

# Instituto Politécnico do Porto

Escola Superior de Estudos Industriais e de Gestão

Ciências e Tecnologias da Documentação e Informação

Ano Lectivo 2006/7

# Bibliorandum

Projecto Final de

Projecto e Implementação de Sistemas de Comunicações e Informação

Discente:  
Júlio dos Anjos  
Nº 9040193

Vila do Conde  
2007

Docente:  
Engº Lino Oliveira

# Índice

1. Introdução .....	4
2. Fundamento do projecto.....	5
2.1. A designação.....	5
2.2. O objectivo.....	5
2.3. As ferramentas .....	5
2.4. Outras considerações .....	6
3. WebDesign.....	7
3.1. Funcionalidades .....	7
3.1.1. Pesquisa.....	8
3.1.2. Notícias.....	9
3.1.3. Blogosfera .....	9
3.1.4. Periódicos .....	11
3.1.5. Akademya.....	11
3.1.6. Feedback.....	12
3.2. Design Gráfico.....	13
3.3. Estética.....	13
3.4. Conteúdo.....	13
4. Bases de dados .....	15
4.1. Na funcionalidade Blogosfera .....	15
4.1.1. Tabela “Themes” .....	15
4.1.2. Tabela “Feeds” .....	16
4.1.3. Tabela “Items” .....	16
4.1.4. Tabela “Feeds_themes” .....	17
4.1.5. Os <i>queries</i> .....	17
4.2. Na funcionalidade Akademya.....	18
5. Conclusão.....	22
5.1. ... e os utilizadores?.....	22
5.2. ... e o futuro ? .....	22
6. Bibliografia .....	24
7. Anexos.....	25
7.1. Ferramentas e tecnologias.....	25

---

7.1.1.	OAI-PMH (Protocolo).....	25
7.1.2.	RSS.....	28
7.1.3.	Google Custom Search Engine.....	28
7.1.4.	PHP.....	30
7.1.5.	MySQL.....	30
7.1.6.	APACHE.....	31
7.1.7.	Open Archives Initiative - Repository Explorer.....	31
7.1.8.	PKP Open Archives Harvester.....	32
7.1.9.	A actualização da informação da “blogosfera”.....	32
7.1.10.	A escolha de repositórios em “Akademya”.....	33
7.1.11.	A actualização da informação de “Akademya”.....	34
7.2.	Listas.....	35
7.2.1.	Repositórios.....	35
7.3.	Imagens.....	36
7.4.	Anexos soltos.....	40
7.4.1.	Relatórios Google Analytics.....	40

# 1. Introdução

O presente relatório documenta a construção do sítio internet a que foi dada a designação de trabalho “**Bibliorandum**”, desenvolvido como concretização do trabalho final da disciplina de Projecto de Implementação de Sistemas de Comunicação e Informação

Este relatório focará os seguintes aspectos: o objectivo do sítio e funcionalidades planeadas e concretizadas; soluções em termos de *webdesign* implementadas; a utilização feita de sistemas de gestão de bases de dados e SQL, tanto trabalho original como aproveitamento das bases de dados mantidas por outros programas que alimentam várias das funcionalidades. Após a conclusão, onde termina o relatório das actividades directamente ligadas com as competências em avaliação, é feita uma análise, nalguns casos aprofundada, dos protocolos subjacentes a algumas funcionalidades bem como das ferramentas utilizadas para a colecção de dados.

## 2. Fundamento do projecto

### 2.1. A designação

A designação “**bibliorandum**” foi, como muitos outros projectos hoje em dia, ditada por ser a única designação contendo o sufixo “*randum*” que tivesse uma raiz relacionada com as profissões da documentação-informação (esta escolha devida ao ponto 2. do enunciado), e que tivesse nome de domínio livre a nível de *.com*, *.net* e *.org*,<sup>1</sup>

### 2.2. O objectivo

O sítio desenvolvido tem por objectivo servir como um “Sistema de Informação Corrente em Biblioteconomia e Ciências da Informação”, frase onde encontramos as duas directrizes orientadoras das funcionalidades desenvolvidas: O produto é “Informação Corrente”, e o público-alvo é “Biblioteconomia e Ciências da Informação”. Portanto o objectivo do **bibliorandum** é prestar serviços de informação corrente que a membros da comunidade com interesse em Informação Corrente na área da Biblioteconomia e Ciências da Informação

### 2.3. As ferramentas

O **bibliorandum** foi desenvolvido na sua esmagadora maioria em linguagem de programação **PHP** (7.1.4) executada sob um servidor http **Apache**(7.1.6), sendo o SGBD utilizada o **MySQL**(7.1.5). Estas ferramentas encontram-se instaladas num servidor **Linux** (distribuição **Ubuntu**, uma variação de Debian). Apesar de ter sido necessário durante o desenvolvimento mudar para um computador completamente estranho ao ambiente de desenvolvimento original foi possível fazê-lo praticamente qualquer contratempo. Estas ferramentas e outras mais específicas, que serão focadas ao longo do relatório, encontram-se documentadas no ponto 7.1 dos anexos, na página 25 e seguintes. Uma única funcionalidade é dependente de um *script* em linguagem **PERL**

Outras ferramentas e programas usados como colectores de informação serão focados em tempo oportuno conforme a sua aplicação em cada funcionalidade.

---

<sup>1</sup> Em poucas palavras: “**informandum.com**” já estava ocupado.

## 2.4. Outras considerações

Os nomes de domínio foram registados através do serviço **godaddy** sendo o sítio acessível pelo endereço <http://www.bibliorandum.net>

## 3. WebDesign

### 3.1. Funcionalidades

Será quase um lugar comum afirmar que estamos no meio de uma explosão de informação à disposição de qualquer pessoa com ligação aos recursos da internet. É, aliás, praticamente impossível verificar quem foi a primeira pessoa a declarar que saímos da info-escassez e entramos na info-abundância (será que o termo é “info-obesidade”). Não só a quantidade de informação disponível é cada vez maior mas a quantidade de canais que a disponibilizam não pára de aumentar. Veja-se os números de sítios disponíveis na internet<sup>2</sup> bem como de blogues<sup>3</sup> editados nas mais variadas línguas sobre (atrevo-me a afirmá-lo sem qualquer base ou citação) todos os interesses possíveis da espécie humana (para não falar de todas as perspectivas possíveis sobre esses interesses).

Um dos mecanismos que podem ser aplicados para diminuir o nível de ruído incontornável no uso da internet é a criação de condições de surdez selectiva, ou seja a pré selecção de recursos. Os técnicos de documentação fazem-no à séculos: a procura, selecção e organização de recursos informativos específicos a um determinado público é, por assim dizer, uma segunda (e para alguns uma primeira) definição de profissional da informação.

Assim as funcionalidades do **bibliorandum** representam diferentes maneiras de criar esta surdez selectiva (selecção de recursos) . Na funcionalidade **pesquisa** apenas determinados sítios da internet é recuperado, na funcionalidade **notícias** apenas um conjunto de notícias que contém determinadas palavras é apresentado, na funcionalidade **blogosfera** apenas as notícias de alguns blogues são apresentados, na funcionalidade **periódicos** é feita uma selecção prévia de revistas das quais a apresentar e, por fim, na funcionalidade **Akademya** é feita uma selecção de repositórios digitais, e nalguns destes de colecções de documentos, cujas referências são apresentadas. Todas as selecções são subjectivas e da responsabilidade do autor do trabalho. Mas na verdade um dos trabalhos do profissional de informação é precisamente decidir o que tem valor suficiente para ser incluído de modo a que a perda de informação do que fica de fora da colecção

---

<sup>2</sup> Os numeros da Netcraft (empresa especializada na análise de presença de vários http servers na internet) para o mes de Junho da apontam para a existência de 122,000,635 servidores anunciando placidamente que é uma aumento de 3.97 milhões em relação a Maio. Netcraft: June 2007 Web Server Survey. [http://news.netcraft.com/archives/2007/06/08/june\\_2007\\_web\\_server\\_survey.html](http://news.netcraft.com/archives/2007/06/08/june_2007_web_server_survey.html)

<sup>3</sup> O veneável Technorati apontava a existência de 70 milhoes de blogues e 17 notícias publicadas por dia Technorati Weblog: The State of the Live Web, April 2007 ; <http://technorati.com/weblog/2007/04/328.html>

seja diminuída. Nesse aspecto este projecto está perfeitamente dentro dos parâmetros (ver referencial de profissões francês e referencial de competências ID português).

Aos escolher recursos (sítios, blogues, repositórios, etc) considerando as necessidade informativas de um publico específico estou a criar uma biblioteca especializada num determinado tema, um Centro de Documentação e Informação.

### 3.1.1. Pesquisa

Esta é a funcionalidade de página de entrada e programaticamente corresponde a duas páginas. Uma, que é apresentada quando se acede ao sítio por <http://www.bibliorandum.net>, que não tem outra função que apresentar o formulário de recolha de pesquisa, e outra página (`resultado.php`) em que os resultados são apresentados e o formulário é re-exibido com os termos preenchidos na página inicial.

Esta funcionalidade usa o serviço **Custom Search Engine**<sup>4</sup> da **Google**, que era o único sistema, à data de arranque do projecto, a permitir esta funcionalidade<sup>5</sup>. O botão de pesquisa em português é acrescentado pelo autor.

