REPOX – Uma Infra-estrutura XML para a Base de Dados Bibliográfica Nacional

José Borbinha[1], Nuno Freire[2]

INESC-ID – Instituto de Engenharia de Sistemas e Computadores, Rua Alves Redol 9, Apartado 13069, 1000-029 Lisboa, Portugal
[1]jlb@ist.utl.pt
[2]Nuno.Freire@bn.pt

Resumo. A PORBASE – Base de Dados Bibliográfica Nacional, é um catálogo colectivo em linha de 150 bibliotecas Portuguesas, incluindo a BN – Biblioteca Nacional, a entidade que tem a responsabilidade da sua coordenação. O sistema integrado de gestão de bibliotecas que suporta a PORBASE armazena os dados num sistema de gestão de base de dados relacional. A PORBASE é uma importante fonte de informação cujo potencial tem sido subaproveitado, pretendendo a BN por isso desenvolver novos serviços. Tal tem-se mostrado no entanto difícil devido ao facto da actual estrutura da base de dados, que está definida para um fim específico nem sempre compatível com os novos serviços que se pretendem desenvolver. O sistema REPOX – Repositório XML da PORBASE pretende ser uma infra-estrutura para resolver esse problema. O objectivo é permitir ter toda a informação da PORBASE num repositório XML, em redor do qual se poderão desenvolver os novos serviços. Outra mais valia é a persistência do historial diário dos dados e suas alterações, que não é possível manter actualmente. Ainda, o REPOX representará assim uma cópia de segurança expressa em XML dos dados da PORBASE, que assim ficarão independentes de qualquer sistema específico. Este artigo descreve a infra-estrutura do sistema REPOX, os resultados e conclusões da sua utilização em ambiente de produção de apoio a uma nova gama de serviços, e ainda os planos de desenvolvimentos futuros.

1 Introdução

A PORBASE – Base de Dados Bibliográfica Nacional [9], é um catálogo colectivo em linha das bibliotecas portuguesas, sendo actualmente a maior base de dados bibliográficos do país, reflectindo as colecções de mais de 150 bibliotecas Portuguesas, incluindo a BN – Biblioteca Nacional [1], a entidade que tem a responsabilidade da sua coordenação. O sistema integrado de gestão de bibliotecas (SIGB) que suporta a PORBASE armazena os dados num sistema de gestão de base de dados relacional (SGBD) da Sybase [11]. É também sobre este SGBD que assenta o catálogo em linha (OPAC - *Online Public Access Catalogue*) que disponibiliza a PORBASE para consulta através da Internet (Fig. 1).

A PORBASE é uma importante fonte de informação cujo potencial tem sido subaproveitado, problema que a BN pretende encarar desenvolvendo novos serviços.

Mas para isso é necessário considerar outra infra-estrutura para os dados da PORBASE. O esquema actual da base de dados suporta um sistema de informação proprietário (sistema HORIZON [2]), e está por isso definida segundo os requisitos tradicionais de suporte às actividades de catalogação desse sistema e de pesquisa por OPAC [19] (Fig. 1). Esse esquema não é assim sempre é compatível com os requisitos dos novos serviços que se pretendem desenvolver, nem será boa prática alterar o esquema pelo risco de conflito potencial que tal implicaria com a empresa responsável pela sua manutenção. Para além disso a possível utilização do mesmo servidor que é usado para catalogação e pesquisa pelo público iria sobrecarregar esse para tarefas que poderá muitas vezes ser perfeitamente feitas em *batch* noutro servidor. Finalmente, a replicação dos dados, codificados num esquema aberto e armazenados numa infra-estrutura objectiva, será uma resposta às questões de segurança e preservação.

