

Proposta de infraestrutura para a gestão de conhecimento científico sensível ao contexto geográfico

Alaor Bianco Rodrigues¹, Walter Abrahão dos Santos², Sidnei João Siqueira Sant'Anna³, Corina da Costa Freitas³

¹Programa de Mestrado em Computação Aplicada – CAP
Instituto Nacional de Pesquisas Espaciais – INPE

²Laboratório de Matemática e Computação Aplicada, LAC
Instituto Nacional de Pesquisas Espaciais – INPE

³Divisão de Processamento de Imagens, DPI
Instituto Nacional de Pesquisas Espaciais – INPE

alaorbianco@yahoo.com.br, walter.abrahao@lac.inpe.br, {sidnei, corina}@dpi.inpe.br

Abstract. *This work discuss how the area of e-Science has been exploited to develop an infrastructure capable of helping the management of scientific knowledge produced in the Image Processing Division at INPE, focusing on but not limited to, geospatial artifacts, applying a case study using as inputs several studies conducted by researchers at INPE in the area of the Tapajos National Forest.*

Resumo. *Neste trabalho é abordado como a área de e-Science foi explorada para o desenvolvimento de uma infraestrutura capaz de auxiliar a gestão do conhecimento científico produzido na Divisão de Processamento de Imagens do INPE, com foco em, mas não limitado a, artefatos sensíveis ao contexto geográfico, aplicado um estudo de caso usando como insumos diversos trabalhos realizadas por pesquisadores do INPE na região da Floresta Nacional do Tapajós.*

Palavras-chave: *Gestão do Conhecimento, Gerenciamento de Conteúdo, CMS, CMIS, OGC*

1. Introdução

No início da criação de um novo conhecimento, o esforço de um pesquisador parte daquilo que foi construído anteriormente por outros pesquisadores, ou seja, recorre à literatura de sua especialidade, e, ao fim, divulga os resultados de sua pesquisa por meio dos veículos de comunicação apropriados à sua área de conhecimento.

Percebe-se assim a importância da comunicação, informar ao mundo científico seus feitos, resultados e etc. Meadows (1999) diz que a comunicação reside no coração da ciência, sendo tão vital quanto a própria pesquisa. No entanto apenas uma fração do que é produzido durante uma pesquisa é publicado, ou seja, é formalmente comunicado a comunidade. Braga (1985) ressalta que a comunicação formal é responsável por apenas 20% de todas as comunicações no processo de geração do conhecimento. Sendo que as demais são constituídas de processos informais, e uma grande parcela desse

conhecimento encontra-se em um formato que poderia ser explícito, como anotações, planilha de resultados, registros de experimentos, resultados parciais, etc.

Os recursos computacionais e o ambiente *Web* muito contribuem para um cenário de compartilhamento e comunicação. Os recursos computacionais facilitam o trabalho em rede, podendo manter os conhecimentos descentralizados junto aos locais em que são mais gerados e/ou utilizados (Davenport et al, 1998) e melhorando o grau de interatividade do usuário com os registros de conhecimentos (Davenport et al, 1998). A computação é efetivamente útil para a gestão do conhecimento, se for empregada utilizando-se uma sistemática interferência (interatividade) humana (Davenport, 2001).

2. Gestão do Conhecimento

Gestão do conhecimento é um tema relativamente novo, multidisciplinar e muito explorado em diversas pesquisas, mas quase sempre seu foco são as organizações empresariais. No entanto há iniciativas da aplicação destes conceitos da gestão de conhecimento sob o âmbito do conhecimento científico como dissertado, principalmente, em (Leite, 2006).

Nonaka e Takeuchi (1997) forneceram uma grande contribuição para o assunto, sendo suas obras as maiores referências atualmente. Estes autores realizaram uma importante distinção entre os tipos de conhecimento humano, classificando-os em conhecimento tácito e conhecimento explícito. Sendo os conhecimentos explícitos aqueles estruturados capazes de serem verbalizados, facilmente transmitido, sistematizado e comunicado. Já os conhecimentos tácitos são aqueles inerentes às pessoas, isto é, o conhecimento pessoal incorporado à experiência individual, crenças e valores. É difícil de ser articulado na linguagem formal e transmitido por se tratar da parcela não estruturada do conhecimento.

Nonaka e Takeuchi (1997), ainda, consideram que um trabalho efetivo com o conhecimento somente é possível em um ambiente em que possa ocorrer a contínua conversão entre esses dois formatos. Segundo estes autores são 4 os processos de conversão entre os dois tipos de conhecimento: socialização, externalização, combinação e internalização.

