DSPACE VERSÃO 1.4: UMA ANÁLISE DAS FACILIDADES RELACIONADAS AO ASSUNTO

Milton Shintaku - IBICT/UnB
shintaku@ibict.br
Marisa Bräscher – UnB
marisab@unb.br

Resumo

A versão 1.4 do DSpace trouxe, entre outras novidades, a possibilidade de recuperação dos documentos depositados por assunto, essa facilidade está relacionada com o preenchimento do metadado descritivo palavra chave. Este trabalho pretende analisar a utilização do vocabulário controlado, implementado na mesma versão, para facilitar o preenchimento do metadado palavra-chave, evitando variações como a de número e grafia e a busca por assunto em alguns repositórios que possivelmente ainda não implementaram um vocabulário controlado. Uma análise do vocabulário controlado fornecido pelo DSpace, em norueguês, revela que, em relação à área da Ciência da Informação esse vocabulário controlado apresenta-se deficiente, pois hierarquicamente possui apenas dois níveis e com apenas seis termos, não representado os assuntos cobertos pela área. Comparações feitas dos termos do vocabulário controlado do DSpace com os termos dos artigos dos últimos três anos da revista "Ciência da Informação", publicada pelo IBICT, demonstra que há pouca relação entre os dois, o que indica uma certa deficiência e carece de maior estudo. Os repositórios que migraram ou foram criados na versão 1.4 do DSpace possuem a recuperação de documentos por assunto, baseados nas palavras-chaves inseridas durante o processo de submissão, esta facilidade evidencia os problemas de variações terminológicas que provocam a um estudo melhor desta facilidade e de como melhor descrever um documento para facilitar a recuperação. Dentre os metadados descritivos, a palavra-chave é um dos únicos que matêm certa independência entre o conteúdo descrito no documento e o metadado. Título e autor, por exemplo, devem ser os mesmos no documento e no metadado. Pode-se utilizar a palavra-chave para facilitar a organização e recuperação dos documentos pelo assunto.

Palavras-chaves

Repositórios institucionais; DSpace; Metadados; Dublin Core; Palavras-chaves; Vocabulário Controlado

Abstract

DSpace version 1.4 begins a new possibility to use a controlled vocabulary to fill keywords in subject form and retrieve item by subject. This paper intent to analyze the suggested controlled vocabulary in DSpace and recommended by Dublin Core Metadata Initiative for keyword and retrieve documents by subject option. The analysis of suggested controlled vocabulary in DSpace, in Norwegian, for Information Science have only three levels and six terms, impossible to cover the subjects researched by this knowledge area. Comparing the terms of controlled vocabularies recommended by Dublin Core Metadata Initiative and the terms most frequently used by journal "Ciência

da Informação" – last three years, demonstrate a few coincidences. An analysis in repositories created or migrated to DSpace version 1.4 in the browse by subject demonstrate some problems with terminology. Variations in terms, like use of uppercase or plural, retrieve different documents even for a same term. The metadata keyword has an independent rule to filling, title and authors needs to be the same in the document and metadata, but is not proper to keyword, this metadata could be used to retrieve and organize the information. This is a preliminary analysis of the tool.

keywords

Institucional repository; DSpace; Metadata; Dublin Core; Keyword; Controled vocabulary

Introdução

O repositório institucional é um sistema informatizado que possui como unidade de armazenamento o conjunto dos objetos digitais e seus metadados, denominado de Item, criado como uma opção para divulgação da produção científica de uma instituição (Linch, 2003). Utilizando o auto-arquivamento, permite que o próprio autor submeta o trabalho, desde o preenchimento dos metadados até o depósito do arquivo (objeto digital). Essa liberdade de procedimentos, porém, gera algumas dificuldades em relação à qualidade dos metadados fornecidos pelo autor. Para evitar grandes variações e padronizar entradas, a utilização de vocabulário controlado para metadados permite restringir as possibilidades de preenchimento do assunto.

