Teses e Dissertações Eletrônicas (ETD): preservação e disseminação de informações científicas

Jayme Leiro Vilan Filho
Professor do Departamento de Ciência da Informação e Documentação (CID) da Universidade de Brasília (UnB)
Campus Darcy Ribeiro – Asa Norte
Brasília – DF – Brasil Cx Postal 04561
CEP 70919-970
jleiro@unb.br

Resumo

Iniciado em 1999 o Projeto Teses do CID tem como objetivos preservar a memória técnica e aumentar o acesso à produção científica do CID/UnB, mais especificamente em relação às teses de doutorado e dissertações de mestrado produzidas pelo CID desde 1980. Um acervo eletrônico de teses e dissertações, também conhecidas como "e-theses" ou "electronic theses and dissertations" (ETD), em hipertexto no formato PDF foi criado a partir de 2001 por meio de uma metodologia que está em constante evolução. São abordados aspectos relacionados com a criação de acervos de ETD, incluindo estruturas de “links” e metodologia desenvolvida com ferramentas de mercado. São mostrados resultados do uso da última versão da metodologia de criação de ETD e apresentadas as ferramentas de editoração, OCR, autoria e acesso a documentos eletrônicos estruturados como hipertexto. Dificuldades operacionais são relatadas, bem como novas metas para a próxima fase do projeto. Conclui que um documento com média de 172 páginas pode ser transformado em uma ETD em aproximadamente 8 horas de trabalho, resultando em um arquivo no formato PDF texto com tamanho médio de 2 MB que pode ser navegado e aceita pesquisa por palavras do conteúdo.

Palavras-Chave: biblioteca digital; teses e dissertações; editoração eletrônica; autoria de hipertexto; ETD; e-these; documento eletrônico; digitalização.

1. Introdução

O Departamento de Ciência da Informação e Documentação (CID) da Universidade de Brasília (UnB) mantém atualmente dois cursos de pós-graduação em ciência da informação nos níveis de mestrado e doutorado. De 1980 a setembro de 2004, foram depositados no CID 197 trabalhos de conclusão de curso sendo 32 teses de doutorado e 165 dissertações de mestrado. Tais trabalhos constituem um acervo localizado no próprio CID, cujo acesso é bastante limitado, com cópias existentes na biblioteca central da UnB e em outras instituições como a biblioteca do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) em Brasília.

As teses e dissertações são documentos com características específicas que dificultam a sua preservação e disseminação:

- pequeno número de exemplares em locais públicos,

- a baixa qualidade editorial,
- a deterioração causada por manipulação durante o processo de reprografia, e
- pequeno número de títulos publicados amplamente.

Uma das alternativas para a preservação e disseminação de informações bibliográficas é a publicação eletrônica em diversos suportes como disquete, CD-ROM e Internet, que tem como principais vantagens[1]:

- redução de custos,
- maior capacidade de disseminação da informação, e
- maior eficiência na administração de coleções na forma de banco de dados.

O termo "e-theses" e, mais recentemente, a expressão "electronic theses and dissertations (ETD)" são amplamente utilizados para designar as teses e dissertações eletrônicas e dentre as razões e vantagens de sua elaboração temos[2]:

- maior liberdade dos autores na demonstração dos resultados de suas pesquisas,
- maior flexibilidade na apresentação de teses,
- inclusão de links ativos para outras pesquisas e fontes eletrônicas,
- inclusão de ilustrações com som e/ou movimentos,
- maior canal de feedback,
- melhor armazenagem da biblioteca digital,
- melhor acesso público à pesquisa corrente,
- disponibilidade do documento a qualquer momento, e
- menos cópias físicas para tratamento, não sendo necessário o pessoal para (re)colocar material.

Além disso, não podemos deixar de destacar o ganho importante na preservação do acervo impresso por:

- evitar danos físicos, tanto pela menor manipulação dos originais quanto pelo acesso mais controlado, e
- evitar desaparecimento de documentos.

2. Objetivos

O Projeto Teses em Dissertações do CID tem como objetivos a preservação da memória do CID e o aumento do acesso à sua produção científica, especificamente teses e dissertações.

Iniciado em 1999, o projeto produziu catálogos bibliográficos automatizados em bancos de dados[3 4 5 6] e em hipertexto[7] e, a partir de 2001, iniciou a digitalização de documentos visando a formação de um acervo de ETD[8 9].