3. Gerenciamento de Conteúdo

O conceito de *Enterprise Content Management* (ECM) compreende "as estratégias, ferramentas, processos e habilidades que uma organização precisa para gerenciar seus ativos de informação durante o seu ciclo de vida", incluindo todos os ativos digitais, como documentos, dados, relatórios e páginas da *web* (Smith e McKeen 2003). O Meta Group o define como a tecnologia que fornece os meios para criar, capturar, gerenciar, armazenar, publicar, distribuir, pesquisar, personalizar, apresentar e imprimir qualquer conteúdo digital (imagens, texto, relatórios, vídeo, áudio, dados transacionais, catálogo, código). Estes sistemas se concentram na captura, armazenamento, recuperação e disseminação de arquivos digitais para uso corporativo. (Meta Group, em Weiseth et al. 2002, p. 20).

Enterprise Content Management System (ECMS), ou simplesmente *Content Management System* (CMS) é a expressão utilizada para descrever ferramentas que promovem meios de gerenciamento, publicação e manutenção destes ativos de informação. Esta categoria de sistemas ainda incluem funcionalidades de fórum, listas de discussões, *workflows*, controle de acesso, associações, classificação e categorização, o que cria um ambiente propício para gestão do conhecimento uma vez que facilitam a existência, manutenção e crescimento dos processos de transformação citados em Nonaka e Takeuchi (1997). Assumindo não apenas o papel de uma infraestrutura para tal, mas também criando condições ambientais e motivacionais que faça com que as pessoas vivam e reforcem estes ciclos de transformação, por:

1) Estimular o processo de **socialização** do conhecimento uma vez que a diversidade de formatos em que as informações podem existir criam condições favoráveis à assimilação do conhecimento. O resultado é uma transferência da informação e do conhecimento mais efetiva, pois muito do conhecimento científico gerado por um pesquisador não é possível de ser comunicado por meios formais e transforma parte do conhecimento que antes era puramente tácito em conhecimento explícito. Ainda sob a ótica da socialização, é estimulada a interação informal entre pesquisadores interessados em um mesmo assunto, possibilitando discussões e compartilhamento de ideias e esboços para coleta de sugestões e comentários enriquecendo as pesquisas e intensificando a troca de experiências.

2) Ser instrumento de **externalização** do conhecimento tácito que, segundo Nonaka e Takeuchi (1997), trata-se do processo de criação do conhecimento perfeito, ao fornecerem a possibilidade de armazenar múltiplos formatos desse conhecimento. As publicações científicas são formais e desta forma formatam o conhecimento e de certa forma limita seus horizontes. Uma infraestrutura capaz de armazenar os conhecimentos informais aproxima os demais pesquisadores aos elementos que compõe o estado do conhecimento de seu autor. Neste cenário, parte do conhecimento tácito é transformado em uma estrutura comunicável permitindo que esta seja processada, armazenada e recuperada.

3) Permitir a transformação de um determinado conjunto de conhecimento explícito, por meio de agrupamento, acréscimo, categorização e classificação, criando um novo conjunto de conhecimento ou criando e/ou acrescentando um novo conhecimento, constituindo, assim, o processo de **combinação**.

4) Facilitar o processo de **internalização** por criar condições favoráveis para que o conhecimento explícito armazenado seja convertido em conhecimento tácito do indivíduo.

4. Revisão de Literatura

Alguns autores veem desenvolvendo trabalhos sobre o tema de gestão de conhecimento científico e estudando ferramentas e alternativas para auxiliar e facilitar os processos envolvidos em tais atividades. Leite e Costa (2006) discutem a adequação e aplicabilidade de repositórios institucionais como uma ferramenta para tal, abordando as peculiaridades do conhecimento científico, bem como o ambiente no qual se dão os processos de sua criação, compartilhamento e uso.

Contexto semelhante foi explorado por Cañete et al. (2010) ao desenvolver um sistema de informações de biodiversidade baseado em banco de dados, API do Google Maps e o sistema R, que permite catalogar dados a respeito de espécimes coletadas, analisá-los e apresentá-los num mapa.

Este trabalho se diferencia por adotar plataformas abertas e consolidadas no mercado, reduzindo customizações e sendo muito aderente a padrões existentes. Visa não apenas ser um repositório de dados, mas uma plataforma que permita que os processos de transformação do conhecimento ocorram e sejam incentivados. Ainda, por manusear dados matriciais (raster) e prover um barramento de serviços sobre estes.

5. Metodologia

Na fase de levantamento, foram realizadas entrevistas com alguns usuários da Divisão de Processamento de Imagens do INPE (Pesquisadores) que representavam os demais usuários. Foram elencadas suas necessidades e criando uma lista de requisitos, conforme pode ser observado na Tabela 1.

