Estudo comparativo entre sistemáticas de digitalização de documentos: formatos HTML e PDF

André Raabe
Omer Pohlmann Filho

Resumo

Este artigo apresenta o resultado de experimentos realizados pelo Laboratório de Biblioteca Digital da PUCRS voltados para a captura e conversão de documentos a partir do formato tradicional (papel) para o formato digital. São apresentadas e avaliadas as principais etapas envolvidas no processo de digitalização utilizando duas sistemáticas diferentes: uma baseada na conversão para HTML; a outra baseada na geração de arquivos PDF usados pelo ***software*** *Adobe Acrobat Reader.*
São abordados também fatores essenciais aos trabalhos de digitalização tais como tecnologias de Reconhecimento Óptico dos Caracteres (OCR) e avaliação das características do acervo a ser digitalizado. Por fim, é realizado um comparativo entre as duas sistemática estudadas, apontando pontos positivos e negativos que devem ser considerados na escolha de uma diretriz de trabalho.

Palavras-chave

Conversão de documentos do formato tradicional para o digital; Sistemáticas de conversão para HTML; Geração de arquivos PDF; Tecnologias de reconhecimento óptico dos caracteres.

INTRODUÇÃO

A Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), mediante convênio com a IBM, participa do projeto IBM Global Campus, que prevê a colaboração entre instituições de ensino superior de diferentes países, no sentido de pesquisar e desenvolver políticas, abordagens, metodologias e recursos tecnológicos para projetar e implantar universidades com *campus* de abrangência global.

A proposta de trabalho do projeto Campus Global PUCRS visa a desenvolver estudos sobre universidade virtual, centrando seu foco de atenção em pesquisas sobre metodologias e recursos tecnológicos na área de educação à distância. Neste contexto, trabalha-se com o conceito de Educação à Distância (EAD), como uma forma de educação na qual alunos e professores se encontram separados fisicamente, sendo o processo de interação multidirecional, apoiado por tecnologia de comunicação, em que o aluno é o protagonista de seu aprendizado e o professor um facilitador deste.

Tendo em vista esta proposta, o projeto Campus Global foi estruturado a partir de quatro frentes de pesquisa, a saber, educação à distância e colaborativa, bibliotecas digitais, trabalho cooperativo, gerência de recursos Internet.

Neste contexto, o Laboratório de Biblioteca Digital vem pesquisando o desenvolvimento de tecnologias para permitir o acesso a informações de conteúdo bibliográfico à distância. Uma das alternativas pesquisadas aponta para a digitalização de documentos e sua disponibilização por meio da Internet.

Para tanto, o Laboratório de Biblioteca Digital voltou-se inicialmente para a pesquisa de *software* e desenvolvimento de sistemáticas para a captura e transformação de documentos do formato tradicional (papel), para o formato digital. Foram avaliadas duas sistemáticas distintas para realização do trabalho, uma delas baseada no reconhecimento ótico dos caracteres e conversão para HTML detalhada em (Pohlmann[1]); a outra baseada no formato digital Portable Document Format (PDF).

DIRETRIZES DE TRABALHO

Dentre as alternativas pesquisadas para o processo de digitalização de documentos, avaliaram-se duas diretrizes genéricas:

1. digitalização da obra como imagens e conversão destas em textos mediante reconhecimento óptico de caracteres (OCR);

2. criação de arquivos de imagens (JPG), contendo as páginas da obra e mantendo o leiaute original da publicação, sem conversão para texto.

A escolha do formato de arquivo JPG deveu-se ao fato de este possuir uma alta taxa de compressão, permitindo o armazenamento de imagens com qualidade em arquivos de tamanhos reduzidos, sendo, por este motivo, amplamente utilizado na rede Internet.

Para exemplificar a relação entre formato de arquivo e espaço de armazenagem, foi realizado um teste comparativo permitindo verificar a relação entre os tamanhos dos arquivos gerados no contexto estudado – digitalização de documentos a partir do formato papel. É importante salientar que foram utilizadas rigorosamente as mesmas configurações de compressão e qualidade disponíveis em formatos de arquivos como o JPG e GIF.

No teste, foram utilizadas uma página do livro História da PUCRS ilustrada na figura 1, por esta ser composta de imagem e texto sem cores, e a capa da mesma obra, por ser colorida, figura 2.

A tabela 1 apresenta, a seguir, os resultados do teste comparativo considerando os formatos de arquivos de imagens mais utilizados.

Para avaliar as vantagens e desvantagens relacionadas a cada uma das diretrizes estudadas (conversão para texto e disponibilização como imagem), tomou-se como base (Haigh[2]), que considera, para a escolha do processo de digitalização, os seguintes pontos:

• necessidade de reutilização, edição ou reformatação do texto;

• disponibilidade do texto para pesquisas *full-text* ;

• posterior codificação do texto no formato HTML;

• recursos disponíveis para realização do processo;

• tamanho dos arquivos para armazenagem e transmissão.

Pela análise realizada, chegou-se às seguintes conclusões sobre cada um dos processos:

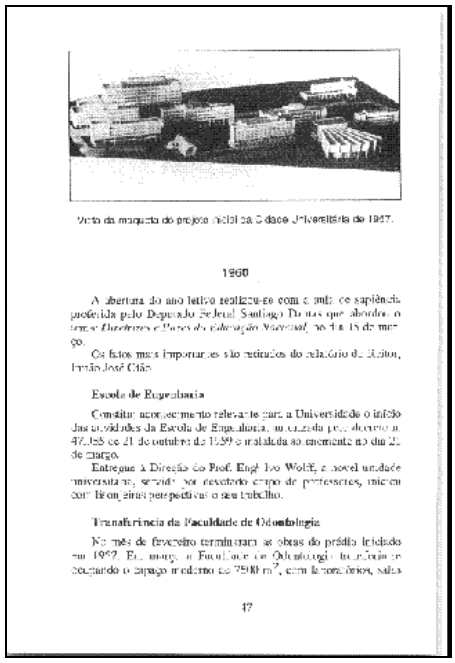
• *Com conversão para texto:*

– possibilidade de edição e manipulação do texto das obras;

– possibilidade de realização de pesquisas *full-text*;

– processo de digitalização é demorado e trabalhoso;

FIGURA 1



Vista da maquete do projeto inicial da Cidade Universitária de 1957.

