Comunicado Técnico 71
Novembro, 2006
Campinas, SP

ISSN 1677-8464

Indexação Textual no Sistema de Bases de Dados da Pesquisa Agropecuária

Isaque Vacari[1]
Marcos Cezar Visoli[2]

A Recuperação de Informações (IR - Information Retrieval) em Tecnologia da Informação (TI) constituiu-se um vasto campo multidisciplinar que de forma geral lida com a organização e acesso aos itens da informação. Pode-se dizer que a IR é a arte e a ciência de buscar documentos, informações em documentos ou metadados que descrevam documentos.

Nesse contexto diversos tipos de organizações, classificações, divisões etc., foram criadas para facilitar a recuperação de informações. Um exemplo prático, é o sistema decimal de DEWEY, descrito e tratado em Markey (1984), criado para categorizar artigos em uma coleção de bibliotecas de forma hierárquica. Com a explosão da internet e a criação de repositórios de dados eletrônicos, algumas companhias, tal como Yahoo (Yahoo, 2006), fez da organização e classificação de informações o seu negócio. Outro exemplo, do uso da organização e IR, como oportunidade de negócio, é a área de bibliotecas digitais, incentivado pelo investimento de milhões de dólares em centenas de projetos de pesquisa e desenvolvimento nos Estados Unidos, Europa e outros países do mundo na última década.

Devido ao aumento do uso da internet, a quantidade de dados disponíveis para busca ficou tão vasto que o método baseado em navegações por categorias e subcategorias tornou-se ineficiente e longo para encontrar informações rapidamente. Métodos mais eficientes de busca foram criados para recuperar informações de forma dinâmica e rápida. Um exemplo é o método de busca por palavras-chave aplicado pelo sítio de busca Google (Google, 2006) que recupera conteúdo à partir de expressões de busca em Imensas bases de dados.

A necessidade de localizar informações em enormes bases de dados rapidamente não é apenas uma tarefa dos sítios de busca e sistemas digitais disponíveis na internet, mas também dos computadores pessoais, que possuem uma grande capacidade de armazenar dados. Além disso os computadores pessoais já não são usados para habilidades de computação crua, pois eles também atuam como dispositivos de armazenamento e reprodutores de multimídia. Portanto, uma nova demanda de desenvolvimento de ferramentas de busca tem surgido, não só com a finalidade de encontrar dados armazenados em banco de dados, em sistema de arquivos e diretórios (como documentos de texto, HTML e PDF), mas também em repositórios de imagens, vídeos, áudios e outros formatos multimídia/hipermídia.

[1] *Tecnólogo em Processamento de Dados, Analista da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: isaque@cnptia.embrapa.br)*
[2] *Bacharel em Ciência da Computação, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: visoli@cnptia.embrapa.br)*

Embrapa

Bases de Dados da Pesquisa Agropecuária

As Bases de Dados da Pesquisa Agropecuária - BDPA - (Embrapa Informática Agropecuária, 2006) têm por objetivo contribuir para o cumprimento da missão institucional da Empresa Brasileira de Pesquisa Agropecuária - Embrapa, que é "Viabilizar soluções para o desenvolvimento sustentável do espaço rural, com foco no agronegócio, por meio da geração, adaptação e transferência de conhecimentos e tecnologias, em benefício dos diversos segmentos da sociedade brasileira". A base de dados mais significativa para consulta é o Acervo Documental Embrapa (Embrapa Informática Agropecuária, 2006b), que atualmente é composto por aproximadamente 435.000 registros. A consulta ao acervo e demais bases de dados foram desenvolvidas, inicialmente, com o indexador proprietário Rubicon (Tamarack Associates, 2006), atuando obrigatoriamente sobre o sistema operacional proprietário Windows (marca registrada de Microsoft Inc.). Devido o aumento do acesso ao sítio de busca da BDPA, e a demanda por novos recursos de busca, e as ações para o uso de software livre nas aplicações desenvolvidas pela Empresa, a arquitetura citada, apresentou deficiências como solução tecnológica para evolução do sistema da BDPA. As principais carências são relatadas à seguir:

- sistema operacional proprietário e suscetível a invasões, portanto inseguro para ser usado em um ambiente web;
- elemento indexador proprietário, com dificuldades para inserir melhorias e adaptações;
- uso da linguagem de programação proprietária Delphi (Borland Software Corporation, 2006) para o desenvolvimento do sistema de busca, que não consta mais nas ferramentas de desenvolvimento utilizadas pela Embrapa Informática Agropecuária.