O metadado palavra-chave, entre outros pertencentes ao esquema de metadados Dublin Core, é um caso que se encaixa na categoria dos metadados para os quais é recomendado o uso de vocabulário controlado (Borbinha,2000). Os metadados, além de fornecer informações sobre o objeto digital, servem de ponto de recuperação a esses objetos. Para documentos textuais, porém, na maioria dos casos, há uma coincidência entre o conteúdo do documento e o metadado. Título, autores e resumo, por exemplo, são os mesmos no documento e nos metadados. Para repositórios baseados no DSpace que fazem uso da indexação de texto completo, essa repetição não ajuda em criar formas de recuperação do documento mais otimizados. O metadado palavra-chave possui certa independência entre o metadado e o documento. Esse metadado cumpre a função de organizar, classificar e hierarquizar os documentos no repositório e facilitar a recuperação por assunto, agrupando os documentos que possuem relação de assunto.

Uma análise terminológica preliminar nas palavras-chaves em artigos em repositórios (Repositorium[1], BDJUR[2] e MIT[3]) e em periódico (Ciência da Informação) demonstra que problemas de variações nos termos podem dificultar o acesso aos documentos. Problemas simples como o de grafia, que provocam a recuperação de documentos diferentes, revelam a necessidade de padronização e de um estudo mais profundo sobre vocabulários controlados que auxiliem na qualidade dos metadados para a organização e recuperação da informação.

Neste trabalho serão analisados dois aspectos relativos ao metadado assunto. Primeiramente, foi efetivada uma análise dos termos fornecidos por alguns vocabulários

[1] Repositório institucional da Universidade do Minho, endereço eletrônico: http://repositorium.sdum.uminho.pt/

[2] Biblioteca Digital Jurídica do Supremo Tribunal de Justiça, endereço eletrônico: http://bdjur.stj.gov.br/dspace

[3] Repositório do Instituto de Tecnologia de Massachusetts, endereço eletrônico: http://dspace.mit.edu/

controlados. O utilizado pelo DSpace versão 1.4 para a área de Ciência da Informação e os sugeridos pelo Dublin Core, comparando-os com os termos extraídos das palavras-chaves da revista "Ciência da Informação". Foi realizada também uma análise do preenchimento do metadado assunto nos repositórios Repositorium, BDJUR e MIT, para identificar aspectos relativos à organização de assuntos.

Referencial teórico

Em 1999, com a consolidação dos conceitos de arquivos abertos, a comunicação científica iniciou uma nova maneira de divulgação científica (OAI, 1999). A diponibilização na web da produção científica permite, entre outras coisas, a possibilidade de acesso pela comunidade científica ou leiga, incrementando significamente a abrangência das informações. A Internet, com a democratização de acesso às informações, necessita de padronização para que iniciativas isoladas possam ter a compatibilidade necessária para a interoperatividade, isso significa definir um conjunto mínimo de metadados (Dublin Core), formato do arquivo usado no intercâmbio de informações (XML) e programas utilizados na comunicação entre as iniciativas (protocolos, conversores, ferramentas para validar metadados etc.) (Triska e Café, 2001). Os requerimentos necessários para implementar os arquivos abertos podem ser feitos de várias formas e utilizandas diversas facilidades, que permitem a flexibilidade necessária aos diversos tipos de necessidades. Duas grandes iniciativas dos arquivos abertos são as publicações digitais e os repositórios institucionais, que implementam o conceitos dos arquivos abertos e promovem serviços diferenciados e confiáveis na web.

Os repositórios institucionais são estruturas informatizadas que possibilitam a disponibilização na web da produção científica de uma instituição. Inicialmente implementados para documentos pós e pré-prints (Lynch, 2003) e depois extendidos para outros tipos de documentos. Implementam os conceitos de interoperabilidade: auto-arquivamento, tipos de submissão e provedores de dados, entre outros, definidos pelos arquivos abertos e fornecem facilidades de recuperação dos documentos arquivados.

O DSpace é um software mantido pela Instituto Tecnológico de Massachussets – MIT e pela Hewllet Packard – HP desenvolvido para facilitar a criação de repositórios institucionais. Baseado no conceito de software livre de código aberto permite a utilização sem ônus e a possibilidade de alteração dos programas. Utilizado por várias instituições, de maioria acadêmica, possui atualmente uma comunidade que desenvolve facilidades e que provê solução para os novos desafios encontrados. Como o DSpace organiza-se em comunidades e coleções, um item, a princípio, deve pertencer a uma coleção. Dessa forma, podem-se listar todos os itens pertencentes a uma coleção. Elencar os itens pelo assunto permite sua representação em várias listas, conforme os inter-relacionamentos de conteúdo. Caso sejam utilizadas estruturas hierárquicas para as palavras-chaves, pode-se também hierarquizar as listas de Itens por assunto.

