

Histórias de Vida + Processamento Estrutural = Museu da Pessoa

Alberto Manuel Simões and José João Almeida

Museu da Pessoa
Departamento de Informática
Universidade do Minho
{albie@alfarrabio.ljj@}di.uminho.pt
<http://natura.di.uminho.pt>

Resumo Este artigo apresenta a arquitectura actual do Museu da Pessoa, contemplando a forma como os documentos estão a ser editados, catalogados, arquivados, e processados para a criação das estruturas necessárias ao Museu.

1 Introdução

O Museu da Pessoa nasceu no Brasil, S. Paulo, de um grupo de historiadores que resolveram compilar a história do país usando depoimentos do cidadão comum. Depois de iniciado o projecto começaram a surgir outras ideias, e outros projectos. Foram-se construindo acervos não só da história do país, mas também da história de determinadas profissões, festas populares, instituições ou empresas.

Entretanto um docente da Universidade do Minho, depois de ter assistido a uma apresentação do Museu da Pessoa - Brasil, resolveu criar um núcleo em Portugal: o Núcleo Português do Museu da Pessoa[1].

Em Portugal, o projecto está a ser mantido pelo Departamento de Informática que tenta manter o espírito do Museu da Pessoa original:

- é urgente recolher histórias da pessoa comum;
- sempre que possível, uma história deve ser contada na primeira pessoa, com toda a sua riqueza adicional;
- as histórias estar acessíveis em vários media (Internet, réplicas, livros, etc.), uma vez que se trata de um património comum;
- as regras de ética e direitos de autor têm de ser respeitadas;
- usar abordagens o mais metódicas possível;

No entanto, ao estar sediado num departamento de informática um dos principais objectivos é criar uma base informática para a gestão do acervo.

O projecto está a tentar evoluir para uma rede de Museus da Pessoa a nível mundial.

A secção 2 apresenta uma introdução ao ciclo de vida ou modelo de tratamento das histórias de vida recolhidas. De seguida, na secção 3, apresenta-se as tecnologias usadas para arquivar os documentos recolhidos enquanto que as secções seguintes apresentam as várias tecnologias usadas para a geração dos vários media.

2 Tratamento de Histórias de Vida

As histórias de vida são recolhidas e tratadas segundo o modelo apresentado na figura 1.

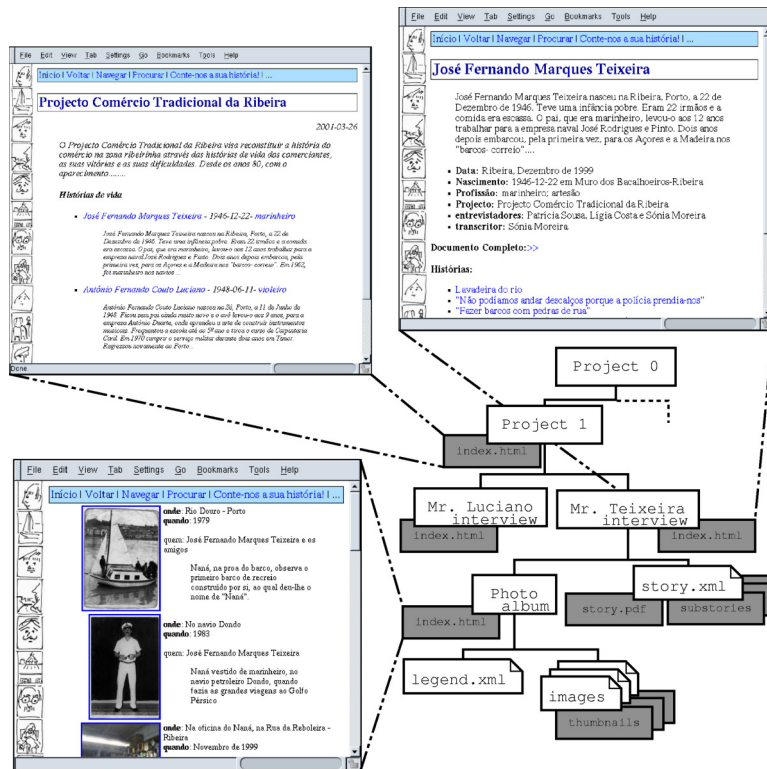


Figura 1. Modelo de tratamento da História de Vida

Resumidamente:

- entrevista e recolha de documentos junto do público — inclui a entrevista e gravação vídeo (ou apenas sonora) da entrevista, registo dos dados biográficos (e questões jurídicas, como direitos de autor) e recolha de material vídeo ou fotográfico como sejam cartões, fotografias e outros;
- digitalização dos documentos encontrados — digitalização de vídeos e dos documentos, tratamento gráfico, legendagem em XML e inclusão no catálogo multimédia;
- transcrição do documento para XML — conversão do som da entrevista para formato electrónico de forma a facilitar a transcrição em computador utilizando *software* desenvolvido no Museu;

