistemas de fidelidade de cor foram propostos, para fazer corresponder as cores em diferentes dispositivos, e endereçar o problema das cores não representáveis.

Um sistema, proposto pelo *International Color Consortium* prescreve que os dispositivos indiquem seus perfis de comportamento quanto à cor, em um formato padronizado, e especifica técnicas para traduzir as cores entre equipamentos diferentes [ICC 03]. Outro sistema, endereçado especificamente para a *Internet* e para a visualização de

imagens em monitores, envolveu a criação de um espaço de cores padronizado, chamado *sRGB* [STOKES 96].

Stone explica detalhadamente as questões envolvidas no processamento de imagens coloridas, bem como os diferentes sistemas de fidelidade de cor [STONE 01]. McCarthy descreve um fluxo de trabalho para usuários e instituições interessadas em trabalhar com gerência de fidelidade de cor [MCCARTHY 02].

3.3 Conclusões do capítulo

A facilidade no uso corrente da informação digital provoca, freqüentemente, a ilusão de que o usuário detém sua custódia completa, e de que os recursos disponíveis para sua manipulação são, em si, flexíveis e de alta acessibilidade. Uma análise mais esclarecida demonstra, entretanto, que os dados digitais são fortemente dependentes de todo um contexto tecnológico/social para que se obtenha mesmo o mínimo grau de intelegibilidade.

Ainda assim, a preservação dessa informação se impõem como uma necessidade premente, devido não apenas ao crescente patrimônio documental cujos originais provém diretamente de computadores, mas também ao uso da digitalização como técnica de reformatação para preservação em casos onde técnicas convencionais não apresentam bons resultados.

O uso da digitalização como forma principal ou colateral de preservação de acervos, requer um cuidadoso ajuste de parâmetros de qualidade, e decisões que variam desde os equipamentos de aquisição, até o formato de armazenamento dos arquivos.

Acima de tudo, é preciso compreender que a representação numérica consiste em uma síntese radical do artefato original — como no caso da informação de cor, em que toda a riqueza do espectro luminoso refletido pelo objeto é traduzido em apenas três números correspondentes, de forma aproximada, à percepção humana. É preciso avaliar quais os aspectos mais importantes a capturar dos documentos, com que grau de fidelidade, e garantir que o processo de conversão atenda as metas de qualidade estabelecidas.

4 Acesso

No universo dos acervos físicos e das técnicas convencionais de conservação, preservação e acesso são dimensões que não só se distinguem, como freqüentemente se opõem. “No mundo do papel e do filme, a preservação e o acesso são atividades relacionadas, porém distintas. É possível atender às necessidades de preservação de uma coleção de manuscritos, por exemplo, sem resolver os problemas de acesso.” [CONWAY 97]

Em se tratando de documentos frágeis, esta oposição torna-se mais dramática, como aponta Mustardo:

> “Mais sensível que a maioria dos documentos sobre papel, as fotografias têm uma química complexa que deve ser levada em consideração, caso se pretenda preservá-las para o futuro. (...) Indiscutivelmente, a maioria dos danos infligidos às fotografias são causados por **seres humanos**. Existem incontáveis exemplos de danos causados por manuseio, falta de cuidado, negligência, acidentes evitáveis, tentativas de conservação desastradas ou mal-informadas e até mesmo danos intencionais.” [MUSTARDO 97] (grifo original)

Para artefatos como registros cinematográficos e fonográficos, o compromisso entre preservação e acesso é ainda mais severo, tal a fragilidade dos artefatos. Normalmente só se obtém consulta a estas coleções sob as vistas e orientação cuidadosas de um técnico, em entrevistas marcadas com antecedência.

Com a aplicação da tecnologia digital, esse cenário é radicalmente transformado, pois não tanto se rompe o compromisso entre preservação e acesso, como se tornam essas dimensões relacionadas e cooperantes:

> “Recentes estratégias de gerenciamento de preservação, contudo, consideram que uma ação de preservação deverá ser aplicada a um item com o objetivo de torná-lo disponível para uso. Nesta perspectiva, gerar uma cópia de preservação de um livro deteriorado, em microfilme, sem tornar possível sua localização (...) é um desperdício de dinheiro. *A preservação no universo digital descarta toda e qualquer noção dúbia que entenda preservação e acesso como atividades distintas.*” [CONWAY 97] (grifo nosso)

Assim, o desenvolvimento de uma estratégia de recuperação eficaz da informação é necessário para que a digitalização de um acervo possa ser considerada uma atividade de preservação. Isso se coaduna com a idéia de que a memória tem uma importância social que

ultrapassa o simples arquivamento de artefatos. Já em 1939, o arquivista Binkley, defendia que “o objetivo da política de arquivos em um país democrático não pode ser a simples guarda de documentos. Deve ser nada menos que promover o enriquecimento da consciência histórica dos povos como um todo.” [BINKLEY 39]

Um grande desafio aos sistemas de informação multimídia que açambarcam grandes volumes documentais é impedir que o acervo degenere numa massa amorfa de informações mal classificadas e mal indexadas. Quando a base de dados não possui uma boa estrutura informacional, seu valor para os usuários tende a diminuir dramaticamente com seu crescimento, pois cada consulta resulta em uma grande quantidade de respostas espúrias, mescladas à informação útil:

> “O grande problema de hoje é ensinar às pessoas a ignorar o irrelevante, a recusar saber as coisas, antes que se vejam sufocadas. Pois informações em excesso são tão danosas quanto nenhuma informação.” (W. H. Auden)

Assim, o acervo de dados digitais deve ser protegido contra sua própria desmesura, o que é feito através de mecanismos de arranjo e indexação e da criação de instrumentos de pesquisa que garantam a possibilidade de recuperação da informação.

4.1 Coleções e arranjo

Entre os extremos de se considerar o acervo inteiro de uma instituição como um todo articulado, e cada item documental em sua individualidade, o conceito de *coleção* vem crescendo de progressivamente de importância, tanto para fins de organização quanto de descrição dos itens [MILLER 00].

Numa visão extremista, qualquer subconjunto de documentos pode merecer este nome. O significado do termo *coleção* varia muito de acordo não só com o tipo de acervo, mas principalmente com a natureza da instituição que o custodia.

Nos acervos de Arquivos, as coleções obedecem universalmente, ao *princípio da proveniência*, ou *princípio de respeito aos fundos*. Este princípio se tornou central na arquivologia moderna, estabelecendo que a integridade de um item documental depende, em

parte, do conhecimento de onde ele veio. Desta forma, uma determinada coleção, vinda de uma instituição ou herdada de uma personalidade será tratada como um *fundo*, e seus documentos não serão misturados aos de outros fundos. Na medida do possível, procura-se também preservar a organização interna do *fundo* como ela se encontrava em sua forma original (o chamado princípio de respeito *interno* aos fundos), pois isso ajuda a manter os documentos em seu contexto original, preservando melhor seu significado [DUCHEIN 86]. Para cumprir esses objetivos, as coleções de arquivos são hierárquicas ou multinível, seguindo as subdivisões convencionais *fundo*, *série, subsérie* e *dossiê* [SWEET 00].

Em Museus, onde um mesmo documento pode ganhar diferentes perspectivas em diferentes contextos, as coleções adquirem uma notável flexibilidade. A coleção pode consistir no acervo completo do museu, ou apenas nas obras de um determinado artista (e.g., *Coleção Manet*), ou apenas nas obras de uma determinada época, local ou escola (e.g., *Coleção Séc. XIX*, *Coleção Impressionista*), ou um tipo de material ou técnica (e.g., *Coleção de Gravuras*), ou uma certa disciplina (e.g., *Coleção de Paleontologia Mamífera*), ou ainda serem definidas de acordo com o seu curador ou doador (e.g., *Coleção Augusto de Lima*) ou sua finalidade (e.g., *Coleção Educativa*) [DUNN 00]. Depreende-se que as coleções não são estanques, e nem mesmo puramente hierárquicas, um mesmo documento pode pertencer a diversas coleções.

A noção de coleção é mais recente em Bibliotecas, talvez porque um livro seja um tipo de documento autocontido em seu significado, em contraste com um documento arquivístico ou uma peça museológica. O conceito de coleção mais antigo encontrado nestas instituições talvez seja o utilizado para definir diferentes políticas de acesso (e.g., *coleção reserva, obras de referência, obras para empréstimo…*). Não obstante, um novo conceito de coleção surge para as bibliotecas a partir da criação e do compartilhamento de catálogos, e do interesse na criação de *diretórios de recursos*, propiciados sobretudo pelo sucesso dos grandes portais de pesquisa padronizados por sistemas como o Z39.50. Com o advento desses diretórios, o acervo de uma biblioteca, ou suas divisões temáticas, passam a corresponder a coleções de um sistema mais amplo [MILLER 00].

A divisão do acervo em coleções permite dar-lhe um *arranjo* - i.e., separá-lo em categorias menores, auto-contidas e relacionadas. O arranjo é um poderoso auxiliar no processo de recuperação da informação, pois divide um acervo muito grande em porções mais

manuseáveis e inteligíveis, permitindo ao usuário recuperar as informações que lhe são de interesse.

4.2 Indexação

A indexação consiste na atribuição de termos de pesquisa aos itens do acervo. Esses termos podem ser palavras-chave controladas ou não, ou até mesmo códigos de classificação. Os termos podem ser atribuídos não apenas aos itens documentais, mas também aos elementos do arranjo, permitindo uma indexação mais sumária ou mais analítica da coleção.

A forma mais rudimentar de indexação consiste na atribuição de palavras-chave não controladas aos itens do acervo. Este método, embora simples e facilmente aplicável, apresenta severas desvantagens:

a) **Sinonímia:** ocorre quando termos diferentes são utilizados para denotar mesmo significado (e.g., *carro* e *automóvel*).

b) **Polissemia:** ocorre quando o mesmo termo é utilizado para denotar significados diferentes (*e.g.*, *prato* - vasilha, instrumento musical, a cidade italiana…).

c) **Dissociação semântica:** ocorre por não haver relação alguma entre termos de significados correlatos (*e.g.*, agricultura e agronomia são correlatos, mas o sistema de busca não tem noção disso).

d) **Especificidade irregular:** ocorre quando alguns termos de indexação são muito específicos e outros muito gerais, resultando em pesquisas ora em poucos itens, ora em itens demais. (*e.g.* usar ao mesmo tempo *vertebrado* e *invertebrado* para animais — muito geral; e *rosácea*, *iridácea*, *cariofilácea* para plantas — muito específico).

No intuito de superar essas dificuldades, a indexação pode ser feita através de, em ordem crescente de sofisticação, vocabulários controlados ou tesauros.

4.2.1 Vocabulários controlados

No primeiro degrau de sofisticação, temos os vocabulários controlados, que tentam endereçar principalmente o problema da sinonímia, da polissemia, e em parte da especificidade irregular.

Nessa técnica os termos não são escolhidos livremente pelo indexador, mas a partir de uma ou mais bases léxicos. Desta forma, ficam decididos, *a priori*, quais termos podem participar do vocabulário de indexação. Os vocabulários devem ser cuidadosamente construídos para evitar sinônimos, e os termos devem conter notas que esclareçam a acepção preferida, eliminando assim a polissemia.

Os termos nos vocabulários são controlados, mas não relacionados - ignorando o problema da dissociação semântica, e resolvendo apenas em parte a questão da especificidade irregular.

Se houver a necessidade de estender o vocabulário, os novos termos devem ser introduzidos cuidadosamente, de forma a preservar a integridade semântica do léxico, evitando, por exemplo, gerar sinônimos. Essa tarefa de manutenção às vezes se revela bastante difícil.

4.2.2 Tesauros e tesauros facetados

Os tesauros avançam a sofisticação dos vocabulários controlados, provendo relacionamentos entre os termos de indexação.

Os tesauros adicionam um nível de sofisticação aos vocabulários controlados, permitindo relacionar semanticamente os termos de indexação. Desta forma, é possível endereçar os problemas da dissociação semântica e da especificidade irregular.

Os termos de um tesauro podem ser *autorizados* ou *não autorizados*. Os primeiros são os termos preferenciais, utilizados na descrição dos itens. Os segundos são fornecidos apenas para fins de consulta, indicando termos equivalentes.

Essas relações são:

a) **Termo Sinônimo:** relação entre termos autorizados, diferentes que denotam exatamente o mesmo significado (e.g., *automóvel* e *viatura*, sendo que ambos participam do tesauro).

b) **Termo Equivalente:** termo autorizado que denota exatamente o mesmo significado de um termo não autorizado (e.g., *automóvel* é termo equivalente de *carro*, se o primeiro é autorizado e o segundo não).

c) Termo Abrangente ou Genérico: termo cujo significado é mais geral (e.g., *veículo* é termo geral de *automóvel*).

d) **Termo Específico:** termo cujo significado é mais específico (e.g., *caminhão* é termo específico de *veículo*).

e) **Termo Relacionado:** relação entre termos que participam da mesma área semântica (e.g,, *automóvel* e *motor a explosão*).

Uma sofisticação adicional se encontra nos chamados tesauros facetados. Nessa técnica, o universo de classificação é dividido em múltiplos aspectos, denominados *facetas*. As facetas devem ser ortogonais e mutuamente exclusivas, e uma combinação delas é utilizada na indexação.

Os tesauros facetados permitem a combinação de termos de forma a localizar documentos que atendam a critérios bem precisos.

Os tesauros tem sido usados em tarefas de indexação dirigidas por humanos e em instrumentos de pesquisa convencionais, em papel. Associados porém ao processamento eletrônico, eles se tornam extremamente úteis, adicionando uma dimensão semântica aos termos de uma coleção. Assim, é possível combiná-los com algoritmos especiais de indexação automática, ou ainda a mecanismos de busca semântica que permitem devolver ao usuário um conjunto de resultados semanticamente relacionados à sua pesquisa, classificados por relevância [TUDHOPE 02].

4.3 Busca no conteúdo

Os métodos de busca baseada no conteúdo permitem a localização dos itens documentais através da investigação direta do conteúdo dos dados, ao contrário dos métodos vistos anteriormente, que se baseiam nos metadados.

As técnicas aplicáveis dependem da natureza dos documentos que compõem o acervo.

4.3.1 Conteúdo textual codificado

Chamamos de conteúdo textual codificado aquele em que o texto é representado em nível de caractere. Nesse tipo de informação, é possível fazer a busca por palavras ou frases, utilizando algoritmos clássicos de buscas em seqüência, listas invertidas, e árvores digitais.

A simples busca por seqüências de caracteres — o recurso que a maioria dos sistemas oferece como *busca em texto-livre*, apresenta os mesmos problemas dos vocabulários não controlados: termos sinônimos fazem com que parte a informação relevante não seja encontrada, e a polissemia provoca o retorno de respostas espúrias.

A busca textual, entretanto, pode ser associada a tesauros, à análise da *informação semântica latente*. Esses métodos permitem, com mais precisão, que o usuário encontre os documentos de seu interesse [DUMAIS 88].

4.3.2 Conteúdo de imagens de texto

Nas imagens de texto, não temos os caracteres codificados individualmente. Nesses casos, para fazer uma busca por palavras e frases, é preciso antes fazer uma interpretação do conteúdo da imagem, na tentativa de reconhecer os caracteres ali **desenhados**. (Seção 3.2.3)

A conversão para caracteres, quando se deseja apresentar ao usuário o texto convertido, requer taxas de reconhecimento bastante elevadas, mas quando a conversão é utilizada apenas para fins de localização, não é necessária tanta precisão. Uma técnica denominada *OCR sujo*[1] aplica a busca aproximada em seqüências ao texto parcialmente reconhecido por um OCR. Uma vez localizadas as seqüências desejadas, apresenta-se ao usuário a imagem de varredura original, ao invés do texto com erros.

[1] Do inglês: *Dirty-OCR*

4.3.3 Conteúdo visual

Uma das aplicações mais interessantes da tecnologia digital em acervos iconográficos é a possibilidade de utilizar métodos de busca que interpretam diretamente o conteúdo das imagens, ao invés de simplesmente confiarem nos metadados a elas associados.

Esses métodos de busca extraem características que tanto podem corresponder às dimensões de percepção humana como cor, textura e forma, quanto ao resultado de operações matemáticas como médias, histogramas, transformadas espectrais, etc. Em seguida eles utilizando métodos estatísticos ou de inteligência artificial para classificar as imagens do acervo. O usuário faz a pesquisa especificando uma *imagem chave*, e o sistema retorna todos os itens similares a ela.

Métodos de busca baseado em conteúdo costumam ser mais bem sucedidos em acervos especializados, em que as características visuais das imagens provocam uma classificação significativa. Assim, em um acervo que contenha apenas moedas, as características de textura, forma e cor são bons classificadores, permitindo ao usuário localizar itens de interesse. Em acervos muito variados, entretanto, esses métodos são menos bem sucedidos, pois um grande número de imagens não relacionadas irá apresentar características extraídas similares.

Um outro fator que ajuda o sucesso desses métodos é a regularidade das imagens. Se em um acervo de moedas, as fotografias forem feitas com o mesmo fundo monocromático, sob condições de foco e iluminação similares, o conteúdo de interesse (a moeda) irá ser o fator mais expressivo de diferença entre as imagens, e a classificação tenderá a agrupar os itens similares. Se, ao contrário, as imagens tiverem fundos texturizados e distintos, ângulos de iluminação diversos, equilíbrios de cor muito variados, esses elementos tenderão a interferir na classificação.

A busca no conteúdo visual já é fornecida em alguns sistemas comerciais, com bons resultados em acervos especializados, respondendo a critérios simples e específicos (e.g., localizar todos os carros vermelhos de um acervo de automóveis). Estão sendo estudados métodos cada vez mais sofisticados, que irão permitir no futuro, fazer buscas muito mais interessantes, por elementos temáticos e estilísticos (e.g., localizar todas as fotografias em que aparece determinada pessoa, localizar todos os quadros de um determinado pintor, etc.) [DEL BIMBO 99].

4.4 Conclusões do capítulo

O impacto da tecnologia digital sobre o acesso provoca uma completa mudança de perspectiva na preservação dos acervos, já que se antes as duas dimensões eram ortogonais ou até opostas, agora elas se tornam cooperantes, e mesmo interdependentes.

Por isso, a implantação de projetos digitais em instituições de custódia documental vai além da simples aquisição de equipamentos e sistemas, e considerações de ordem tecnológica — é preciso se preparar para as inevitáveis questões que o novo instrumento de acesso ao acervo suscita: repriorização das atividades de indexação convencional , compatibilização dos antigos instrumentos de pesquisa com os novos, questões de cessão de imagens e controle de *copyright*, impactos políticos e econômicos dos novos serviços sobre a organização, *etc*.