Durante os últimos anos a metodologia de criação de ETD[10 11] passou por melhoramentos sucessivos visando a criação de uma linha de montagem no CID/UnB com equipamentos e sistemas de uso simples.

Além das ETD, a criação de um catálogo automatizado é fundamental tanto para possibilitar a identificação de referências das teses por vários critérios de recuperação quanto para facilitar o acesso aos documentos convencionais e eletrônicos.

3. Metodologia

O projeto foi iniciado em 1999 pelo CID e contou com a colaboração metodológica do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) em 2001.

Recursos obtidos junto à Faculdade de Economia, Administração, Contabilidade e Ciência da Informação e Documentação (FACE), a partir de 2003, possibilitaram a aquisição de novas versões de sistemas, equipamentos e mobiliário, além da contratação de serviços.

Os equipamentos[12] usados atualmente pelo projeto são do Laboratório de Editoração Eletrônica[13] do CID/UnB.

Nessa fase, foram consideradas prioritárias:

- a atualização da metodologia, incluindo equipamentos e sistemas, e
- estudo operacional da linha de produção.

Foram processados dez novos documentos que se somaram aos produzidos em etapas anteriores, totalizando 34 ETD em setembro de 2004[14].

A metodologia representada na Figura 1, mostra a entrada de dois tipos de documentos[15]:

- os arquivos eletrônicos de editores de texto fornecidos pelos autores e convertidos automaticamente para um formato interno de editoração, e

- os originais impressos por meio de scanners.

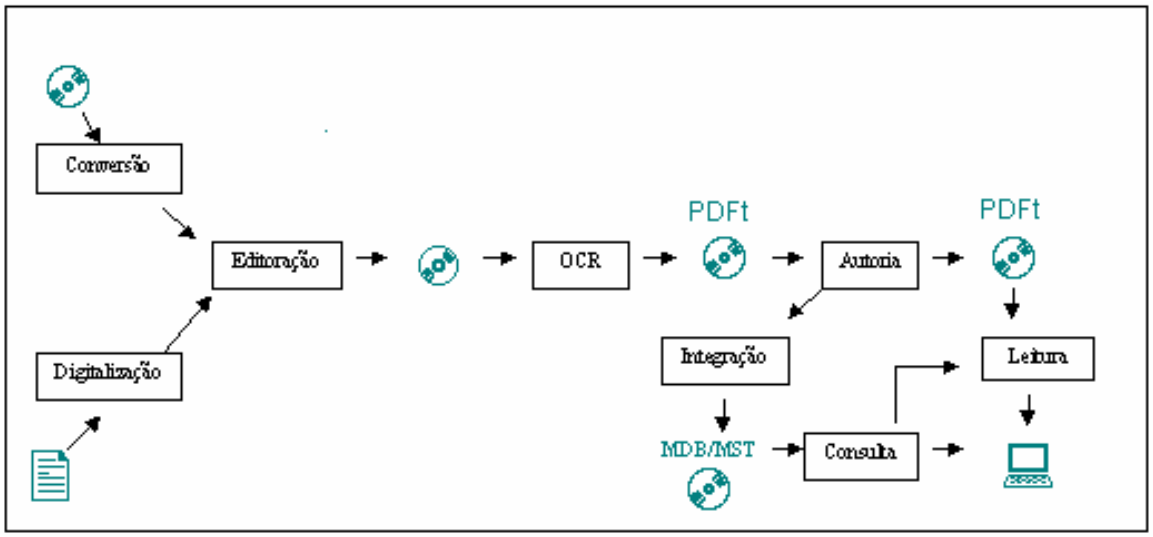

Figura 1 – Visão geral da metodologia de criação de ETD.

Na etapa de digitalização os documentos não encadernados são processados mais rapidamente quando é usado o alimentador automático de folhas soltas. Os parâmetros do scanner devem ser modificados quando há mudança no tipo de informação em um mesmo documento: texto, imagem em tons de cinza e imagens coloridas. Isto exige atenção a cada página processada, e às vezes o reprocessamento é inevitável para obter uma imagem legível. Um ponto interessante é que a maior parte dos documentos não é o original e muitos são cópias reprográficas de baixa qualidade, dificultando o processamento.

