Vocabulário controlado e palavras-chave em repositórios digitais: relato de experiência do repositório institucional da FGV

Márcia Nunes Bacha (FGV) - marcia.bacha@fgv.br

Maria do Socorro G. de Almeida (FGV) - Maria.Socorro@fgv.br

Resumo:

Este estudo propõe a utilização das linguagens natural e controlada como forma de melhoria da pesquisa por assunto dentro de um repositório digital. Foi traçado um paralelo entre as atribuições de assuntos feitas em bibliotecas tradicionais e repositórios digitais. Estas últimas, disponíveis na web, requerem uma descrição informal do assunto. Em contrapartida, a hierarquização dos assuntos é também de extrema importância. O fato mais importante observado neste estudo é que práticas emergentes de melhoria na criação e descrição de metadados de assunto, usando mutuamente as duas linguagens, enriquece as pesquisas e facilita o acesso a objetos digitais.

Palavras-chave: *Linguagem natural. Linguagem controlada. Bibliotecas digitais*

Área temática: *Temática I: Tecnologias de informação e comunicação – um passo a frente*

Vocabulário controlado e palavras-chave em repositórios digitais: relato de experiência do repositório institucional da FGV

Resumo:
Este estudo propõe a utilização das linguagens natural e controlada como forma de melhoria da pesquisa por assunto dentro de um repositório digital. Foi traçado um paralelo entre as atribuições de assuntos feitas em bibliotecas tradicionais e repositórios digitais. Estas últimas, disponíveis na web**,** requerem uma descrição informal do assunto. Em contrapartida**,** a hierarquização dos assuntos é também de extrema importância. O fato mais importante observado neste estudo é que práticas emergentes de melhoria na criação e descrição de metadados de assunto, usando mutuamente as duas linguagens, enriquece as pesquisas e facilita o acesso a objetos digitais.

Palavras-chave: Linguagem natural. Linguagem controlada. Repositórios digitais.

Área temática: Tecnologias de informação e comunicação – um passo a frente

1 INTRODUÇÃO

Comparando com as bibliotecas tradicionais, os repositórios digitais estão disponíveis através da web para que qualquer usuário acesse. Muitas vezes, uma atribuição informal do assunto satisfaz uma busca. Porém**,** como num repositório digital o enriquecimento dos metadados permite uma maior recuperação do documento**,** a ideia de que palavra-chave e vocabulário controlado devem andar juntos, foi o desafio na criação do repositório digital da FGV.

Tendo em mente que o acesso físico a um documento digital não se faz possível, a recuperação da informação por assunto é crucial para um repositório digital de sucesso.

Para Greenberg (2005), metadados são dados estruturados sobre um objeto que suporta funções associadas do objeto designado. São usados em bibliotecas digitais para organizar a informação e sua recuperação efetiva através das pesquisas.

Para o controle de terminologia de assuntos, o Repositório Institucional da FGV se baseou no Catálogo de Autoridades da Rede Bibliodata, que utiliza vocabulário controlado, e tem como parâmetro a LCSH (Lista de Cabeçalhos de Assunto da Library of Congress), mas permite também que seus usuários utilizem a palavra-chave, ou texto livre, conforme vemos abaixo:

a) Vocabulário controlado – que consiste em uma lista formalmente mantida de termos;
b) Palavra-chave, ou, texto livre – que se baseiam em termos numa linguagem natural.

A alta qualidade na descrição dos metadados de assunto é fundamental para organizar o acesso. Observa-se que as práticas emergentes para melhoria na criação de metadados de assunto relevam que uns documentos devem ser descritos com valores mutuamente complementares em campos de metadados de vocabulário controlado e texto livre de assunto. (ZAVALINA, 2008, tradução nossa).

2 Revisão de Literatura

Inúmeros estudos sobre o uso da linguagem controlada e natural na recuperação da informação têm se concentrado na utilização conjunta das duas linguagens na estratégia de busca, comprovando que o uso simultâneo dessas linguagens proporciona melhor desempenho nos resultados. (LOPES, 2002)

Neste relato de experiência é necessário conceituar a linguagem natural e controlada e demonstrar que estas podem e devem caminhar juntas dentro de um repositório digital.