Com o conhecimento e experiência da Embrapa Informática Agropecuária no uso da plataforma Java e de Sistemas Gerenciadores de Bancos de Dados (SGBDs) livres, o problema para a migração e evolução da BDPA passou a ser a escolha de indexadores textuais.

Com base nisto, foram realizados estudos sobre indexadores textuais livres e gratuitos, discutidos em Vacari & Visoli (2006), associado a banco de dados ou não, com o objetivo de subsitituir a plataforma proprietária, inserir novas funcionalidades e criar um ambiente mais adequado para futuras inovações.

Em resumo, nesse estudo foram avaliados os Sistemas Gerenciadores de Banco de Dados MySQL (MySQL AB, 2006) e PostgreSQL (The PostgreSQL Global Development Group, 2006), pelo motivo de ambos terem uma camada de indexação e busca nativa ao banco de dados. Essa camada dispensa o desenvolvimento de software para criação e manutenção da base de dados indexada, deixando essa atividade transparente para a aplicação e o desenvolvedor. A avaliação centrou-se em verificar os recursos de busca disponíveis, a facilidade e o tempo para criação dos índices, a flexibilidade e o tempo de resposta em operações de busca. Após testes efetuados com os dois SGBDs sobre os dados da BDPA, concluiu-se que ambos estão em processo de evolução quanto a sua atuação como mecanismo de busca textual, e para a nova BDPA não seria uma boa escolha.

Decidiu-se, então, ampliar o horizonte dos estudos, para as ferramentas destinadas puramente à indexação textual de documentos estáticos, como: Lucene, SWISH-E - Simple Web Indexing System for Humans – Enhanced (Swish-e, 2006), OpenFTS - FULL TEXT Search Engine (XWare, 2006). De modo, a construir uma solução híbrida, onde a gerência dos dados fica por conta do SGBD, e os dados indexados para busca ficam sob responsabilidade do mecanismo indexador. Essa solução demanda um custo maior em termos de desenvolvimento de software, pois além de gerar duplicidade de dados no banco de índices, é necessário escrever um programa para sincronizar os dados armazenados no banco de dados com aqueles guardados no banco de índices, de forma que, todo registro inserido no banco de dados possui um elemento vinculado no banco de índices. A situação ideal seria automatizar esse trabalho, deixando também, por conta do SGBD, as tarefas relativas à indexação, no entanto, essa solução não está amadurecida suficientemente.

Lucene

Dentre os mecanismos de busca, que atuam sobre documentos estáticos, citados anteriormente, optou-se pelo Lucene (The Apache Software Foundation, 2006b) para atuar como camada de indexação, devido a facilidade de uso, a capacidade de conexão com qualquer SGBD, a disponibilidade para desenvolvimento de aplicações em várias linguagens de programação e ser um software livre e gratuito. Posteriormente, será apresentado a biblioteca Lucene, demonstrado um estudo sobre o uso do Lucene como solução de busca textual para o sistema da BDPA com ênfase para o Acervo Documental Embrapa.

O indexador Lucene foi escrito inicialmente por Doug Cutting em 1997 como fruto de sua experiência com IR, como uma camada de apoio que indexa e torna buscável qualquer dado que possa ser convertido em formato texto. Escrito originalmente na linguagem de programação Java (Sun Microsystems, 2006), visa adicionar as aplicações à capacidade de indexação[3] e busca[4] de texto por palavras-chave, de maneira escalável e com alto desempenho. O Lucene pode atuar como uma camada intermediária entre a aplicação e o banco de dados, desvinculando as tarefas de indexação e busca textual do SGBD para uma camada mais poderosa em termos de recursos oferecidos e performance.