Esta funcionalidade foi implementada para demonstrar a possibilidade de usar um motor de pesquisa não específico (horizontal) numa situação específica (vertical), uma vez devidamente parametrizado; a outra demonstração realizada é a capacidade de configuração e *branding* que se revela possível ao poder trazer essas capacidades ao nosso produto embutindo segmentos de código (*javascript* e *html*) as páginas por nós produzidas.

Sendo os resultados da pesquisa filtrados contra uma lista de endereços de sítios e de páginas os resultados apresentados têm origem em sítios que serão à partida mais relevantes para a temática a que a escolha do filtro obedeceu, neste caso: sítios de bibliotecas e arquivos bem como sítios relacionados com estes<sup>6</sup>.

---

<sup>4</sup> Doravante referenciado como CSE

<sup>5</sup> Existiam outros que não permitiam a apresentação remota do formulário de pesquisa ou dos resultados como o **Rollyo** ([www.rollyo.com](http://www.rollyo.com)), e de então para cá a **yahoo** lançou o **Yahoo Search Web Services** (<http://developer.yahoo.com/search/>). Este último apesar de considerado na altura da sua divulgação pública foi preterido depois de se constatar haver grandes hiatos no leque de sítios portugueses, relevantes para o **bibliorandum**, que não eram, na altura, *crawled* pelos *bots* da yahoo.

<sup>6</sup> IPLB – agora: Direcção-Geral do Livro e das Bibliotecas, RBE - Rede de Bibliotecas Escolares.etc não são sítios de bibliotecas mas são sítios relevantes para o público alvo

Para uma descrição mais completa dos passos necessários para configurar este CSE consultar o ponto 7.1.3 na página 28.

### 3.1.2. Notícias

Esta funcionalidade materializa-se uma única página em que são apresentadas notícias recentes respeitantes a bibliotecas que tenham sido apanhadas na selecção de serviços noticiosos *online* dos dois maiores agregadores de notícias *online* disponíveis ao público em geral, o Google e o Yahoo.

Para isso foi construída uma pesquisa que depois de adaptada às idiossincrasias de notação booleana de cada um dos serviços foi capturada em formato RSS<sup>7</sup>. A linha de execução do pedido de informação foi depois transformada por um utilitário web de conversão de RSS em *javascript* e o código resultante foi embutido na página “Notícias” (*noticias.php*).

Ambos os “canais” (um de notícias descobertas pelo google outro de notícias descobertas pelo yahoo) são apresentados lado a lado.

Esta funcionalidade foi implementada com o objectivo de demonstrar que é possível personalizar um agregador de notícias de modo a criar um ‘canal virtual’ e que os resultados podem ser embutidos numa página *html* sem qualquer programação a nível de servidor, pois todo o esforço é realizado pelo *javascript* fornecido pelo serviço conversor<sup>8</sup>.

As pesquisas implementadas são apenas demonstrativas estando previsto reformular o serviço noutros moldes.

### 3.1.3. Blogosfera

Esta funcionalidade faz a agregação e apresentação de notícias publicadas em blogues da área das ciências da informação em Portugal apresentando aos utilizadores os 7 dias mais recentes de informação ou as 30 notícias mais recentes,, o que for maior.

O elenco de blogues processados é da inteira responsabilidade do autor do projecto sendo resultado de um levantamento realizado a título pessoal seguindo as teias de relações inter-blogues

---

<sup>7</sup> Acrónimo com vários significados conhecidos *Really Simple Syndication* (RSS 2.0), *RDF Site Summary* (RSS 1.0 e RSS 0.90) e *Rich Site Summary* descrito em mais pormenor no ponto 7.1.2, na página 28

<sup>8</sup> No nosso caso o serviço é o **Feed2JS** disponível em <http://feed2js.org/>

explicitamente colocadas nas funcionalidades de *blogroll* e analisando a relevância do conteúdo para o público alvo do **bibliorandum**. Grosso modo, a inclusão nesta lista era determinada conforme cada blogue em análise não apresentava uma única entrada, nas 10 mais recentes, sobre temática biblioteconómica/arquivística/ciências da documentação/ciências da informação (caso dos blogues de “promoção da leitura”, blogues de professores bibliotecários mais preocupados com a docências que com a biblioteconomia, blogues de editores ou livreiros, etc).

Os blogues encontram-se organizados em categorias generalistas sem pretensões de representar uma taxionomia com qualquer fundamento científico.

As entradas publicadas são apresentadas por ordem decrescente de data e hora de edição, sendo separadas visualmente em blocos diários.

São disponibilizados canais RSS específicos a cada categoria para sindicância por qualquer interessado e todas as categorias apresentam a possibilidade de exportar a lista dos blogues nelas incluídas num ficheiro OPML.

Algumas categorias apresentam material (blogues e respectivas entradas) não-português entrelinhado com material português; outras categorias não apresentam mesmo nenhuma entrada portuguesa, tendo a presença da categoria um propósito mais pedagógico que útil/funcional, isto é, como convite ao aparecimento de blogues portugueses na categoria. No entanto a categoria CPLP<sup>9</sup> é permanentemente dedicada a agrupar os blogues de profissionais de outros países que não Portugal.

Foi também criado um CSE especificamente dedicado aos blogues sendo alimentado em exclusivo com endereços dos blogues Portugueses presentes nesta funcionalidade. A este CSE foi agregada a definição que permite registar este CSE como *plugin* de pesquisa instalável em IE 6.0+ e Firefox 2.0+. Para isso foi investigada<sup>10</sup> a especificação OpenSearch 1.1, o que resultou na construção do seguinte ficheiro *xml*:

---

<sup>9</sup> CPLP: Comunidades de Países de Língua Portuguesa

<sup>10</sup> OpenSearch em <http://www.opensearch.org/>

```
<?xml version="1.0" encoding="UTF-8"?>
<OpenSearchDescription xmlns="http://a9.com/-/spec/opensearch/1.1/">
  <ShortName>Bibliorandum Blog Search</ShortName>
  <Description>Pesquisa na Biblio'Blogosfera Portuguesa</Description>
  <Tags>blogs portugueses</Tags>
  <Contact>julio-9040193@janjos.com</Contact>
  <Url type="text/html"
  template="http://www.bibliorandum.net/blog_resultado.php?cx=000675947084576693848:zsikhca
  rode&amp;q={searchTerms}&amp;sa=Search&amp;cof=FORID%3A10&amp;ie=utf-8#975" />
  <LongName>Bibliorandum Blog Search Engine via Google</LongName>
  <Image height="16" width="16"
  type="image/vnd.microsoft.icon">http://www.bibliorandum.net/favicon.ico</Image>
  <Query role="example" searchTerms="test" />
  <Developer>Julio dos Anjos</Developer>
  <Attribution>
    Google Co-op Custom Search Engine
  </Attribution>
  <SyndicationRight>open</SyndicationRight>
  <AdultContent>>false</AdultContent>
  <Language>pt-pt</Language>
  <OutputEncoding>UTF-8</OutputEncoding>
  <InputEncoding>UTF-8</InputEncoding>
</OpenSearchDescription>
```

Existe também uma caixa de texto que convida a registar novos blogues<sup>11</sup>.

### 3.1.4. Periódicos

Esta funcionalidade apresenta o conteúdo de revistas publicadas por mecanismos de acesso livre, ou pelos menos exploráveis por OAI-PMH pois algumas podem expor os metadados em OAI mas não permitir o acesso directo aos conteúdos.

No entanto ao longo do trabalho foi notável que este serviço não é viável pois as políticas de metadados de cada repositório, e por vezes entre várias revistas no mesmo repositório (o sciello brasileiro é um caso impossível) não permitem o tratamento de todas as revistas de igual modo ao nível unidade de cada edição/*issue* (ou sua metáfora digital), exigindo trabalho manual para afinar os diversos dados. Este serviço a curto prazo será de alguma maneira incorporado na funcionalidade **akademya**.

Durante o desenvolvimento do projecto funcionou com uma instalação autónoma do

### 3.1.5. Akademya

Esta funcionalidade apresenta os documentos (títulos, autores e resumos) recentemente registados em repositórios de acesso livre. Os registos são recolhidos por meio de protocolo OAI-

---

<sup>11</sup> Os endereços e dados blogues aqui registados são neste momento enviados ao autor por email que os regista manualmente na base de dados subjacente ao serviço. Está prevista a integração com uma plataforma de catalogação de recursos *online* (<http://blogssubmit.bibliorandum.net>) numa fase posterior

PMH<sup>12</sup> de uma lista seleccionada de repositórios de acesso livre, sendo que nalguns dos repositórios apenas alguns segmentos temáticos são recolhidos. A apresentação é feita por ordem decrescente de data de registo mas como a maior parte dos repositórios apenas tem granularidade a nível de dia, e não de hora, na verdade a ordenação é por data de registo e número de série dentro do programa de recolha<sup>13</sup>. Os registos apresentados podem ser filtrados por registos em língua portuguesa, por registos em línguas latinas e finalmente por registos em línguas não latinas<sup>14</sup>. É ainda possível aceder a um canal RSS que exporá, a qualquer altura que seja consultado, os 10 registos mais recentes da totalidade dos documentos registados.

Foi previsto mas não implementada a possibilidade de subscrever alertas por email para um relatório diário de novos documentos. Estamos em crer que os mecanismos de alerta por RSS estão a ter mais aceitação de momento e os utilizadores podem, junto do agregador/leitor da sua preferência ou de outros serviços grátis disponíveis na internet, criar mecanismos de alerta com a mesma funcionalidade ou melhor ainda<sup>15</sup>.