A linguagem natural é aquela que o próprio usuário define sem intervenção de um indexador.

Para Lancaster (1993, p.200). “a expressão normalmente se refere às palavras que ocorrem em textos impressos, considerando-se como seu sinônimo a expressão “texto livre”.

Num repositório digital, os metadados mais importantes são aqueles fornecidos pelos próprios autores e incluem informações detalhadas dos documentos. (TERRA, et al., 2005).

A linguagem controlada ou mais especificamente vocabulário controlado é uma lista de termos pré-definidos, autorizados e que foram pré-selecionados para organizar o conhecimento**,** visando a precisão na recuperação posterior ao documento.

Segundo Lancaster (2004), um vocabulário controlado é essencialmente uma lista de termos autorizados. Em geral o indexador somente pode atribuir a um documento termos que constem na lista adotada pela instituição.

Assim, os campos de resumo, de títulos, de identificadores, de descritores ou cabeçalhos de assunto e de códigos de classificação podem ser amplamente

utilizados visando à obtenção de um resultado mais satisfatório, independentemente da verificação, no momento de operação da busca, de qual dessas linguagens terá melhor desempenho. O foco, portanto, está na obtenção de resultados satisfatórios, e não no instrumento utilizado para alcançar esses resultados.

3 Metodologia

Dentro dos processos técnicos de uma biblioteca, a indexação é uma das tarefas que exige maior cuidado e análise, isto serve também para um repositório digital. A metodologia para criação da linguagem documentária foi dividida em duas etapas. A primeira etapa consiste em uma revisão de literatura e análise, baseada em artigos e livros que tratassem da temática indexação e vocabulário controlado, inicialmente foi acertado o uso do vocabulário controlado por ser utilizado no processamento técnico da FGV e é a instrumentalização que contribui efetivamente para melhor aproveitamento das possibilidades da representação dos conteúdos dos documentos. Numa segunda etapa, foi definido que além do vocabulário controlado a linguagem natural, utilizada pelos autores, também seria adotada.

A combinação das duas abordagens (vocabulário controlado e linguagem natural) veio resolver questões importantes: uma era atender as reivindicações dos autores que sentiam a necessidade de reconhecer os termos utilizados por eles, outra, era melhorar a eficácia dos sistemas de recuperação, sistemas de navegação Web, e outros ambientes que buscam identificar e localizar o conteúdo desejado através de algum tipo de descrição utilizando uma linguagem de armazenamento e informação.

A Biblioteca da Fundação Getulio Vargas (FGV), utiliza o vocabulário controlado da Rede Bibliodata para indexar seus documentos em seu catálogo físico.

Para implantar o repositório digital foi necessário desenvolver um estudo que fizesse a correlação dos campos MARC**,** já utilizados na biblioteca tradicional**,** com DUBLIN CORE**,** que é esquema de dados que descreve os objetos digitais.

Ao montar o workflow a ser utilizado no repositório digital, a FGV optou pelo auto-arquivamento dos documentos**.** Assim, neste fluxo de trabalho**,** as palavras-chave inseridas pelo autor no ato da submissão seriam trocadas, pelo indexador, por assuntos autorizados.

Alguns autores, com razão, questionaram este procedimento, pois o Dublin Core, embora tenha campo específico para informar os assuntos da LCSH, não permite uma relação entre os termos, como acontece com os softwares de bibliotecas tradicionais,

O Dublin Core é uma descrição simples e muito restrita que não permite hierarquizações de assuntos, utilização de remissivas, etc. Sendo assim, a utilização de uma linguagem natural viria para sanar esta lacuna.

Para solucionar este problema, foi realizada uma estratégia, onde foi instalado no DSpace o Dublin Core não somente com os 15 principais identificadores, mas sim com seus qualificadores também. Isso significa dizer que o metadado dc.suject que anteriormente era preenchido para designar um assunto, agora podia ser subdividido.