1960

A abertura do ano letivo realizou-se com a aula de sapiência proferida pelo Deputado Federal Santiago Dantas que abordou o tema "Diretrizes e Bases da Educação Nacional", no dia 15 de março.

Os fatos mais importantes são retirados do relatório do Reitor, Irmão José Otão.

Escola de Engenharia

Constitui acontecimento relevante para a Universidade o início das atividades da Escola de Engenharia, autorizada pelo decreto n. 47.185 de 21 de outubro de 1959 e instalada solenemente no dia 21 de março.

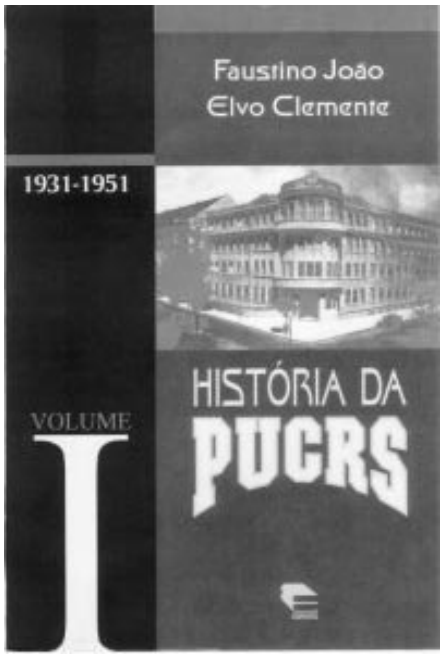
Entregue à Direção do Prof. Engº Ivo Wolff, a novel unidade universitária, servida por excelente corpo de professores, iniciou com galhardia perspectiva o seu trabalho.

Transferência da Faculdade de Odontologia

No mês de fevereiro terminaram as obras do prédio iniciado em 1962. Ele abriga a Faculdade de Odontologia, ocupando o espaço moderno de 7500 m², com laboratórios, salas

17

FIGURA 2



Páginas utilizada como teste comparativo

TABELA 1
Comparativo entre os tamanhos de arquivo

Formato do Arquivo	Pagina P&B 11,50 cm X 17,80 cm	Capa (reduzida) 7,14 cm X 10,68 cm
JPG (Joint Picture Experts Group)	34 Kb	16 Kb
TIF (Tagged Image File)	74 Kb	43 Kb
GIF (Graphics Interchange Format)	76 Kb	40 Kb
PDF (Portable Document Format)	76 Kb	34 Kb
PSD (PhotoShop)	82 Kb	62 Kb
PCX (Zsoft Paintbrush)	97 Kb	64 Kb
BMP (Windows Bitmap)	280 Kb	60 Kb
WMF (Windows Meta File)	282 Kb	61 Kb
EPS (Encapsuled Post Script)	573 Kb	135 Kb

– necessita pouco espaço para armazenagem das obras digitalizadas.

• *Disponibilização como imagem:*

– impossibilidade de edição e manipulação do texto das obras;

– impossibilidade de realização de pesquisas *full-text* ;

– processo de digitalização simplificado e rápido;

– ocupa grande espaço para armazenagem da obra digitalizada (aproximadamente 20 vezes mais que textos).

No contexto geral do projeto, a realização de pesquisa *full-text* se faz necessária e é um objetivo a ser alcançado. Além disso, outro fator determinante a favor da conversão para texto é que a velocidade de transmissão de dados no Brasil ainda não atinge os padrões desejados para transferência de arquivos de imagem. No caso dos arquivos texto, a velocidade de transmissão não é um fator crítico, pois estes possuem tamanho bem inferior.

Portanto, decidiu-se proceder inicialmente à digitalização de obras mediante o reconhecimento ótico dos caracteres (OCR) e posterior transformação no formato HTML.

SISTEMÁTICA DE DIGITALIZAÇÃO HTML

Para realização dos trabalhos voltados à definição de uma sistemática de captura e conversão de documentos para o formato HTML, foi utilizado como instrumento de testes a publicação da Faculdade de Medicina da PUCRS denominada Acta Médica Volume 1. Os trabalhos foram realizados mediante os recursos disponíveis, ou seja, scanner HP Scanjet II, *software* de OCR (Reconhecimento Ótico de Caracteres) OmniPage Pro 5.0, editor de textos MS-*Word* 6.0 e o conjunto de *softwares* do Netscape Communicator 4.0 (Pohlmann[1]).

Inicialmente, são apresentadas as principais características das Actas Médicas, para que se possa ter uma idéia do contexto estudado e da adequação dos recursos utilizados.

Como um compilado de artigos de alunos (trabalhos de conclusão), estas publicações começaram a ser editadas em 1977. Em sua maioria, são documentos antigos que precisam passar pela função de criação e captura, ou seja, conforme (Pohlmann[1]) são documentos que têm de ser necessariamente digitalizados.

O objeto inicial desta pesquisa foi a primeira edição da Acta Médica, editada em 1977. Este documento possui um leiaute de texto complexo contendo:

– texto dividido em duas colunas;

