SOFTWARE LIVRE PARA IMPLEMENTAÇÃO DE REPOSITÓRIOS DIGITAIS E PROVEDORES DE SERVIÇOS: EXPERIÊNCIA DA EMBRAPA INFORMÁTICA AGROPECUÁRIA

Isaque Vacari[(1)], Marcos Cezar Visoli[(2)], Fernando César Lima Leite [(3)], Sabrina Déde de Castro Leite Degaut Pontes [(4)], Massayuki Franco Okawachi [(5)], Victor Paulo Marques Simão[(6)], Luís Eduardo Gonzales [(7)] e Maria Goretti Gurgel Praxedes [(8)]

[(1)]Embrapa Informática Agropecuária. Email: isaque@cnptia.embrapa.br, [(2)]Embrapa Informática Agropecuária. Email: visoli@cnptia.embrapa.br, [(3)]Universidade de Brasília. Email: fernandodfc@gmail.com, [(4)]Embrapa Informação Tecnológica. Email: sabrina.pontes@sct.embrapa.br, [(5)]Embrapa Informação Tecnológica. Email: massayuki.okawachi@sct.embrapa.br, [(6)]Embrapa Meio Ambiente. Email: victor@cnpma.embrapa.br, [(7)]Embrapa Informática Agropecuária. Email: eduardo@cnptia.embrapa.br, [(8)]Embrapa Informática Agropecuária. Email: goretti@cnptia.embrapa.br

RESUMO

O trabalho apresenta a experiência técnica da Embrapa Informática Agropecuária (CNPTIA), instituição de pesquisa em tecnologia da informação para agricultura, no contexto de *software* livre para implementação de repositórios digitais e provedores de serviços. Mais especificamente, serão relatados os estudos e as soluções de *software* livre escolhidas para viabilizar a inserção da Embrapa no modelo alternativo de comunicação científica denominado Acesso Aberto (*Open Access Initiative*) [1], que visa promover o acesso livre e irrestrito à produção científica desenvolvida pelos pesquisadores e instituições, e que tem na implantação de repositórios digitais, sejam eles institucionais ou temáticos, uma de suas estratégias principais.

PALAVRAS-CHAVES

Acesso aberto; Software livre; Recuperação de informação; Indexação textual; DSpace; JOAI; Solr; Empresa Brasileira de Pesquisa Agropecuária; Embrapa; Comunicação científica; Gestão da informação científica; Repositório institucional; Repositório digital; Provedor de serviços.

1 INTRODUÇÃO

No contexto de organização, tratamento e disseminação da informação na Internet, o Sistema Embrapa de Bibliotecas (SEB), em cooperação com o projeto Acesso Aberto à Informação Científica na Embrapa (liderado pela Embrapa Informação Tecnológica) e a Embrapa Informática Agropecuária, definiram como objetivo a construção dos repositórios digitais: Infoteca (Informação Tecnológica em Agricultura) e Alice (Acesso Livre à Informação Científica da Embrapa) e do provedor de serviços: Sabiia (Sistema Aberto e Integrado de Informação em Agricultura), com o propósito de favorecer a transferência de tecnologias produzidas pela Embrapa, o aumento da visibilidade da produção científica dos pesquisadores, e da própria instituição, por meio da maximização do acesso à sua própria produção intelectual, e por consequência, a ampliação dos resultados de pesquisa realizados na Embrapa, como também, acessar fontes de informações externas de interesse da Embrapa. O presente trabalho relata a construção dos repositórios digitais Infoteca e Alice com o *software* livre DSpace [2] e o desenvolvimento do provedor de serviços Sabiia com os *softwares* livres JOAI (*Java Open Archives Initiative*) [3] e Solr *Enterprise Search Server* [4]. Conclui-se com relatos dos resultados obtidos com as ferramentas de *software* livre escolhidas, bem como, as principais linhas que se irão prosseguir no desenvolvimento dos repositórios digitais e do provedor de serviços.

2 *SOFTWARE* LIVRE PARA CRIAÇÃO DE REPOSITÓRIOS DIGITAIS: DSPACE

Após a decisão da criação de repositórios digitais para comunicar, divulgar e socializar o conhecimento produzido pela Área de Pesquisa e Desenvolvimento (P&D) da Embrapa, o CNPTIA realizou um levantamento de alternativas livres então existentes para construção de repositórios digitais. Como resultado do levantamento foi escolhido o *software* livre DSpace para construção da primeira versão de repositório digital para a Embrapa. A escolha da ferramenta DSpace deu-se pela maior experiência da Embrapa Informática em atuar com sistemas desenvolvidos com a linguagem de programação Java [5], o sistema gerenciador de banco de dados PostgreSQL [6] e a ferramenta de indexação textual Lucene [7], bem como o uso do padrão de metadados Dublin Core [8] para descrever os recursos digitais, o processo descentralizado de submissão de novos documentos, a facilidade de customização da interface, a capacidade de armazenamento de vários formatos digitais (texto, pdf, som, vídeo etc) e exemplos de repositórios digitais implementados com a ferramenta DSpace no Brasil, como: Biblioteca Digital do Senado Federal, Repositório Digital da Universidade do Rio Grande do Sul, Instituto Antônio Carlos Jobim etc.

