em questão
E-ISSN 1808-5245
Revista da Faculdade de Biblioteconomia e Comunicação da UFRGS
v. 21, n. 1 – Jan./Abr. 2015

Tecnologias da Web Semântica para a recuperação de dados agrícolas: um estudo sobre o *International Information System of the Agricultural Science and Technology (AGRIS)*

Fábio Mosso Moreira
Mestrando; Universidade Estadual Paulista (UNESP);
fabiomoreira@tupa.unesp.br

Ricardo César Gonçalves Sant'Ana
Doutor; Universidade Estadual Paulista (UNESP);
ricardosantana@marilia.unesp.br

José Eduardo Santarem Segundo
Doutor; Universidade de São Paulo (USP);
santarem@usp.br

Silvana Aparecida Borsetti Gregorio Vidotti
Doutora; Universidade Estadual Paulista (UNESP);
svidotti@gmail.com

Resumo: As informações provenientes da agricultura estão distribuídas entre diversos agentes e instituições. Muitas dessas informações são disponibilizadas em formato digital, podendo ser acessadas por intermédio de tecnologias como a web. Contudo, a diversidade na cobertura conceitual dos temas da área e a heterogeneidade léxica dos termos utilizados para pesquisas são entraves na recuperação de dados e informações agrícolas. As tecnologias da Web Semântica podem contribuir no sentido de padronizar a representação semântica dos recursos informacionais disponíveis na web reduzindo estes entraves. Nesse sentido, este trabalho tem como objetivo descrever as tecnologias semânticas utilizadas no *International Information System of the Agricultural Science and Technology* – AGRIS, e apontar o papel destas no processo de recuperação de dados e informações agrícolas. Por meio de uma pesquisa bibliográfica e uma análise exploratória no portal web do AGRIS, foram descritas as tecnologias utilizadas pelo sistema e suas principais contribuições.

Palavras-chave: Informações agrícolas. Tecnologias da Web Semântica. Recuperação de dados e informação.

1 Introdução

Na Ciência da Informação, são recentes os estudos que abordam as contribuições das novas Tecnologias da Informação e Comunicação (TIC) no âmbito da agricultura, tanto na questão do uso das TIC por parte dos produtores e habitantes rurais quanto na questão tecnológica de geração e disseminação de informações agrícolas.

Viero e Silveira (2011) ressaltam que as possibilidades proporcionadas pelas TIC para o meio rural são diversas, destacando seu papel na ampliação de horizontes e incorporação de novas expectativas; na constituição de grupos de comercialização; nas novas políticas públicas; nas estimativas de safras e desempenhos nas bolsas de valores e *commodities*; em serviços bancários; cooperativas de crédito e de produção; educação à distância e assistência técnica. Neste sentido, o meio rural não pode mais ser visto como um local distante e atrasado pelos que vivem no meio urbano e industrial, mas sim, como um ícone de diversidade que está em constante desenvolvimento, demandando cada vez mais informações atualizadas e constantes.

As informações provenientes da agricultura estão distribuídas entre diversos agentes e instituições como produtores, universidades, institutos de pesquisa, serviços de extensão rural, iniciativa privada e organizações não governamentais (ONGs). Muitas dessas informações são disponibilizadas em formato digital e podem ser acessadas por meio da *World Wide Web* (WWW), uma tecnologia de compartilhamento de informações que foi difundida amplamente nas últimas décadas. Para Berners-Lee (2010), a web permite um canal mundial de comunicação com fluxo contínuo acessível a partir de qualquer tipo de dispositivo que possa conectar-se com a internet. Os recursos informacionais são disponibilizados por intermédio de mídias e páginas da web que armazenam estoques de dados e informações seguindo a lógica de hipertexto.

Apesar da web constituir-se numa importante ferramenta tecnológica para a comunicação e disseminação de conhecimentos, ainda persistem problemas na recuperação por conteúdos informacionais sobre agricultura na web, tais como: heterogeneidade na definição dos domínios e ambiguidade léxica. Salokhe, Sini e Keizer (2007) destacam a existência de heterogeneidade na cobertura conceitual dos

temas da área, na qual alguns conteúdos abordam a agricultura genericamente enquanto que outros a abordam por meio de subdisciplinas específicas como nutrição, saúde animal e saúde das plantas. Roussey et al. (2010) ressaltam a presença de uma ambiguidade léxica que pode ser observada na utilização de diferentes termos para representar a mesma ideia em processos de buscas por dados e informações sobre agricultura.

Propõe-se que os problemas encontrados na recuperação de dados e informações agrícolas podem ser amenizados por meio da utilização de tecnologias da Web Semântica. Segundo Hendler e Berners-Lee (2009), a web semântica vem sendo desenvolvida nas últimas décadas e oferece suporte para um conjunto de tecnologias que exploram a padronização da representação semântica dos recursos informacionais disponíveis na web. Para os autores, a web semântica proporciona um mecanismo para a formatação de dados de uma forma interpretável por máquinas, ligando propriedades individuais destes dados com esquemas acessíveis globalmente, propiciando inferências sobre os dados em atividades e modos escalonáveis.