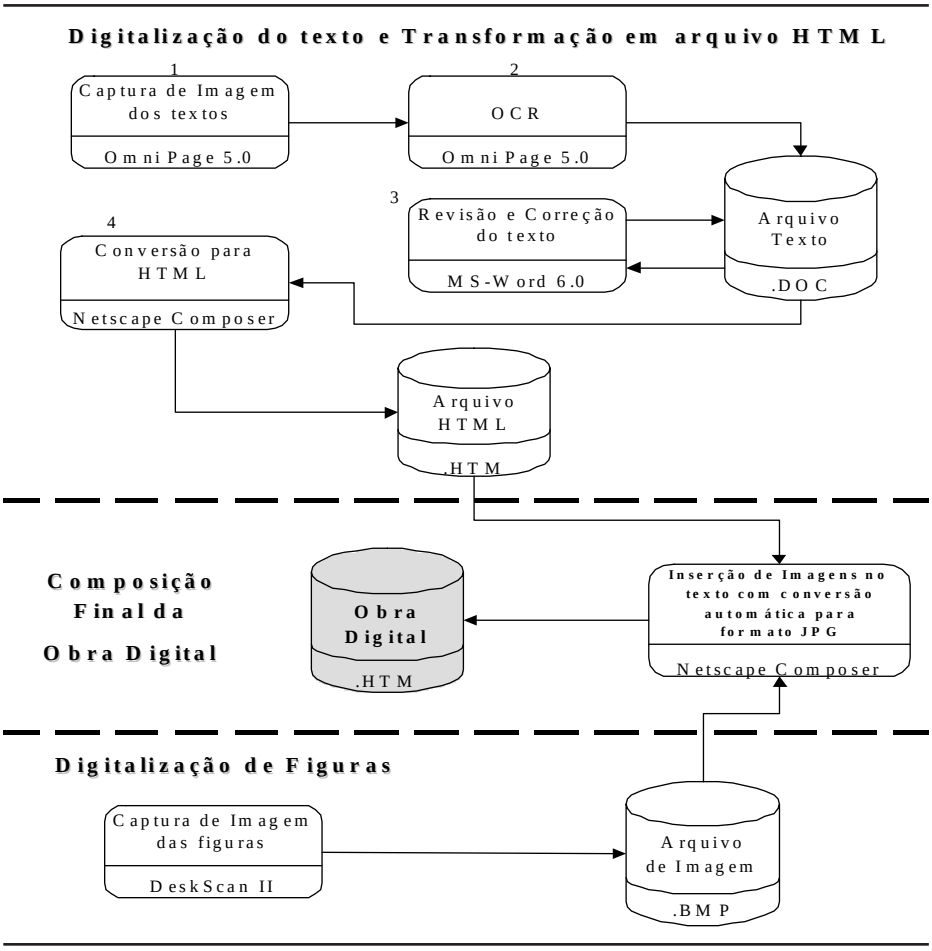
– tamanho de letra muito pequeno (aproximadamente tamanho *times new roman* 10);

– fórmulas matemáticas manuscritas em meio ao texto;

– seqüência de texto não linear (em alguns capítulos, o texto inicia pela coluna da direita);

– figuras e imagens;

FIGURA 3
Representação Esquemática do Processo de Digitalização HTML



– notas de rodapé;

– subdivisão de itens por meio de chaves;

– manchas de tinta e falhas na impressão;

– textos e figuras somente na cor preta ou tons de cinza.

Descrição do Processo de Digitalização

O objetivo do processo foi a transformação da obra para formato digital e sua publicação em formato HTML. Para tanto, dividiu-se o trabalho em quatro etapas:

1) leitura ótica das páginas da obra;

2) reconhecimento do texto por meio de *software* de OCR;

3) revisão e correção do texto por intermédio de editor de texto; (MS-*Word* 6.0);

4) conversão para formato HTML.

A descrição detalhada de cada uma destas etapas pode ser vista no endereço

http://www.cglobal.pucrs.br/bibdigital/artigos/art3.htm.

Para cada capítulo do livro, foi criado um arquivo HTML diferente para facilitar a posterior ligação com indexadores, *hiperlinks* e *softwares* de gerência de bibliotecas.

O espaço total em disco, ocupado pelos 16 capítulos digitalizados, contendo 241 páginas de texto e 72 figuras, foi de 2,41 *megabytes*, comprovando eficiência em termos de economia de espaço de armazenagem e conseqüente agilidade no acesso aos documentos *full-text*, via Internet.

Um resumo deste processo é apresentado, acima, esquematicamente, na figura 3.

Na tabela 2 são apresentados também os tempos médios verificados na execução de cada uma destas etapas. Os tempos apresentados são para um número padrão de 50 páginas e 12 figuras.

Busca de melhores resultados no OCR

A elaboração da sistemática HTML foi feita com a utilização do *software* de OCR Omni Page Pro 5.0. Os resultados deste experimento apontaram um tempo total de conversão muito alto, principalmente pela necessidade de realização de uma revisão e correção meticulosa dos erros gerados pelo processo de reconhecimento óptico dos caracteres (OCR).

Dando continuidade ao trabalho, desejava-se verificar a utilização de uma versão mais atual do *software*, o Omni Page Pro 8.0, a fim de identificar melhorias no processo de reconhecimento de caracteres que reduzissem o trabalho de revisão e correção a patamares aceitáveis, dentro do escopo de um projeto de digitalização em larga escala.

Para realização deste trabalho comparativo, escolheu-se um informativo de publicação interna na PUCRS chamado "PUCRS Informação". O mesmo foi escolhido por possuir uma diagramação elaborada, com fotos e textos distribuídos de forma não-linear, permitindo a comparação entre os procedimentos de definição automática das zonas de texto de ambas as versões do OmniPage Pro.

O processo de definição de zonas de texto pode ser realizado de forma manual, conforme descrito em (Pohlmann[1]), produzindo um resultado melhor em termos de fidelidade ao leiaute da obra original, no entanto esta atividade envolve muita interação do usuário tornando o processo lento. A utilização da definição automática das zonas do texto é uma tentativa de reproduzir o leiaute da obra original sem a interação do usuário. No entanto, os resultados ficam aquém dos esperados.

TABELA 2
Tempos médios para realização das etapas

ETAPAS	TEMPO MÉDIO
Captura das imagens dos textos e execução do programa de reconhecimento ótico de caracteres – OCR (com a criação do arquivo texto)	65 minutos
Revisão e correção do texto	400 minutos
Conversão dos arquivos texto para arquivos HTML	15 minutos
Captura de imagens e criação de arquivos BMP	20 minutos
Inserção de imagens no texto e composição final da obra	10 minutos
Tempo médio para transformação de um texto de 50 páginas, com 12 figuras, do formato convencional (em papel), para o formato digital, segundo a sistemática proposta	510 minutos (aprox. 8,5 horas)

Observação: Cumpre salientar que estes tempos foram estimados contando com a participação de duas pessoas para sua realização. Obviamente, quanto maior a equipe, menor o tempo consumido. Também os recursos de *hardware* utilizados, principalmente o *scanner* que não possuía recurso ADF (Automatic Document Feeder), não são os recomendados para este tipo de trabalho. A utilização de recursos mais apropriados tende a melhorar as *performances* observadas, principalmente nas etapas de captura de imagens, revisão e correção de texto, que são críticas neste processo.

Outra característica a ser salientada é a alta qualidade (qualidade laser) de impressão do informativo, bem como a utilização de fontes padrão (arial), o que, segundo (Caere[3]), levaria o Omni Page Pro 8.0 a atingir uma taxa de acerto no reconhecimento dos caracteres superior a 99%.

