

NAVEGAÇÃO SEMÂNTICA SOBRE UMA BASE DE METADADOS RDF IMPLEMENTADA NO PROJECTO OMNIPAPER*

Teresa Susana Mendes Pereira, Ana Alice Baptista

Universidade do Minho
Campus de Azurém, 4800-058, Guimarães, Portugal
{tpereira, analice}@dsi.uminho.pt

Resumo

O projecto OmniPaper (*Smart Access to European Newspapers*) é um projecto do programa IST (*Information Society Technologies*) da comissão europeia, que visa fomentar um acesso distribuído a diferentes tipos de fontes de informação. Com este projecto pretende-se chegar a um protótipo de um sistema que permita aos utilizadores (quer ocasionais, quer profissionais) um acesso estruturado, personalizado e multilingue a todo um conjunto de artigos de notícias disponibilizados pelas empresas parceiras (a que denominamos por arquivos locais). Este artigo pretende descrever o trabalho desenvolvido na implementação do protótipo RDF no âmbito do projecto OmniPaper, focando a utilização dos *IPTC Subject Codes* na definição de uma camada de navegação semântica sobre as descrições de metadados codificados em RDF/XML.

Keywords: Metadados, Base de Metadados, *Resource Description Framework* (RDF), *IPTC Subject Codes*, Ontologia, e Navegação.

1. INTRODUÇÃO

Nas últimas décadas o crescimento da informação digital foi exponencial. O mesmo se verifica no crescimento da Internet. A informação está cada vez mais disponível em formato digital e a sua acessibilidade através da Internet tem vindo a aumentar rapidamente. Este crescimento e disponibilidade contribui para a necessidade de agrupar a informação a nível semântico, uma vez que tanto o seu acesso como a comparação com outras fontes que se encontram geograficamente dispersas é fisicamente suportado pela Internet.

Um dos importantes desafios dos utilizadores da Web, reside na descoberta dos recursos electrónicos de que efectivamente necessitam. Uma das formas de facilitar essa descoberta consiste em organizá-los previamente.

É neste contexto, que surge o projecto OmniPaper, que pretende investigar formas de promover o acesso distribuído a diferentes tipos de fontes de informação.

Um dos aspectos fundamentais deste projecto consiste na definição de uma camada de metadados do sistema, utilizados na descrição dos diferentes géneros de artigos, promovendo uma pesquisa mais eficiente na Web.

Conceptualmente a arquitectura do projecto OmniPaper é constituída por duas camadas de metadados: (1) uma primeira camada, denominada por *Local Knowledge Layer* é adicionada aos arquivos locais e tem como principal objectivo proporcionar uma descrição de todos os artigos existentes; e (2) uma segunda camada denominada por *Overall Knowledge Layer*, encontra-se a um nível de abstracção mais elevado, e recorre á primeira para proporcionar um ambiente integrado e estruturado de navegação e pesquisa, possibilitando quando possível, uma ligação a um ambiente multilingue. Esta camada pretende adaptar novas funcionalidades, nomeadamente a navegação integrada através da informação relativa aos diferentes arquivos.

No âmbito do projecto OmniPaper foram implementados em paralelo dois protótipos utilizando também duas tecnologias distintas de manipulação de metadados, tanto na camada *Local Knowledge Layer* como na camada *Overall Knowledge Layer*: a tecnologia *Resource Description Framework* (RDF) [1] e a tecnologia *Topic Maps* (TM) [2].

Este artigo pretende descrever o trabalho elaborado na implementação do protótipo RDF desenvolvido no âmbito do projecto OmniPaper, focando a utilização dos *IPTC Subject Codes* (IPTC-SC) [3], na definição de uma camada de navegação semântica sobre as descrições de metadados estruturados em RDF/XML, promovendo uma maior eficácia na pesquisa de artigos de notícias sobre a Web.

Este artigo apresenta o trabalho conduzido na representação hierárquica dos IPTC-SC, implementado no protótipo RDF. A estrutura do artigo está organizada com as seguintes secções: na Secção 2 será apresentado o termo Ontologia e serão analisadas

algumas das principais linguagens de representação destas. Na secção 3 será apresentado o trabalho conduzido na implementação da Ontologia que representa a estrutura hierárquica dos IPTC-SC. Por fim será apresentado na Secção 4 as conclusões e o trabalho futuro.

