A representação do assunto por estrutura profunda[1]

Maria Aparecida Lourenço Santana (UFMG)
Eduardo Wense Dias (UFMG)

Resumo: Ao estudar a representação do assunto de recursos eletrônicos sob o desafio do problema da construção de hierarquias semânticas para interpretação tanto por pessoas como para processamento por máquinas, encontrou-se na Teoria Geral da Indexação de Assuntos, o esquema teórico da estrutura profunda. Aliou-se esta estrutura aos procedimentos de análise e projeto orientado ao objeto consolidada nos esquemas XML/ RDF que são fundamentados no modelo de representação em tripla. Observou-se que a representação em tripla pareceu estruturalmente adequada para a descrição do assunto por estrutura profunda, o que proporcionou a concepção do modelo DEPAm-OR. Recursos eletrônicos indexados, disponibilizados na Internet foram representados segundo o modelo proposto. A descrição do assunto usando o modelo DEPAm-OR foi feita usando ferramenta de árvore hiperbólica. Concluiu-se pela recomendação do modelo, observando que sua aplicação permitia grande explicitação semântica, determinada pela estrutura profunda.

Palavras-chave: Análise de assunto. Análise orientada ao objeto. Estrutura profunda. Linguagens de indexação.

Abstract: When studying the representation of the subject of electronic resources under the challenge problem of the semantic hierarchies' construction able to interpretation by people or machine processing, it was found in General Theory of the Indexation of Subjects, the theoretical outline of the deep structure. It was formed an alliance between this structure with the analysis procedures and project object-oriented consolidated in the outlines XML / RDF that are based in triple representation model. It was observed that the triple representation seemed appropriated to deep structure subject description, what provided the conception of DEPAm-OR model. Indexed electronic resources, made available in the Internet, were represented according to the proposed model. The subject description using DEPAm-OR model was made using hyperbolic tree software. It was recommended the model use, observing that its application allowed great semantic explicitness, mainly obtained because of deep structure.

Keywords: Subject analyses. Object-oriented analyses. Deep structure. Indexing language.

[1] Comunicação oral apresentada ao GT-02 - Organização e Representação do Conhecimento

1 INTRODUÇÃO

A busca por modelos de representação de informação para análise de assunto tem orientado muitas pesquisas na Ciência da informação. Desde muito cedo, com a invenção dos computadores, verificou-se intensos esforços para representar e dar acesso a conteúdos da classificação bibliográfica e da indexação de periódicos utilizando dispositivos computacionais. A década de 1970 foi especialmente marcada por estudos que uniam teorias da ciência da informação e tecnologias de computadores, como os sistemas PRECIS e POPSI registrados pela literatura (FUJITA, 1988), (RIVIER, 1992), (LANCASTER, 1993).

Ao observar as ferramentas computacionais percebem-se desenvolvimentos para representação de conteúdos de forma a permitir a adição sistemática de semântica aos recursos, com o objetivo de se obter maior relevância e precisão nos resultados de busca por informação. Os projetos orientados ao objeto, especialmente observados nesse estudo, permitem conceitualizar a realidade em categorias, ou seja, abstrações gerais que formam os objetos, então considerados os construtos básicos da análise. Por outro lado, na ciência da informação, a formulação do assunto na forma de palavras-chave ou da frase de indexação - ambas em linguagem natural, não pode ser considerada uma prática que permita a construção de estruturas que cooperem com explicitação semântica necessária à interpretação tanto por pessoas como por programas de computadores.

Tal verificação direcionou as pesquisas para a busca de um modelo de representação que permitisse a conjunção do modelo orientado ao objeto da computação com um modelo adequado à representação do assunto com a possibilidade de também especificar a sua estrutura semântica. Essa busca permitiu encontrar a estrutura profunda descrita dentro da Teoria Geral da Indexação de Assuntos, formulada na Índia, na década de 1970, descrita a seguir.