Para descrever os Itens existentes no repositório, o DSpace utiliza com o padrão o Dublin Core (DSPACE). Pode-se fazer uso de outros esquemas, facilidade também apresentada na versão 1.4, mas isso requer alterações nos programas e um grande esforço em programação para adequar os formulários de submissão de documentos e para responder à coleta automática de metadados (harvesting).

O vocabulário controlado apresenta-se no DSpace como uma estrutura hierárquica de termos. Definida como um arquivo XML, não possui interface para alterações. Para

fazer qualquer manutenção precisa-se do profissional de informática. Por ter a forma árvore, ao selecionarmos um elemento hierarquicamente inferior, todos os acima hierarquicamente serão selecionados. O DSpace provê um vocabulários controlado em norueguês com diversas áreas, o fragmento para a Ciência da Informação apresenta a hierarquia e termos fornecidos para a área, com três níveis e seis termos, conforme ilustrado na figura 1.

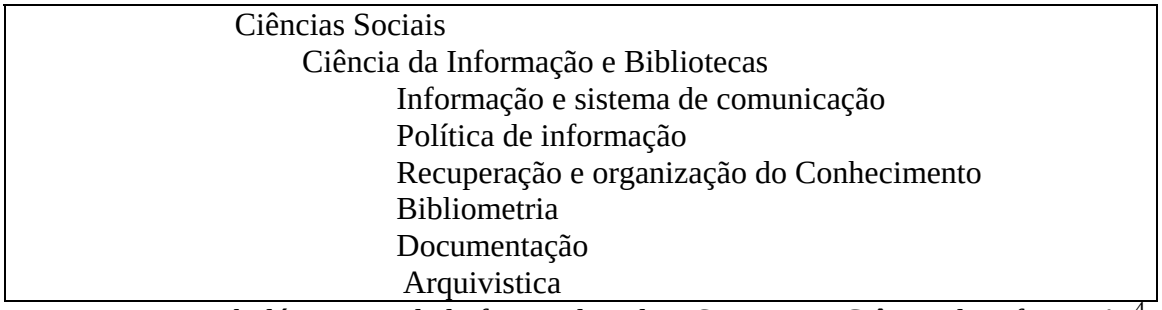

Figura 1 – Vocabulário controlado fornecido pelo DSpace para Ciência da Informação[4]

Para o vocabulário controlado da figura 1, caso seja selecionado o termo "Documentação", os temos "Ciência da Informação e Bibliotecas" e "Ciências Sociais" também serão selecionados. Desta forma, o documento aparecerá nas três listas por assunto. Isso permite generalizar ou refinar listas de documentos por assunto.
As palavras-chaves são termos e estes podem ser uma palavra (termo simples), grupo de palavras (termo composto), sintagma, símbolo ou fórmula que designam um conceito de uma área específica (Pavel e Nolet, 2002). Essa definição pode ser extendida para conter as abreviaturas (Faulstich e Abreu, 2003). Os termos podem sofrer variações como: a) formas utilizadas geograficamente, exemplo: "sistema operacional de rede" no Brasil e "sistema operativo de rede", ou o clássico: "mandioca", "macaxeira" e "aipim" nas diferentes regiões do Brasil; b) formas de tradução diferenciadas, exemplo: "acesso aberto" e "acesso livre" os dois termos possuem o mesmo conceito e são traduções diferentes para o português do original "open access"; c) mudança na ordem, exemplo: "lista de endereços eletrônicos" e "lista eletrônica de endereços" - os dois termos possuem o mesmo conceito, mas apresentam variação na ordem das palavras; d) apagamento de temos intermediários, por exemplo: "repositório digital institucional" e "repositório institucional" - referem-se ao mesmo conceito, o apagamento da palavra digital não altera o significado.

Metodologia

A análise das palavras-chaves incluídas nos artigos disponíveis em ferramentas de divulgação digital na web, fornece um panorama sobre os assuntos abordados em uma área. Nesse trabalho a análise focou apenas na perspectiva terminológica, quais os impactos das palavras-chaves inseridas pelo autor na organização dos documentos em um repositório e nas possibilidades de recuperação. Os repositórios e publicações digitais são instrumentos dinâmicos e a análise efetuada é um corte momentâneo - maio de 2007 - que pode ser modificado pelas necessidades das instituições e usuários e evolução da tecnologia.