Soulignac e colaboradores (2009[1] apud ROUSSEY et al., 2010) destacam que o desenvolvimento de um portal web para recuperação de informação dedicado à agricultura que incorpore em seu sistema o uso de tecnologias da Web Semântica pode ser considerado uma possível solução para aperfeiçoar a disseminação de dados e informações agrícolas na web. A Organização das Nações Unidas para Alimentação e Agricultura (FAO), que tem como compromisso reforçar a agricultura e promover o desenvolvimento sustentável combatendo a fome e a pobreza, liderou nos anos 70 uma iniciativa que criou uma ampla cooperação para o acesso compartilhado em ciência e tecnologia agrícola, resultando posteriormente no desenvolvimento de um sistema de recuperação de dados e informações especializado neste segmento, o *International Information System of the Agricultural Science and Technology* – AGRIS (FAO, 2014b).

O AGRIS destaca-se por possuir um dos mais amplos sistemas de recuperação de dados e informação agrícola na web, apoiado por uma base de dados de domínio público global com mais de sete milhões de registros bibliográficos

estruturados em ciência e tecnologia agrícola, provenientes de mais de 150 repositórios institucionais e abertos, localizados em mais de 65 países. Os recursos armazenados no AGRIS envolvem artigos, dados, estatísticas e conteúdos multimídia que cobrem diversos temas relacionados à agricultura, pecuária, sistemas florestais, pesca e serviços aquáticos e nutrição humana (FAO, 2014b).

Neste sentido, o objetivo deste trabalho consiste em descrever as tecnologias da Web Semântica utilizadas no AGRIS, apontando o papel destas no processo de recuperação de dados e informações sobre agricultura, além de propor um novo olhar sobre a contribuição destas tecnologias em cada uma das fases do Ciclo de Vida dos Dados (SANT'ANA, 2013).

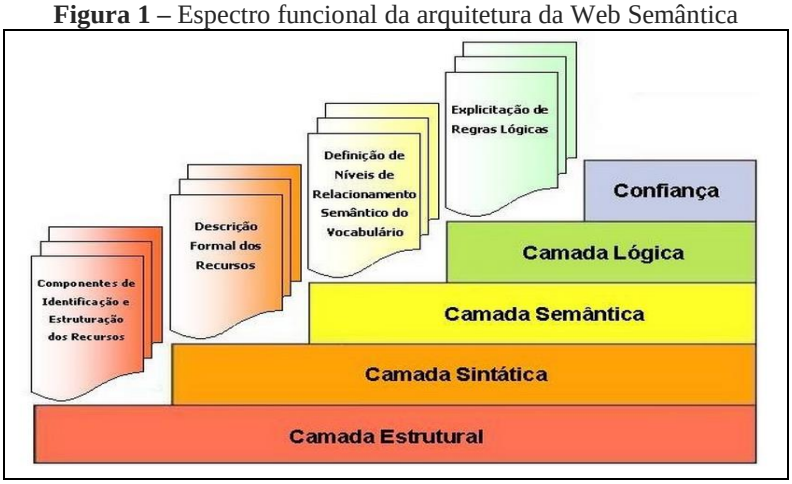
O artigo está estruturado em cinco seções: a presente introdução; uma seção com a descrição das tecnologias da Web Semântica; a metodologia utilizada para o desenvolvimento da pesquisa; os resultados obtidos com a descrição das tecnologias semânticas utilizadas no AGRIS e suas contribuições para a recuperação de dados e informações agrícolas; e por fim, as considerações finais do trabalho.

2 Tecnologias da Web Semântica

Para embasar o desenvolvimento das tecnologias semânticas de forma padronizada, Tim Berners-Lee (2005) e o World Wide Web Consortium (W3C) idealizaram uma estrutura padrão para compor a arquitetura da Web Semântica. Essa arquitetura é segmentada por camadas sobrepostas e escalonáveis, sendo que cada camada abarca recursos tecnológicos com funções específicas.

Ao estudar a arquitetura da Web Semântica, Ramalho, Vidotti e Fujita (2007) propuseram um "Espectro Funcional" para descrever, sob o olhar da Ciência da Informação, as camadas desta arquitetura e seus principais objetivos. Como ilustrado na Figura 1, os autores destacam cinco camadas principais: Camada Estrutural, Camada Sintática, Camada Semântica, Camada Lógica e Camada de Confiança.

Figura 1 – Espectro funcional da arquitetura da Web Semântica



Fonte: Ramalho, Vidotti e Fujita (2007, documento eletrônico não paginado).

A base desta arquitetura é composta pela Camada Estrutural, que tem como princípio a identificação de cada recurso informacional disponível na web de forma padronizada. A identificação dos recursos é dada por meio de uma *Uniform Resource Identifier* (URI), que, segundo Berners-Lee (2005), consiste em uma sequência compacta de caracteres que identifica um recurso informacional, abstrato ou físico, de modo uniforme, podendo ser classificada como um localizador – *Uniform Resource Locator (URL)*, um nome – *Uniform Resource Name (URN)*, ou uma combinação de ambos. Desta forma, a URI permite que pessoas, lugares e elementos do mundo físico possam ser referenciados no ambiente da web a partir de identificadores únicos.

A Camada Sintática tem como fundamento viabilizar a descrição dos recursos informacionais por meio da definição e validação de regras sintáticas. A principal tecnologia empregada para definir a sintaxe desta descrição é a linguagem *eXtensible Markup Language (XML)*, que consiste em um formato de texto simples e flexível, utilizado para representar documentos, conjuntos de dados, configurações, livros, transações, faturas, entre outros tipos de conteúdos por intermédio de *tags* de marcação (W3C, 2014a).