2 REVISÃO DE LITERATURA

Segundo o dicionário Houaiss (2001), estrutura profunda tem o mesmo sentido de estrutura subjacente, proveniente da área da gramática gerativa (LYONS, 19730), e é definida como “representação da frase em nível abstrato, na qual se estabelecem as relações semânticas básicas entre os itens lexicais, cuja ordem linear pode ser modificada com aplicação das transformações que forem necessárias, para derivar a estrutura superficial, mantendo as relações semânticas iniciais”.

Partiu de Ranganathan a teoria de que todas as linguagens de indexação eram estruturas de superfície de uma estrutura profunda. Bhattacharyya (1979, 1981) elaborou essa idéia na teoria geral da indexação de assunto. A teoria formulada opera sobre uma linguagem de indexação de assuntos (LIA). Nas palavras de Bhattacharyya:

> Uma LIA é uma linguagem artificial desenvolvida baseada em estruturas semânticas intrínsecas, elementos artificialmente postulados, e com estruturas sintáticas de proposição de assuntos (...) A estrutura de uma LIA específica deve ser suposta como sendo a estrutura de superfície de uma estrutura profunda de LIA´s. (BHATTACHARYYA, 1979, p. 24).

Sob a orientação de Ranganathan, Bhattacharyya enuncia a existência da estrutura profunda nas linguagens de indexação de assunto. A estrutura profunda é definida como "composta por constituintes elementares e regras para a formulação de expressões admissíveis, que são usadas para sumarizar em formulações indicativas sobre o que é o conteúdo de uma fonte de informação" (Bhattacharyya, 1981, p. 12).

Os constituintes elementares enunciados por Bhattacharyya são concebidos dentro da estrutura elementar das proposições de assunto, uma vez que o autor vê essas proposições formadas por três tipos de estrutura: semântica, elementar e sintática. A estrutura semântica é responsável pela compreensão e significação. A estrutura elementar é em geral formada por constituintes elementares formulados artificialmente para dar reconhecimento e significado de campo semântico aos substantivos. Finalmente as estruturas sintáticas seriam estruturas lineares, horizontais, das descrições de assunto. Portanto, uma frase se configuraria como uma seqüência de ocorrências dos constituintes elementares.

Devadason (1985a, 1985b) explica as categorias elementares como sendo formadas por (D) disciplina, (E) entidade, (P) propriedade, (A) ação e (m) modificador. A disciplina é a manifestação da categoria elementar para os campos de estudos convencionais, disciplinares. A entidade é a categoria elementar das manifestações conceituais das coisas. A categoria elementar propriedade inclui manifestações denotando os conceitos de atributo qualitativo ou quantitativo. A categoria elementar ação inclui manifestações que denotam o conceito de ação "fazendo". O elemento modificador refere-se a uma idéia usada para qualificação de uma manifestação sem perturbação do seu todo conceitual. Um modificador pode ser um especificador, qualificador, especializador ou diferenciador. Ele modifica a manifestação de qualquer uma das categorias elementares. Modificadores são definidos por Devadason como comuns, quando denotam forma, tempo, ambiente e lugar, ou, especiais,quando atuam sobre as manifestações de disciplina, entidade, propriedade ou ação.

As categorias elementares DEPAm estão inextricavelmente ligadas às categorias PMEST (Personalidade, Matéria, Energia, Espaço e Tempo). Foram elaboradas como uma evolução da estrutura dos sistemas de classificação da época. Foi demonstrado que nas classificações bibliográficas existia uma estrutura subjacente representada pelas categorias elementares (PARKHI, 1964), (BHATTACHARYYA, 1979, 1981), (BISWAS, SMITH, 1988), (LANGRIDE, 1989), (RIVIER, 1991). Sendo que, a lingüística estruturalista nascidas em Saussure influenciou o pensamento e desenvolvimentos de sistemas naquele tempo. O eixo sintático foi representado na ordem das categorias DEPAm e o paradigmático no agrupamento hierárquico das categorias elementares formando tabelas semânticas (SAUSSURE, 1981, DEVADASON, 1985a).