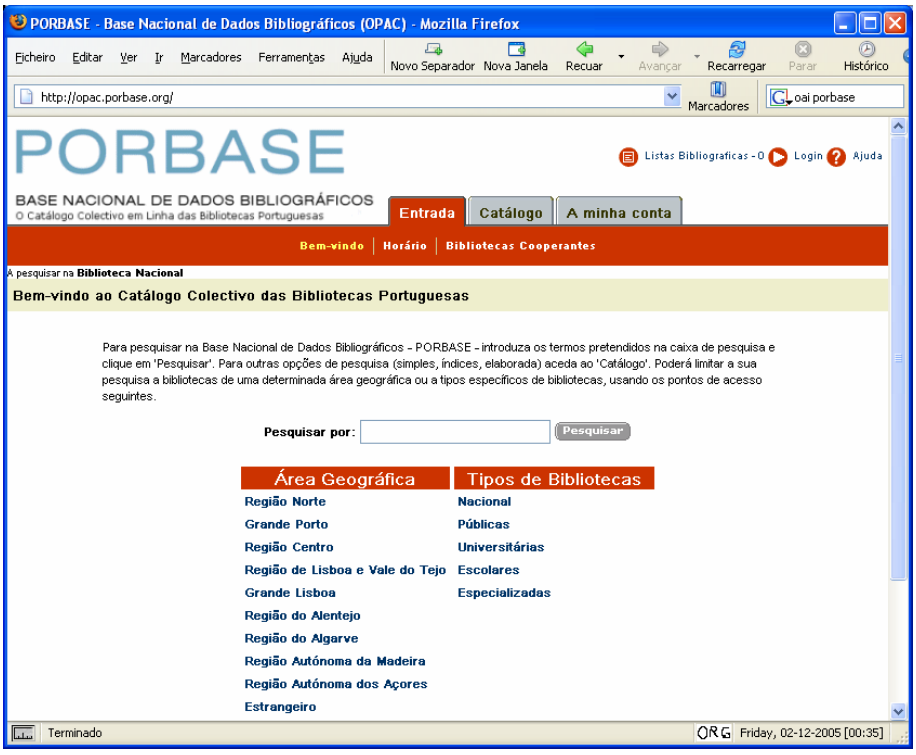

Fig. 1. OPAC da PORBASE.

O objectivo da infra-estrutura REPOX – Repositório XML da PORBASE é assim o de permitir concentrar toda a informação da PORBASE num repositório XML, em torno do qual se podem então planear o desenvolvimento dos novos serviços, tais como de produção de estatísticas, controlo de qualidade, e todo o tipo de serviços úteis de suporte à decisão, explorar novas formas de pesquisa, etc. Nalguns casos pontuais tal poderá ser feito directamente sobre o REPOX, mas o normal é que, de

acordo com os requisitos do serviço em causa, os registos possam ser indexados em índices próprios, importados para bases de dados dedicadas ou sistemas de *data wharehousing*, etc.

Este artigo procede com uma descrição do desenho e arquitectura da solução, seguindo-se uma descrição dos serviços até ao momento já desenvolvidos.

2 Desenho

A infra-estrutura REPOX gere, como primeiro objectivo, o armazenamento em XML de registos da PORBASE, mas o mesmo modelo pode ser igualmente explorado para outros registos da BN. Um registo, na perspectiva REPOX, é uma entidade que se encontra num repositório externo, tem um identificador unívoco nesse repositório, e pode ser representado segundo um esquema XML, tal como ilustrado na Fig. 2. Para o caso da PORBASE, o REPOX gere o armazenamento de duas classes de entidades: os registos bibliográficos e os registos de autoridade.

O REPOX mantém um histórico de todas as versões recolhidas de todos os registos, com a respectiva data, o conteúdo tal como se encontrava nessa data codificado em XML, e ainda uma lista dos eventos que ocorreram sobre esta versão do registo. Um evento pode indicar qual o actor interveniente (um utilizador do repositório externo, ou uma aplicação) e o tipo de acção (criação do registo, alteração, remoção, alteração de registo relacionado, etc.).

Apesar de na fase actual o sistema REPOX estar apenas em funcionamento interno na BN suportando serviços exclusivos da PORBASE, o projecto prevê a existência de vários repositórios externos com dados bibliográficos, sobre autoridades ou mesmo de registos de sistemas de gestão de arquivo (em oposição aos registos documentais da PORBASE). Os registos serão recolhidos periodicamente e guardados num repositório XML, destinando-se a ser posteriormente recuperados e manipulados por serviços externos, indexadores e gestores de colecções (Fig. 3).

Os serviços externos têm acesso ao repositório XML e são notificados quando termina a recolha de um repositório externo, o que despoleta a sua execução. Um exemplo de um serviço já desenvolvido é o sistema QualiCat, que faz o controle de qualidade, com uma periodicidade diária, para os registos bibliográficos da PORBASE segundo os requisitos do formato UNIMARC [14], das Regras Portuguesas de Catalogação, e ainda das especificidades de preenchimento da rede de cooperação PORBASE.