Alternativa como a do Lucene é indicada para aplicações de busca que possuam as seguintes características:

- independência de SGBD para desenvolvimento de aplicações, pois a camada de indexação Lucene é inserida entre a aplicação e o SGBD;

[3] É o processo de conversão que consiste em passar pelos documentos analisando e extraindo informações, tais como as ocorrências de uma palavra, sua localização e freqüência que permitam uma busca rápida sem precisar examinar seqüencialmente todos os documentos. O resultado da indexação é um índice.

[4] É o processo de procurar por palavras no índice visando encontrar os documentos onde elas aparecem.

- suporte a todos os principais recursos presentes em um mecanismo de busca, discutidos em Vacari & Visoli (2006), como: operadores booleanos, truncagem, mascaramento, busca por proximidade, busca por frase, busca por campo específico, agrupamento, palavras de parada (*stopword*), suporte à indexação dos principais tipos de campo (como: texto, número, data, hora e binário), ordenação do resultado da busca por relevância e outros campos definidos pelo usuário, suporte à pesquisa acentuada, marca sobre o texto encontrado etc;
- criação de um banco de índices para busca, cujos dados são provenientes de documentos de texto, HTML, PDF, de banco de dados, ou de qualquer outro formato não texto que possa ser convertido em formato texto através de um filtro adequado;
- linguagens de programação para desenvolvimento de aplicações: Java, Perl, Python, C++, .Net e Ruby;
- sem custos com licença, pois o Lucene é disponibilizado como software livre.
- autonomia tecnológica, pois o Lucene é licenciado como software livre.

A versão binária do Lucene pode ser obtida em The Apache Software Foundation (2006c). No pacote, além de toda documentação, estão duas aplicações de demonstração. O Lucene é apenas um arquivo .jar e basta incluí-lo no *classpath* de *build* e *runtime* da aplicação para começar a usá-lo.

Indexação Textual com o Lucene

A Fig. 1 demonstra como o Lucene indexa documentos de diversos formatos. Antes de indexar documentos HTML, PDF, Microsoft Word (marca registrada de Microsoft Inc.) e outros documentos não textuais, como dados provenientes de um banco de dados, o Lucene converte os dados que farão parte do banco de índices em formato texto. Essa conversão é realizada pela camada de Parser, também conhecida como filtro.

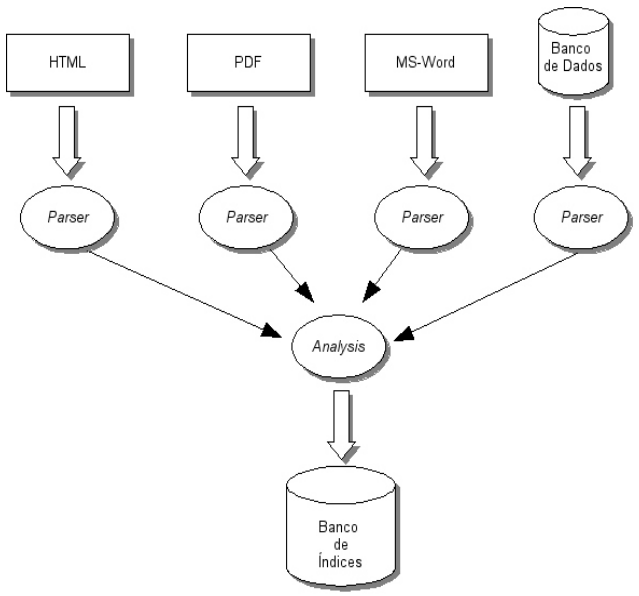

Fig 1. Modelo de indexação implementado pelo Lucene.