[4] Tradução dos autores

Os repositórios analisados foram o Repoistorium - https://repositorium.sdum.uminho.pt/ da Universidade do Minho que possui 9673 termos para pesquisa como assunto, para um pouco mais de 5.500 documentos depositados, um número grande se pensarmos que essa opção deve ser utilizada como ponto de recuperação de documentos que possuem o mesmo assunto. Termos como endereços IP (Internet protocol) são encontrados na opção de recuperação. O repositório do MIT - http://dspace.mit.edu/browse-subject possui 16652 termos em assunto. Uma grande quantidade de números é encontrada, que dificilmente podem ser enquadrados como assuntos. Termos como: "Z7164.F5 HG173" são encontrados e podem ser reconhecidos como significativos em uma área, mas não inseridos contextualmente como assunto. A BDJUR – Biblioteca Digital Jurídica do Supremo Tribunal de Justiça possui 6184 termos em busca por assunto, para um pouco mais de 5000 documentos, por ser uma instituição não acadêmica e possuir um direcionamento para os documentos depositados, os termos na recuperação por assunto refletem esse característica.

Para ter um panorama das palavras-chaves utilizadas pelos artigos da área da Ciência da Informação, escolhido o periódico "Ciência da Informação" período de 2004 a 2006, com oito volumes e 89 artigos. Um total de 478 palavras-chaves foram utilizadas, aqui a utilização da visão terminológica, onde não se contam as palavras, mas os termos. Desta forma "indexação" e "indexação automática" são termos distintos, assim para 478 termos temos 1.156 palavras analisadas. [5]

Quadro 1 – Termos mais freqüentes nas palavras-chaves nos artigos da revista Ciência da Informação

Termo	Ocorrência
Ciência da Informação	8
Comunicação científica	6
Biblioteca digital	6
Bibliometria	6
Informação	6
Biblioteca universitária	5
Acesso aberto	4
Organização do conhecimento	4
Redes sociais	3
Internet	3

Esses dados serviram para verificar quais os assuntos mais freqüentes nos artigos publicados. Outros pontos também foram alvo de verificação, como: a) as variações de número, como em: repositório institucional e repositórios institucionais; b) variação de fontes, como em: Ciência da Informação, Ciência da informação e ciência da informação; e c) variação lingüística, com em: acesso aberto e acesso livre. Todas essas variações, em muitos casos, tornam-se termos distintos, o que refletirá na apresentação e recuperação por assuntos no repositório.

Análise de termos da área de Ciência da Informação

[5] O periódico, "Ciência da Informação", publicado pelo Instituto Brasileiro de Informação em Ciência e Tecnologia – IBICT está disponível na web no endereço: http://www.ibict.br/cionline/. Os termos mais utilizados e sua freqüência são apresentados na quadro 1.

O metadado palavra-chave em repositórios baseados no DSpace é implementado no elemento "subjetct" do esquema de metadados Dublin Core. É um campo de entrada livre, ou seja, o autor pode escrever o que melhor lhe é conveniente. Essa liberdade, porém, tem o custo da falta de padronização, que podem gerar as variações supracitadas. Para Borbinha (2000) o elemento "*subject*" é traduzido como "Assunto ou Palavras-chaves", que no DSpace vai aparecerá nas duas formas: a) como palavra-chave: no formulário de submissão de documento, há campos para a entrada das palavras-chaves, nesse ponto pode-se fazer o uso do vocabulário controlado; b) como assunto: na página inicial há duas possibilidades: 1) visualizar uma lista dos assuntos existentes no repositório, selecionar um assunto e ver os documentos contidos nesse assunto; 2) fazer uma busca utilizando como filtro o assunto. Neste caso uma lista hierárquica dos assuntos é apresentada com uma caixa de opções múltiplas, pode-se fazer a busca com várias opções de assuntos marcadas como filtro. Nas opções "a" e "b2" o vocabulário controlado se faz presente. Para o Dublin Core Metadata Iniciative (DCMI) vários vocabulários controlados são sugeridos, como: *Library of Congress Subject Heading* – LCSH, *Dewey Decimal Classification* – DDC; *Library of Congress Classification* – LCC e Universal Decimal Classification. – UDC. O DSpace na versão 1.4 traz dois vocabulários controlados um em inglês e outro em norueguês.