É também disponibilizado um mecanismo de pesquisa no conteúdo dos registos conhecidos pelo **bibliorandum**<sup>16</sup>

### 3.1.6. Feedback

Esta funcionalidade permite o envio de comentários ao autor do projecto, recolhendo o nome, endereço de correio electrónico e comentário do utilizador. É realizado em três passos: um em que os dados são preenchidos, e outro em que os dados são apresentados sujeitos a correcção ou

---

<sup>12</sup> OAI-PMH: Open Archives Initiative Protocol for Metadata Harvesting, é descrito no ponto 7.1.1, na página 25

<sup>13</sup> O programa de recolha é o PKP Open Archives Harvester descrito no ponto 7.1.8 na página 32. A metodologia de actualização e recolha de registos é descrita no ponto 7.1.11

<sup>14</sup> Teria sido interessante e relevante fazer também a filtragem por país de publicação, no entanto o Dublin Core, que é o formato mínimo obrigatório de exposição de registos em OAI-PMH, não apresenta campo para catalogar o atributo “país de publicação”. Os repositórios que implementam esse campo fazem-no internamente e se o expõem nalgum formato complementar ao Dublin Core cada um usa um formato diferente, pelo que se a tarefa se revela impossível quando estamos a tentar agregar informação dos mais diversos repositórios motorizados pelos mais diversos programas.

<sup>15</sup> Apesar de não os conhecer directamente deve ser possível converter um RSS em Twitter, alertar via SMS ou alimentar um ecrã numa parede no SecondLife.

<sup>16</sup> Esta funcionalidade baseia-se em executar um programa (`labrao_de_xml.php`) ao fim do dia, que analisa todos os documentos conhecidos da base de dados relevante e se não existir uma cópia local do documento em formato XML (em `bibliorandum/oai/`), é criada uma linha num ficheiro *shell* (`bibliorandum/oai/wget_oais.sh`) que realiza a recolha do registo em XML. Ao fim deste programa o programa *shell* por ele criado é executado e os registos bibliográficos são transferidos para o disco local a partir de cada um dos repositórios. Por fim o programa de actualização do motor local de pesquisa de texto integral (`swish-e`) é executado fazendo que o conteúdo dos documentos se torne visível ao realizar a pesquisa.

confirmação e um terceiro de registo em que o *email* com os dados é efectivamente enviado. Neste passo, o utilizador, ao qual se agradece o tempo empregue no feedback é reenviado automaticamente para a página onde se encontrava quando acedeu à funcionalidade *feedback* <sup>17</sup>

### 3.2. Design Gráfico

Foi criado um símbolo para identificação do sítio serviço logo no início do projecto:



**Ilustração 1: A imagem de marca do serviço**

O símbolo foi criado com um gerador de imagens e foi usado para criar uma identificação propositada com as funcionalidades do motor de pesquisa Google.

Em rodapé estão presentes os símbolos da ESEIG, do curso CTDI e da Instituição que ao serviço da qual o sítio será colocado depois de avaliado.

A imagética de *links* é muito simples e os *links* de navegação entre as várias funcionalidades são mantidos constantes ao longo destas devido à partilha da rotina de criação tanto do cabeçalho como do rodapé. Nota-se no entanto que entre as várias funcionalidades não foi mantida uma mesma linha gráfica o que tem raiz na natureza distinta das necessidades de cada funcionalidade.

### 3.3. Estética

O sítio pretende ter impacto pelos conteúdos e funcionalidades únicas apresentados, o que não invalida uma preocupação com a sua estética que será alterada regularmente conforme outras pessoas mais capazes nessa área possam ser envolvidas.

### 3.4. Conteúdo

Dadas as tecnologias utilizadas e todas as funcionalidades serem agregação de material produzido noutros sítios é possível afirmar que este sítio não tem conteúdo próprio absolutamente

---

<sup>17</sup> Por “<meta http-equiv="refresh" content="5;...>”

nenhum, na verdade apenas tem funcionalidades para re-apresentar o conteúdo de terceiros<sup>18</sup>. Isso tem por reverso que está sempre tão vivo e actual como a comunidade que de várias maneiras , entre elas o **bibliorandum**, se aucta e reage de volta à mesma comunidade. Os desenvolvimentos possíveis a nível de conteúdo são vislumbrados principalmente em termos de novas funcionalidades e novas tecnologias de agregação e aperfeiçoamento das funcionalidades presentes. A funcionalidade de registo de novos blogues, já falada, (ou novas páginas de qualquer tipo desde a desactivação do DIRECTÓRIO DE BIBLIOTECAS - PORTUGAL<sup>19</sup>) faz parte desse vector. Também existe o projecto de adaptação do software que motoriza o serviço “do melhor.net” para a criação de um serviço semelhante ao *digg* mas para registar informação, notícias, e porque não artigos, de interesse para o publico-alvo do **bibliorandum**.

---

<sup>18</sup> Incidentalmente o google funciona precisamente no mesmo paradigma

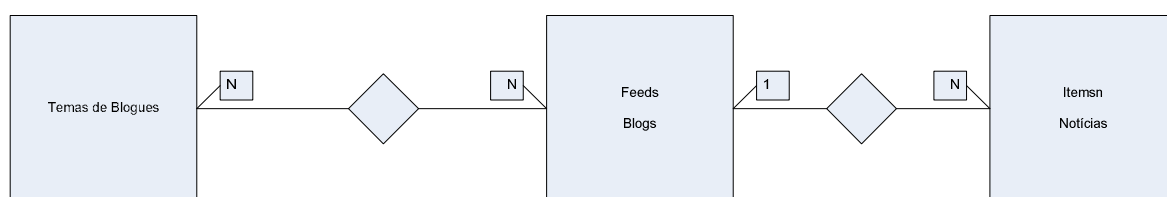
<sup>19</sup> <http://bibliotecas.wetpaint.com/>

## 4. Bases de dados

### 4.1. Na funcionalidade Blogosfera

O trabalho de base de dados realizado nesta funcionalidade apesar de ser o menos complexo do projecto é o único que é original de raiz pelo que é objecto de um especial tratamento no relatório.

Foi criada a base de dados **bibliorandum\_feeds** com quatro tabelas para fazer a gestão da recolha de notícias e entradas publicadas em blogues (e nalguns casos recolhidas de outros agregadores).



**Ilustração 2: Tabelas e Relações**

#### 4.1.1. Tabela “Themes”

Contém os dados relacionados com as categorias de blogues e outros serviços apresentados nesta funcionalidade:

Column Name	Data Type	Primary Key	Not Null	AutoInc	Comment
theme_id	int(10)	YES	YES	YES	
theme_title	varchar(45)	NO	YES	NO	Texto a apresentar para selecção desta categoria
theme_desc	text	NO	YES	NO	Descrição para mouseover sobre o texto acima
google_coop_id	varchar(100)	NO	YES	NO	Valor para atributo Name do elemento Label do ficheiro de contexto do CSE
google_cse_label	varchar(45)	NO	YES	NO	Valor para atributo Title do elemento FacetItem do CSE

Index Name	Columns
PRIMARY	theme_id

### 4.1.2. Tabela “Feeds”

Contem os dados relacionados com cada blogue, ou outro tipo de serviço, apresentado nesta funcionalidade

Column Name	Data Type	Primary Key	Not Null	AutoInc	Comment
feed_id	int(10)	YES	YES	YES	
site_title	varchar(150)	NO	YES	NO	Titulo do Recurso
site_url	varchar(150)	NO	YES	NO	Endereço Humano do Sitio ou Canal
site_feed_url	varchar(150)	NO	YES	NO	Endereço XML do Sítio ou do Canal
feed_enabled	tinyint(1)	NO	YES	NO	Deve ser extraído um feed deste canal? Pode por exemplo estar avariado (servidor que não dá timeout provocando timeout local)
skip_google_coop	tinyint(1)	NO	YES	NO	Deve este sitio ser colocado no CSE quando se pesquisa nos blogues?
google_cse_path	varchar(150)	NO	NO	NO	Qual o "path" a ser requerido ao google? tipo http://www.bn.pt/* , http://www.cm-viladoconde/bib*
lista_main	tinyint(1)	NO	NO	NO	Deve este Blogue ou Sitio ser apresentado na página principal?
portugues	tinyint(1)	NO	NO	NO	Um blogue ou sítio pode ser português e levar a cor de apresentação a azul vivo, no entanto não ser de biblioteconomia, e portanto não ir para a página principal
site_desc	Longtext	NO	YES	NO	Para benefício exclusivo do Catalogador de Blogues

Index Name	Columns
PRIMARY	feed_id
Index 2	site feed url

### 4.1.3. Tabela “Items”

Contem todos os itens publicados nos blogues e serviços enumerados em na tabela feeds. Todas as ocorrências de feed\_id desta tabela têm de existir no campo homónimo da tabela feeds., tendo portanto relação N->1 com a tabela feeds.