2. ONTOLOGIAS

Tradicionalmente o termo Ontologia é descrito como um ramo da metafísica aplicada à classificação exaustiva da natureza dos seres humanos [4].

Nos dias de hoje é normalmente aplicado em vários domínios, em particular na área da Ciência da Informação, designadamente na partilha e reutilização da representação formal do conhecimento [5]. Enquanto que na área da programação lógica as ontologias são definidas com duas funções principais: (1) " proporcionar um modo de ver o mundo, e portanto um processo de organizar a informação "; (2) " permitir a interoperabilidade de vocabulários e na definição de relações entre termos, segundo o seu conteúdo semântico" [5].

As ontologias estão organizadas e estruturadas através de conceitos e não por palavras [6]. As ontologias contêm um conjunto de conceitos semânticos e relações que permitem gerar várias interpretações qualificadas. Algumas ontologias permitem a definição de axiomas ou relações lógicas entre condições com o mesmo objectivo [7]. A selecção da ontologia é efectuada com base na estrutura do conhecimento que se pretende representar. Deste modo, no âmbito do projecto OmniPaper, em particular na abordagem RDF foi desenvolvido um estudo sobre várias linguagens de representação de ontologias, de forma a seleccionar a que melhor se adequa à representação, da estrutura hierárquica dos IPTC-SC.

Nas seguintes subsecções são apresentadas algumas iniciativas de linguagens de representação de ontologias. A selecção das linguagens apresentadas foi efectuada tendo em conta apenas as que estão associadas à Web Semântica.

2.1 RDF-S

O *Resource Description Framework Schema* RDF-S é a primeira linguagem a ser apresentada, atendendo à sua aceitação na comunidade da Web Semântica, e pelo facto de, no fim, ter sido a tecnologia seleccionada para representar a estrutura hierárquica dos IPTC-SC. O RDF-S [8] é uma recomendação oficial da *World Wide Web Consortium* (W3C) [9] desde Fevereiro de 2004.

O *Resource Description Framework* - RDF contém um modelo para expressar semântica. O RDF é um formato para representar conhecimento, que pretende descrever recursos através de metadados. A própria especificação do RDF-S auto denomina-se como *RDF Vocabulary Description Language* [10]. Através do RDF-S podem-se desenhar e implementar de uma forma consistente, vocabulários de metadados específicos. Estes podem ainda ser desenvolvidos no seio de outros projectos gerando, assim uma rede de esquemas de metadados.

O RDF-S define propriedades e classes que podem ser utilizadas para dar valores a estas mesmas propriedades. Contém mecanismos para descrever grupos de recursos assim como relações entre recursos.

2.2. OIL

O acrónimo OIL é apresentado com dois significados distintos: "*Ontology Inference Layer*", ou "*Ontology Interchange Language*" [11].

A sintaxe da linguagem OIL é particularmente direccionada para os leitores e escritores destas ontologias. No âmbito das máquinas, a linguagem OIL utiliza a sintaxe RDF. Esta linguagem explora tanto quanto possível o modelo primitivo do RDF-S. Deste modo a linguagem OIL permite tratar as suas ontologias como extensões do RDF e do RDF-S, disponibilizando-as não só em aplicações OIL, mas também em aplicações baseadas em RDF [11].

2.3. DAML

Oficialmente o programa DARPA *Agent Markup Language* (DAML) teve inicio em Agosto de 2000. A DAML foi criada com o objectivo de desenvolver uma linguagem e ferramentas de modo a facilitar o conceito da Web Semântica [12].

Em Agosto de 2000, surge a linguagem DAML-ONT, caracterizada como uma linguagem simples para expressar de forma mais sofisticada a definição de classes

RDF que o RDF-S. DAML é uma extensão do RDF-S permitindo adicionar semântica aos dados [13]. Mais tarde, o grupo que concebeu a DAML agrupa esforços com a linguagem OIL, resultando na criação da linguagem DAML+OIL. Esta ontologia foi criada com o objectivo de expressar classificações mais sofisticadas e propriedades de recursos, do que permite o RDF-S [12].