Os indexadores são programas que processam os registos armazenados, preparando-os para serem pesquisados de forma eficiente. No caso dos registos bibliográficos da PORBASE, em formato UNIMARC, estes indexam o título, os autores, os identificadores (ISBN, ISSN, cotas dos exemplares e Número de Depósito Legal), assim como outra informação que seja pertinente para os serviços que num dado momento a necessite.

O REPOX organiza os dados recolhidos em colecções, que não são mutuamente exclusivas, ou seja, um registo no REPOX pode pertencer a várias colecções. A atribuição dos registos às colecções é levada a cabo por gestores de colecção, os quais podem avaliar os registos segundo vários critérios. Estes critérios podem ser

informação de assunto (história, ciências, artes, etc.), o local ou língua de publicação, o ano de nascimento dos autores, a data de publicação, etc.

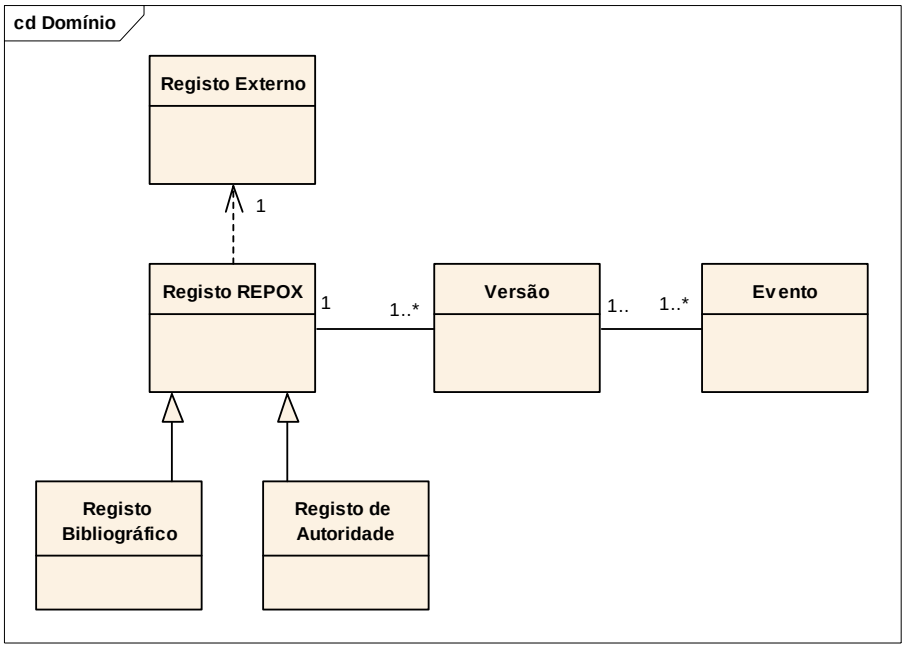

Fig. 2. O conceito de registo na perspectiva do REPOX

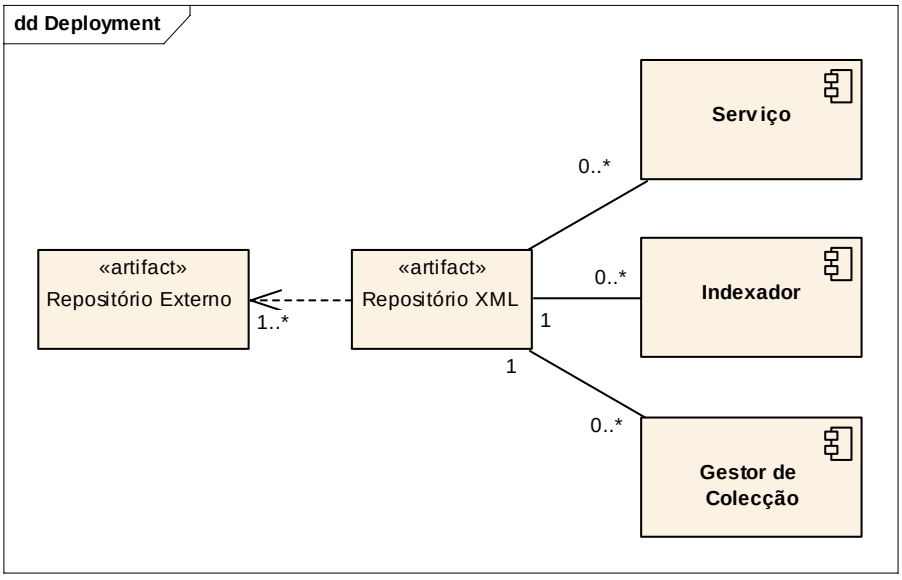

Fig. 3. O repositório XML e principais entidades relacionadas

3 Arquitectura REPOX

A infra-estrutura REPOX é constituída por três componentes principais: o gestor do repositório XML, um SGBD relacional MySql [7] e um servidor de aplicações Tomcat 5.5 [3]. A respectiva arquitectura está representada na Fig. 4.

