



MINISTÉRIO DA EDUCAÇÃO
FUNDO NACIONAL DE DESENVOLVIMENTO DA EDUCAÇÃO – FNDE
DIRETORIA DE TECNOLOGIA – DIRTE



Brasília, 22 de outubro de 2012

NOTA TÉCNICA
Software de Inteligenciamento
Empresa INTELIPAR/VERT

PREGÃO ELETRÔNICO Nº 25/2012.

1. CONTEXTO

Avaliação de conformidade com os requisitos de edital do software de inteligenciamento da empresa INTELIPAR/VERT.

No dia 17/10/2012, a empresa INTELIPAR/VERT apresentou a solução de software de inteligenciamento, requisito do **EDITAL DE PREGÃO ELETRÔNICO Nº 25/2012**.

2. AVALIAÇÃO

Referente à apresentação realizada pela empresa INTELIPAR/VERT, a apresentação ocorreu no ambiente da empresa, utilizando notebook e computador desktop, ambos acessando o servidor web instalado na própria máquina (localhost).

A solução é composta de módulos isolados para cada uma das atividades:

Módulo de inteligenciamento de documentos – PDFIND, com os seguintes recursos:

- Conversão de PDF/arquivos de imagem para PDF's pesquisáveis (inteligenciamento);
- Reconhecimento de caracteres (OCR);
- Geração de conteúdo de marcação de caracteres reconhecidos (OMR); Interface gráfica de correção de caracteres reconhecidos;
- Criação de indexadores de conteúdo textual;
- Indexação e pesquisa textual, full-text, booleana, fuzzy, com repositório baseado em sistema de arquivos e sistema de permissionamento;
- Solução inteiramente web, podendo ser instanciada na forma stand-alone ou distribuída (em rede), ou acesso via internet;
- Solução baseada em Java Server Faces, com interface adaptável;


1



MINISTÉRIO DA EDUCAÇÃO
FUNDO NACIONAL DE DESENVOLVIMENTO DA EDUCAÇÃO – FNDE
DIRETORIA DE TECNOLOGIA – DIRTE

- Utilização de tabelas de banco de dados para armazenamento dos indexadores.

Módulo de pesquisa textual – EASYSEARCH, com os seguintes recursos e características:

- Interface WEB da aplicação no idioma português do Brasil;
- Desenvolvido inteiramente em Linguagem de Programação Java 1.6;
- Compatibilidade, para o ambiente de execução da aplicação, com os sistemas operacionais Linux, com versão Kernel 2.6 ou superior e Windows 2003 Server ou superior, ou Solaris 10 x86 ou Oracle-SPARC;
- Compatibilidade, em relação à infra-estrutura de execução da aplicação, com Apache Web Server e IIS utilizando o contêiner WEB JBoss Application Server;
- Compatibilidade, em relação ao ambiente cliente, estação de trabalho, com navegadores Mozilla, Firefox, Internet Explorer 7 ou superior;
- Capaz de indexar arquivos em um repositório local/remoto de pelo menos os seguintes tipos de arquivos de natureza textual: DOC (Word 97), DOC (Word 2007), DOCX, RTF, PDF, XSL, TXT, PPT, PPS, ODT, ODS, ODP, HTM, SXC, SXD, SXI, SXW, STC, STI, STW, STM, MHT, HTML, TXT, MSG e MP3 (para indexação de tags ID3) e arquivos de imagens com conteúdo predominantemente textual: TIFF, PNG, JPG, JPEG e BMP.
- Demonstra ser capaz de criar um índice otimizado que mantém os documentos indexados provenientes das fontes de indexação.
- Possui módulo de pesquisa que mantém uma única interface Web para busca baseada em palavras-chave sobre o conteúdo indexado.
- O processo de busca é insensível a variações de caixa (maiúsculas / minúsculas);
- Desconsidera palavras muito comuns sem conteúdo semântico ("stopwords"), como preposições e artigos, no processo de busca;
- Destaca ("highlight") os termos pesquisados;
- Suporta os operadores booleanos: AND, OR, NOT e os símbolos + e – representando, respectivamente, a presença e a ausência de um termo de pesquisa;
- Suporta buscas por frase exata;



MINISTÉRIO DA EDUCAÇÃO
FUNDO NACIONAL DE DESENVOLVIMENTO DA EDUCAÇÃO – FNDE
DIRETORIA DE TECNOLOGIA – DIRTE



- Suporta o operador de proximidade, isto é, a busca por duas ou mais palavras desde que elas estejam separadas por “n” palavras quaisquer, sendo “n” um parâmetro definido pelo usuário em tempo de busca;
- Oferece recursos avançados de pesquisa, como busca fonética, busca por radicais e busca por similaridade de grafia, de forma que o usuário tenha ferramentas para aumentar a cobertura de buscas que retornam nenhum, ou poucos, resultados conforme metodologia Fuzzy;
- Na busca fonética considera apenas a carga fonética das palavras de busca informadas pelo usuário e de todo o conteúdo indexado, de forma que palavras escritas de forma diferente, mas com sonoridade idêntica, sejam consideradas iguais.

O Módulo possibilita ainda:

- a. Criação ilimitada de estruturas taxonômicas.
- b. Exibição do total de itens em cada classificação.
- c. Criação de metadados para cada item da taxonomia.
- d. Aplicação de filtros.

Utiliza o motor de busca textual LUCENE, que é:

- a. Um motor de busca “open Source”;
- b. O Apache Lucene é escrito na linguagem de programação Java, com diversos recursos de busca textual.
- c. Contém tecnologia adequada para praticamente qualquer aplicação que requer pesquisa de texto completo, especialmente multi-plataforma.
- d. Mais detalhes em: <http://mac.softpedia.com/get/Developer-Tools/Apache-Lucene.shtml>

O responsável técnico da empresa apresentou também outros recursos utilizados na solução:

1. Tesseract-OCR: Engine de reconhecimento de caracteres open source. <http://code.google.com/p/tesseract-ocr/>
2. XPDF: Pacote open source de manipulação de arquivos PDF, incluindo conversão para diversos formatos. <http://www.foolabs.com/xpdf/about.html>

3



MINISTÉRIO DA EDUCAÇÃO
FUNDO NACIONAL DE DESENVOLVIMENTO DA EDUCAÇÃO – FNDE
DIRETORIA DE TECNOLOGIA – DIRTE

3. iText: Pacote open source de manipulação de arquivos PDF.
<http://itextpdf.com/>
4. PDFBox: Pacote open source de manipulação de arquivos PDF.
<http://pdfbox.apache.org/>
5. Apache Lucene: Pacote <http://mac.softpedia.com/get/Developer-Tools/Apache-Lucene.shtml>

3. RECOMENDAÇÕES DA EQUIPE TÉCNICA

A equipe técnica identificou que os módulos da solução apresentada, incluindo inteligenciamento de documentos, pesquisa textual e motor de reconhecimento de caracteres atendem aos requisitos exigidos no edital, no entanto, faz-se necessário a integração entre os 03 (três) módulos apresentados. Desta forma, solicitamos que a empresa INTELIPAR/VERT realize a integração dos módulos no prazo máximo de 30 (trinta) dias corridos, preferencialmente em um único container (JBoss EAP 6), devendo ser reapresentado à equipe técnica para avaliação e aprovação.

Celso Ribas Junior
Consultor TI – DIRTE/FNDE

Jose Olávio Castelo Branco Felício Junior
CTU – DIRTE/FNDE

Mauro Cândido Moura
Diretor Substituto DIRTE/FNDE