ID	Requisito
RQ001	A solução deve contemplar um sistema de fórum.
RQ002	A solução deve contemplar um sistema de listas de discussões.
RQ003	A solução deve ser de acesso público, mas com recursos de restrições de acesso a determinados conteúdos caso pertinente.
RQ004	A solução deve contemplar mecanismo de armazenamento de arquivos multimídias.
RQ005	A solução deve ser suficientemente configurável de modo que possam ser definidos quais metadados importantes para cada tipo de conteúdo.
RQ006	A solução deve possuir mecanismo de busca por todo o conteúdo textual.
RQ007	A solução deve contemplar mecanismo de classificação de conteúdo por rótulos.
RQ008	A solução deve prover conteúdo geográfico segundo padrões abertos OGC (WMS, WFS, WCF, WPS).
RQ009	A solução deve usar produtos de software livre, preferencialmente de código fonte aberto e na linguagem Java.
RQ010	A solução deve contemplar a manipulação, armazenamento e recuperação de imagens vetoriais (raster).
RQ011	A solução deve contemplar o agrupamento e o relacionamento de conteúdos.
RQ012	A solução deve contemplar o referenciamento geográfico dos conteúdos.
RQ013	A solução deve contemplar a plotagens dos elementos georreferenciados no mapa.
RQ014	A solução deve ser capaz de consumir serviços web de geolocalização, GeoRSS, WPS e BaseMaps.

Tabela 1: Requisitos da solução

Analisando os requisitos, foi possível perceber que grande parte dos requisitos eram elucidados por uma ferramenta de CMS, caso dos requisitos RQ001, RQ002, RQ003, RQ004, RQ005, RQ006, RQ007 e RQ011. Mas ainda assim, havia alguns requisitos que não eram contemplados por esta. Neste caso os requisitos que fogem ao escopo de atuação das ferramentas de CMS são, em essência, relacionados ao contexto espacial. Sendo assim, estes requisitos foram tratados fora do CMS, incluindo na arquitetura da solução um elemento de gerenciamento de conteúdo geográfico.

Esta abordagem implica em integração entre o CMS e o gerenciador de conteúdo geográfico. Esta integração foi realizada utilizando o padrão CMIS - Content Management Interoperability Services, que é um padrão aberto criado para facilitar a interoperabilidade entre sistemas CMSs.

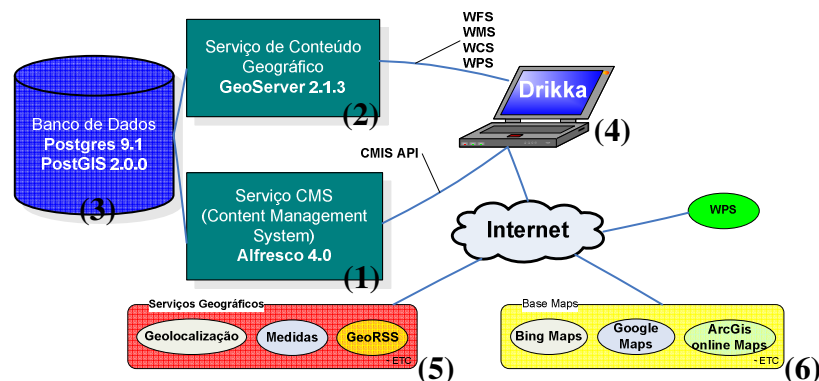


Figura 1: Arquitetura da Solução

A Figura 1 esboça a macro-arquitetura adotada para esta solução. Foi escolhida a ferramenta Alfresco como CMS (1), que é uma ferramenta desenvolvida em Java e possui uma versão Community que é de código aberto. A ferramenta Alfresco é utilizada em várias instituições no mundo, e no Brasil, um grande exemplo de seu emprego é na Dataprev (Empresa de Tecnologia e Informações da Previdência Social) onde é utilizada para facilitar a gestão de documentos durante o processo de compras na empresa.

Como gerenciador de conteúdo geográfico (2), adotou-se o GeoServer, que é um Software livre, de código aberto, mantido pelo Open Planning Project e que é capaz de integrar diversos repositórios de dados geográficos com simplicidade e alta performance. O GeoServer é um servidor de Web Map Service (WMS), Web Coverage Service (WCS) e de Web Feature Service (WFS) completamente funcional que segue as especificações da Open Geospatial Consortium (OGC), além disso ainda provê um barramento de serviços Web Processing Service (WPS), outro padrão OGC para serviços de processamento de dados.

Tanto o Alfresco quanto o GeoServer utilizam um banco de dados relacional (3) para persistência e, uma boa solução que atende a ambos, é o Postgres com a extensão espacial PostGIS em sua versão 2.0. A aplicação cliente foi desenvolvida em Flex e consome os dados tanto do CMS via CMIS quanto do gerenciador de conteúdo geográfico via WMS, WFS e WCS. Além destes serviços, a aplicação cliente ainda consome serviços on-line (web) para geolocalização, GeoRSS e medidas de feições (5) desenvolvidos como plugins e facilmente extensíveis. O plano de fundo do mapa (Base Map) (6) é um outro exemplo de serviço web consumido pela aplicação. Hoje é possível incluir Base Map do Google, Bing e ArcGis On-Line.