A estrutura definida para um arquivo XML pode ser descrita por meio de um XML *Schema*, um modelo de dados que fornece um inventário das possíveis construções de marcação XML e o que estas marcações representam no documento,

assim como a definição dos esquemas de *Namespace* e de seus elementos e atributos (W3C, 2012b). Para Santarém Segundo (2010), com a aplicação de um XML *Schema*, a estrutura e o conteúdo do arquivo XML são descritos por meio de elementos específicos, favorecendo a interoperabilidade na descrição dos recursos informacionais na web.

A linguagem XML viabiliza a representação dos recursos informacionais, contudo, não diz muito sobre o que estes significam. Para tanto, na Camada Semântica encontram-se tecnologias como o *Resource Description Framework (RDF)*, um *framework* usado para expressar declarações sobre os recursos informacionais e suas relações. O RDF tem sua sintaxe serializada na linguagem XML e sua representação baseada em um modelo gráfico de dados estruturado por triplas definidas pelos elementos: sujeito, predicado e objeto (W3C, 2014b).

Nessa representação, o sujeito é composto pelo recurso informacional que se pretende atribuir às declarações, geralmente referenciado pela sua URI; o predicado é a relação que será indicada pela afirmação, que, por sua vez, é caracterizada pelo objeto, atribuindo valor ao relacionamento (W3C, 2014b). Por meio de um RDF *Schema*, são previstos mecanismos para descrever grupos de recursos relacionados e as relações entre os mesmos, além de determinar características de outros recursos, como domínios e faixas de propriedade (W3C, 2014d).

Assim como a tecnologia RDF, no contexto descrito aqui, as ontologias figuram como um tipo de artefato computacional que também busca a inserção de ligações semânticas nos recursos informacionais na web. Ontologias são compostas por grupos de termos específicos projetados para propiciar a expansão de uma hierarquia organizada por esquemas de classificação e relação entre os termos de um domínio específico do conhecimento. Também existem linguagens para o desenvolvimento e compartilhamento de ontologias, sendo a *Web Ontology Language (OWL)* a recomendada pelo W3C para a construção de ontologias (W3C, 2012a).

Para atender aos princípios da Camada Semântica, também podem ser utilizados os tesauros ou vocabulários controlados, que, assim como o RDF e as ontologias, buscam descrever e representar uma área de interesse por meio da

definição de conceitos e relações entre os mesmos. Geralmente, os princípios são aplicados para classificar os termos que poderão ser utilizados em uma aplicação específica, assim como caracterizar as possíveis relações e definir restrições sobre o uso dos mesmos.

Não existe uma divisão clara entre o que pode ser referenciado como uma ontologia ou um vocabulário, já que a tendência é usar o termo "ontologia" para coleções de termos mais complexos e formais, enquanto que "vocabulário" é utilizado quando tal formalismo estrito não é necessariamente considerado (W3C, 2014c).

Para recuperar dados estruturados em arquivos RDF e OWL, pode ser utilizada a linguagem de consulta SPARQL, que permite consultas em uma base de dados específica, ou inclusive, entre diversas bases de dados ao mesmo tempo, tendo como resultado dados em forma de tabelas ou grafos RDF (W3C, 2013).

As camadas superiores – Camada Lógica e Camada de Confiança – ainda não possuem tecnologias consolidadas, contudo, possuem seus objetivos bem definidos. Para Ramalho, Vidotti e Fujita (2007), a Camada Lógica tem como proposta possibilitar a verificação e comprovação da coerência lógica dos recursos informacionais, fazendo com que os aspectos semânticos das informações estejam descritos de maneira consideravelmente adequada, atendendo assim requisitos das camadas inferiores. Com a Camada de Confiança, espera-se garantir que as informações estejam representadas de modo correto, possibilitando certo grau de confiabilidade.

Com as descrições das camadas da Web Semântica, acompanhadas das definições das tecnologias semânticas associadas a cada uma, é possível analisar a aplicação destas em sistemas de recuperação de dados e informação. Na próxima seção deste texto são levantadas as possibilidades e contribuições da utilização das tecnologias da Web Semântica para o acesso e a disseminação de informações agrícolas por meio de um sistema especializado na área da agricultura.

3 Metodologia

A metodologia utilizada para a realização desta pesquisa consistiu de levantamento bibliográfico de textos que abordam a temática da Web Semântica na área da Ciência da Informação. A partir deste levantamento, foi possível descrever a arquitetura da Web Semântica e identificar as principais tecnologias associadas a cada uma das camadas descritas neste trabalho.

Realizou-se, então, um estudo exploratório no portal web do AGRIS (http://agris.fao.org) a fim de caracterizar a estrutura do sistema e demonstrar como as tecnologias da Web Semântica foram aplicadas neste contexto. As informações referentes à descrição do AGRIS foram levantadas em áreas específicas do portal, como *'About', 'How it works', 'AGRIS centers', 'AGROVOC'*, As informações foram complementadas com outras levantadas nos artigos de Anibaldi e colaboradores (2013), que trata sobre o processo de reestruturação da base de dados AGRIS apoiada pelas tecnologias da Web Semântica e de Roussey e colaboradores (2010), que aborda a utilização de ontologias em sistemas de recuperação de informação na área da agricultura, compondo os resultados desta pesquisa.

Com a demonstração de um processo de busca realizada no AGRIS utilizando como termo "agricultura familiar", foi possível observar e descrever como as tecnologias da Web Semântica atuam e podem contribuir para o processo de recuperação de dados e informações agrícolas neste sistema. A fim de garantir um melhor entendimento da atuação das tecnologias semânticas em todas as fazes do processo – coleta, armazenamento e recuperação –, foi aplicado o modelo de Ciclo de Vida dos Dados (SANT'ANA, 2013).