O "PUCRS Informação", composto de 20 páginas, foi digitalizado e armazenado como imagem para posterior reconhecimento dos caracteres e zonas de texto em ambas as versões do OmniPage. Para tanto, foi utilizado o Omni Page Pro 5.0 e um *scanner* de mesas HP Scanjet II, gerando um arquivo de saída no formato proprietário MET contendo as 20 páginas digitalizadas. Este arquivo foi aberto em ambas as versões 5.0 e 8.0, onde foi realizado o reconhecimento óptico dos caracteres (OCR) e a definição automática das zonas de texto, uma vez que ambas as versões possuem esses recursos.

Após realizado o processo, os arquivos de saída contento o texto reconhecido pelo OCR foram salvos no formato DOC do *MS-Word* 6.0, por ser comum a ambas as versões e permitir a utilização de um dicionário ortográfico comum na detecção dos erros de reconhecimento dos caracteres. A comparação entre as taxas de reconhecimento atingidas pelas versões 5.0 e 8.0 do Omnipage Pro partiu de uma análise destes arquivos.

Realizou-se a contagem do número total de palavras na obra. A seguir, realizou-se a contagem das palavras que possuíam incorreções na grafia originadas por um erro no reconhecimento dos caracteres. De posse deste valores, calculou-se o percentual de acertos atingido pelo reconhecimento dos caracteres em ambas as versões. Cumpre salientar que os dados obtidos relacionados à taxa de reconhecimento do processo de OCR consideraram as 20 páginas do informativo na íntegra. A tabela 3, a seguir, ilustra os resultados obtidos.

Concluiu-se que a utilização de uma versão mais atual do *software* Omni Page Pro não promoveu significativa melhoria nos resultados do processo de reconhecimento dos caracteres que pudesse acelerar significativamente os trabalhos de digitalização de um acervo em larga escala, uma vez que a necessidade de revisão do texto permaneceu necessária.

TABELA 3
Comparativo dos resultados do OCR

	Total de palavras	Palavras com erro	Taxa de reconhecimento
Versão 5.0	4833*	308	93,6%
Versão 8.0	4785	168	96,5%

* A diferença observada no número total de palavras deve-se ao fato de a versão 5.0 dividir algumas palavras ao meio, gerando duas novas.

Diretrizes para busca de uma nova sistemática

Segundo (Haigh[2]), a taxa de reconhecimento de um OCR para conversão de documentos deve ser superior a 98%. Caso contrário, é mais eficiente realizar a redigitação do documento.

Esta taxa de reconhecimento é medida considerando o número de edições necessárias (inserções, deleções, substituições) diante do número total de caracteres. Recomenda-se, no entanto, que este dado não seja utilizado como referência para trabalhos de digitalização em larga escala, pois desconsidera todo o trabalho de localização de erros no texto, que muitas vezes demanda uma leitura completa da obra. Além disso, quando o vocabulário utilizado é eminentemente técnico, pode ser necessária a confrontação com a obra original em papel, para identificação da grafia correta de uma palavra.

O que deve ser considerado efetivamente é o volume de tempo despendido por um usuário, ao realizar a correção/conferência de um texto reconhecido pelo OCR.

Enquanto os *software* de OCR não atingirem uma taxa de reconhecimento de 100%, será necessária meticulosa revisão da obra para localização e correção dos erros, atividade essa que torna a realização de trabalhos de digitalização em larga escala altamente custosos, sendo necessária a utilização de grandes equipes com numerosos recursos para que o trabalho não se torne excessivamente demorado.

A integração de dicionários ortográficos ao processo de reconhecimento dos caracteres, como o procede o Omni-Page Pro, auxilia a identificação das palavras consideradas suspeitas. No entanto, dada a impossibilidade de se construir um dicionário eletrônico que abranja todos os termos técnicos específicos de cada área, nos diversos idiomas contemplados pelo acervo da biblioteca da PUCRS, persiste a necessidade de realização de uma leitura cuidadosa, pois uma palavra assinalada como suspeita pode estar correta ou não.

Como um agravante, muitas das obras do acervo da Biblioteca Central da PUCRS (aproximadamente 300 mil obras) não apresentam um estado de conservação adequado à realização da digitalização por meio de *softwares* de OCR, tais como:

• obras com páginas riscadas e com anotações a lápis e a caneta;

• obras com papéis com gramatura muito fina (50 gr/m²), fazendo com que o texto de um lado da página seja visível do outro lado;

• obras com páginas amassadas, manchadas, sujas, deterioradas por mofo, traças, ou mesmo pela própria utilização.

Apesar de possuir um setor específico para recuperação do acervo, muitos dos problemas são irrecuperáveis, e, com uma movimentação diária de 2 600 empréstimos, a tendência é que tais problemas não sejam eliminados.

Aliado a estes fatores, pode-se mencionar também o fato de que a maioria das obras é composta não somente de textos. Também contêm figuras, fórmulas e esquemas gráficos cujos *softwares* de OCR atuais não oferecem um tratamento adequado, sendo necessária a utilização de um outro *software* para captura das imagens e posterior integração. Tal procedimento exige muita interação do usuário e torna o tempo de digitalização de uma obra muito alto, considerando a amplitude do trabalho desejado.

Tais informações levaram os pesquisadores do laboratório de biblioteca digital da PUCRS a buscar outras alternativas de digitalização que envolvessem menor interação do usuário, viabilizando a realização do trabalho em larga escala.

Desta forma, iniciou-se o trabalho de definição de uma nova sistemática de digitalização utilizando o *software* da empresa Adobe, denominado Adobe Acrobat. Este *software* foi cedido pela empresa ao Laboratório de Biblioteca Digital da PUCRS para a realização de testes por um período de 60 dias. Segundo (Adobe[4]), tem como características:

• facilidade de criação e publicação de documentos *on-line*;

• mantém o leiaute original das obras digitalizadas;

• utiliza o formato de arquivo PDF (Portable Document Format), que permite a criação de documentos multiplataforma que podem ser visualizados inclusive em *browsers* (*software* de navegação na Internet);

• possibilidade de captura e conversão de grandes volumes de documentos com um baixo nível de interação do usuário.

Estas características se mostraram bastante adequadas ao tipo de trabalho que se tinha para realizar.