Na etapa de editoração são retiradas manualmente as manchas e imperfeições na imagem para facilitar a etapa seguinte de OCR que pode ser executada simultaneamente com a digitalização: enquanto uma nova página é lida pelo scanner a página anterior pode ser analisada e alterada.

Na etapa de OCR as imagens de textos são transformadas em texto, permitindo a posterior pesquisa por palavras e diminuindo substancialmente o tamanho dos arquivos. Apenas as assinaturas, figuras, gráficos e fluxos permanecem como imagem. Cuidados especiais são importantes para evitar:

- ausência de blocos de informação,
- interpretação indevida de blocos, i.e., imagem interpretada como texto,
- detecção de erros de interpretação de caracteres e sinais, especialmente acentos,

- correção do idioma do texto a ser interpretado etc.

As correções são feitas geralmente pela comparação da imagem lida com o texto interpretado pelo sistema com interferência direta do operador e, eventualmente, o documento original é consultado para dirimir dúvidas.

A Figura 2 mostra uma tela de trabalho típica do sistema de OCR onde podemos notar várias janelas com detalhes de uma folha de rosto de dissertação de mestrado em processamento. A janela da esquerda contém um lote ("Batch") de miniaturas das imagens de páginas digitalizadas, possibilitando ao operador escolher a página a ser visualizada em detalhes nas demais janelas. Os blocos de textos identificados automaticamente dentro de uma página são interpretados e podem ser alterados manualmente na janela "Texto" ou reprocessados individualmente à critério do operador do sistema.

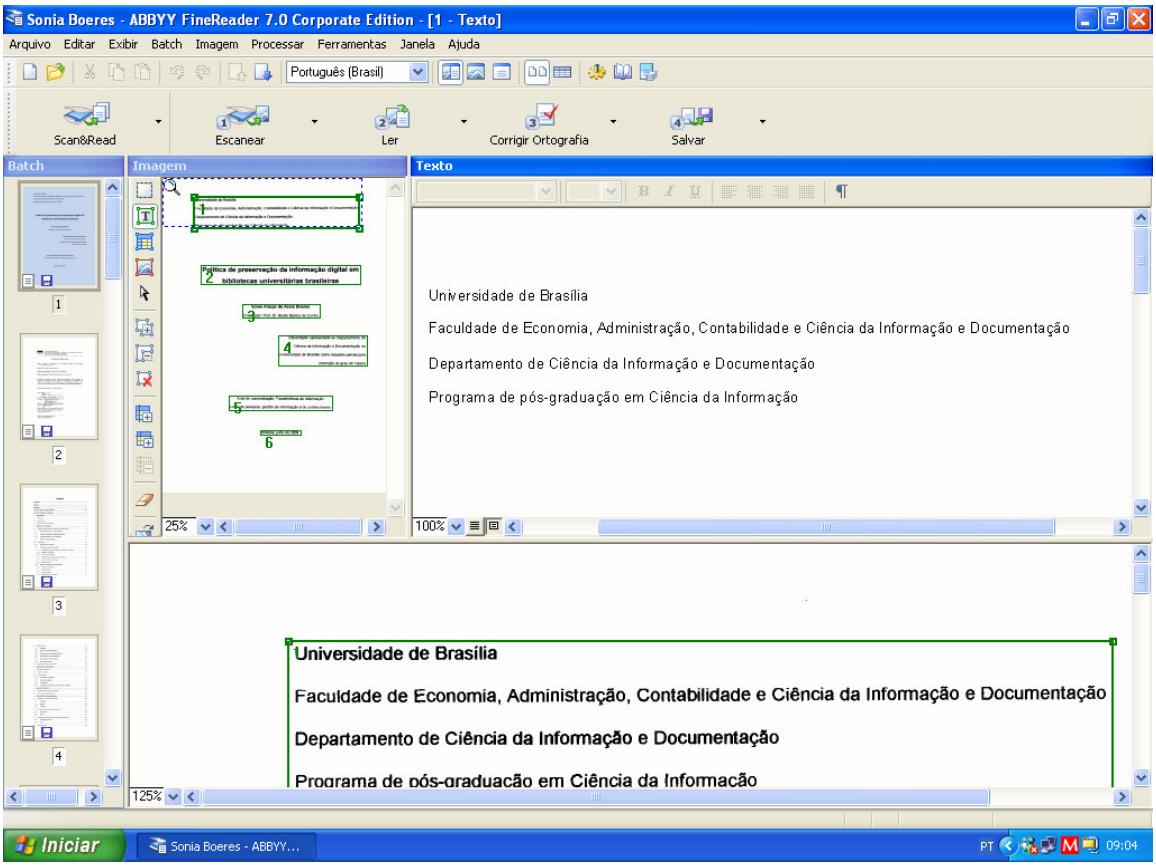

Figura 2 – Tela de trabalho do sistema de OCR

A interpretação prossegue página a página, bloco a bloco, e ao final da sessão de trabalho ou do documento, a ETD é salva em formato PDF texto.