Column Name	Data Type	Primary Key	Not Null	AutoInc	Comment
feed_id	int(10)	YES	YES	NO	A que canal pertence esta entrada
date_published	datetime	YES	YES	NO	Qual a data de publicação
item_url	varchar(255)	YES	YES	NO	Qual a URL, como temos de apontar para a URL da entrada este campo tem de ser parte da chave primária
item_id	int(10)	NO	YES	YES	

date_recieved	datetime	NO	YES	NO	Data em que esta entrada foi registada neste serviço
item_title	varchar(150)	NO	YES	NO	Título da entrada
item_author	varchar(150)	NO	YES	NO	Autor se declarado
enclosure_url	varchar(255)	NO	YES	NO	Endereço do anexo multimédia

Index Name	Columns
PRIMARY	feed_id,date_published,item_url
Index 3	item id
Index 4	date published

#### 4.1.4. Tabela “Feeds\_themes”

Relaciona em N<->N as tabela Feeds e Themes, já um blogue pode ser enquadrado em várias categorias (Themes)

Column Name	Data Type	Primary Key	Not Null	AutoInc	Flags	Comment
feed_id	int(10)	YES	YES	NO	UNSIGNED	Qualquer vaor neste campo tem de estar presente na coluna homónima do da tabela feeds
theme_id	int(10)	YES	YES	NO	UNSIGNED	Qualquer vaor neste campo tem de estar presente na coluna homónima do da tabela themes

Index Name	Columns
PRIMARY	feed_id,theme_id
FK_Feeds_Themes_2	theme_id

#### 4.1.5. Os queries

Comando para retornar todos os blogues portugueses

```
select max(date_published) as mais_recente,
       feeds.*
from items join Feeds_Themes on
           (Feeds_Themes.feed_id=items.feed_id)
       join feeds on
           (feeds.feed_id=items.feed_id)
group by items.feed_id having portugues=1
order by mais_recente desc;
```

Comando para retornar todos os blogues de determinada categoria

```
select max(date_published) as mais_recente, feeds.*
from items join Feeds_Themes
            on (Feeds_Themes.feed_id=items.feed_id)
join feeds
            on (feeds.feed_id=items.feed_id)
where Feeds_Themes.theme_id=20
group by items.feed_id
order by mais_recente desc;
```

Comando para retornar todas as notícias publicadas em blogues nos últimos 7 dias:

```
SELECT *, UNIX_TIMESTAMP(date_published) as rfc822
FROM bibliorandum_feeds.items
      join feeds on (items.feed_id=feeds.feed_id)
where  lista_main=1
      and date_published>=DATE_SUB(NOW(),INTERVAL 7 DAY)
      and date_published<=DATE_ADD(NOW(),INTERVAL 1 DAY)
order by date_published desc;
```

Comando para retornar todas as notícias publicadas em blogues nos últimos 7 dias em determinada categoria:

```
SELECT *, UNIX_TIMESTAMP(date_published) as rfc822
FROM bibliorandum_feeds.items
      join Feeds_Themes on (items.feed_id=Feeds_Themes.feed_id) join feeds on
(items.feed_id=feeds.feed_id)
where  theme_id=5
      and date_published>=DATE_SUB(NOW(),INTERVAL 7 DAY)
      and date_published<=DATE_ADD(NOW(),INTERVAL 1 DAY)
order by date_published desc;
```

Se por acaso o comando acima retornar menos de 30 notícias o comando é substituído por

```
SELECT *, UNIX_TIMESTAMP(date_published) as rfc822
FROM bibliorandum_feeds.items
      join Feeds_Themes on (items.feed_id=Feeds_Themes.feed_id) join feeds on
(items.feed_id=feeds.feed_id)
where  theme_id=5
order by date_published desc
limit 30;
```

A notável linha `and date_published<=DATE_ADD(NOW(),INTERVAL 1 DAY)` deve-se ao estranho facto de por vezes as notícias aparecerem com datas bastante no futuro mesmo nas plataformas de maior reputação. Este discente nunca conseguiu descobrir porque isto acontece mas em termos de usabilidade a tendência humana de olhar para o topo da pilha (e neste caso patentemente *above the fold*) para ver o que á de novo é completamente destruída.

## 4.2. Na funcionalidade Akademya

Os registos recolhidos pelo programa PKP Harvester (ver ponto 7.1.8) veem os seus dados distribuídos em duas tabelas: na base de dados `harvester2`, uma (`records`) com os dados básicos e outra (`entries`) com os dados complementares. É apenas uma questão de construir um comando

SQL que faça a extracção dos 25 registos mais recentes (ou registados nos últimos 7 dias), com ou sem limitação à língua em que os trabalhos originais se encontravam expressos, a partir da tabela records. O comando é apresentado no quadro seguinte:

```
SELECT          harvester2.records.datestamp,
                harvester2.records.record_id,
                harvester2.records.archive_id,
                harvester2.entries.value
FROM ( harvester2.records join harvester2.entries on
(harvester2.records.record_id = harvester2.entries.record_id and raw_field_id=1 )
WHERE harvester2.records.archive_id <> 1
AND harvester2.records.datestamp > DATE_SUB(NOW(),INTERVAL 7 DAY)
ORDER by harvester2.records.datestamp desc , record_id desc
```

Resultado (apenas 5 registos) da selecção dos documentos mais recentemente depositados em repositórios digitais (excepto BN por força da linha harvester2.records.archive\_id <> 1)

datestamp	record_id	archive_id	value
2007-06-27 01:00:00	7604	4	Cooperación frente al caos en Internet: CORC, una propuesta de trabajo
2007-06-27 01:00:00	7603	9	The 1,4-naphthoquinone scaffold in the design of cysteine protease inhibitors
2007-06-27 01:00:00	7602	2	Managing cognitive and affective trust in the conceptual R&D organization
2007-06-27 01:00:00	7601	2	Collaboration in the large: Using video conferencing to facilitate large group interaction
2007-06-27 01:00:00	7600	2	Experimental comparison of 2D and 3D technology mediated paramedic-physician collaboration in remote emergency medical situations

Se seguida os dados sobre cada registo são recolhidos da tabela entries construindo a apresentação do registo no ecrã.

O quadro seguinte apresenta os dados disponíveis para o registo cujo record\_id é 7604:

raw_field_id	value
1	Cooperación frente al caos en Internet: CORC, una propuesta de trabajo
19	2000-01-01
22	es
25	[Spanish abstract]En los entornos académicos cada día es más evidente la necesidad de mediación entre la información de Internet y el usuario final; así surge la demanda de la creación de espacios integradores de recursos y fuentes de información. En el Área de Documentación Científica de la Universidad Politécnica de Valencia surgió esta necesidad y para ello se creó una página web que recoge recursos de calidad en Internet relacionados con las diferentes áreas temáticas que interesan en dicha universidad. Se trata de un proyecto temático basado en la selección de contenidos de alta calidad formal. Durante tres meses el Área de Documentación Científica ha participado en el proyecto CORC (Cooperative Online Resource Catalog), de la OCLC, que es una herramienta para la coordinación de sitios web en los portales institucionales. En el estudio se contraponen la solución individualizada, que es la existente, frente a la cooperativa o consorciada. Se concluye proponiendo la creación de un proyecto de cooperación entre bibliotecas universitarias del ámbito español y europeo que recojan recursos de unas mismas áreas temáticas. [English abstract] Every day it is more evident the need for mediation between the information available on the Internet and the end users in the academic environments, as it is the demand for the creation of gateways to integrate resources and information. The Scientific Documentation Area of the Polytechnic University of Valencia created a web site to collect quality Internet resources within the different subject areas of interest in this university. During three months the Scientific

```
Documentation Area participated in the project OCLC CORC (Cooperative Online
Resource Catalog), a tool for the coordination of web sites in institutional
portals. In this study, individual solutions is compared to the cooperative
or consortium ones, concluding with a proposal for cooperation between
university libraries in Spain and Europe to collect resources in common
subject areas.
```

28 Conference Proceedings

29 <http://eprints.rclis.org/archive/00008089/>

31 I. Information treatment for information services

32 pdf

[http://eprints.rclis.org/archive/00008089/01/Cooperaci%C3%B3n\\_frente\\_caos.pdf](http://eprints.rclis.org/archive/00008089/01/Cooperaci%C3%B3n_frente_caos.pdf)

É tomado o cuidado de cortar os registos que têm descrições (`raw_field_id= 25`) demasiado longas ao último espaço antes do carácter 300 (função `oai_resumo` em `madulos.php`)

Os significados dos identificadores `raw_field_id` é realizado por análise da tabela `raw_fields`:

raw_field_id	name
1	title
16	creator
19	date
22	language
25	description
26	publisher
27	contributor
28	type
29	identifier
30	source
31	subject
32	format
33	relation
34	coverage
35	rights

Só a título de exemplo é apresentado o comando necessário para fazer a selecção de registos segundo determinada língua:

```
SELECT          harvester2.entries1.entry_id,
                harvester2.entries1.record_id,
                harvester2.entries1.value ,
                harvester2.records.archive_id,
                harvester2.records.datestamp
FROM
    harvester2.entries JOIN ( harvester2.entries as entries1 ,
harvester2.records )
    ON ( ( harvester2.entries.record_id = harvester2.entries1.record_id AND
harvester2.entries1.raw_field_id=1)
        AND
        harvester2.entries.record_id = harvester2.records.record_id
    )
WHERE (harvester2.entries.raw_field_id=22 and harvester2.entries.value in
('pt','por')and harvester2.records.archive_id <> 1)
ORDER by harvester2.records.datestamp desc, harvester2.records.record_id
desc
limit 25;
```

A expressão `harvester2.entries.value in ('pt','por')` é usada pois os repositórios são livres de escolher expressar as línguas em códigos de 2 ou de 3 caracteres. O documento regulador do DublinCore<sup>20</sup> afirma realmente que os valores têm de provir de uma das tabelas normativas e aconselha a rfc3066:

```
Term Name: language
URI: http://purl.org/dc/elements/1.1/language
Label: Language
Definition: A language of the resource.
Comment: Recommended best practice is to use a controlled vocabulary such as RFC 3066 [RFC3066].
References: [RFC3066] http://www.ietf.org/rfc/rfc3066.txt
```

A RFC3066 no entanto diz taxativamente, no ponto 2 da secção 2.3:

When a language has both an ISO 639-1 2-character code and an ISO 639-2 3-character code, you MUST use the tag derived from the ISO 639-1 2-character code.