Houve então uma mudança no workflow. Palavras-chave descritas pelos autores seriam preservadas, mesmo em outras línguas, preenchidas no formulário no ato da submissão do autor, estas designadas “subject”. Na parte do workflow em que os indexadores estivessem padronizando os metadados, os assuntos controlados seriam inseridos em “subject LCSH”, uma vez que a FGV utiliza a Library of Congress para padronização de autoridades de nomes e assuntos.

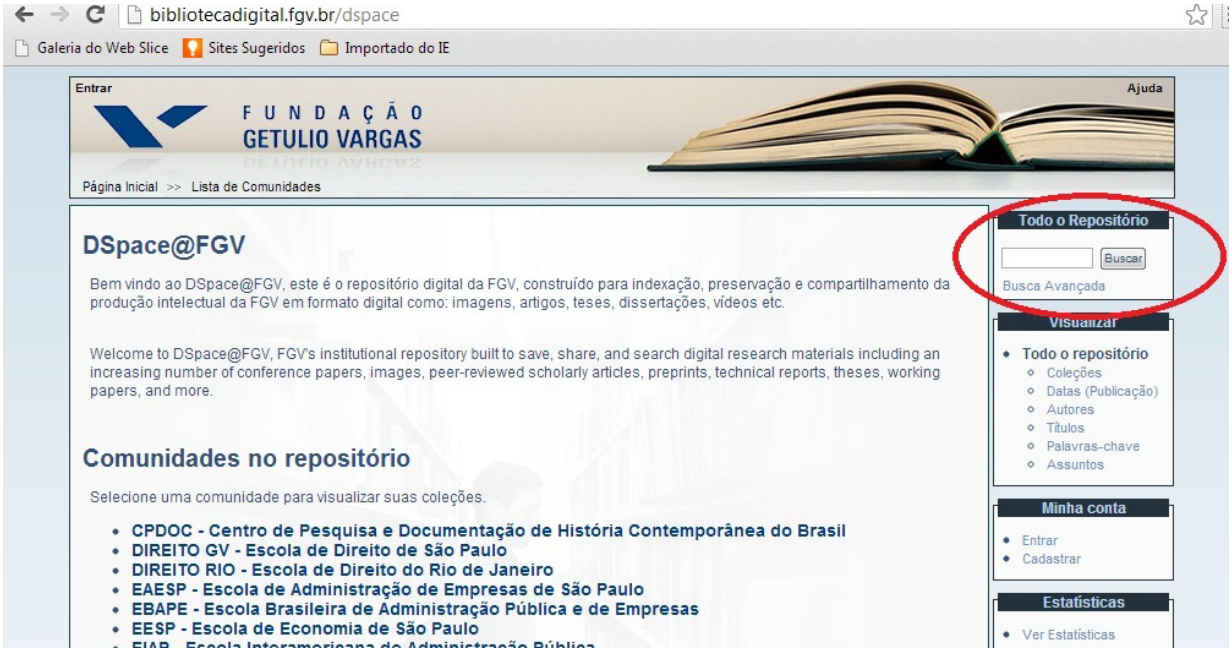
Assim uma lista geral de assuntos foi sendo alimentada com assuntos em linguagem controlada e texto livre.

No que diz respeito à recuperação da informação**,** isso é perfeito**,** pois o acesso a informação fica assegurado. Numa pesquisa por assunto, por exemplo, não é necessário que um pesquisador leigo tenha conhecimentos sobre os assuntos autorizados e utilizados**,** como numa biblioteca tradicional. Tanto o termo livre quanto o vocabulário controlado terão o mesmo peso na hora da busca.

Como estas modificações foram feitas para melhoria das buscas, o repositório digital da FGV implantou um diferencial dentro do DSpace. A separação das listas de palavras-chave e assuntos controlados.

Numa pesquisa em todo o repositório, na opção “Buscar”, aparecem os resultados em todas as listas existentes, sejam elas: título, autor, assuntos, palavras-chave, etc (ver Figura 1).