A etapa seguinte, chamada de Autoria, prevê a montagem dos diversos tipos de links de hipertexto no arquivo PDF que possibilitarão a navegação no documento eletrônico. Cada ETD é estruturada, conforme descrito na Figura 3, com:

- links de contexto, como no sumário e nas listas de figuras e quadros, e
- links fora de contexto em uma estrutura chamada de marcadores ("bookmarks"), que reproduz a estruturação do conteúdo da obra em suas várias partes como folha de rosto, folha de aprovação, sumário, listas de figuras, capítulos, anexos etc.

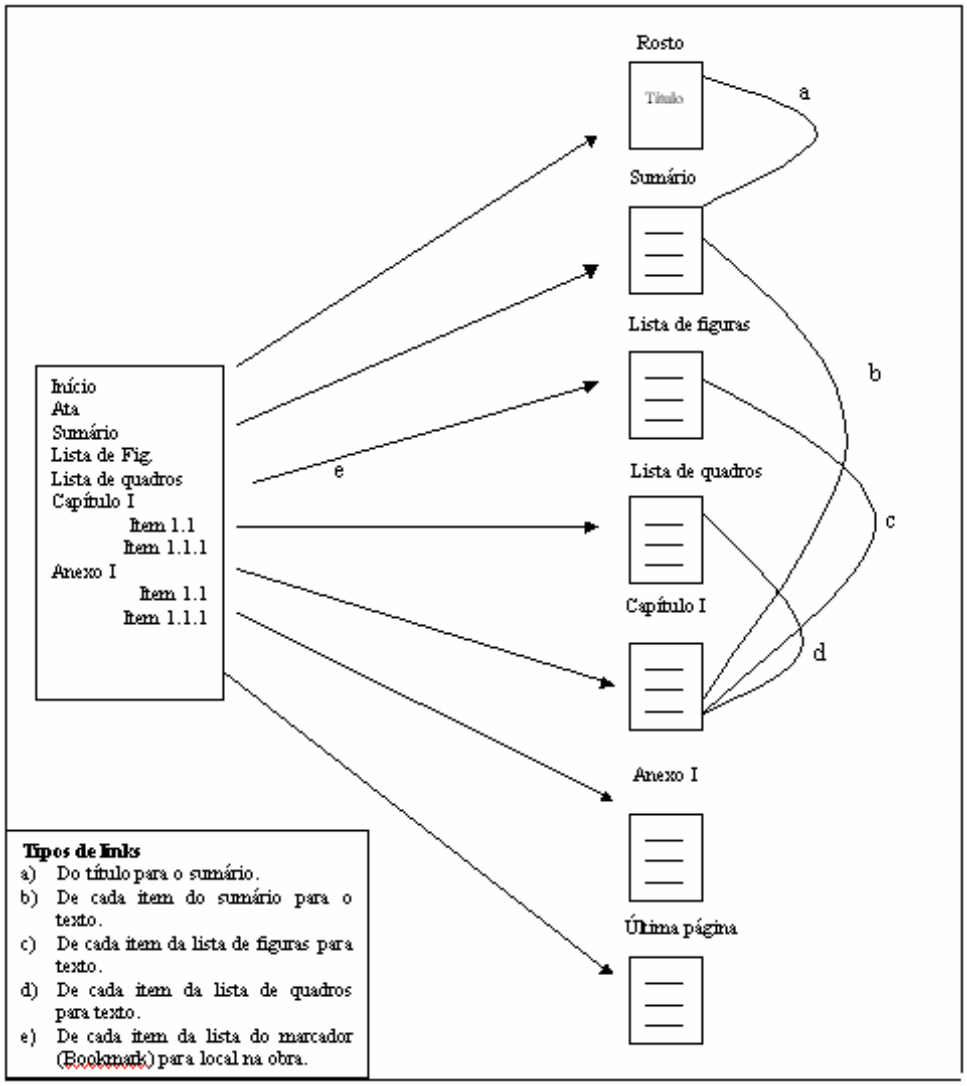

Figura 3 – Estrutura de links de uma ETD

Os demais processos representam as tarefas necessárias para a integração do documento eletrônico ao catálogo bibliográfico em banco de dados e para a leitura e navegação do documento eletrônico pelo usuário final.