A Fig. 1 mostra o esquema da estrutura profunda de linguagem de indexação de assunto. No esquema Bhattacharyya apresenta o conjunto de manifestações elementares e modificadores, desenhando a forma como interagem os modificadores sobre qualquer das manifestações das categorias elementares.

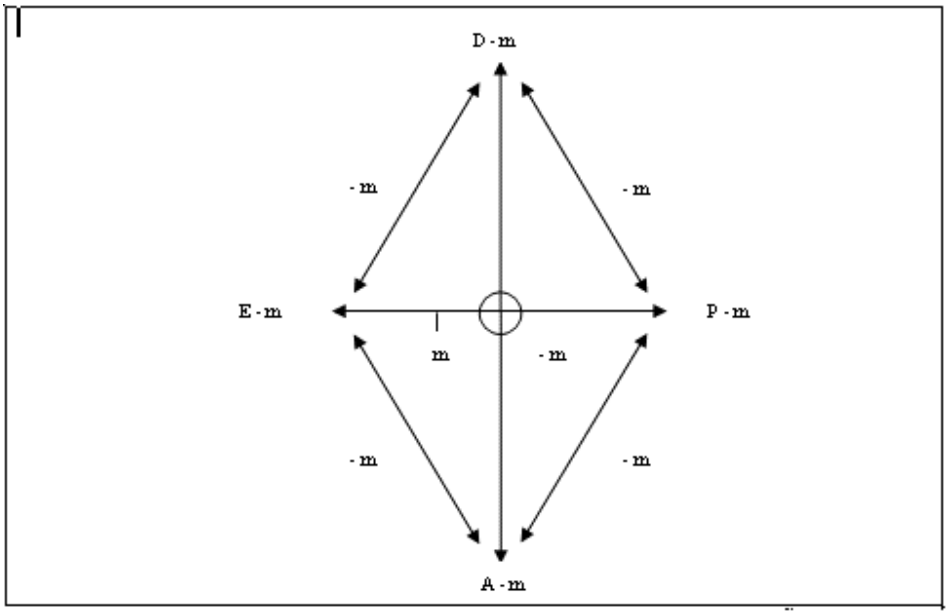

FIGURA 1: *Esquema da estrutura profunda de linguagem de indexação de assunto*
Fonte: Bhattacharyya, 1981, p.12.

Neste artigo apresenta-se a consolidação da estrutura profunda com a modelagem orientada ao objeto. Na análise conceitual orientada ao objeto levou-se em consideração a interseção de princípios de linguagens de indexação e a modelagem das estruturas de representação computacional (PARSAYE, 1989), (SHLAER, MELLOR, 1990). Esse fato permitiu vislumbrar a possibilidade de as estruturas computacionais serem capazes de operacionalizar a concepção teórica da estrutura profunda da linguagem de indexação de assunto de forma bem próxima à idealizada por Bhattacharyya.

Segundo Neelameghan (1992) é muito semelhante o processo de conceituar a realidade, como é feito nas linguagens de indexação e na modelagem orientada ao objeto. O autor parte do pressuposto de que a conceitualização estaria preocupada, basicamente, com elementos a serem representados, não *per si*, mas por categorias.

Heaney (1995), concluiu que a modelagem orientada ao objeto parece, em muitos aspectos ideal para implementação na área da ciência da informação. Para o autor a modelagem orientada ao objeto não é um sistema de base de dados, e não estaria, por isso, diretamente ligada à programação de computadores ou linguagens de programação. Seria uma ferramenta de modelagem conceitual para orientar o pensamento sobre os objetos. Para Heaney qualquer coisa é um objeto, possui atributos e um conhecido conjunto de operações das quais pode participar, possuindo também características que permitem sua associação, agregação ou agrupamento em classes.

Em Buckland (1991) e Alvarenga (2001) são elaboradas as idéias de informação como coisa e objeto respectivamente. Essas abordagens são importantes porque mostram uma percepção de orientação ao objeto dentro ciência da informação e da área de tratamento – sendo um dentre seus diversos processos.