4 International Information System of the Agricultural Science and Technology (AGRIS)

A partir da análise exploratória realizada no Portal AGRIS e das informações levantadas nas pesquisas de Anibaldi e colaboradores (2013) e Roussey e colaboradores (2010), identificou-se que o sistema atribui um identificador único

(URI) para cada um dos recursos informacionais armazenados. Cada registro indexado é identificado por um AGRIS *Record Number* (ARN), contendo uma estrutura pré-definida baseada na sigla do país de origem do repositório fornecedor, no ano de criação do conteúdo e no código progressivo atribuído pelo sistema no momento da indexação.

Por exemplo, o recurso com o ARN "ES2011001090" refere-se a um registro de um conteúdo criado no ano de 2011 por um repositório fornecedor de dados localizado na Espanha (ES), cujo código primário no sistema AGRIS é "001090". Esse identificador, além de distinguir todos os recursos de forma única, permite que o AGRIS faça um *link* formal dos resultados do sistema com outros recursos disponíveis em bases de dados externas, por exemplo, os recursos provenientes da *DBpedia*.

Os metadados dos conteúdos provenientes dos repositórios fornecedores do AGRIS são formatados em arquivos XML e armazenados na base de dados AGRIS XML. O processo é baseado em operações de transformação e enriquecimento de dados, e requer a utilização de sistemas com filtros para checar a correção de algumas informações e adicionar outros dados utilizando fontes de dados pré-definidas.

Para incorporar ligações semânticas aos recursos indexados, é realizada a conversão dos arquivos XML para o formato RDF, estes então são armazenados na base AGRIS *Record*. Esse processo envolve a definição de vocabulários padrões e propriedades necessárias para a modelagem dos dados; para tanto, são utilizados vocabulários tradicionais como BIBO, FOAF e Dublin Core, e um tesauro específico para área da agricultura, o *AGROVOC*.

O *AGROVOC* é um vocabulário controlado agrícola multilíngue com mais de 32.000 conceitos traduzidos para mais de 20 idiomas, cobrindo assuntos na agricultura, pecuária, sistemas florestais e pesca. Atuando como um "*backbone*", o *AGROVOC* permite realizar interligações dos recursos AGRIS com bases de dados externas, tais como *DBpedia, World Banck, Google Custom Search API, Nature Open Search, FAO Geopolitical Ontology – Country profiles, Global Biodiversity*

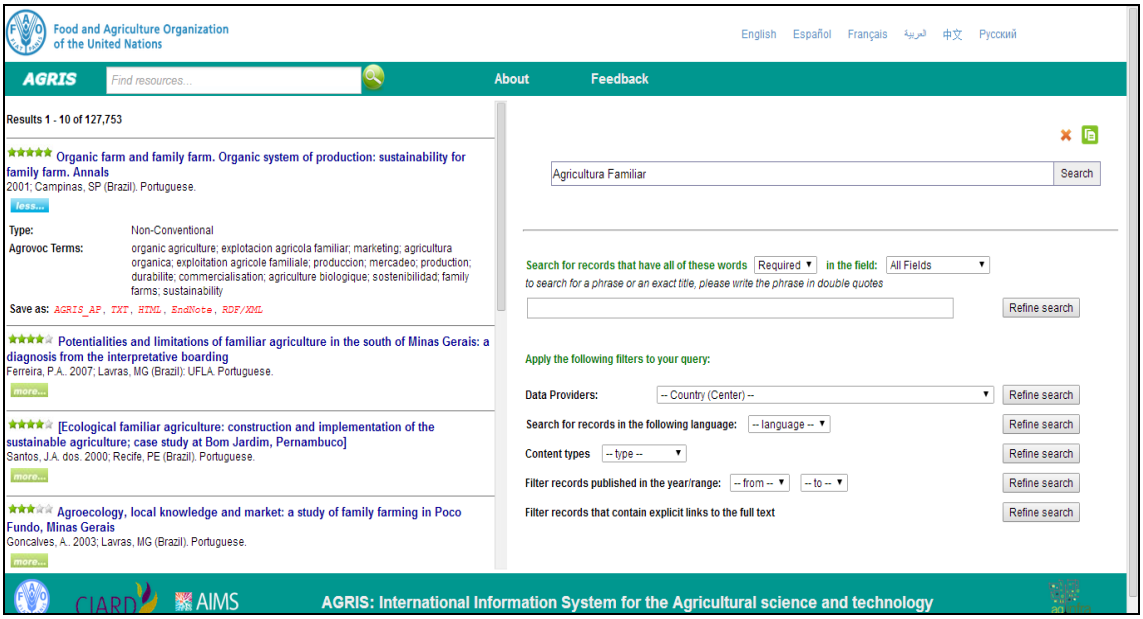
Information Facility, International Food Policy Research Institute – IFPRI, FAO Fisheries and Aquaculture fach sheets API e *Bioversity International*.

Assim como a base de dados AGRIS XML e AGRIS *Record*, também integram o sistema a base de dados AGRIS *Serials*, responsável por armazenar dados sobre periódicos agrícolas mundiais, e a base de dados do *AGROVOC*, responsável por interligar os recursos armazenados com as bases de fontes externas.

