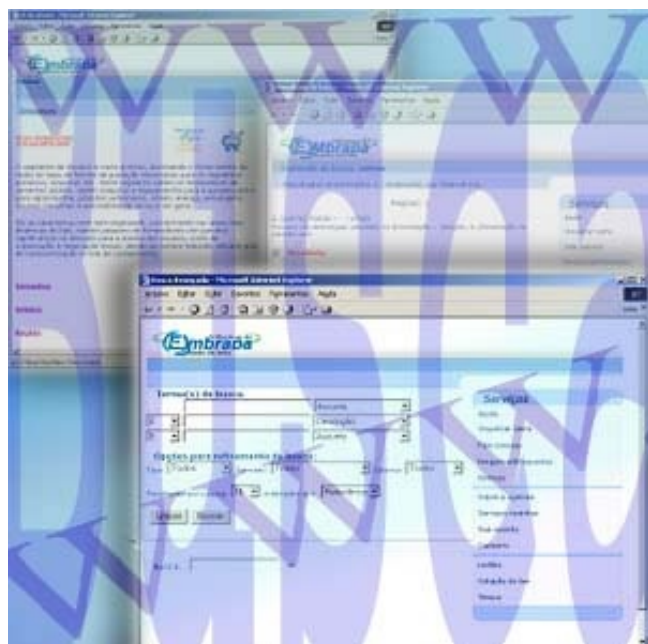


ISSN 1677-8464

Implantação de um Serviço de Busca em Site da WWW

Sergio Aparecido Braga da Cruz¹



A navegação por páginas de um *site* na WWW (*World Wide Web*) ou simplesmente *Web*, é direcionada pelo modo como os projetistas ou as ferramentas utilizadas na sua construção determinam. Dependendo do tamanho, do número, e da maneira como as páginas do *site* estão inter-relacionadas, a busca através dos *hyperlinks* por uma informação desejada pode ser extremamente ineficiente. A utilização de mecanismos de busca automáticos é uma alternativa para a obtenção da informação desejada sem a necessidade de navegação através de *hyperlinks*. Estes mecanismos tentam identificar páginas do *site* relevantes para o usuário baseados em seu conteúdo.

O projeto *Infra-estrutura Tecnológica para a Agência de informação de Pecuária de Corte* é um dos projetos sendo desenvolvidos pela Embrapa Informática Agropecuária em parceria com a Embrapa Informação Tecnológica e Embrapa Gado de Corte, que tem como objetivo construir ferramentas para manutenção do conteúdo de um *site* da *web* que organiza informações técnico-científicas resultantes dos projetos de pesquisa da Empresa Brasileira de Pesquisa Agropecuária - Embrapa. Este *site* é conhecido como Agência de Informação Embrapa, ou simplesmente Agência. Para facilitar a localização de informações desejadas pelos usuários neste *site*, além de outros recursos de navegação, implantou-se um serviço de busca automática.

O objetivo deste documento é descrever como foi selecionado e implantado este serviço de busca no *site* Agência. No próximo item é descrito o funcionamento básico de um mecanismo de busca. Em seguida são comparados alguns mecanismos de busca selecionados para uso no *site* da Agência e no final como foi realizada a implantação do serviço.

Funcionamento de um mecanismo de busca

Um mecanismo de busca automático é um programa que seleciona itens de informação de um conjunto de informações a partir de um critério de busca fornecido pelo usuário. O critério de busca pode ser avaliado diretamente sobre cada item do conjunto de informações, ou, pode ser avaliado usando índices. O uso de índices torna a busca mais eficiente e consiste na extração de informações básicas de cada item do conjunto que auxiliam o processo de recuperação e servem como um apontador para o item como um todo (Kowalski, 1997). No contexto da WWW os itens correspondem a cada documento acessível através dos servidores conectados na *Web*. A indexação, originalmente conhecida como catalogação, é utilizada a muito tempo em bibliotecas na organização manual de