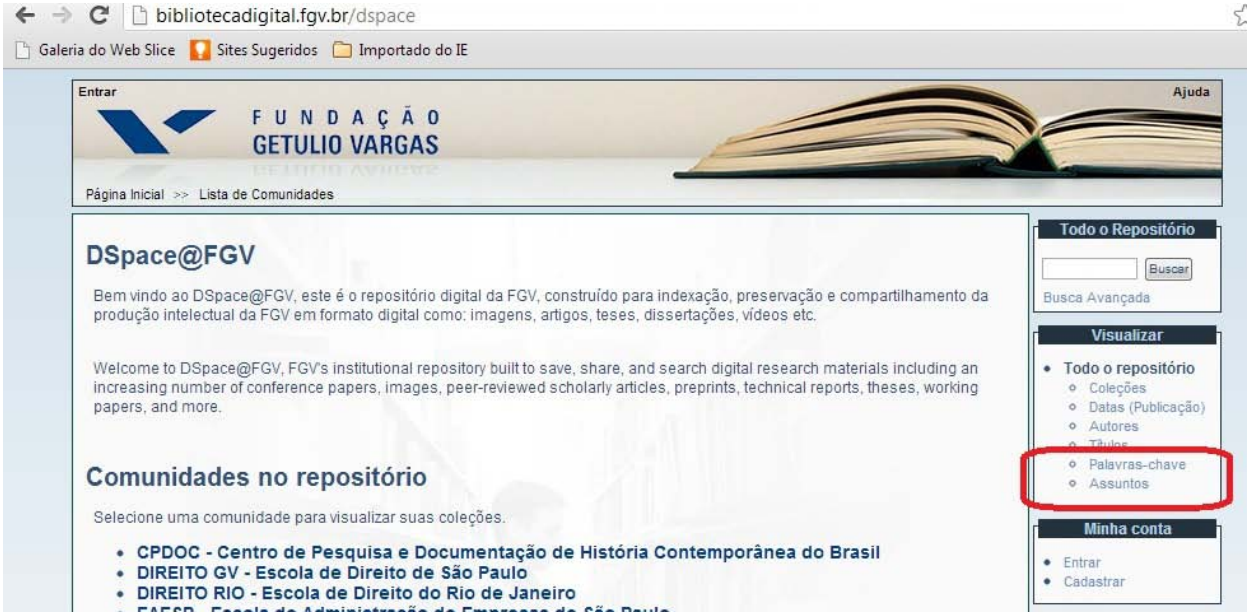
Figura 1 – Busca em todo repositório



Fonte: Repositório DSpace@FGV (2013)

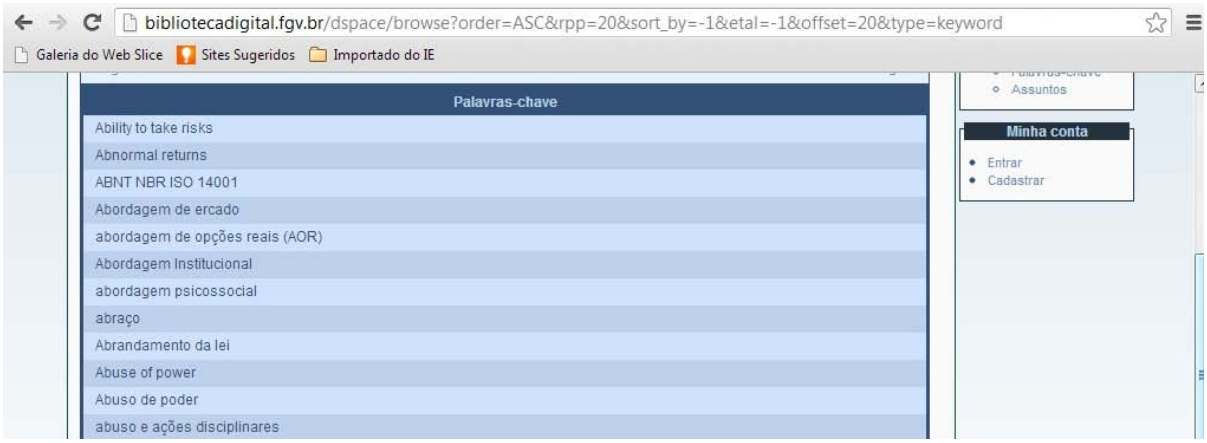
Com a separação das listas (ver Figura 2), o pesquisador pode navegar pela lista de palavras-chave (ver Figura 3) ou pela lista de assuntos (ver Figura 4), conforme for mais interessante para suas pesquisas. Na lista de palavras-chave não há interferência dos indexadores, não são alteradas a forma de escrita, são aceitas palavras em outras línguas, as palavras são descritas exatamente como os autores a submeteram. Na lista por assunto, são aquelas atribuídas pela lista de Autoridades de Assunto da Rede Bibliodata.

