Museu Virtual da Emigração e das Comunidades

Projecto Integrado de Engenharia de Linguagens

Relatório Intermédio 2

Bruno Azevedo e Miguel Costa

Mestrado em Engenharia Informática,
Departamento de Informática,
Universidade do Minho

De: 7 de Novembro de 2011
Até: 18 de Março de 2012

Resumo

Este documento representa toda a documentação feita para o desenvolvimento do Museu Virtual da Emigração e das Comunidades. Neste momento contém essencialmente o caso de estudo definido, análise de requisitos, alguma especificação UML e as principais decisões tomadas até ao momento.

Conteúdo

1 Introdução

No âmbito da UCE30 Engenharia de Linguagens do Mestrado em Engenharia Informática, é pretendido para o Projecto Integrado que se desenvolva um Museu Virtual (MV).

Este MV terá de seguir normas internacionais para catalogação de peças de arte e terá uma Ontologia[1] associada. Através desta Ontologia será possível implementar uma característica importante no projecto, que é gerar automaticamente salas de exposição. Assim, o utilizador poderá fazer uma visita orientada por tema/história.

Os conhecimentos que foram e serão adquiridos nos vários módulos desta UCE30, quando aplicados neste projecto irão permitir-nos fazer ferramentas mais genéricas, eficientes e permitir que este MV possa comunicar com outros já existentes. Isto porque irão ser adoptados modelos internacionais para a recolha e partilha de informação.

[1] Ontologia é um termo que foi adoptado pela comunidade de Inteligência Artificial para referir conceitos e termos para descrever alguma área do conhecimento ou construir uma representação desse conhecimento.

2 Contextualização

Hoje em dia, devido ao grande processamento que os computadores possuem, é possível processar enormes quantidades de informação. O problema talvez seja agora arranjar mecanismos de comunicação entre serviços, isto é, encontrar uma linguagem para partilhar informação que qualquer programa consiga entender sem grande esforço.

Devido à dificuldade de transferência de informação, o XML acabou por se tornar uma óptima opção e é considerado o standart para a resolução deste problema, logo, no nosso Museu Virtual (MV), terá de ter mecanismos de recolha de informação de um ficheiro XML de peças para serem expostas e, ainda, exportar essa informação também em XML.

Os Museus para exporem as suas peças, tem de definir regras de catalogação para o tipo de peça que é pretendido descrever, devido à necessidade de uniformizar estas regras, foram criadas diversas normas para os mais diversos objectos que podem ser apresentados num museu, por exemplo, desde monumentos até cartas ou roupa. No nosso caso, vamos adoptar essencialmente os modelos internacionais do CDWA[2] e do CCO[3]. Estes modelos contém uma grande quantidade de elementos que se podem usar, em que alguns deles são obrigatórios, mas para além destes, de acordo com o que precisamos, seleccionámos aqueles que nos pareceram mais relevantes e úteis para o nosso caso de estudo.

3 Objectivos pretendidos

O objectivo da realização deste projecto é desenvolver um Sistema para Geração Automática de Salas de Exposições em Museus Virtuais com base em Ontologias, ou seja, criação de um Web Site que permita ser gerido como um Museu, seja possível alguns utilizadores colocarem peças para exposição, outros com direitos de administração gerirem as salas e o museu e, como finalidade óbvia, mostrar a qualquer pessoa que entre no Web Site, as peças que estão arquivadas.

Por outras palavras, os objectivos principais são:

- arquivar o espólio do museu,
- gerar automaticamente os espaços de aprendizagem na forma de salas de exposição do museu,
- guiar as visitas recebendo delas informação.

Quando falamos em "arquivar o espólio", estamos a referir-nos a receber meta-informação sobre cada objecto, classificá-lo tendo em conta a ontologia e armazenar essa meta-informação depois de classificada.

O MV vai ter várias salas de exposição, em que cada uma delas transmite o conhecimento que relacionada todos os objectos que pertencem à sala. O tema ou conhecimento que é pretendido ser transmitido, será descrito por uma ontologia. É esta ontologia que define os objectos a serem expostos e as relações entre eles.