A versão actual recolhe diariamente os registos bibliográficos e de autoridade da PORBASE. Esta é gerida pelo sistema integrado de gestão de bibliotecas HORIZON, utilizando um SGBD relacional SYBASE [11]. O REPOX gere um serviço de sincronização que recolhe os registos através de uma ligação ao SGBD por JDBC [7]. Cada registo recolhido é guardado num ficheiro XML segundo um esquema definido para o REPOX. Este formato contém uma secção onde são codificadas as várias versões do registo, segundo um esquema próprio para cada tipo de registo. O formato para a troca de registos bibliográficos utilizado quase exclusivamente em Portugal, e consequentemente na PORBASE, é o UNIMARC, sendo por isso utilizado o formato MARCXML [6] para codificar esses registos no REPOX, numa estrutura como se mostra no exemplo da Fig. 5.

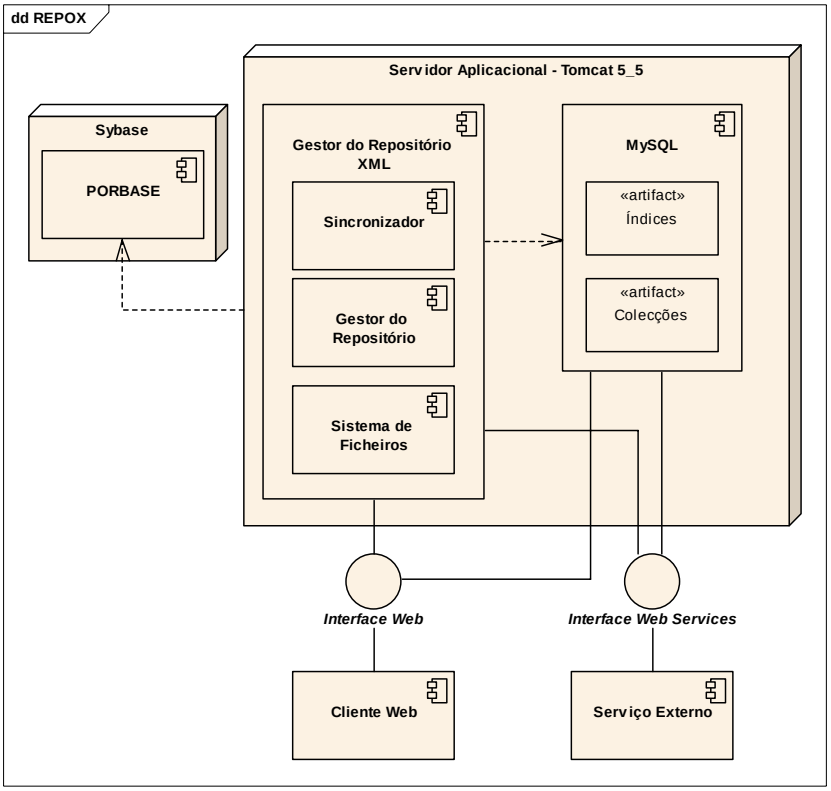

Fig. 4. Os principais componentes do sistema REPOX

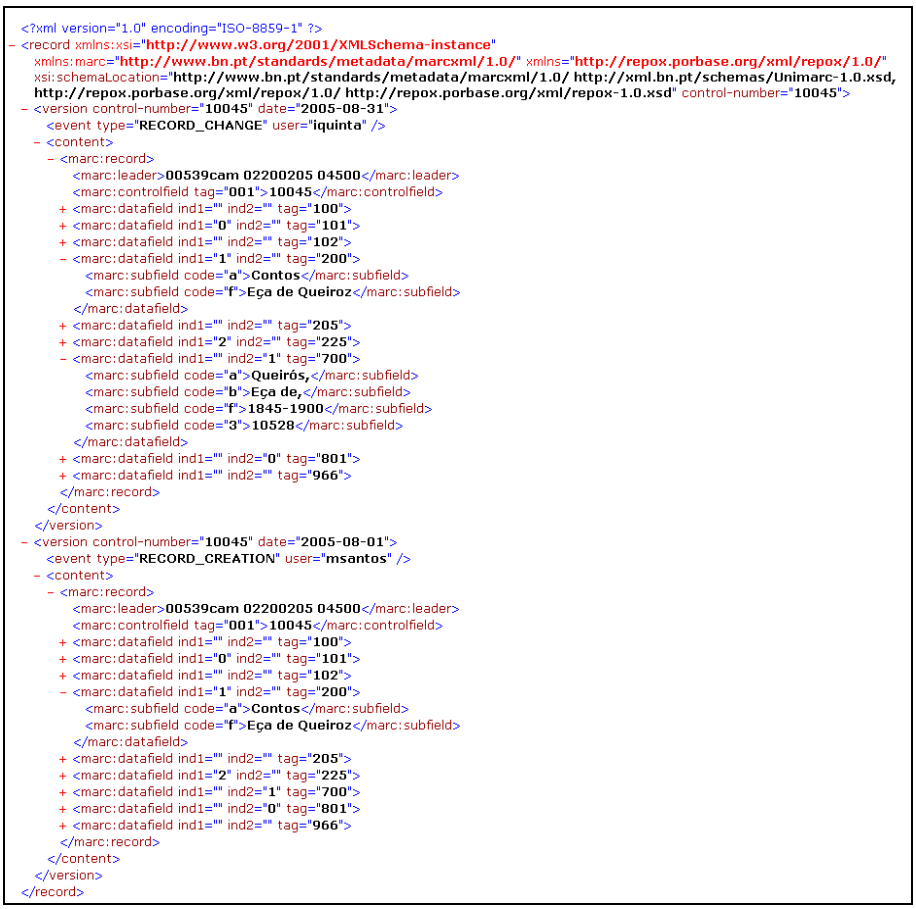

Fig. 5. Um registo bibliográfico em MARCXML

Todos os ficheiros XML do REPOX são guardados no sistema de ficheiros do servidor. Para efeitos de eficiência, é utilizado um SGBD relacional MySql para guardar e indexar alguns dados dos registos com o objectivo de permitir a sua pesquisa. Os indexadores do REPOX processam os registos extraindo e preparando esses dados que sejam relevantes para pesquisa.

Os gestores de colecções utilizam, tal como os indexadores, uma base de dados no mesmo SGBD, na qual são inseridas e indexadas as relações entre os registos e as colecções a que pertencem.

Existem ainda várias ferramentas de gestão do repositório. Estas, através de uma interface por linha de comandos, permitem ao administrador criar, actualizar ou eliminar índices e colecções.