- revisão da entrevista — correcção de erros ortográficos, edição de histórias, marcação XML de zonas interessantes. Criação de uma mini-biografia do depoente. Indexação da entrevista no catálogo geral; Este conceito de revisão leva à existência de mais do que um documento por entrevista, como seja um em formato de entrevista e um outro na primeira pessoa, como se o próprio depoente o escrevesse.
- publicação web do material recolhido — transformação de XML em HTML, divisão das histórias em pequenas porções, criação de álbum de fotografias, extracção de pequenos vídeos e várias outras operações.
- publicação em papel — criação de colectâneas de histórias de vida, transformação do XML para \LaTeX , criação de livros em PostScript para *download* ou impressão.

3 Suporte Informático

Embora o projecto conte actualmente com mais de dois anos de existência, o seu suporte informático ainda não foi totalmente decidido. De facto, ao longo do seu desenvolvimento foram tidos em conta os seguintes objectivos:

- utilizar formatos abertos, de modo a facilitar o intercâmbio de histórias entre ferramentas, permitir independência de plataformas e de aplicações;
- disponibilização dos documentos em formatos ricos, capazes de estabelecer relações entre os mesmos;
- utilizar estruturas classificativas comuns entre todos os documentos;
- disponibilização de ferramentas construídas e de formatos propostos;
- privilegiar a automatização;

No entanto, só alguns dos pontos relativos à escolha do suporte informático foram completamente definidos, enquanto que outros vão evoluindo à medida que a necessidade surge.

3.1 Armazenagem

Em relação ao formato em que se deveria armazenar o acervo, a sua escolha não foi complicada. A possibilidade de se definir estruturas complexas em XML e de ser um formato aberto e facilmente processável fez-lo a escolha do Museu.

O XML permite que vários dos objectivos de desenvolvimento de Museu da Pessoa se concretizem mas, por si só, não resolve todos os problemas. Em particular, foi necessário definir uma forma de arquivar os vários documentos em algum sítio. Com este fim surgiram duas hipóteses:

- utilizar uma base de dados, estilo Oracle que, nas suas versões mais recentes, inclui suporte para campos de tipo XML, permitindo a realização de *queries* XQL sobre estes;
- criar um documento XML por entrevista e, de alguma forma, utilizar o sistema de ficheiros do sistema operativo para guardar estes documentos;

A primeira hipótese obriga à utilização de ferramentas mais pesadas e, na sua maioria, comerciais. Sem dúvida que a utilização de uma base de dados permite ao museu crescer facilmente, mas a falta de fundos obrigou ao uso temporário de uma estrutura mais ou menos rígida sobre o sistema de ficheiros.

Um outro problema surgiu ao usar esta abordagem: várias pessoas podem vir a editar o mesmo documento ao mesmo tempo ou, determinada edição pode não fazer sentido, sendo necessário quer gerir conflitos entre as várias edições, quer ter capacidade de recuar edições.

Para resolver este tipo de situações optou-se por incluir todos os documentos XML num repositório CVS (Concurrent Version System) que permite de uma forma bastante eficiente permitir que vários utilizadores possam editar documentos mantendo sempre o histórico das alterações. Inclui também, a capacidade de juntar versões paralelas e só em casos extremos requerer a intervenção humana.

A organização do acervo no sistema de ficheiros obrigou à definição de determinadas regras para manter limpeza e coerência na sua arrumação:

- cada directoria na raiz do repositório corresponde a um projecto;
- directorias a níveis superiores podem corresponder a sub-projectos ou a depoimentos dentro desse projecto;
- cada directoria de projecto (ou sub-projecto) pode conter um documento XML com uma pequena sinopse com uma introdução e objectivos do projecto, várias directorias de depoentes ou sub-projectos, e uma directoria com fotografias;
- cada directoria de depoente pode ter várias versões da entrevista em XML, um documento com a identificação do depoente e uma directoria com fotografias;
- Cada directoria de fotografias contém imagens e um documento XML com a legendagem das mesmas;

A figura 2 mostra esta estrutura de uma forma gramatical.

$$\begin{aligned} \text{Projecto} &\leftarrow \text{Projecto}^* \times \text{Entrevista}^* \times \text{Album} \times \text{SinopseXML} \\ \text{Entrevista} &\leftarrow \text{DocXML}^* \times \text{BiXML} \times \text{Album} \\ \text{Album} &\leftarrow \text{LegendaXML} \times \text{Foto}^* \end{aligned}$$

Figura 2. Estrutura de directorias do repositório

3.2 Etiquetação

A informação relativa ao depoente está toda a ser guardada na sua directoria dentro do projecto em que se insere. Esta informação está dividida em três partes:

- mini-biografia e dados pessoais, como sejam o nome, data e local de nascimento, e profissão. Esta informação é colocada num documento próprio, denominado de BI (Bilhete de Identidade — na figura 2, denominado por BiXML);

- várias versões da entrevista, em documentos XML. Cada nome do documento deve reflectir o seu tipo (entrevista, editado);
- informação relativa a documentos digitalizados encontra-se no documento de legenda dentro da directoria de fotografias;

Estes documentos são processados automaticamente sempre que é necessário voltar a publicar determinado projecto. Este processamento constrói o catálogo (índice) de documentos do projecto e extrai um conjunto extra de informação.

Bilhete de Identidade

Como foi referido, o bilhete de identidade do depoente está a ser armazenado em XML. No entanto contém informação de tal forma rígida que não é mais do que uma tabela de base de dados armazenada em formato XML.

O seu conteúdo é:

- nome, profissão, data e local de nascimento;
- mini-biografia para ser apresentada como resumo da história de vida;
- informação sobre uma fotografia *escolhida* para ser mostrada juntamente com a mini-biografia;

A figura 3 mostra um bilhete de identidade a ser consultado na Internet.

```

1 <?xml version="1.0" encoding="ISO-8859-1"?>
2 <!DOCTYPE bi SYSTEM "http://alfarrabio.di.uminho.pt/mp/dtd/bi.dtd">
3 <bi>
4 <projecto>Memórias do Trabalho</projecto>
5 <depoente>José Vieira</depoente>
6 <biografia>
7   José Vieira nasceu a 16 de Outubro de 1920 em Resende. Filho de
8   lavradores passou uma infância difícil. Tinha 7 anos quando o pai
9   ficou paralisado. Dois anos depois foi morar com os tios que eram
10  lavradores. Mais tarde, foi com a mãe e os irmãos para o Porto
11  trabalhar na lavoura. Foi feitor na Quinta do Ramalho e trabalhou
12  ainda como serralheiro na F. Brindley. Esteve sempre envolvido nas
13  reivindicações dos trabalhadores.
14 </biografia>
15 <profissao>lavrador; feitor; serralheiro; jardineiro</profissao>
16 <nascimento onde="Resende" mes="10" dia="16" ano="1920"/>
17 </bi>

```

História de Vida

As etiquetas usadas nos documentos de histórias de vida dividem-se em três grupos fundamentais:

marcação de estrutura do mesmo género das etiquetas estruturais do HTML, permitem que se definam parágrafos bem como estrutura em poemas (poema, estrofe e verso);

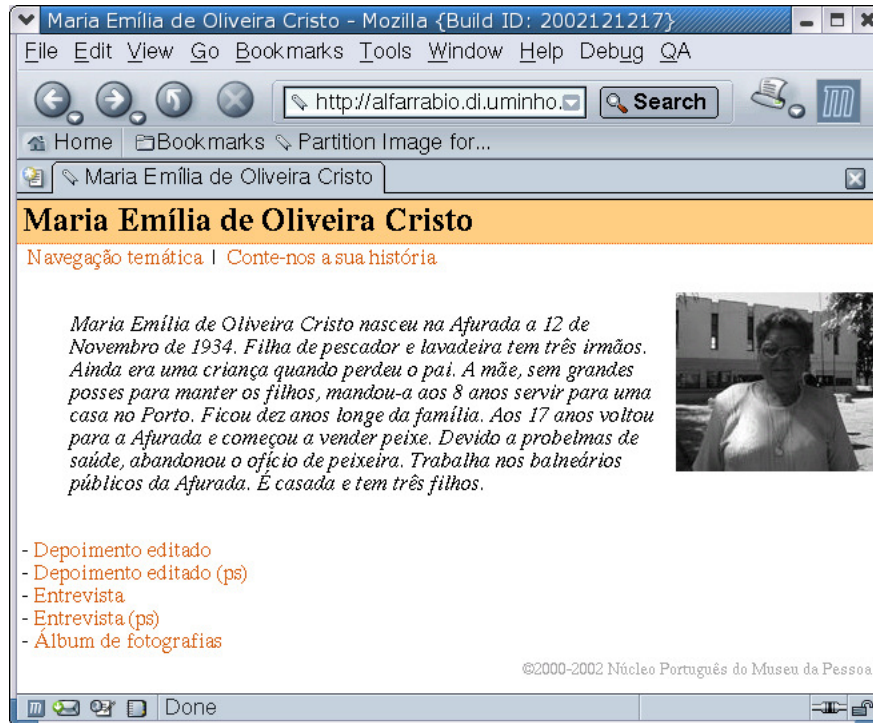


Figura 3. Consulta de um Bilhete de Identidade

```

1 | <poema>
2 |   <estrofe>
3 |     <verso>Hoje que Deus me levou</verso>
4 |     <verso>Não chorem que é o Destino. </verso>
5 |     <verso>Ele já ficou traçado </verso>
6 |     <verso>Dos meus tempos de menino. </verso>
7 |   </estrofe>
8 |   <estrofe>
9 |     <verso>Fui um poeta romântico </verso>
10 |    <verso>E cumpri a minha sina. </verso>
11 |    <verso>Fui um jovem tão feliz </verso>
12 |    <verso>E formei uma família. </verso>
13 |   </estrofe>
14 | </poema>