Uma análise preliminar dos vocabulários fornecidos para a Ciência da Informação, revela que: a) para a CDU (CDU,1977) não há uma subclasse para a Ciência da Informação, sendo utilizada a classe 0 " Generalidades". Por não ter a finalidade de representar assuntos específicos, mas para classificar, seria muito complicado implementar esse tipo de vocabulário.

b) para o LCC (LCC) há a subclasse: "Z *Books* (General). *Writing. Paleography*. *Book industries and trade. Libraries. Bibliography*" contendo outra subclasse "ZA *Information resources (General)*" com 11 elementos como o "ZA4050-4480 *Electronic information resources*". Isso permitiria três níveis hierárquicos para o repositório, mas sem muitas especificações necessárias aos assuntos da área.

c) para o DCC os assuntos relacionados à área da Ciência da Informação estão em generalidades "*000 Generalities*" como em "*026 Libraries for specific subjects*", sem muita especificação e com a possibilidade de criar apenas dois níveis hierárquicos.

d) para o vocabulário controlado do DSpace em norueguês a área da Ciência da Informação é contemplado sendo hierarquicamente dependente da Ciências Sociais e possui seis subdivisões (figura 1 apresentada anteriormente). Vemos que apenas grandes áreas são apresentadas, algumas nem sendo próprias da Ciência da Informação, mas áreas correlatas.

e) para o vocabulário controlado do DSpace em inglês o assunto Ciência da Informação não aparece, sendo que o termo "*information science*" aparece apenas na seguinte seqüência: "*FORESTRY, AGRICULTURAL SCIENCES and LANDSCAPE PLANNING" → Area economics → Information science.* Que contextualmente não condiz com o entendimento para a área da Ciência da Informação.

Ama análise das palavras-chaves mais utilizadas nos artigos do periódico "Ciência da Informação" revela que assuntos agregados no vocabulário aparecem separados nos artigos como assuntos distintos, o termo "Recuperação e organização do Conhecimento", do vocabulário controlado, pode ser desmembrado em "recuperação do conhecimento" e "organização do conhecimento" e, dessa forma, aparecem como assunto de artigos distintos na revista. Mesmo possuindo correlação, os termos utilizados para recuperação do conhecimento são: "recuperação da informação", "sistemas para recuperação da informação", "sistema de busca", "mecanismos de busca" e "busca de informação". Essas variações nos levam a inferir quanto a

necessidade de estudo mais aprofundado para a criação de modelos de vocabulário controlado que sejam realmente significativas às áreas cobertas pelo repositório.

Análise do metadado assunto em repositórios

A análise das palavras-chaves nos repositórios mostrou os problemas causados pela falta de padronização na entrada dos metadados, causados possivelmente pelo auto-arquivamento, o que ressalta a necessidade de vocabulários controlados que permitam, além da padronização, a possibilidade de organização que, em segundo plano, facilitaria a recuperação por assunto. Uma análise preliminar dos termos utilizados para a recuperação por assunto e quais documentos recuperam mostra que as variações interferem na efetividade dessa facilidade implementada nos repositórios baseados no DSpace. O quadro 2 a seguir fornece um panorama dos problemas observados

Quadro 2 – Análise em relação às variações

Repositório	Variação terminológica	Variação de gênero ou número	Variação gráfica	Variação de língua
MIT	Sim	Sim	Sim	Não observado
Repositorium	Sim	Sim	Sim	Sim
Bdjur	Sim	Sim	Sim	Não observado