Muitos dos artefatos a serem manipulados por esta solução são imagens matriciais (raster) que são armazenadas no PostGIS. Para armazená-los no banco de dados é utilizada a função `raster2pgsql`, carregando-as em uma tabela. Cada dado raster é carregado em uma tabela própria. Após o carregamento dos dados raster, é criada uma representação vetorial (polígono) de sua área de extensão, isso é feito utilizando a função: `"SELECT ST_AsBinary(ST_Buffer(ST_Union(rast::geometry), 0.000001)) FROM raster_table"`. A estratégia de criar uma representação vetorial para os dados matriciais é, principalmente, por questões de performance em buscas e não causa efeito colateral, uma vez que visualmente uma imagem matricial propicia poucas informações relevantes.

O modelo de dados para os dados geográficos é muito simples, a princípio têm-se apenas classes de feição para pontos de interesse e áreas (polígonos) de interesse, com

identificador, descrição simples e um atributo específico para relacionar o elemento geográfico a um conteúdo no CMS. Como supramencionado, a estratégia de separar o conteúdo geográfico do CMS implica em uma integração entre estes dados. No CMS são cadastrados os metadados de cada um dos conteúdos e cada conteúdo armazenado no CMS possui um identificador único. Caso este conteúdo tenha componente geográfica a ferramenta possibilita que seja criada ou selecionada uma geometria de ponto ou polígono que se relacionará com o conteúdo no CMS através de seu identificador único.

Os conteúdos multimídias, tais como arquivos de texto, planilhas, vídeos, fotos e etc, são armazenados no CMS e utilizam o modelo de metadados Dublin Core. Dublin Core pode ser definido como sendo o conjunto de elementos de metadados planejado para facilitar a descrição de recursos eletrônicos, sendo um dos padrões mais conhecidos e tradicionalmente adotados em sistemas gerenciadores de conteúdo. Os dados matriciais utilizam um versão estendida do Dublin Core, adicionando alguns atributos específicos para tal. Os usuários do sistema optaram por especificar quais atributos são importantes para os dados matriciais ao invés de usar algum metadado existente, como FGDC (Federal Geographic Data Committee) ou ISO 19115, por entender que estes modelos apresentam uma quantidade muito grande de atributos que nem sempre são utilizados e que em geral representa um desincentivo ao uso.

6. Trabalhos Futuros

Como continuidade deste trabalho, está sendo desenvolvido um barramento de serviços WPS que de início proverá uma série de algoritmos de classificação para os dados matriciais. Contará ainda com melhorias no mecanismo de buscas geográficas para os conteúdos vinculados ao CMS.

7. Referencias

- BRAGA, G. M. Informação, ciência da informação: breves reflexões em três tempos. *Ciência da Informação*, v. 24, n. 1, p. 84-88, 1985.
- CAÑETE, S. C.; TAVARES, D. L. M.; ESTRELA, P. C.; FREITAS, T. R. O.; HENKIN R.; GALANTE, R., FREITAS, C. M. D. S.. Integrando visualização e análise de dados em sistema de gerenciamento de dados de biodiversidade. IV e-Science Workshop (SBC), 2010.
- DAVENPORT, T. H., PRUSAK, L.. *Conhecimento empresarial*. Rio de Janeiro: Campus, 1998.
- DAVENPORT, T. *Ecologia da informação: porque só a tecnologia não basta para o sucesso na era da informação*. São Paulo: Futura, 1998. 316p.
- DAVENPORT, T. H. Data to knowledge to results: building an analytic capability. *California Management Review*, v. 43, n. 2, p. 117-138, Winter 2001
- LEITE, F. C. L. *Gestão do Conhecimento Científico no Contexto Acadêmico: Proposta de um Modelo Conceitual*, 2006

- LEITE, F. C. L.; COSTA, S. Repositórios institucionais como ferramentas de gestão do conhecimento científico no ambiente acadêmico. 2006
- MEADOWS, A. J. A comunicação científica. Brasília: Briquet de Lemos, 1999. 268p.
- NONAKA, I.; TAKEUCHI, H. Criação do conhecimento nas empresas: Como as empresas japonesas geram a dinâmica da inovação. Rio de Janeiro, 1997. 358p.
- Smith, H. A.; McKeen, J. D. Developments in Practice VIII: Enterprise Content Management Communications of the AIS, 2003, pp. 647-659.
- Weiseth, P. E.; Olsen, H. H.; Tvedte, B.; Kleppe, A. eCollaboration Strategy 2002-2004, Statoil, Trondheim/Stavanger, 2002.