Dizer "guiar as visitas recebendo delas informação", é a nossa solução basear-se em apresentar o sítio web com base na ontologia da exposição e permitir que o visitante possa fazer um comentário e até fornecer conhecimento que pode ser arquivada no MV, quer acrescentando novo objectos ou ligações, quer corrigindo informações incorrectas.

O nosso repositório para o MV tem de seguir o modelo de referência Internacional OAIS[4] que indica como se deverá estruturar a Figura 1.

Neste repositório terá de ser possível que haja três grandes processos:

- ingestão - processo que permite ingerir conteúdo no sistema.

[2]Categories for the Description of Works of Art (CDWA) - `http://www.getty.edu/research/publications/electronic_publications/cdwa/introduction.html`

[3]Cataloging Cultural Objects (CCO) - `http://cco.vrafoundation.org/`

[4]Open Archivel Information System - consiste numa organização de pessoas e sistemas que tem a responsabilidade de preservar a informação e torna-la disponível para uma comunidade designada.

- administração - gestão de utilizadores, gestão dos arquivos armazenados...
- disseminação - processo que permite ver o conteúdo armazenado. Pode ser através de uma página web ou da criação de um ficheiro XML.

3.1 Modelo de referência internacional OAIS

O modelo que se vai usar para a recepção, disponibilizar e gerir a informação é o OAIS e possui a estrutura que é ilustrada na Figura 1.

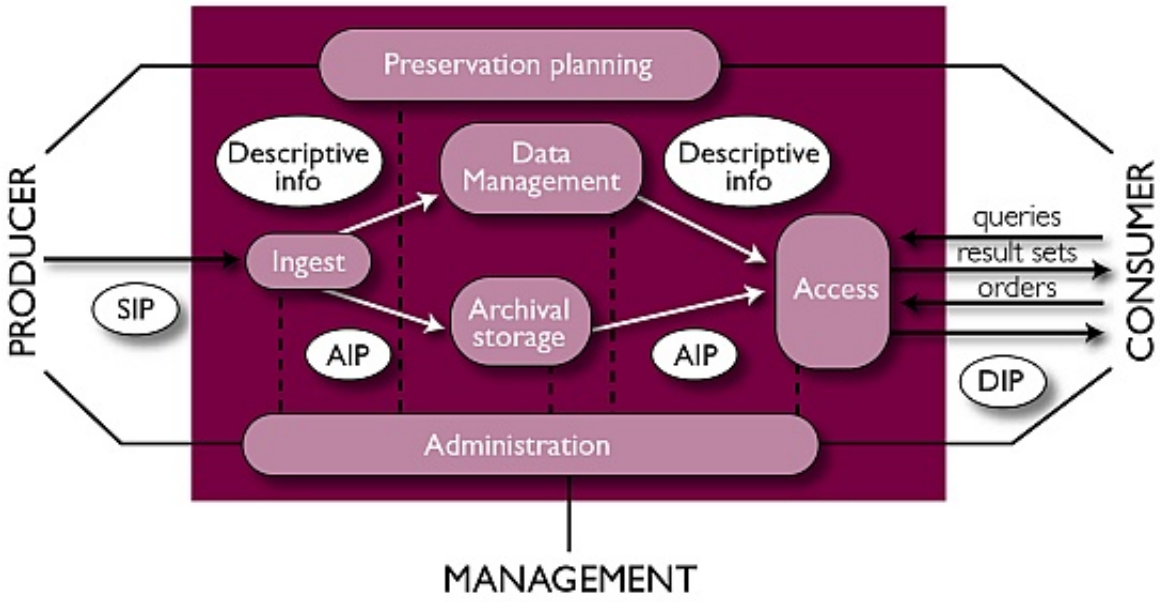

Figura 1: Modelo de referência internacional OAIS

A estrutura baseada neste modelo inclui três actores distintos:

- Produtor - corresponde aos utilizadores que irão submeter informação no repositório.
- Administrador - realiza todas as tarefas de manutenção do sistema.
- Consumidor - utilizador que consulta e pesquisa a informação do repositório.

3.1.1 SIP e o Processo de Ingestão

Neste processo, o ideal é conseguir enriquecer o repositório através do preenchimento de um formulário e ainda através de um ficheiro XML que respeite o Schema do CDWA-Lite.