O que patentemente não é sempre cumprido<sup>21</sup>. Isto obrigou o autor a fazer a verificação manual das diversas línguas, e em boa hora o fez pois a selecção das várias línguas latinas tem de ser feita por: `[...] entries.value in ('es','esp','it','ita','fr','fra')[...]`.

---

<sup>20</sup> **Dublin Core Metadata Element Set, Version 1.1: Reference Description. (2004).** 28 June 2007 <<http://www.dublincore.org/documents/dces/>>.

<sup>21</sup> O mesmo ênfase é dado em vários documentos normativos incluindo **W3C i18n article: Language tags in HTML and XML** ( <http://www.w3.org/International/articles/language-tags/Overview.en.php> )

## 5. Conclusão

Em jeito de auto avaliação tenho sinceros remorsos de não ter podido cumprir a tempo todas as funcionalidades<sup>22</sup> que foram explanadas na proposta de trabalho final. Tenho também receio de não ter conseguido fazer justiça ao esforço que o docente colocou na elevação do nosso senso de usabilidade e estética. A obrigatoriedade de o trabalho ser individual não permitiu a criação de sinergias que pudessem proporcionar um resultado visualmente mais cuidado

Por outro lado este sítio deu-me muito prazer a desenvolver. Primeiro pelo desafio de programação e procura de soluções de arquitectura de dados que resolvessem aparentes impossibilidades. Em segundo lugar pela possibilidade que proporcionou de familiarização com muitas tecnologias que têm até agora sido olhadas apenas muito de longe no curso em que o projecto se insere. Estou em crer que todos os discentes vão ter de conhecer as ferramentas aqui usadas (ou suas semelhantes e descendentes) em menos de 5 anos. Em terceiro lugar, foi uma experiência muito edificante verificar como, um pouco de experiência em programação e o contacto prático<sup>23</sup> com menos de uma mão cheia de protocolos e normas, pode fazer nascer uma série de funcionalidades que parecem fazer falta à comunidade.

### 5.1. ... e os utilizadores?

Entre 7 de Abril e o dia de hoje o **bibliorandum**, recebeu 861 visitas com 1936 visualizações de páginas. 184 das visitas foram directas e das visitas por referência de um motor de pesquisa 235 foram pelo vocábulo “bibliorandum”. Das 861 visitas apenas existem 367 visitantes únicos o que representa que os utilizadores estão a retornar e tornaram-se ‘clientes’.

### 5.2. ... e o futuro ?

A maior parte dos projectos conhecidos realizados no âmbito da avaliação da disciplina PISCI em edições anteriores foram abandonados pouco depois da sua missão de exposição das competências adquiridas no decurso do ano lectivo terem sido avaliadas. O **bibliorandum** foi no entanto desenvolvido com o objectivo de ser uma plataforma estável de fornecimento de serviços ao longo do tempo, para o que foi necessário prescindir da prerrogativa de funcionar no servidor interno da disciplina.

---

<sup>22</sup> Agregação de multimédia

<sup>23</sup> Em vernáculo seria possível dizer mesmo “sujar as mãos”

Assim, após avaliação, este projecto será mantido em funcionamento e em permanente desenvolvimento, sob a égide da **INCITE: Associação Portuguesa para a Gestão de Informação**, como serviço que esta presta à comunidade de profissionais de Informação-Documentação e às diversas comunidades de prática, investigação, docência e discência em Ciências da Informação.

## 6. Bibliografia

SAUERS, M.P., **Blogging and RSS: a librarian's guide**. 2006, Medford, N.J: Information Today, Inc.

TENNANT, R., **XML in libraries**. 2002, New York: Neal-Schuman Publishers. xi, 213 p.

MARQUES, Francisco; MENDES, Ana – **FrontPage XP**, Lousã: FCA, 2004, 423 p.

Além de muitas soluções pontuais de diversos problemas de programação os inseparáveis manuais de PHP, MySQL, e SWISH-E nos mais variados formatos (HTML, CHM, em linha, páginas impressas, etc).

Os manuais dos diversos programas utilizados:

PARMAN, RYAN ; SNEDDON, GEOFFREY, **SimplePie**: [Em Linha] [EUA?] [Consultado 2007-06-28] disponível em WWW: < URL: <http://simplepie.org/wiki/start>>

SIMON FRASER UNIVERSITY , **Open Archives Harvester**: [Em Linha] Vancouver, Canadá [Consultado 2007-06-28] disponível em WWW: <URL: [http://pkp.sfu.ca/harvester\\_documentation](http://pkp.sfu.ca/harvester_documentation)>

**Custom Search Engine Documentation** [em Linha] EUA [Consultado 2007-06-28] disponível em WWW: < URL: <http://www.google.com/coop/docs/cse/>>

E por fim a melhor formação possível: horas e horas de a ver como dezenas de soluções de repositórios e periódicos electrónicos implementavam tanto o protocolo OAI como as estruturas de classificação usando:

SULEMAN, H., **Open Archives Initiative - Repository Explorer**. [Em linha] 2006, U of Cape Town: Cape Town. África do Sul [Consultado 2007-06-28] disponível em WWW: < URL: <http://re.cs.uct.ac.za/>>

O sincero reconhecimento do trabalho feito por VILARINHO, FERNANDO. **Directório de Bibliotecas- Portugal**. [Em Linha] Vila do Conde, Portugal [Consultado 2007-06-28] disponível em WWW: < URL: <http://bibliotecas.wetpaint.com/>>

## 7. Anexos

### 7.1. Ferramentas e tecnologias

#### 7.1.1. OAI-PMH (Protocolo)

A definição em português mais adequada vem do site Luso DSpace<sup>24</sup>:

“Protocolo OAI para disseminação de meta-dados. Uma forma de os repositórios (*data providers*) partilhar (exponer) os seus metadados para serem recolhidos (*harvested*) por serviços (*service providers*) que permitem a pesquisa por entre vários repositórios OAI-Compliant.”<sup>25</sup>

No mesmo sítio, OAI é definido como:

“Open Archives Initiative: Lançada em 1999 com o objectivo de criar uma plataforma simples para permitir a interoperabilidade e a pesquisa de publicações científicas de diversas disciplinas. Esta iniciativa surgiu no seio da comunidade dos “eprints” e partiu de uma abordagem essencialmente técnica (de que resultou o protocolo OAI-PMH), sem grande preocupação “filosófica”[...]”

Em termos práticos é um protocolo de comunicações a nível de aplicação em que um serviço cliente dispõe de 6 verbos para pedir informação a um serviço servidor, por norma um repositório. Os pedidos são colocados como parâmetros num pedido *http* e o resultado é fornecido em XML. Por norma um dos formatos em que os registos requeridos são apresentados é precisamente o formato Dublin Core, podendo cada repositório optar por expor dados segundo outros protocolos/DTD.

Um exemplo prático:

---

<sup>24</sup> Sítio da comunidade de língua portuguesa que usa o sistema [DSpace](#)

<sup>25</sup> Citado de <http://lusospace.sdum.uminho.pt:8080/pt/glossary.jsp>

<http://eprints.rclis.org/perl/oai2?verb=Identify>

```
<?xml version="1.0" encoding="UTF-8" ?>
<?xml-stylesheet type='text/xsl' href='/oai2.xsl' ?>

<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <responseDate>2007-06-28T02:24:51Z</responseDate>
  <request verb="Identify"
resumptionToken="">http://eprints.rclis.org/perl/oai2</request>
  <Identify>
    <repositoryName>E-LIS</repositoryName>
    <baseURL>http://eprints.rclis.org/perl/oai2</baseURL>
    <protocolVersion>2.0</protocolVersion>
    <adminEmail>mailto:eprints@dois.it</adminEmail>
    <earliestDatestamp>0001-01-01</earliestDatestamp>
    <deletedRecord>persistent</deletedRecord>
    <granularity>YYYY-MM-DD</granularity>
    <description>
      <oai-identifier xmlns="http://www.openarchives.org/OAI/2.0/oai-identifier"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai-identifier
http://www.openarchives.org/OAI/2.0/oai-identifier.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
        <scheme>oai</scheme>
        <repositoryIdentifier>eprints.rclis.org</repositoryIdentifier>
        <delimiter>:</delimiter>
        <sampleIdentifier>oai:eprints.rclis.org:23</sampleIdentifier></oai-
identifier></description>
      <description>
        <eprints xmlns="http://www.openarchives.org/OAI/1.1/eprints"
xsi:schemaLocation="http://www.openarchives.org/OAI/1.1/eprints
http://www.openarchives.org/OAI/1.1/eprints.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
          <content>
            [...]
          </content>
        </eprints>
      </description></Identify></OAI-PMH>
```

Temos portanto a certeza que o repositório tem interface OAI-PMH e toda a informação necessária em termos de responsabilidade, autorização de uso e eventuais limitações ‘legais’ impostas.

