



Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz



CURSO DE ESPECIALIZAÇÃO EM INFORMAÇÃO CIENTÍFICA E TECNOLÓGICA EM SAÚDE

IMPLANTAÇÃO & CUSTOMIZAÇÃO DE VOCABULÁRIO CONTROLADO NO SISTEMA ARCA

por

Vitor Hugo da Silva Martins

Fundação Oswaldo Cruz/ Instituto de Comunicação e Informação Científica e Tecnológica em Saúde

Projeto apresentado ao Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, da Fundação Oswaldo Cruz, como requisito parcial para obtenção do título de Especialista em Informação Científica e Tecnológica em Saúde.

Orientador: Luis Fernando Sayão
Doutor em Ciência da Informação

Rio de Janeiro, novembro de 2009

SUMÁRIO

1	INTRODUÇÃO.....	3
2	CONTEXTUALIZAÇÃO.....	5
2.1	RECUPERANDO A INFORMAÇÃO ATRAVÉS DOS METADADOS.....	5
2.2	DSPACE E OUTRAS FERRAMENTAS DE IMPLANTAÇÃO.....	6
2.3	ESQUEMAS DE METADADOS.....	7
2.4	IMPLANTANDO VOCABULÁRIOS CONTROLADOS EM REPOSITÓRIOS.....	10
3.	O PROBLEMA	11
3.1	O SISTEMA ARCA.....	11
3.2	ARQUITETURA / INTERFACE DO SISTEMA.....	12
3.3	NECESSIDADE DE UM VOCABULÁRIO CONTROLADO.....	16
4	OBJETIVOS	21
4.1	OBJETIVO GERAL	21
4.2	OBJETIVOS ESPECÍFICOS	21
4.3	JUSTIFICATIVA.....	21
4.4	RESULTADOS ESPERADOS.....	22
5	ETAPAS DE DESENVOLVIMENTO.....	23
6	REFERÊNCIAS CONSULTADAS	24
7	CRONOGRAMA	26
8	ORÇAMENTO	27

1 INTRODUÇÃO

Mesmo com os sistemas de indexação e recuperação da informação estando cada vez mais sofisticados, num ambiente de constante mudança como a Internet, a relevância e a precisão nas respostas que os sistemas devolvem aos utilizadores ainda estão longe de alcançarem níveis satisfatórios. Grande parte dos usuários ao buscar informações seja na Web, em repositórios ou bases de dados, se depara com resultados de buscas que pouco, ou até mesmo nada, tenham a ver com o assunto pelo qual procuravam, tornando uma prática que deveria ser relativamente fácil em uma busca complexa e muitas vezes sem sucesso pela tal desejada informação. Como se não bastasse a dificuldade nos mecanismos de busca, há também a ausência da certeza da existência ou não da informação e de acordo com Campos e Gomes (2004), existe uma insatisfação de grande parte dos usuários com relação à recuperação da informação disponibilizada em rede e na Internet, devido à freqüente falta de ferramentas de acesso adequadas.

De modo a melhorar a eficácia e a eficiência dos serviços de informação, torna-se necessária à utilização de metadados. No entanto, com a utilização de diferentes metadados e estruturas de metadados pelos mais diversos tipos de serviços e aplicações, o problema se mantém. Com a evolução das bases e repositórios institucionais, verificou-se que como em outras áreas, a padronização será um elemento-chave para a utilização eficaz dos metadados.

Para Crow (2002), os repositórios institucionais são arquivos digitais que reúnem, preservam, divulgam e permitem o acesso à produção científica de comunidades universitárias. Podem ser criados de forma individualizada ou por grupos de instituições que trabalhem de forma cooperante. Já para Linch (2003) o repositório institucional é uma plataforma informatizada que possui como unidade de armazenamento o conjunto dos objetos digitais e seus metadados, criado como uma alternativa para divulgação da produção científica de uma instituição.

Neste contexto, o sistema ARCA, que é atualmente o repositório do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT) da Fiocruz (Fundação Oswaldo Cruz), recentemente implantado e customizado, dará continuidade ao processo de preservação, acessibilidade e disseminação dos trabalhos acadêmicos.

O sistema foi desenvolvido com base na solução *DSpace*, que é um *software*¹ livre, e que utiliza o padrão de metadados *Dublin Core(DC)*, e oferece dentre outras, a funcionalidade do auto-arquivamento, isto é, que o próprio autor ou pesquisador realize a submissão do trabalho na plataforma, desde o preenchimento de seus metadados, até o depósito final do arquivo.

É justamente neste instante que a falta de padronização no preenchimento dos metadados traz como conseqüência, além de uma possível redundância, duplicidade e ambigüidade dos dados, provocando em seguida, obstáculos significantes no processo de recuperar a informação.