O DSpace é uma alternativa tecnológica gratuita criada pelo *Massachusetts Institute of Technology* (MIT) [9] em parceria com a *Hewlett Packard* (HP Labs) [10], para construção de repositórios digitais dedicados ao gerenciamento da produção intelectual de uma instituição contemplando a reunião, armazenamento, organização, preservação, recuperação e, sobretudo, a ampla disseminação da informação científica. O sistema foi disponibilizado gratuitamente em Novembro de 2002 sob a licença *BSD open source license* [11] de acordo com o padrão *OASIS reference model* [12] e tem sido largamente utilizado no MIT, em diversas universidades dos Estados Unidos, Europa e instituições da América do Sul.

3 IMPLEMENTAÇÃO DOS REPOSITÓRIOS DIGITAIS INFOTECA E ALICE

A implementação dos repositórios digitais Infoteca e Alice com o DSpace iniciou-se em Abril de 2008. O plano de implantação constitui-se em 5 etapas principais:

3.1 Instalação, configuração, tradução e customização

Decidiu-se usar a versão 1.5.1 do DSpace com o modelo de interface JSPUI para construção dos repositórios digitais. O processo de instalação e configuração do sistema DSpace seguiu as instruções do manual de Instalação e Configuração do DSpace [13], nessa etapa definiu-se os requisitos de *hardware* (capacidade de armazenamento em disco, memória, processador etc) e *software* (sistema operacional, servidor de aplicações *web*, sistema gerenciador de banco de dados, sistema de *backup* etc) necessários para o funcionamento do sistema. Após a instalação e configuração do sistema DSpace, e realizada uma análise mais detalhada de todas as suas funcionalidades por um grupo restrito de bibliotecários, iniciaram-se as tarefas de evolução do *software*, entre as melhorias destacam-se: a tradução existente para o português do Brasil, o ajuste e a personalização da interface gráfica, a alteração do esquema padrão de metadados Dublin Core, a implementação de *thumbnails*, a implementação da busca no texto completo, a implementação e a tradução do *add-on* de estatística da Universidade do Minho [14] para o português do Brasil, a implementação da navegação no formato de árvore hiperbólica [15] e a correção de *bugs* existentes na versão 1.5.1 do DSpace. Essa fase ocorreu entre Abril de 2008 e Agosto de 2009.

3.2 Carregamento automático de documentos digitais

Para viabilizar a disponibilização da produção técnico-científica digital da Embrapa em um único ambiente institucional com o DSpace, a Embrapa Informática, criou um *software* específico (baseado no módulo de importação e exportação de documentos do DSpace) para coletar e armazenar no DSpace toda literatura espalhada nos sítios *web* de suas 43 unidades de pesquisa. Essa ação permitiu o resgate e o carregamento automático de aproximadamente 11.500 publicações no DSpace. Essa fase ocorreu de Setembro de 2008 à Outubro de 2008.

3.3 Abertura ao público externo

A atividade de carregamento automático de documentos digitais favoreceu o lançamento da primeira versão do repositório digital. Em outubro de 2008, com o nome de Repositório Digital Embrapa, a empresa disponibilizou todos os trabalhos técnico-científicos, em formato digital, gerados pela área de Pesquisa, Desenvolvimento & Inovação na Internet, abrangendo livros, folhetos, capítulos de livros, trabalhos apresentados em eventos, além dos artigos das revistas Pesquisa Agropecuária Brasileira e Cadernos de Ciência & Tecnologia. De imediato, o Repositório Digital Embrapa trouxe os seguintes benefícios:

- Armazenamento, preservação e disponibilização da produção intelectual da Embrapa em um único local, dispensando a manutenção de servidores e sistemas locais diferentes em cada unidade da Embrapa;
- Maior garantia de acesso e disponibilidade do sistema e dos recursos digitais para a sociedade, uma vez que a infraestrutura de acesso à Internet da Embrapa Informática tem-se mostrado mais eficiente em relação às demais unidades da empresa.