O comando mais informativo que se segue é a aplicação do verbo `ListSets` que expõe a organização interna do repositório:

```

http://eprints.rclis.org/perl/oai2?verb=ListSets

<?xml version="1.0" encoding="UTF-8" ?>
<?xml-stylesheet type='text/xsl' href='/oai2.xsl' ?>

<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <responseDate>2007-06-28T02:32:34Z</responseDate>
  <request verb="ListSets"
resumptionToken="">http://eprints.rclis.org/perl/oai2</request>
  <ListSets>
    <set>
      <setSpec>7374617475733D756E707562</setSpec>
      <setName>Status = Unpublished</setName></set>
    <set>
      <setSpec>7374617475733D707562</setSpec>
      <setName>Status = Published</setName></set>
    <set>
      <setSpec>7374617475733D696E7072657373</setSpec>
      <setName>Status = In Press</setName></set>
    <set>
      <setSpec>7375626A656374733D472E</setSpec>
      <setName>Subject = G. Industry, profession and education.</setName></set>
  [...]
</ListSets></OAI-PMH>

```

Neste ponto há que explorar o resultado completo e verificar se há um ou mais *sets* que nos interessem anotando o seu “setSpec” como , por exemplo o *setSpec* “7375626A656374733D482E:4853” correspondente ao assunto “*Subject = H. Information sources, supports, channels.: HS. Repositories.*”. Podemos agora construir um comando que apresente os indicadores (cabeçalhos) ou

```

http://eprints.rclis.org/perl/oai2?verb=ListIdentifiers&from=2007-06-01&metadataPrefix=oai\_dc&set=7375626A656374733D482E%3A4853

```

```

<?xml version="1.0" encoding="UTF-8" ?>
<?xml-stylesheet type='text/xsl' href='/oai2.xsl' ?>

<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <responseDate>2007-06-28T02:42:10Z</responseDate>
  <request verb="ListIdentifiers" metadataPrefix="oai_dc"
set="7375626A656374733D482E:4853"
resumptionToken="">http://eprints.rclis.org/perl/oai2</request>
  <ListIdentifiers>
    <header>
      <identifier>oai:eprints.rclis.org:10196</identifier>
      <timestamp>2007-06-20</timestamp>
      <setSpec>7374617475733D707562</setSpec>
      <setSpec>7375626A656374733D492E:4945</setSpec>
      <setSpec>7375626A656374733D482E:4853</setSpec></header>
  [...]
</ListIdentifiers></OAI-PMH>

```

os registos completos conforme o verbo usado<sup>26</sup>:

<sup>26</sup> ListIdentifiers ou ListRecords

### 7.1.2. RSS

O RSS é um formato específico de XML orientado para a transferência de um conjunto simplificado de dados. Teve até hoje várias edições e quase tantos significados como edições. Dada a sua simplicidade é usado para a transferência (“pull”) de listas de notícias com actualização frequente. Dada a ubiquidade do protocolo e a sua simplicidade é usado para muitas outras finalidades deste a apresentação de valores bolsistas até ao estado do tempo.

O **bibliorandum** é tanto consumidor como produtor de RSS. Na **blogosfera** recebe, agrega e redistribui conteúdos em RSS enquanto que na **Akademya** recebe em OAI-PMH (outro formato XML encapsulando Dublin Core), agrega e redistribui em formato RSS.

### 7.1.3. Google Custom Search Engine

Esta funcionalidade do **Google** permite explorar as capacidades de indexação e pesquisa do google criando situações de cegueira selectiva no motor de pesquisa, isto é, apenas são apresentados resultados que estejam numa lista de endereços previamente ‘ensinada’<sup>27</sup>. É necessário dispor de uma conta pessoal no google e saber que o serviço tem de ser acedido pelo endereço <http://www.google.com/cse>.

Assim, para a funcionalidade de pesquisa em páginas de bibliotecas, depois de criado um novo “Search Engine” este foi configurado para permitir a apresentação de resultados apenas das páginas ou sítios indicados. A língua de apresentação dos detalhes de navegação e mensagens de serviço foi configurada para a língua portuguesa (Ilustração 3)

Os sítios pesquisados foram seleccionados usando os directórios de páginas e sítios web de bibliotecas e de organismos relevantes para o público-alvo. Esta lista de sítios, foi inicialmente alimentada linha a linha na **interface de administração do CSE** (Ilustração 4) sendo este método rapidamente substituído pela exportação de um ficheiro (Ilustração 5) com os sítios já carregados, sua análise (Ilustração 6) e desenvolvimento de um *workbook* em MS-Excel que criasse o XML necessário para poder fazer a carga de toda a colecção de endereços de volta ao CSE. Na mesma interface de administração foram recolhidos os segmentos de *html* e *javascript* que foram embutidos nas páginas do **bibliorandum**.

---

<sup>27</sup> É também possível utilizar este serviço pelo sentido inverso, ou seja pela apresentação de endereços que desejamos que sejam excluídos dos resultados da pesquisa.

```

<!-- Google CSE Search Box Begins -->
<form id="searchbox_000675947084576693848:vhyuavnqlr4"
action="http://www.bibliorandum.net/resultado.php">
  <input type="hidden" name="cx" value="000675947084576693848:vhyuavnqlr4" />
  <input type="hidden" name="cof" value="FORID:9" />
  <input name="q" type="text" size="40" />
  <input type="submit" name="sa" value="Search" />
</form>
<script type="text/javascript"
src="http://www.google.com/coop/cse/brand?form=searchbox_000675947084576693848%3Avhyuavnq
lr4"></script>
<!-- Google CSE Search Box Ends -->

```

```

<!-- Google Search Result Snippet Begins -->
<div id="results_000675947084576693848:vhyuavnqlr4"></div>
<script type="text/javascript">
  var googleSearchIframeName = "results_000675947084576693848:vhyuavnqlr4";
  var googleSearchFormName = "searchbox_000675947084576693848:vhyuavnqlr4";
  var googleSearchFrameWidth = 600;
  var googleSearchFrameborder = 0;
  var googleSearchDomain = "www.google.com";
  var googleSearchPath = "/cse";
</script>
<script type="text/javascript"
src="http://www.google.com/afsonline/show_afs_search.js"></script>
<!-- Google Search Result Snippet Ends -->

```

Para a funcionalidade de pesquisa nos recurso cobertos pela funcionalidade Blogosfera o processo foi mais complexo pois fez-se uso da capacidade de busca em contexto (nalguma documentação referenciada como “Facetada”<sup>28</sup>, e noutra como “Refinamentos”<sup>29</sup>).

Para fazer uso desta funcionalidade os endereços internet que desejamos filtrar devem ser acompanhados dos identificadores das categorias em que o recurso se encontra inserido. Como se pode ver na Ilustração 7, o endereço de um sítio além de ser agregado a um motor de pesquisa por `Label name="_cse_zsikhcarode"` é também agregado a um determinado tema por `Label name="blogs_municipais"`. O valor deste atributo é exactamente extraído da tabela “themes” (descrita no ponto 4.1.1). O programa que gera automaticamente o ficheiro para importar no CSE com estes parâmetros para todos os Blogues, a partir de um *join* da tabela “themes” com a tabela “feeds” (cf. 4.1.2), é acessível em `BIBLIORANDIUM_URLS.php`. O programa complementar é `BIBLIORANDIUM_SITE.php` que gera uma definição de motor de pesquisa com os parâmetros que ligam a este motor as diversas categorias (tabela “themes”) e indica qual a forma como devem ser colocadas à disposição do utilizador. Na Ilustração 8 podemos ver precisamente um elemento `FacetItem` ao qual é agregado um elemento `Label`. O atributo *title* do *facetitem* é o texto (campo `google_cse_label` da tabela `Themes`) que o utilizador terá no resultado da pesquisa e o *label*

<sup>28</sup> Nos ficheiros XML por exemplos que são exportados e importados para parametrização da funcionalidade

<sup>29</sup> Em todo o interface do CSE que modifica os parâmetros desta funcionalidade

associado fará com que apenas os sítios que foram qualificados com este mesmo *label* aparecerão nos resultados de pesquisa se aquela categoria for escolhida para refinar a pesquisa. As possibilidades do sistema são muito mais vastas e confessamos não as ter explorado na totalidade, podendo mesmo serem criados *label's* que ao invés de restringirem ou excluïrem um grupo de sítios dão diferentes pesos a cada grupo de sítios no resultado. Infelizmente o CSE apresenta sempre todas as categorias como disponíveis mesmo que a uma ou mais categorias não tenham resultado nenhum possível pelo que não é possível considerar isto um sistema capaz de fazer *clustering* de resultados.