```

marcação de meta-informação delimitam zonas do depoimento que devem de algum modo ser processadas de forma especial. Exemplo desta etiquetação são as datas do documento, bem como nomes de instituições ou expressões regionais (e respectiva explicação).

```

1 | (...)
2 | Não quer ir para a pesca do bacalhau, não trabalha mais, vá trabalhar

```

```

3 | para onde quiser, mas aqui não trabalha mais" - e ele passou muita
4 | fome até resolver esse problema. Isto era no tempo da
5 | <ref tipo="Instituição">PIDE</ref>,
6 | no tempo da
7 | <expressao
8 |   significado='Quando as pessoas se referiam à PIDE, chamavam sempre
9 |     "outra senhora" e não PIDE, com medo de represálias.'>
10 |   outra senhora</expressao>,
11 | no tempo do
12 | <ref tipo="personalidade">Cardeal Cerejeira</ref>
13 | que dizia que - "Para o povo ser humilde tem que
14 | passar fome" - é verdade, é verdade.
15 | (...)

```

marcação de histórias embora também possam ser consideradas etiquetas de meta-informação, estas delimitam pequenas porções da história de vida que podem ser lidas independentemente do depoimento completo. Desta forma, permitimos que uma história de vida possa dar origem a histórias temáticas. Como exemplo temos mais de uma dúzia de etiquetas como: “*namoro*”, “*ofício*”, “*casamento*”, “*infância*”. Além destas, existe uma genérica denominada “*episódio*” para ser usada caso não exista uma etiqueta para o tipo de história a anotar. Poderíamos ter usado sempre esta, para todas as histórias a anotar. No entanto, ao definir um conjunto de etiquetas específicas para este propósito, lembramos os transcritores da sua existência e marcamos *slots* a preencher.

```

1 | <ascendencia>
2 | <p>
3 | O meu pai era José Teixeira Bonifácio, era tintureiro na <ref
4 | tipo="empresa">Fábrica dos Carrinhos</ref>, onde ganhava 12$50 por
5 | dia. O meu pai não sabia ler nem escrever, fui eu que lhe ensinei a
6 | escrever o nome. A minha mãe, Emília Tomásia Pereira da Silva, era
7 | doméstica mas tinha ocasiões em que vendia peixe. No Inverno havia o
8 | período das traineiras encostarem, durante 3 meses. Nessa altura não
9 | havia peixe. Os meus pais parece que tiveram 12 filhos, mas alguns
10 | morreram com a meningite. Eu era o mais velho.
11 | </p>
12 | </ascendencia>

```

A imagem 4 apresenta o aspecto de uma história de vida a ser consultada na Internet.

Legendas

A legenda inclui uma secção para cada fotografia ou documento digitalizado com o nome do ficheiro e uma legenda composta não só pela descrição da imagem mas também uma data e, sempre que possível, nomes dos intervenientes. A figura 5 mostra um álbum fotográfico.

```

1 | <fotos>
2 | <foto ficheiro="004-F-07.jpg">