¹ M.Sc em Engenharia Elétrica, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP.
(e-mail: sergio@cnpia.embrapa.br)

documentos por assunto, autor, etc. O processo de indexação automática pode ser classificado em dois tipos: indexação por termos e indexação por conceitos (Kowalski, 1997). Na indexação por termos podem ser utilizadas medidas estatísticas que indicam quão relevante para o usuário é um item de informação contendo um termo específico. A indexação por termos pode também ser realizada por meio da interpretação de informações contidas nos itens de informação utilizando técnicas de processamento de linguagem natural. Na indexação por conceitos os documentos a serem indexados são analisados através de técnicas de processamento de linguagem natural e deles são criados um conjunto de conceitos representados de forma matemática. Segundo Frakes & Baeza-Yates (1999), existe uma grande dificuldade na classificação de serviços de busca devido a diversidade de técnicas que uma mesma ferramenta pode adotar.

A maior parte dos mecanismos de busca automática utiliza a indexação por termos devido a sua facilidade de implementação e eficiência na busca. No caso da WWW, a indexação por termos geralmente gera um arquivo, na estrutura de arquivo invertido, que armazena em que páginas uma determinada palavra pode ser encontrada. Para montagem deste arquivo todas as páginas de um *site* são analisadas e delas são extraídas as palavras de interesse e a sua localização.

Comparação entre mecanismos de busca

Existem inúmeros mecanismos de busca disponíveis na WWW, cada um com as suas peculiaridades (Search Tools Consulting, 2003). A Tabela 1 apresenta a comparação de alguns mecanismos de busca *open source* que foram selecionados dos *sites* indicados pelo Open Source Initiative (2003), levando em conta a sua última data de atualização, estado de desenvolvimento, plataformas de execução, nível de atividade de desenvolvimento e simplicidade. Os *sites* indicados pela *OSI* que foram consultados são *Sourceforge.net* (Open Source Development Network,

2003a) e *FreshMeat.net* (Open Source Development Network, 2003b).

A ferramenta Lucene é fornecida na forma de uma biblioteca Java, sobre a qual deve ser construída uma aplicação. A ferramenta Swish-e permite um uso mais imediato, com menos esforço de desenvolvimento, pois já está disponível sob a forma de uma aplicação.

Existem inúmeras ferramentas que podem ser utilizadas na implementação de um mecanismo de busca em páginas de um *site* (Search Tools Consulting, 2003). No projeto Agência (Cruz et al., 2003), desenvolvido na Embrapa Informática Agropecuária, foi necessária a escolha de um mecanismo de busca que satisfizesse as seguintes necessidades:

- > ferramenta de baixo custo;
- > suporte à busca por texto livre e por campos em metadados - no projeto Agência é necessário que os documentos a serem disponibilizados via *Web* sejam catalogados usando o padrão de metadados desenvolvido pelo Dublin Core Metadata Initiative (2003) e o mecanismo escolhido deve suportar a busca utilizando os seus elementos;
- > uso em ambiente *Web*, podendo também ser migrado para aplicação *stand-alone* - esta característica possibilita que as páginas disponíveis no *site* possam ser armazenadas em CD mantendo a funcionalidade de busca;
- > uso de operadores booleanos - permite que buscas mais elaboradas sejam realizadas;
- > suporte à acentuação - devido à necessidade de indexação e busca de documentos em português;
- > suporte à indexação de documentos em formato MS Word, PDF além dos documentos em formato texto simples.

Tabela 1. Comparação entre mecanismos de busca.