À medida que as modernas técnicas de busca baseada no conteúdo — com possibilidades tão interessantes quanto encontrar um trecho de melodia assoviado, todas as fotografias em que aparece determinado rosto, ou todos os manuscritos que falam de determinado assunto (através de uma análise da linguagem natural, não de simples palavras-chave) — forem se transmutando de promessas de pesquisa em realidades aplicáveis, o cotidiano do profissional de acervos, e os modelos de interação consulente/documentos praticados hoje terão de ser profundamente revistos.

5 Preservação

Para efeitos de preservação, a maior vantagem dos dados digitais é sua perfeita replicabilidade, que se explica por sua natureza numérica. Enquanto os dados analógicos estão sujeitos às imperfeições do mundo físico, que impedem a fidelidade da replicação, cada cópia digital é um clone, indistinguível do original. Em teoria, poderíamos regenerar a informação digital de um suporte ao outro eternamente, à medida que os originais fossem envelhecendo, enquanto ruídos e atenuações impediriam a informação analógica de gozar desta mesma perenidade.

A prática, entretanto, demonstra que a preservação digital apresenta desafios formidáveis, muito maiores que os da preservação convencional, devido não apenas à fragilidade de suas mídias, mas também à obsolescência da tecnologia que permite interpretá-los [CONWAY 97].

Neste capítulo exploramos os três suportes da informação digital mais usados contemporaneamente: os discos rígidos, as fitas magnéticas e os discos ópticos. Oferecemos não apenas informações sobre sua confiabilidade e durabilidade, mas também recomendações para utilizá-los de forma correta e segura.

Descrevemos, em seguida, alguns procedimentos capazes de prolongar a vida útil dos materiais digitais, combatendo simultaneamente a obsolescência e o envelhecimento dos suportes.

Finalizamos com algumas considerações importantes sobre a segurança de sistemas e informações.

5.1 Suportes da informação digital

Talvez por causa das campanhas publicitárias das grandes empresas de tecnologias, as pessoas tendem a acreditar que tanto os suportes magnéticos (fitas e discos) quanto os ópticos (*CD-ROMs*, *DVDs*) são muito mais longevos do que a realidade aponta. Ao contrário do que diz o bordão acerca dos *compact discs* — "Som Perfeito, Para Sempre"[1] — estudos apontam que as maioria das mídias digitais, em condições especiais, têm uma vida útil máxima de 10 ou 20 anos. Em condições mais comuns, a vida útil pode ser tão baixa quanto 5 anos [CONWAY 97] [VAN BOGART 95].

Os suportes digitais são frágeis por natureza. Dadas as enormes densidades de gravação, pequenos defeitos nas mídias podem tornar grandes massas de informação indisponíveis. Essa alta densidade, aliada à delicadeza dos processos de leitura e gravação, faz com que os suportes físicos da informação digital tenham uma expectativa de vida média de no máximo umas poucas décadas.

Os arquivistas estão tradicionalmente acostumados a lidar com suportes cuja durabilidade se mede em séculos, e não em anos. Se a disseminação do papel de polpa de madeira química — cuja constituição acídica o leva a uma lenta autodestruição ao longo de 100 ou 150 anos — já foi motivo de grave preocupação, as mídias digitais, com suas curtíssimas vidas, por vezes de menos de uma década, representam um desafio que não se poderia conceber no passado.

Conway demonstra que ao longo dos séculos, a escolha de suportes de informação tem privilegiado o aumento de densidade em detrimento da durabilidade, fazendo com que hoje cheguemos à situação denominada por ele "o dilema dos suportes modernos". (Figura 5.1) [CONWAY 97].

1 "*Perfect Sound, Forever*" - bordão de lançamento do *compact disc* de áudio, pelas companhias *Sony* e *Philips*, na década de 1980.

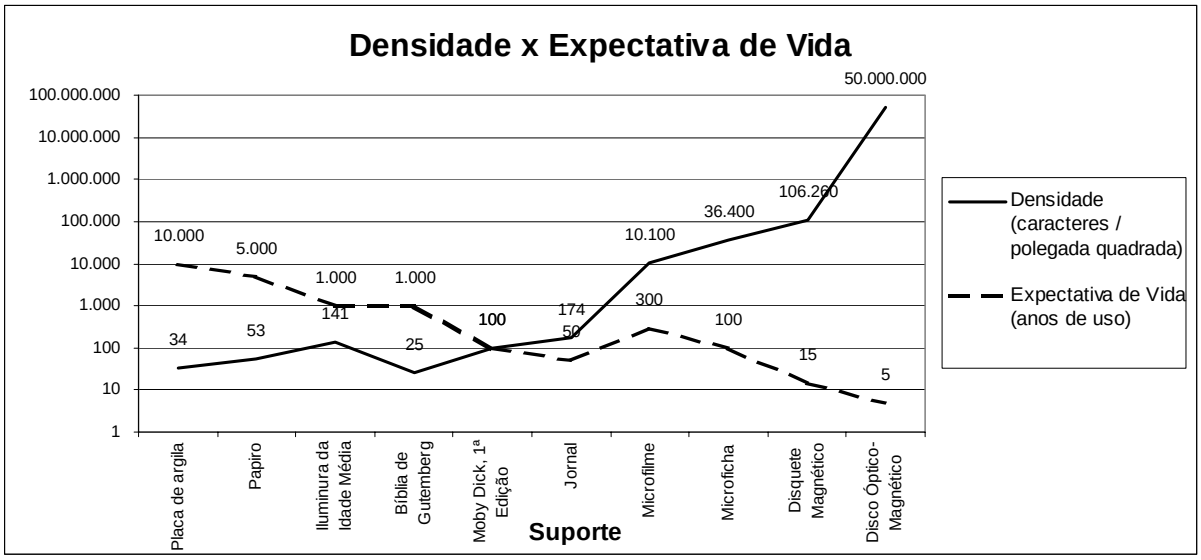


Figura 5.1: O dilema dos suportes modernos
Fonte — CONWAY 97

5.1.1 Armazenamento secundário x terciário

Os suportes de armazenamento podem se classificar, segundo a disponibilidade mais ou menos imediata da informação para o usuário, em *secundários* ou *online*, e *terciários* ou *offline*.[1]

O armazenamento secundário caracteriza-se pela imediata disponibilidade dos dados, sem a necessidade de intervenção do operador. O usuário pode solicitar ao sistema de informação um arquivo ou registro, e acessar seu conteúdo sem nenhuma operação manual intermediária.

O armazenamento terciário caracteriza-se pela necessidade de intervenção do operador antes que os dados se tornem disponíveis ao usuário. É preciso inserir a mídia da informação em um dispositivo de leitura/gravação e efetuar alguns procedimentos (chamados procedimentos de *montagem*) para permitir o acesso aos dados. Em alguns dispositivos simples — e.g., disquetes comuns, *CD-ROMs* — os procedimentos de montagem são triviais e podem ser executados pelo próprio usuário. Em outros — e.g., fitas magnéticas em rolo — a montagem é complexa, e requer a assistência de um operador treinado.

Apesar dos inconvenientes trazidos pela necessidade da montagem, os suportes terciários têm aplicações importantes, devido às vantagens trazidas pela desvinculação entre a

[1] Há também os meios de armazenamento *primários* ou a *memória primária* - que é usada apenas para fins imediatos de processamento: o conteúdo da memória primária é perdido sempre que o computador é desligado.

mídia da informação e o dispositivo de leitura/gravação. De forma geral, a mídia terciária é mais portátil e mais econômica que a mídia secundária.

A maior parte dos sistemas de informação complexos associa os dois tipos de suporte para garantir uma melhor preservação dos dados por um custo mais reduzido. Utiliza-se normalmente o critério da *freqüência de acesso* para determinar quais dados serão armazenados em mídias secundárias, e quais serão delegados às mídias terciárias. Ficam nas primeiras, informações habitualmente acessadas, como arquivos de uso corrente, bancos de dados *online*, código executável de aplicativos e do sistema operacional, etc. Para as segundas vão as informações de acesso mais casual, como cópias de segurança (*back-ups*), versões antigas de arquivos, arquivos grandes de uso infreqüente, e o arquivo-morto.

5.1.2 Discos rígidos - *SLED* e *RAID*

A tecnologia dos discos rígidos é, no momento, a principal representante dos suportes de armazenamento secundário. Estes discos se caracterizam por acoplarem a mídia dos dados ao dispositivo de leitura/gravação em uma unidade, permitindo grandes velocidades de transferência de dados, confiabilidade em operação contínua, e dispensabilidade da montagem.

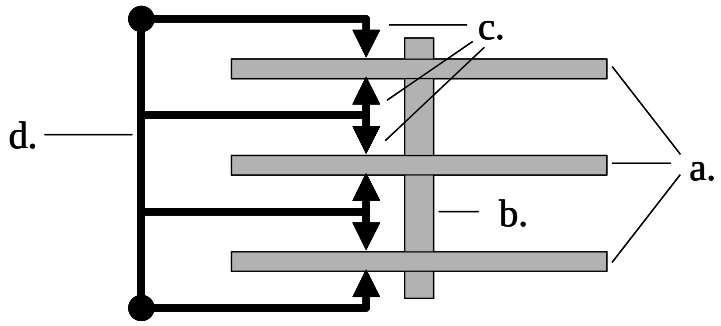


Figura 5.2: Perfil esquematizado de um disco rígido, mostrando os *a)* discos montados sobre o *b)* eixo, as *c)* cabeças de leitura/gravação e o *d)* braço de acionamento das cabeças.

Os discos rígidos são formados por um ou mais discos de material rígido, cobertos de material magnetizável, que giram unidos a um eixo, em alta velocidade (da ordem de 5.000 RPM). Sobre cada uma das faces dos discos, uma cabeça de leitura e gravação move-se radialmente, medindo ou alterando seu campo magnético. (Figura 5.2, Figura 5.3). As cabeças flutuam sobre os discos a distâncias micrométricas, suspensas sobre colchões de ar: a

proximidade garante confiabilidade nas operações, enquanto a ausência de contato real evita o desgaste das peças [BRAIN 03a].

Figura 5.3: Aspecto interno de um disco rígido, evidenciando os múltiplos discos e a cabeça de leitura/gravação da superfície superior. Há outras 5 cabeças de leitura/gravação, não visíveis nesta imagem.
Fonte — BRAIN.03a

Ao aplicar os discos rígidos no armazenamento secundário, o projetista tem basicamente duas opções.

A primeira consiste em utilizar unidades de disco rígido de alto desempenho e alta capacidade, de forma que um único dispositivo seja capaz de armazenar todos os dados. Este esquema, nomeado sob o acrônimo *SLED — um único disco grande e caro*[1], é conceitualmente simples e não requer suporte especial de *hardware* ou *software*. Se houver dados demais para um único disco, a solução é simplesmente adicionar mais dispositivos, sem nenhum arranjo especial — num esquema chamado, algo pejorativamente, de *JBOD — somente um monte de dispositivos*[2]. Os problemas desta abordagem são óbvios: no caso de falha de um dos discos, todos os dados nele armazenados serão perdidos.

A segunda opção associa diversos dispositivos, dividindo entre eles a tarefa de armazenar os dados e provendo alguma redundância, sendo assim tolerante a falhas. Este esquema é chamado de *RAID — arranjo redundante de discos baratos*[3], e embora seja mais difícil de configurar e requeira suporte especial de *hardware* e *software*, é muito mais seguro contra falhas internas dos dispositivos.

---

[1] *SLED* - do inglês *Single Large Expensive Disk*
[2] *JBOD* - do inglês *Just a Bunch of Drives*
[3] *RAID* - do inglês *Redundant Array of Inexpensive Disks*

Existem diversos modos de configurar um arranjo *RAID*, com diversos compromissos entre segurança, desempenho e espaço ocupado pelos dados redundantes. A arquitetura do sistema *RAID* é um fator crítico para a segurança do sistema de informação e deve ser decidida criteriosamente

Um dos arranjos mais populares hoje é o chamado *RAID-5* que associa três ou mais dispositivos. O *RAID-5* apresenta bom desempenho tanto para leitura quanto para escrita. Sua segurança e custo são dependentes do número de unidades envolvidas: se o arranjo for composto de n discos ($n \geq 3$), estarão disponíveis para o usuário o espaço equivalente a apenas n-1 discos. Entretanto, se qualquer um dos discos falhar, é possível reconstruir completamente as informações do arranjo a partir das informações dos discos restantes. O *RAID-5* não é, entretanto, tolerante à falha simultânea de dois ou mais discos.

Chen et al. descrevem em detalhes diversos possíveis esquemas *RAID*, seus custos, desempenhos, aplicações e até detalhes técnicos de implementação [CHEN 94].

Devido à sua precisão, discos rígidos são artefatos delicados. Uma queda de alguns centímetros de altura é suficiente para arruiná-los. Um acidente comum é o choque de uma das cabeças de leitura/gravação sobre a superfície de um disco, que dadas as distâncias mínimas de seu sobrevôo, pode ser disparado por uma simples partícula de poeira. O choque arranha a superfície magnética, adulterando informações e desprendendo mais partículas — o que pode provocar novos choques. Por causa das altas velocidades de operação, falhas nos motores e na lubrificação também podem ocorrer.

Não se atribui aos discos rígidos uma expectativa de vida superior a 4 ou 5 anos. Por causa de sua fragilidade, deve-se minimizar seu transporte e mantê-los sob operação em ambientes controlados, livres de vibrações ou flutuações de temperatura. Sob essas condições, pode-se esperar deles uma operação contínua e confiável, durante sua curta vida útil.

5.1.3 Fitas magnéticas

As fitas magnéticas são, ao lado dos discos ópticos, a principal representante dos suportes de armazenamento terciário. Sendo, talvez, o suporte de dados mais antigo ainda amplamente utilizado em sistemas de informação, elas sofreram diversas evoluções desde seu advento, no início da década de 1950.

Quando comparadas aos discos ópticos e óptico-magnéticos, as vantagens das fitas são a grande capacidade de armazenamento, o baixo custo por unidade armazenada, a longa expectativa de vida e a confiabilidade na retenção dos dados ao longo de sua vida útil. Suas desvantagens são o acesso seqüencial (as fitas requerem um moroso avanço e retrocesso para que sejam acessados os dados desejados), a necessidade de treinar o operador ou usuário para sua manipulação correta, o elevado custo dos dispositivos de leitura/gravação e a maior fragilidade.

Fitas estão disponíveis em rolos, cassetes ou cartuchos (Figura 5.4). Fitas em rolos, a forma mais antiga, requerem cuidadosos procedimentos de montagem, mas são baratas e permitem bastante controle do operador. Fitas em cassete embutem um rolo doador e um rolo receptor em um único invólucro e são hoje em dia as mais difundidas. Cartuchos possuem um único rolo: a fita se apresenta ou como um laço sem fim (de forma que um único rolo possa atuar como doador e receptor) ou com uma guia inicial que é adaptada a um segundo rolo embutido no dispositivo de leitura/gravação. Cassetes e cartuchos são muito mais simples de montar [BOSTON 00a].

(a)

(b)

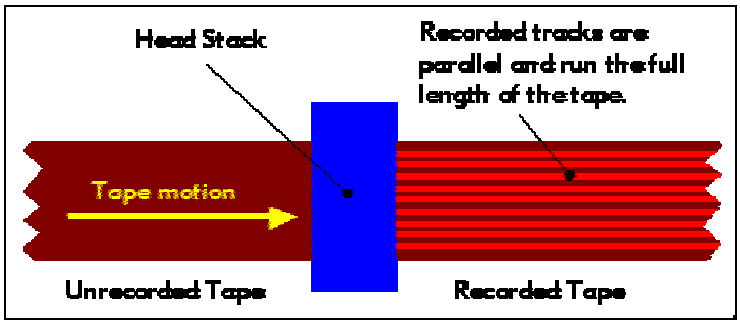
(c)

Figura 5.4: Invólucros disponíveis para fitas magnéticas. *a)* rolo; *b)* cassete; *c)* cartuchos

Existem basicamente duas tecnologias de gravação em fitas magnéticas: a longitudinal e a helicoidal. A primeira utiliza uma cabeça estática, que grava trilhas de dados paralelas ao sentido de deslocamento da fita. A segunda utiliza cabeças rotativas, acopladas a um tambor que gira em alta velocidade, gravando trilhas de dados diagonais ao sentido da fita (Figura 5.5). A tecnologia helicoidal permite uma densidade de gravação muito maior que a

longitudinal, mas impõe um severo desgaste tanto sobre a mídia quanto sobre o equipamento, por causa do atrito do tambor giratório, que chega a alcançar velocidades de 2.000 RPM.

(a)



(b)

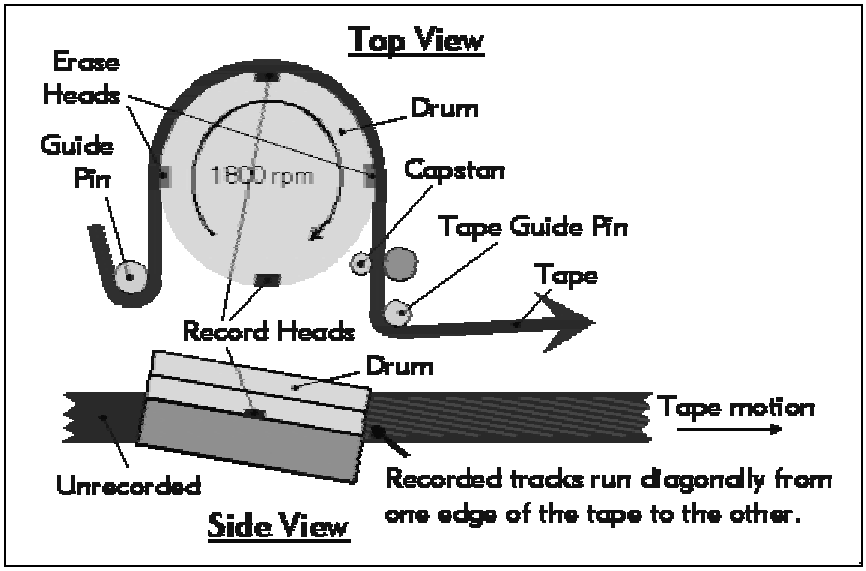


Figura 5.5: Sistemas de gravação de fita magnética. *a)* Linear; *b)* Helicoidal
FONTE — [VAN BOGART 95]

Um exemplo da tecnologia helicoidal é a *DDS*[1], uma fita de 4mm em cassete, introduzida pela *Sony* e pela *Hewlett-Packard*, que utiliza a mesma tecnologia da fita *DAT*[2]. Em sua versão mais recente, o *DDS-4*, essas fitas têm capacidades nativas de 20 GB, chegando a 40 GB em modo comprimido. Por causa do desgaste mecânico, os fabricantes destas fitas garantem sua confiabilidade por apenas 2.000 passagens pela cabeça de leitura/gravação, em condições ideais. Como em uma única operação da fita normalmente provoca mais de uma passagem pelo mesmo local, os fabricantes recomendam que a mesma fita seja usada em apenas cerca de 100-150 operações de cópia — em condições ideais. A

[1] Do inglês *Digital Data Storage*
[2] Do inglês *Digital Audio Tape*

cabeça de leitura do dispositivo sofre também desgastes, e tem uma expectativa de vida de 2.000 horas de uso [ANDERSON 02a].