3.4 Constituição de Unidades piloto para alimentação do Repositório Digital Embrapa

A quarta etapa do processo, constitui-se da definição de unidades de pesquisa piloto, com o objetivo de alimentar o Repositório Digital Embrapa com novos documentos e verificar sua integração com o sistema de administração de bibliotecas existente, denominado Ainfo [16]. Com base em critérios de localização geográfica e capacidade de acesso à Internet, relevância e quantidade de documentos técnico-científicos produzidos e disponibilidade imediata em atuar como unidade piloto foram selecionadas e convidadas quatro centros de pesquisa (Embrapa Florestas, Embrapa Informática, Embrapa Meio Ambiente e Embrapa Semiárido) e duas unidades centrais (Embrapa Informação Tecnológica e Embrapa Sede) para participar do projeto piloto.

Para cada unidade da Embrapa criou-se uma comunidade no Repositório Digital Embrapa, e cada comunidade recebeu uma coleção denominada Memória Técnica, com exceção da Embrapa Informação Tecnológica que definiu suas coleções. Nessa etapa definiu-se as políticas de alimentação (alimentação realizada pelo bibliotecário, ou seja, os bibliotecários assumem o papel de mediadores do depósito de documentos no repositório digital junto aos autores) e acesso ao Repositório Digital Embrapa, o *workflow* e os atributos mínimos (título, autor, palavra-chave e ano de publicação) para submissão de documentos.

Apesar de boa parte dos documentos terem sido carregados em lote, como se pode constatar na Figura 1, nota-se o aumento frequente de novos documentos submetidos diretamente no repositório digital.

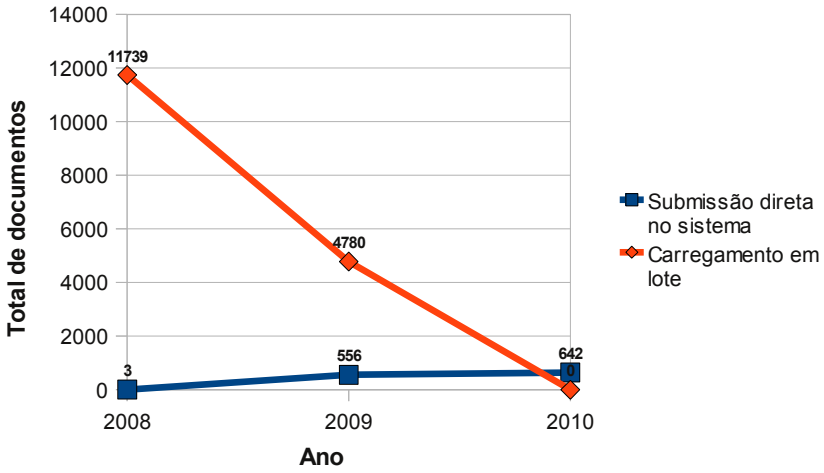

Figura 1: Documentos por tipo de submissão

3.5 Segmentação do Repositório Digital Embrapa para públicos apropriados, criação dos repositórios digitais Infoteca e Alice

A implementação do Repositório Digital Embrapa respondeu satisfatoriamente ao anseio institucional de organizar toda produção técnico-científica da empresa em único ambiente, dispensando a disponibilização dos arquivos eletrônicos nas páginas das unidades de pesquisa. No entanto, verificou-se que além da informação científica, que é produzida por pesquisadores e destinada à outros pesquisadores, a Embrapa produz intensamente informação tecnológica, que é produzida por pesquisadores e destinada ao diversos segmentos da produção rural com estrutura diferenciada e linguagem apropriada (não científica) para transferência de tecnologia.

A partir dessa constatação decidiu-se segmentar o Repositório Digital Embrapa em dois novos repositórios digitais para públicos distintos. A Infoteca (Informação Tecnológica em Agricultura), destinada a produtores rurais, técnicos agrícolas e profissionais da área, abrangendo informações sobre tecnologias produzidas a partir da pesquisa científica realizada na instituição como vídeos, áudios, cartilhas, *folders* e diversas publicações das linhas editoriais de transferência de tecnologia. E o repositório institucional Alice (Acesso Livre à Informação Científica da Embrapa), em fase de construção, destinado a comunidade e a comunicação científica, contendo artigos em periódico, publicações em eventos, teses e dissertações, capítulos de livros e livros.

A Infoteca está disponível no endereço http://www.embrapa.br/infoteca desde setembro de 2009, enquanto o repositório institucional Alice encontra-se em fase de definição e implementação do *designer* gráfico (logomarca, cores e interface) e sua abertura para o público externo está prevista para o segundo semestre de 2010.

4 *SOFTWARE* LIVRE PARA CRIAÇÃO DE PROVEDORES DE SERVIÇOS

O modelo de acesso aberto da Embrapa [17] lida tanto com a informação que resulta de suas atividades de P&D quanto com as necessidades de informação de seus pesquisadores para executar tais atividades. Ou seja, além de permitir a gestão e o aumento da visibilidade de sua informação científica por meio do repositório institucional Alice, o modelo aberto da Embrapa sistematiza também o acesso à informação científica externa proveniente de áreas de interesse da instituição.