#### 7.1.4. PHP

PHP (um acrónimo recursivo para "PHP: Hypertext Preprocessor") é uma linguagem de script de fonte aberta, de uso geral, muito utilizada e especialmente apetrechada para o desenvolvimento de aplicações Web, sendo interlineável com HTML.<sup>30</sup>

O sistema usado tem instalada a versão Version 5.1.2

#### 7.1.5. MySQL

MySQL é um sistema de gestão de base de dados relacional bastante usado em ambientes \*nix e começa a ser uma alternativa de Fonte Aberta e custo gratuito às ferramentas tradicionais em servidores Windows como o SQL server.<sup>31</sup>

Versão instalada e em uso (incidentalmente em servidor separado):

---

<sup>30</sup> Adaptado de [http://pt.php.net/manual/pt\\_BR/introduction.php](http://pt.php.net/manual/pt_BR/introduction.php)

<sup>31</sup> Traduzido de <http://www.mysql.org/doc/refman/5.0/en/what-is-mysql.html>

```
mysql> status;
-----
mysql Ver 14.12 Distrib 5.0.24a, for pc-linux-gnu (i486) using readline 5.1

Connection id:          122262
Current database:
Current user:           root@localhost
SSL:                    Not in use
Current pager:          stdout
Using outfile:          ''
Using delimiter:        ;
Server version:         5.0.24a-Debian_9-log
Protocol version:      10
Connection:             Localhost via UNIX socket
Server characterset:    latin1
Db characterset:        latin1
Client characterset:    latin1
Conn. characterset:     latin1
UNIX socket:            /var/run/mysqld/mysqld.sock
Uptime:                 79 days 16 hours 19 min 0 sec

Threads: 6  Questions: 21377350  Slow queries: 15  Opens: 8274  Flush tables: 1  Open
tables: 64  Queries per second avg: 3.105
-----
```

Para além do servidor no equipamento da Incite foram usadas ferramenta de acesso gráfico **MySQL Query Browser**<sup>32</sup> e **MySQL Administrator**<sup>33</sup>, ambos parte do pacote **MYSQL GUI Tools**

### 7.1.6. APACHE

O Projecto **Apache HTTP Server** é um esforço de desenvolvimento colectivo de uma plataforma robusta, fiável, de acesso livre e fonte aberta, rica em capacidades para servir informação sob o protocolo http<sup>34</sup>

A versão instalada é, conforme exposto pela extensão *server\_info*<sup>35</sup>:

```
Server Version: Apache/2.0.55 (Ubuntu) PHP/5.1.2 mod_ssl/2.0.55 OpenSSL/0.9.8a
mod_perl/2.0.2 Perl/v5.8.7
```

### 7.1.7. Open Archives Initiative - Repository Explorer

Disponível no endereço <http://re.cs.uct.ac.za/> permite explorar repositórios para os quais se saiba o endereço do interface OAI-PMH. Facilita a exploração e experimentação pois permite

---

<sup>32</sup> <http://dev.mysql.com/doc/query-browser/pt/index.html>

<sup>33</sup> <http://dev.mysql.com/doc/administrator/pt/index.html>

<sup>34</sup> Traduzido de [http://httpd.apache.org/ABOUT\\_APACHE.html](http://httpd.apache.org/ABOUT_APACHE.html)

<sup>35</sup> A extensão está documentada em [http://httpd.apache.org/docs/2.2/mod/mod\\_info.html](http://httpd.apache.org/docs/2.2/mod/mod_info.html). O endereço de acesso neste computador não é aqui indicado por constituir um risco de segurança

aceder e testar os diversos verbos OAI directamente de um interface sem ter de construir a lista de parâmetros manualmente ou por *copy/paste* num editor de texto

### 7.1.8. PKP Open Archives Harvester

O PKP Open Archives Harvester é um sistema de indexação de metadados, de acesso livre e fonte aberta, desenvolvido pelo Public Knowledge Project<sup>36</sup> com o apoio de fundos federais ,com o objectivo de melhorar o acesso aos resultados de investigação científica.<sup>37</sup>

A versão instalada é a 2

### 7.1.9. A actualização da informação da “blogosfera”

Durante uma fase inicial a recolha de novas entradas na blogosfera esteve a cargo de uma ferramenta denominada **FEED ON FEEDS**. No entanto com o decorrer do tempo e à medida que a colecção de blogues a capturar de diversificou e cresceu começaram a aparecer deficiências não tanto no programa, mas na maneira como reagia a erros de configuração na informação RSS, principalmente discrepâncias nos conjuntos de caracteres indicados no *header http*, no *header xml* e realmente praticados no corpo das entradas. Ou seja o programa não tinha “condescendência” nenhuma com XML mal formado. Estes problemas tinham passado despercebidos até então porque, apesar de fazer a recolha de 445 canais RSS, estes eram todos em língua inglesa. Ao investigar maneiras de corrigir o programa (que em termos de processamento de formatos de data foi extensivamente alterado no decorrer da fase inicial do **bibliorandum**) foi localizada uma biblioteca de interpretação de RSS, **SimplePie**<sup>38</sup>, com bastantes análises e críticas positivas publicadas em fóruns relevantes na área do PHP. Uma vez experimentada esta biblioteca o *feedonfeeds* foi abandonado e optou-se por um sistema de recolha de informação completamente realizado de raiz que pode ser encontrado no programa *simplepie.php* que é repetido todos os 10 minutos por um comando na tabela *contrab*

```
00,10,20,30,40,50 * * * * wget -O - http://www.bibliorandum.net/SimplePie.php -q
```

---

<sup>36</sup> <http://pkp.sfu.ca/>

<sup>37</sup> Open Archives Harvester: Public Knowledge Project em <http://pkp.sfu.ca/?q=harvester>

<sup>38</sup> SimplePie: *Super-fast, easy-to-use, RSS and Atom feed parsing in PHP*. Disponível em <http://simplepie.org/>

Incidentalmente o comando <http://www.bibliorandum.net/SimplePie.php>, quando executado num computador comum, num navegador comum, apresenta o progresso da captura e eventuais problemas existentes.

### 7.1.10. A escolha de repositórios em “Akademya”

A escolha dos repositórios a tratar focou-se precisamente na área das ciências da informação obedecendo a dois critérios: ou um repositório é completamente dedicado às ciências da informação (caso do E-LIS) ou dispõe de uma frequência de registo, que se possa qualificar subjectivamente entre o “ocasional” e o “frequente”, de material relevante para ao público-alvo. É também essencial que haja mecanismos que permitam identificar e isolar essa material relevante do resto dos registos pelo mecanismo “*selective harvesting*” de uma determinada *setspec*<sup>39</sup>. É assim que as teses da Biblioteca Nacional tiveram de ser abandonadas, pois não há *set* relevante para o assunto da tese; também o mesmo aconteceu com a produção da Universidade do Minho pois os *sets* disponíveis não são temáticos mas sim orgânicos e a produção que seria relevante para o **bibliorandum** encontra-se muito dispersa por todos os departamentos

A escolha de repositórios foi feita começando por consultar a página “**LIS open access resources**”<sup>40</sup> do serviço “E-LIS - Eprints for LIS”. Daqui foi possível aceder ao sítio da METALIS<sup>41</sup> que faz o *harvesting* de vários repositórios. Estes repositórios foram analisados com a ferramenta **Open Archives Initiative - Repository Explorer**. Os sítios **OpenDOAR - Directory of Open Access Repositories**<sup>42</sup>, **Registered Data Providers do Open Archives Initiative**<sup>43</sup> e **Directory of open access journals**<sup>44</sup>

---

<sup>39</sup> Um *set* (conflação de **Especificação de Conjunto**) corresponde por norma a um “assunto”, mas repositórios há, como o repositoriUM, que usam os *sets* para agregar a produção científica por departamentos e unidades orgânicas. O *selective harvesting* é uma funcionalidade do OAI-PMH que permite requerer que o servidor apenas exponha os dados (identificadores ou registos completos) dos registos que obedecem a determinada condição, e duas delas são obrigatórias: que são mais recentes que uma determinada data e/ou que fazem parte de um determinado *set*

<sup>40</sup> Antonella De Robbio, Imma Subirats Coll. “LIS open access resources.” (2002). 28 June 2007 <<http://eprints.rclis.org/resources.html>>.

<sup>41</sup> <http://metallic.cilea.it/>

<sup>42</sup> <http://www.opendoar.org/index.html>

<sup>43</sup> <http://www.openarchives.org/Register/BrowseSites>

<sup>44</sup> <http://www.doaj.org/>

### 7.1.11. A actualização da informação de “Akademya”

A recolha de informação dos repositórios OAI foi inicialmente pensada para ser realizada manualmente (isto é, por um programa desenvolvido de raiz no âmbito do projecto). No entanto, ao investigar as melhores práticas para reagir programaticamente a um comando `resumptionToken`, foi detectada a existência de um programa Open Source, o **Open Archives Harvester** já referenciado no ponto 7.1.7, que não só indicava conter uma boa implementação do comando como, pela documentação, fazia tudo o que era necessário: isto é: Permitia a construção de uma lista de repositórios, recolha de informação desses repositórios e pesquisa na informação recolhida. Depois de instalado e parametrizado (ver Ilustração 9, na página 39 ou consultar directamente <http://pkp.janjos.com>) verificou-se que fazia as funções indicadas apesar de toda a recolha ser despoletada manualmente ou ciclicamente por linha de comando. Não dispunha também de mecanismos de gestão de *sets*<sup>45</sup>. No entanto tendo já alguma experiência na localização de *subsets* e parametrização de recolha dos respectivos registos foi fácil criar os mecanismos automáticos de recolha de registos em repositórios que é realizada com o seguinte comando (de exemplo):

```
php -c /etc/php5/apache2/php.ini \  
/var/www/harvester/tools/harvest.php \  
5 \  
verbose \  
skipExistingEntries \  
  from=2007-04-01 \  
  set=7375626A656374733D61737263:343030303030:343030323030:343030323031
```

Este comando ordena a leitura do repositório número 5, em modo verboso<sup>46</sup>, não substituindo os registos que já estejam replicados localmente, tratando apenas de registos registados depois de 1 de Abril de 2007 e que estejam agregados ao *set* com a identidade indicada. Para evitar picos de acesso à internet e de uso do CPU este comando especificamente é executado às 8:45, 13:45, 15:45, 17:45 e 21:45<sup>47</sup>.