Como meio de consultar os recursos indexados pelo AGRIS interligando os conjuntos de dados do sistema com fontes externas, foi desenvolvida uma plataforma de consulta baseada na tecnologia SPARQL. A plataforma *Open* AGRIS é um *webservice* integrado ao AGRIS que permite, a partir de um registro recuperado pelo sistema, visualizar estatísticas, mapas, gráficos, tabelas dinâmicas, indicadores do Banco Mundial, dados do *DBPedia*, e demais artigos relacionados ao conteúdo do recurso selecionado.

O AGRIS pode ser acessado pelo endereço eletrônico <http://agris.fao.org/>, que direciona para a página inicial do portal, a qual contém do uma "*search box"* localizada no canto superior esquerdo da interface, onde o usuário insere sua expressão de busca definida por um termo específico que expressa sua necessidade informacional, como ilustrada na Figura 2.

Figura 2 – Sistema de busca do AGRIS exibindo resultados para pesquisa sobre "Agricultura Familiar"



Fonte: FAO (2014a).

O resultado da pesquisa é exibido em uma página contendo, no lado esquerdo, os dados bibliográficos dos recursos recuperados e, no lado direito, ferramentas para aperfeiçoamento de busca como a atribuição de parâmetros específicos para formulação de busca avançada (fornecedor de dados, país de origem do recurso, tipo do conteúdo a ser recuperado). Na lista de resultados são exibidas as palavras-chave do *AGROVOC* atribuídas ao recurso recuperado junto aos dados de registro bibliográfico como título, demais dados sobre o país de origem, idioma, ano e tipo do recurso.

Ao selecionar um dos recursos da lista de resultados, o usuário é direcionado para a plataforma *Open* AGRIS, que realiza a consulta nas bases do sistema por meio de um terminal SPARQL e exibe dados e informações do AGRIS associadas às fontes externas da web, ativando os atalhos semânticos entre os elementos ligados pelas triplas RDF. Se o recurso selecionado refere-se a um artigo de periódico científico, o sistema consulta a base de dados AGRIS *Serials*, fornecendo também dados sobre outros artigos e periódicos que cobrem temas semelhantes. Porém, nem sempre os registros resgatados pelo AGRIS possuem o *link* para o arquivo com o

conteúdo original do recurso, contudo, são fornecidos dados e informações sobre a fonte de origem no qual o usuário pode proceder com sua busca.

Por intermédio da plataforma *Open* AGRIS, o usuário tem a possibilidade de visualizar o arquivo RDF do registro utilizando-se de uma aplicação que exibe as triplas contendo as relações de propriedade e valor do recurso selecionado. O visualizador de RDF pode ser acessado por meio de um atalho disponível na plataforma, localizado abaixo do título do recurso, como ilustrado na Figura 3. Outra opção disponibilizada pela plataforma é a ferramenta '*LodLive*', uma tecnologia que consiste de um visualizador de recursos RDF baseada na representação gráfica das relações de propriedade e objeto (valor), capaz de conectar-se a recursos externos. A partir de um círculo central (sujeito), o usuário pode interagir com a ferramenta exibindo as relações semânticas entre diversos recursos por meio de links entre bases de dados externas interligadas com as bases AGRIS.

Figura 3 – Interface da plataforma *Open* AGRIS após a seleção de um recurso na lista de resultados para a pesquisa sobre "Agricultura Familiar"

Fonte: FAO (2014a).

5 Contribuições das tecnologias da Web Semântica no AGRIS

Os resultados obtidos com a pesquisa exploratória realizada no AGRIS e estruturada em função do espectro funcional da arquitetura da Web Semântica permitiram identificar tecnologias semânticas aplicadas no sistema AGRIS e suas principais funcionalidades. No Quadro 1 encontram-se relacionadas as camadas da arquitetura da Web Semântica usadas pelo AGRIS acompanhadas da descrição das tecnologias semânticas identificadas na pesquisa e suas principais contribuições para o funcionamento do sistema de recuperação.

Observou-se que as primeiras camadas são aplicadas no AGRIS mediante a utilização de tecnologias como URI, XML, RDF, vocabulários controlados e utilização de um terminal SPARQL para consulta. Contudo, não foram verificados indícios de que o AGRIS utiliza-se de tecnologias para garantir os objetivos da Camada Lógica e da Camada de Confiança.

Quadro 1 – Arquitetura e tecnologias da Web Semântica no AGRIS

Arquitetura da Web Semântica	Tecnologias semânticas no AGRIS	Contribuições para o AGRIS
Camada Estrutural	O AGRIS atribui um identificador único (URI) para cada recurso armazenado em sua base de dados. Definido como ARN (*AGRIS Record Number*), o código possui uma estrutura pré-definida que é composta pela sigla do país de origem do repositório fornecedor do recurso, o ano em que o mesmo foi criado e o código representativo do registro atribuído pelo sistema na indexação.	A atribuição de um ARN permite que cada recurso armazenado no AGRIS possua um identificador único na web e provê meios para representar cada recurso descrevendo seu mecanismo de acesso primário ou sua localização.
Camada Sintática	Os recursos provenientes dos repositórios fornecedores do AGRIS são traduzidos para o formato XML. O processo é baseado em operações de reestruturação dos metadados por meio da definição de regras sintáticas e do enriquecimento de seus dados com outros dados de fontes externas.	Ao converter os metadados dos conteúdos indexados para o formato XML, é possível aplicar filtros para correção de algumas informações, além de adicionar outros dados utilizando fontes de dados pré-definidas. O formato XML também proporciona ao AGRIS um ganho de eficiência no processo de recuperação dos recursos, facilitando a identificação de dados e permitindo, inclusive, a localização e combinação de fontes externas.

Continua.

Conclusão.