A camada de Parser deve ser implementada, pois cada formato de arquivo exige um filtro adequado para o Lucene realizar as tarefas de indexação. Vários parsers foram escritos para trabalhar com documentos de texto (HTML, PDF, Microsoft Word, XML etc.) e estão disponíveis em LUCENE (The Apache Software Foundation, 2006d). A camada *Analysis* é responsável por extrair e converter os textos filtrados para o formato entendido pelo Lucene.

Lucene no Sistema das Bases de Dados da Pesquisa Agropecuária

Para avaliar e validar o uso do Lucene como elemento indexador do sistema da BDPA, decidiu-se elaborar uma validação sobre o Acervo Documental Embrapa. O processo de validação foi elaborado com a finalidade de comparar o indexador Lucene com o indexador proprietário Rubicon, em termos de: performance de indexação e busca, recursos oferecidos e plataforma de desenvolvimento. À seguir será detalhado a infra-estrutura de hardware e software, o conjunto de consultas realizados, e os resultados obtidos com a ferramenta Lucene.

Comparação de desempenho

Infra-estrutura de hardware e software

Para uma análise mais detalhada, em termos de tempo de resposta de busca, foi implementada a tabela de dados do Acervo Documental Embrapa (composta por aproximadamente 435.000 registros e 25 campos, dos quais 22 campos são indexados) no Lucene, para isso foi usado as seguintes ferramentas para desenvolvimento das aplicações necessárias:

- Sistema operacional GNU/Linux **debian-br-cdd-1.0pre5 32 bits** com 1GB de memória RAM e processador AMD Sempron(tm) 2500+;
- Linguagem de programação **Java** (Sun Microsystems, 2006) com o ambiente de programação **Eclipse** (The Eclipse Foundation, 2006) para desenvolvimento da aplicação de criação da base de índices do Acervo Documental Embrapa no formato **Lucene**, bem como, a interface de busca para efetivação dos testes;
- Mecanismo de indexação e busca Lucene (The Apache Software Foundation, 2006b) versão 2.0;
- **Apache TomCat Web Server** (The Apache Software Foundation, 2006a) versão 5.5.17.

Conjunto de consultas, testes e resultados

Como exemplo, foi elaborado um pequeno conjunto de buscas para verificar o desempenho dos indexadores Rubicon e Lucene. Foram escolhidas expressões de busca, que atuam sobre operadores booleanos, frases, em campos específicos, truncagem e busca por proximidade. Para execução dos testes, foi criado uma aplicação Web usando a linguagem de programação Java Server Pages (JSP), essa aplicação além de apresentar os registros encontrados, também, exibe o tempo da busca em segundos. Todas as buscas efetuadas foram replicadas na estrutura da BDPA com o indexador proprietário Rubicon, com o propósito de compará-lo ao indexador Lucene. Vale ressaltar que a estrutura da BDPA com Rubicon está implementada em um

computador distinto, com configurações de hardware e software diferentes da infra-estrutura implementada para os testes com o Lucene (Tabela 1).

Tabela 1. Testes efetuados com tempo de resposta (em segundos) dos mecanismos de busca.

Expressão de busca	*Rubicon (segundos)*	*Lucene (segundos)*
Procurar por (soja ou soybeans ou soybean) em todos os campos	0,407	0,173
Procurar por (soja ou soybeans ou soybean) e trigo e prt em todos os campos	0,343	0,321
Procurar por (soja ou soybeans ou soybean) nos campos título e palavras-chave	Implementa, mas não disponível	0,801
Procurar por assad e (umt ou upc) e "produção agrícola" em todos os campos	Tempo limite de busca esgotado	1,549
Procurar por arroz proximo feijao	Tempo limite de busca esgotado	2,22
Procurar por bras* e agri* em todos os campos	3,032	2,376

Ponderações

Durante a implementação dos mecanismos de busca textual Rubicon e Lucene, foram observadas algumas ponderações importantes presentes nos dois indexadores, relatados na Tabela 2.

Tabela 2. Ponderações dos mecanismos de busca.