A fita *DLT*[1], uma fita de meia polegada em cartucho, patenteada pela *Quantum Corporation*, exemplifica a tecnologia longitudinal. Na versão *DLT-IV*, estas fitas têm capacidades nativas de 40 GB (80 GB em modo comprimido). Um mecanismo especial reduz tanto o desgaste das fitas, quanto das cabeças de leitura do dispositivo. Em condições ideais, as fitas resistem a 1.000.000 de passagens, ou cerca de 10.000 operações de cópia, enquanto a expectativa de vida da cabeça pode chegar a 30.000 horas [ANDERSON 02a] [SUPER DLT].

Estruturalmente as fitas magnéticas são formadas por uma base coberta por uma superfície de gravação — um polímero onde está disperso o pigmento magnético (como óxidos de ferro ou de cromo). Normalmente adiciona-se a esta superfície um componente lubrificante. A fita pode ter uma cobertura traseira, para proteção e redução de atrito. (Figura 5.6).

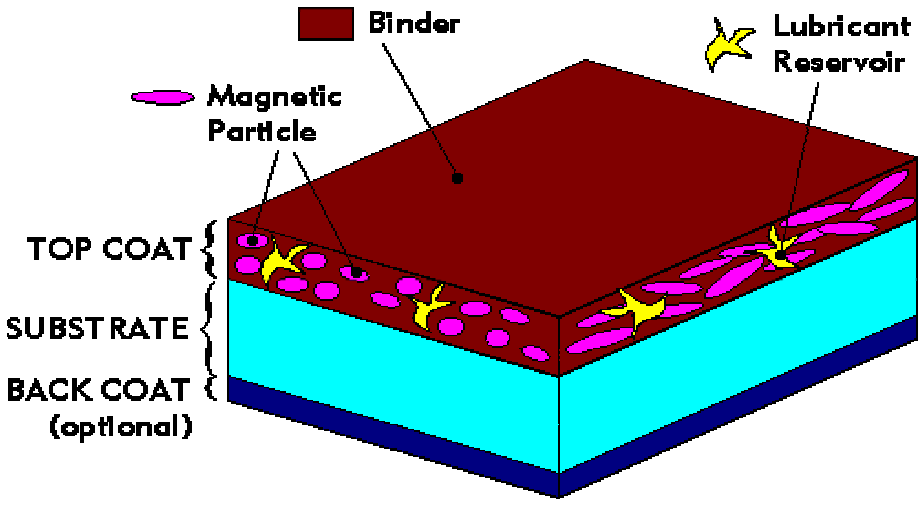


Figura 5.6: Corte de uma fita magnética, mostrando seus componentes: a base; a camada magnética com as partículas dispersas no polímero; o lubrificante; e a cobertura traseira opcional
FONTE — [VAN BOGART 95]

A durabilidade e confiabilidade da fita magnética estão condicionadas à saúde de todos estes componentes. Nas fitas modernas, a base é de poliéster muito resistente e quimicamente estável e os pigmentos magnéticos são óxidos metálicos estáveis. O primeiro elemento a se degradar quase sempre é o polímero de dispersão, responsável — dentre outras

[1] Do inglês *Digital Linear Tape*

coisas — pela adesão da superfície de gravação à base. A umidade atmosférica provoca no polímero uma reação conhecida como hidrólise, deteriorando suas propriedades. A fita atacada por hidrólise pode apresentar separação entre as camadas de gravação e de base, ou ainda a *síndrome da fita grudenta*[1] em que a superfície magnética se torna pegajosa e adere à cabeça de leitura/gravação, por vezes impedindo completamente a recuperação dos dados [ROSS 99] [VAN BOGART 95].

Em alguns casos, a superfície de gravação não é composta de pigmentos dispersos em polímero, mas de uma finíssima camada metálica depositada diretamente sobre a base. Embora estas fitas não estejam sujeitas à hidrólise, elas são extremamente sensíveis à poluição e umidade atmosféricas, que atacam o metal.

Procedimentos corretos para manipulação e armazenamento de fitas magnéticas são essenciais para garantir sua longevidade. Basicamente, as fitas devem ser armazenadas em condições de baixa temperatura e umidade relativa do ar, longe de poluição, poeira, tabaco e gases corrosivos. Elas devem ser protegidas da exposição acidental a campos magnéticos fortes, como detectores de metais, autofalantes, motores elétricos, etc. As fitas devem ser sempre armazenadas em posição vertical, de forma que com o tempo, o rolo não se apóie sobre um dos lados do carretel, danificando a borda da fita quando esta for desenrolada. Algumas fitas precisam ter seus rolos retensionados periodicamente, após longos períodos sem uso, o que é feito rebobinando-os em velocidades controladas. As fitas não devem sofrer quedas ou choques violentos, nem grandes variações de temperatura, e somente devem ser manipuladas por usuários treinados, em ambientes limpos. Para que as mídias não sejam danificadas durante a operação, os dispositivos de leitura/gravação devem estar sempre cuidadosamente limpos e regulados, especialmente os rolos tensores, os guias da fita e a cabeça de leitura/gravação [BOSTON 00a] [ROSS 99] [VAN BOGART 95].

Ao contrário dos discos rígidos, as fitas magnéticas não toleram uso contínuo: o desgaste das mídias provocado cada passagem pelo mecanismo limita o número de operações. Algumas tecnologias de fitas (e.g. *DLT*) provêem redundância e uma capacidade de corrigir pequenos erros nos dados (chamados *soft errors* — erros leves). Ao final de cada operação, o usuário é informado de quantos erros leves foram encontrados e corrigidos. Um aumento

[1] Do inglês *sticky tape syndrome*

nesse número é sinal de que a fita deve ser substituída antes que ocorram erros irrecuperáveis (chamados *hard errors* — erros graves).

Com os cuidados devidos, a expectativa de vida de uma fita pode alcançar três décadas, freqüentemente ultrapassando a própria obsolescência de sua tecnologia.

5.1.4 Discos ópticos e óptico-magnéticos

Discos ópticos e óptico-magnéticos são um avanço relativamente recente na tecnologia de suportes de informação, tendo sido introduzido em larga escala apenas na década de 1980.

O primeiro meio óptico digital foi o *compact disc*, ou *CD*, um disco de policarbonato de 12 cm revestido por uma finíssima superfície refletora de alumínio e por uma camada de verniz protetor. A base de policarbonato é mecanicamente estampada de modo a formar pequenos ressaltos sobre a superfície refletora, que ora dispersam, ora refletem um raio *laser* usado na leitura das informações do disco (Figura 5.7) [BRAIN 03b]. O *CD* foi originalmente concebido para o registro de áudio digital. Mais tarde surgiram variações para gravação de outros tipos dados, incluindo *CD-ROM* ou *disco compacto de memória apenas para leitura*[1] para uso em computadores. Todos eles, entretanto, são gravados em processo mecânico que se torna econômico somente para centenas de unidades.

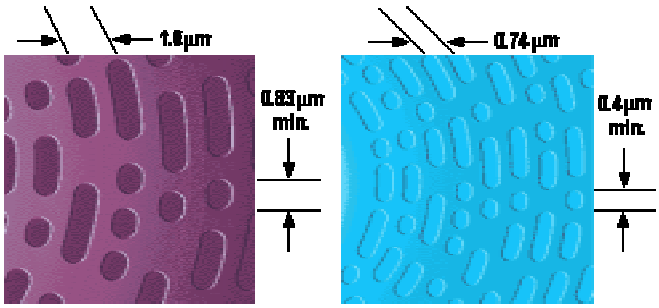


Figura 5.7: Os discos ópticos, tanto o CD (à esquerda) quanto o DVD (à direita) utilizam pequenos ressaltos para refletir e dispersar a luz do *laser*, representando assim os padrões binários
FONTE — [ANDERSON 03]

Para permitir a gravação de *CD*'s em pequena escala, foram propostas as tecnologias *CD-MO*, *CD-R* e *CD-RW*. O *CD-MO* é um disco *óptico-magnético:* sua leitura é feita através de raio *laser*, mas sua gravação é feita por uma combinação de aplicação de *laser*

[1] *Compact Disc Read-Only Memory*

e de um campo magnético, que altera as propriedades refletoras da superfície de registro. O *CD-MO* permite múltiplas gravações. Exclusivamente ópticos, o *CD-R* e o *CD-RW*, apresentam uma camada adicional entre a superfície refletora e a base de policarbonato. Durante o processo de gravação, as propriedades desta camada são alteradas por um *laser* de gravação, que altera suas propriedades ópticas. O *CD-R* permite apenas um registro, mas múltiplas leituras — enquanto o *CD-RW* pode ser regravado [ANDERSON 03].

A tecnologia dos *compact discs* foi desenvolvida pela *Philips Electronics* e pela *Sony Inc.*, com a participação de outros parceiros. As diferentes mídias e formatos de dados aplicáveis são descritos nos chamados *rainbow books* (**Tabela 5.1**).

Tabela 5.1: Os múltiplos formatos de *compact discs*

Mídia	Formato de Dados	Descrição	Especificação	Lançamento
CD	CD-DA	Áudio digital	*Red Book*	1980
	CD-ROM	Dados de computador	*Yellow Book*	1984
	Video CD	Vídeo digital	*White Book*	1993
CD-MO		Óptico-magnético / Regravável	*Orange Book* - Parte I	1990
CD-R		Óptico / Gravável	*Orange Book* - Parte II	1993
CD-RW		Óptico / Regravável	*Orange Book* - Parte III	1996

FONTE — Dados compilados de [ANDERSON 02b], [ANDERSON 03], [BRAIN 03b] e [MCFADDEN 03]

Um *compact disc* pode chegar a uma capacidade de 700 MB — um número impressionante para a década de 1980, mas insuficiente para as necessidades do século XXI. Discos ópticos para aplicações mais agressivas, como vídeo digital de alta qualidade e armazenamento de grandes volumes de dados precisavam ser criados.

O *DVD* ou *disco digital versátil*[1] foi a resposta da indústria. Esses discos ópticos apresentam diversas inovações em relação aos *CDs*: uma maior densidade de dados, uma codificação mais eficiente, a possibilidade de duas camadas de dados em cada lado do disco, e a possibilidade de gravação de dados nos dois lados do disco. Em sua versão mais simples, com um único lado de uma única camada, o *DVD* suporta 4,7 GB de dados — aproximadamente 7 vezes mais que um *CD*. Com dois lados de duas camadas o DVD chega a alcançar 17 GB de capacidade (Figura 5.7) [ANDERSON 02c].

[1] Do inglês *Digital Versatile Disc*

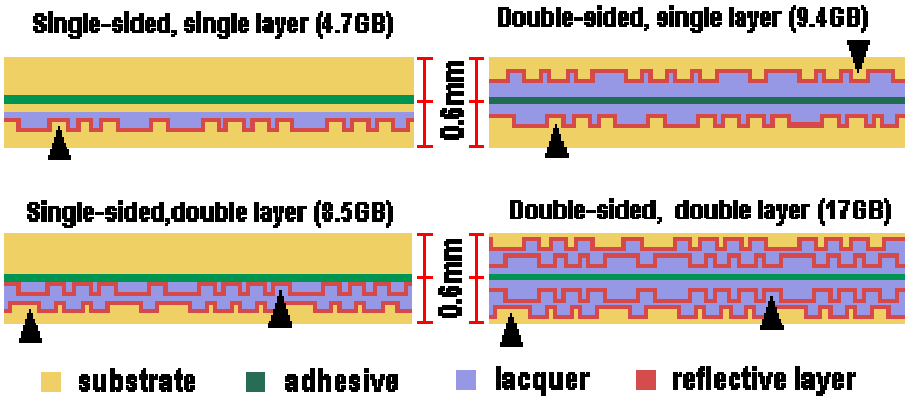


Figura 5.8: Corte dos diferentes tipos de DVD, revelando sua estrutura.
FONTE — [ANDERSON 03]

Em sua versão original, o *DVD*, como o *CD*, é um formato de distribuição às centenas. Para atender as demandas de gravação em pequena escala, logo surgiram versões graváveis e regraváveis. Infelizmente, este processo gerou uma série de mídias ligeiramente incompatíveis entre si (**Tabela 5.2**).

Tabela 5.2: Diferentes tecnologias de DVD e suas compatibilidades

Tecnologia	Aplicação	Lançamento	Mídias que Lê	Mídias que Grava
DVD-ROM	Apenas leitura	1996	ROM, *R*, *RW**, *RW+**	N/A
DVD-R	Gravável	1997	ROM, R, RW, *RW+**	R
DVD-RAM	Regravável	1998	ROM, R, RAM, *RW**, *RW+**	RAM
DVD-RW	Regravável	2000	ROM, R, RW*, *RW+**	RW
DVD+RW	Regravável	2001	ROM, R, *RW**, RW+	RW+

* Mídias que podem apresentar problemas de compatibilidade na leitura
FONTE — Dados obtidos de [ANDERSON 02c]

Os discos ópticos tornaram-se a mídia terciária mais popular, por uma série de razões. Seu procedimento de montagem é extremamente rápido e simples, podendo ser executado pelo próprio usuário, sem necessidade de treinamento especial. Os discos permitem acesso direto, i.e., a localização imediata dos dados desejados sem a necessidade de lentas operações de avanço e retrocesso e — em comparação com fitas magnéticas, disquetes ou discos rígidos — são suportes muito mais resistentes.

Um *CD-ROM* ou *DVD* pode cair da altura de uma mesa, ser coberto por impressões digitais, ou encharcado de café e, após uma limpeza, sobreviver a esses acidentes, que certamente inutilizariam a melhor das fitas. Os discos puramente ópticos apresentam a vantagem adicional de serem imunes a campos magnéticos.

Essa relativa resistência das mídias ópticas tem seus limites. Arranhões às vezes invisíveis a olho nu podem desviar o raio *laser* provocando desalinhamentos e erros de leitura. Nesses casos, a base de policarbonato pode ser polida, para tentar eliminar os arranhões, mas é um procedimento trabalhoso e falível. A camada de verniz (oposta à face de leitura) é ainda mais frágil, especialmente nos discos graváveis e regraváveis, onde, por vezes, ela pode ser descascada com a unha. Ao contrário da base de policarbonato, que admite tentativas de restauração, qualquer dano à camada de verniz resulta em completa ruína dos dados subjacentes. Ambas as superfícies, portanto, devem ser protegidas contra poeira excessiva, abrasão, umidade e ataques de fungos. Muito critério deve ser utilizado ao se selecionar o tipo de caneta, etiqueta ou tinta de estampagem a ser aplicado sobre discos ópticos, pois substâncias inadequadas encurtam muito a vida útil do verniz. Os discos são sensíveis a deformações dimensionais, que podem ser provocadas tanto por exposição a altas temperaturas, quanto por tensões mecânicas [BOSTON 00b] [MCFADDEN 03] [ROSS 99].

A expectativa de vida dessas mídias tem sido alvo de muita controvérsia. A camada de policarbonato é quimicamente estável, e armazenada em condições de temperatura e umidade razoáveis, permanece perfeita por séculos. A camada refletora, quando de alumínio, está sujeita à oxidação, pois algumas moléculas de oxigênio conseguem romper a barreira protetora do verniz — limitando assim a vida útil dos discos a 10 ou 20 anos. Para remover este obstáculo, alguns fabricantes substituíram o alumínio — um material altamente reativo — por metais nobres como prata e ouro, estendendo assim a vida útil deste componente indefinidamente. A camada de verniz está sujeita à hidrólise em condições de excessiva umidade, ou no caso de imersão prolongada dos discos em água. Além disso, ela pode se decompor em contato com adesivos ou tintas inadequadas para o uso em mídias ópticas.

Presentes apenas nas mídias graváveis e regraváveis, os componentes das camadas de gravação são o ponto mais polêmico na discussão da expectativa de vida dos discos. Nos *CD-R*'s, são usados os pigmentos *Cyanine*, o *PhtaloCyanine* (e sua versão mais recente, o *Advanced PhtaloCyanine*), o *Metallized Azo* e o *Formazan*. Cada fabricante debate sobre as vantagens relativas destes componentes. Aparentemente, o *Cyanine* é mais tolerante a pequenas variações no processo de gravação, enquanto, o *PhtaloCyanine* e o *Advanced PhtaloCyanine* são mais resistentes para armazenamento de muito longo prazo [MCFADDEN 03].

O que é certo é que as mídias apresentam enormes variações de durabilidade de acordo com os fabricantes. Mídias de baixa qualidade, muitas vezes comercializadas sem marca, podem deteriorar-se completamente em menos de um ano, às vezes com completa separação entre as camadas — um fenômeno chamado *delaminação*[1]. Enquanto isso, alguns fabricantes, como a *Kodak* garantem uma durabilidade de 100 anos para suas mídias, se armazenadas em condições razoáveis de temperaturas e umidade relativa [KODAK 01] [STINSON 95]. Como todas as mídias ópticas são muito recentes, e os testes de durabilidade são resultados de ensaios de envelhecimento acelerado — e não de observação direta de situações reais — muitos arquivistas têm uma certa prevenção contra esses números [CONWAY 97].

Estudos apontam para o fato de que existe uma forte correlação entre a qualidade imediata dos discos e sua durabilidade em longo prazo. Os *CD*s são gravados com muita redundância e métodos de correção de erros, porém as mídias que apresentam uma grande necessidade de correção de erros logo no início de sua vida útil parecem apresentar uma durabilidade mais curta. É possível, através do uso de *hardware* ou *software* especiais, acompanhar a qualidade dos discos logo após a sua gravação e reservar para armazenamento de longo prazo apenas aqueles cujo registro tenha alcançado qualidade excepcionalmente boa [HARTKE 01] [PARKER 96].