No processo de abstração de tipos de dados, uma classe define ambos, a estrutura e o comportamento de tipos abstratos de dados. O paralelo da classe na estrutura profunda de linguagens de indexação é a categoria. Os objetos podem ser vistos como as manifestações de entidade, e as manifestações de disciplinas contidas na idéia de um domínio. Um domínio remete ao contexto de ocorrência das classes de objetos, e dá a idéia de especialização, não devendo ser entendido como área do conhecimento. Conforme descreve o Quadro 1 que resume as equivalências entre a orientação ao objeto e a estrutura profunda da teoria geral de indexação de assunto.

QUADRO 1: *Equivalência entre processos de modelagem conceitual orientada ao objeto e estrutura profunda de linguagem de indexação – DEPAm*

Equivalência de processos de conceitualização	
Orientação ao objeto	DEPAm
Classe	Categorias
Objetos	Entidades
Atributos	Propriedades
Comportamentos	Ações

No final da década de 1990, os desenvolvimentos de estruturas para metadados permitiram vislumbrar a operacionalização da proposta de junção da estrutura profunda com projeto orientado ao objeto (SWETLAND, 2000). O padrão descritivo de recursos RDF, sigla de *Resource Description Framework*, formado por um modelo para apresentação de metadados e uma linguagem para especificação da sintaxe definida no modelo, formaram um conjunto adequado de recursos computacionais para implementação da estrutura profunda. Isso é possível porque o RDF implementa o conceito de tripla (AHMED et a., 2001). Uma tripla é a expressão simples formada por três elementos: um recurso, sua propriedade e o valor da propriedade. A melhor expressão didática da tripla é através de diagramas com nós e arcos, chamados de grafos. A Fig. 2 mostra como é o esquema simplificado de tripla. Um recurso tem uma propriedade e essa propriedade é descrita pelo valor. Convencionalmente, o formato oval exprime um recurso, o arco (na verdade uma seta), o nome de uma propriedade ou relacionamento, e o quadrado exprime uma expressão literal.

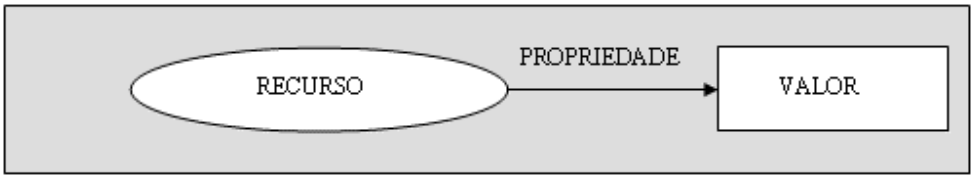

FIGURA 2: *Esquema de tripla expresso em grafos*
Fonte: Ahmed et al., 2001, p. 115, adaptado.

Uma sintaxe para o RDF pode ser construída através da representação em XML – *Extended Markup Language* formando o par RDF/XML. O modelo RDF e a sintaxe XML são especialmente importantes porque potencializam a metarrepresentação baseada na teoria geral da indexação de assunto e torna possível o casamento entre estruturas profundas de linguagem de indexação, estruturas de análise orientada ao objeto e permitem, ao final, o desenvolvimento prático do conjunto teórico pesquisado (DODEBEI, 2002), (SANTANA, 2005).

3 PROCEDIMENTOS METODOLÓGICOS

Percebeu-se que, nos casos estudados, os recursos digitais acessíveis pela *web*, apesar da tentativa de estruturação de seus contextos de representação, não chegavam a atingir o nível do assunto. Pode-se afirmar que o assunto não tem tido uma adequada representação semântica que permita seu processamento por computador, pois até que evoluam os programas para interpretação de sintagmas nominais, verbais, predicações e suficiente mapeamento conceitual, a descrição semântica de um recurso será tarefa humana.

Assim, para o estudo da representação semântica do assunto, em estruturas orientadas ao objeto, utilizaram-se os métodos de modelagem e de análise de conteúdo por trabalho intelectual. Segundo Kaplan, um modelo é "algo eminentemente digno de imitação, exemplar ou ideal" (KAPLAN, 1975, p. 265), durante a ordenação dos dados, ao passar de uma observação a outra, um modelo promove a noção de perseguição de uma idéia, enquanto se espera que algo aconteça.