Ponderações	*Rubicon*	*Lucene*
Restrições de campos de tabela	Não há.	Campos do tipo data e número devem ser tratados de forma diferenciada.
Licença	Proprietário, o que eleva o custo da aplicação e distribuição da BDPA para novos interessados. Como também, exige o pagamento de licença em melhorias e/ou correções de erros implementadas pela empresa mantedora do Rubicon.	Livre e gratuito.
Ambiente de desenvolvimento	Disponível somente para o sistema operacional Windows com a linguagem de programação proprietária Delphi e o SGBD relacional gratuito Firebird (Firebird Project, 2006).	Independe de SGBD para gerência dos dados, mas a aplicação de criação dos índices e busca de informações deve ser construída usando uma das seguintes linguagens de programação: Java, Perl, Python, C++, .Net ou Ruby.

Nova BDPA

Fruto dos ótimos resultados obtidos com o indexador Lucene em termos de desempenho, performance, recursos de busca e facilidade de desenvolvimento de aplicações, decidiu-se utilizar o Lucene para implementar uma versão atualizada para a BDPA. Também, para garantir a segurança, estabilidade e performance do novo sistema, foram realizados testes de sobrecarga. Foram simulados diversos ambientes de busca, todavia todos os ambientes foram projetados para executar inúmeras buscas por minuto com diversos usuários conectados simultaneamente ao novo sistema. Ao fim da execução dos testes de sobrecarga o novo sistema da BDPA estava em condições normais de uso e o tempo das buscas efetuadas foi altamente satisfatório (buscas complexas foram executadas em menos de 0,5 segundos). O uso desse processo de desenvolvimento de software foi de grande importância para garantir a qualidade do novo sistema de busca da BPDA para seus usuários.

O novo sistema da BDPA está disponível em <http://www.bdpa.cnptia.embrapa.br>, com a seguinte arquitetura:

- sistema operacional: FreeBSD;
- linguagem de programação: Java com suporte a Java Server Pages;
- mecanismo de indexação: Lucene;
- servidor web: Apache TomCat Web Server.

Como citado anteriormente é necessário converter os dados presentes no banco de dados para o formato de texto. Para realizar essa conversão foi criado o software **LIBBD** (Lucene Indexação Busca Banco de Dados) que tem por objetivo extrair os dados armazenados no banco de dados e convertê-los para o formato texto, e também, fornecer uma linguagem de busca amigável para o usuário recuperar os registros desejados. Posteriormente, foi criado o software **AinfoAnalyzer**, cujo papel principal é coletar os dados filtrados pelo LIBBD e transformá-los no formato Lucene. Além disso, ele resolve problemas de acentuação para o conjunto de caracteres ISO-8859-1 e descarta *stopwords* do banco de índices. A seguir será apresentado a nova arquitetura da BDPA com o indexador Lucene (Fig. 2).

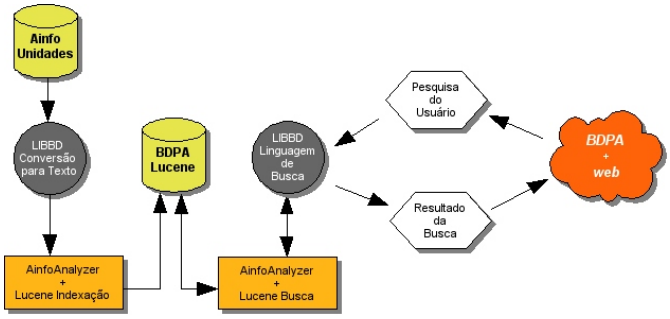

Fig. 2. Arquitetura da nova BDPA com o indexador Lucene.

A experiência positiva com a BDPA desencadeou a criação de novos sítios de busca, todos baseados no mecanismo de busca Lucene, são eles: **Produção Científica Embrapa** (Embrapa Informática Agropecuária, 2006c) e **Rede de Bibliotecas da Área de Engenharia** (Embrapa Informática Agropecuária, 2006d). As Fig. 3, 4 e 5 apresentam a interface dos novos sítios de buscas produzidos com a ferramenta Lucene.