Quando se inserir alguma informação no MV, essa informação irá sofrer um processo até ficar completamente armazenado no sistema para se conseguir relacionar com outras obras presentes no museu.

3.1.2 AIP e armazenamento de projectos

Neste repositório, vamos ter uma solução híbrida, com a metainformação guardada na Base de Dados relacional e a outra informação que corresponde a outros ficheiros incluídos no SIP que estejam relacionados com as obras, como imagens, documentos word, etc, guardados num Sistema de Ficheiros.

3.1.3 DIP e a disseminação/publicação de conteúdos

Um DIP pode ser disponibilizado de duas formas, ou através de um WebSite criado para o efeito ou então através de um ficheiro XML.

No Web Site será possível ver o espólio do museu, ver a sua informação e ainda entrar em salas que serão geradas automaticamente para mostrar obras que tinham algum fio de ligação entre elas.

4 Caso de Estudo

Depois de um grande impasse sobre o Museu que iríamos tratar, optamos por um projecto já com algum desenvolvimento que é o Museu da Emigração e das Comunidades de Fafe.

Dentro deste Museu existe uma grande variedade de peças de arte, em que grande parte estão relacionadas com a emigração de portugueses no início do século XX. Muitas das relíquias pertenceram a "Brasileiros retorna viagem", pessoas que emigraram para o Brasil, enriqueceram e quando voltaram para Portugal queria ser importantes na sua terra, para isso mandavam construir edifícios (escolas, hospitais...) e alguns até tinha o seu próprio jornal/almanaque.

Do vário material que poderíamos tratar, optámos pelas **fotografias** e pelas **zincogravuras**.

As fotografias foram tiradas pelo fotógrafo e pintor Gérald Bloncourt, de nacionalidade Haitiana, que tirou muitas fotografias ao dia a dia dos emigrantes portugueses em França.

Figura 2: Bidonville à Aubervilliers - 1957

Quanto às zincogravuras, elas foram usadas para diversos fins, como cartons em almanaques, capas de jornais, fotografias de famílias e retratos, logótipos...

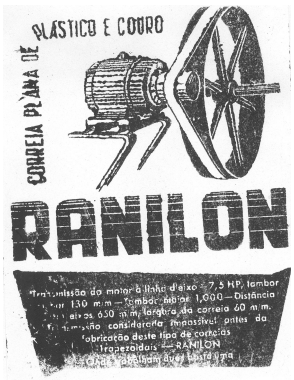

Figura 3: RANILON - Correia plana de plástico e couro

Figura 4: "Director deste Almanaque, esposa e seus 10 filhos" - O Desforço

5 Levantamento de Requisitos

Na explicação do que era pretendido fazer, já descrevemos um pouco os requisitos necessários, fazendo uma síntese dos objectivos temos então:

1. Basear a especificação no estudo e análise das normas internacionais: CCO e CDWA.
2. Criar um modelo para uma Base de Dados Relacional que permita guardar a informação proveniente de documentos com os formatos internacionais já mencionados.
3. Tendo em conta o caso de estudo, encontrar uma solução para inserir a informação no repositório de forma a que as diferentes obras se possam relacionar entre si.
4. Construir um Museu Virtual com salas geradas automaticamente para apresentar as relíquias do repositório aos seus visitantes.

5.1 Especificação dos Requisitos

5.1.1 Dados

Os vários documentos que nos foram fornecidos para analisar, tinham uma estrutura muito equivalente entre as várias peças, por isso, depois de analisar o seu conteúdo tivemos de pensar na forma mais correcta de passar essa informação para os modelos internacionais de catalogação.

Essa transformação nos documentos foi toda feita através dos nossos conhecimentos de scripting (na linguagem Perl).

