ACESSO ABERTO A INFORMAÇÃO CIENTÍFICA AGROPECUÁRIA NA INTERNET: CASO DO SISTEMA ABERTO E INTEGRADO DE INFORMAÇÃO EM AGRICULTURA (SABIIA)

ISAQUE VACARI[1]
MARCOS CEZAR VISOLI[2]
LUÍS EDUARDO GONZALES[3]

RESUMO: Este trabalho apresenta o Sistema Aberto e Integrado de Informação em Agricultura (Sabiia) como mecanismo de busca automatizado que coleta e centraliza metadados de provedores de dados OAI-PMH científicos de acesso aberto, previamente selecionados. O sistema reúne informações sobre agricultura e áreas afins, possibilitando o acesso ao texto integral (livros, capítulos de livros, artigos em periódicos, folhetos, teses, anais e *proceedings* de eventos, entre outros) de milhares de publicações científicas disponíveis em diversas instituições nacionais e internacionais.

PALAVRAS-CHAVE: Acesso Aberto, Provedor de Serviços OAI-PMH, Recuperação de Informação, Software Livre, Pesquisa Agropecuária.

OPEN ACCESS TO INFORMATION AGRICULTURAL SCIENCE ON THE INTERNET: CASE OPEN INTEGRATED INFORMATION SYSTEM IN AGRICULTURE (SABIIA)

ABSTRACT: This work presents The Open Integrated Information System in Agriculture (Sabiia) is an automated search engine that collects and centralizes metadata from previously selected open access scientific data providers OAI-PMH. This interface gathers information on agriculture and related fields, providing access to the full text (books, book chapters, journal articles, pamphlets, theses, records and proceedings of events, among others) of thousands of scientific papers in various national and international institutions.

KEYWORDS: Open Access, Service Provider OAI-PMH, Information Retrieval, Open Source, Agricultural Research.

1. INTRODUÇÃO

O acesso aberto ao conhecimento científico tem se constituído em um paradigma emergente da comunicação científica em todo o mundo. Como uma reação ao modelo tradicional de comunicação da ciência - cuja lógica favorece que editores comerciais atribuam preços exorbitantes e imponham barreiras de permissão e acesso a publicações de resultados de pesquisas que são amplamente financiadas com recursos públicos - o acesso aberto tem sido visto como uma nova filosofia de comunicação, como observam (COSTA; MOREIRA, 2003). JOHNSON (2002) afirma que o modelo tradicional de comunicação científica limita, mais do que expande, a disponibilidade e legibilidade (*readership*) da maior parte da pesquisa científica (ao tempo que obscurece suas origens institucionais). O fato é que a insatisfação da comunidade científica aliada ao desenvolvimento de tecnologias resultam no que assinala (PROSSER, 2003) ao sugerir que a convergência entre os preços crescentes impostos por

[1] Analista, Tecnólogo em Processamento de Dados, FATEC, E-mail: isaque@cnptia.embrapa.br
[2] Pesquisador, Bacharel em Ciência da Computação, UFSC, E-mail: visoli@cnptia.embrapa.br
[3] Analista, Bacharel em Análise de Sistemas, PUC/Campinas, E-mail: eduardo@cnptia.embrapa.br

editores, a limitação de recursos orçamentários das bibliotecas acadêmicas e as tecnologias aplicadas à comunicação estão promovendo um ambiente apropriado para uma transformação significativa em como se compartilham informação no seio das comunidades científicas. Ao cenário desenhado por (PROSSER, 2003), (LEITE, 2006) acrescenta que, além de desencadear em um modelo de comunicação científica, o acesso aberto, provê de modo consistente, uma infraestrutura social, política, cultural e tecnológica apropriada e necessária à gestão da informação e do conhecimento científico.