Com uma seleção das melhores mídias, um controle de qualidade logo após a gravação, o armazenamento em condições ambientais razoáveis (menos de 30º C, menos de 60% de umidade relativa do ar, com pouca flutuação desses parâmetros) e limpeza no armazenamento e na manipulação, é certo atingir uma longevidade de ao menos uma década, e — se os ensaios de envelhecimento acelerado forem confiáveis — talvez um século. Controles de qualidade periódicos podem permitir avaliar a deterioração progressiva da mídia, e substituí-la em tempo, enquanto houver dados redundantes suficientes para a atuação dos mecanismos de correção de erros.

[1] Do inglês *delamination*

5.2 Volatilidade tecnológica

O maior problema da preservação dos dados digitais não é a fragilidade dos suportes de armazenamento e sim a rápida mudança de tecnologia. Não basta garantir a sobrevivência e longevidade das mídias: é necessário garantir que as informações possam ser recuperadas e entendidas.

Ao contrário de uma página de papel, que pode ser lido diretamente, ou de um microfilme, que na pior das hipóteses irá necessitar de uma vela e uma lupa, arquivos digitais requerem sofisticados sistemas cooperantes de *hardware* e *software* para tornar a informação inteligível. Considerando a velocidade com que esses sistemas são superados e abandonados por seus fabricantes, deixando de receber qualquer suporte, é possível chegar a uma situação em que vastos acervos digitais sejam perdidos, não sendo mais possível encontrar os equipamentos e programas necessários para acessá-los [DOLLAR 92].

As dependências da informação digital em relação à sua tecnologia de origem são discutidas mais detalhadamente na Seção 3.1.

A única forma de superar essa dificuldade é aceitar que os dados digitais, sem intervenção humana freqüente, não sobrevivem. É preciso a cada poucos anos (5 ou 10), reavaliar o cenário tecnológico e tomar as medidas para garantir a sobrevivência do acervo nas novas condições. Isso pode ser feito através da migração dos dados (convertendo os arquivos digitais para formatos de programas mais modernos) ou da emulação (criando simuladores dos ambientes operacionais antigos, nas plataformas novas). Ambas as medidas, requerem como forma de apoio, a operação mais simples do refrescamento (transferência dos dados para outras novas mídias) [BESSER 00].

5.2.1 Refrescamento

O refrescamento consiste na transferência dos arquivos de um suporte para outro, sem alteração do seu conteúdo. Ele é feito tanto pelos efeitos da obsolescência, quanto pela degradação física das mídias [TFADI 96].

O envelhecimento das mídias digitais precisa ser acompanhado cuidadosamente, porque ao contrário do que acontece com as gravações analógicas, que vão perdendo a qualidade progressiva e perceptivelmente, as mídias digitais vão sofrendo degradações de

sinal a princípio aceitáveis pelos seus mecanismos de leitura e, portanto, invisíveis para o operador — até o dia em que a leitura falha completamente e já não é mais possível recuperar os dados. Algumas tecnologias de fitas magnéticas e discos ópticos permitem fazer mensurações periódicas das taxas de correção de erro de leitura — um indicador de que o dispositivo já está começando a enfrentar problemas para interpretá-las — e tomar medidas corretivas antes que seja tarde demais.

O refrescamento entre mídias de mesma natureza — e.g., de um *CD-R* para outro — é uma simples operação de cópia, e não envolve outras considerações a não ser os cuidados habituais em qualquer trabalho de reformatação: o controle da retirada dos materiais de seus depósitos permanentes, a separação cuidadosa entre as mídias que já foram refrescadas e as que ainda vão o ser, a identificação correta dos novos materiais com etiquetas, o controle de qualidade da gravação, *etc*…

Entretanto, freqüentemente o refrescamento é feito entre suportes de tecnologias diferentes, à medida que os antigos vão deixando de ter suporte da indústria, e os novos vão se estabelecendo como padrões. Esta operação em nada afeta a natureza dos dados digitais, *i.e.*, o arquivo copiado no novo suporte (e.g., DVD-R) é, em todos os aspectos, idêntico ao original no suporte antigo (e.g., CD-R). Entretanto, o refrescamento para novas mídias enfrenta uma dificuldade adicional, que é o aumento da densidade dos dados (e.g., um DVD-R equivale a mais de 5 CD-Rs; um CD-R equivale a mais de quatrocentos disquetes). Isso faz com que, um grande número de itens físicos distintos seja condensado em itens unificados.

O impacto disso sobre a preservação está na indexação dos dados, essencial para que se possa recuperá-los depois. Uma determinada instituição poderia ter os arquivos *“0001.TIF”* e *“0002.TIF”* referentes às páginas do documento *X* armazenados no disquete *“D-120”*, e outros arquivos *“0001.TIF”* e *“0002.TIF”* referentes às páginas do documento *Y* armazenados no disquete *“D-350”*. No processo de reformatação para CD-R, os disquetes *“D-001”* até *“D-450”* são copiados para o CD-R *“C-001”*. O que deve ser feito? Separar os arquivos no CD-R em pastas distintas é uma possibilidade, renomear os arquivos para *“120-0001.TIF”* e *“350-0001.TIF”*, etc., é uma outra. Em qualquer caso, será necessário alterar as referências ao arquivo original, de forma que elas reflitam a nova localização do arquivo, e/ou, criar uma tabela de equivalência entre as referências antigas e as novas, para

que as referências que não puderem ser atualizadas (por estarem em material impresso, por exemplo) possam ser transliteradas para o novo esquema.

5.2.2 Migração

O refrescamento só resolve o problema de preservação dos dados, enquanto não houver mudanças na plataforma de *software* e *hardware* que ameacem a possibilidade de interpretar e visualizar esses dados. Mudanças tecnológicas mais profundas exigem a *migração* dos dados para os formatos suportados pelos novos *softwares*.

Um exemplo seria a transferência de um arquivo do *Wordstar* para o *Microsoft Word*. Outro seria converter um arquivo gráfico GIF para PNG, ou TIFF 5.0 para TIFF 6.0. A migração, como o refrescamento, é uma operação periódica, e seu propósito é preservar a integridade dos dados digitais, mantendo a possibilidade de que eles sejam visualizados e utilizados pelos usuários. Ao contrário do refrescamento, porém, a migração é um procedimento complexo e custoso, que modifica a estrutura dos arquivos digitais.

Como qualquer processo de tradução, a migração sujeita os dados a uma certa distorção de significado. É freqüente, por exemplo, que no processo de conversão de textos de um processador para outro, parte da formatação original seja perdida ou alterada. A conversão de valores numéricos entre formatos diferentes pode provocar a perda de dígitos significativos, com conseqüências mais ou menos sérias, dependendo da natureza do original.

Lawrence estabelece um método de levantamento de riscos para a análise da migração, permitindo avaliar, *a priori*, a chance de a conversão de um formato a outro provocar perdas significativas de informação. Seu método procura avaliar o impacto da reformatação sobre todos os aspectos do dado digital: conteúdo, imutabilidade, contexto, referência, etc. O levantamento de riscos leva em consideração também questões administrativas, como custos, recursos humanos, gestão de *copyright*, etc. [LAWRENCE 00].

Lawrence usa uma escala para a medida dos riscos que combina a probabilidade de um evento ocorrer, e o impacto desse evento sobre a coleção, caso ocorra (**Tabela 5.3**).

Tabela 5.3: Escalas de Medida do Risco de Migração

Probabilidade do Risco		
Probabilidade	**Valor**	**Descrição**
Muito Alta	5	Probabilidade estimada entre 26-99%
Alta	4	Probabilidade estimada entre 11-25%
Moderada	3	Probabilidade estimada entre 6-10%
Baixa	2	Probabilidade estimada entre 1-5%
Muito Baixa	1	Probabilidade estimada abaixo de 1%

Impacto do Risco		
Impacto	**Valor**	**Descrição**
Catastrófico	E	Completa e irreversível perda de dados. Dados não podem ser obtidos de outras fontes — impressas ou digitais
Muito Sério	D	Parcial e irreversível perda de dados. Dados não podem ser obtidos de outras fontes
Sério	C	Completa perda de dados. Dados podem ser totalmente reconstruídos de outras fontes
Significante	B	Parcial perda de dados. Dados podem ser totalmente reconstruídos de outras fontes
Pequeno	A	Perda parcial ou completa de dados. Dados podem ser copiados de outros arquivos digitais

FONTE — [LAWRENCE 00]

Medidos os riscos, utiliza-se uma tabela de decisão para determinar o plano de ação (Figura 5.9). Se todos os riscos são considerados baixos (região branca da tabela), o processo tem poucas chances de prejudicar significativamente a coleção e, com os devidos cuidados, a migração pode ser executada. Se alguns riscos são considerados altos (região cinza-clara da tabela), a migração tem grandes chances de causar adulteração dos dados, e recomenda-se adiá-la até que a probabilidade de ocorrência desses eventos diminua. Se algum dos riscos é considerado inaceitável (região cinza-escura da tabela), o processo de migração deve ser descartado como mecanismo de preservação dos dados digitais.

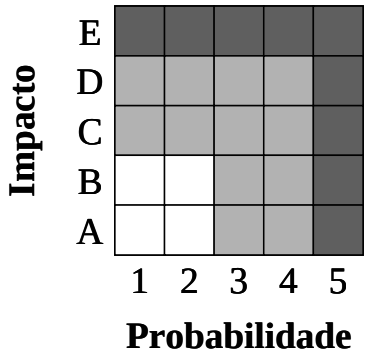


Figura 5.9: Tabela de decisão para os riscos de migração. A região branca indica que a migração pode prosseguir; a região cinza-clara indica que é melhor adiar a migração; a região cinza-escura indica que a migração não deve ser conduzida

5.2.3 Emulação

O processo de migração apresenta dificuldades para a preservação dos dados digitais. Além dos erros de tradução a que ela sujeita os dados, a migração é um processo cuja necessidade é impossível de prever com antecedência, uma vez que as mudanças de paradigma da indústria de informática seguem uma agenda completamente estranha ao interesses das instituições de custódia. Mais do que isso, a migração é um procedimento que envolve um certo senso de urgência, pois se um documento não é convertido em tempo hábil, após uma mudança de paradigma, pode ser impossível fazê-lo mais tarde, quando seu *software* ou *hardware* já se houverem tornado obsoletos [ROTHENBERG 98].

Um processo que pode ser usado em conjunto ou em substituição à migração, é a emulação, que propõem manter os documentos em seus formatos originais, enquanto cria ambientes operacionais que simulam, nas novas plataformas, o comportamento daquelas que se tornaram obsoletas [BESSER 00].

Migração e emulação podem ser utilizadas em conjunto, num esforço de ao mesmo tempo prover os documentos num formato mais acessível e conveniente, e resguardar seus originais em toda a sua essência. Em analogia com as técnicas convencionais de preservação, onde embora seja útil fazer traduções para novas línguas, ou conversão de formato em documentos históricos, todo esforço é feito no sentido de preservar os originais, não só pela possibilidade de verificar se o significado foi adulterado (proposital ou inadvertidamente) na transcrição, mas também por razões acadêmicas e estéticas [ROTHENBERG 98].

Para que um documento digital se torne legível, uma cooperação bastante complexa entre componentes deve tomar parte (Figura 5.10). A seqüência de *bits* precisa ser lida do seu suporte físico, através de um dispositivo de leitura. O dispositivo, para ser utilizado, requer um suporte de *software* chamado *acionador de dispositivo*, que faz a interface entre ele e o sistema operacional. O sistema operacional é executado sobre uma plataforma de *hardware* e coordena aplicativos e dispositivos, permitindo que os primeiros sejam executados e que acessem as funcionalidades fornecidas pelos segundos. Por fim, um programa aplicativo precisa interpretar os *bits* do documento, tornando-o visível e utilizável.

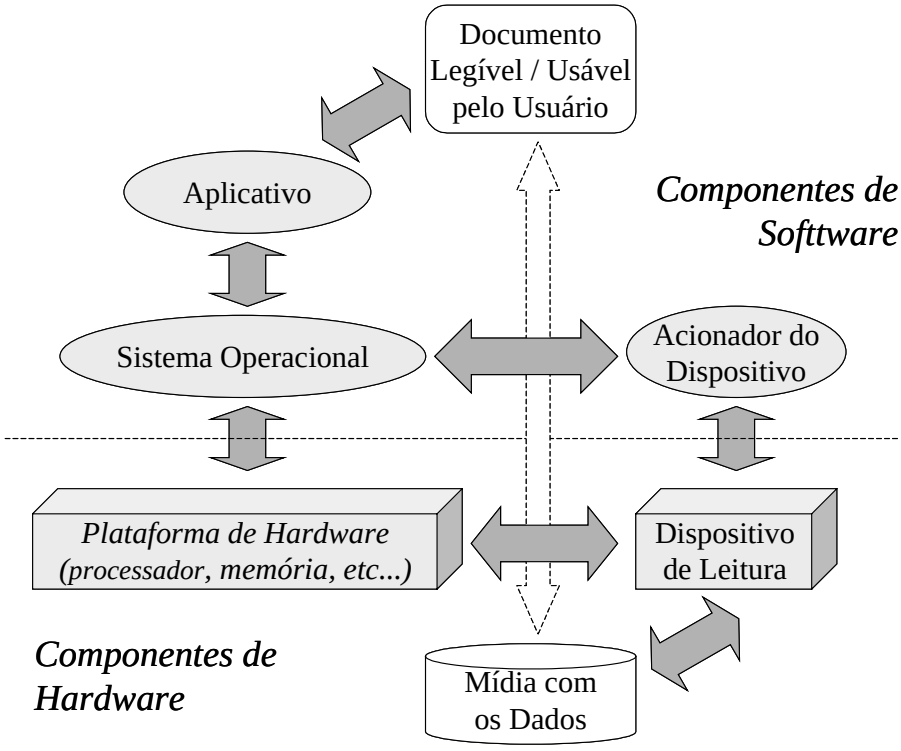


Figura 5.10: Para que o usuário possa acessar o conteúdo de um documento (seta pontilhada) uma complexa interação entre componentes deve tomar parte (setas cheias).

É teoricamente possível emular tanto os componentes de *software* quanto os de *hardware*, mas por várias razões, reproduzir o comportamento do aplicativo, ou do sistema operacional são tarefas muito mais árduas do que emular o *hardware* subjacente. O *hardware*, quase sempre, é mais simples do que o *software*, mas mais importante do que isso, suas interfaces são normalmente muito mais bem especificadas e documentadas por seus fabricantes [ROTHENBERG 98].

Numa solução de emulação, o primeiro esforço é salvaguardar o conteúdo físico das mídias originais, o que é feito através do refrescamento. É preciso também tomar todas as providências para salvar os arquivos digitais que compõem o sistema operacional, o programa aplicativo, acionadores de dispositivo, etc… A plataforma de *hardware* é então simulada, através de um programa que é executado no novo sistema e *hardware* correntes. O simulador emula todos os comportamentos do *hardware* original, inclusive fazendo a ponte para a seqüência de *bits* preservada por refrescamento.

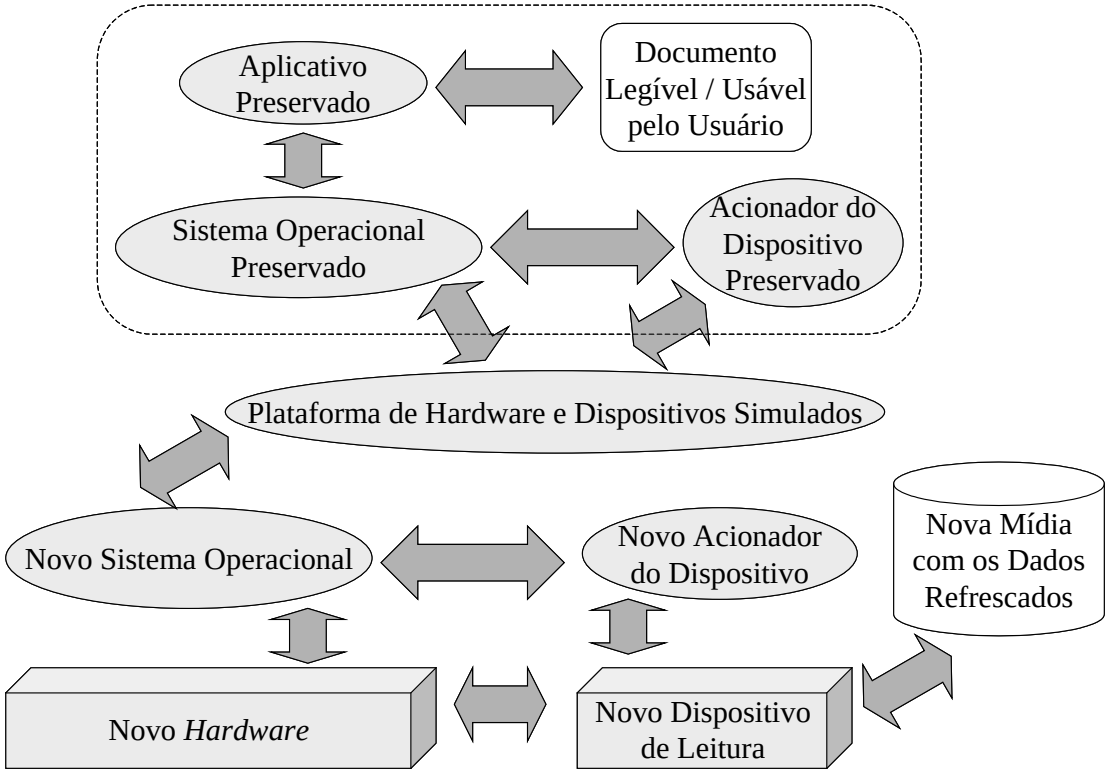


Figura 5.11: Num sistema emulado, um simulador cria um ambiente virtual (retângulo pontilhado) que permite executar os *softwares* necessários para interpretar o documento. O simulador também faz a ponte entre os dados salvaguardados e o ambiente emulado.

O processo de emulação tem um fluxo de trabalho e uma estrutura de custos distinta do processo de migração. Enquanto este último tem seus custos bastante variáveis, dependendo do volume e diversidade de formatos do acervo original, o último apresenta um custo fixo, relacionado ao desenvolvimento do emulador (acrescido naturalmente, dos custos de refrescamento, que são variáveis). Entretanto, criar um emulador é uma tarefa tecnicamente muito mais complexa, e cara, do que desenvolver um conversor entre formatos de arquivos, ou utilizar os conversores disponíveis no mercado, quando disponíveis.

Como um único emulador de *hardware* permite preservar a legibilidade de todos os formatos de arquivos executáveis sobre aquela plataforma, é possível que as instituições de custódia dividam os custos de seu desenvolvimento [ROTHENBERG 98].