Na Figura 1 é ilustrado a linguagem DAML+OIL como extensão ao RDF-S:

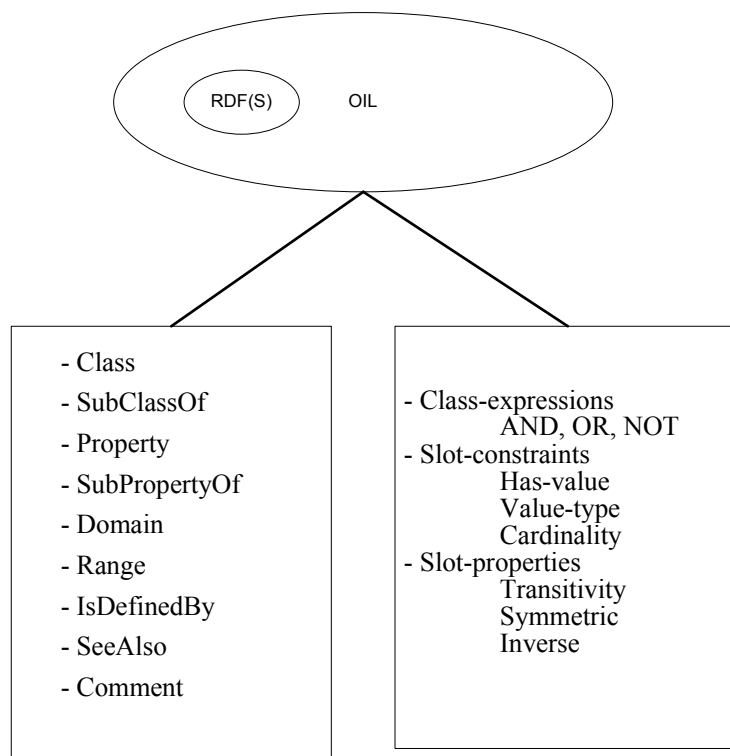


Figura 1. A linguagem DAML+OIL como extensão ao RDF-S (adaptado de [14]).

2.4. OWL

A *Web Ontology Language* (OWL) é uma linguagem semântica de anotação, com o objectivo de publicar e partilhar ontologias na *World Wide Web*. A OWL é desenvolvida como uma extensão de vocabulário RDF, e é proveniente da linguagem de ontologias DAML+OIL [11].

A linguagem OWL foi criada com o objectivo de descrever classes e relações entre documentos e aplicações Web [15].

Como qualquer conjunto de triplos RDF, pode ser representada de diversas formas sintácticas [16].

Porque é uma extensão do RDF, uma ontologia OWL pode incluir conteúdo RDF arbitrário, que é tratado de forma consistente com o seu tratamento pelo RDF. A OWL atribui significado especial a determinados triplos RDF.

Atendendo às principais características que expressam as linguagens de ontologias da Web Semântica, apresentadas acima, e considerando a simplicidade da estrutura hierárquica dos conceitos e relações representado nos IPTC-SC, o RDF-S foi a seleccionada para utilizar no âmbito do projecto OmniPaper. Esta escolha foi baseada no facto de a estrutura dos IPTC-SC apresentada ser tão simples que o uso de uma linguagem mais expressiva não traria benefícios adicionais.

3. IMPLEMENTAÇÃO

O trabalho de investigação efectuado na implementação do protótipo RDF desenvolvido no âmbito do projecto OmniPaper teve como objectivo a realização das seguintes tarefas (1) definição de um perfil de aplicação com todos os elementos de metadados necessários à descrição dos artigos de notícias [17]; (2) definição de uma base de metadados que armazena a meta-informação dos artigos de notícias [18]; (3) definição de uma linguagem de ontologias para representar os conceitos hierárquicos apresentados no IPTC-SC. O desenvolvimento destas tarefas conduziu à implementação de dois protótipos RDF. No primeiro foram executados os dois primeiros passos. O segundo protótipo inclui o terceiro passo, que é implementado com o objectivo de introduzir valor acrescentado ao primeiro protótipo RDF desenvolvido.

Na implementação do primeiro protótipo RDF, procedeu-se à catalogação dos artigos de notícias digitais de acordo com os vocabulários normalizados de metadados [18] de modo a facilitar a pesquisa sobre a camada de metadados definidos em RDF/XML, e armazenados na base de metadados. Este primeiro protótipo RDF foi desenvolvido na camada *Local Knowledge Layer* e permite um acesso estruturado e uniforme a todo um conjunto de artigos de notícias disponibilizados pelas empresas parceiras (a que denominamos por arquivos locais).

A camada *Overall Knowledge Layer* foi desenvolvida na implementação do segundo protótipo RDF, adicionando mecanismos de navegação semântica sobre a árvore de conceitos, representada na estrutura hierárquica dos IPTC-SC.