Neste sentido, este projeto tem como objetivo geral implantar um vocabulário controlado customizado para o preenchimento do metadado que represente o conteúdo temático dos documentos (assunto), possibilitando, assim, uma maior eficiência na recuperação dos mesmos já que o repositório abrigará conteúdos de outras áreas fora da saúde. Para isso ele deverá ser customizado de tal forma que atenda aos propósitos do Instituto e não seja somente focado em temas relacionados à saúde.

¹ Produto desenvolvido pela Engenharia de Software, e inclui não só o programa de computador propriamente dito, mas também seus respectivos manuais e especificações.

2 CONTEXTUALIZAÇÃO

Na contextualização abordaremos os seguintes assuntos:

- Recuperando a informação através dos metadados;
- DSpace e outras ferramentas de implantação de repositórios;
- Esquemas de metadados;
- Implementando vocabulários controlados em repositórios.

2.1 RECUPERANDO A INFORMAÇÃO ATRAVÉS DOS METADADOS

Segundo Shintaku e Brascher (2006), os metadados, além de fornecer informações sobre o objeto digital, servem de ponto de recuperação a esses objetos, já que o princípio fundamental dos metadados é integrar coleções digitais, heterogêneas, em múltiplos formatos e de várias instituições, permitindo que elas sejam acessíveis a qualquer pessoa, em qualquer lugar ou hora.

Podemos verificar que em repositórios baseados no *DSpace*, o metadado assunto, cumpre a função de organizar, classificar e hierarquizar os documentos no repositório, e facilitar a posterior recuperação por assunto. É através deste metadado assunto que iremos adaptar uma ferramenta de vocabulário controlado customizado para aprimorar esta recuperação da informação.

2.2 DSPACE E OUTRAS FERRAMENTAS DE IMPLANTAÇÃO

Existem algumas ferramentas disponíveis no mercado para implantação de Repositórios ou bases de dados, sejam eles institucionais, temáticos, centrais ou de preservação. Na tabela a seguir listo algumas das mais utilizadas.

Institucionais	Temáticos	Centrais	Preservação
Archimede	Greenstone	Greenstone	LOCKSS
ARNO			DAITSS
CDSware			
DSpace	DSpace	DSpace	DSpace
Driver			
Eprints	Eprints	Eprints	
Fedora		Fedora	Fedora

Figura 1: Tipos de Repositórios

Aqui foi abordado O *DSpace* (solução que serviu como base para criação do ARCA), que é um *software* livre desenvolvido pelo MIT (*Massachusetts Institute of Technology*) e pelos Laboratórios da HP (*Hewlett-Packard Company*), intencionalmente implementado para servir de repositório.

2.3 ESQUEMAS DE METADADOS

Os metadados são organizados de uma forma específica denominada esquema. Existem diversos tipos de esquemas de metadados, utilizados em repositórios ou bases de dados, abaixo listo alguns exemplos e seus significados:

CEDARS	CURL Exemplars in Digital ARchiveS
NEDLIB	Networked European Deposit Library
METS	Metadata Encoding and Transmission Standard
RKMS	Recordkeeping Metadata Schema
DC	Dublin Core
ISAD	International Standard Archival Description
MODS	Metadata Object Description Schema
CSDGM	Content Standard for Digital Geospatial Metadata
GILS	Government Information Locator Services
VRA	Visual Resources Association

Seção 1.01 Figura 2: Esquemas de Metadados

Abordaremos neste projeto o *DC (Dublin Core)* que é o esquema utilizado pela ferramenta *Dspace*, e que serviu como base para o desenvolvimento do sistema ARCA do Instituto de Informação Científica e Tecnológica da Fiocruz (ICICT).

2.3.1 DUBLIN CORE (DC)

Para descrever os itens existentes no repositório, o *DSpace* utiliza o padrão de metadados *Dublin Core Resource Description* (DC). O nome "*Dublin*" se refere a Dublin, Ohio, U.S., onde o trabalho se originou a partir de um workshop sediado em 1995 pela *OCLC*². O "*Core*" se refere ao fato de que o conjunto de elementos de metadados é composto de uma lista básica, porém expansível.

O *Dublin Core* tem como característica ser um formato menos estruturado e mais flexível que propicia duas formas de utilização: **simples e qualificada**.

No *Dublin Core Simples* são usados 15 elementos padrão descritos conforme a tabela que se segue:

CONTEÚDO	PROPRIEDADE INTELLECTUAL	INSTÂNCIAS
TITLE: TÍTULO	CREATOR: CRIADOR	DATE: DATA
SUBJECT: ASSUNTO	PUBLISHER: PUBLICADOR	FORMAT: FORMATO
DESCRIPTION: DESCRIÇÃO	CONTRIBUTOR: CONTRIBUIDOR	IDENTIFIER: IDENTIFICADOR
TYPE: TIPO	RIGHTS: DIREITOS	LANGUAGE: IDIOMA
SOURCE: ORIGEM		
RELATION: RELAÇÃO		
COVERAGE: ABRANGÊNCIA		

Figura 3: Metadados do Dublin Core Simples

Cada elemento *Dublin Core* é de caráter opcional e pode ser repetido. O *DCMI* (*Dublin Core Metadata Initiative*) estabeleceu maneiras padronizadas para refinar os elementos e encorajar o uso de esquemas de codificação e vocabulário, portanto não há ordem no *Dublin Core* para apresentar ou utilizar os elementos.