5.2.4 Arqueologia digital

Refrescamento, migração e emulação são medidas preventivas de conservação da informação digital. Quando as mídias originais já se degradaram além da legibilidade em seus dispositivos originais, ou a tecnologia de interpretação dos arquivos já foi há muito esquecida

é preciso tomar ações corretivas para preservar, tanto quanto possível, os registros digitais remanescentes. Esse esforço, numa analogia à delicadeza dos seus métodos — e à imprevisibilidade dos seus resultados — é chamado de *arqueologia digital*.

Ao tomar contato com mídias degradadas, ou arquivos não reconhecíveis, não se deve sumariamente descartar esse conteúdo. Alguns procedimentos simples muitas vezes conseguem uma nova leitura de disquetes, fitas magnéticas e discos ópticos deteriorados, permitindo transferir seu conteúdo para um suporte mais seguro. As técnicas consistem desde o uso de estufas para fitas com síndrome-da-fita-grudenta, até a microscopia magnética, passando pela tentativa de ler as mídias em dispositivos de qualidade industrial, normalmente mais robustos e tolerantes do que os de qualidade comercial.

Contudo, a não ser que os registros sejam realmente valiosos, os custos desses procedimentos podem se revelar um impedimento insuperável. Como na conservação de materiais convencionais, as medidas preventivas são muito preferíveis às corretivas, e têm um custo muito mais baixo.

Ross e Gow fazem uma discussão bastante completa sobre arqueologia digital, discutindo desde a recuperação de mídias danificadas, até o uso de técnicas que facilitam a interpretação de dados de formatos obsoletos. Eles relatam, dentre outros estudos de caso, como as fitas magnéticas que registravam o comportamento do ônibus espacial *Challenger* puderam ser resgatadas, mesmo depois de imersas em água do mar; e como importantes registros eletrônicos do governo americano puderam ser salvos, após a passagem do furacão *Marilyn* pelas Ilhas Virgens, em 1995 [ROSS 99].

5.3 Outros Fatores

5.3.1 Destruição acidental

A fragilidade dos suportes e a extrema precisão dos mecanismos de leitura sujeitam os dados digitais à possibilidade de destruição acidental em uma escala muito maior que os convencionais. É certo que água, fogo e fumaça, poluentes, choques e vibrações, são danosos para qualquer acervo, mas em coleções digitais, um pouco de umidade, uma elevação

de temperatura acima de 80º C, ou a fuligem provocada por um incêndio são suficientes para provocar perdas extensas, e muitas vezes, irrecuperáveis.

À parte dos mecanismos de prevenção, e da tomada correta e imediata de procedimentos corretivos na ocorrência de incêndios, enchentes e outros desastres, a única maneira realmente confiável, e de custo compatível para a prevenção da destruição acidental de acervos digitais é o uso da redundância.

O fato de os dados serem perfeitamente replicáveis, permite que haja mais de uma cópia das matrizes digitais. É extremamente importante, que essas cópias sejam feitas com regularidade, inclusive mantendo uma réplica dos dados fora do edifício da instituição.

5.3.2 Sabotagem, vandalismo e destruição intencional

Os dados digitais também ficam mais expostos à destruição intencional do que os convencionais. Um imã forte, escondido no bolso, é suficiente para que um funcionário vingativo — ou subornado para suprimir evidências — arruíne um arquivo de fitas magnéticas.

Com o advento das redes de computadores, e a interligação completa dos sistemas, a possibilidade de ataques externos fica grandemente ampliada. Além das falhas intrínsecas dos sistemas de segurança, provocadas por erros nos *softwares*, usuários ilegais contam com diversos métodos para conseguir com que os legítimos revelem, inadvertidamente, suas senhas (no que é chamado, no jargão da segurança de sistemas eletrônicos, de *engenharia social*) [TENENBAUM 96].

Certas medidas para proteção do acervo contra ataques, embora óbvias, muitas vezes são desconsideradas. O acesso físico aos computadores centrais (servidores) e aos depósitos de dados digitais deve ser rigorosamente controlado. Os usuários devem ser constantemente educados sobre senhas, autenticação e demais procedimentos de segurança da rede, para que não se tornem alvos fáceis da engenharia social. Os *softwares* devem ser sempre atualizados e deve-se prestar especial atenção aos boletins de sistemas operacionais e aplicativos, que são publicados regularmente, documentando novas falhas de segurança descobertas e prescrevendo medidas para sua correção.

As cópias de segurança são um recurso importante para corrigir os efeitos da destruição intencional, quando ela ocorre. Por isso mesmo, especial atenção deve ser tomada

para que as mídias contendo as cópias de segurança não sejam, elas também, sujeitas à ação dos sabotadores.

5.3.3 Falsificação e adulteração intencional

A informação digital possui facilidades de modificação que comprometem a segurança de sua integridade. Mesmo dados armazenados em mídias não alteráveis, como CD-ROMs e CD-Rs, são passíveis de sofrerem adulterações nos processos de refrescamento e migração. Para garantir a preservação, é preciso criar mecanismos para que os artefatos não possam ser modificados ou suprimidos sem deixar evidências.

Para resolver esse problema, existe uma gama de mecanismos, derivados das técnicas de criptografia, que associam aos objetos digitais marcas e assinaturas que garantem sua autenticidade. Entretanto, faltam ainda normas e padrões bem difundidos para que o uso dessas técnicas seja efetivo, e não ameace as perspectivas de preservação do conteúdo dos dados a longo prazo [TFADI 96].

Tenenbaum descreve, detalhadamente, diversos esquemas de criptografia, baseados em chaves privadas e públicas, e como esses esquemas podem ser adequados para prover assinaturas digitais [TENENBAUM 96].

5.4 Conclusões do capítulo

São três as dimensões a considerar no planejamento para a preservação da informação digital: o desgaste dos suportes, a obsolescência da tecnologia e os fatores humanos/ambientais.

A primeira, embora não deva ser negligenciada, é provavelmente a menor das preocupações. Não só mídias cada vez mais longevas estão sendo propostas, quanto a mais simples das operações de preservação digital — o refrescamento entre mídias de mesma natureza — basta para combater os efeitos do desgaste.

A segunda dimensão é muito mais séria, uma vez que os suportes atuais comumente sobrevivem muitos anos a mais do que sua própria tecnologia. Soma-se a ela o fato de que não basta ter os dispositivos de leitura/gravação das mídias para garantir a sobrevivência da informação que elas contém: é preciso tornar o arquivo inteligível por seres humanos, o que em geral depende de delicados sistemas de *hardware* e *software*.

A dimensão de mais difícil previsibilidade, entretanto, continua sendo a humana. A freqüência com que são noticiadas violações de acesso em sistemas supostamente "à prova de falhas" provoca um certo ceticismo de que seja possível garantir, apenas através de medidas de controle, a segurança dos grandes repositórios de informação. Considerando ainda a possibilidade de desastres como incêndios, inundações, falhas de refrigeração, quebra de encanamentos e outros eventos capazes de arruinar em poucos minutos todo um acervo digital, a criação de réplicas (cópias de segurança) adeqüadamente mantidas e atualizadas, com uma cópia fora do edifício da instituição, e procedimentos bem documentados de resgate, parece ser a única alternativa realmente segura de se precaver contra esses fatores.

6 Gestão Documental e Fluxo de Trabalho

6.1 Metadados

O termo *metadado* significa, literalmente, *dado acerca do dado*, e é utilizado para denotar os atributos associados a um documento digital. Os metadados adicionam valor ao documento, fornecendo informações importantes acerca do seu conteúdo, formato e história administrativa (**Tabela 6.1**). Um definição ampla de metadado seria "qualquer conhecimento a respeito de um objeto de informação, em qualquer nível de agregação destes". Nesse contexto, objeto de informação é qualquer artefato que possa ser manipulado por uma pessoa ou sistema como uma entidade discreta [GILL 98].

Além de terem um papel fundamental no acesso aos dados digitais, os metadados contribuem para sua preservação, registrando informações acerca do seu formato, contexto e dependências de *software* [BESSER 00].

Várias iniciativas têm procurado normalizar a descrição de metadados, de forma a permitir às instituições o compartilhamento das informações acerca dos seus acervos.

A mais antiga dessas iniciativas talvez seja o MARC, criado no final da década de 1960, para permitir a representação eletrônica dos catálogos de bibliotecas [TLOC 02a]. O MARC tornou-se um sistema complexo, com centenas de campos possíveis, englobando diversos tipos de metadados, entre descritivos, administrativos e de preservação.

No extremo oposto o DCMI — *Dublin Core Metadata Initiative*, procurou normalizar os metadados descritivos, pautando-se pela simplicidade. O *Dublin Core* é formado por um conjunto de apenas 15 metadados, que procuram capturar apenas os dados essenciais para a descrição de um objeto informacional. O *DC* foi concebido principalmente para o uso em recursos digitais e seus grandes repositórios [DCMI 03].

Tabela 6.1: Categorias de metadados

Tipo	Descrição	Exemplos de Metadados
Administração	Metadados usados na gerência e administração de recursos de informação	• Informação de aquisição • Acompanhamento de direitos e reprodução • Documentação de requisitos legais para acesso • Informação sobre localização • Critério de seleção para digitalização • Controle de versões e diferenciação entre objetos de informação similares • Auditorias criadas por sistemas de arquivamento
Descritivo	Metadata usados para descrever ou identificar recursos de informação	• Catálogos e registros de catálogos • Instrumentos de pesquisa • Índices especializados • Relacionamento de hiperligação entre recursos • Anotações feitas por usuários • Metadados para sistemas de arquivamento gerados pelos criadores dos artefatos
Preservação	Metadadados relacionados à gestão de preservação dos objetos de informação	• Documentação das condições de conservação dos artefatos físicos • Documentação das ações tomadas para rpeservar as versões digitais e físicas do recurso, e.g., referescamento e migrações
Técnico	Metadados relacionados aos comportamentos do sistema, incluindo suas funções e dados	• Documentação de hardware e software • Informações de digitalização, e.g., formatos, compressão, processo de escaneamento • Acompanhamento do desempenho do sistema • Dados de atenticação e segurança, e.g., chaves de criptografia, senhas
Uso	Metadados relacionados	• Registro de exibições • Acompanhamento de uso e usuários • Informações sobre reúso de conteúdo e múltiplas versões information

FONTE — [GILL 98]

A experiência do projeto *Making of America II* revelou a importância dos metadados estruturais e de comportamento, em adição aos descritivos e administrativos. Baseado nas conclusões do trabalho, a *Digital Library Federation* promoveu uma especificação de como um conjunto completo de metadados pode ser descrito em um arquivo XML, para troca entre repositórios de informação. Esse esquema foi chamado de METS – padrão para codificação e transmissão de metadados [TLOC 02b].

Um outro padrão de metadados expresso em XML é o EAD — descrição arquivística codificada, criado especialmente para atender às demandas de descrição multinível da comunidade de Arquivos. O EAD, permite a representação dos instrumentos de pesquisa arquivísticos de forma padronizada, de forma a permitir seu compartilhamento [TLOC 03c] [TSAA 00a] [TSAA 00b].

Para garantir que os metadados descritivos pudessem ser representados de forma adequada nos documentos da *World Wide Web*, o W3C promoveu um padrão denominado RDF — formato para descrição de recursos. O RDF usa um esquema extensível, permitindo a adição de novos tipos de metadados [MILLER 03].

6.2 Gestão documental

Sistemas de gestão documental oferecem funcionalidades como classificação, edição, armazenamento, transmissão e mecanismos de consulta aos documentos. Além disso, o sistema deve capturar a estrutura lógica dos documentos. Isso pode ser feito através de um catálogo de *tipos de documentos*, compondo uma hierarquia de generalização, composta de uma classe geral seguida por sucessivas especializações [FERREIRA 93].

Sistemas de gestão de documentos aparecem nos mais variados contextos, como instituições financeiras, acadêmicas, empresariais. Entretanto, um SGD para acervos permanentes deve conciliar as necessidades imediatas com as possibilidades de uso futuro do acervo. Assim, é preciso fornecer meios para a indexação e descrição dos itens, segundo as mais recentes normas internacionais, mas também facilidades para a gestão dos dados digitais e garantia de sua permanência.

6.3 Fluxo de trabalho

Nos trabalhos das instituições de custódia, tanto com os acervos convencionais, quanto com os digitais, freqüentemente surge a necessidade de haver procedimentos bem documentados e com bom suporte ferramental para as mais diversas atividades: reformatação, indexação, restauração, descrição dos itens, etc…

Como analisamos na Seção 3.2 o grau de planejamento, cuidado operacional e controle de qualidade na digitalização de acervos para fins permanentes, está muito acima daquele necessário para documentos comuns, de instituições comerciais. Sem ferramentas que dêem suporte às múltiplas e complexas decisões do arquivista, a reformatação de um acervo volumoso pode rapidamente tornar-se inviável.

Além da digitalização, muitas outras atividades estão envolvidas na gestão de documentos permanentes, incluindo a aquisição de metadados, a geração de vocabulários controlados, a indexação dos documentos, a criação de coleções, a criação de instrumentos de pesquisa, o acompanhamento de esforços de higienização, microfilmagem e restauração do acervo, etc…

Sistemas de fluxo de trabalho (do inglês, *workflow*) permitem a gestão controlada de atividades complexas, documentando procedimentos, acompanhando a execução das tarefas e emitindo relatórios de *status*. Os mais sofisticados permitem modelagem e reengenharia de processos, especificação detalhada das etapas que compõem os processos, controle e coordenação da execução das tarefas, e monitoramento e supervisão dos resultados [CICHOCKI 98].

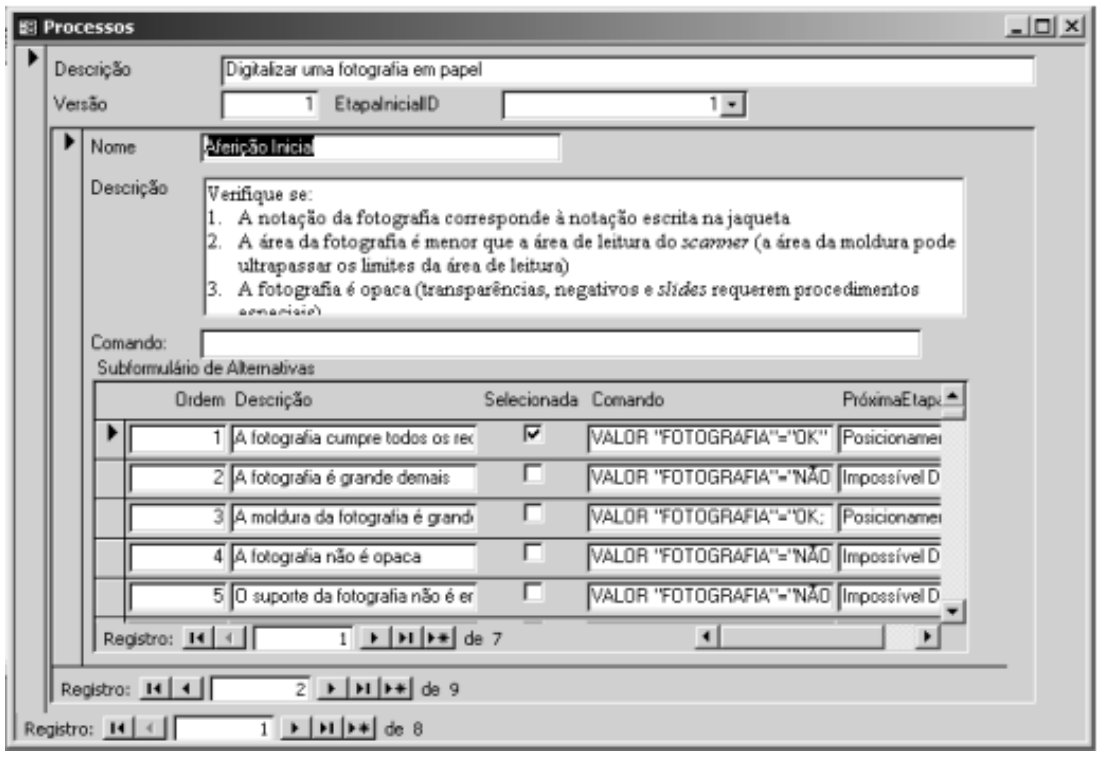


Figura 6.1: Tela de descrição de processo do subsistema de fluxo de trabalho implementado no Arquivo Público Mineiro

Os sistemas de gestão de fluxo de trabalho armazenam uma representação dos processos da organização, automatizando-os no que for possível, e distribuindo as tarefas para os usuários. O sistema também controla o fluxo de informações, registrando metadados administrativos, e encaminhando documentos entre usuários e repositórios. (Figura 6.1, Figura 6.8)

6.4 Modelo para sistemas de informação em acervos permanentes

Uma grande aliada na tarefa de desenvolvimento do sistema de informação é a orientação por objetos. Sistemas complexos de gestão de documentos e fluxo de trabalho modelam-se muito mais fácil e naturalmente usando os novos paradigmas de objetos que as velhas abordagens estruturadas da década de 70.

As abordagens orientadas por objetos permitem uma descrição conceitual do sistema de informação, e integram de forma extremamente natural os conceitos de generalização e especialização através do mecanismo de herança. Utilizando análise, desenho e programação OO, a arquitetura de documentos pode ser mantida extremamente flexível.

Hoje a indústria reconhece a utilidade das metodologias OO, que passaram de curiosidade acadêmica à prática *de facto*. O sucesso comercial da linguagem e da plataforma *Java*, da *Sun Microssystems* e da linguagem de modelagem *UML* [BOOCH 99] atestam a popularidade da metodologia.

A *UML*, ou *Unified Modelling Language*, é o resultado da convergência de uma série de metodologias orientadas por objetos. A *UML* propicia a diminuição do chamado *gap semântico* entre as etapas de análise, desenho e implementação do *software*, provendo representações diagramáticas do sistema através de *modelos*. Esses modelos procuram capturar, de forma ao mesmo tempo precisa e intuitiva, todos os aspectos estruturais e comportamentais do *software* que eles descrevem. Neste trabalho, a *UML* foi amplamente utilizada tanto na análise quanto na construção do sistema de informação proposto.

6.4.1 Gestão do acervo digital

As funcionalidades de acervo digital compõem o núcleo do sistema de informação, incluindo as facilidades para inclusão de documentos e seus metadados, vocabulários controlados, consultas ao acervo e geração de instrumentos de pesquisa. Incluem-se aqui também as funcionalidades necessárias para representação do arranjo dos documentos.