Arquitetura da Web Semântica	Tecnologias semânticas no AGRIS	Contribuições para o AGRIS
Camada Semântica	Os arquivos da base AGRIS XML são convertidos para o formato RDF. Esse processo envolve a definição de vocabulários padrões tradicionais que visam facilitar a interligação de seus dados com outros conjuntos de dados disponíveis em bases externas. Utilizando o tesauro *AGROVOC* (vocabulário específico do domínio da agricultura) é possível realizar consultas na base de triplas AGRIS RDF de forma interligada com bases externas, esta consulta é realizada utilizando um terminal SPARQL disponibilizado pela plataforma *Open* AGRIS, um *webservice* integrado ao sistema.	O formato RDF permite a modelagem dos dados com base em vocabulários padrões tradicionais como o BIBO e o FOAF e Dublin Core, inserindo ligações semânticas nos recursos AGRIS. A utilização do *AGROVOC* permite realizar interligações com bases de dados externas utilizando as palavras-chave atribuídas ao recurso, providenciando o máximo de dados possíveis referentes a um determinado tópico pesquisado. A plataforma *Open* AGRIS, por intermédio de um terminal SPARQL, ativa as ligações semânticas entre os elementos ligados pelas triplas RDF, que torna possível a consulta dos resultados em forma de estatísticas, mapas, indicadores e demais artigos relacionados ao assunto.

Fonte: elaborado pelos autores.

Para entender melhor como as tecnologias semânticas utilizadas no AGRIS atuam de forma complementar e escalonável seguindo os princípios da arquitetura da Web Semântica, aplicou-se um modelo para facilitar o entendimento de como os dados podem ser disponibilizados e acessados com base nos fundamentos da Ciência da Informação. O Ciclo de Vida dos Dados (SANT'ANA, 2013) que propõe uma estrutura para estudo e acompanhamento das atividades envolvidas no acesso, manutenção e disponibilização dos dados é composto por quatro fases: coleta, armazenamento, recuperação e descarte.

A fase da coleta envolve ações relacionadas ao planejamento de como serão obtidos, filtrados e organizados os dados, definindo-se a estrutura, formato e meios de descrição a serem utilizados.

A fase de armazenamento envolve atividades relacionadas ao processamento, transformação, inserção, migração, transmissão e toda e qualquer ação que vise à persistência dos dados.

A fase de recuperação constitui-se no acesso aos dados pelo usuário e

envolve a consulta e a visualização dos dados. Neste ciclo, considera-se também a fase de descarte dos dados, que pode ocorrer por meio da migração da base de dados para outro contexto/ambiente, ou, simplesmente, com a eliminação destes dados depois de cumpridas suas finalidades.

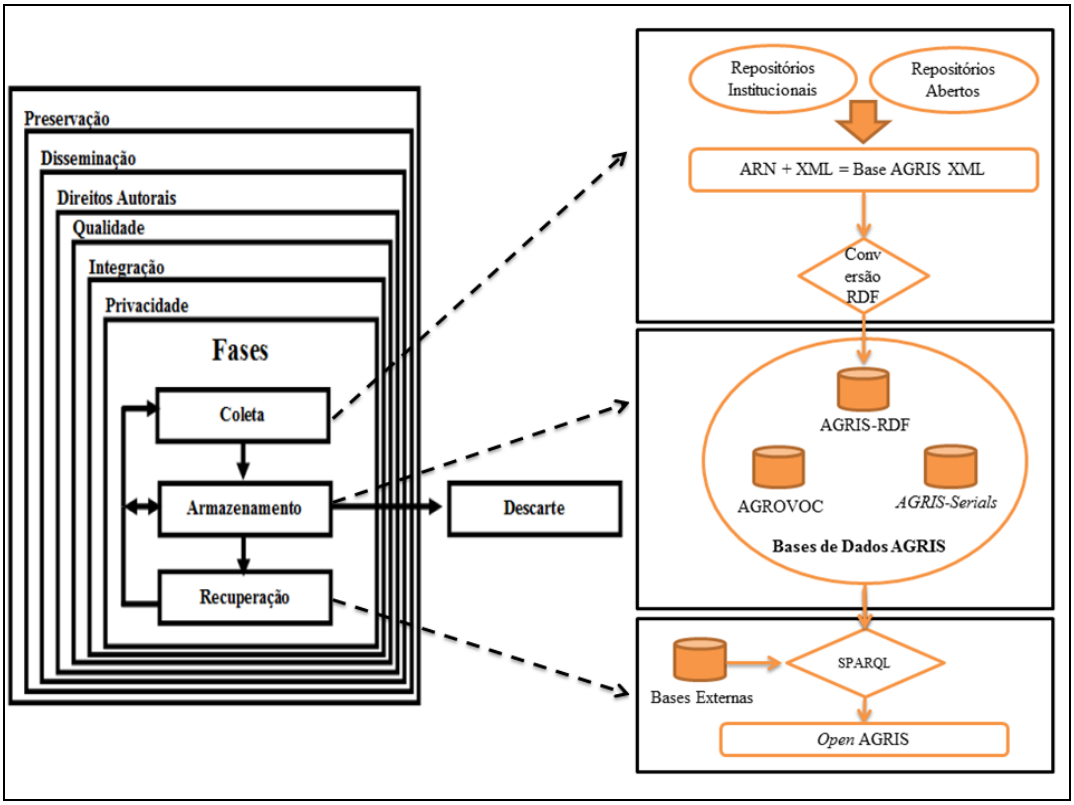
Conhecidas as fases envolvidas no Ciclo de Vida dos Dados, aplicaram-se as quatro fases o modelo no processo de descrição das tecnologias semânticas utilizadas no sistema AGRIS, conforme descrito no esquema ilustrado na Figura 4.