Característica	Swish-e	ht://dig	Lucene	Aspseek
Unix	sim	sim	sim	sim
Windows	sim	não	sim	não
Suporte a tags meta definidas pelo usuário	sim	não, somente são identificadas tags meta de controle da ferramenta	sim, através de módulo extrator adequado	não
Operadores booleanos	sim	sim	sim	sim
Acentuação	sim	sim, caracteres acentuados são tratados como não acentuados.	sim	sim
Tratamento de arquivos em outros formatos	sim, através de filtros adequados	sim, através de filtros adequados	sim, através de filtros adequados	sim, através de conversores externos
Versão <i>stand-alone</i>	sim	sim	não (biblioteca java)	sim
Última versão	SWISH-E RELEASE 2.4.0-pr4 em 03/10/2003	ht://dig 3.1.6 em 01/02/2002	lucene 1.3 rc1 Em 20/03/2003	aspseek 1.2.10 em 19/07/2002
Site de referência	Site do projeto apache http://www.apache.org	-	http://www.bobdylan.com	Site de busca Aspseek http://www.aspseek.com
Site da ferramenta	http://swish-e.org	http://www.htdig.org	http://jakarta.apache.org/lucene	http://www.aspseek.org/

Baseado nestas restrições e avaliando ferramentas utilizadas pelos sites The Apache Software Foundation (2003) e Librarians' Index to the Internet (2003), foi escolhida a ferramenta Swish-e (Swish-e - Simple Web Indexing System for Humans - Enhanced, 2003).

A Swish-e além de suportar todos os requisitos especificados, possui interface simples com outras linguagens, tais como C, Perl, C++ e PHP. Outra característica interessante do Swish-e é que ele é capaz de tratar documentos em formato XML, identificando a sua estrutura e possibilitando a construção da busca a partir de seus elementos. A ferramenta Htdig (The http://DigGroup, 2003), que foi inicialmente selecionada, não suporta a busca por metadados personalizados pelo usuário.

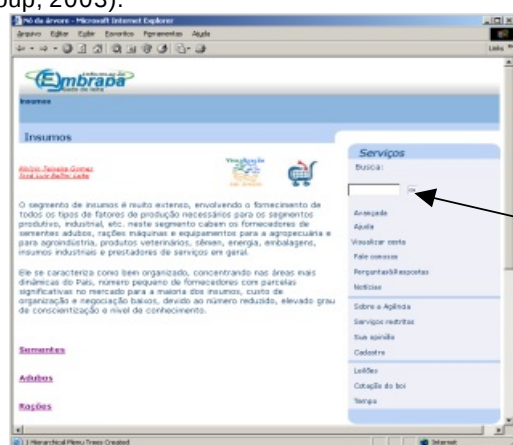
Implantação de um mecanismo de busca na Agência

Para que a ferramenta Swish-e fosse utilizada no projeto Agência foi necessária a construção de formulários de especificação de critério de busca e *scripts* para execução de busca e formatação de resultados. Os formulários e o formato para apresentação dos resultados foram construídos seguindo orientações dos técnicos em catalogação de acordo com as necessidades da Agência. A integração da ferramenta Swish-e com a Web permitindo a execução da busca pelo site da Agência foi realizada usando a linguagem PHP (The PHP Group, 2003).

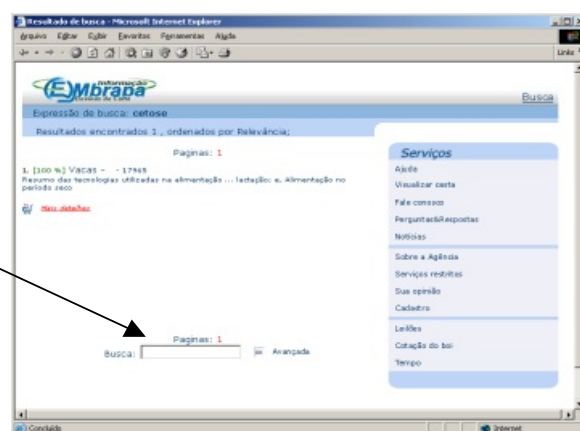
Interface do mecanismo de busca

O projeto Agência conta com um mecanismo de busca que auxilia na identificação de recursos contendo informações de interesse. Os critérios de busca podem ser especificados por meio de formulário simples ou avançado. O formulário avançado permite um maior detalhamento na descrição do critério de busca e o formulário de busca simples conta com somente um campo para especificação do critério de busca. No site produzido pelas ferramentas do projeto Agência este campo para busca simples está disponível nas páginas de resultado de busca, de formulários de busca e nas páginas descrevendo informações de conteúdo de nó, que são páginas sintetizando informações básicas sobre determinado assunto organizado pelo site, Fig. 1.