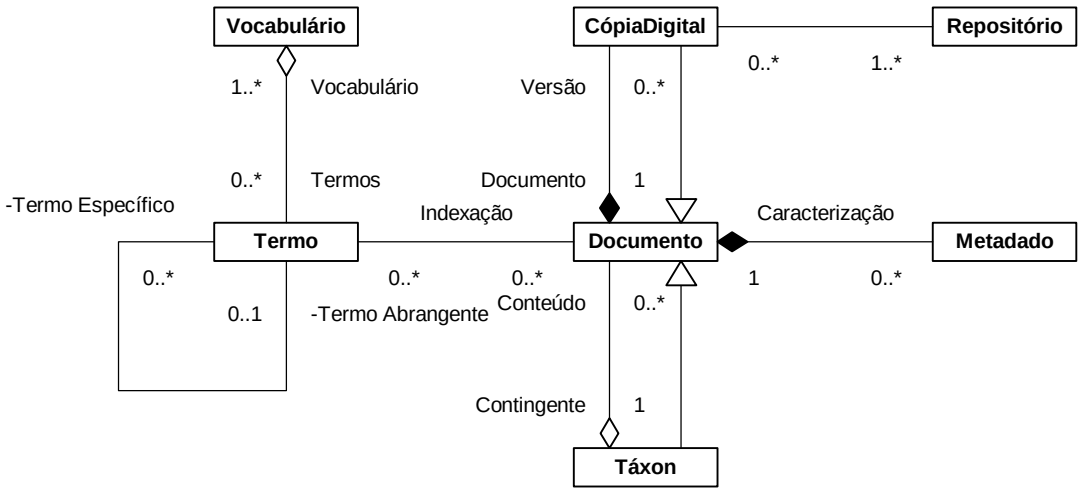


Figura 6.2: Diagrama das classes envolvidas na gestão do acervo digital (em *UML*)

No diagrama de classes da Figura 6.2 é possível apreender a estrutura da solução do sistema de informação para o tratamento do acervo digital. A unidade central é composta pelos *documentos*, associados aos *metadados* que os caracterizam. Para cada documento são então geradas zero ou mais *cópias digitais*, que ficam guardadas em diversos *repositórios* (em memória secundária ou terciária). Os *termos descritores*, agrupados *vocabulários controlados*, são então associados aos documentos para permitir as operações de busca, na tarefa de *indexação*. Por fim, um sistema de *táxons* (classes) hierárquicos permite a representação do arranjo das coleções (repare que os táxons em si são considerados documentos, através da relação de generalização da *UML*).

6.4.2 Gestão da arquitetura de documentos

A gestão da arquitetura de documentos permite formalizar a estrutura lógica dos documentos tratados pelo sistema informação, catalogando os diversos tipos de documentos, seus atributos e sua hierarquia de generalização/especialização.

O sistema de informação permite muita flexibilidade na definição dos tipos de documentos e seus metadados, de forma a garantir a sobrevivência da aplicação face às necessidades mutantes do acervo.

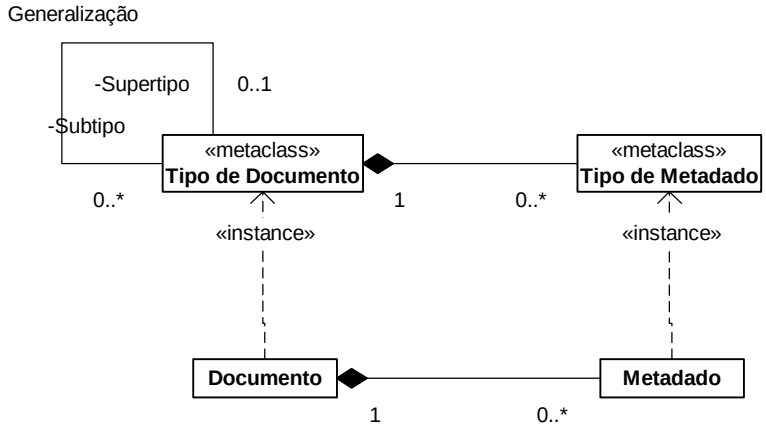


Figura 6.3: Diagrama das classes envolvidas na gestão da arquitetura de documentos (em *UML*)

6.4.3 Fluxo de trabalho

O fluxo de trabalho é gerido de forma a permitir o planejamento, controle e acompanhamento das atividades que envolvem os documentos do acervo. O sistema de informação armazena uma descrição detalhada (roteiro) dos procedimentos a serem utilizados em cada atividade documental, e no decorrer da execução dos procedimentos, registra todos os eventos relevantes, de forma a permitir rastrear o *status* das atividades, identificar possíveis problemas ou mesmo rastrear um histórico das manipulações sofridas pelos artefatos que compõem as coleções.

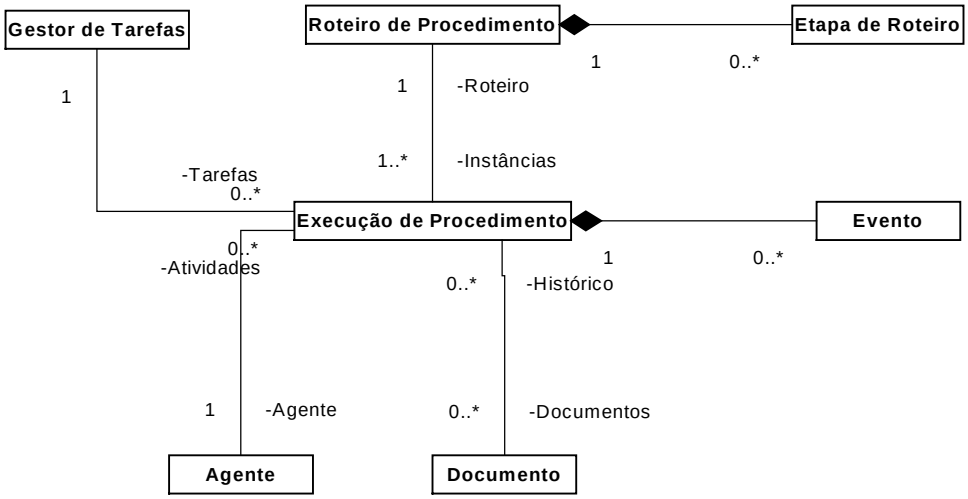


Figura 6.4: Diagrama das classes envolvidas no controle do fluxo de trabalho (em *UML*)

6.4.4 Estrutura da Aplicação

A estrutura simplificada do sistema de informação multimídia pode ser vislumbrada através do diagrama da Figura 6.5, que demonstra como a implementação em camadas procura isolar o núcleo da aplicação dos mecanismos de persistência e apresentação.

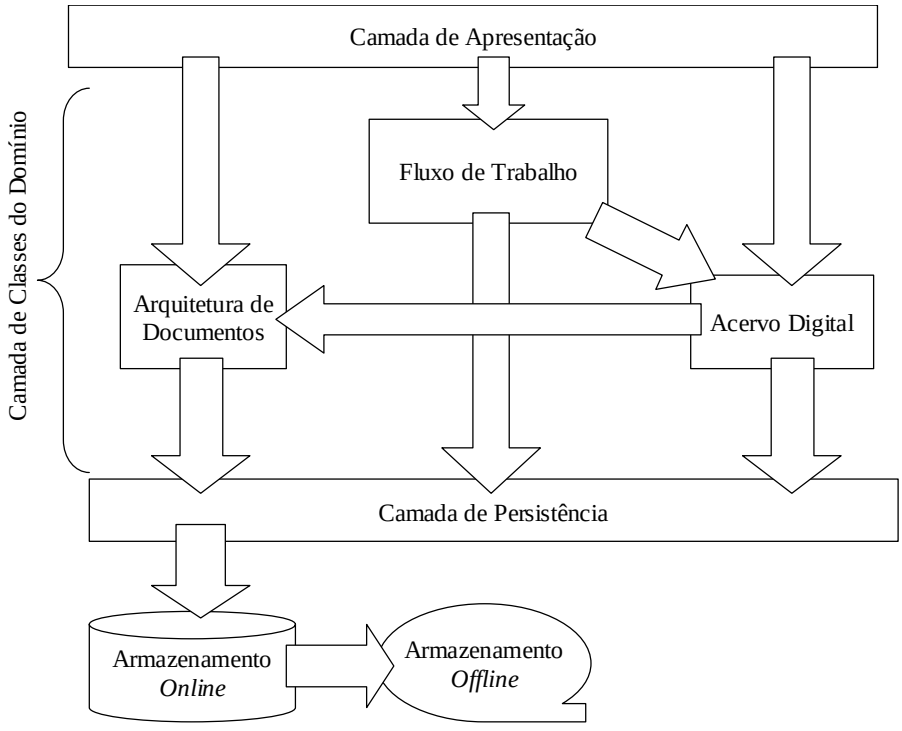


Figura 6.5: Estrutura do sistema de informação multimídia (as setas indicam dependência funcional, não fluxo de dados)

A camada de persistência encapsula diversos mecanismos de persistência: um banco de dados relacional, acesso direto a arquivos do sistema operacional e acesso à *streaming media* (áudio e vídeo sob demanada). O trabalho da camada de persistência foi baseado nas diretrizes de Ambler [AMBLER 00a] [AMBLER 00b].

6.5 Implantação do sistema no Arquivo Público Mineiro

6.5.1 Planejamento das atividades

As atividades de implantação do sistema foram divididas em quatro linhas principais: desenvolvimento da aplicação, tratamento do acervo, implantação da infra-estrutura de apoio e implantação de uma política de gestão tecnológica para a instituição do Arquivo. O planejamento procurou explorar as possibilidades de paralelismo entre as atividades, que podem ser vistas no diagrama da Figura 6.6.

O planejamento da construção do sistema seguiu as etapas prescritas pelo *PRAXIS*, um processo de *software* concebido para aplicações interativas: *ativação*, *levantamento de requisitos*, *análise*, *planejamento*, *desenho de alto nível*, *desenho de testes*, *implantação das liberações executáveis* (que contempla *desenho detalhado*, *codificação* e *testes de unidade*), *testes alfa* (testes executados no ambiente do desenvolvedor), *testes beta* (testes executados no ambiente definitivo) e *operação piloto* (operação "vigiada" do sistema). Os detalhes precisos dessas operações e dos artefatos produzidos e consumidos em cada uma delas podem ser verificados em [PAULA 01].

Tão ou mais complexa que a construção do sistema é a preparação dos dados que serão manipulados por ele. Só para esta etapa do projeto, estão sendo digitalizadas mais de 14.000 fotografias, dispersas em diferentes tamanhos e formatos. Procuramos analisar a realidade do acervo, identificando o perfil de tamanhos das fotografias. Além disso, procuramos estimar o volume total dos dados digitalizados, de forma a dimensionar o *hardware* de armazenamento secundário (discos magnéticos, *RAID*, etc.) e terciário (fitas magnéticas, *etc.*). Logo após a implantação da primeira fase da infra-estrutura tecnológica,

seguiram em paralelo os trabalhos de digitalização das imagens, e de indexação dos documentos e digitação dos metadados.

De forma a prover as funcionalidades requeridas com um grau aceitável de desempenho, uma plataforma de *hardware* e *software* robusta é necessária. Além dos equipamentos que sustentam o sistema de informação propriamente dito, foi preciso adquirir, instalar e manter *scanners* para aquisição de imagens (em negativos ou cópias).

A tecnologia digital requer gestão permanente, para evitar os problemas trazidos pela obsolescência, procedimentos inadequados e riscos de perdas de dados. Uma das atividades contempladas pelo projeto trata da implantação de uma política de gestão permanente dos acervos digitais, que visa a determinar responsabilidades sobre as operações de conservação do acervo, bem como um planejamento estratégico de longo prazo para a tecnologia no Arquivo.

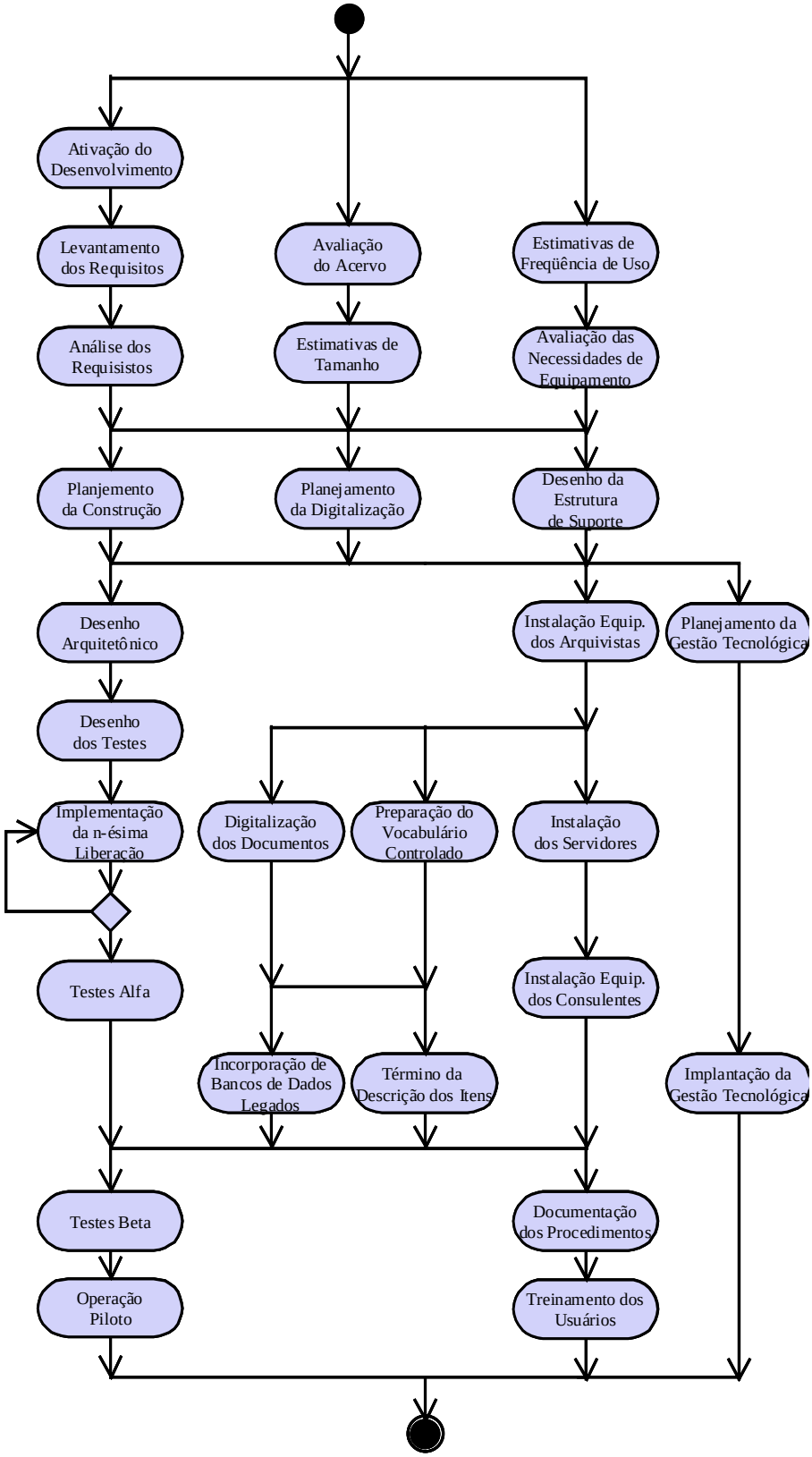


Figura 6.6: Planejamento das atividades de implantação do sistema de informação no Arquivo Público Mineiro
FONTE — [VALLE 02a]

6.5.2 Descrição e indexação

Descrição e indexação são duas das tarefas mais trabalhosas no trabalho com acervos. Na implantação do nosso sistema, essas etapas foram cumpridas paralelamente à digitalização das fotografias. O sistema de informação possui um módulo para a coleta de metadados de descrição, que incluem os atributos do documento e os termos de indexação, baseados no vocabulário controlado. (Figura 6.7)

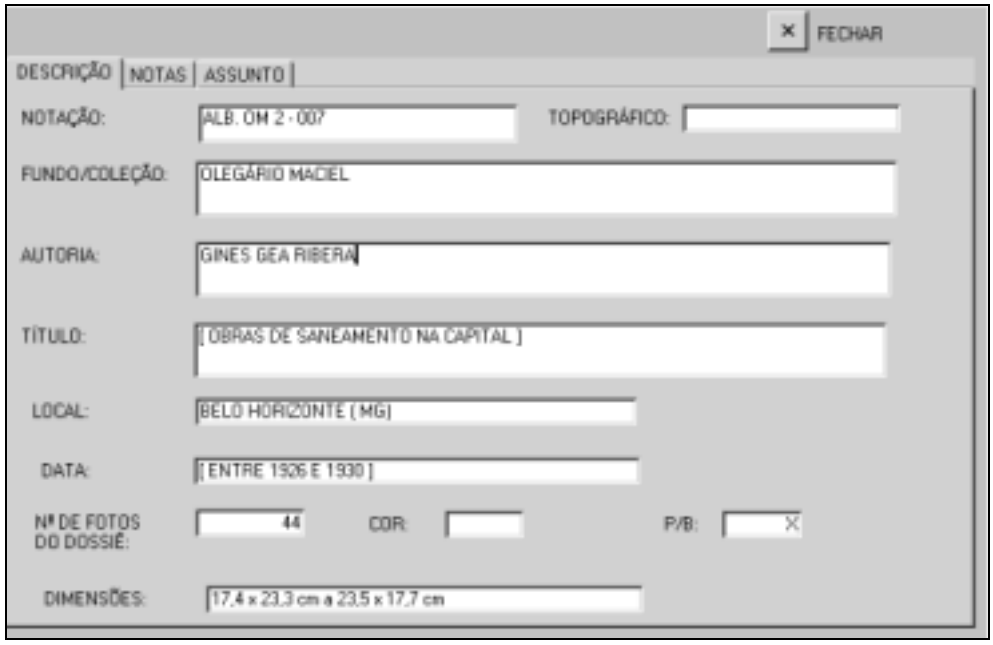


Figura 6.7: Tela de entrada dos metadados de descrição do subsistema de gestão documental implementado no Arquivo Público Mineiro

6.5.3 Digitalização

Para digitalização do acervo fotográfico, adquirimos dois *scanners*: um para filmes e outro para as cópias. O *scanner* de filmes escolhido foi o *Kodak RFS 3570+*, capaz de trabalhar com os filmes de médio formato (6×6 cm e 6×7 cm) que compõem uma parte expressiva do acervo. Para as cópias, adquirimos um *scanner* de mesa, o *HP 6300c*, capaz de atingir a resolução de 2400 dpi.