Portanto, na medida em que visa remover barreiras que restringem o acesso à informação científica, o acesso aberto maximiza o acesso à pesquisa propriamente dita, como sugerem resultados de estudos empíricos realizados (BRODY et al., 2007), (BRODY; HARNARD, 2004) e (HAJJEM et al., 2005). Desse modo, concorda-se com (BRODY; HARNARD, 2004) ao afirmarem que o acesso aberto maximiza e acelera o impacto das pesquisas e, por esta razão, acelera também sua produtividade, progresso e recompensas. (LAWRENCE, 2001) demonstra a afirmação de (BRODY; HARNARD, 2004) ao apresentar resultados de sua pesquisa sobre o aumento do impacto de pesquisa em ambiente de acesso aberto. Seus resultados evidenciaram que houve um aumento médio de 336% nas citações a artigos disponíveis em ambiente de acesso aberto, em relação a artigos restritos, publicados na mesma fonte, e conclui assinalando que "para maximizar o impacto, minimizar a redundância e acelerar o progresso científico, autores e editores deveriam tornar a pesquisa fácil de ser acessada". Essa tarefa tem sido, em última análise, objeto dos avanços tecnológicos promovidos pela *Open Archives Initiative* (OAI - http://www.openarchives.org).

A *Open Archives Initiative* – Iniciativa de Arquivos Abertos, desde seu princípio, exerce importante papel, e sem precedentes, ao ter como missão o desenvolvimento e a promoção de padrões de interoperabilidade entre arquivos digitais, alcançados, sobretudo, com o estabelecimento do protocolo OAI-PMH (*Protocol for Metadata Harvesting*) com o propósito de promover a disseminação eficiente e ampla da informação científica. Tais padrões têm sido amplamente adotados pela comunidade científica mundial. Os padrões de interoperabilidade preconizados foram inicialmente definidos a partir do estabelecimento de recomendações e mecanismos para facilitar a cooperação entre arquivos digitais e o oferecimento de serviços de informação com valor agregado.

Na perspectiva da Iniciativa de Aquivos Abertos existem dois papéis principais: os provedores de dados (*Data Providers*) e os provedores de serviços (*Service Providers*). Os provedores de dados são "arquivos digitais" (repositórios institucionais ou temáticos e periódicos científicos) mantidos por instituições científicas. São eles que oferecem os mecanismos de submissão, armazenamento e preservação de documentos e também expõem os metadados descritivos dos objetos digitais por meio de protocolos abertos. Provedores de serviços, por sua vez, são os responsáveis pela coleta de metadados expostos por provedores de dados distribuídos geograficamente e pela criação de serviços de informação de valor agregado.

2. OBJETIVO

Nesse contexto de organização, tratamento e disseminação da informação na Internet utilizando software livre e padrões abertos, 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 estratégico o desenvolvimento do provedor de serviços Sabiia (Sistema Aberto e Integrado de Informação em Agricultura), com o propósito de oferecer acesso a toda produção intelectual de fontes de informação (periódicos nacionais, periódicos estrangeiros, repositórios institucionais e temáticos) abertas em agricultura.

No que tange aos objetivos específicos do sistema Sabiia destacam-se: i) identificação, seleção e coleta de metadados de provedores de dados da área de pesquisa agropecuária e áreas afins disponíveis em ambiente de acesso aberto no Brasil e no mundo; ii) disponibilização dos metadados coletados em um sistema de recuperação de informação (mecanimo de busca); iii) exposição dos metadados coletados para outros provedores de serviços ou sistema de informação que porventura tenham interesse em coletá-los; iv) monitoramento permanentemente a fim de identificar o surgimento de novos provedores de dados; v) uso prioritário de software livre e padrões abertos para desenvolvimento do sistema.

3. MATERIAL E MÉTODOS

O sistema Sabiia é um mecanismo de busca automatizado, baseado no protocolo OAI-PMH, que coleta e centraliza metadados de provedores de dados científicos de acesso aberto, previamente selecionados. O sistema reúne informações sobre agricultura e áreas afins, possibilitando o acesso ao texto integral (como livros, capítulos de livros, artigos em periódicos, folhetos, teses, anais e *proceedings* de eventos, entre outros.) de milhares de publicações científicas disponíveis em diversas instituições nacionais e internacionais.

O protocolo OAI-PMH provê interoperabilidade não imediata (ou seja, não é um protocolo para consultas *on-line*) entre provedores de dados ou qualquer outro serviço na rede que queira tornar visível os metadados de documentos nele armazenados com vistas à facilitar a recuperação e a disseminação eficiente de conteúdos produzidos pelas comunidades científicas. Um exame preliminar identificou um total de 261 provedores de dados baseados no protocolo OAI-PMH em áreas de interesse da pesquisa agropecuária. 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.