Com vistas a permitir o acesso rápido e facilitado aos diferentes provedores de dados de acesso aberto externos, está em andamento a construção de um provedor de serviços compatível com o protocolo de comunicação OAI-PMH (*Protocol for Metadata Harvesting*) [18]. O protocolo OAI-PMH foi criado com o objetivo de desenvolver e promover padrões de interoperabilidade entre repositórios digitais com vistas à facilitar a recuperação e a disseminação eficiente de conteúdos produzidos pelas comunidades científicas. O provedor de serviços sistematiza o fluxo e canaliza toda a produção científica em áreas de interesse institucional disponíveis em ambiente de acesso aberto. Um exame preliminar identificou um total de 261 provedores de dados em áreas de interesse da Embrapa. Dentre eles, 52 periódicos nacionais, 74 periódicos estrangeiros, 27 repositórios institucionais e temáticos, 4 repositórios de conferências, e 104 periódicos nacionais e estrangeiros disponíveis no SciELO. Através de uma única interface de busca, espera-se que os usuários internos e externos realizem pesquisas em todos os periódicos e repositórios institucionais e temáticos, previamente selecionados, e coletados de acordo com as políticas estabelecidas. A definição e a escolha das soluções de *software* livre para implementação do provedor de serviços foram resultados de uma análise que contemplou as seguintes etapas:

4.1 Identificação de *softwares* livres existentes para implementação do módulo de coleta de dados

Nessa etapa foram identificadas seis alternativas livres:

- PKP Metadata Harvester (http://pkp.sfu.ca);
- MOD OAI (http://www.modoai.org);
- OAI Harvester (http://webservices.itcs.umich.edu/mediawiki/dlxs14/index.php/OAI_Harvester/);
- OAI Arc (http://oaiarc.sourceforge.net);
- JOAI Harvester (http://www.dlese.org/dds/services/joai_software.jsp);
- OAI Havester OCLC (http://www.oclc.org/research/software/oai/harvester2.htm).

A identificação de *softwares* livres existentes para implementação do módulo de coleta de dados ocorreu por meio de pesquisas no sítio *web* SourceForge [19]. Em seguida, realizou-se a instalação e configuração de cada uma das ferramentas identificadas. Essa etapa ocorreu entre os meses de maio e junho de 2009.

4.2 Análise preliminar dos software livres selecionados

Nessa etapa foram analisadas as seis soluções de *software* livre identificadas na etapa anterior mediante os critérios: i) frequência de atualização e disponibilização de novas versões; ii) mecanismo de coleta de dados (interface *web* ou linha de comando); iii) interface *web* para administração de provedores dados e iv) documentação e suporte apropriado para instalação e configuração do *software*. Baseado nos critérios acima, foram selecionados três *softwares* para testes mais sofisticados, apresentados na tabela 1.

Características	OAI Arc	JOAI Harvester	PKP (Publick Knowledge Project)
Sítio *web*	http://oaiarc.sourceforge.net	http://www.dlese.org/dds/services/joai_software.jsp	http://pkp.sfu.ca
Download	http://sourceforge.net/projects/oaiarc	http://sourceforge.net/project/showfiles.php?group_id=198325&package_id=269889	http://pkp.sfu.ca/harvester_download
Demonstração	http://www.rcaap.pt	http://www.dlese.org/library/index.jsp	http://harvesters.sfu.ca/demo/index.php/browse
Patrocinador (*Sponsor*)	Old Dominion University.	Digital Library for Earth System Education (DLESE).	University of British Columbia.
Licença	GNU/GPL.	GNU.	GNU.
Última versão (*Last release*)	1.0 (15-Out-2006)	3.0.14 (02-Fev-2009)	2.3.0 (18-Fev-2009)
Suporte (Features, Bugs, Fórum etc)	**Bugs:** http://sourceforge.net/tracker/?atid=497555&group_id=61532&func=browse	**Documentação:** http://www.dlese.org/oai/docs/harvester.jsp	**Bugs:** http://pkp.sfu.ca/bugzilla **Fórum:** http://pkp.sfu.ca/support/forum/index.php
Linguagem de Programação	Java.	Java.	PHP.
Formato de armazenamento dos dados coletados	Banco de dados MySQL.	Arquivos XML, um arquivo XML para cada registro.	Banco de dados PostgreSQL.
Sistema de coleta	Linha de comando e/ou interface *web* JSP.	Linha de comando e/ou interface *web* JSP.	Linha de comando e/ou Interface *web* PHP.
Sistema de busca	Interface *web* JSP.	Interface *web* JSP.	Interface *web* PHP.

Tabela 1: Softwares Livres para coleta de dados OAI-PMH

4.3 Instalação e Configuração dos softwares selecionados

O processo de instalação e configuração de cada uma das ferramentas selecionadas, na etapa anterior, seguiu o manual de instalação e configuração disponibilizado junto com o *software*.