```

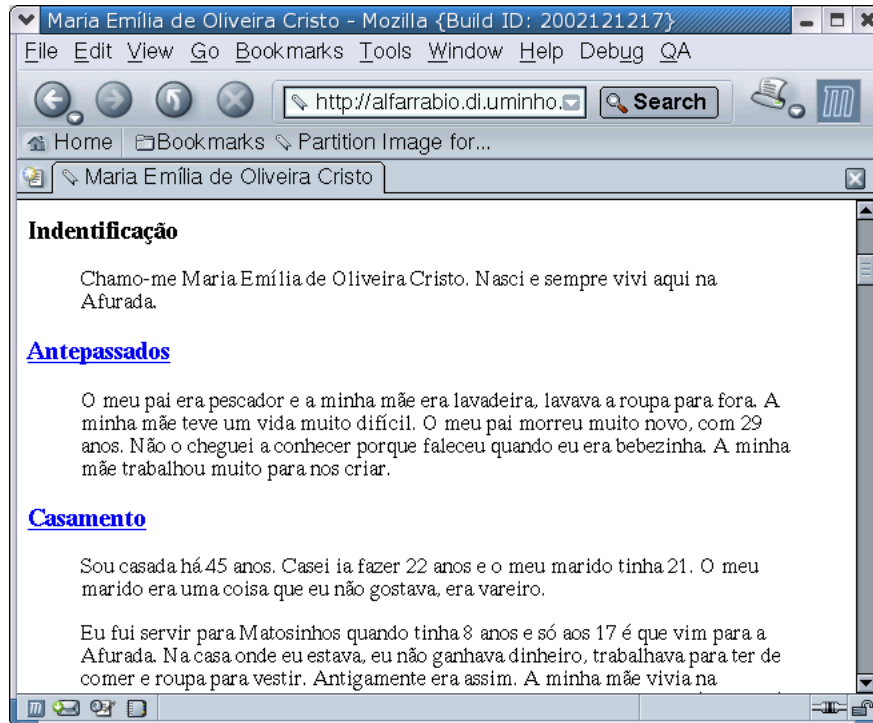


Figura 4. Consulta de História de vida

```

3   <onde>Ribeira, Porto</onde>
4   <quando>1950</quando>
5   <quem>
6     Do lado esquerdo, Francisco da Cruz Lopes, o marido de Maria de
7     Lurdes Pereira Vásquez, e do lado direito, Francisco Moreira dos
8     Santos, o sogro.</quem>
9   <facto>
10    As cheias de 1950 que atingiram a Ribeira. O rio Douro invadiu o
11    mercado.</facto>
12 </foto>
13 <foto ficheiro="004-F-06.jpg">
14   <quando>1925</quando>
15   <quem>Rosalina Pereira, sentada com a filha Maria de Lurdes Pereira
16     Vásquez ao colo, acompanhada pelo marido Isidro Jorge Vásquez.</quem>
17   <facto>Baptizado de Maria de Lurdes Pereira Vásquez. </facto>
18 </foto>
19 </fotos>

```

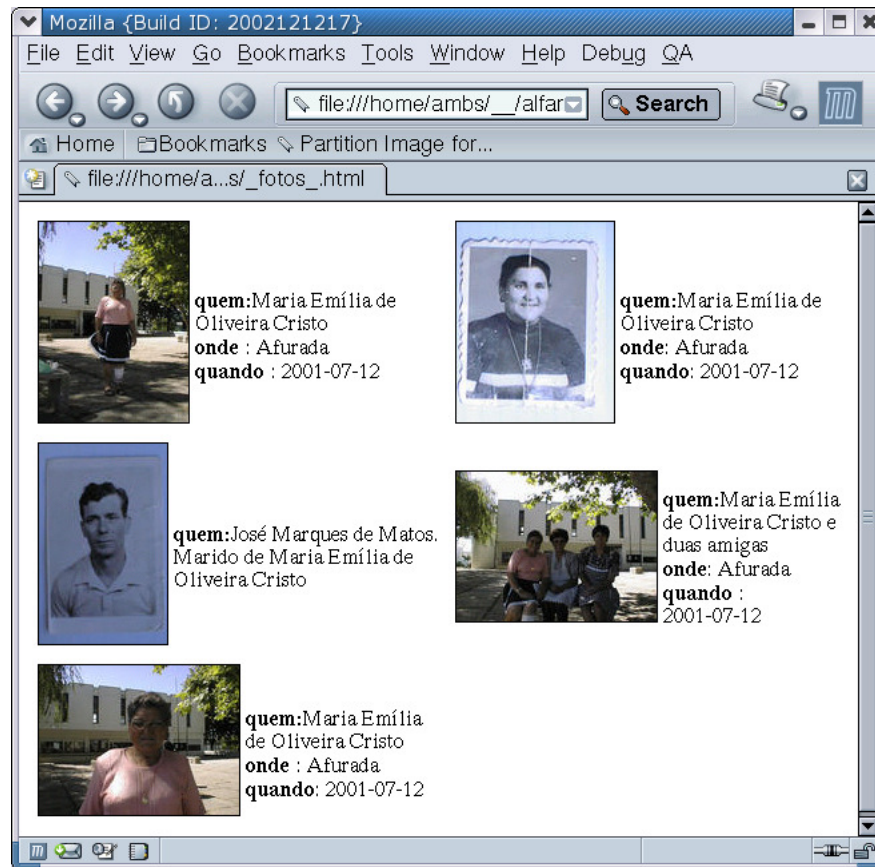


Figura 5. Álbum fotográfico

4 Geração Web

Como o museu da pessoa é virtual, a sua presença é feita especialmente na Internet. Para isso é necessário criar algum tipo de navegação sobre as entrevistas dando ao utilizador vários métodos de encontrar a informação que lhe interessa.