² OCLC (Online Computer Library Center) é um consórcio de grandes bibliotecas principalmente dos EUA que detém o maior banco de dados catalográfico do mundo.

Já o *Dublin Core Qualificado* é apenas uma extensão ao *Dublin Core* original e seus 15 elementos (*Dublin Core Metadata Element Set, DCMES*), onde alguns dos elementos são acompanhados por um qualificador que os refina, tornando-os mais restritos (expos: *Date.Created, Date.Available, Date.Modified, Creator.Degree, Contributor.Editor, Subject.Other*).

Os qualificadores ou termos adicionais são identificados, geralmente em grupos de trabalho do *DCMI*, e julgados pelo *DCMI Usage Board* para estar em conformidade com os princípios da boa prática na qualificação dos elementos de metadados *Dublin Core*.

Em adição ao refinamento de elementos, o *Dublin Core Qualificado* inclui um conjunto de esquemas de codificação recomendados, desenhados para ajudar na interpretação dos valores de metadados. Estes esquemas incluem o vocabulário controlado, notações formais e regras de *parsing*³. Se um esquema de codificação não é compreendido por uma aplicação, o valor ainda deve ser útil para o leitor humano. (e.g., "2000-12-31" como expressão padrão para data).

³ Processo de analisar uma sequência de entrada (lida de um arquivo de computador ou do teclado, por exemplo) para determinar sua estrutura gramatical segundo uma determinada gramática formal. Essa análise faz parte de um compilador, junto com a análise léxica e análise semântica.

2.4 IMPLEMENTANDO VOCABULÁRIOS CONTROLADOS EM REPOSITÓRIOS

Para evitar grandes variações e padronizar entradas, a utilização de vocabulário controlado para metadados permite restringir as possibilidades de preenchimento do assunto. A literatura dá sinais de que um vocabulário estruturado melhora a experiência do usuário. Segundo Muramaki (2005), o uso de vocabulários controlados melhora a qualidade e disponibilidade como suporte para buscas em várias bases de dados. A utilização do vocabulário controlado dentre outras coisas trará confiabilidade na entrada de dados, com a diminuição de erros.

Verificou-se que a plataforma *DSpace* permite esta adaptação de modo que o desenvolvimento do vocabulário controlado local se dê utilizando a linguagem *XML*⁴ com os termos hierarquicamente e hidentamente organizados formando uma estrutura de vocabulário controlado.

⁴ XML por definição significa *Xtensible Markup Language* – Linguagem de Marcação Extensível. É um subtipo de SGML (acrônimo de *Standard Generalized Markup Language*, ou *Linguagem Padronizada de Marcação Genérica*) capaz de descrever diversos tipos de dados. Seu propósito principal é a facilidade de compartilhamento de informações através da Internet e sua principal característica é a de criar uma infraestrutura única para diversas linguagens permitindo assim a interoperabilidade entre sistemas.

3 O PROBLEMA

3.1 O SISTEMA ARCA

Buscando reunir, registrar e preservar a memória do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT), foi desenvolvido, baseado na solução tecnológica *DSpace*, o Repositório Institucional **ARCA**.

Os objetivos principais do repositório são o de armazenar, gerir, preservar e divulgar, a nível internacional, via Internet, a produção científica do Instituto. Deste modo, propiciará a proteção e valorização da memória intelectual do mesmo e também servirá para criar um aumento da visibilidade, acessibilidade e impacto do seu valor científico e cultural.

3.2 ARQUITETURA / INTERFACE DO SISTEMA

3.2.1 ARQUITETURA

A arquitetura básica de funcionamento do *DSpace* está baseada em três camadas segundo o sítio oficial (<http://www.dspace.org>):

- **Aplicação:** esta camada realiza a comunicação das outras com o exterior. Dentre algumas funções podemos destacar: a interoperabilidade através da simples importação e exportação dos *Schemas XML* e também a exportação dos *METS*⁵. Esta camada é formada basicamente da interface do sistema e do protocolo *OAI-PMH*⁶.
- **Negócio:** camada de controle do sistema e que administra as informações do sistema. Ela é composta pelo sistema de indexação e busca *Lucene*⁷, pela administração dos tipos de usuário do sistema, ferramentas administrativas em geral e também engloba o *handle*.
- **Armazenamento:** cuida do armazenamento físico dos dados no sistema, dos RDBMS (*relational database management system*) ou seja, banco de dados (que também gerem os metadados) e *bitstreams*.