Para acesso ao repositório por serviços externos, o REPOX disponibiliza ainda uma interface por *Web Services* [15], a qual oferece várias funções, tais como:

- Obter a versão mais recente de um registo;
- Obter um registo tal como esse se encontrava em determinada data;
- Obter todo o histórico de um registo;
- Obter uma lista de todos os registos alterados entre duas datas;
- Obter listas de registos através de pesquisas nos índices.

O REPOX disponibiliza ainda uma interface WEB para os utilizadores. Esta permite pesquisar, navegar e consultar registos individualmente, mostrando toda a informação existente, incluindo todas as versões dos registos, os eventos ocorridos e identificação exacta das alterações aos registos entre versões.

O repositório XML da PORBASE conta neste momento com 2,6 milhões de registos (entre registos bibliográficos e de autoridade) correspondendo a cerca de 6 milhões de ficheiros XML que totalizam em tamanho aproximadamente 25 GBytes.

A base de dados relacional que suporta a pesquisa nos dados do repositório contém 21 índices (como o títulos, autores, datas de publicação, editores, ISBN, entre outros). São também mantidas na base de dados 18 colecções de registos bibliográficos e 2 colecções de registos de autoridade. Dependendo da utilização, são frequentemente criados índices e colecções adicionais temporárias. Tal ocorre sempre que, no contexto de alguma actividade da PORBASE, surge a necessidade de processar dados para os quais não existem índices, ou se pretende exportar registos segundo critérios que definem uma colecção virtual adicional.

4 Serviço URN.PORBASE.ORG

Vários serviços da Biblioteca Nacional tiram já proveito eficazmente dos dados do REPOX estarem em XML.