4.4 Testes de performance de coleta de dados

A etapa de testes de performance de coleta de dados subsidiou a escolha do *software* livre mais adequado para a construção do provedor de serviços. Para construção dos testes de performance, a equipe do Projeto Acesso Aberto à Informação Científica na Embrapa selecionou dez provedores de dados (com aproximadamente 16.000 registros) de interesse da Embrapa, com o objetivo de simular um ambiente robusto de coleta de dados. Além dos testes de performance, foram investigados, de forma mais aprofundada, os recursos oferecidos por cada um dos *softwares*. A tabela 2 ilustra o resultado dos testes de performance realizados.

	OAI Arc	**JOAI Harvester**	**PKP Harvester**
Coleta de dados			
Interface de administração para coleta de dados	Sim, mas a coleta de dados é realizada por linha de comando por meio do *script* **harvester.sh**.	Sim.	Sim, com a possibilidade de efetuar a coleta de dados por meio da interface *web* ou através de linha de comando utilizando o *script* **harvester.php**.
Formato de metadados suportados para coleta	Somente oai_dc.	O *software* suporta o formato de metadados definido pelo provedor de dados, geralmente oai_dc.	O *software* suporta o formato de metadados definido pelo provedor de dados, geralmente oai_dc.
Metadados coletados	Todos, mas somente os atributos principais são apresentados: *title*, *creator*, *subject* etc.	Todos.	Todos.
Formato de armazenamento dos dados coletados	Banco de dados relacional MySQL.	Arquivos no formato XML.	Banco de dados relacional MySQL ou PostgreSQL.
Registro de *log* dos dados coletados	Na interface de administração de coleta de dados é apresentado a informação *Last Harvester Time*, e durante a execução do *script* **harvester.sh** são apresentadas mensagens de execução do programa.	Além de apresentar o *log* de coleta de dados mais recente é possível consultar os *logs* de coleta de dados anteriores.	Somente durante a execução do *script* **harvester.php**.
Coleta de dados por comunidade/coleção	Sim.	Sim.	Sim.
Coleta de dados de dois ou mais repositórios digitais simultaneamente	Não.	Sim, por meio do recurso *Scheduler* presente na interface de administração	Sim, através da execução em paralelo do *script* **harvester.php**.

	OAI Arc	JOAI Harvester	PKP Harvester
		de coleta de dados.	
Recoleta de dados de repositórios digitais coletados anteriormente	Sim, por linha de comando por meio do *script* **harvester.sh**.	Sim, por meio da interface de administração de coleta de dados.	Sim.
Coleta de dados por linha de comando	Sim, por meio do *script* **harvester.sh**.	Sim, por meio do *script* **harvester**.	Sim, através do *script* **harvester.php**.
Bloqueio de acesso ao módulo de coleta de dados	Sim, por meio do arquivo de administração de usuários do servidor *web* Apache Tomcat.	Sim, por meio do arquivo de administração de usuários do servidor *web* Apache Tomcat.	Sim, por meio de usuário e senha definidos durante a instalação da ferramenta.
Ativação e desativação de repositórios digitais	Sim.	Sim, pode-se optar pela exclusão definitiva, ou então, pela desativação do repositório digital.	Sim, mas o repositório digital desativado é excluído definitivamente do sistema de coleta de dados.
Recursos de busca			
Mecanismo de indexação e busca	MySQL FULL TEXT.	Lucene.	Zend Search Lucene.
Interface *web* para busca	Sim, por meio de uma interface *web* JSP.	Sim, por meio de uma interface *web* JSP.	Sim, por meio de uma interface PHP.
Busca por repositório digital/coleção	Sim.	Não.	Sim.
Busca por campo	Sim, somente pelos campos autor, título e *abstract*.	Não.	Sim.
Suporte para busca com operadores booleanos	Sim, suporta os seguintes operadores booleanos: AND, OR, + e -.	Sim, suporta os seguintes operadores booleanos: AND, OR, + e -.	Sim, suporta os seguintes operadores booleanos: AND, OR, + e -.
Agrupamento do resultado da busca (*facets*)	Sim, oferece agrupamento do resultado da busca por repositório coletado, ano e assunto.	Não.	Não.
Opções de ordenação do resultado da busca	Sim, oferece as opções de ordenação do resultado da busca por: relevância, data e assunto.	Não.	Não.
Opções de filtro para o resultado da busca	Sim, oferece as seguintes opções para filtro de busca: repositório, coleção, assunto e data.	Não.	Sim.
Navegação (*browse*) por repositório digital	Sim.	Não.	Sim.
Outras formas de navegação	Não.	Não.	Não.
Visualização do registro original	Sim.	Sim.	Sim.
Highlight sobre o resultado da busca	Sim.	Sim.	Não.
Reindexação da base de dados para busca	Sim, por meio do *script* **parse.sh**.	Sim, por meio da interface de administração de coleta de dados.	Sim, por meio da interface de administração de coleta de dados.
Recursos adicionais			
Estatística de acesso por documento (consulta e	Não.	Não.	Não.