SISTEMÁTICA PDF

Digitalização por meio da captura de documentos

Para a definição de uma sistemática de digitalização por meio da captura de documentos utilizando o *software* Adobe Acrobat, escolheu-se, como objeto de teste, a obra história da PUCRS. Esta escolha deveu-se a adequação da obra ao trabalho que estava proposto e pela liberação dos direitos autorais da obra por parte dos autores, permitindo aliar sua disponibilização à comemoração do cinqüentenário da universidade, ocorrida em 1998, época em que esta pesquisa estava em andamento.

A obra possui somente uma coluna de texto, as letras são Times New Roman de tamanho 12. São encontradas fotografias e ilustrações em meio ao texto, no entanto nenhuma utiliza cores, somente tons de cinza.

Para digitalização da História da PUCRS, foi utilizado o *scanner* HP Network Scanner 5, destacando-se pela velocidade de digitalização e pela presença de uma bandeja para entrada automática de papel ADF (Automatic Document Feeder), apesar de suportar somente a digitalização de imagens em tons de cinza, fato que não atrapalhou a definição da sistemática.

Para a inserção das páginas da obras na bandeja de entrada de papel, foi necessário realizar um corte rente a parte onde as folhas estão presas para que estas se soltassem. Ao final, a aplicação cliente do *scanner* gera automaticamente um arquivo contendo as páginas digitalizadas como imagens, ou seja, uma imagem para cada página, agrupadas em um só arquivo PDF.

TABELA 4
Tempo para digitalização das obras utilizando o Scanner HP Network 5

Volume 1	159 páginas	16 minutos
Volume 2	295 páginas	36 minutos
Total	454 páginas	52 minutos

Média: 8,73 páginas por minuto

O tempo levado para digitalização dos dois volumes da História da PUCRS está explicitado na tabela 4.

Um dos objetivos previstos era a possibilidade de realização de pesquisas *full-text* na obra. Para tanto, seria necessário que as imagens digitalizadas passassem por um processo de reconhecimento ótico de caracteres (OCR), ou seja, a transformação da imagem em texto.

A realização de OCR em arquivos no formato PDF é feita pelo *software* Adobe Acrobat Exchange, o qual possui uma interface adequada para a tarefa, sendo possível a realização do processo sobre todas as páginas de uma vez só.

No entanto, foi frustrante descobrir que o *software* em sua versão 3.01 não possui dicionário para reconhecimento das palavras e caracteres da língua portuguesa, o que inviabilizou a realização do OCR e, por conseqüência, a transformação da imagem capturada via *scanner* em um texto.

Apesar disso, duas características interessantes relacionadas ao OCR aplicado pelo Acrobat Exchange devem ser ressaltadas:

1) as palavras ou conjuntos de caracteres não reconhecidos foram mantidos como imagens, de forma que a leitura do texto continuou sendo possível;

2) quando da aplicação do OCR nas páginas da obra, os arquivos reduziram o seu tamanho em aproximadamente quatro vezes;

Dado a inviabilidade de realização do OCR nas páginas da História da PUCRS, alguns dos objetivos traçados no início da pesquisa precisaram ser abandonados. Em especial a possibilidade de realização de pesquisas no texto desta obra.

Neste momento, decidiu-se dar seqüência ao trabalho de digitalização mantendo as páginas da obra como imagens, pois julgou-se importante avaliar as características e o comportamento dos arquivos PDF contendo imagens das páginas digitalizadas.

Partiu-se, então, para a exploração de recursos do formato PDF que permitiriam aprimorar a navegabilidade e aparência da obra, tais como a criação de *links* e *bookmarks*, a separação dos capítulos, a adequação do tamanho e da visualização das páginas da obra. Estes ajustes foram realizados utilizando o Adobe Acrobat Exchange. Uma descrição detalhada deste procedimentos pode ser encontrada em **http:\\www.cglobal.pucrs.br\bibdigital\kits\kit3.html**.

A seguir, será apresentado o resumo esquemático do trabalho de digitalização da História da PUCRS (figura 4).

São apresentados também os tempos médios verificados na execução de cada uma destas etapas. Os tempos apresentados correspondem à digitalização do primeiro volume da obra, com 159 páginas (tabela 5).

Digitalização e da conversão de documentos

Tendo em vista a inadequação do OCR, decidiu-se pesquisar alternativas que tornassem possível a realização de pesquisas *full-text*. Uma possibilidade encontrada foi a conversão de documentos já existentes no formato digital para o formato PDF, evitando, dessa forma, a necessidade de reconhecimento óptico dos caracteres, uma vez que o documento original não é gerado mediante a captura de imagens com uso de um *scanner.*

Para a realização deste novo trabalho, foi escolhido como instrumento de testes a dissertação de mestrado do professor Omer Pohlmann Filho, co-autor deste artigo, pela facilidade de negociação dos direitos autorais e pela adequação da obra ao trabalho proposto.

A dissertação em questão consiste de um conjunto de arquivos digitais com 247 páginas divididos entre textos elaborados no *Word* 2.0 contendo grande número de tabelas, fórmulas e figuras, e oito tabelas elaboradas no *Excel* 3.0. Pelas características das tabelas, seria muito trabalhoso realizar a conversão para HTML, uma vez que os recursos de tabela disponíveis na linguagem não são satisfatórios para tanto.

O trabalho de conversão consistiu na carga destes arquivos, por meio da utilização de versões mais recentes do *Word* e do *Excel*, e posterior impressão dos mesmos utilizando o *driver* de impressão denominado *Adobe PDF Writer*. Este *driver* se encarrega de capturar a impressão e convertê-la para um arquivo PDF.

Foi necessário agrupar estes arquivos em um só, baseando-se na ordenação encontrada na publicação original (em papel).