Com a definição do OAI-PMH como protocolo para intercâmbio de dados e a identificação preliminar dos provedores de dados agropecuários, o desafiou concentrou-se em desenvolver o sistema de coleta de metadados e busca. Devido a maior experiência da Embrapa Informática Agropecuária em ferramentas de software livre, priorizou-se o uso de tecnologias livres para construção do sistema. Foram realizados estudos sobre software livre para coleta de metadados baseados no protocolo OAI-PMH discutidos em (BERTIN, 2010). Fundamentalmente, o estudo apresenta o software jOAI (Digital Library for Earth System Education - http://www.dlese.org/dds/services/joai_software.jsp) como solução tecnológica para coleta de metadados OAI-PMH. Com a definição do software jOAI, os esforços seguintes para construção do provedor de serviços concentraram-se na escolha do mecanismo de indexação e busca textual e implementação do projeto gráfico de interface.

O mecanismo de indexação e busca textual Lucene (Apache Lucene – http://lucene.apache.org) tem sido largamente utilizado em projetos *open source*. Entretanto, procurou-se selecionar 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 – http://java.sun.com/javaee). O resultado dessa investigação levou a escolha do software livre Solr (Apache Solr - http://lucene.apache.org/solr) integrado ao Sistema Gerenciador de Banco de Dados PostgreSQL (PostgreSQL Global Development Group – http://www.postgresql.org).

A interface gráfica de busca do sistema Sabiia foi implementada com tecnologias aderentes a arquitetura Java EE versão 6 e disponibilizada no servidor Web Tomcat (Apache Tomcat – http://tomcat.apache.org).

Em resumo, o esquema funcional do sistema Sabiia é composto das seguintes etapas:

1. Identificação, seleção e coleta de metadados de provedores de dados da área de pesquisa agropecuária e afins disponíveis em ambiente de acesso aberto no Brasil e no mundo. Os metadados são coletados com o software livre jOAI e armazenados em arquivos XML (Extensible Markup Language) no formato OAI-PMH;
2. Em seguida, executa-se um software específico, desenvolvido no âmbito do projeto, que processa os arquivos XML no formato OAI-PMH e efetua a gravação dos dados coletados no Sistema Gerenciador de Banco de dados Relacional PostgreSQL;
3. A próxima etapa consiste em converter os dados relacionais em dados mais adequados para recuperação de informação. Para realizar essa atividade foi criado mais um software específico capaz de converter os dados armazenados no banco de dados PostgreSQL para o formato do software livre de indexação e busca textual Solr;
4. Por fim, foi desenvolvido uma interface de consulta que disponibiliza os metadados coletados em um sistema de recuperação de informação (mecanismo de busca). A interface de consulta criada interage diretamente com os dados indexados armazenados no software livre Solr;
5. Além de oferecer uma interface sofistica de busca, o sistema Sabiia expõe os metadados coletados para outros provedores de serviços ou sistema de informação que porventura tenham interesse em coletá-los. Os metadados são expostos por meio do software livre OAICat (OCLC - The world's libraries. Connected - http://www.oclc.org/research/activities/oaicat/default.htm) na forma de serviço OAI-PMH, para disponibilizar os metadados no software OAICat foi necessário criar um novo software específico capaz de converter os dados armazenados no banco de dados relacional para arquivos XML no formato OAI-PMH.

É importante ressaltar que, todo o processo, mesmo combinando vários softwares livres e específicos, é executado de forma automática, desde a coleta de dados até a disponibilização dos metadados para consulta ou exposição dos metadados para coleta por outros sistemas de informação interoperável.

4. RESULTADOS E DISCUSSÃO

O conjunto de metodologias e tecnologias disponibilizados pela Open Archives Initiative (OAI), permitem desenvolver novas alternativas de sistemas de informação. Uma série de novos serviços baseados em reuso de metadados podem ser concebidos, incluindo redes cooperativas, sistemas de informação regionais entre outros. Aos poucos as comunidades envolvidas com publicações digitais, estão conhecendo, avaliando e utilizando o potencial de interoperabilidade do protocolo OAI-PMH.