---

<sup>45</sup> A nível de recolha podia usar *subsets* para recolha selectiva de informação, mas a nível de interface não tinha mecanismos que aproveitassem o registo dos *subsets*

<sup>46</sup> Permite a visualização quando executado manualmente. Quando executado automaticamente os resultados são descartados

<sup>47</sup> De início a colheita era feita num minuto determinado de todas as horas mas notou-se que a frequência de depósito nos repositórios não o exigia.

## 7.2. Listas

### 7.2.1. Repositórios

DITED	48	<a href="http://dited.bn.pt/">http://dited.bn.pt/</a>
DLIST		<a href="http://dlist.sir.arizona.edu/">http://dlist.sir.arizona.edu/</a>
ALIA e-prints		<a href="http://www.alia.org.au/">http://www.alia.org.au/</a>
E-LIS		<a href="http://eprints.rclis.org/">http://eprints.rclis.org/</a>
QUT ePrints Archive	49	<a href="http://eprints.qut.edu.au/">http://eprints.qut.edu.au/</a>
@rchiveSIC		<a href="http://archivesic.ccsd.cnrs.fr/">http://archivesic.ccsd.cnrs.fr/</a>
TEL French eTheses - ©HAL	50	<a href="http://tel.archives-ouvertes.fr/">http://tel.archives-ouvertes.fr/</a>
Library Student Journal		<a href="http://informatics.buffalo.edu/org/ljsj/">http://informatics.buffalo.edu/org/ljsj/</a>
Open Research Online	51	<a href="http://libeprints.open.ac.uk/">http://libeprints.open.ac.uk/</a>
CNR Bologna Research Library		<a href="http://biblio-eprints.bo.cnr.it/">http://biblio-eprints.bo.cnr.it/</a>
Computer Science Technical Reports	52	<a href="http://eprints.cs.vt.edu/">http://eprints.cs.vt.edu/</a>
@Virginia Tech		

---

<sup>48</sup> O repositório de teses da Biblioteca Nacional de Portugal é refrescado apesar de ser excluído no momento de fazer a selecção de registos para apresentar na funcionalidade *akademya*

<sup>49</sup> Apenas dois *sets*

<sup>50</sup> Apenas um *set*

<sup>51</sup> Apenas um *set*

<sup>52</sup> Apenas três *sets*

## 7.3. Imagens

### Basic information

Your search engine's name and description will be shown on its Google [homepage](#).

Search engine name:

Search engine description:

Keywords describe the content or subject of your search engine. These keywords are used to tune your search engine results. [Learn more](#).

Search engine keywords:

e.g. climate "global warming" "greenhouse gases"

Search engine language:

### Preferences

How to search included sites:  Search only included sites.  
 Search the entire web but emphasize included sites.

Who can collaborate:  Anyone may volunteer to contribute to this search engine.  
[Learn more](#).  Only people I invite may contribute to this search engine.

Specify whether your search engine is for a non-profit, university, or government website that should not have advertising on the results pages.

Advertising status:  Show ads on results pages.  
 Do not show ads on results pages (for non-profits, universities, and government agencies only).

### Ilustração 3: A configuração do motor de pesquisa no CSE

**Control panel: Bibliotecas Portugesas**  
[Basics](#) | [Sites](#) | [Refinements](#) | [Look and feel](#) | [Code](#) | [Collaboration](#) | [Make money](#) | [Advanced](#) | [Preview](#)

**Included sites** « Previous 20 Viewing 21 - 40 of 258 Next 20 »

URL contains:

- [cmmatosinhos.wiremaze.com/document/816725/859770.pdf](http://cmmatosinhos.wiremaze.com/document/816725/859770.pdf)
- [cmmatosinhos.wiremaze.com/pagegen.asp?SYS\\_PAGE\\_ID=830186](http://cmmatosinhos.wiremaze.com/pagegen.asp?SYS_PAGE_ID=830186)
- [cmmatosinhos.wiremaze.com/pagegen.asp?SYS\\_PAGE\\_ID=850423](http://cmmatosinhos.wiremaze.com/pagegen.asp?SYS_PAGE_ID=850423)
- [cmmbib.cm-moura.pt/](http://cmmbib.cm-moura.pt/)
- [cmmbib.cm-moura.pt/nova/](http://cmmbib.cm-moura.pt/nova/)
- [del.icio.us/post](http://del.icio.us/post)
- [estarreja.libware.net/PortalWeb/portal/alias\\_PortalWeb/lang\\_pt-PT/tabID\\_3352/DesktopDefault.aspx](http://estarreja.libware.net/PortalWeb/portal/alias_PortalWeb/lang_pt-PT/tabID_3352/DesktopDefault.aspx)
- [estarreja.libware.net/portalweb/](http://estarreja.libware.net/portalweb/)
- [feiradolivrorodao.blog.com/](http://feiradolivrorodao.blog.com/)
- [figueira.net/patrimonio/cultural/biblioteca/](http://figueira.net/patrimonio/cultural/biblioteca/)
- [fundao.libware.net/opac/](http://fundao.libware.net/opac/)
- [hemerotecadigital.cm-lisboa.pt/](http://hemerotecadigital.cm-lisboa.pt/)
- [ilhavo.libware.net/opac/default.htm](http://ilhavo.libware.net/opac/default.htm)
- [oazemeis.libware.net/opac/](http://oazemeis.libware.net/opac/)
- [oleiros.netvidade.com/conteudos/biblioteca-municipal.php?header=infra-estruturas](http://oleiros.netvidade.com/conteudos/biblioteca-municipal.php?header=infra-estruturas)
- [opapalagui.blogspot.com/](http://opapalagui.blogspot.com/)
- [revelarlx.cm-lisboa.pt/gca/index.php?id=1002](http://revelarlx.cm-lisboa.pt/gca/index.php?id=1002)
- [revelarlx.cm-lisboa.pt/gca/index.php?id=1003](http://revelarlx.cm-lisboa.pt/gca/index.php?id=1003)
- [revelarlx.cm-lisboa.pt/gca/index.php?id=1004](http://revelarlx.cm-lisboa.pt/gca/index.php?id=1004)
- [revelarlx.cm-lisboa.pt/gca/index.php?id=1005](http://revelarlx.cm-lisboa.pt/gca/index.php?id=1005)

**Excluded sites**  
 You have not excluded any sites - [Exclude sites](#)

**Ilustração 4: A Interface de edição de sítios do CSE**

**Download annotations**

You can download all of your current annotations in the [TSV format](#) or [XML format](#).

**⚠ URLs from which links are extracted will not appear in TSV format.**

**Ilustração 5: A descarga de endereços**

```

- <GoogleCustomizations>
  - <Annotations>
    - <Annotation about="212.55.147.58/*">
      <Label name="_cse_vhyuavnqr4"/>
    </Annotation>
    - <Annotation about="213.13.161.67/*">
      <Label name="_cse_vhyuavnqr4"/>
    </Annotation>
    - <Annotation about="62.48.200.223/*">
      <Label name="_cse_vhyuavnqr4"/>
    </Annotation>

```

Ilustração 6: O formato de exportação/importação do CSE (segmento)

```

- <GoogleCustomizations>
  - <Annotations>
    - <Annotation about="boamemoria.blogspot.com/*/*">
      <Label name="_cse_zsikhcarode"/>
      <Label name="blogs_municipais"/>
    </Annotation>

```

Ilustração 7: Formato de Importação Exportação do CSE com categorias

```

<?xml version="1.0" encoding="UTF-8" ?>
- <CustomSearchEngine version="1.0" volunteers="true" keywords="blogs portuguese" Title="TESTE:
  BIBLIORANDUM : BLOGS" Description="Teste dos mecanismos Disponíveis" language="pt">
- <Context>
  - <Facet>
    - <FacetItem Title="Municipal">
      <Label name="blogs_municipais" mode="FILTER" Rewrite=""
        IgnoreBackgroundLabels="false" />
    </FacetItem>
    - <FacetItem Title="Escolar">
      <Label name="blogs_escolares" mode="FILTER" Rewrite=""
        IgnoreBackgroundLabels="false" />
    </FacetItem>

```

Ilustração 8: Definição de Facetas como Categorias de Blogues

PUBLIC KNOWLEDGE PROJECT

# Open Archives Harvester<sub>2</sub>

HOME
ABOUT
SEARCH
BROWSE
HELP

---

Home > **Harvester2**

## Harvester2

---

**Welcome to the Public Knowledge Project's metadata archive...**

To improve the accuracy of searching within the PKP System, authors have been asked to index their work, where applicable, by discipline(s), topics, genre, method, coverage, and sample. This allows you to search for "empirical" versus "historical" studies, for example, under "index terms." You can also view a document's index terms by selecting the complete record from among the search results.

**HARVESTER STATS**

Harvester2 currently has **7604** papers from **11** archive(s) indexed.

**ADD YOUR ARCHIVE**

[Click here](#) to add your system to our index.

**CONTENT**

Search

© 2005-2006 [Public Knowledge Project](#)

**Ilustração 9: o PKPHarvester em <http://pkp.janjos.com>**

## **7.4. Anexos soltos**

### **7.4.1. Relatórios Google Analytics**

- Lealdade do visitante
- Conteúdo principal
- Cobertura Geo Map