Em todos os repositórios foram observados problemas de variações nos termos utilizados como assunto, termos equivalentes recuperam documentos distintos. A análise verificou quatro tipos de variações, apenas a variação em relação ao idioma foi verificada no Repositorium, enquanto os outros tipos de variações foram constantes em todos os repositórios. Em relação a variação de língua, termos como “Brasil” e “Brazil” recuperam documentos distintos no Repositorium, mesmo possuindo equivalência de sentido, apenas em línguas diferentes. A possibilidade de entrar com documentos em língua estrangeira deve ser analisada. Os metadados estarão na língua de origem do documento, ou podem ser traduzidos, ficando o documento no formato original, mas a descrição em metadados na língua vernácula. As variações de número e gráficas são constantes, termos no plural apresentam-se em grande número e termos no singular recuperam documentos distintos dos termos no plural é muito comum nos três repositórios. As variações gráficas apresentaram-se em duas formas, sendo a mais comum a diferença entre iniciais maiúsculas e minúsculas, como em “*Object Detection*’, ‘*Objet detection*’ e ‘*object detection*’ no repositório do MIT, apesar de todos recuperarem os mesmo documentos, apresentam-se como três assuntos distintos, outra maneira de variação gráfica, apresenta-se na BDJUR em que os termos “obrigação de não fazer” e “obrigação de não-fazer” recuperam documentos distintos. Variações como o apagamento de parte do termo sem alterar o significado pode ser visto no repositório do MIT em que o termo “digital institutional repository” recupera um Item, enquanto o termo “intitutional repository” recupera cinco itens, a equivalência do significado e o apagamento do termo “digital” pode ser comprovada com os títulos recuperados (quadro 3).

Quadro 3 – Relação dos documentos recuperados pelo termo repositório do MIT

Termo	Seq.	Titulo recuperado
Digital institutional repository	1	Implementing an Institutional Repository: The DSpace Experience at MIT
Institutional repository	1	The DSpace Institutional Digital Repository System: Current Functionality
	2	Building a Business Plan for DSpace, MIT Libraries Digital Institutional Repository
	3	DSpace: An Open Source Dynamic Digital Repository
	4	DSpace as an Open Archival Information System: Current Status and Future Directions
	5	The DSpace Open Source Digital Asset Management System: Challenges and Opportunities

Pelos títulos recuperados, podemos ver que os termos “digital institutional repository” e “institutional repository” possuem equivalência semântica, mas foram tratados como assuntos distintos e ,portanto, a função de recuperar Itens, não será cumprinda a tarefa eficientemente.

Outro ponto relacionado ao assunto a ser analisado é a hierarquia dos termos. Um termo mais genérico (hiperônimo) deve conter os termos mais específicos (hipônimo). Desta forma, se usar a taxonomia proposta por Hawkin (Hawkin, Larson e Caton, 2003), como ilustração ao conceito de hierarquização, podemos ver no recorte (figura 2)

1. Pesquisa em Ciência da Informação (tradução dos autores);
    1.1. Conceitos básicos, definições, teorias, metodologias e aplicações;
    1.2. Propriedades, necessidades, qualidade e valor da informação;
    1.3. Estatísticas e medições
        1.3.1. Bibliometria, Análise de citação, Cienciometria e Infometria
    1.4. Pesquisa em recuperação de informação
        1.4.1. Técnica de busca (booleano, Fuzzy e língua natural), O processo de busca
    .....

Figura 2

No recorte da figura 2 podemos verificar a hierarquização dos termos, assim o assunto assinalado por 1 abrange todos os termos inferiores 1.1, 1.2, 1.3, 1.3.1 ..... Desta forma se um Item tivesse como assunto “bibliometria” poderia ser recuperado por: “bibliometria”, “estatísticas e medições” e “pesquisa em Ciência da Informação”. Ao ser especificado um assunto, todos os assuntos hierarquicamente superiores também devem ser assinalados, pois são inclusivos. Um problema freqüente nas recuperações por assunto pode ser visto em relação a essa hierarquização, o termo mais abrangente não se relaciona com o termos menos abrangentes. Esta falta de hierarquia entre as palavras-chaves inseridas pelo autor dificulta na organização por assunto. O quadro 4 mostra alguns problemas de hierarquização apresentados nos repositórios

Quadro 4 – Análise em relação aos problemas de hierarquia

Repositório	Problemas de hierarquia	Observações
MIT	Hiperônimo isolado do hipônimo	O assunto “conhecimento” recupera três itens diferentes do assunto “aquisição de conhecimento”; o assunto guerra da Coréia traz um Item e o assunto “guerra da Coréia – 1950 a 1963” recupera outro Item.
Repositorium	Hiperônimo isolado do hipônimo	O Assunto “Brazil” recupera 1 documento e “Brazilian Portuguese” recupera 2, distintos; O assunto “Brasil” recupera 5 Itens. O Assunto “Português” recupera 1 Itens e o assunto “Português do Brasil” recupera 2 Itens, todos distintos.
BDJUR	Hiperônimo isolado do hipônimo	O Assunto “obrigação” recupera um Item enquanto o assunto “obrigação de fazer” recupera 4 Itens