(MARCONDES, 2002) previa que o intercâmbio de dados entre provedores de dados e provedores de serviços previsto pelo protocolo OAI-PMH possibilitariam a criação de novos serviços de informação com valor agregado. O Sabiia é um caso prático dessa visão, onde metadados expostos em escala planetária por diversos provedores de dados OAI-PMH do setor agropecuário e área afins são centralizados, unificados e disponibilizados para consulta em uma única interface, constituindo um novo serviço de informação.
A adoção de padrões abertos e interoperáveis, associada ao acesso livre e irrestrito aos dados e informações configuram uma oportunidade altamente significativa para construção de aplicações digitais e democratização no acesso aos resultados de pesquisas e do conhecimento em geral. Lançado oficialmente no aniversário da Embrapa em Abril de 2011, o sistema Sabiia está disponível no endereço: http://www.embrapa.br/sabiia com aproximadamente 270.399 documentos indexados em 128 provedores de dados.

5. REFERÊNCIAS

BERTIN, P. R. B.; VACARI, I.; SIMÃO, V. P. M.; VISOLI, M. C.; LEITE, F. C. L. **An Open Access Approach to scientific information management at the brazilian agricultural research corporation. Scholarly and Research Communication**, v. 1, 1, p. 1-12, 2010. Disponível em: <http://ainfo.cnptia.embrapa.br/digital/bitstream/item/17246/1/1-130-2-PB.Open.pdf>. Acessado em: 13 mai. 2011.

BRODY, T., CARR, L., GINGRAS, Y., HAJJEM, C., HARNAD, S. and SWAN, A. Incentivizing the Open Access Research Web:
Publication-Archiving, Data-Archiving and Scientometrics. **CTWatch Quarterly**, v.3 n.3, 2007.

BRODY; T.; HARNAD, S. **The research impact cycle**. Disponível em <http://opcit.eprints.org/feb19oa/harnad-cycle.ppt>. Acesso em: set. 2004.

COSTA, S. M. S.; MOREIRA, A. C. S. The diversity of trends, experiences and approaches in electronic publishing: evidences of a paradigm shift on communication. In: COSTA, S. M. S.; CARVALHO, J.A.C.; BAPTISTA, A.A.; MOREIRA, A.C.S. **From information to knowledge: proceedings of the 7th ICCC/IFIP International Conference on Electronic Publishing.** Guimarães: Universidade do Minho, 2003, p. 5-9.

HAJJEM, C., HARNAD, S. and GINGRAS, Y. (2005) Ten-Year Cross-Disciplinary Comparison of the Growth of Open Access and How it Increases Research Citation Impact. **IEEE Data Engineering Bulletin** 28(4) pp. 39-47. Disponível em: <http://eprints.ecs.soton.ac.uk/12906>. Acesso em: 14 mai. 2011.

JOHNSON, R. K. **Partnering with faculty to enhance scholarly communication. D-Lib Magazine**, v. 8, n. 11, nov. 2002. Disponível em <http://www.dlib.org/dlib/november02/johnson/11johnson.html>. Acesso em: 13 mai. 2005.

LAWRENCE, S. Institutional repositories: enhancing teaching, learning and research. **EDUCAUSE Envolving Technologies Committee.** Disponível em: <http://sitemaker.umich.edu/dams/files/etcom-2003-repositories.pdf>. Acesso em: jul. 2004.

LEITE, F. C. L.; MORENO, F. P.; MARDERO A., MIGUEL A. **Acesso livre a publicações e repositórios digitais em Ciência da Informação no Brasil.** Perspectivas em Ciência da Informação, v. 11, n. 1, p. 255-269, 2006. Disponível em: <http://repositorio.bce.unb.br/handle/10482/623>. Acesso em: 13 mai. 2011.

MARCONDES, C. H.; SAYÃO, L. F. **Documentos digitais e novas formas de cooperação entre sistemas de informação em C&T.** Ciência da Informação, set 2002, vol.31, no.3, p.42-54. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652002000300005&lng=en&nrm=iso>. Acesso em: 13 mai. 2011.

PROSSER, D. **Information revolution: can institutional repositories and open access transform scholarly communications? The ELS Gazette**, v. 15, jul. 2003.