	OAI Arc	JOAI Harvester	PKP Harvester
download)			
Tradução para o Português do Brasil	Não, mas é possível realizar a tradução da interface para o Português do Brasil alterando o arquivo **ArcResourceBundle**. Possui tradução para os idiomas: Inglês, Espanhol e Francês.	Não, mas é possível realizar a tradução da interface para o Português do Brasil alterando o código fonte das interfaces de busca.	Não, mas é possível realizar a tradução da interface para o Português do Brasil alterando o arquivo **locale.xml**. Possui tradução para o Inglês.
Bugs			
	Não identificado.	Não identificado.	Coleta parcial de registros em determinados repositórios. Erro: The metadata index could not be update. The following error(s) occurred.
Teste de performance de coleta de dados (exemplos)			
Animal Physiology and Livestock Systems Archive **(340 registros)**	18/06/2009 – 340 registros coletados em 39s – Média: **8,71** registros coletados por segundo.	18/06/2009 – 340 registros coletados em 46s – Média: **7,39** registros coletados por segundo.	17/06/2009 – 340 registros coletados em 223s – Média:**1,52** registros coletados por segundo.
IBSS Repository **(787 registros)**	18/06/2009 – 787 registros coletados em 120s – Média: **6,55** registros coletados por segundo.	18/06/2009 – 787 registros coletados em 59s – Média: **13,33** registros coletados por segundo.	17/06/2009 – 1 registro coletado em 7s, exibindo erro “Invalid Character”
Nature Precedings **(1.177 registros)**	18/06/2009 – 1.210 registros coletados em 15s – Média: **80,66** registros coletados por segundo.	18/06/2009 – 1.209 registros coletados em 17s – Média: **71,11** registros coletados por segundo.	16/06/2009 – 1.204 registros coletados em 599s – Média: **2,01** registros coletados por segundo.
NWISRL Publications **(2.630 registros)**	18/06/2009 – 2.640 registros coletados em 600s – Média: **4,4** registros coletados por segundo.	18/06/2009 – 2.640 registros coletados em 581s – Média: **4,54** registros coletados por segundo.	17/06/2009 – 1.320 registros coletados em 1.297 s – Média: **1,02** registros coletados por segundo.
Organic Eprints **(9.170 registros)**	18/06/2009 – 9.221 registros coletados em 1.461s – Média: **6,31** registros coletados por segundo.	18/06/2009 – 9.221 registros coletados em 1.806s – Média: **5,10** registros coletados por segundo.	17/06/2009 – 9.208 registros coletados em 6.796s – Média:**1,35** registros coletados por segundo.
SERPENT Image & Video Database **(1.179 registros)**	18/06/2009 – 1.179 registros coletados em 86s – Média: **13,70** registros coletados por segundo.	18/06/2009 – 1.179 registros coletados em 113s – Média: **10,43** registros coletados por segundo.	16/06/2009 - 1.179 registros coletados em 294s – Média: **4,01** registros coletados por segundo.
UNITAU - Departamento de Ciências Agrárias **(103 registros)**	18/06/2009 – 103 registros coletados em 15s – Média: **6,86** registros coletados por segundo.	18/06/2009 – 103 registros coletados em 20s – Média: **5,15** registros coletados por segundo.	18/06/2009 – 103 registros coletados em 134s – Média: **0,77** registros coletados por segundo.

Tabela 2: Resultado dos testes de performance de coleta de dados

4.5 Avaliação dos *softwares* livres de coleta de dados

As três soluções livres analisadas e testadas apresentaram recursos sofisticados de coleta de dados, como: coleta de dados por interface *web* ou linha de comando, agendamento de coleta de dados, armazenamento dos dados coletados em banco de dados relacionais ou arquivos etc. No entanto, recursos mais sofisticados de busca presentes na interface, como: busca por campo, agrupamento do resultado da busca, ordenação do resultado da busca, filtro de busca, cesta de itens, tradução para o português do Brasil precisavam ser melhorados ou implementados.

Devido a maior experiência da Embrapa Informática em atuar com sistemas desenvolvidos em Java e os ótimos resultados obtidos com o processo de coleta de dados, optou-se pelo uso do *software* livre JOAI para coleta de dados e pelo desenvolvimento da interface de busca do sistema.