Para preencher o campo de busca simples basta preencher com as palavras que se deseja procurar. Caso mais de uma palavra seja preenchida, elas serão concatenadas por meio do operador E, ou seja, os documentos satisfazendo a busca precisam ter ocorrências simultâneas das palavras preenchidas. Para que sejam tratadas como uma frase a ser procurada nos documentos, as palavras preenchidas devem estar delimitadas por " (aspas dupla). A existência do caracter * (asterisco) após o primeiro caracter da palavra indica que no local onde está o *, o mesmo pode ser substituído por qualquer sequência de caracteres.

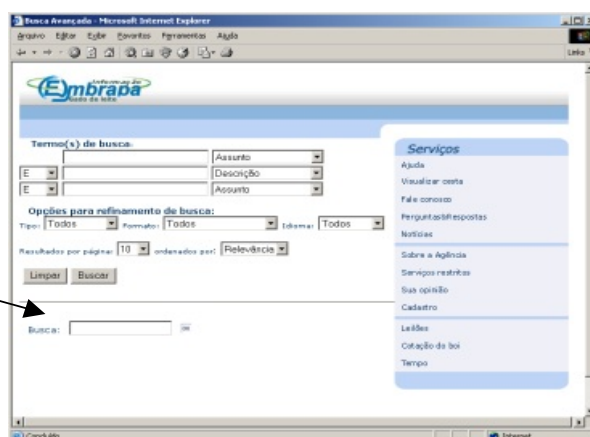


(a)



(b)

campo para
busca simples
simples



(c)

Fig. 1. (a) Página de conteúdo de nó (b) Página de resultado de busca (c) Página de formulários de busca simples e avançada.

Exemplos:

gado - procura por documentos contendo a palavra *gado*

gado pastagem - procura por documentos contendo as palavras *gado* e *pastagem* simultaneamente, mas não necessariamente juntas ou nesta ordem

"verminose bovina" - procura por documentos contendo o trecho *verminose bovina* como preenchido no campo

*ver** - procura por documentos contendo palavras com o prefixo *ver*

O formulário de busca avançada está disponível na página de formulários de busca conforme ilustra a Fig. 2.

O formulário de busca avançado apresenta um número

maior de campos que permitem definir com mais detalhes o critério de busca e o modo de apresentação dos resultados.

Na Fig. 2 os campos indicados por B devem ser preenchidos com as expressões de busca e seguem as mesmas regras descritas para o preenchimento do campo na busca simples. Os campos indicados por C selecionam em que parte do documento a expressão deve ser procurada. Os locais que podem ser selecionados são *Título*, *Descrição*, *Assunto*, *Criador/Autor* e *Todos* e correspondem aos elementos de metadados do padrão Dublin Core utilizados na catalogação dos documentos da Agência e que devem ser utilizados na busca. Os campos indicados por A definem como os critérios de busca preenchidos devem ser logicamente concatenados. Para estes campos existem as opções *E*, *OU* e *NÃO*. A opção *E* indica que os critérios de busca devem ser satisfeitos simultaneamente em um mesmo documento.

formulário de
busca
avançada

Fig. 2 . Página com formulário de busca avançada.

A opção *OU* indica que pelo menos um dos critérios de busca deve ser satisfeito pelo documento. A opção *NÃO* indica que o critério de busca preenchido não deve ser satisfeito no documento e implicitamente será concatenado com os outros campos de busca através do operador *E*. Os campos indicados por D não são obrigatórios e selecionam o tipo, formato e idioma que os documentos devem satisfazer para pertencerem ao resultado da busca. Os campos indicados por E determinam como os resultados da busca devem ser

apresentados. Os resultados podem ser apresentados em grupos com 10, 20, 50 ou 100 registros por página ordenados por relevância ou título. A relevância é determinada pela ferramenta de busca a partir do número de ocorrências dos critérios de busca no documento.