⁵ O Metadata Encoding & Transmission Standard (METS) é um XML Schema elaborado sob o patrocínio da Digital Library Federation e mantido pela biblioteca do congresso norte-americano que permite criar uma rica estrutura capaz de registrar, não apenas os múltiplos tipos de metadados usados para descrever o acervo de uma biblioteca digital, como também, os próprios objetos digitais, seja para gerenciá-los ou para permitir o intercâmbio entre diferentes instituições.

⁶ O protocolo OAI-PMH (Open Archive Initiative – Protocol for Metadata Harvesting) é um mecanismo para transferência de dados entre repositórios digitais. É uma interface que um servidor de rede pode empregar para que os metadados de objetos residentes no servidor estejam disponíveis para aplicações externas que desejem coletar esses dados.

⁷ Criado por Doug Cutting em 2000, o Lucene é uma das mais famosas e mais usadas bibliotecas de programação para indexação e consulta de textos, disponível em código aberto. Sob o domínio da Apache Foundation, a biblioteca, escrita em java, pode ser utilizada em qualquer aplicativo J2SE ou J2EE, de código aberto ou não.

Abaixo podemos visualizar as três camadas do sistema e suas principais funcionalidades.

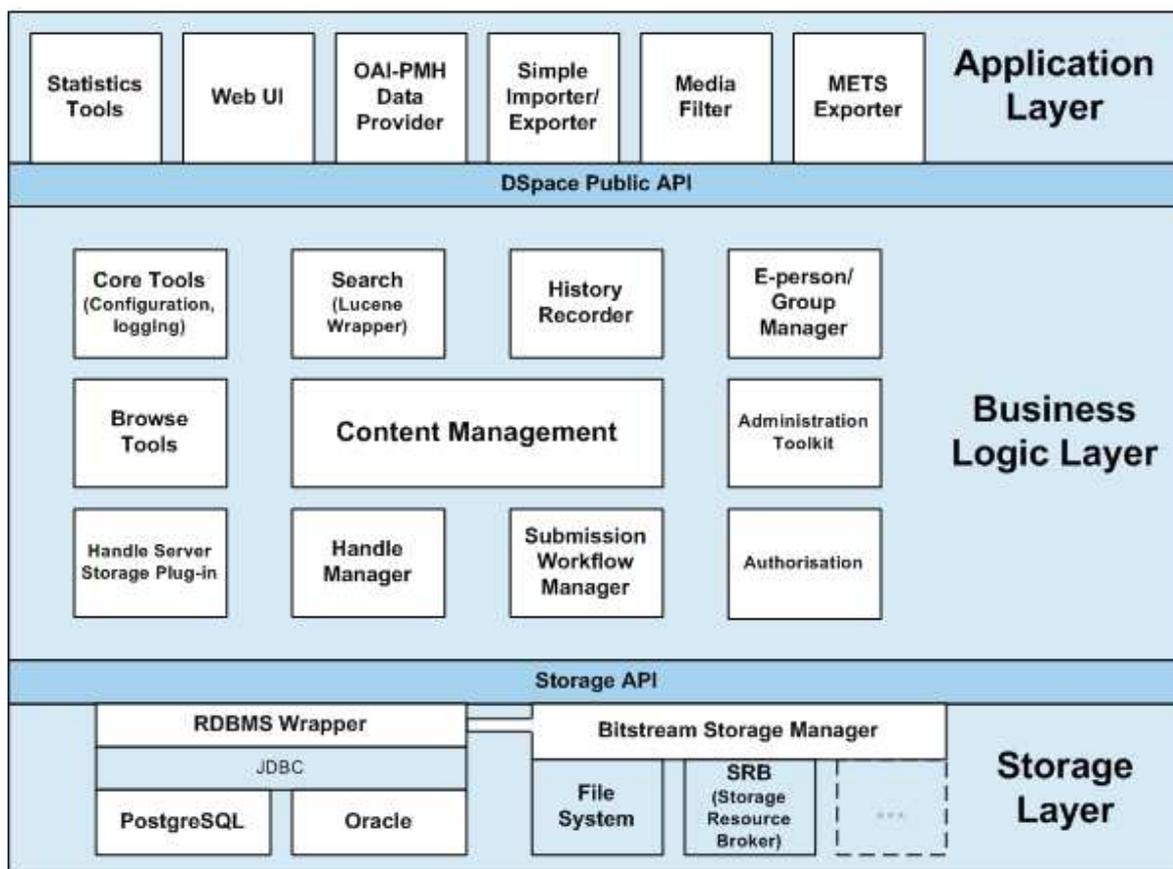


Figura 4: Camadas da Arquitetura do sistema (<http://www.dspace.org>)

3.2.2 INTERFACE DO SISTEMA

A interface do ARCA foi completamente customizada para estar de acordo com as demandas do Instituto e está em constante aprimoramento.

O *browser*⁸ utilizado para a visualização das figuras abaixo foi o *Mozilla Firefox 5.0*.



Figura 5: Repositório Institucional da Fiocruz – ARCA – Página Inicial

⁸ Browser ou Web browser é um programa de computador que habilita seus usuários a interagirem com documentos virtuais da Internet, também conhecidos como páginas e que estão hospedadas num servidor Web.