5 IMPLEMENTAÇÃO DO PROVEDOR DE SERVIÇOS SABIIA

Com a definição do JOAI como solução tecnológica de coleta de dados no formato OAI-PMH, os esforços seguintes para construção do provedor de serviços SABIIA (Sistema Aberto e Integrado de Informação em Agricultura) concentraram-se na escolha do mecanismo de indexação e busca textual, na definição e implementação do projeto gráfico de interface.

5.1 Escolha do mecanismo de indexação e busca textual

O mecanismo de indexação e busca textual Lucene tem sido largamente utilizado em projetos *open source* em todo mundo. Entretanto, para o projeto Sabiia, objetivou-se encontrar soluções livres capazes de estender e melhorar os recursos originais da ferramenta Lucene, como: integração com bases de dados relacionais, recursos mais sofisticados de indexação e busca, e maior integração com a arquitetura Java EE (*Java Platform, Enterprise Edition*) [20]. O resultado dessa investigação levou a escolha do *software* livre Solr.

O Solr é um mecanismo de indexação e busca textual de código aberto baseado na biblioteca Lucene oferecendo recursos sofisticados de indexação e busca textual, como: busca com operadores booleanos, busca por campo, *highlighting* sobre o resultado da busca, paginação do resultado da busca, *facets* sobre o resultado da busca, *caching* de busca, integração com banco de dados relacionais, replicação de bases de dados, interface de administração *web* etc.

O Solr é disponibilizado na foma de uma aplicação *web*, o sistema cliente pode ser construído com as linguagens de programação: Ruby, PHP, Java, Python, .NET, C# e Perl.

5.2 Arquitetura de *software* do provedor de serviços Sabiia

Além dos *softwares* livres JOAI e Solr, o sistema utiliza a arquitetura Java EE (*Java Server Pages* e *Servlets*) para implementação da interface *web* de busca e o servidor de aplicações Apache Tomcat [21], conforme ilustra a figura 2:

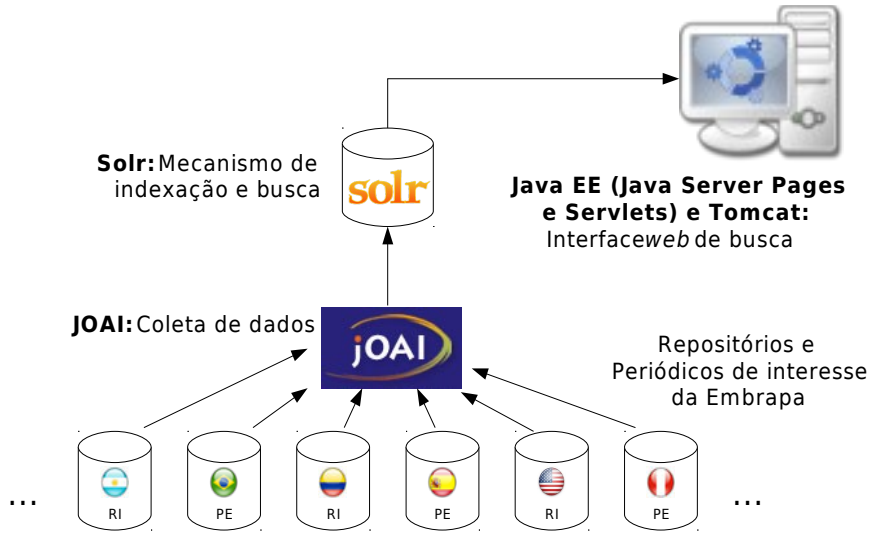

Figura 2: Arquitetura do provedor de serviços Sabiia

5.3 Abertura ao público externo

Com aproximadamente 101.000 documentos coletados em 30 provedores de dados, o sistema foi disponibilizado com acesso restrito à Embrapa para avaliação. Atualmente, o provedor de serviços Sabiia encontra-se em fase de construção e implementação do projeto gráfico de interface, bem como, implementação de novos repositórios de interesse da Embrapa identificados pelo projeto Acesso Aberto à Informação Científica na Embrapa. Assim, como o repositório institucional Alice, espera-se que o provedor de serviços Sabiia seja disponibilizado para o público externo no segundo semestre de 2010.

6 RESULTADOS ALCANÇADOS

O uso de *software* livre tem-se consolidado e expandido em diversas áreas do conhecimento, como: geotecnologias, matemática computacional, banco de dados, engenharia de *software* etc. O presente trabalho relatou a experiência da Embrapa Informática no uso de tecnologias livres para a área de gestão da informação técnico-científica, mais especificamente para o movimento Acesso Aberto. A tecnologia livre DSpace mostrou-se amplamente adequada para construção dos repositórios Infoteca e Alice, ambos destinados à disseminação da informação produzida pela área de P&D da Embrapa. Enquanto a combinação, das ferramentas livres JOAI e Solr, favoreceram a criação do provedor de serviços Sabiia, caracterizado como elemento responsável pela integração de todos os fornecedores de dados (repositórios institucionais, periódicos científicos, bibliotecas digitais, e outros), tanto internos quanto externos de interesse da Embrapa.