Fig. 3. Sítio de busca da Bases de Dados da Pesquisa Agropecuária.

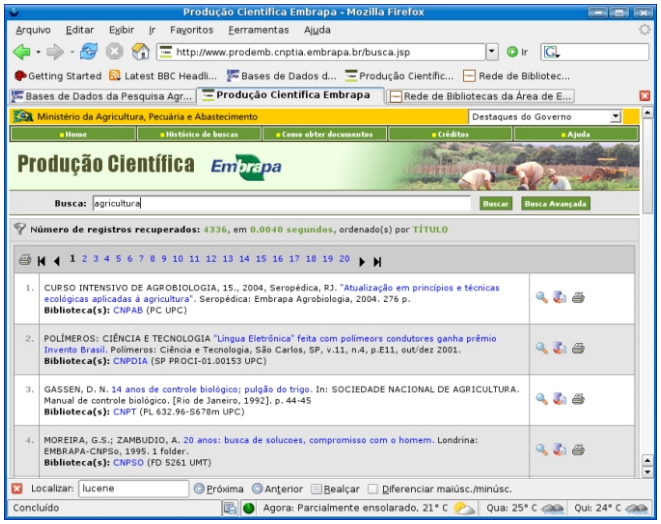

Fig. 4. Novo sítio de busca da Produção Científica Embrapa.

Fig. 5. Novo sítio de busca da Rede de Bibliotecas da Área de Engenharia.

Considerações Finais

Baseado no conjunto de testes realizados, ficou evidente o potencial da ferramenta Lucene como mecanismo de indexação e busca textual. Além de possuir todos os principais recursos (busca booleana, truncagem, mascaramento etc.) de uma ferramenta de indexação e busca textual, ele oferece uma interface de integração amigável com os principais SGBDs existentes no mercado. Suas principais características estão centradas no processo de indexação de diversos formatos de documentos e no tempo de recuperação da busca desejada, onde é possível encontrar documentos armazenados em bases de dados gigantescas em questão de milisegundos, como é o caso da BDPA. Outra característica, não menos importante, é a flexibilidade para o desenvolvimento de aplicações. Com o Lucene é possível escrever softwares para ambiente desktop e Web em diversas linguagens de programação. Vale ressaltar que o Lucene é uma ferramenta livre e gratuita.

A experiência com ferramentas livres, como o Lucene e demais softwares livres utilizados nesse estudo, permitiu à Embrapa Informática Agropecuária ter o domínio tecnológico no uso de indexadores textuais. Como também a arquitetura empregada para o desenvolvimento do sistema da BDPA aliada a boas práticas de desenvolvimento de software (como testes de automatizados de sobrecarga, uso de CVS para controle de versão e processo de desenvolvimento aberto com a utilização do banco de projetos da Embrapa Informática Agropecuária), possibilitou a escalabilidade para a construção de novos sítios web de busca de informações, como Produção Científica Embrapa e Rede de Bibliotecas da Área de Engenharia com garantia de qualidade.

Referências Bibliográficas

THE APACHE SOFTWARE FOUNDATION. **Apache Tomcat**. Disponível em: <http://tomcat.apache.org>. Acesso em: 11 set. 2006a.

THE APACHE SOFTWARE FOUNDATION. **Lucene**. Disponível em: <http://lucene.apache.org>. Acesso em: 11 set. 2006b.

THE APACHE SOFTWARE FOUNDATION. **Apache download mirrors - The Apache Software Foundation**. Disponível em: <http://www.apache.org/dyn/closer.cgi/lucene/java/>. Acesso em: 11 set. 2006c.

THE APACHE SOFTWARE FOUNDATION. **Apache Lucene - contributions**. Disponível em: <http://lucene.apache.org/java/docs/contributions.html>. Acesso em: 11 set. 2006d.