A digitalização foi guiada por procedimentos que estabelecemos *a priori*, e cujo desempenho acompanhamos em uma primeira amostra para testes. As fotografias em cópia foram digitalizadas a 600 dpi, em cores, com 8 *bits* por cor. Os negativos, preto e brancos, foram digitalizados à resolução máxima do *scanner*, em escala de cinza.

Estabelecemos um fluxo de trabalho preciso para a digitalização, que incluía normas para pré-avaliação do material, nomeação dos arquivos, avaliação dos resultados, e coleta de metadados técnicos e administrativos. Para o sucesso dessa atividade, foi essencial o uso das facilidades de gestão do fluxo de trabalho, que permitiam simultaneamente orientar o operador passo-a-passo da digitalização, e coletar a maior parte dos metadados de forma automatizada (Figura 6.8) [VALLE 02b].

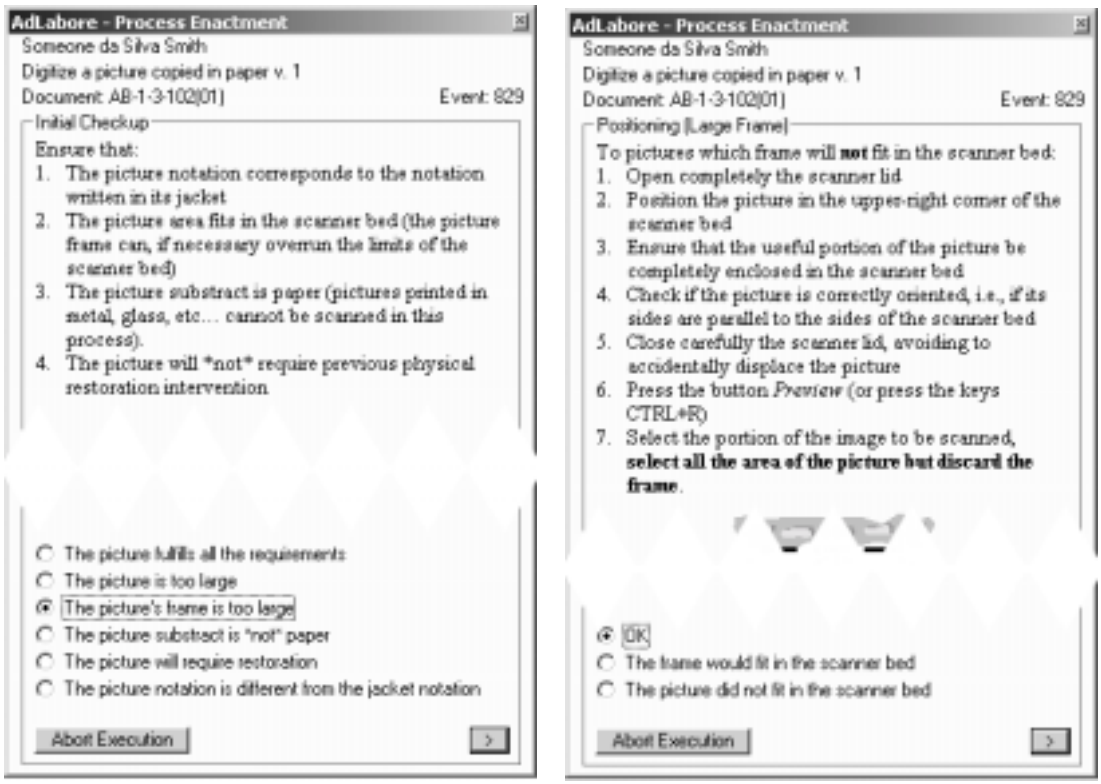


Figura 6.8: Duas telas sucessivas do subsistema de acompanhamento do fluxo de trabalho implementado no Arquivo Público Mineiro
FONTE — [VALLE 02b]

6.5.4 Interface com os consulentes

A disponibilização do acervo para os usuários do acervos envolveu a criação não apenas do sistema de informação, mas também de uma página da *web* que focalizasse os interesses dos diferentes públicos do arquivo: acadêmicos, produtores de documentação no Executivo do Estado, estudantes, *etc.* (Figura 6.9)

Figura 6.9: *Homepage* do Arquivo Público Mineiro
FONTE — [APM 03]

A consulta aos dados digitais foi então incorporada à esta página, como um dos serviços providos pelo arquivo (Figura 6.10).

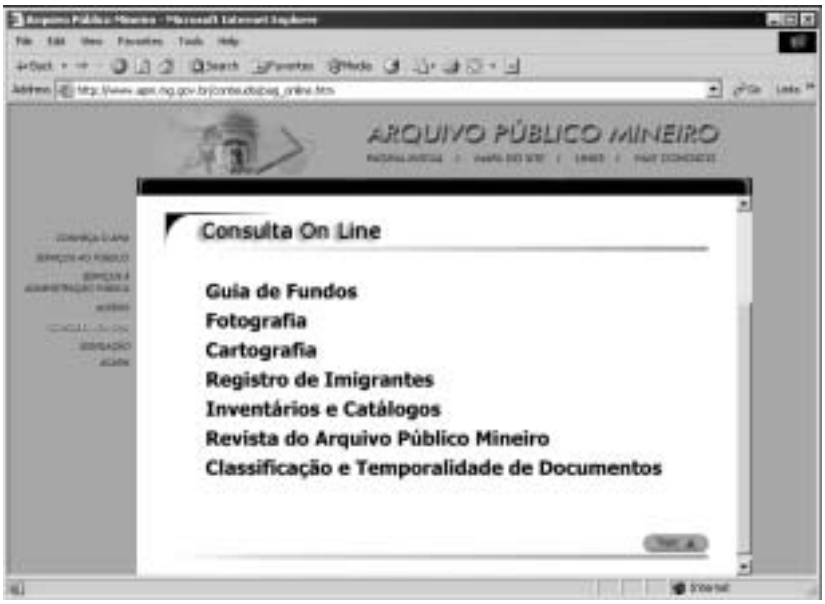


Figura 6.10: Serviços de consulta
FONTE — [APM 03]

(a)

(b)

(c)

Figura 6.11: Consulta ao acervo fotográfico: *a)* interface de pesquisa, *b)* lista de resultados; *c)* dados do resultado escolhido / FONTE — [APM 03]

Para localizar uma fotografia, o consulente pode fazer uma busca de texto-livre em todos os campos da fotografia, ou especificar seus critérios em relação aos metadados *Título*, *Autoria*, *Data* e *Local*. O consulente também fazer uma busca por *Assunto*, utilizando os termos de um vocabulário controlado. (Figura 6.11)

Figura 6.12: Visualização detalhada da imagem da fotografia
FONTE — [APM 03]

Uma vez localizado o documento desejado, é exibida uma lista completa dos seus metadados, e uma pequena miniatura da fotografia (Figura 6.11.c). O usuário pode então, se quiser, fazer uma visualização detalhada da imagem da fotografia. (Figura 6.12)

7 Conclusão

Abrangendo desde métodos de realce e restauração de imagens até novos paradigmas de bancos de dados complexos, a tecnologia digital traz novas e impressionantes possibilidades para o universo da preservação de acervos documentais. Entretanto, sua aplicação em artefatos de valor permanente deve ser conduzida com cuidado, acompanhada de uma política de gestão da tecnologia, sob pena de colocar o acervo à mercê da fragilidade da informação digital.

A preservação da informação digital é uma tarefa delicada, devido à própria natureza dessa informação e aos rápidos ciclos de obsolescência da tecnologia que a suporta. Entretanto, é possível estender a sua vida útil, identificando e endereçando os fatores que levam à sua degradação: principalmente a perda do seu contexto, e a perda da capacidade de visualizá-la.

Apesar do risco trazido por esses fatores, a digitalização de acervos é utilizada como meio ou coadjuvante de preservação, na reformatação de acervos. Nesses casos, é preciso prestar especial atenção aos critérios de qualidade da imagem digital gerada, e do seu formato de armazenamento.

A aquisição das imagens em alta qualidade gera arquivos muito grandes, que nem sempre são convenientes para se armazenar no sistema *online*, ou para serem utilizados diretamente. Mas é possível criar uma matriz de digitalização e derivar dela os arquivos de menor tamanho necessários. A matriz pode então ser guardada em fitas magnéticas ou discos ópticos, para fins de preservação, e os arquivos menores podem ser mantidos no sistema de informação para fins de acesso e reprodução.

O acesso é uma das dimensões mais importantes de todo projeto de preservação em meio digital, pois não se pode considerar a digitalização como meio de conservação se não for possível aos usuários encontrarem os dados de seu interesse em meio à enormidade do acervo.

Para que a pesquisa no acervo possa ser bem efetuada, é preciso descrever o conteúdo dos itens documentais e dos elementos do arranjo. Essa descrição freqüentemente é feita através de palavras-chave ou descritores agrupados em vocabulários controlados, que tem de ser geridos pelo sistema. Mesmo assim a consulta precisa ser feita não apenas por filtragem através de palavras-chave, mas também pela navegação através do arranjo.

O acesso às grandes bases de dados digitais é hoje alavancada pela quase ubiqüidade das redes de computadores.Considerando a base instalada e a taxa de crescimento atuais da *Internet*, pode-se afirmar, sem grande temeridade, que a penetração da informação digital será num futuro próximo tão grande quanto hoje é a da televisão. Essa possibilidade de dar acesso remoto a um grande público tem sido um grande estímulo para que arquivos, bibliotecas e museus procurem digitalizar seus acervos.

Os desafios da preservação dos acervos digitais são melhor compreendidos quando examinamos suas três dimensões: o desgaste dos suportes da informação digital, a obsolescência da tecnologia que permite esta informação se tornar inteligível pelos usuários, e o conjunto de fatores sócio/ambientais a que o acervo está sujeito (abrangendo desde riscos de desastres naturais, até problema de vandalismo e adulteração intencional). As três dimensões devem ser tratadas simultanea e satisfatoriamente para que um acervo digital tenha chance de sobrevivência a longo prazo.

Este trabalho permitiu experimentar algumas tecnologias relativamente recentes, como modelagem orientada por objetos, distribuição de informações através da *web*, processamento de imagens, consultas a sistemas de informação multimídia et cetera.

Foi também uma oportunidade interessante de aplicar os conceitos de engenharia de *software* em uma aplicação relativamente sofisticada (todo um sistema de *software* e *hardware* destinado a gerenciar um acervo documental e permitir sua ampliação), para garantir que a construção dos módulos programados viesse a atender as expectativas de qualidade da comunidade de usuários.

Por fim, e talvez mais importante, foi uma oportunidade ímpar de cooperar para a conservação da memória fotográfica de Minas Gerais e de abrir, possivelmente, o caminho para a implantação desse modelo de preservação em outras entidades.

“C’est une fable bien ancienne, bien universelle, que celle de la montagne qui, ayant effrayé tout le pays par ses clameurs en travail d’enfant, fut sifflée de tous les assistants, quand elle ne mit au monde qu’une souris. Le parterre n’était pas philosophe. Les siffleurs devaient admirer. Il était aussi beau à la montagne d’accoucher d’une souris, qu’à la souris d’accoucher d’une montagne. Un rocher qui produit un rat est quelque chose de très prodigieux; et jamais la terre n’a a vu rien qui approche d’un tel miracle. Tous les globes de l’univers ensemble ne pourraient pas faire naître une mouche. Là où le vulgaire rit, le philosophe admire; et il rit où le vulgaire ouvre de grands yeux stupides d’étonnement.”

Voltaire

8 Referências

8.1 Bibliografia

[ADLIB 03] AdLib Information Systems Ltd. *Homepage*, 2003.
Disponível na WWW: <http://www.adlibsoft.com/> 14/jan/2003.

[ADOBE 92] Adobe Developers Association. *TIFF Specification - Revision 6.0*. Adobe Systems Inc., Junho / 1992.
Disponível na WWW: <http://partners.adobe.com/asn/developer/pdfs/tn/TIFF6.pdf> 14/jan/2003.

[AMBLER 00a] AMBLER, Scott. *Mapping Objects To Relational Databases*. AmbySoft, 2000.
Disponível na WWW: <http://www. AmbySoft.com/mappingObjects.pdf> 14/jan/2003.

[AMBLER 00b] AMBLER, Scott. *The Design of a Robust Persistent Layer for Relational Databases*. AmbySoft, 2000.
Disponível na WWW: <http://www.AmbySoft.com/persistenceLayer.pdf> 14/jan/2003.

[ANDERSON 02a] ANDERSON, Dave. Storage: Tape. In: *PCTechGuide*, 2002.
Disponível na WWW: <http://www.pctechguide.com/15tape.htm> 14/jan/2003.

[ANDERSON 02b] ANDERSON, Dave. Storage: CD-ROM. In: *PCTechGuide*, 2002.
Disponível na WWW: <http://www.pctechguide.com/08cd-rom.htm> 14/jan/2003.

[ANDERSON 02c] ANDERSON, Dave. Storage: DVD. In: *PCTechGuide*, 2002.
Disponível na WWW: <http://www.pctechguide.com/10dvd.htm> 14/jan/2003.

[ANDERSON 03] ANDERSON, Dave. Storage: CD-R and CD-RW. In: *PCTechGuide*, 2003.
Disponível na WWW: <http://www.pctechguide.com/09cdr-rw.htm/> 14/jan/2003.

[ANDRADE 98a] ANDRADE, Nélson S. *Sistema de Informações Multimídia*. Dissertação de Mestrado. Belo Horizonte, Fundação João Pinheiro e Universidade Federal de Minas Gerais, 1998.

[ANDRADE 98b] ANDRADE, Nélson S. et al. A multimedia information system for governamental historical documents. In: Museums and the Web - An International Conference, 1998, Toronto, Canada. *Proceedings of (CD-ROM)…*

[APM 03] Arquivo Público Mineiro. *Homepage*, 2003.
Disponível na WWW: <http://www.apm.mg.gov.br/> 14/jan/2003.

[APP 03] Arquivo Público do Paraná. *Homepage*. 2003
Disponível na WWW: < http://www.pr.gov.br/arquivopublico/> 14/jan/2003.

[ARAÚJO 92] ARAÚJO, Arnaldo A. et al. An image data base system of brazilian historical documents. In: 2nd. International Conference on Automation, Robotics and Computer Vision, 1992, Singapore. *Proceedings of the… vol. 2*. p. cv-16.4.1.-5.

[ARAÚJO 93] ARAÚJO, Arnaldo A. et al. Image coding for storing historical documents in a database. In: Picture Coding Symposium - PCS, 1993, Lausanne, Switzerland. *Proceedings of the…* p.
Disponível na WWW: <http://www.npdi.dcc.ufmg.br/pesquisa/publicacoes/sol01.pdf> 14/jan/2003.

[BECK 97] BECK, Ingrid. (Coord.) *Reformatação*. Rio de Janeiro: Arquivo Nacional, 1997. 40 p. Disponível na WWW: <http://www.cpba.net/pdf_cadtec/44_47.pdf> 14/jan/2003.

[BESSER 00] BESSER, Howard. Digital Longevity. In: SITTS, Maxine (ed.). *Handbook for Digital Projects:* a management tool for preservation and access. Andover MA: Northeast Document Conservation Center, 2000. p.155-166.

[BINKLEY 39] BINKLEY, Robert C. Strategic Objectives in Archive Politics. *American Archivist*, p. 162-168, julho 1939. apud [CONWAY 97]

[BN 03] Biblioteca Nacional. *Homepage*. 2003.
Disponível na WWW: <http://www.bn.br/> 14/jan/2003.

[BOOCH 99] BOOCH, Grady, RUMBAUGH, James e JACOBSON, Ivar. *The Unified Modelling Language - User Guide*. Reading - MA, Addison-Wesley Longman Inc., 1999.

[BOSTON 00a] BOSTON, George and SCHÜLLER, Dietrich. Magnetic Carriers. In: [UNESCO 00].
Disponível na WWW: <http://webworld.unesco.org/safeguarding/en/all_magn.htm> 14/jan/2003.

[BOSTON 00b] BOSTON, George and SCHÜLLER, Dietrich. Optical Carriers. In: [UNESCO 00].
Disponível na WWW: <http://webworld.unesco.org/safeguarding/en/all_opti.htm> 14/jan/2003.

[BRAIN 03a] BRAIN, Marshall. How Hard Disks Work. *How Stuff Works*, HowStuffWorks Inc., 2003.
Disponível na WWW: <http://www.howstuffworks.com/hard-disk.htm> 14/jan/2003.

[BRAIN 03b] BRAIN, Marshall. How CDs Work. *How Stuff Works*, HowStuffWorks Inc., 2003.
Disponível na WWW: <http://www.howstuffworks.com/cd.htm> 14/jan/2003.

[CAND 94] Comissão Ad-hoc para as Normas de Descrição. *ISAD(G) - Normas gerais internacionais de descrição em arquivo*. Estocolmo, Conselho Internacional de Arquivos, 1994.
Disponível na WWW: <http://www.arquivonacional.gov.br/pub/virtual/virtual_cp.htm#> 14/jan/2003.

[CAND 95] Comissão Ad-hoc para as Normas de Descrição. *ISAAR (CPF) - Norma Internacional de Registro de Autoridade Arquivística para Entidades Coletivas, Pessoas e Famílias*. Paris, Conselho Internacional de Arquivos, 1995.
Disponível na WWW: <http://www.arquivonacional.gov.br/download/ISAAR.zip> 14/jan/2003.

[CICHOCKI 98] CICHOCKI, Andrzej et al. *Workflow and Process Automation: Concepts and Technology*. Kluwer Academic Publishers, Boston, 1998.

[CLIR 98] Council on Library and Information Resources. The Making of America, Part 2. *DLF Standards and Practices*. Digital Library Federation, 2001.
Disponível na WWW: < http://www.diglib.org/standards/dlfmoaii.htm> 14/jan/2003.

[CLIR 00] Council on Library and Information Resources. *Guides to Quality in Visual Resource Imaging*. Research Libraries Group, 2000.
Disponível na World Wide Web <http://www.rlg.org/visguides/index.html> 14/jan/2003.

[CONLAN 88] CONLAN, Roberta. (Coord.). *A Revolução dos Computadores*. São Paulo, Nova Cultural, 1988. 64p. (Entenda o Computador)

[CONWAY 97] CONWAY, Paul. *Preservação no universo digital*. Coord. Ingrid Beck, Trad. Olga Marder. Rio de Janeiro, Arquivo Nacional, 1997. 24p. (Tradução de *Preservation in the digital world*).
Disponível na World Wide Web (em inglês):
<http://www.clir.org/pubs/reports/conway2/index.html> 14/jan/2003.