4.1 Navegação por projecto

Dado que os projectos estão organizados hierarquicamente (como descrito na secção 3.1) no sistema de ficheiros torna-se simples a construção de uma árvore e a sua navegação.

A figura 6 mostra uma página típica de um projecto onde, do lado esquerdo, se pode ver um conjunto de ligações para sub-projectos.

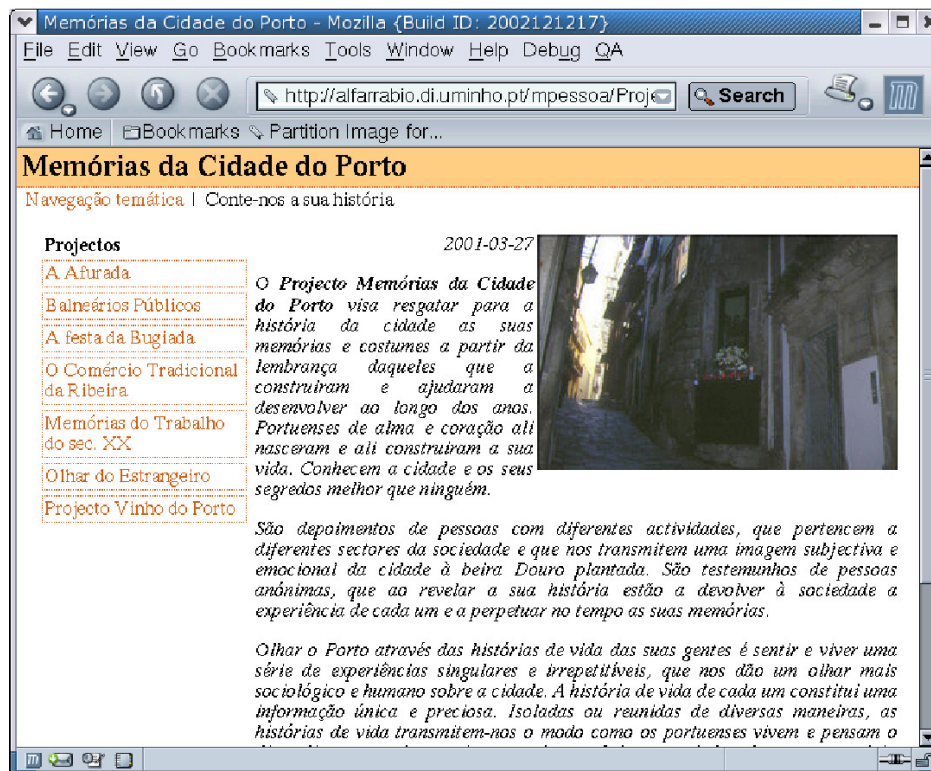


Figura 6. Navegação por projecto

4.2 Navegação temática

Para navegação temática foi utilizado um módulo Perl que utiliza um formato semelhante ao ISO para Thesaurus Multilingue. Este módulo Perl denominado `Biblio::Thesaurus` e alguns outros de suporte a bibliotecas digitais[4] estão a suportar a ontologia de navegação sobre o Museu.

No entanto, a definição de uma ontologia nem sempre é simples. Ou se tem uma equipa a trabalhar na construção de uma que contemple toda a realidade, ou torna-se difícil a sua manutenção.

Uma das vantagens de se ter utilizado XML e portanto, uma etiquetação definida de acordo com as necessidades do projecto, é o de se ter convencionado um conjunto de etiquetas para indicar termos de catalogação (ver a secção 3.2). Deste modo, pode ser feita a extracção automática destas entidades. No caso da navegação conceptual, são extraídos regularmente os termos usados nas histórias para ser criada uma lista de termos não contemplados na ontologia do projecto.

Além da ontologia é usado um catálogo de histórias (construído a partir da etiquetação das histórias) e que faz a ponte entre os termos apresentados na ontologia e os respectivos documentos. Podemos ver o par ontologia/catálogo como uma forma mais flexível para o que poderia ser implementado com Topic Maps.

A figura 7 mostra um resultado da navegação temática do Museu que inclui uma vista da ontologia mas também a possibilidade de uso directo de termos de pesquisa sobre o catálogo.

4.3 Outros recursos on-line

Além da navegação sobre as histórias e a sua consulta, o Museu pretende disponibilizar um conjunto de recursos relacionados com os depoimentos. Para seguir as directivas do Museu é crucial que estes recursos existam mas que sejam automatizados seus processos de construção.

Calendário

Qualquer história que seja recolhida pelo Museu acaba por conter um conjunto razoável de datas: nascimento, baptizado, casamento, volta da guerra, catástrofe e outras. Dados que todas estas datas devem ser correctamente etiquetadas na história, é possível construir ferramentas de extracção automática das datas e construção de um calendário mensal com as efemérides registadas no Museu. A figura 8 é o calendário deste mês, no dia 6 de Fevereiro de 2003.