O conjunto teórico formado pela estrutura profunda DEPAm e a modelagem orientada ao objeto, apresentados na revisão de literatura, foram os dois marcos que permitiram a elaboração do modelo DEPAm-OR (SANTANA, 2005), mostrado na Fig. 3.

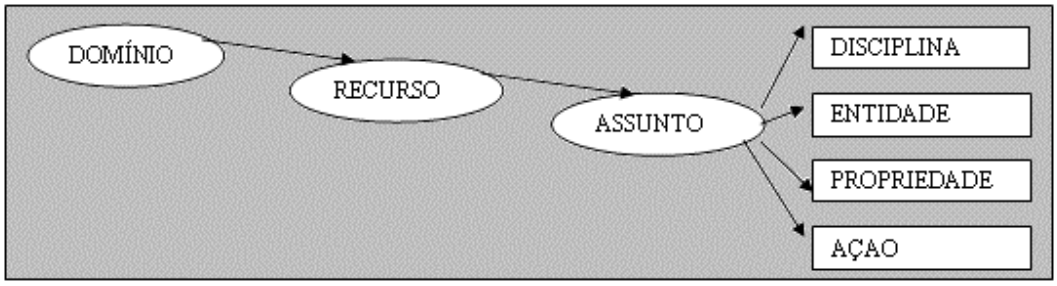

FIGURA 3: *Esquema explicativo do modelo de representação proposto – DEPAm-OR*
Fonte: Elaboração própria

No modelo articulam-se quatro tipos de estruturas de informação. O primeiro é o domínio, ou seja, o ambiente que mantém o recurso; o espaço de criação, manutenção e gestão do recurso. A segunda é o próprio recurso, com os atributos que lhe são peculiares, como nome, autor, título, URL, publicador, conteúdo, dentre outros. A terceira estrutura refere-se ao assunto, objeto do problema especificado na pesquisa. A quarta estrutura, representada por quatro retângulos (disciplina, entidade, propriedade, ação), que é a proposição para a representação de assuntos da estrutura profunda. Ressalta-se que o modelo DEPAm-OR tão somente acrescenta aos construtos teóricos encontrados as integrações que os tornam adequados aos novos padrões de desenvolvimento das descrições semânticas disponíveis.

Em resumo, o modelo DEPAm-OR representa uma estrutura de mapeamento conceitual adequada à representação do assunto, como resultado das atividades do indexador e potencial de interpretação, tanto por pessoas como por programas de computador.

Adicionalmente ao modelo proposto, a pesquisa recorreu ao método de análise de conteúdo para examinar padrões de conceitos existentes nas definições de atinência de recursos. Segundo Kim (1996), a análise de conteúdo compreende estudos quantitativos de recursos ou outras formas de comunicação que examinam freqüências/padrões de palavras, frases, conceitos, imagens, temas, caracteres, papeis, etc.".

Aplicou-se a análise de conteúdo a recursos indexados por serviços de informação, que utilizavam uma linguagem de indexação para representar o assunto de recursos eletrônicos. Foram elaborados estudos sobre representações que utilizavam sistemas como a Classificação Decimal de Dewey, a Classificação Decimal Universal e a Library of Congress Classification.

Os recursos eletrônicos estudados foram observados e sobre eles extraíram-se conceitos de maneira a preencher os requisitos dados pelo modelo DEPAm-OR proposto.

A coleta de dados utilizou-se de motores de busca para *web*. Dados gerais foram recolhidos sobre cada recurso como o endereço do site, o nome do sistema de informação fonte, a linguagem de indexação utilizada, a estratégia de busca utilizada para recuperação, o endereço URL, o nome do recurso, o código de classificação no qual o recurso foi indexado, as observações e o pesquisador que realizou a busca.