[CHEN 94] CHEN, P. M. et al. RAID: High-Performance, Reliable Secondary Storage. *ACM Computing Surveys*. v.26, n.2, p.145-185, 1994.
Disponível na WWW: <http://citeseer.nj.nec.com/chen94raid.html> 14/jan/2003.

[DCMI 03] Dublin Core Metadata Initiative. Homepage. 2003.
Disponível na WWW: < http://dublincore.org/> 14/jan/2003.

[DEL BIMBO 99] DEL BIMBO, Alberto. *Visual information retrieval*. Morgan Kaufmann, 1999. 270p.

[DIGITAL 02] *Digital Photography vs Traditional Photography (How to compare)*. TheImage.com. Junho/2002.
Disponível na WWW: <http://www.theimage.com/photography/index.htm> 14/jan/2003.

[DOLLAR 92] DOLLAR, Charles M. *Archival theory and information technologies:* the impact of information technologies on archival principles and methods. Macerata, University of Macerata Press, 1992. apud [CONWAY 97]

[DUCHEIN 86] DUCHEIN, Michel. O respeito aos fundos em arquivistica: Princípios teóricos e problemas práticos. *Arq. & Adm*. Rio de Janeiro, Agosto/1986.

[DUMAIS 88] DUMAIS, Susan et al. Using Latente Semantic Information to improve Access to Textual Information. In: Conference on Human Factors in Computing Systems CHI'88, 1998. *Proceedings of the…* p. 281-285.
Disponível na WWW: < http://superbook.bellcore.com/~std/papers/CHI88.ps> 14/jan/2003.

[DUNN 00] DUNN, Heather.Collection Level Description - the Museum Perspective. *D-Lib Magazine*. v. 6, n. 9. Setembro/2000.
Disponível na WWW: <http://www.dlib.org/dlib/september00/dunn/09dunn.html> 14/jan/2003.

[EPRINTS 02a] EPrints.org — Self-Archiving and Open Archives. *Homepage*, 2003.
Disponível na WWW: <http://www.eprints.org> 14/jan/2003.

[EPRINTS 02b] EPrints. *Self-Archiving FAQ*. University of Southampton, 2002.
Disponível na WWW: <http://www.eprints.org/self-faq/> 14/jan/2003.

[EPRINTS 03] GNU EPrints 2. *Homepage*, 2003.
Disponível na WWW: <http://software.eprints.org> 14/jan/2003.

[FERREIRA 93] FERREIRA, Benedito de Jesus Pinheiro. *Projeto e Implementação de Um Sistema de Gerência de Documentos Segundo o Paradigma de Objetos*. Dissertação de Mestrado. Universidade Federal do Rio Grande do Sul, Porto Alegre, 1993.

[GALLAGHER 00] GALLAGHER, Len, CARNAHAN, Lisa. "A General Purpose Registry/Repository Information Model". 2nd Draft. National Institute of Standards and Technology, 2000.

[GILL 98] GILL, Tony, Anne Gilliland-Swetland, and Murtha Baca. *Introduction to Metadata*; Pathways to digital information. Version 2.0. Getty Research Institute. 47 p.
Disponível na WWW: <http://www.getty.edu/research/institute/standards/intrometadata/> 14/jan/2003.

[GORMAN 78] GORMAN, Michael e WINKLER, Paul. *Código de Catalogação Anglo-Americano*. 2ª Edição. Federação Brasileira de Associações de Bibliotecários, 1978.

[GREENSTONE 03] Greenstone Digital Library Software. *Homepage*, 2003.
Disponível na WWW: < http://www.greenstone.org/> 14/jan/2003.

[HARTKE 01] HARTKE, Jerome L. Measures of CD-R Longevity. Media Sciences Inc., 2001.
Disponível na WWW: <http://www.mscience.com/longev.html> 14/jan/2003.

[HURLEY 99] HURLEY, Bernard et al. *The Making of America II Testbed Project*: A Digital Library Service Model. Massachussets: Council on Library and Information Resources, 1999.
Disponível na WWW: <http://www.clir.org/pubs/reports/pub87/pub87.pdf> 14/jan/2003.

[ICC 03] International Color Consortium. *Homepage*. 2003.
Disponível na WWW: <http://www.color.org> 14/jan/2003.

[JPEG 03] Joint Picture Expert Group. *Homepage*. 2003.
Disponível na WWW: <http://www.jpeg.org/> 14/jan/2003.

[KENNEY 97] KENNEY, Anne and CHAPMAM, Stephen. *Requisitos de resolução digital para textos*: métodos para o estabelecimento de critérios de qualidade de imagem. Coord. Ingrid Beck, Trad. José L. Pedersoli. Rio de Janeiro: Arquivo Nacional, 1997. 26 p. (Tradução de *Tutorial — Digital Resolution Requirements for Replacing Text-based Materials: Methods for Benchmarking Image Quality*).
Disponível na WWW: <http://www.cpba.net/pdf_cadtec/51.pdf> 14/jan/2003.

[KODAK 01] Kodak Ultima Lifetime Discussion. Eastman Kodak Company, 2001.
Disponível na WWW: <http://www.kodak.com/global/en/service/cdrMedia/index.jhtml> 14/jan/2003.

[LAHANIER 02] LAHANIER, Christian et al. CRISATEL: high definition spectral digital imaging of paintings with simulation of varnish removal. In: ICOM-CC 13th Triennial Meeting, 2002, Rio de Janeiro. *Annals of…*p. 295-300.

[LAWRENCE 00] LAWRENCE, Gregory et al. *Risk Management of Digital Information:* a file format investigation. Council on Library and Information Resources, 2000.

[LILLEY 03] LILLEY, Chris. *PNG (Portable Network Format).* World Wide Web Consortium, 2003.
Disponível na WWW: <http://www.w3.org/Graphics/PNG/> 14/jan/2003.

[LOOTS 02] LOOTS, Michel, CARMAZAN Dan and WITTEN, Ian. *From Paper to Collection.* University of Waikato, 2002.
Disponível na WWW: <http://prdownloads.sourceforge.net/greenstone/Paper.pdf> 14/jan/2003.

[MCCARTHY 02] MCCARTHY, Ann. *Color Imaging Workflow Primitives*. International Color Consortium, 2002.
Disponível na WWW: <http://www.color.org/primitivelong.pdf> 14/jan/2003.

[MCFADDEN 03] MCFADDEN, Andy. *CD-Recordable FAQ*.
Disponível na WWW: <http://www.cdrfaq.org/> 14/jan/2003.

[MELLO 99] MELLO, C. e LINS, R. A Comparative Study on OCR Tools. Vision Interface'99, Quebéc, Canadá, 1999.
Disponível na WWW: <http://www.cin.ufpe.br/~nabuco/Projeto/publicacoes/vi99.ps> 14/jan/2003.

[MHS 03] Minnesota Historical Society. *Homepage*. 2003
Disponível na WWW: <http://www.mnhs.org> 14/jan/2003.

[MILLER 00] MILLER, Paul. Collected Wisdom: Some Cross-domain Issues of Collection Level Description. *D-Lib Magazine*. v. 6, n. 9. Setembro/2000.
Disponível na WWW: <http://www.dlib.org/dlib/september00/miller/09miller.html> 14/jan/2003.

[MILLER 03] MILLER, Eric, et al. *Resource Description Framework (RDF).* World Wide Web Consortium, 2003.
Disponível na WWW: <http://www.w3.org/RDF/ > 14/jan/2003.

[MSA 02a] Minnesota State Archives. *Electronic Records Management Guidelines Homepage*. Minnesota Historical Society, 2002.
Disponível na WWW:
<http://www.mnhs.org/preserve/records/electronicrecords/erguidelines.html> 14/jan/2003.

[MSA 02b] Minnesota State Archives.*Metadata*. Minnesota Historical Society, 2002.
Disponível na WWW: <http://www.mnhs.org/preserve/records/metadata.html> 14/jan/2003.

[MSA 03] Minnesota State Archives. *Electronic Records Management Guidelines*. Version 2. Minnesota Historical Society, 2003.
Disponível na WWW:
<http://www.mnhs.org/preserve/records/electronicrecords/erguidelinestoc.html> 14/jan/2003.

[MUSTARDO 97] MUSTARDO, Peter e KENEDY, Nora *Preservação de fotografias: métodos básicos para a salvaguardar suas coleções*. Coordenação de Ingrid Beck, tradução de Olga de Souza Marder. Rio de Janeiro, Arquivo Nacional, 1997.

[OAI 03] Open Archives Initiative. *Homepage*, 2003.
Disponível na WWW: <http://www.openarchives.org/> 14/jan/2003

[PARKER 96] PARKER, Dana J. and STARRETT, Robert. Testing, Testing… CD-R. *CD-ROM Professional*, Fevereiro/1996.
Disponível na WWW: <http://www.cdpage.com/dstuff/BobDana296.html> 14/jan/2003.

[PAULA XX] PAULA Filho, Wilson. Manual do Engenheiro de Software.

[PJN 00] Projeto Joaquim Nabuco. *Homepage*. 2000.
Disponível na WWW: <http://www.cin.ufpe.br/~nabuco/> 14/jan/2003.

[PUTS 01] PUTS, Erwin. *Exploring the limits of 35mm BW Photography*. imX Photosite. 21/Maio/2001.
Disponível na WWW: <http://www.imx.nl/photosite/technical/highres.html> 14/jan/2003.

[RITTER 97] RITTER, Niles. *The Unofficial TIFF Home Page*. 1997.
Disponível na WWW: <http://home.earthlink.net/~ritter/tiff/> 14/jan/2003.

[ROELOFS 03] Roelofs, Greg. *Portable Network Format Home Site*. 2003.
Disponível na WWW: <http://www.libpng.org/pub/png/> 14/jan/2003.

[ROSA 94] ROSA, Luciano G. Nabuco: Uma Base de Dados para Documentos Históricos. Dissertação de Mestrado. Departamento de Informática, Universidade Federal de Pernambuco, Recife, 14/12/1994.
Disponível na WWW: < http://www.cin.ufpe.br/~nabuco/Projeto/publicacoes/tese1.html> 14/jan/2003.

[ROSS 99] ROSS, Seamus and GOW, Ann. *Digital Archaelogy:* rescuing neglected and damaged data resources. London, Library Information Technology Centre, 1999.
Disponível na WWW:
<http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/p2con.pdf> 14/jan/2003.

[ROTHENBERG 98] ROTHENBERG, John. *Avoiding technological quicksand*: finding a viable technical foundation for digital preservation. Concil on Library and Information Resources, 1998. 35p.
Disponível na WWW: <http://www.clir.org/pubs/reports/rothenberg/contents.html> 14/jan/2003.

[SCHELLENBERG 73] SCHELLENBERG, Theodore R. *Arquivos modernos; princípios e técnicas*. Trad. de Nilza T. Soares. Rio de Janeiro, Fundação Getúlio Vargas, 1973. 345p. (Tradução de *Modern archives: principles and techniques*).

[SHAW 97] SHAW, Elizabeth and BLUMSON, Sarr. Making of America; Online Searching and Page Presentations at the University of Michigan. *D-Lib Magazine*, July/August 1997.
Disponível na WWW: <http://www.dlib.org/dlib/july97/america/07shaw.html> 14/jan/2003.

[STINSON 95] STINSON, David et al. Lifetime of KODAK Writable CD and Photo CD Media. Eastman Kodak Company, 1995.
Disponível na WWW: <http://www.cd-info.com/CDIC/Technology/CD-R/Media/Kodak.html> 14/jan/2003.

[STOKES 03] STOKES, Michael et al. *A Standard Default Color Space for the Internet*. Version 1.1. International Color Consortium, 1996.
Disponível na WWW: <http://www.color.org/sRGB.html > 14/jan/2003.

[STONE 01] STONE, David. Color Matching. *Extreme Tech*. Ziff Davis Media: June / 2001.
Disponível na WWW: <http://www.extremetech.com/article2/0,3973,15467,00.asp> 14/jan/2003.

[SUPER DLT] *Super DLT tape*. Quantum Corporation.
Disponível na WWW: <http://www.dlttape.com//> 14/jan/2003.

[SWEET 00] SWEET, Meg and Thomas, David. Archives Described at Collection Level. *D-Lib Magazine*. v. 6, n. 9. Setembro/2000.
Disponível na WWW: <http://www.dlib.org/dlib/september00/sweet/09sweet.html> 14/jan/2003.

[TEI 02] TEI Consortium. The TEI Guidelines. *Text Encoding Initiative*, 2002.
Disponível na WWW: < http://www.tei-c.org/Guidelines2/index.html> 14/jan/2003.

[TENENBAUM 96] TENENBAUM, Andrew. *Computer networks*. 3rd edition. Upple Saddle River, New Jersey: Prentice Hall PTR. 1996.

[TFADI 96] Task Force on Archiving of Digital Information. *Preserving Digital Information*. The Comission of Preservation and Access and The Research Libraries Group, 1996. 64p.

[TLOC 03a] The Library of Congress. *Homepage*. 2003.
Disponível na WWW: <http://www.loc.gov/> 14/jan/2003.

[TLOC 03b] The Library of Congress. *Digital Collections and Programs: Library Functions*. 2003.
Disponível na WWW: <http://www.loc.gov/> 14/jan/2003.

[TLOC 02a] The Library of Congress. *MARC Standards*. 2002.
Disponível na WWW: <http://www.loc.gov/marc/> 14/jan/2003.

[TLOC 02b] The Library of Congress. *METS —Metadata Encoding & Transmission Standard*. 2002.
Disponível na WWW: <http://www.loc.gov/standards/mets/> 14/jan/2003.

[TLOC 03c] The Library of Congress. *Encoded Archival Description (EAD)*. 2003.
Disponível na WWW: <http://lcweb.loc.gov/ead/> 14/jan/2003.

[TSAA 00a] The Society of American Achivists. *Descripción Archivística Codificada;* directrices de aplicación. Versión 1.0. Fundación Histórica Tavera, 2000. 313p.
Disponível na World Wide Web (em inglês): <http://www.loc.gov/ead/ag/aghome.html> 14/jan/2003.

[TSAA 00b] The Society of American Achivists. *Descripción Archivística Codificada;* repertorio de etiquetas. Versión 1.0. Fundación Histórica Tavera, 2000. 272p.
Disponível na World Wide Web (em inglês): <http://www.loc.gov/ead/tglib/index.html> 14/jan/2003.

[TUDHOPE 02] TUDHOPE, Douglas et al. Representation and Retrieval in Faceted Systems. In: Seventh International ISKO Conference, 10-13/Julho/2002, Granada, Espanha. *Proceedings of…* (Advances in Knowledge Organization; Vol 8). p. 191-197.
Disponível na WWW: <http://www.glam.ac.uk/soc/research/hypermedia/publications/presentationdocs/ISKOn.doc> 14/jan/2003.

[UNESCO 00] UNESCO. *Safeguarding our documentary heritage*. UNESCO – "Memory of the World" Programme, 2000.
Disponível na WWW: <http://webworld.unesco.org/safeguarding/> 14/jan/2003.

[USCB 00] United States Census Bureau. Computer Use and Ownership. *Current Population Survey*, 1984, 1993, 1997 and 2000.
Disponível na WWW:
<http://www.census.gov/population/www/socdemo/computer.html> 14/jan/2003.

[VAKULENKO 00] VAKULENKO, Alex. *Removing Moiré Patterns*. 2000.
Disponível na WWW: <http://www.oberonplace.com/dtp/moire/index.htm> 14/jan/2003.

[VALLE 02a] VALLE, Eduardo A. e ARAÚJO, Arnaldo A. Preserving Historical Collections Using Multimedia Information Systems. VII Brazillian Symposium on Multimedia and Hypermedia Systems — SBMIDIA, Thesis and Dissertations Workshop, Fortaleza - CE, Brasil, 2002. *Proceedings of the…* p. 317-324.
Disponível na WWW: <http://www.npdi.dcc.ufmg.br/pesquisa/publicacoes/edujr02.pdf> 14/jan/2003.

[VALLE 02b] VALLE, Eduardo A. et al. A Tool for Workflow Management in the Composition of Multimedia Databases from Preexisting Documents. VII Brazillian Symposium on Multimedia and Hypermedia Systems — SBMIDIA, Tools and Applications Workshop, Fortaleza - CE, Brasil, 2002. *Proceedings of the…* p. 317-324.
Disponível na WWW: <http://www.npdi.dcc.ufmg.br/pesquisa/publicacoes/edujr01.pdf> 14/jan/2003.

[VAN BOGART 95] VAN BOGART, John W. *Magnetic tape storage and handling: a guide for libraries and archives*. Washington D.C., Comission on Preservation and Access, 1995.

[WATERS 97] WATERS, Donald. *Do microfilme à imagem digital*: como executar um projeto para estudo dos meios, custos e benefícios de conversão para imagens digitais de grandes quantidades de documentos preservados em microfilme. Rio de Janeiro: Arquivo Nacional, 1997. 37 p.
Disponível na WWW: <http://www.cpba.net/pdf_cadtec/49.pdf> 14/jan/2003.

[WEBER 97] WEBER, Hartmut, DÖRR, Marianne. *Digitization as a Means of Preservation?*. European Comission on Preservation and Access, 1997.

[WILLIS 97] WILLIS, Don. *Uma abordagem de sistemas híbridos para a preservação de materiais impressos*. Rio de Janeiro: Arquivo Nacional, 1997. 61 p.
Disponível na WWW: <http://www.cpba.net/pdf_cadtec/50.pdf> (Tradução de *A Hybrid Systems Approach to Preservation of Printed Materials*)

[WITTEN 01] WITTEN, I., BAINBRIDGE, D., and BODDIE, S. Greenstone: Open-source Digital Library Software with end-user collection building. *Online Information Review*, v. 25, n. 5, 2001. p. 288-198.
Disponível na WWW: <http://www.cs.waikato.ac.nz/~nzdl/publications/2001/01IHW-DB-SB-Greenstoneopn.pdf> 14/jan/2003.

8.2 Créditos de imagens

Figura 3.3 —Imagem original: *Cavendish Laboratory, University of Cambridge*
Digitalizada a partir de: Superinteressante. Super Fotos Especial. Número especial, p. 46. 1990. São Paulo: Editora Abril

Figura 3.4 —Imagem original: *AFP - Agence France-Press*
Digitalizada a partir de: Jornal Estado de Minas. n. 22,309 p. 15. 26 / Janeiro / 2003. Belo Horizonte: Estado de Minas S/A.