Na fase de coleta, ocorreu o processo de indexação dos dados provenientes dos repositórios institucionais abertos dos fornecedores do AGRIS, no qual os recursos recebem um identificador único (ARN) e são traduzidos para linguagem XML.

Na fase de armazenamento que compreende a preservação em formato digital dos dados, o conteúdo XML é convertido para RDF e armazenado na base AGRIS *Record* (que contém os arquivos RDF dos recursos), na AGRIS *Serials* (que armazena informações sobre os periódicos agrícolas mundiais) e na base do *AGROVOC*, compondo assim o conjunto total de bases de dados que integram o AGRIS.

Na fase de recuperação dos dados é disponibilizada a linguagem SPARQL para consulta nas bases de dados do AGRIS e fontes de dados externas de forma interligada. Esse processo é realizado pela plataforma *Open* AGRIS, por meio da qual é possível recuperar os dados das bases AGRIS *Record*, *AGROVOC*, AGRIS *Serials* e demais bases externas de forma relacional. A plataforma *Open* AGRIS também disponibiliza tecnologias para a visualização dos arquivos RDF de forma integrada ao sistema.

Figura 4 – Aplicação do Modelo de Ciclo de Vida dos Dados no processo de disponibilização e acesso a dados no AGRIS em função das tecnologias semânticas utilizadas



Fonte: elaborado pelos autores com base em Sant'Ana (2013).

O Ciclo de Vida dos Dados também envolve uma série de objetivos que permeiam todas as suas fases. Os objetivos são relacionados ao atendimento de critérios relacionados à privacidade, integridade, qualidade, direitos autorais e disseminação e preservação dos dados. Desenvolver ações no sentido de garantir esses critérios no AGRIS poderia significar um avanço em relação à concretização das camadas da arquitetura da Web Semântica que ainda não foram consolidadas no sistema (Camada Lógica e Camada de Confiança).

6 Considerações finais

O AGRIS e sua plataforma *Open* AGRIS figuram como opções para auxiliar na disseminação de dados e informações sobre agricultura na web. O sistema incorpora tecnologias semânticas em sua estrutura, buscando reduzir fatores que interfiram no

processo de recuperação de dados e informações sobre agricultura como a ambiguidade léxica e a heterogeneidade na cobertura conceitual dos temas relacionados à agricultura.

A utilização de um tesauro específico da agricultura na indexação dos registros configura-se como um diferencial do sistema AGRIS. Esse tesauro, o *AGROVOC*, permite interligar os dados do AGRIS com outros conjuntos de dados existentes em bases de dados externas. Esse processo é realizado por meio da plataforma *Open* AGRIS que objetiva expandir a busca de dados referentes a um tópico ou recurso bibliográfico pesquisado no AGRIS por meio do uso de tecnologias da Web Semântica.

Com base no Ciclo de Vida dos Dados foi possível observar como as tecnologias semânticas presentes no AGRIS atuam de forma complementar e escalonável, proporcionando uma nova abordagem sobre as funcionalidades destas tecnologias para o processo de recuperação, desde a coleta dos dados dos repositórios fornecedores, passando pelo armazenamento até a disponibilização dos recursos.

Considera-se importante que outras pesquisas sejam desenvolvidas no sentido de analisar, também, ações e políticas que possam garantir o atendimento de critérios relacionados aos objetivos (preservação, disseminação, direitos autorais, qualidade, integração e privacidade) que permeiam as fases do Ciclo de Vida dos Dados no sistema AGRIS.

Referências

ANIBALDI, Stefano et al. Migrating bibliographic datasets to the semantic web: the AGRIS case. **Semantic Web**: Interoperability, Usability, Applicability an IOS Press Journal, v.2, p.1-9, 2013. Disponível em: <http://www.semantic-web-journal.net/system/files/swj463.pdf>. Acesso em: 10 jan. 2015.

BERNERS-LEE, Tim. **Uniform Resource Identifier (URI)**: Generic Sintax. 2005. Disponível em: <http://tools.ietf.org/html/rfc3986>. Acesso em: 23 abr. 2014.

BERNERS-LEE. Tim. Long Live the Web. **Scientific American**, New York, v. 303, n.6, p.80-85, 2010. Disponível em:

<http://www.cs.virginia.edu/~robins/Long_Live_the_Web.pdf>. Acesso em: 14 abr. 2014.

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS (FAO). **AGRIS:** International Information System for the Agricultural Science and Technology. Roma: FAO, 2014a. Disponível em: <http://agris.fao.org/agris-search/index.do>. Acesso em: 10 jan. 2014.

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS (FAO). **Knowledge and information sharing through the AGRIS Network**. Roma: FAO, 2014b. Disponível em: <http://agris.fao.org/knowledge-and-information-sharing-through-agris-network>. Acesso em: 10 jan. 2014.

HENDLER, Jim; BERNERS-LEE, Tim. From the Semantic Web to social machines: a research challenge for AI on the World Wide Web. **Artificial Intelligence Journal**, v. 174, p. 156-161, 2009. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0004370209001404>. Acesso em: 10 jan 2015.