Um desses casos é o serviço URN– Acesso à PORBASE por Identificadores Unívocos [10], que disponibiliza os registos da PORBASE para as bibliotecas portuguesas em vários formatos através de uma interface por http. Na Fig. 7 encontra-se um exemplo da interface para utilizadores deste serviço. O serviço URN tira partido dos índices existentes no REPOX para permitir que os registos sejam obtidos através de vários identificadores (ISBN, ISSN, número de depósito legal, etc.). Os registos são disponibilizados em vários esquemas (UNIMARC, Dublin Core [12]) e em vários formatos de codificação (XML, HTML, texto, etc.). Estes formatos são gerados em tempo real através de transformações XSLT [16] a partir do registo armazenado no REPOX.

Este serviço é acedido principalmente por sistemas de informação de gestão de bibliotecas, especialmente de bibliotecas portuguesas, que importam os registos directamente da PORBASE para as bases de dados locais, facilitando assim o trabalho de catalogação aí levado a cabo.

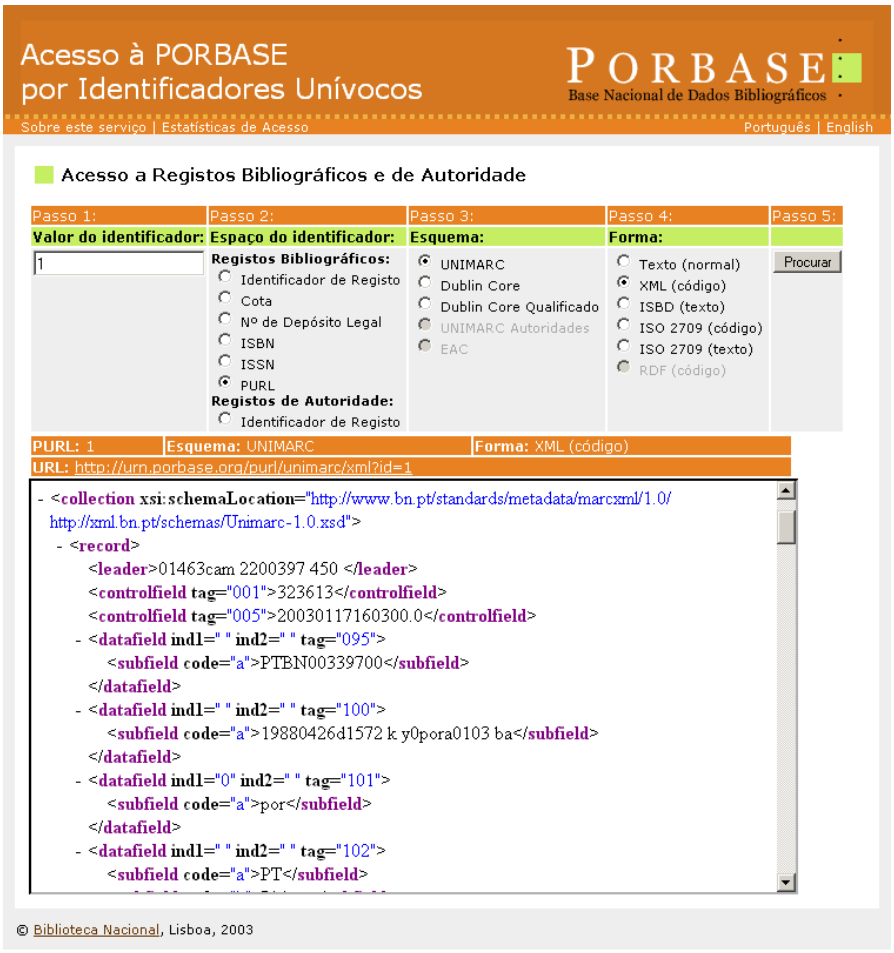

Fig. 6. O serviço URN (http://urn.porbase.org)

5 Serviço OAI.PORBASE.ORG

Um outro serviço suportado pelo REPOX é o serviço OAI-PMH – *Open Archives Iniciative Protocol for Metadata Harvesting* [8].

Este serviço disponibiliza os registos da PORBASE em projectos de cooperação, permitindo que cópias de registos existentes num outro local sejam mantidas actualizadas. É assim por exemplo através deste serviço que o portal da TEL – *The European Library* [13], recolhe regularmente os registos da PORBASE, e os disponibiliza para pesquisa e acesso (Fig. 8).