Na Figura 4 podemos ver a exibição da versão final de um documento eletrônico na tela de computador, (no lado direito) com seus respectivos marcadores (no lado esquerdo da tela) representando sua estrutura.

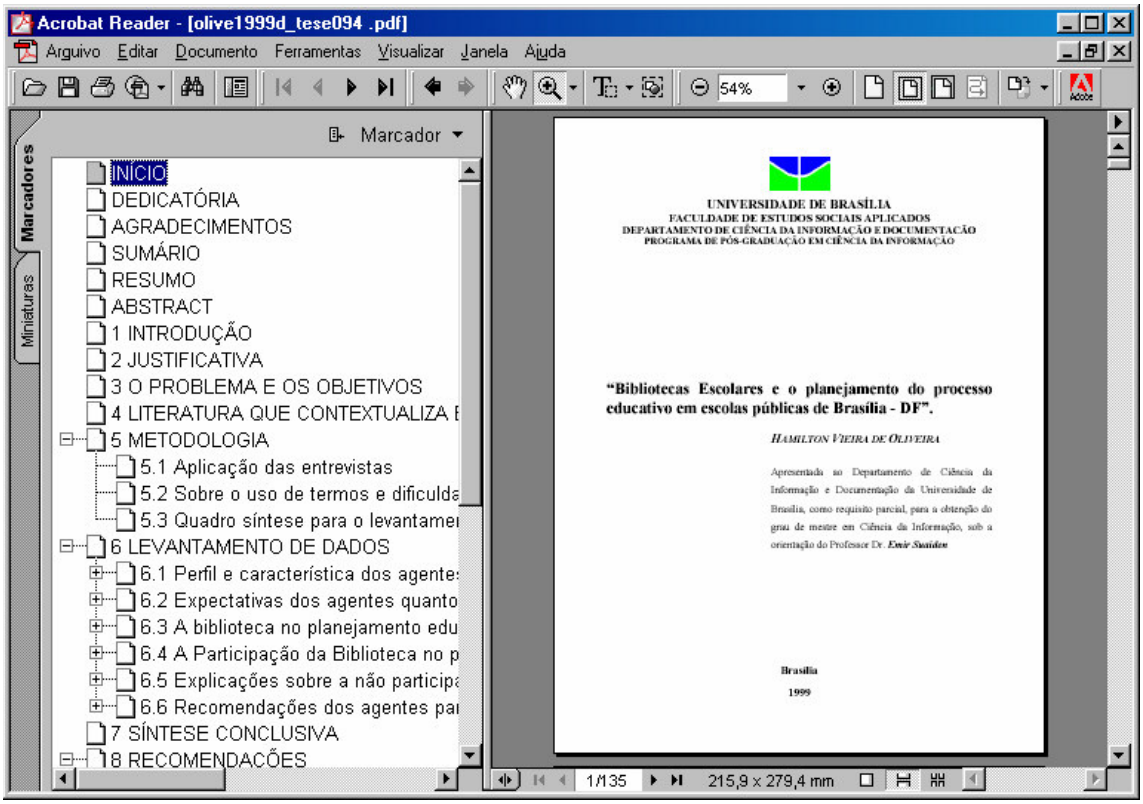

Figura 4 – Exibição da estrutura e da folha de rosto de uma ETD

A estrutura dos marcadores ("bookmarks") inclui links para todas as partes da obra, incluindo folha de rosto, ata, resumo, abstract, lista de figuras, lista de quadros, sumário, agradecimentos, capítulos e anexos. Tanto os capítulos quanto os anexos incluem entradas hierarquizadas para cada seção e subseção. Desta forma, o usuário tem a esquerda da tela toda a estrutura detalhada da obra, podendo exibir na janela da direita a imagem da página escolhida com apenas um toque no mouse.