Figura 2 - Busca por palavras-chave e assuntos



Fonte: Repositório DSpace@FGV (2013)

Figura 3 – Índice de palavras-chave



Fonte: Repositório DSpace@FGV (2013)

Figura 4 – Índice de vocabulário controlado

Fonte: Repositório DSpace@FGV (2013)

3 Considerações Finais

Os resultados apresentados neste estudo mostram o esforço para tirar proveito da riqueza inerente à semântica**,** a fim de proporcionar uma recuperação da informação mais eficiente dentro de repositórios digitais.

A experiência de utilizar a linguagem natural e linguagem controlada simultaneamente dentro do repositório digital da FGV, demonstrou que quanto mais detalhado o conjunto de metadados de assunto mais completa a representação de conteúdo intelectual dos objetos digitais, e finalmente, a melhoria do acesso por assunto para os usuários.

Deve-se ressaltar que numa instituição como a FGV tradicional por tratar seus registros de autoridades no catálogo convencional de forma hierarquizada, os desafios com a implantação do repositório digital foram grandes. A nova forma em descrever os documentos digitais, a reflexão sobre as formas de recuperação, só vieram a acrescentar conhecimento aos indexadores.

Em um catálogo tradicional há também a preocupação com a recuperação e o fácil acesso às informações. Porém num repositório digital isto se potencializa, pois as limitações do Dublin Core fazem perceber que a linguagem natural e a controlada podem e devem caminhar juntas, uma completando a outra, visando única e exclusivamente o acesso à informação, mantendo a simplicidade e facilidade de uso.

Em linhas gerais, podemos concluir que uma recuperação da informação por assunto eficiente é crucial para um repositório digital de sucesso.

REFERÊNCIAS

GREENBERG, J. Metadata and the World Wide Web. In: **Encyclopedia of Library and Information Science**. New York: Marcel *Dekker,* 2005*.* p. 1876-1888. Disponível em < http://www.ils.unc.edu/mrc/pdf/greenberg03metadata.pdf> Acesso em: 12 mar. 2013.

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

____. ____ 2. ed. rev. atual. Brasília: Briquet de. Lemos, 2004.

LOPES, I.L. Uso das linguagens controlada e natural em bases de dados: revisão de literatura. **Ci. Inf.**, Brasília, v. 31, n. 1, p. 41-52, jan./abr. 2002. Disponível em <http://www.scielo.br/pdf/ci/v31n1/a05v31n1.pdf> Acesso em: 02 mar. 2013.

TERRA, J.C.C.; et al. **Taxonomia**: elemento fundamental para a gestão do conhecimento. 2005. Disponível em <http://www.terraforum.com.br> Acesso em: 17 de mar. 2013.

ZAVALINA, O.L. **Collection-level metadata and its role in subject access to digital libraries**. 2008. 84f. Dissertação (mestrado) – University of Illinois at Urbana-Champaign. Disponível em <http://www.academia.edu/467441/COLLECTION-LEVEL_SUBJECT_ACCESS_METADATA_APPLICATION_AND_USERS> Acesso em: 20 mar. 2013.