O vocabulário controlado que constitui os IPTC-SC é constituído por uma árvore hierárquica de três níveis, que descrevem o conteúdo de um conjunto de termos. Os tópicos apresentados ao nível dos termos *Subject* contêm uma descrição editorial do conteúdo das notícias; ao nível do *SubjectMatter* contêm uma descrição a um nível semântico mais preciso, e finalmente o *SubjectDetail* contêm um nível semântico mais específico do conteúdo das notícias.

Para representar os IPTC-SC, várias linguagens foram analisadas e estudadas de forma a seleccionar a que melhor se adapta à sua representação hierárquica de conceitos. No entanto os IPTC-SC sob o ponto de vista semântico não são assim tão ricos. Deste modo, atendendo à sua simplicidade, uma vez que apenas era necessário definir os seus conceitos hierárquicos, verificou-se que o RDF-S era a linguagem de representação mais adequada para descrever a estrutura hierárquica representada nos IPTC-SC.

Após a descrição dos IPTC-SC através da linguagem de representação RDF-S, procedeu-se ao seu armazenamento numa base de metadados. A ligação aos elementos apresentados na árvore hierárquica dos IPTC-SC é efectuada através do elemento de metadados "*dc:subject*". Deste modo, na definição do perfil de aplicação, é indicado que o "*rdfs:range*" do elemento de metadados "*dc:subject*" são os IPTC-SC.

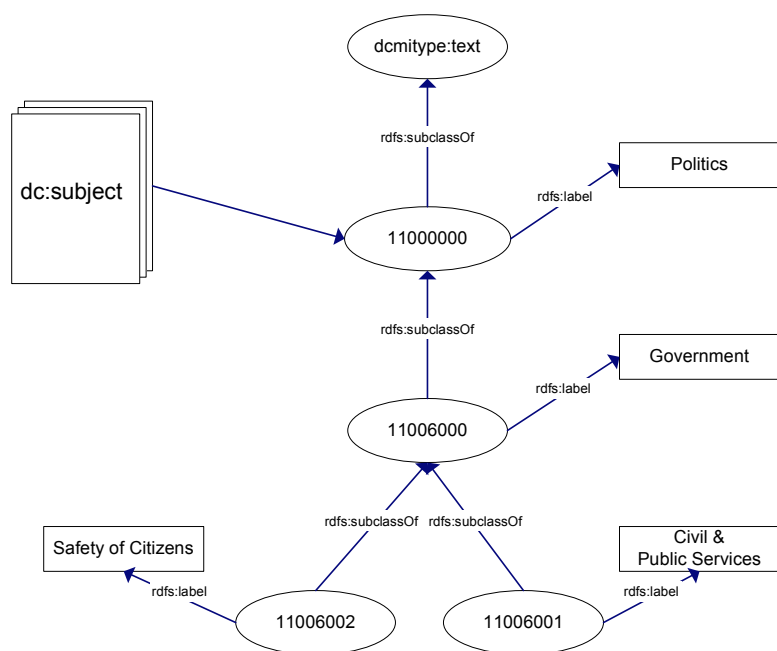


Figura 2. Exemplo do termo '*Politics*' modelado em RDF-S

O protótipo para a manipulação da camada RDF foi realizado utilizando o RDF Gateway [19]. Esta é uma ferramenta que conjuga os poderes do servidor HTTP com o sistema de Gestão de bases de dados nativas RDF.

O RDF Gateway é simultaneamente um cliente e um servidor Web constituído por uma base de dados nativa RDF para administrar a informação. O acesso aos dados é feito via HTTP.

4. CONCLUSÃO

Este artigo pretende introduzir a representação de uma ontologia pequena, utilizando o RDF-S. Numa primeira fase, procedeu-se ao estudo e análise de um conjunto de ontologias expressas na Web Semântica, seleccionando aquela que melhor se adaptava às nossas necessidades de descrição. O estudo elaborado sobre ontologias, apresentado acima, permitiu concluir que o RDF-S era a linguagem de representação adequada à descrição da estrutura hierárquica dos IPTC-SC. Em particular, foi também analisada a origem e a noção do termo ontologia verificando-se várias interpretações deste termo, tendo sido apresentado o que melhor se adequava à nossa perspectiva.

De modo a introduzir valor acrescentado ao primeiro protótipo RDF, desenvolvido no âmbito do projecto OmniPaper, foram adicionados mecanismos de navegação sobre conceitos representados na árvore hierárquica dos IPTC-SC.