Assim, o leitor poderá chegar a uma determinada parte da obra, selecionando o item desejado, diretamente da página do sumário ou a partir da

estrutura de marcadores, ou ainda pesquisando por palavra ou expressão por meio de comando específico.

O conjunto catálogo e acervo de ETD, mesmo incompleto, está pronto para ser disponibilizado no CID meio de CD-ROM e intranet. Existe ainda a possibilidade de disseminação externa via internet e CD-ROM que depende de questões jurídicas relacionadas com direitos autorais. Uma estratégia interessante para esta questão é liberar o acesso de cada ETD a medida que o autor autorize formalmente a sua disseminação externa.

Em relação à forma de trabalho a maior parte do projeto foi executada por meio de trabalhos finais de conclusão de curso de bacharéis em biblioteconomia, alunos voluntários e prestação de serviços de ex-alunos.

4. Resultados e Discussão

Na Tabela 1 vemos os resultados considerando o conjunto de documentos processados nesta etapa.

Número de documentos processados	10
Número médio de páginas	172
Número médio de páginas coloridas	6
Número médio de links	188
Tempo médio para montagem dos links	101 min.
Tamanho médio dos arquivos sem páginas coloridas	1,66 MB
Tamanho médio dos arquivos com páginas coloridas	2,56 MB
Tamanho médio dos arquivos	2 MB

Tabela 1 – Resultados desta etapa

Pode-se observar que os documentos que possuem páginas coloridas são 54% maiores, sendo que a média geral é de seis páginas coloridas por documento.

Em relação à metodologia anterior as vantagens foram:

- diminuição do tempo de processamento por página de 3,1 para 2,79 minutos;
- diminuição do número de sistemas usados, que passou de três para apenas dois, um para editoração e OCR e outro para autoria dos links;
- tamanho médio dos arquivos de ETD em PDF é cerca de 37% do tamanho médio obtido em fases anteriores do projeto também em PDF.

Para a implementação de uma linha de produção de ETD com duas ilhas de edição, capaz de atender ao CID e a outros departamentos da UnB, concluímos

que a manipulação de originais encadernados exige a atuação de um operador de scanner, além dos técnicos de editoração. Acredita-se que um operador de scanner pode atender a duas ilhas de edição proporcionando grande economia de tempo na leitura das imagens.

5. Conclusão

A disseminação das ETD por meio de CD-ROM, intranet e internet incluindo a BDTD[16], possibilitará o amplo acesso às informações científicas produzidas pelo CID/UnB, especialmente à medida que as bibliotecas digitais universitárias[17 18] se consolidem, preservando os documentos impressos.

Para os documentos processados nessa etapa do projeto conclui-se que um documento com média de 172 páginas é transformado em uma ETD em aproximadamente 8 horas de trabalho, resultando em um arquivo no formato PDF texto com tamanho médio de 2 MB que pode ser navegado e aceita pesquisa por palavras do conteúdo.

A diminuição substancial no tamanho médio dos arquivos de ETD, de 5,39 MB para 2 MB, possibilita a sua disseminação pela Internet de maneira mais adequada.

A grande quantidade de links facilita a navegação no documento tanto pelos links de contexto, incluídos no próprio texto, quanto pelos links da estrutura auxiliar de marcadores ("bookmarks"). Tais links dão uma visão geral do documento permitindo deslocamentos mais rápidos entre as partes da obra. São necessários procedimentos específicos para controle da qualidade dos links visando principalmente manter a uniformidade dos links no acervo de ETD.

A possibilidade de localização de palavras permite a pesquisa direta no texto facilitando muito a identificação de trechos específicos pelo leitor.

Para subsidiar projetos e estudos de criação de ETD com a estrutura de links apresentada podemos considerar, após analisar dados das duas últimas etapas do projeto, que cada página de ETD leva cerca de três minutos para ser produzida usando-se a metodologia e a infra-estrutura do CID/UnB. Para uma dissertação com 100 páginas, por exemplo, podemos projetar uma duração aproximada de 300 minutos, ou cinco horas, desde o documento impresso até a ETD pronta para navegação.

A diminuição do número de sistemas proporciona ganhos substanciais no tempo de treinamento dos técnicos em editoração e diminui gastos com aquisição de sistemas.

A próxima etapa prevê a implementação de uma linha de produção por meio da duplicação do número de equipamentos, permitindo a criação de duas ilhas de edição, capaz de processar não só o acervo de teses e dissertações do CID, mas também de outros departamentos e faculdades. A equipe de produção será integrada principalmente por alunos bolsistas do CID, além de alunos voluntários.