A seguir estão alguns exemplos de preenchimento de formulários de busca avançada.

(a)

(b)

(c)

(d)

Fig. 3. Exemplos de preenchimento de formulário de busca avançada.

O preenchimento ilustrado pela Fig. 3 (a) determina uma busca por documentos contendo *pastagem* no *Título* com apresentação de resultados em grupos de 100 em 100 registros por ordem de relevância. A Fig. 3 (b) busca por documentos contendo *pastagem* no *Título* e *gado* na *Descrição* simultaneamente, com resultados apresentados de 100 em 100 registros por ordem de relevância. A Fig. 3 (c) busca por documentos contendo *pastagem* no *Título*, ou *gado* na *Descrição* e resultados apresentados de 100 em 100 registros por ordem de relevância. A Fig. 3 (d) busca por documentos contendo o trecho *verminose bovina* no *Título* e *gado* na *Descrição* sendo os resultados apresentados de 10 em 10 registros por ordem de *Título*.

Integração Swish-e com PHP

A funcionalidade de busca propriamente dita é suportada pela ferramenta Swish-e. Esta ferramenta está disponível na forma de um programa executável com interface por linha de comando, biblioteca C, ou através da linguagem Perl, sendo que esta última é construída com base na biblioteca C existente. Para integração com o servidor Web foi escolhida a versão executável com interface por linha de comando, a qual é executada por meio de *scripts* na linguagem PHP (The PHP Group, 2003). O uso da versão por linha de comando, simplifica a codificação da integração, pois não é necessário entender o código de implementação do Swish-e, somente suas opções de execução. Isto propiciou um desenvolvimento mais rápido da integração. O PHP é uma linguagem interpretada que possui uma série de funcionalidades que facilitam o tratamento de requisições da Web e a geração dinâmica de documentos em HTML. Com isto torna-se adequada para implementação de pequenas aplicações para a Web, resultando em desenvolvimento rápido e em facilidade de modificação de *layouts* de páginas HTML.

O Swish-e apresenta opções diferenciadas para execução da indexação ou da busca, além de utilizar um arquivo de configuração. O arquivo de configuração do swish-e apresenta vários parâmetros que são detalhadamente explicados em seu manual disponível no *site* da ferramenta Swish-e (Swish-e - Simple Web Indexing System for Humans Enhanced, 2003). Tais parâmetros permitem determinar o comportamento da ferramenta e especificar a localização de arquivos a serem tratados.

O mecanismo de busca implementado no projeto Agência utiliza índices do Swish-e obtidos segundo o comando:

```
swish-e c swish.conf -f AG01 -i diretorio_raiz_agencia
```

O qual indica que o swish-e deve indexar o diretório *diretorio_raiz_agencia*, gerando o arquivo de índices AG01 e utilizando o arquivo de configuração *swish.conf*.

No arquivo de configuração utilizado, a linha mais significativa é:

```
MetaNames id criador descricao titulo assunto tipo
formato idioma categoria
```