A falta de hierarquização dos termos implica na impossibilidade de recuperação pelo assunto mais genérico dos Itens que tratam dos assuntos mais específicos. No caso do Repositorium, podemos ver o termo “Português” como língua recuperando um Item (termo mais genérico) e o termo “português do Brasil” (termo mais específico) recuperando dois Itens, sendo que os Itens recuperados pelos termos são totalmente distintos. Se uma hierarquia fosse criada, possivelmente o termo “português” recuperaria três Itens (a,b e c) e o termo “português do Brasil” recuperaria dois (b e c). Pela análise do documento recuperado pelo termo “português” seria melhor a criação do termo “português europeu” ou “português de Portugal” inexistente no Repositorium como termos para recuperação por assunto.

Todos os problemas apresentados anteriormente podem ser classificados em: variação e hierarquização e foram gerados pelas palavras-chaves inseridas pelo autor no auto-arquivamento e que podem ou não ter sido alvo de críticas na validação dos metadados.

Na submissão de um documento, os formulários são, na maioria dos casos, campos livre, onde o autor transcreve passagens do texto. Essa liberdade causa alguns problemas, principalmente na recuperação por assunto e influencia na política do repositório. No caso de documentos em língua estrangeira, os metadados devem ser traduzidos ou mantidos na língua original? Essa decisão influenciará na qualidade da recuperação por assunto, como visto anteriormente, pois irá inserir termos em língua estrangeira sem relação como os termos em língua vernácula. Desta forma seria uma boa prática ter os metadados uniformizados em relação ao idioma. Isso facilitaria inclusive na recuperação pelas ferramentas de busca. O termo recuperaria tantos os Itens em língua vernácula quanto os em língua estrangeira. Essa opção pode ser feita pelo autor durante o processo de submissão do documento ou pela edição dos metadados pelo administrador do repositório.

O vocabulário controlado é uma opção para solucionar os problemas apresentados nos repositórios em relação ao assunto, pois forneceria termos padrão para a inserção de termos no campo palavra-chave, implementado de forma hierárquica no DSpace. Esta facilidade carece de estudos para a criação de um vocabulário controlado condizente com as áreas. Importante mencionar que essa facilidade não engessa a entrada de termos, pode-se mesclar as duas formas, usar o vocabulário controlado mais geral e entrar com outros termos mais específicos. A utilização de vocabulário controlado irá atuar em dois pontos no repositório: na inserção dos termos no campo palavra-chave e na busca por assunto, note que difere da recuperação por assunto mencionada

anteriormente. A busca por assunto apresenta uma árvore hierárquica com os assuntos existentes no vocabulário controlado e um campo para entrar com o termo para a busca. Escreve-se o termo que se deseja buscar e marca-se o assunto a qual deseja restringir a busca. O assunto torna-se um filtro que restringe a busca.
A opção pela utilização do vocabulário controlado dá-se por meio de customização do repositório e a criação de um arquivo no formato XML com os termos hierarquicamente organizados formando a estrutura de vocabulário controlado. O DSpace fornece dois exemplos, um em inglês e outro em norueguês (recorte na figura 1), ambos com termos em várias áreas destinados aos repositórios institucionais. Por serem bastante diversificados, não contemplam com profundidade nenhuma área.