RAMALHO, Rogério; VIDOTTI, Silvana; FUJITA, Mariângela. Web Semântica: uma investigação sob o olhar da Ciência da Informação. **DataGramaZero – Revista de Informação**, Rio de Janeiro, v. 8, n. 6, 2007. Disponível em: <http://www.dgz.org.br/dez07/Art_04.htm>. Acesso em: 10 jan. 2015.

ROUSSEY, Catherine et al. Ontologies in Agriculture. In: INTERNATIONAL CONFERENCE ON AGRICULTURAL ENGINEERING, 11., 2010, Shanghai. **Proceedings...** Clermont-Ferrand: Ed. LIRIS, 2010. p. 178-188.

SALOKHE, Gauri; SINI, Margherita; KEIZER, Johannes. **Case Study:** the semantic web for the agricultural domain, semantic navigation of food, nutrition and agriculture journal. 2007. Disponível em: <http://www.w3.org/2001/sw/sweo/public/UseCases/FAO/>. Acesso em: 04 fev. 2014.

SANT'ANA, Ricardo. Ciclo de Vida dos Dados e o papel da Ciência da Informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 14., 2013, Florianópolis. **Anais...** Rio de Janeiro: Associação Nacional em Pesquisa e Pós-Graduação em Ciência da Informação (ANCIB), 2013.

SANTAREM SEGUNDO, José Eduardo. **Representação Iterativa:** um modelo para repositórios digitais. 2010. Tese (Doutorado em Ciência da Informação) – Programa de Pós Graduação em Ciência da Informação, Universidade Estadual Paulista, Marília, 2010. Disponível em: <http://www.marilia.unesp.br/Home/Pos-Graduacao/CienciadaInformacao/Dissertacoes/santaremsegundo_je_do_mar.pdf >. Acesso em: 20 jan. 2015.

VIERO, Verônica; SILVEIRA, Ada. Apropriação de Tecnologias de Informação e Comunicação no meio rural brasileiro. **Cadernos de Ciência e Tecnologia – Embrapa**, Brasília, v. 28, n. 1, p. 257-277, 2011. Disponível em: <http://seer.sct.embrapa.br/index.php/cct/article/view/12042>. Acesso em: 17 jan. 2015.

WORLD WIDE WEB CONSORTIUM (W3C). **Ontologies.** Massachusetts: W3C/MIT, 2014c. Disponível em: <http://www.w3.org/standards/semanticweb/ontology>. Acesso em: 11 abr. 2014.

WORLD WIDE WEB CONSORTIUM (W3C). **OWL 2 Web Ontology Language Document Overview (Second Edition).** Massachusetts: W3C/MIT, 2012a. Disponível em: <http://www.w3.org/TR/owl2-overview>. Acesso em: 01 maio 2014.

WORLD WIDE WEB CONSORTIUM (W3C). **OWL 2 Web Ontology Language:** semantics and abstract syntax. Massachusetts: W3C/MIT, 2014a. Disponível em: <http://www.w3.org/TR/owl-semantics>. Acesso em: 01 maio 2014.

WORLD WIDE WEB CONSORTIUM (W3C). **RDF 1.1 Concepts and Abstract Syntax**. Massachusetts: W3C/MIT, 2014b. Disponível em: <http://www.w3.org/TR/rdf11-concepts>. Acesso em: 11 abr. 2014.

WORLD WIDE WEB CONSORTIUM (W3C). **RDF Schema 1.1**. Massachusettes: W3C/MIT, 2014d. Disponível em: <http://www.w3.org/TR/rdf-schema/#ch_introduction>. Acesso em: 22 jan. 2015.

WORLD WIDE WEB CONSORTIUM (W3C). **SPARQL 1.1 Overview**. Massachusetts: W3C/MIT 2013. Disponível em: <http://www.w3.org/TR/2013/REC-Sparql11-overview-20130321/>. Acesso em: 11 abr. 2014.

WORLD WIDE WEB CONSORTIUM. **W3C XML Schema Definition Language (XSD) 1.1 Part 1**: Structures. Massachusetts: W3C/MIT, 2012b. Disponível em: <http://www.w3.org/TR/xmlschema11-1/>. Acesso em: 17 jan. 2015.

Semantic Web technologies for retrieval of data and agricultural information: a study of *International Information Systemm of the Agricultural Science and Technology (AGRIS)*

em questão E-ISSN 1808-5245
Revista da Faculdade de Biblioteconomia e Comunicação da UFRGS
v. 21, n. 1 – Jan./Abr. 2015

Abstract: The agricultural information is distributed in different actors and institutions. A lot of this information is available in digital format and can be accessed with web technologies. However, the diversity in conceptual coverage of the topics in the area of lexical heterogeneity of terms used for research are barriers in data retrieval of agricultural information. The Semantic Web technologies can contribute in reducing those barriers by standardizing the semantic representation of information resources available on the web. Thus, this paper aims to describe the semantic technologies used in the Information System of the International Agricultural Science and Technology - AGRIS, and point out their role in retrieval of data and agricultural information. Through a literature review and an exploratory analysis on the AGRIS web portal we describe how technologies can be used by the system and its main contributions.

Keywords: Agricultural information. Semantic Web. Retrieval of data and information.

---

[1] Soulignac, V. et al. Knowledge management and innovative design, state of the art. INTERNATIONAL CONFERENCE ON THE MODERN INFORMATION TECHNOLOGY IN THE INNOVATION PROCESSES OF THE INDUSTRIAL ENTERPRISES, 11., Bergamo, Italy, 2009. **Proceedings...** Bergamo, 2009.

Recebido: 17/09/2014
Aceito: 20/02/2015