a qual especifica que, nos arquivos HTML a serem indexados, os textos definidos para as *tag meta* com os nomes indicados devem ser indexados. Esta linha permitirá que a busca seja realizada por campos de metadado relativos a estes parâmetros. Nos arquivos HTML estes campos são definidos como ilustrado no trecho HTML a seguir:

```
<META name="id" content="RE200.0.70.20362002154821">
<META name="titulo" content="O sistema de produção de gado de
corte implantado no CNPGC">
<META name="criador" content="Corrêa, Afonso Simões">
<META name="criador" content="Arruda, Zenith João de">
<META name="criador" content="Corrêa, Eduardo Simões">
<META name="criador" content="Kessler, Raul Henrique">
<META name="criador" content="Figueiredo, Geraldo Ramos de">
<META name="criador" content="Ruiz, Manuel Enrique">
<META name="criador" content="Carvalho, Elusio Guerreiro de">
<META name="descricao" content="Características físicas da área;
Caracterização do sistema de produção; Pastagens; Rebanho bovino;
Alimentação e manejo do rebanho; Fase de cria; Recria e engorda;
Suplementação mineral Controle sanitário; Seleção e descarte de
animais; Administração e mão-de-obra; Resultados de cinco anos
(1983/84 a 1987/88); Resultados zootécnicos; Resultados econômico-
financeiros.">
<META name="assunto" content="Zootecnia; Veterinária;
Bovinocultura">
<META name="tipo" content="Texto">
<META name="formato" content="aplicação/pdf">
<META name="formato" content="texto/html (.html, .htm)">
<META name="idioma" content="pt">
```

A partir do arquivo indexado, seguindo os parâmetros de execução e de configuração especificados, o papel da integração PHP consiste em:

- > obter valores preenchidos nos formulários de busca e formatá-los de acordo com uma expressão de busca válida, ou seja, compor uma expressão válida para o parâmetro -w do swish-e;
- > selecionar o arquivo de índices adequado;

- > definir valores para os parâmetros m, -b e parâmetro de ordenação;
- > invocar a execução do Swish-e;
- > capturar e analisar o resultado da busca obtido, formatando-o adequadamente.

As linhas de código principais correspondendo à invocação do Swish-e através do PHP são:

```
$strCommandLine = $searchengine.escapeshellcmd(" -m
$resxpag -b $inicio -f $arqindice -p $propriedades
$optordem -w ").$swishexpressao;
```

```
exec($strCommandLine, $resultado, $return);
```

onde *\$strCommandLine* é uma variável PHP que indica a linha de comando devidamente montada para invocação do Swish-e. A invocação é realizada pela função PHP *exec*, a qual captura a saída do comando na variável PHP em *\$resultado* e o estado final da execução do comando em *\$return*.

A Fig. 4 apresenta uma página formatada com resultado de consulta pelo texto *gado* no campo de metadado *Título*. Nesta Fig., toda a formatação HTML, montagem de *hyperlinks* e destaque de termo de busca são realizados pelas funções PHP implementadas.