Figura 4
Esquema de digitalização da história da PUCRS

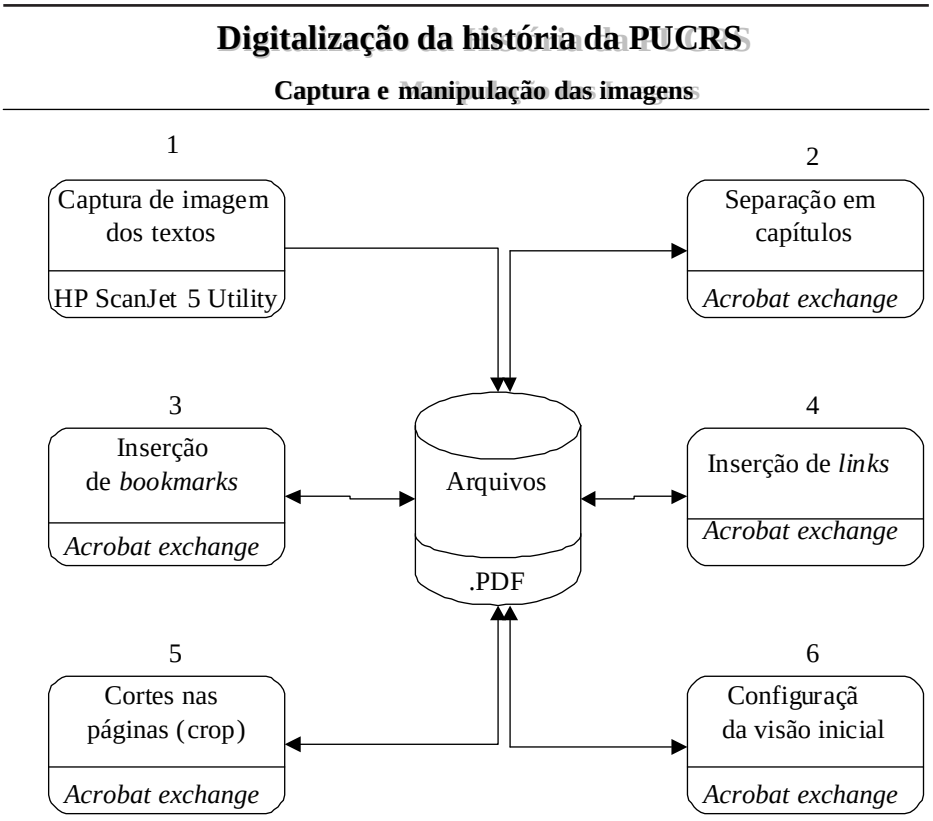


TABELA 5
Os tempos de realização das etapas do processo de digitalização

ETAPAS	TEMPO MÉDIO
Captura da imagens dos textos no *scanner* criação dos arquivos PDF (159 páginas)	16 minutos
Separação dos Arquivos em capítulos	30 minutos
Criação dos índices *link* no índice da obra	40 minutos*
Criação das bookmarks	30 minutos*
Cortes nas páginas (crop)	20 minutos*
Configuração da visão inicial	1 minuto
Tempo médio para transformação de um texto do formato convencional (em papel) para o formato digital, segundo a sistemática proposta (PDF)	137 minutos (2 horas e 17 min.)

* Estas etapas são opcionais ao processo de digitalização e podem variar de acordo com as características da obra.

A conversão produziu um resultado considerado excelente. Sem maiores dificuldades, foi possível agrupar todas as partes que formavam a dissertação (arquivos de *Word* e *Excel*) em um mesmo arquivo no formato digital, sem preocupações maiores com as versões dos *softwares* utilizados.

A seguir, será apresentado o resumo esquemático do trabalho de conversão da dissertação de mestrado (figura 5).

São apresentados também os tempos médios verificados na execução de cada uma destas etapas (tabela 6).

COMPARATIVO ENTRE AS SISTEMÁTICA APRESENTADAS

Ao final deste trabalho, foram considerados positivos os resultados obtidos com a sistemática PDF, uma vez que, na comparação com a sistemática HTML pesquisada anteriormente, esta apresentou vantagens significativas conforme ilustra a tabela 7. Os tempos apresentados são para um número padrão de 50 páginas e 12 figuras.

O principal ponto a favor da sistemática PDF é o tempo total demandado para transformação de maneira totalmente confiável, de uma obra em papel para o formato digital. Este ponto merece especial importância, pois torna viável a realização do processo em larga escala influenciando também na quantidade de recursos humanos necessários para estruturação de um núcleo para realização desta tarefa.

Segundo (Pohlmann[5]), no contexto do projeto de criação de uma biblioteca digital na PUCRS, está prevista a estruturação de um núcleo de digitalização de documentos. Este núcleo terá por objetivo a produção de acervo digital para a biblioteca digital da PUCRS, devendo contemplar os recursos necessários para a realização desta tarefa incluindo *hardware*, *software*, instalações e recursos humanos. Tais recursos humanos receberão treinamento e orientações a partir dos resultados apontados por esta pesquisa.

FIGURA 5
Esquema de conversão da dissertação de mestrado

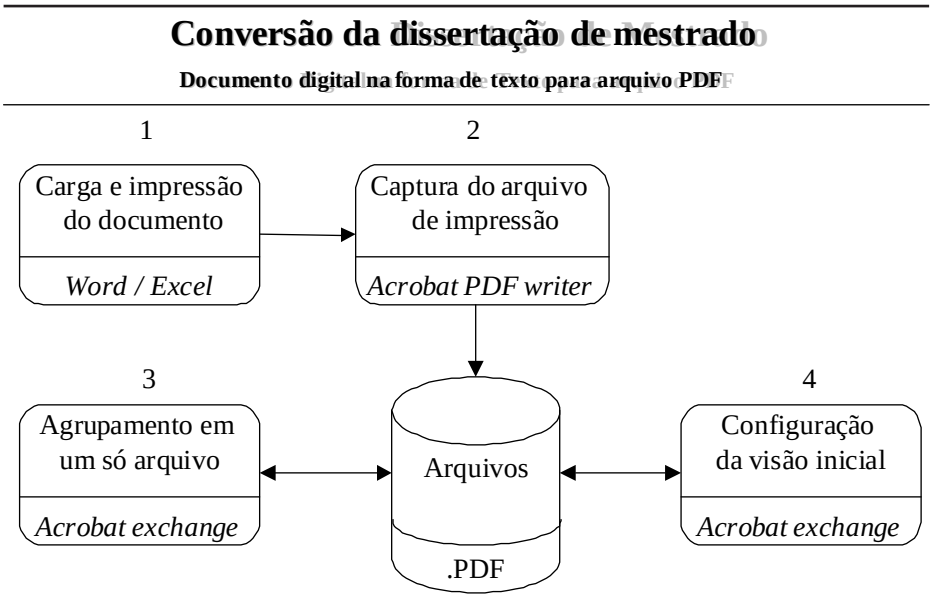


TABELA 6
Os tempos de realização das etapas do processo de conversão

ETAPAS	TEMPO MÉDIO
Carga e solicitação de impressão dos arquivos (1 Texto em MS-Word e 8 tabelas em MS-Excel)	10 minutos*
Captura da Impressão e Conversão para PDF	10 minutos*
Agrupamento dos arquivos	15 minutos*
Configuração da visão inicial	1 minuto
Tempo total de conversão	36 minutos

* Tempos que podem variar de acordo com as características da obra.