A tela abaixo exibe os registros de metadados do sistema, mostrando a utilização do esquema de metadados *Dublin Core Qualificado*.

Ministério da Saúde
FIOCRUZ
FUNDAÇÃO OSWALDO CRUZ

ICICT
Instituto de Comunicação e Informação Científica e Tecnológica em Saúde

ARCA
MEMÓRIA DIGITAL INSTITUCIONAL
INFORMAÇÃO CIENTÍFICA E TECNOLÓGICA EM SAÚDE

ARCA >
Administrador >

Registro e Campo de Metadados

[Esquema](#) | [Ajuda...](#)

Nota: Adicionar um novo campo para o Registro não corresponde a adicionar campos nos formulários de submissão!

	Identificação	Elemento	Qualificador	Nota de Abrangência		
77	contributor	advisor		Use primarily for thesis advisor.	Modificar	Remover...
78	contributor	advisorco		co-orientador	Modificar	Remover...
79	contributor	author			Modificar	Remover...
80	contributor	editor			Modificar	Remover...
81	contributor	event		Evento	Modificar	Remover...

Comunidades/
Coleções
E-people
Grupos
Items
Registros de Metadados
Registros no formato binário
Fluxo
Autorização
Editar notícias
Edita Licença Padrão
Supervisores
Estatísticas
Ajuda
sair

Figura 6: Repositório Institucional da Fiocruz – ARCA – Página do Esquema de Metadados DC

3.3 NECESSIDADE DE UM VOCABULÁRIO CONTROLADO

Conforme estudos do software *DSpace*, a partir do sítio da ferramenta⁹, foi constatado que a solução não vem com um vocabulário controlado implantado¹⁰, sendo necessário seu desenvolvimento e customização, para que como dito anteriormente, possa atender diversas áreas do conhecimento específicas do Instituto e não somente áreas relacionadas à saúde como será mostrado no vocabulário controlado *DeCS* mais abaixo.

3.3.1 ESTUDO DE CASO: ACM Computing Classification System

O *ACMCCS*, assim como é chamado o *Association for Computing Machinery Computing Classification System* (Sistema de Classificação da Computação da ACM) é um vocabulário controlado desenvolvido em XML que teve toda sua estrutura adaptada aos moldes da solução *DSpace* através de técnicos de computação do Departamento de Sistemas de Informação (*DSI - Department of Information Systems*) da Universidade do Minho em Portugal. É um vocabulário controlado específico da área de computação e sistemas de informação com aproximadamente 2900 registros e que serve como referência inicial onde concluímos que qualquer vocabulário mesmo fora do formato padrão do *DSpace* pode ser incorporado e customizado para a solução.

⁹ DSpace Official Website - <http://www.dspace.org>

¹⁰ A versão do Dspace 1.4.2 utilizada para desenvolver o sistema ARCA ainda não continha o vocabulário controlado desenvolvido. Vinha apenas adaptada para receber algum vocabulário controlado, sendo necessário ajustes no código-fonte para tal. Há estudos que a versão 2.0 com previsão de lançamento em 2010 possa já vir com esta funcionalidade incorporada.

Na figura a seguir podemos ver como o ACMCCS se comporta no ARCA no momento de submeter um item para o sistema.

The screenshot shows the ARCA system interface for submitting an item. At the top, there is a navigation bar with buttons: **Descrever** (highlighted in red), **Carregar**, **Verificar**, **Licenciar**, and **Completar**. Below the navigation bar, the main heading is **Depósito: Descrever seu item**. The instructions state: "Preencha as informações sobre seu documento. (Ajuda...)" and "Entre com as palavras chave ou frases relevantes." The **Palavras-chave** field contains the text "ACMCCS98::Genera" and "Categorias de assunto". A red arrow points to the "Categorias de assunto" link. To the right of the text input is a "Mais" button. Below the keywords field, there are three more input fields: **Resumo** (with the instruction "Entre com o(s) resumo(s) do Item."), **Patrocinador** (with the instruction "Entre com o patrocinador, agência de Fomento ou nome da bolsa."), and **Descrição** (with the instruction "Entre com outra descrição ou comentários neste campo.").

Seção 1.02 Figura 7: Tela de submissão de dados no sistema ARCA fazendo uso do ACMCCS

Ao clicar na opção Categorias de Assunto, o sistema exibe toda a estrutura do vocabulário controlado e também permite que se faça uma busca dentro do vocabulário para se encontrar o termo desejado.



Figura 8 : Tela de escolha do assunto fazendo uso do ACMCCS

3.3.2 DeCS ou MeSH

O vocabulário estruturado e trilingue *DeCS*¹¹ foi criado pelo Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (BIREME), a partir do *MeSH - Medical Subject Headings* da *U.S. National Library of Medicine*, com o objetivo de permitir o uso de uma terminologia controlada, tanto na indexação (entrada de dados), como na pesquisa de assuntos (recuperação) da literatura técnica e científica.