A análise dos dados passou pela elaboração da atinência dos recursos, pela reconstrução da classificação segundo a estrutura profunda DEPAm, pela formalização gráfica da representação do assunto, conforme o modelo DEPAm-OR e finalmente pela tradução do modelo para a linguagem XML/RDF. Implementou-se, assim, uma representação de assunto legível para pessoas e programas de computador.

A Fig. 4 mostra como ficou a representação de um caso (Caso 2, SANTANA, 2005) do modelo DEPAm-OR na estrutura no software livre HiperNavegador. As letras DEPAm indicam a classificação segundo o modelo DEPAm-OR.

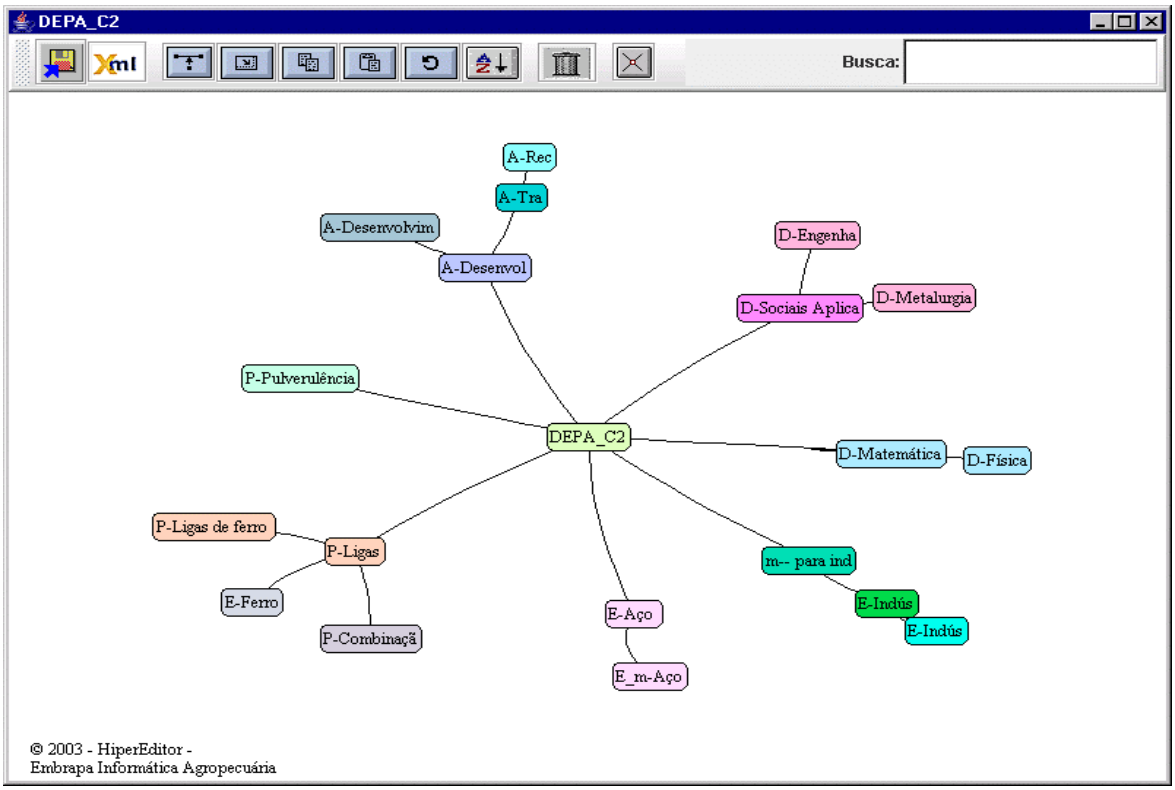

FIGURA 4: *Representação DEPAm-OR no HiperNavegador – Caso 2 Desenvolvimento de aços livres de intersticiais – IF – via recozimento em caixa para indústria automobilística – Classificado por CDU*

Não houve dificuldades para se elaborar as descrições, pelo formato de descrição de recursos RDF/XML. Estas se revelaram atividades simples de serem executadas. O formato de tripla, descrito na revisão teórica, mostrou-se eficiente e simples de usar, podendo ser considerado um instrumento a ser facilmente implementado pelo indexador.