O servidor OAI-PMH da PORBASE obtém os registos através do REPOX, organizados por colecções que são geridas pelos gestores de colecções. No caso da

colecção para o portal TEL, os registos em MARCXML do REPOX são transformados para o formato de dados da TEL, no perfil para bibliotecas do formato Dublin Core, o DC-Lib [5]. Esta transformação é naturalmente obtida por XSLT.

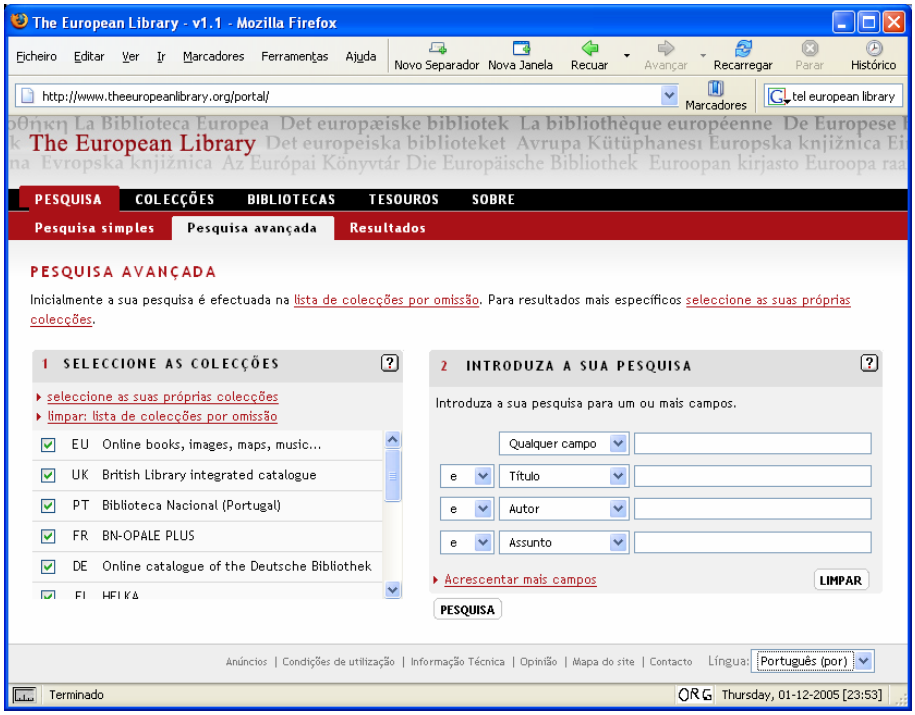

Fig. 7. Serviço TEL, onde a PORBASE está disponível para pesquisa (mostra-se a interface em Português).

Relatório resumido de: 2005-11-25

Ver relatório detalhado

Utilizador	Bibliográficos Criados	Bibliográficos Alterados	Bibliográficos Apagados	Autoridades Criadas	Autoridades Alteradas	Autoridades Apagadas	Itens Criados	Itens Alterados	Itens Apagados	Total
aalves	0	28	0	10	4	0	0	0	0	42
acarol	0	13	0	14	14	0	0	0	0	41
apires	0	18	0	11	7	0	0	0	0	36
asantos	33	1	0	0	0	0	61	0	0	95
barbara	0	44	0	15	3	0	0	25	0	87
bpmp02	0	15	0	3	0	0	30	1	0	49
bpmp03	2	0	0	0	0	0	27	0	0	29
bpmp04	3	0	0	1	0	0	10	0	0	14
bpmp05	0	0	0	0	0	0	4	0	0	4
bpmp06	8	5	0	1	0	0	30	23	0	67
caetano	0	8	0	10	12	0	0	8	0	38
dfontes	0	9	0	2	0	0	0	0	0	11
ecalapa	0	14	0	10	0	0	0	10	0	34
fatima	22	0	0	0	0	0	49	0	0	71
freitas	5	1	0	0	0	0	10	0	0	16
goreti	4	8	0	1	0	0	8	2	0	23
icosta	0	10	0	4	3	0	0	3	0	20
ines	0	73	0	0	1	0	0	1	0	75
ipuga	0	6	0	13	8	0	0	0	0	27

Fig. 8. Exemplo de um relatório de actividade diária na PORBASE criado pelo REPOX.

6 Serviço de Relatórios de Actividade da PORBASE

Finalmente, outro serviço criado baseado na infra-estrutura REPOX é o serviço de relatórios de actividade diária da PORBASE.

Este serviço notifica diariamente os utilizadores do sistema HORIZON de todas as operações sobre os registos, enviando um e-mail contendo um relatório das operações efectuadas nas 24 horas anteriores. A mesma informação fica ainda acessível em linha, num servidor da rede interna, como ilustrado na Fig. 9.