O controle do acervo de teses e dissertações do CID/UnB por meio de catálogos automatizados e ETD permitirá não só a preservação da memória e a disseminação fácil e ampla das informações contidas nos documentos, mas também proporcionará condições de serem realizados trabalhos mais avançados de indexação automática, bibliometria e metodologia científica, beneficiando toda a comunidade científica interna e externa à UnB.

6. Notas e Referências

[1] PACKER, Abel L. Publicações eletrônicas, controle bibliográfico e recuperação de informação: um enfoque integrado. In: Congresso Regional de Informação em Ciências da Saúde, 3, 1996, Rio de janeiro. **Anais...** Disponível: http://www.bireme.br/cgi-bin/cricS3/text0?id=cricS3-mr1.2-mr1.2.2-04.

[2] MCMILLAN, Gail. Electronic theses and dissertations : merging perspectives. **Cataloging and classification quarterly**, v. 22, n° 3/4, p. 105-125, 1996.

[3] COLOMBELLI, C.M. **Sistema de controle bibliográfico do acervo de teses de doutorado e dissertações de mestrado produzidas no CID**. 1999. Monografia (Bacharelado em Biblioteconomia) – Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília.

[4] MEDEIROS, R.U.F. de. **Indexação e resumo dos documentos do acervo de tese do CID**. 2000. Monografia (Bacharelado em Biblioteconomia) – Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília.

[5] OLIVEIRA, P. H. N. de. **Inclusão de resumos na base de tese do CID**. 2000. Monografia (Bacharelado em Biblioteconomia) – Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília.

[6] MELO, R.de O. **Criação de um aplicativo de teses do CID no Winisis**. 2001. Monografia (Bacharelado em Biblioteconomia) – Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília.

[7] ZIMBA, H. **Implementação do hipercatálogo de tese do CID**. 2000. Monografia (Bacharelado em Biblioteconomia) – Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília.

[8] RODRIGUES, A.M. **Acervo eletrônico de tese do CID**. 2001. Monografia (Bacharelado em Biblioteconomia) – Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília.

[9] ARISAWA, Elisângela Dourado. **Digitalização de tese do CID**: uma nova metodologia. 2002. Monografia (Bacharelado em Biblioteconomia) – Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília.

[10] VILAN FILHO, Jayme Leiro. E-theses do CID. In: Congresso Brasileiro de Biblioteconomia Documentação e Ciência da Informação, 20, 2002, Fortaleza. **Anais**...

[11] VILAN FILHO, Jayme Leiro. E-theses do CID: resultados da nova metodologia. In: Seminário Nacional de Bibliotecas Universitárias, 12, 2002, Recife. **Anais**...

[12] Estão sendo usados microcomputadores com processador AMD Duron 1,3 MHz, 256 MB RAM, HD com 20 GB, leitora/gravadora de CD-ROM, monitor de 17", scanner HP 549C com alimentador de folhas soltas , sistema operacional Windows xp Professional, sistema de OCR ABBYY FineReader OCR 7.0 Corporate Edition e sistema de autoria Adobe Acrobat 6.0 Standard.

[13] Recursos da FINEP pelo convênio FUBRA/FUB/FINEP referente ao Edital CTInfra01/2001 do Ministério da Ciência e Tecnologia (MCT).

[14] ARISAWA, Elisângela Dourado & VILAN FILHO, Jayme Leiro. **Digitalização de tese do CID**: relatório de atividades de janeiro a junho de 2004. 2004. Relatório técnico – Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília.

[15] ARISAWA, Elisângela Dourado. **Digitalização de tese do CID**: manual operacional. 2004. Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília.

[16] Biblioteca Digital Brasileira de Teses e Dissertações (BDTD) coordenado pelo MCT/IBICT. Disponível em: http://bdtd.ibict.br/bdtd/

[17] CUNHA, Murilo Bastos da. Desafios na construção da biblioteca digital. **Ciência da Informação**, v.28, n.3, p.255-266 set./dez. 1999.

[18] CUNHA, Murilo Bastos da. Construindo o futuro: a biblioteca universitária brasileira em 2010. **Ciência da Informação**, v.29, n.1, p. 71-89, jan./abr. 2000.