As árvores hiperbólicas, construídas através do *software* HiperNavegador, facilitaram a implementação da representação da fatoração como objetos que se relacionam entre si. Esse *software* livre é disponibilizado pelo Ministério do Desenvolvimento, Indústria e Comércio em www.agrolivre.gov.br. A oportunidade de representar o modelo em software trouxe a garantia da plausibilidade da conjunção teórica dos dois objetos: estrutura profunda e modelagem orientada ao objeto.

Os nós da árvore hiperbólica implementam, de maneira muito tranqüila, a estrutura DEPAm. Mas a explicitação dos relacionamentos, conforme as possibilidades previstas pelo modelo E-R (entidade relacionamento), não foi garantida. Esse tipo de descrição deveria ter ocorrido nos arcos de ligação entre os nós. Para que não houvesse prejuízo do aspecto teórico, explicitou-se a categoria junto com o nome do atributo, que são as letras que antecedem cada nome nos nós da árvore.

4 RESULTADOS E CONCLUSÕES

O modelo pôde ser considerado fundamental para se verificar a viabilidade da melhor representação temática dos recursos estudados por esta pesquisa. Pela sua simplicidade e facilidade de implementar, o modelo proposto poderá contribuir consideravelmente para os processos de indexação temática legíveis por pessoas e por computadores.

Ao analisar os dados, concluiu-se que as representações que utilizam os sistemas de classificação CDD, CDU, LCSH e tesauros deixaram perder elementos supostamente coletados durante a determinação da atinência e úteis à representação pelo modelo DEPAm. Não se encontraram as expressões de ação. Para adequar as representações encontradas ao modelo de estrutura profunda, foi preciso voltar ao recurso e recuperar a expressão da ação, a partir do título, do resumo ou de uma leitura rápida de todo o recurso.

A representação do assunto não encontrou problemas quanto às técnicas de fatoração e descrição com o uso da estrutura DEPAm. Dado que as estruturas DEPAm auto-organizam os conceitos e os agrupam em categorias, tanto pessoas quanto programas de computadores estariam potencialmente habilitados a afirmar que um conceito ou subcategoria, dentro de outra categoria, na estrutura DEPAm, pertencem à categoria hierarquicamente superior, o que faz gerar informação nova, por inferência entre níveis hierárquicos.

Também se confirmou que a representação do assunto fatorado e modelado por estrutura profunda é passível de ser descrita por modelagem orientada ao objeto, adequada à representação e à recuperação pelo padrão RDF/XML, como especificado pelo *World Wide Web Consortium-W3Con* para descrição de conteúdos na *web*.

Conclui-se que o modelo DEPAm-OR é uma ferramenta de uso fortemente desejável para melhorar problemas de estruturação e representação do assunto dos recursos eletrônicos, com a finalidade de construir campos semânticos, de semântica profunda, na *web*, e operacionalizar facilidades de busca e acesso a recursos eletrônicos a partir de especificações da sua semântica profunda.

O estudo do modelo DEPAm-OR para representação de recursos eletrônicos fundamentados pelos conhecimentos sólidos e validados na Ciência da Informação evidenciou o fato de existir muito conhecimento a ser aproveitado para auxiliar as pesquisas sobre a representação de informações no ambiente virtual. A Ciência da Informação possui a competência consolidada pelos anos de desenvolvimento teórico e prático, seja para representar um recurso em relação a uma coleção, seja para representar tópicos dentro de um recurso.

REFERÊNCIAS

AHMED, Kal et al. *Professional XML Meta Data*. Chicago: Wrox, 2001. 568p.

ALVARENGA, Lídia. A teoria do conceito revisitada em conexão com ontologias e metadados no contexto de bibliotecas tradicionais e digitais. *Datagramazero*, v. 2, n. 6, dez. 2001.

BHATTACHARYYA. G. Some significant results of current classification research in India. *International Forum on Information an Documentation*, v. 6, n. 1, p. 11-18, Jan. 1981.