Mensalmente é ainda criado e enviado aos utilizadores e responsáveis de áreas e serviços um relatório consolidado.

7 Conclusão

O sistema REPOX entrou em funcionamento operacional no dia 1 de Outubro de 2005. De imediato melhorou consideravelmente a disponibilidade e o desempenho do acesso aos dados pelos variados serviços da Biblioteca Nacional e das bibliotecas cooperantes da PORBASE. O sistema de gestão de bibliotecas da PORBASE ficou também livre de processar os acessos aos registos por parte de outras bibliotecas e de vários serviços que o sobrecarregavam já de modo bastante significativo.

Para além destes efeitos imediatos, o REPOX abre novas perspectivas. A possibilidade de estabelecer uma ligação ao REPOX através de um URL que identifica o estado do registo numa determinada data, por exemplo, permite a simplificação de relatórios sobre registos da PORBASE que são gerados em outros sistemas. Estes sistemas já não necessitam de guardar localmente cópias dos registos, bastando-lhes guardar uma referência para o registo na data pretendida.

Devido a esta solução, tem também sido possível fazer análise dos registos da PORBASE sobre campos que não estão indexados no sistema HORIZON. É possível descobrir assim com mais celeridade falhas de preenchimento dos registos bibliográficos, bem como determinar as formas de as corrigir automaticamente sem necessidade de intervenção humana.

Um outro factor de extraordinária importância é o facto de desta forma o REPOX representar ainda uma cópia de segurança dos dados da PORBASE, expressa em XML. Neste momento tal representa já um total de cerca de 2,6 milhões de registos, que assim estão preservados, independentes de qualquer software ou hardware específico.

A utilização de tecnologia XML no repositório permite ainda que formatos de registos heterogéneos sejam geridos pelo REPOX de forma relativamente transparente. Isto permite a sua fácil adaptação a outros esquemas de dados. O próximo repositório de dados a ser assim integrado no REPOX será o do Arquivo da Cultura Portuguesa Contemporânea (http://acpc.bn.pt), cujos metadados estão representados num outro sistema de informação, num esquema definido segundo as ISAD(G) [17]. Neste caso os registos deverão vir a ser codificados segundo o esquema EAD [18].

Finalmente, estão em fase de análise de requisitos e desenho de serviços de estatísticas mais avançados da PORBASE (e do REPOX em geral), de controlo de

qualidade (com por exemplo validações sintácticas e semânticas de acordo com as regras do UNIMARC), bem como o estudo de soluções de *data warehouse* para suportar serviços gerais de *data mining* e especificamente de suporte à decisão.

Referências

[1] Biblioteca Nacional: http://www.bn.pt
[2] Horizon: http://www.novabase.pt/ConteudosHTML/Horizon.pdf
[3] Jakarta Tomcat: http://tomcat.apache.org/
[4] JDBC: http://java.sun.com/products/jdbc/
[5] Library Application Profile: http://dublincore.org/documents/library-application-profile/
[6] MARCXML. MARC 21 XML Schema: http://www.loc.gov/standards/marcxml/
[7] MySQL: http://www.mysql.com
[8] OAI-PMH. Open Archives Initiative – Protocol for Metadata Harvesting http://www.openarchives.org/
[9] PORBASE – Base Nacional de Dados Bibliográficos: http://www.porbase.org
[10] *Serviço URN:* http://urn.porbase.org
[11] SyBase: http://www.sybase.pt/gvsview/gvs/sybase-pt/home/index.html
[12] The Dublin Core Metadata Initiative: http://dublincore.org
[13] The European Library: http://www.theeuropeanlibrary.org/
[14] UNIMARC http://www.unimarc.info
[15] *Web Services*: http://www.w3.org/2002/ws/
[16] XSLT: http://www.w3.org/TR/xslt
[17] ISAD(G) - General International Standard Archival Description: http://www.ica.org/biblio/cds/isad_g_2e.pdf
[18] EAD – Encoded Archival Description: http://www.loc.gov/ead/
[19] IFLA Guidelines for Online Public Access Catalogue (OPAC) Displays - Final Report May 2005. IFLA Series on Bibliographic Control; vol. 27. Saur, 2005. ISBN 3-598-24276-X (rascunho disponível em http://www.ifla.org/VII/s13/guide/opacguide03.pdf)