Por fim, o uso de *software* livre tem-se consolidado nas atividades de pesquisa, desenvolvimento e inovação da Embrapa frente aos novos cenários e desafios propostos.

7 REFERÊNCIAS

[1] OAI. **Open Archives Initiative.** Disponível em: <http://www.openarchives.org> Acesso em: 11 de maio de 2010.

[2] THE DSPACE FOUNDATION. **DSpace.** Disponível em: <http://www.dspace.org>. Acesso em: 10 de maio de 2010.

[3] DIGITAL LIBRARY FOR EARTH SYSTEM EDUCATION. **jOAI.** Disponível em: <http://www.dlese.org/dds/services/joai_software.jsp>. Acesso em: 11 de maio de 2010.

[4] THE APACHE SOFTWARE FOUNDATION. **Apache Solr.** Disponível em: <http://lucene.apache.org/solr>. Acesso em: 11 de maio de 2010.

[5] ORACLE SUN DEVELOPER NETWORK. **Java Technology.** Disponível em: <http://java.sun.com>. Acesso em: 10 de maio de 2010.

[6] THE POSTGRESQL GLOBAL DEVELOPMENT GROUP. **PostgreSQL.** Disponível em: <http://www.postgresql.org>. Acesso em: 10 de maio de 2010.

[7] THE APACHE SOFTWARE FOUNDATION. **Lucene.** Disponível em: <http://lucene.apache.org>. Acesso em: 10 de maio de 2010.

[8] DUBLIN CORE METADATA INITIATIVE. **Dublin Core.** Disponível em: <http://www.dublincore.org/documents/usageguide>. Acesso em: 11 de maio de 2010.

[9] MASSACHUSETTS INSTITUTE OF TECHNOLOGY. **MIT.** Disponível em: <http://http://web.mit.edu/>. Acesso em: 11 de maio de 2010.

[10] HEWLETT PACKARD DEVELOPMENT COMPANY, **HP Labs.** Disponível em: <http://http://www.hpl.hp.com>. Acesso em: 11 de maio de 2010.

[11] BERKELEY STANDARD DISTRIBUTION LICENSE. **BSD open source license.** Disponível em: <http://www.opensource.org/licenses/bsd-license.php>. Acesso em: 10 de maio de 2010.

[12] THE CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS. **Reference Model - Reference Model for an Open Archival Information System - OAIS.** Disponível em: <http://public.ccsds.org/publications/RefModel.aspx>. Acesso em: 11 de maio de 2010.

[13] THE DSPACE FOUNDATION. **DSpace 1.5.1 Manual.** Disponível em: <http://www.dspace.org/1_5_1Documentation>. Acesso em: 11 de maio de 2010.

[14] INSTITUTIONAL REPOSITORY OF UNIVERSIDADE OF MINHO. **Statistics AddOn**

[15] AGROLIVRE - REDE DE SOFTWARE LIVRE PARA AGROPECUÁRIA. HiperNavegador - Navegador Hiperbólico. Disponível em: <http://http://repositorio.agrolivre.gov.br/projects/hipernavegador>. Acesso em: 11 de maio de 2010.

[16] EMBRAPA INFORMÁTICA AGROPECUÁRIA. **Ainfo.** Disponível em: <http://www.ainfo.cnptia.embrapa.br>. Acesso em: 11 de maio de 2010.

[17] LEITE, FERNANDO CÉSAR LIMA and BERTIN, PATRÍCIA ROCHA BELLO and VACARI, ISAQUE and SIMÃO, VICTOR PAULO MARQUES and VISOLI, MARCOS CÉZAR. **Implementação de estratégias de acesso aberto em uma instituição de pesquisa de grande porte na área de agricultura: a experiência da Embrapa.**, 2009 . In XV Reunión de la Asociación Interamericana de Bibliotecarios, Documentalistas y Especialistas en Información Agrícola, Lima, Peru.

[18] OPEN ARCHIVES INITIATIVE. **The Open Archives Initiative Protocol for Metadata Harvesting.** 2008. Disponível em: <http://www.openarchives.org/OAI/openarchivesprotocol.html>. Acesso em: 11 de maio de 2010.

[19] SOURCEFORGE.NET. SourceForge. Disponível em: <http://sourceforge.net>. Acesso em: 14 de maio de 2010.

[20] ORACLE SUN DEVELOPER NETWORK. **Java EE.** Disponível em: <http://java.sun.com/javaee>. Acesso em: 11 de maio de 2010.

[21] THE APACHE SOFTWARE FOUNDATION. **Apache Tomcat.** Disponível em: <http://tomcat.apache.org>. Acesso em: 10 de maio de 2010.