Assim como vimos no exemplo do vocabulário da *ACMCCS*, existe a possibilidade de incorporar o vocabulário *DeCS* ou *MeSH* no sistema *ARCA*, bastando para isso estudar e adaptar seu código *XML*, o que não é trivial, para que o *DSpace* entenda seu funcionamento.

Na figura abaixo podemos ver como o *DeCS* se apresenta virtualmente:

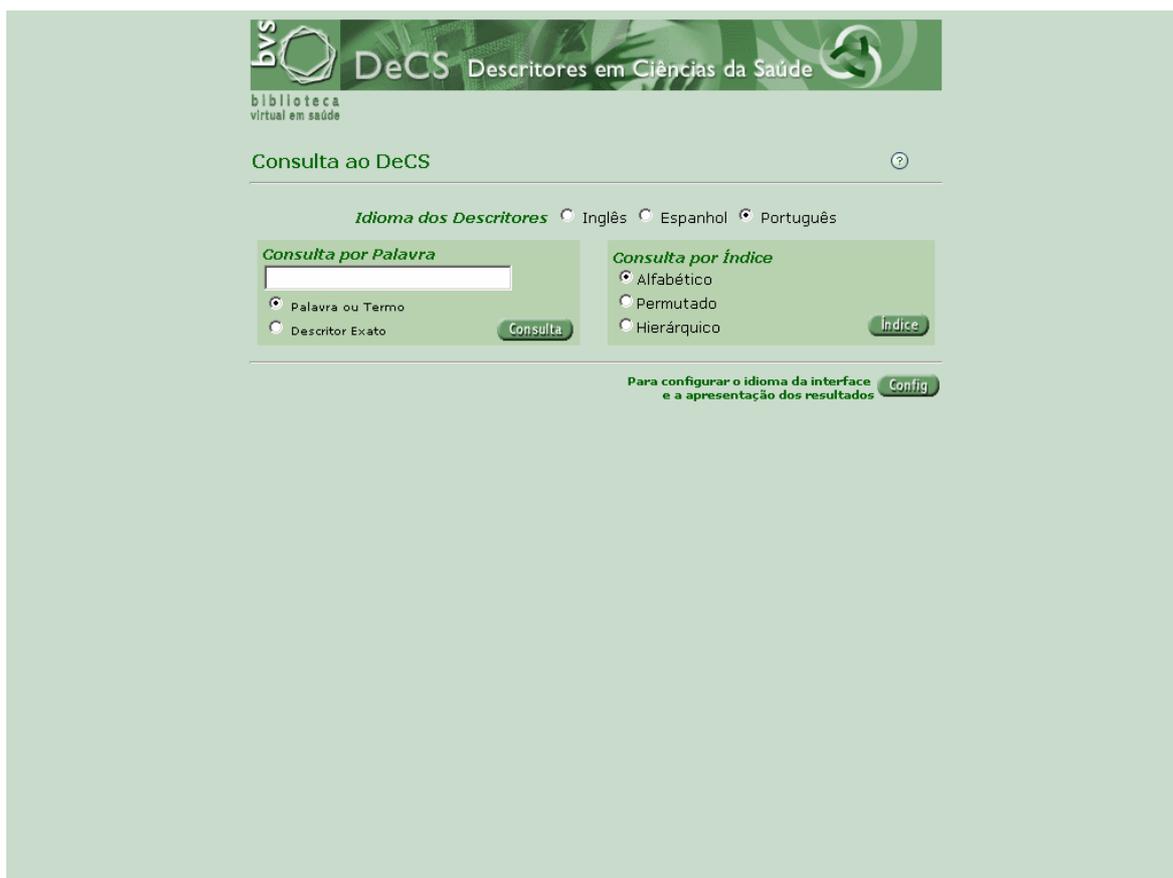


Figura 9: Tela inicial de consulta do DeCS

¹¹ Pode ser acessado em: <http://decs.bvs.br>

3.3.3 VOCABULÁRIO CONTROLADO CUSTOMIZADO

É possível através do elemento **Subject Other** do conjunto metadados *Dublin Core Qualificado* se implantar um vocabulário controlado customizado local, que contemple diversas áreas do conhecimento e que também ofereça a opção de utilizar um ou mais vocabulários de forma independente.

Apenas deve-se ter o cuidado de obedecer a estrutura do XML da ferramenta *DSpace* para evitar posteriores erros de leitura pelo sistema.

A estrutura é baseada no modelo XML a seguir:

```
<node id="nome do vocabulário" label="sigla">
<isComposedBy>
  <node id="A." label="Tema A">
    <isComposedBy>
      <node id="A.0" label="Tema A.0"/>
      <node id="A.1" label="Tema A.1"/>
      ...
    </isComposedBy>
  </node>
  ...
</isComposedBy>
</node>
```

Conforme dito anteriormente, a idéia contida neste vocabulário customizado é criar um vocabulário que não atenda somente aos tópicos relacionados à área da saúde. Um dos objetivos é possibilitar que outras áreas também sejam contempladas, tornando o repositório mais abrangente e funcional suprimindo às demandas do Instituto tanto na submissão quanto na recuperação da informação.