Eixos Cronográficos

Utilizando as datas extraídas para o calendário e de eventos sociais, políticos e religiosos torna-se possível construir um eixo ou friso cronográfico. Neste eixo marcam-se as datas relacionadas com as histórias para as relacionar com outros eventos.

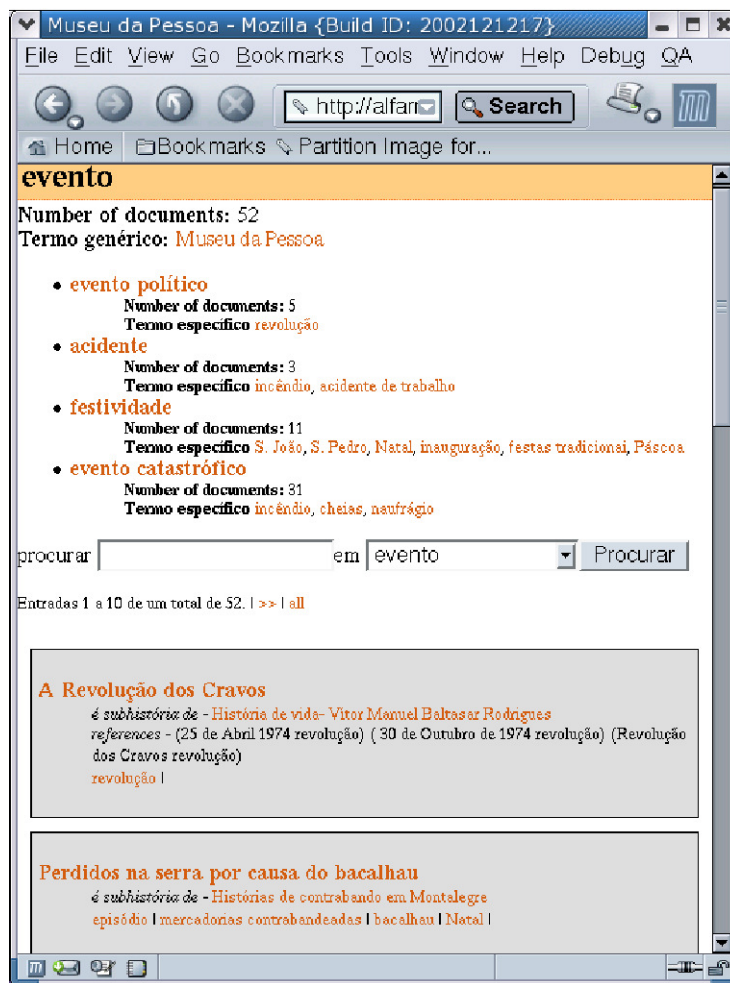


Figura 7. Navegação temática

Fevereiro 2003						
D	S	T	Q	Q	S	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	

13 Fevereiro >>>
A 13 de Fevereiro de 1916
nasceu o Sr. Fausto Ferreira
Gomes

Figura 8. Calendário a 6 de Fevereiro de 2003

Enciclopédia

Determinadas figuras ou personalidades, eventos sociais, políticos e religiosos, ferramentas ou instrumentos típicos, localidades, instituições e outros objectos são anotados para que se possam extrair e, desta forma, construir uma enciclopédia que os explique.

Glossário

Dada a diversidade de expressões usadas nas várias regiões do país é provável que nem todas as pessoas as conheçam. Com a marcação destas expressões e seu significado nas várias histórias recolhidas, estas são extraídas automaticamente para a construção de glossários.

5 Publicação em papel

Embora o Museu da Pessoa seja virtual um dos objectivos é poder publicar histórias noutros suportes como sejam CD-ROMs ou livros. Estes livros são especialmente de dois tipos: livros de depoimentos e colectâneas.

Para a publicação em papel optou-se por gerar \LaTeX a partir dos documentos XML. Esta escolha permite que, por um lado, a geração dos livros possa ser automática (sem necessidade de utilizar ferramentas interactivas), e por outro, que se beneficie de toda a panóplia de módulos e ferramentas que o \LaTeX disponibiliza.

Para as colectâneas tornou-se necessário definir que depoimentos devem ser introduzidos, e por que ordem, no livro resultante. Para isso, definiu-se um novo

formato XML que especifica não só que história deve ser incluída e em que sítio, mas também qual o título, autores, e outras partes fixas como introdução e comentários.

6 Processamento estrutural

Para o processamento de toda estes documentos armazenados no sistema de ficheiros foi desenvolvido um sistema denominado DAG[5] (Directory Attribute Grammars).