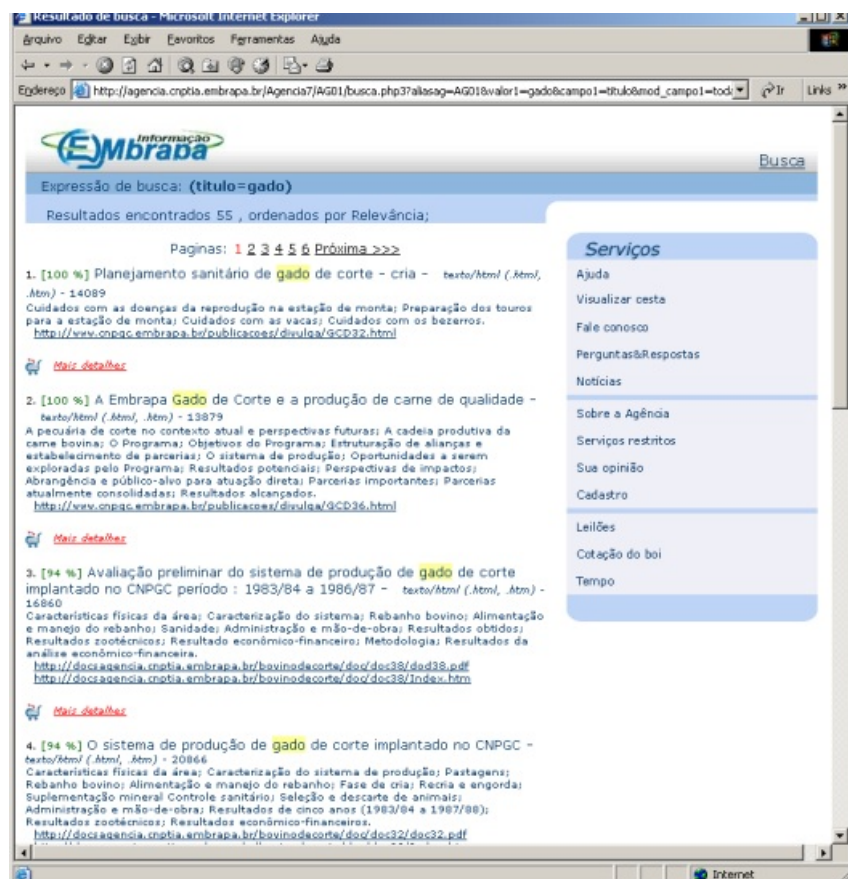


Fig. 4. Exemplo de resultado de busca.

Resultados e discussão

O mecanismo de busca descrito neste comunicado foi implementado na Agência e se encontra funcional desde 3 de abril de 2002. A solução pôde ser rapidamente construída e modificações nos *layouts* HTML são facilmente implementados devido à grande flexibilidade do PHP na integração com a linguagem HTML e com o servidor HTTP. A função PHP *string exec(string comando, string [saida], int [return_var])* é a base para integração de programas por linha de comando com a *Web*, permitindo a sua execução pelo sistema operacional, porém seu uso impõe certos cuidados. Desde que a montagem do comando swish-e é formada por parâmetros fornecidos pelo usuário através de formulários HTML da aplicação de busca, é obrigatório que este comando seja devidamente verificado e formatado de modo que não seja possível a execução de comandos indesejados. A formatação pode ser feita através da função *EscapeShellCommand* do PHP. A montagem de comandos imprevistos através de parâmetros vindos de formulários HTML é uma técnica comum utilizada por usuários mal intencionados para explorar falhas de segurança em aplicações *Web*. O processamento de parâmetros para montagem de uma linha de comando e a execução do comando como ilustrada nesta implementação da busca para o projeto Agência pode ser facilmente implementado em PHP, e serve como um exemplo para outros tipos de programas executáveis por linha de comando. Esta é uma maneira rápida de tornar aplicações que executam por linha de comando disponíveis para o ambiente *Web* sem a necessidade de qualquer modificação de seu código fonte.

A formatação de resultados da execução de programas por linha de comando também pode ser facilmente realizada em PHP. A saída do comando executado pela função PHP *exec* é capturada e armazenada numa *array* de *strings*, onde cada elemento da *array* corresponderá a uma linha de saída. Cada linha de saída obtida pode então ser processada para extração da informação de interesse que posteriormente pode ser apresentada numa página HTML gerada dinamicamente pelo PHP.

Conclusões

Este trabalho tem como principais conclusões:

1. O uso do Swish-e em sua versão por linha de comando mostra-se de fácil uso, sem que seja necessário o conhecimento de seu código fonte para seu pleno uso e integração com a *Web*.
2. A ferramenta Swish-e possui grande flexibilidade no tratamento de documentos a serem indexados, os quais podem ser arquivos texto simples ou estruturados tais como HTML e XML. Em arquivos estruturados o swish-e é capaz de identificar campos e utilizá-los tanto na indexação quanto na busca. Em arquivos HTML o swish-e é capaz de associar campos para algumas *tags* específicas e identificar *tags* do tipo META contendo informações de metadado. Esta capacidade de tratamento de informações sobre metadados armazenadas em arquivos HTML atende às necessidades do projeto, sendo um requisito fundamental no caso da Agência.

3. A linguagem PHP pode ser facilmente utilizada para execução de programas invocados por linha de comando permitindo que sobre eles seja possível a construção de uma nova camada de interface em ambiente *Web*.
4. A integração implementada pode ser facilmente personalizada em outros *sites* quando for necessária a construção de um mecanismo de busca devido à facilidade de configuração da ferramenta Swish-e e pela versatilidade da linguagem PHP na formatação adequada de texto em HTML.