TABELA 7
Comparativo entre a Sistemática HTML e as Sistemáticas PDF

Característica	Sistemática HTML Captura	Sistemática PDF – Captura (imagem)	Sistemática PDF – Conversão (texto)
Mantém o leiaute original da obra	Não	Sim	Sim
Possibilidade de manipulação do texto	Sim	Não	Sim
Possibilidade de realização de pesquisas *full-text*	Sim	Não	Sim
Espaço de armazenagem	Pequeno (texto)	Aproxim. 7 vezes maior	Aproxim. 4 vezes maior
Revisão e correção do texto	*400 min	Não há	Não há
Tempo de transmissão via rede	Baixo	7 vezes maior	4 vezes maior
Tempo total aproximado de transformação de uma obra de 50 páginas e 12 figuras	510min	46min	6min

* Observação: Cumpre salientar que o tempo relacionado a sistemática HTML foi determinado, com o trabalho sendo realizado por duas pessoas. Principalmente, a etapa de revisão e correção de texto pode ser agilizada agregando-se mais uma pessoa à equipe de trabalho.

A seguir, apresenta-se uma sugestão de recursos mínimos necessários e os respectivos custos iniciais envolvidos para a formação de um núcleo de digitalização de documentos. Isto visa a permitir a comparação entre as sistemáticas apresentadas, levando em conta também a questão financeira. As sistemáticas de digitalização e conversão de documentos baseadas no formato PDF necessitam dos mesmos recursos e foram, por este motivo, agrupadas na mesma coluna. Os recursos indicados, bem como custos envolvidos, consideram a realidade da PUCRS.

Verifica-se que os valores diferem somente na aquisição das licenças dos *software* e ainda assim possuem valores aproximados. No entanto, deve-se realizar uma comparação relevando o custo relativo aos recursos humanos envolvidos na realização da tarefa nas diferentes sistemáticas, uma vez que a principal diferença identificada foi o tempo dispendido.

Utilizando o mesmo exemplo da tabela 7, é possível verificar o custo dos recursos humanos envolvidos. Para tanto, foi calculado o valor da hora trabalhada do profissional considerando 160 horas mensais (R$ 2.255,90 por profissional / 160 horas/mês = R$ 14,10/ hora por profissional)

Os pontos negativos da sistemática PDF são menos críticos para um projeto em larga escala, tais como o espaço de armazenagem, tempo de transmissão em rede, impossibilidade de realização de pesquisas *full-text* (somente para o caso da captura) e manipulação do texto.

A possibilidade de realização de conversões de documentos já existentes no formato digital para o formato PDF mostrou-se muito eficaz, em especial pelo fato de reproduzir o conteúdo dos documentos exatamente como estes seriam impressos e por facilitar a mesclagem de documentos, constituídos de diversos arquivos de *software* diferentes, gerando um só arquivo PDF.

TABELA 8
Recursos para estruturação do núcleo

	Sistemática HTML		Sistemáticas PDF (Digitalização e Conversão)	
Recurso	Especificação	Preço*	Especificação	Preço*
Hardware	01 Computador: Pentium II 400 Mhz , 64 Mb RAM , HD 6,2 Gb IDE, CD 24x, Monitor 17", Placa de Rede 10/100 Mbps, Porta USB, Windows NT 4.0 WS	6.158,00	01 Computador: Pentium II 400 Mhz , 64 Mb RAM , HD 6,2 Gb IDE, CD 24x, Monitor 17", Placa de Rede 10/100 Mbps, Porta USB, Windows NT 4.0 WS	6.158,00
	01 Scanner: HP ScanJet 6250: conexão USB, Bandeja ADF, Resolução 1200X 999.999 DPI	1.598,00	01 Scanner: HP ScanJet 6250: conexão USB, Bandeja ADF, Resolução 1200X 999.999 DPI	1.598,00
Software	01 Licença Windows NT 4.0	Incluída no computador	01 Licença Windows NT 4.0	Incluída no computador
	01 Licença Caere Omni Page 8.0	549,00	01 Licença Adobe Acrobat 3.01	500,00
Instalações	01 Aparelho de Ar Condicionado 18.000 btus	1.177,51	01 Aparelho de Ar Condicionado 18.000 btus	1.177,51
	02 mesas para microcomputadores	188,00	02 mesas para microcomputadores	188,00
	02 luminárias de 03 lâmpadas com refletores	100,00	02 luminárias de 03 lâmpadas com refletores	100,00
	02 pontos de rede	166,00	02 pontos de rede	166,00
	02 cadeiras com rodízios	156,00	02 cadeiras com rodízios	156,00
Recursos Humanos	02 Remuneração com Encargos Sociais (Bibliotecário Júnior)	4.511,80	02 Remuneração com Encargos Sociais (Bibliotecário Júnior)	4.511,80
Total		14.604,31		14.555.31

* Os valores de referência estão em Reais, cotados na época a 1,71 em relação ao dólar americano

TABELA 9
Comparativo entre despesas com recursos humanos

	Sistemática HTML Captura	Sistemática PDF – Captura (imagem)	Sistemática PDF – Conversão (texto)
Tempo dispendido	510min	46min	6min
Cálculo	R$ 14.10 X 8.5 horas X 2 pessoas	R$ 14.10 X 0.76 horas X 2 pessoas	R$ 14.10 X 0.1 horas X 2 pessoas
Despesa com recursos humanos	R$ 239,70	R$ 21,43	R$ 2,82

Tais características indicam maior facilidade de formação de um acervo contendo documentos recentes (que teoricamente já existem em meio digital), possibilitando inclusive a realização de pesquisas *full-text* nos arquivos que foram convertidos a partir do formato texto, além de favorecer a padronização das publicações digitais e, por conseqüência, o posterior armazenamento, recuperação e manipulação.