_____. POPSI Its fundamentals and procedure based on a general theory of subject indexing languages. *Library Science*, v.16, n.1, p.1-34, Mar. 1979.

BISWAS, Subal C. SMITH, Fred. Computerized deep structure indexing system: a critical appraisal. *International Classification*, v.15, n.1, p.2-12, 1988.

BUCKLAND, Michael K. Information as thing. *JASIS*, v. 42, n. 5, p. 351-360, 1991.

DEVADASON, F. J. Computerization of deep structure based indexes. *International Classification*, v.12, n.2, p.87-94, 1985a.

_____. On line construction of alphabetic classaurus: a vocabulary control and indexing tool. Information *Processing and Management*, v.21, n.1, p.11-26, 1985b.

DODEBEI, Vera Lúcia Doyle. *Tesauro*: linguagem de representação da memória documentária. Niterói: Intertexto; Rio de Janeiro: Interciência, 2002. 119p.

HIPEREDITOR-HIPERNAVEGADOR. www.agrolivre.gov.br. 2004. [software livre].

HOUAISS, Antônio. VILLAR, Mauro S. *Dicionário Houaiss da Língua Portuguesa.* Rio de Janeiro, Objetiva, 2001.

FUJITA, Mariângela. Sistemas de Indexação: PRECIS I: perspectiva histórica e técnica do seu desenvolvimento e aplicação. *Rev. Bras. de Bibl. e Doc.*, v.21, n.1/2, p. 21-45, jan./jun. 1988.

HEANEY, Michael. Object-oriented cataloguing. *Information Tecnology and Libraries*, p. 135-153, september, 1995.

KAPLAN, Abraham. *A conduta na pesquisa*: metodologia para as ciências do comportamento. 2. ed. São Paulo: EPU, Edusp, 1975. 440 p.

KIM, Mary T. Research record. Journal of Education for Library & Information Science, n. 37, p. 376-382, 1996 apud POWELL, Ronald R. Recent trends in research: a methodological essay. *Library & Information Science Research*, v. 21, n.1, p.91-119, 1999.

LANCASTER, F.W. *Indexação e resumos*: teoria e prática. Brasília: Briquet de Lemos/Livros. 1993.

LANGRIDGE, D. W. *Subject analysis:* principles and procedures. London: Bowker-Saur, 1989. 146p.

LYONS, JOHN. *As idéias de Chomsky*. São Paulo: Cultrix.1973. 121 p.

NEELAMEGHAN, A. Application of Ranganathan's general theory of knowledge classification in designing specialized databases. *Libri*, v. 42, n. 3, p. 202-226, 1992.

PARSAYE, Kamran; CHIGNELL, Mark; KHOSHAFIAN, Setrag; WONG, Harry. *Intelligent databases:* object-oriented, deductive hypermedia technologies. New York: Willey, 1989. 479p.

PARKHI, R.S. *Decimal classification and colon classification in perspective*. London: Asia Publishing House, 1964.

RIVIER, Alexis. Construção de linguagens de indexação: aspectos teóricos. *Rev. Esc. Bibl. da UFMG*, v.21, n.1, p.56-99, jan./jun. 1992.

SANTANA, Maria A. L. A indexação temática de recursos fundamentada por estrutura profunda e abordagem objeto-relacionamento [manuscrito]. Belo Horizonte: UFMG, Escola de Ciência da Informação. 2005. [dissertação].

SAUSSURE, Ferdinand. *Curso de lingüística geral*. São Paulo: Cultrix, 1981.

SHERA, Jesse H. Pattern, structure, and conceptualization in classification. In: International study conference on classification for information retrieval. *Proceedings.* London: Aslib, 1957, p. 15-27.

SHLAER, Sally; MELLOR, Stephen J. *Análise de sistemas orientada para objetos*. São

SWETLAND, Anne J. G. *Introduction to metadata: setting the stage*. Disponível em < http://www.slis.kent.edu/~mzeng/metadata/Gilland.pdf>. 07/05/2000.
Paulo: McGraw-Hill, 1990. 178 p.