Considerações finais

É de extrema importância destacar, aqui, que esse trabalho não pretende criticar, nem a ferramenta (DSpace), nem a facilidade (recuperação por assunto) e nem a implementação feita pelas instituições (Universidade do Minho, MIT e STJ), apenas oferece uma análise preliminar que verifica a necessidade de estudos mais aprofundados na área de qualidade de metadados e como a Ciência da Informação possui um papel fundamental na gestão da informação contida nos repositórios. Por ser uma facilidade necessária, mas ainda incipiente na sua implementação, entendemos que estamos ainda iniciando o estudo em recuperação por assunto nos repositórios institucionais baseados no DSpace, com o objetivo de indicar como os vocabulários controlados permitiriam facilitar essa recuperação e padronização dos termos utilizados. Há necessidade da criação de modelos de vocabulários controlados por área que permitam refletir os assuntos cobertos pelas pesquisas atuais. Isso padronizaria a descrição do metadado palavra-chave dos documentos e permitira uma recuperação por assunto mais eficaz. O estudo baseou-se apenas na identificação dos problemas e verificação de sua ocorrência em repositórios. A análise dos vocabulários controlados indicados pelo DSpace, em relação à Ciência da Informação, teve por objetivo verificar a representatividade dos assuntos das pesquisas feitas no Brasil em relação aos termos fornecidos pelo DSpace.

Referências

BDJUR. Biblioteca Digital do Supremo Tribunal de Justiça. Disponível em: http://bdjur.stj.gov.br/dspace . Acesso em: 29/05/2007

BORBINHA, J. L. Biblioteca nacional, 2000. Disponível em: http://purl.pt/201/1/

IBICT, Ciência da Informação V 33 Nº 1, Nº 2, Nº 3; 2004

IBICT, Ciência da Informação V 34 Nº 1, Nº 2, Nº 3; 2005

IBICT, Ciência da Informação V 35 Nº 1, Nº 2; 2006

CDU, Classificação Decimal Universal; edição-padrão internacional em língua portuguesa / UDC Consortium; tradução de Francisco F. L. de Albuquerque e Maria Thereza G. F. de Albuquerque; revisão de Antonio Agenor Briquet de Lemos – Brasília : Instituto Brasileiro de Informação em Ciência e Tecnologia, 1977.

DCC - Dewey Decimal Classification, disponível em: http://www.oclc.org/dewey/

DCMI – Dubin Core Metadata Inciative, disponível em: http://www.dcmi.org

DSPACE, DSpace System Documentation, disponível em: http://www.dspace.org/technology/system-docs/ Acesso em: 29/05/2007

DSPACE, Introducing DSpace, disponível em: Acesso em:

http://dspace.org/introduction/index.html29/05/2007
FAULSTICH, E. ABREL, S. P. lingüística aplicada à terminologia e à lexicografia. Cooperação internacional: Brasil Canadá. UFGRS. Porto Alegre. 2003.
HAWKINS, T. D. LARSON, S. E. CATON, B. Q. Information Science Abstracts: Tracking the Literature of Information Science. Part 2: A New Taxonomy for Information Science IN JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, disponível em: http://www.ugr.es/~alozano/Translations/3ATrackingtheliterature2.pdf, 2003
LYNCH, C. A. Institutional repositories: essential infrastructure for scholarship in the digital age. ARL Bimonthly Report, 26, 2003. Disponível em: http://www.arl.org/newsltr/226/ir.html
LCC. Library of Congress Classification. Disponível em: http://www.loc.gov/catdir/cpso/lcco/
LYNCH, C. A. Institutional repositories: essential infrastructure for scholarship in the digital age. ARL Bimonthly Report, 26, 2003. Disponível em: <http://www.arl.org/newsltr/226/ir.html>. Acesso em: maio 2007.
MIT – DSpace at MIT. Disponível em: http://dspace.mit.edu/ acesso em: 29/05/2007
OPEN Archives Initiative - OAI. Disponível em: <http://www.openarchives.org/>.
PAVEL, S. NOLET, D. Manual de Teminologia. Tradução de Enilde Faulstich. Bureau de la traduction, Montreal. Canadá. 2002
REPOSITORIUM. Repositório da Universidade do Minho. Disponível em: https://repositorium.sdum.uminho.pt/ . Acesso em: 29/05/2007
TRISKA, R. CAFÉ, L. *Arquivos abertos: subprojeto da Biblioteca Digital Brasileira* Ci. Inf. vol.30 no.3 Brasília Sept./Dec. 2001

Título: DSPACE VERSÃO 1.4: UMA ANÁLISE DAS FACILIDADES RELACIONADAS AO ASSUNTO

Autores:

Shintaku, M. [1]
Brascher, M.[2]

1. Instituto Brasileiro de Informação em Ciência e Tecnologia – IBICT / Universidade de Brasília – UnB
2. Universidade de Brasília - UnB