Esta ferramenta utiliza um conjunto de regras semelhantes às usadas na escrita de gramáticas de atributos, onde é especificada a estrutura de todo o acervo e, sobre ela, um conjunto de funções de processamento. Na figura 2, apresentada na secção 3.1, podemos ver a semelhança existente entre uma especificação da estrutura arbórea e uma gramática tradicional.

Todos os ficheiros das páginas de Internet, e publicações em papel (documentos PostScript) são construídos por este sistema. Executando a ferramenta, esta irá detectar quais os ficheiros desactualizados e recalculá-los.

A conversão de XML para HTML é feita utilizando um módulo escrito em Perl[6] denominado XML::DT[2].

Uma das principais razões da utilização de métodos de programação convencional para o processamento dos documentos XML deve-se à flexibilidade demonstrada em comparação com ferramentas específicas para o manuseio destes documentos como seja o XSL. Em particular, algumas das funções definidas utilizam ferramentas existentes no sistema operativo, o que nos permite não reinventar a roda.

Por exemplo, para que as imagens possam ser mostradas num álbum fotográfico, as *scripts* de processamento das legendas vai, para cada ficheiro, criar um *thumbnail*. Da mesma forma, está a ser desenvolvido um método que permita aos membros do Museu da Pessoa verificar a ortografia dos documentos, utilizando para esse efeito o corrector ortográfico e analisador morfo-sintáctico Jspell[3].

Tamanho da árvore	283 Mbytes
Tamanho da árvore decorada	534 Mbytes
ficheiros XML	267 ficheiros
ficheiros JPG	524 ficheiros
ficheiros HTML	14 ficheiros
atributos HTML	1 745 ficheiros
atributos JPG	1 056 ficheiros
atributos EPS	533 ficheiros
atributos PS	223 ficheiros
Total de atributos	4 926 ficheiros
Atravessar a árvore	2 minutos
Recalcular a árvore	2:30 horas

A tabela mostra o número de ficheiros, atributos e tempos de geração do *site*. No entanto, é de frisar que estes valores não são da actual versão do museu. Além de outras razões, o tempo que demoraria a recalculer todo o *site* desde o início seria bastante elevado. No entanto, em relação aos tamanhos e números de ficheiros, estes duplicaram estando a árvore decorada a ocupar, actualmente, cerca de 1 GByte.

7 Conclusões e trabalho futuro

Sem dúvida que o projecto Museu da Pessoa é importante para o conhecimento social da população e do país. No entanto, do lado técnico podemos concluir:

- é possível que pessoas de áreas menos ligadas à informática tenham a capacidade de utilizar editores estruturados de forma eficiente;
- embora uma solução temporária, a utilização do sistema de ficheiros para armazenagem dos documentos tem vindo a mostrar-se mais fiável e manuseável do que o esperado;
- o método de cálculo de todo o *site* é bastante eficiente (visto não recalculer ficheiros desnecessariamente) e, ao gerar páginas estáticas, permite que a navegação sobre o Museu não obrigue à geração dinâmica de páginas;
- a publicação de pequenos livros por depoente ajuda não só aos editores de histórias para a sua correcção em qualquer sítio, como veio causar grande entusiasmo aos entrevistados;
- os métodos de navegação e pesquisa, embora em fase de desenvolvimento, permitem a consulta temática do acervo, muito apreciado pelos utilizadores;

Actualmente, centra-se o desenvolvimento em:

- recolha de histórias;
- preparar o acervo audio-visual do Museu para publicação, com a sua digitalização (de filmes e de som) e respectivo tratamento;
- criação de um suporte de gestão integrada do Museu utilizando a Web como Interface;
- análise e desenho de novas estruturas de armazenagem que venham a possibilitar a escalabilidade de todo o *site*.

Referências

1. J. João Almeida, J. Gustavo Rocha, P. Rangel Henriques, Sónia Moreira, and Alberto Simões. Museu da pessoa — arquitectura. In *ABAD*, 2000.
2. J.J. Almeida and José Carlos Ramalho. XML::DT a perl down-translation module. In *XML-Europe'99, Granada - Espanha*, May 1999.
3. Alberto Manuel Simões and José João Almeida. `jspell.pm` — um módulo de análise morfológica para uso em processamento de linguagem natural. In *Actas da Associação Portuguesa de Linguística*, 2001.
4. Alberto Manuel Simões and José João Almeida. `Library::*` — a toolkit for digital libraries. In *ElPub 2002 - Technology Interactions*, 2002.

5. Alberto Manuel Simões, José João Almeida, and Pedro Rangel Henriques. Directory attribute grammars. In *VI Simpósio Brasileiro de Linguagens de Programação*, pages 297–308, 2002.
6. Wall, Larry & Christiansen, Tom & Schuartz, Randal. *Programming Perl*. O'Reilly & Associates, Inc.