Deve-se registrar que os problemas que impõem dificuldades à realização do OCR, sejam eles causados pelo estado de conservação do acervo, tais como manchas, amassados, riscos e anotações, ou ligados a ineficiência dos *softwares* de OCR para tratamento de características, como fórmulas matemáticas, figuras, trechos manuscritos, letras muito pequenas ou borradas, podem ser contornados pela digitalização utilizando a sistemática PDF. Para tanto, basta que as obras sejam digitalizadas como imagens, e será possível realizar a leitura das mesmas, por meio de um arquivo PDF, conforme a aparência original no momento da digitalização.

Neste sentido, é importante que se desenvolvam ferramentas para realização de tratamento óptico nos arquivos de imagens PDF, que filtrem automaticamente as características indesejáveis – adulterações, manchas, amassados entre outros –, melhorando a aparência das obras.

CONSIDERAÇÕES FINAIS

Inicialmente, os trabalhos de construção do acervo da Biblioteca Digital da PUCRS serão realizados com a utilização do *software* Adobe Acrobat para digitalização das obras existentes no formato tradicional (papel) e também para conversão dos documentos já existentes em um formato digital diferente do HTML. Os documentos que já estiverem no formato HTML serão mantidos, pois este formato permite a realização de pesquisas *full-text*, possui tamanho inferior ao PDF e atende às diretrizes definidas para este trabalho apontadas no item Diretrizes de Trabalho.

A disponibilização das obras na Internet será feita, em primeiro momento, pelo *software* ALEPH, que gerencia a catalogação e consulta do acervo da Biblioteca Central.

O ALEPH possui uma interface que permite a realização de consultas pela Internet, possibilitando aos usuários verificar a existência das obras no acervo, bem como a sua disponibilidade para empréstimo. Os recursos de pesquisa do ALEPH baseiam primariamente em autores, títulos e assuntos, mas é possível realizar consultas avançadas acessando qualquer informação constante no registro de cadastramento da obra. Pode-se ainda combinar diversos argumentos de pesquisa mediante a utilização de lógica booleana.

No caso de as obras já existirem no formato digital, o ALEPH fornece um *link* para acesso ao documento na íntegra, permitindo assim que o usuário possa ler a obra digital pela Internet, sem precisar se deslocar até a biblioteca.

As pesquisas *full-text* serão realizadas com recursos de pesquisa do Adobe Acrobat Reader, uma vez que o ALEPH não consegue indexar os textos existentes no formato PDF. No entanto, para acessar os recursos de pesquisa *full-text,* o usuário deve realizar o *download* do documento (copiá-lo para sua máquina) e abri-lo por meio do Adobe Acrobat Reader, pois a consulta ao arquivo PDF realizada com auxílio do *browser* não oferece o recurso de pesquisa *full-text* .

Com o crescimento do acervo digital, torna-se necessário a utilização de outros *softwares* que possuam recursos mais adequados à recuperação e acesso a este acervo, permitindo a utilização de técnicas avançadas de pesquisa com a utilização de linguagem natural, utilização de parâmetros fonéticos e técnicas de inteligência artificial. Para tanto, está prevista a utilização do *software* IBM Digital Library, que recentemente foi disponibilizado e encontra-se em fase de instalação em nosso laboratório.

Como prosseguimento deste trabalho, será realizado, em conjunto com a Faculdade de Biologia e com o suporte da Biblioteca Central, a digitalização de obras de formatos diversos tais como fotografias, sons, textos e imagens. Esses recursos serão utilizados para a elaboração de materiais didáticos, servindo como fonte de pesquisa para a avaliação das características do Adobe Acrobat Reader para suporte a estas mídias, comparando-as com a utilização da linguagem HTML para o mesmo propósito.

A continuidade deste trabalho objetiva definir novas sistemáticas que sejam adequadas à digitalização de obras e materiais existentes nos mais diversos formatos, servindo assim para orientar a estruturação do núcleo de digitalização de documentos e produção de materiais digitais de cunho didático.

REFERÊNCIAS BIBLIOGRÁFICAS

1. POHLMANN, Omer F. Campos, Márcia B. Raabe, André L. John, Fabiana. Viera, Sônia. Em Direção a Criação de uma Biblioteca Digital na Pontifícia Universidade Católica do Rio Grande do Sul: - Uma experiência Prática. II Seminário Internacional de Bibliotecas associadas a UNESCO, Cienfuegos – Cuba. 23 a 27 de maio de 1998.

2. HAIGH, Susan. Optical Character Recognition (OCR) as a Digitization Technology. [Citado em 10 jan. 1998]. Disponível em WWW: [http://collection.nlc-bnc.ca/100/201/301/netnotes/netnotes-h/notes37.htm]

3. CAERE Corporation. A Quantum Leap in Accuracy. [Citado em 11 jan. 1998]. Disponível em WWW: [http://www.caere.com/live/content/products/amaretto/amaretto.htm]

4. ADOBE Acrobat 3.0 Product Information. [Citado em 14 jul. 1998]. Disponível em WWW: [http://www.adobe.com/prodindex/Acrobat/prodinfo.html]

5. POHLMANN, Omer F. Raabe, André L. Direito Autoral no Contexto de Bibliotecas Digitais. III Congresso Internacional de (Tele) Informática Educativa, Santa Fe – Argentina. 14 a 17 de abril de 1999.

Comparative study between systematics of digitisation of documents: Formats HTML and PDF

Abstract

This article presents the resulting experience of Digital Library Group of PUCRS University, for the process of capture and conversion of existing documents from traditional format (paper) to a digital format. The major steps of the process are presented and evaluated using two different systematics: one based on HTML conversion; and other based on the creation of PDF files for Adobe Acrobat Reader software.
Critical issues such as Optical Character Recognition (OCR) and characteristics evaluation of the collection to be converted are approached also.
At the end, is presented a comparative study between the two systematics, identifying positive and negative characteristics to be considered for choosing a work direction.

Keywords

Conversion of documents from the traditional to the digital format; Systematics of conversion to HTML; Creation of PDF files; OCR technologies.

André Luís Alice Raabe

Bacharel em Informática, PUCRS, 1997. Mestrando em Informática, PUCRS 1998.

Omer Pohlmann Filho

Bacharel em Administração de Empresas, PUCRS, 1979. Especialista em Análise de Sistemas, PUCRS, 1981. Mestre em Informática, PUCRS, 1996.

{araabe, omer}@cglobal.pucrs.br