Referências bibliográficas

APACHE SOFTWARE FOUNDATION. The Apache Software Foundation [home page]. Disponível em: <http://www.apache.org>. Acesso em: 13 ago. 2002.

CRUZ, S. A. B. da; MOURA, M. F.; SANTOS, A. D. dos; LEITE, M. A. de A. Arquitetura do gerenciador do conteúdo da Agência de Informação Embrapa. In: CONGRESSO BRASILEIRO DE COMPUTAÇÃO-CBComp, 3.; WORKSHOP DE INFORMÁTICA NA SAÚDE, 3.; WORKSHOP DE INFORMÁTICA APLICADA AO MEIO AMBIENTE, 1.; FÓRUM DE INFORMÁTICA APLICADA À PESSOAS PORTADORAS DE NECESSIDADES ESPECIAIS, 2., 2003, Itajaí. Anais... Itajaí: Univali, 2003. p. 1068-1081. ref. SIN255.

DUBLIN CORE METADATA INITIATIVE. Dublin Core Metadata Initiative (DCMI). Disponível em: <http://www.dublincore.org>. Acesso em: 13 ago. 2003.

FRAKES, W. B.; BAEZA-YATES, R. Information retrieval: data structures & algorithms. In: ERICKSON, J. [Ed.]. Dr. Dobb's essential books on algorithms and data structures. San Mateo: Miller Freeman, 1999. CD-ROM.

THE HT://DIG GROUP. ht://Dig - Internet search engine software. Disponível em: <http://www.htdig.org>. Acesso em: 10 set. 2003.

KOWALSKI, G. Information retrieval systems: theory and implementation. Boston: Kluwer Academic, 1997. 282 p. (The Kluwer International Series on Information Retrieval).

LIBRARIANS' INDEX TO THE INTERNET. Librarians' Index to the Internet - lii.org. Disponível em: <http://lii.org>. Acesso em: 13 ago. 2003.

OPEN SOURCE DEVELOPMENT NETWORK. SourceForge.net. Disponível em: <http://sourceforge.net>. Acesso em: 21 out. 2003a.

OPEN SOURCE DEVELOPMENT NETWORK. freshmeat.net. Disponível em: <http://software.freshmeat.net/>. Acesso em: 21 out. 2003b.

OPEN SOURCE INITIATIVE. Open Source Initiative OSI. Disponível em: <http://opensource.org>. Acesso em: 21 out. 2003.

THE PHP GROUP. PHP: Hypertext Preprocessor. Disponível em: <http://www.php.net>. Acesso em: 14 ago. 2003.

SEARCH TOOLS CONSULTING. Search tools - information, guides and news. Disponível em: < <http://www.searchtools.com/index.html> > . Acesso em: 13 ago. 2003.

SWISH-E - Simple Web Indexing System for Humans - Enhanced. Disponível em: < <http://swish-e.org> > . Acesso em: 13 ago. 2003.

Comunicado Técnico, 50

**Ministério da Agricultura,
Pecuária e Abastecimento**
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 - 2003

Ó Todos os direitos reservados.

Comitê de Publicações

Presidente: *Luciana Alvim Santos Romani*
Membros Efetivos: *Carla Geovana Macário, Ivanilde Dispatto, Marcia Izabel Fugisawa Souza, Marcos Lordello Chaim, Suzilei Almeida Carneiro*
Suplentes: *Carlos Alberto Alves Meira, Eduardo Delgado Assad, José Ruy Porto de Carvalho, Maria Angelica Andrade Leite, Maria Fernanda Moura, Maria Goretti Gurgel Praxedis*

Expediente

Supervisor editorial: *Ivanilde Dispatto*
Normalização bibliográfica: *Marcia Izabel Fugisawa Souza*
Editoração eletrônica: *Área de Comunicação e Negócios*