O protótipo RDF será o primeiro estudo apresentado sobre o contributo dos metadados e da tecnologia RDF na descrição e navegação sobre os recursos de informação digital, desenvolvido no âmbito do projecto de OmniPaper. É também importante referenciar, o desempenho da linguagem definida para representar os IPTC-SC e os seus contributos para adicionar ao protótipo uma camada de navegação semântica sobre os recursos de informação.

Numa fase seguinte, foi incluída na camada de navegação do protótipo uma versão RDF (versão 1.6) de um *thesaurus* léxico – o WordNet. A ligação à base de metadados foi efectuada através do elemento de metadados “*omni:key-list*”. A adição do WordNet permite, não só introduzir uma granularidade mais fina na navegação de conceitos como a expansão (automática ou manual) de *queries* de pesquisa [20].

Na próxima fase pretende-se implementar o processo automático de extracção e armazenamento de *links*, de artigos de notícias e de artigos relacionados.

REFERÊNCIAS

1. Lassila, O.R.S., Ralph, *Resource Description Framework (RDF) Model and Syntax Specification*. 1999.
From <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>.
2. Ontolingua, *Knowledge Systems Laboratory Stanford University. Ontolingua*.
From <http://www.ksl.stanford.edu/software/ontolingua/>
3. *IPTC Subject Codes*. 2003.
From <http://www.nitf.org/site/nitf-documentation/subject-codes.html>.
4. *NetLing - Dictionary*.
From <http://www.linktotal.net/tp.htm?http://www.onelook.com/>
5. Miller, L., *Ontologies and Metadata*.
From <http://ilrt.org/discovery/2000/11/lux/>

6. *Ontology (computer science)*.
From [http://en.wikipedia.org/wiki/Ontology_\(computer_science\)](http://en.wikipedia.org/wiki/Ontology_(computer_science))
7. Mika, P., *Applied Ontology-based Knowledge Management: A Report on the State-of-the-Art*, in Master. 2002, Vrije Universiteit: Amsterdam.
From <http://www.cs.vu.nl/~pmika/thesis/pmika-thesis-full.doc>
8. *RDF Vocabulary Description Language 1.0: RDF Schema*. 10 February 2004.
From <http://www.w3.org/TR/rdf-schema/>
9. *W3C World Wide Web Consortium*.
From <http://www.w3c.org/>
10. Guha, D.B.R.V., *RDF Vocabulary Description Language 1.0: RDF Schema*.
From <http://www.w3.org/TR/2002/WD-rdf-schema-20021112/>
11. Horrocks, F.v.H.I., *Questions and answers on OIL: the Ontology Inference Layer for the Semantic Web*.
From <http://www.ontoknowledge.org/oil/oil-faq.html>
12. Ouellet, U.O.R., *DAML Reference*. 01 May 2002.
From <http://www.xml.com/lpt/a/2002/05/01/damref.html>
13. Garshol, L.M., *Topic maps, RDF, DAML, OIL. A comparison*.
From <http://www.ontopia.net/topicmaps/materials/tmrdfoildaml.htm>
14. Stuckenschmidt, H., *DAML+OIL Overview*.
15. McGuinness, M.K.S.D.L., *Web Ontology Language (OWL) Guide Version 1.0*. 4 November 2002.
From <http://www.w3.org/TR/2002/WD-owl-guide-20021104/>
16. Beckett, D., *Coments to Journal archives*. Journalblog, 2003.
17. Teresa Pereira, A.A.B., Tomoko Yaginuma. *Perfil de Aplicação e Esquema RDF dos Elementos de Metadados do Projecto Omnipaper*. in *CLME'2003 - 3º Congresso Luso-Moçambicano de Engenharia*. 2003. Maputo, Moçambique.
18. Tomoko Yaginuma, T.P., Ana Alice Baptista. *Metadata Elements for Digital News Articles in the Omnipaper Project*. in *ELPUB 2003 - 7th ICC/IFIP International Conference on Electronic Publishing*. June 2003. University of Minho - Guimarães, Portugal.
19. *RDF Gateway. A plataafform for the semantic Web*.
From <http://www.intellidimension.com/>

20. Baptista, A.A., *Searching and Browsing Using RDF-Encoded Metadata - The Case of OmniPaper*. Canadian Journal of Communication, 2004.

(*). Este artigo foi submetido e apresentado na Conferência ELPUB 2004 no Brasil em versão inglesa.