4 OBJETIVOS

4.1 GERAL

Implantar no sistema ARCA (Repositório da Memória Digital do Instituto de Informação Científica e Tecnológica da Fundação Oswaldo Cruz) um vocabulário controlado para que se possa padronizar o metadado assunto e recuperar com maior precisão os documentos depositados na plataforma.

4.2 ESPECÍFICOS

A realização do estudo inclui os seguintes objetivos específicos:

- a. Ampliar o escopo do repositório para além das informações em saúde;
- b. Aumentar a qualidade da representação das informações no repositório;
- c. Aumentar o nível de padronização do repositório;
- d. Consolidar um vocabulário que garanta a preservação da memória da Instituição e facilite a gestão do repositório.

4.3 JUSTIFICATIVA

O vocabulário controlado aparece como opção para sanar os problemas apresentados pelo sistema em relação ao metadado assunto, pois ofereceria termos pré-estabelecidos para serem inseridos na plataforma. É importante mencionar que o vocabulário controlado não engessa a entrada de termos, mas para que isso não ocorra é necessário uma atualização constante do vocabulário para que ele possa cada vez mais abranger o conjunto de termos que se deseja alcançar em uma determinada área.

Uma outra solução encontrada é inserir na plataforma mais de um vocabulário controlado para que se alcance um número maior de termos. Porém deve-se ter o cuidado de não transformar o sistema numa coleção de vocabulários controlados gerando assim uma dificuldade no preenchimento e escolha dos termos. A utilização do vocabulário controlado irá atuar diretamente nesta inserção dos termos no campo assunto e na respectiva busca dentro do Repositório Institucional ARCA.

4.4 RESULTADOS ESPERADOS

Os resultados esperados abrangem a implementação e a customização de uma solução de vocabulário controlado no Sistema ARCA, tendo como finalidade aperfeiçoar o sistema de busca atual e aprimorar a gestão da informação técnico-científica gerada pelos pesquisadores do ICICT.

Em consequência deste aperfeiçoamento se espera também a disseminação e o acesso à produção científica da instituição, ampliando assim, a visibilidade da mesma e de seus pesquisadores garantindo a preservação da memória científica e tecnológica.

5 ETAPAS DE DESENVOLVIMENTO

A metodologia de desenvolvimento deste trabalho consiste em uma abordagem tecnológica do sistema ARCA que aborda uma revisão da literatura especializada sobre a solução *DSpace* que é a tecnologia por detrás do sistema, e testes de diferentes tipos de vocabulários controlados para obtenção de resultados no processo da recuperação da informação respeitando os seguintes passos:

1. Efetuar testes com vocabulários controlados padronizados para a solução;
2. Efetuar testes de compatibilidade do *DeCS* com a solução;
3. Customizar os arquivos *XML* de estrutura do vocabulário para adaptar no ARCA;
4. Avaliar o processo de recuperação da informação no ARCA com e sem a utilização do vocabulário controlado.

6 REFERÊNCIAS CONSULTADAS

1. ACM – Association for Computing Machinery. Disponível em: <http://www.acm.org>. Acesso em: 21 ago. 2009.
2. CAFÉ, Lígia; MELO, Bianca Amaro de; BARBOSA, Elza Maria Ferraz et al. **Repositórios institucionais**: nova estratégia para publicação científica na Rede. In: INTERCOM - CONGRESSO BRASILEIRO DE CIÊNCIAS DA COMUNICAÇÃO, 26., 2003, Belo Horizonte. **Anais...** Belo Horizonte: [s.n.], 2003.
3. CAMPOS, M. L. A.; GOMES, HAGAR E. Tesouro e normalização terminológica: o termo como base para intercâmbio de informações. **Datagramazero**, Rio de Janeiro, v. 5, n. 6, 2004. Disponível em: http://www.dgz.org.br/dez04/Art_02.htm. Acesso em: 15 ago. 2007.
4. CROW, Raym. **The Case for Institutional Repositories**: A SPARC position paper. The Scholarly Publishing & Academic Resources Coalition 21 Dupont Circle. Washington, DC 2003. Disponível em: <http://www.arl.org/sparc/IR/ir.html>. Acesso em: 15 set. 2007.
5. DCMI – Dublin Core Metadata Initiative. Disponível em: <http://www.dcmi.org>. Acesso em: 13 set. 2009.
6. DSPACE WEB SITE. Disponível em: <http://www.dspace.org>. Acesso em: 13 set. 2009.
7. DSPACE ONLINE TECH FÓRUM. Disponível em: <https://lists.sourceforge.net/lists/listinfo/dspace-tech>. Acesso em: 25 ago. 2009.
8. GNU, O que é software livre? Disponível em: <http://www.gnu.org/philosophy/free-sw.pt.html>. Acesso em 05/08/2009.
9. HALEVY, A. Y.; IVES, Z. G.; MORK, P. Piazza: Data Management Infrastructure for Semantic Web Applications. Anais da 2ª Conferência Internacional da World Wide Web (WWW 03), W3C, Budapeste, Hungria, Maio, 2003.
10. LÉVY, Pierre. **As tecnologias da inteligência**: o futuro do pensamento na era da informática. Rio de Janeiro: Ed. 34, 1993.
11. LOPES, Ilza Leite. Estratégia de busca na recuperação da informação: revisão da literatura. **Ci. Inf.**, Ago 2002, vol.31, no.2, p.60-71. ISSN 0100-1965
12. 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 29/09/2007.