BORLAND SOFTWARE CORPORATION. **Borland IDE**: Delphi Windows .NET application development too w/ object Relational mapping. Disponível em: <http://www.borland.com/delphi>. Acesso em: 15 set. 2006.

THE ECLIPSE FOUNDATION. **Eclipse.org home**. Disponível em: <http://www.eclipse.org>. Acesso em: 11 set. 2006.

EMBRAPA INFORMÁTICA AGROPECUÁRIA. **Bases de dados da pesquisa agropecuária**. Disponível em: <http://www.bdpa.cnptia.embrapa.br>. Acesso em: 11 set. 2006a.

EMBRAPA INFORMÁTICA AGROPECUÁRIA. **Bases de dados da pesquisa agropecuária: Acervo Documental Embrapa**. Disponível em: <http://www.bdpa.cnptia.embrapa.br/index.jsp?url=acervo.jsp&baseDados=ACERVO>. Acesso em: 11 set. 2006b.

EMBRAPA INFORMÁTICA AGROPECUÁRIA. **Produção Científica Embrapa**. Disponível em: <http://www.prodemb.cnptia.embrapa.br/>. Acesso em: 18 set. 2006c.

EMBRAPA INFORMÁTICA AGROPECUÁRIA. **Rede de Bibliotecas da Área de Engenharia**. Disponível em: <http://www.rebae.cnptia.embrapa.br>. Acesso em: 18 set. 2006d.

FIREBID PROJECT. **Firebird - relational database for the new millenium**. Disponível em: <http://www.firebirdsql.org/>. Acesso em: 11 set. 2006.

GOOGLE. **Google Brasil**. Disponível em: <http://www.google.com.br>. Acesso em: 13 set. 2006.

MARKEY, K. **The Dewey Decimal Classification as a library user´s tool in an online catalog**. White Plains: American Society for Information Science by Knowledge Industry, 1984.

MYSQL AB. **MySQL**: the world's most popular open source database. Disponível em: <http://www.mysql.com>. Acesso em: 11 set. 2006.

THE POSTGRESQL GLOBAL DEVELOPMENT GROUP. **PostgreSQL:** the word's most advanced open source database. Disponível em: <http://www.postgresql.org>. Acesso em: 11 set. 2006.

SUN MICROSYSTEMS. **Java technology**. Disponível em <http://java.sun.com>. Acesso em: 11 set. 2006.

SWISH-E. **Swish-e**: home page. Disponível em: <http://swish-e.org>. Acesso em: 11 set. 2006.

TAMARACK ASSOCIATES. **Rubicon full text search**. Disponível em: <http://www.tamaracka.com>. Acesso em: 14 set. 2006.

VACARI, I.; VISOLI, M. C. **Indexação textual usando sistemas gerenciadores de banco de dados livres**. Campinas: Embrapa Informática Agropecuária, 2005. 6 p. (Embrapa Informática Agropecuária. Comunicado Técnico, 70). Disponível em: <http://www.cnptia.embrapa.br/modules/tinycontent3/content/2005/comtec70.pdf>. Acesso em: 11 set. 2006.

XWARE TEAM. **OpenFTS:** open source full text search engine. Disponível em: <http://openfts.sourceforge.net>. Acesso em: 11 set. 2006.

YAHOO! DO BRASIL INTERNET LTDA. **Yahoo! Brasil**. Disponível em: <http://br.yahoo.com>. Acesso em: 13 set. 2006.

Comunicado Técnico, 71

Ministério da Agricultura, Pecuária e Abastecimento

BRASIL UM PAÍS DE TODOS GOVERNO FEDERAL

Embrapa Informática Agropecuária
Área de Comunicação e Negócios (ACN)
Endereço: Caixa Postal 6041 - Barão Geraldo
13083-970 - Campinas, SP
Fone: (19) 3789-5743
Fax: (19) 3289-9594
e-mail: sac@cnptia.embrapa.com.br

1ª edição on-line - 2006