13. MARCONDES, Carlos Henrique. Metadados: descrição e recuperação de informação na Web. In: MARCONDES, Carlos H. et al. (orgs.). **Bibliotecas digitais: saberes e práticas**. Salvador/Brasília: UFBA/IBICT, 2005. p. 97-113.
14. MARCONDES, Carlos Henrique; SAYÃO, Luis Fernando. Documentos Digitais e novas formas de Cooperação entre Sistemas de Informação em C&T. **Ciência da Informação**, Brasília, v. 31, n. 3, p.42-54, set./dez. 2002.
15. MÁRDERO ARELLANO, M. A. **Repositórios Institucionais DSpace**. Brasília: Instituto Brasileiro de Informação em Ciência e Tecnologia, 2007. Disponível em: http://dspace.ibict.br/index.php?option=com_content&task=view&id=12&Itemid=97. Acesso em: 20 abril 2008.
16. MURAMAKI, Tiago Rodrigo Marçal. Tesouros e a World Wide Web. 2005. 92f. Trabalho de Conclusão de Curso - (Escola de Comunicações e Artes) - Universidade de São Paulo, 2005.
17. NOY, N. F.; McGUINNESS, D. L. Ontology Development 101: A Guide to Creating Your First Ontology. Disponível em: http://www.ksl.stanford.edu/people/dlm/papers/ontology101/*.html. Acesso em 15/09/2008.
18. OPEN ARCHIVES INITIATIVE - OAI. Disponível em: <http://www.openarchives.org>. Acesso em: 12 set. 2009.
19. SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, Belo Horizonte, v.1, n.1, p.41-62, jan./jun.1996.
20. SAYÃO, Luis Fernando. Integração e interoperabilidade no acesso a recursos informacionais em C&T: a proposta da Biblioteca Digital Brasileira. **Ciência da Informação**, Brasília, v. 30, n. 3, p. 24-33, set./dez. 2001. Disponível em: <http://www.scielo.br/pdf/ci/v30n3/7283.pdf>. Acesso em: 29 jul. 2009.
21. SHINTAKU, M.; BRÄSCHER, M. Dspace versão 1.4: uma análise das facilidades relacionadas ao assunto. In: SEMINÁRIO INTERNACIONAL DE BIBLIOTECAS DIGITAIS BRASIL, 2007, São Paulo. **Anais...** São Paulo, [s.n.], 2007.
22. SOUZA, R.R.; ALVARENGA, L. A Web Semântica e suas contribuições para a ciência da informação. **Ciência da Informação**, Brasília, v.33, n.1, p.132-141, jan./abril 2004.
23. XML, site oficial. Disponível em: <http://www.w3.org/XML>. Acesso em: 31/02/2006.
24. W3C, World Wide Web Consortium. Disponível em <http://www.w3.org>. Acesso em 02/02/2007.

8 ORÇAMENTO

O orçamento abaixo relaciona os recursos financeiros a serem utilizados ao longo de todo o projeto. Os itens relacionados e o material de consumo serão utilizados aqueles disponíveis no ICICT.

PROPOSTA DETALHADA DE ORÇAMENTO

DESCRIÇÃO	ITEM	QUANTIDADE	VALOR (R\$)	JUSTIFICATIVA
Bolsa de Pesquisa	Bolsa	1 (12 meses)	14.400,00	Bolsa por 6 meses, para pesquisador que irá desenvolver a pesquisa
Bolsa de Estágio	Bolsa	1 (12 meses)	6.200,00	Bolsa por 6 meses, para 1 estagiário de nível superior na área de análise de sistemas
Computador	Material Permanente	2	3.500,00	Aquisição de 2 computadores para o processamento técnico
Impressora	Material Permanente	1	300,00	Aquisição de 1 impressora
Banner	Serviço de Terceiros	2	200,00	Impressão de banner para divulgação e andamento do projeto, bem como, para ser usado em trabalhos científicos
Passagens	Passagem (total)	1	900,00	Passagens para apresentar resultados em evento científico.
Diárias	Diárias Totais	2	250,00	Diárias para apresentar resultados em evento científico