Exemplo da descrição de uma fotografia dos documentos fornecidos:

```
Inventário              CGB001
Classificação Genérica  Fotografia
Objecto/Documento       Fotografia P/B
Título / Id. pelo       Autor1111/4
Autor/Produção          Gérald Bloncourt
Material/Suporte        Papel
Técnica                 Fotografia P/B - Impressão digital
Dimensões               42 x 59,5 cm
Legenda                 Emigrante português lendo um jornal do seu país num
                        estaleiro de obras nos subúrbios de Paris, 1967.
Datação                 Impressão em 2009
Lugar de Produção       França
História do objecto     Sem moldura. Não foi exposto neste Museu.
Incorporação            Doação do autor ao Museu das Migrações no ano de 2009.
```

Já as zincogravuras, tinham a seguinte estrutura:

```
Denominação/Título:     Zincogravura
Autor/Produção:
Datação:
Dimensões:              12,8 x 10 cm
N° de Inventário:       GI0004
Descrição:              Publicidade: RANILON - Correia plana de plástico e couro
                        Impressa no "Almanaque Ilustrado" 1965:50; 1966:108;
                        1967:25; 1968:76; 1969:114; 1970:108; 1971:90; 1973:24;
                        1975:6; 1976:96; 1977:92; 1978:88; 1980:50; 1981:57;
                        1982:82; 1983:75; 1985:86; 1986:84; 1987:60
Proveniência/Incorporação: "O Desforço"
```

A metainformação das fotografias são mais ricas que as zincogravuras, no entanto é possível cobrir os campos obrigatórios das normas internacionais.

5.1.2 Análise das normas internacionais para a catalogação de obras de arte

Analisando os normas necessários e a sua estrutura, definimos que o modelo a iria ser baseado no XML Schema CDWA-Lite. Este Schema possui praticamente todos os campos necessários para as nossas obras, apenas tivemos de acrescentar mais um ou outro.

Passamos agora a explicar alguns dos campos que irão fazer parte da descrição de uma obra no nosso MV. Há dois tipos de descrição presente, a `Descriptive Metadata` e a `Administrative Metadata`. A primeira é onde colocamos toda a metainformação relacionado com a obra, a segunda é usada para colocar alguns ficheiros que possam estar relacionados com o objecto, como por exemplo imagens e documentos.

Descriptive Metadata:

1. *Object/Work Type Wrapper* - Onde colocamos o tipo de objecto que é a obra (obrigatório).

```
<objectWorkTypeWrap>
    <objectWorkType>Fotografia</objectWorkType>
</objectWorkTypeWrap>
```

2. *Title Wrapper* - Aqui podemos colocar os vários títulos que a obra poderá ter (obrigatório).

```
<titleWrap>
    <titleSet>
        <title>1111/4 - Paris 1967</title>
    </titleSet>
</titleWrap>
```

3. *Display Creator* - Breve informação biográfica do criador (obrigatório).

```
<displayCreator>Gérald Bloncourt</displayCreator>
```

4. *Indexing Creator Wrapper* - Local para colocar toda a informação dos criadores da obra (obrigatório).

```
<indexingCreatorWrap>
    <indexingCreatorSet>
        <nameCreatorSet>
            <nameCreator type="personalName">Gérald Bloncourt</nameCreator>
        </nameCreatorSet>
        <nationalityCreator>Haitian</nationalityCreator>
        <vitalDatesCreator birthdate="1926">1926-</vitalDatesCreator>
        <genderCreator>masculino</genderCreator>
        <roleCreator>Fotógrafo</roleCreator>
        <roleCreator>Pintor</roleCreator>
    </indexingCreatorSet>
</indexingCreatorWrap>
```

5. *Display Measurements* - Breve informação sobre as dimensões e/ou escala da obra.

```
<displayMeasurements>42 x 59,5 cm</displayMeasurements>
```

6. *Indexing Measurements Wrapper* - Toda a informação sobre as dimensões, escala, volume, peso e/ou área da obra.

```
<indexingMeasurementsWrap>
    <indexingMeasurementsSet>
        <measurementsSet value="42" unit="cm" type="width"/>
        <measurementsSet value="59,5" unit="cm" type="height"/>
    </indexingMeasurementsSet>
</indexingMeasurementsWrap>
```

7. *Display Materials/Techniques* - Indicação sobre o material, substância ou técnica de produção usadas na obra (obrigatório).

```
<displayMaterialsTech>Fotografia P/B</displayMaterialsTech>
```

8. *Indexing Materials/Technique Wrapper* - Detalhes sobre a constituição da obra.

```
<indexingMaterialsTechWrap>
    <indexingMaterialsTechSet>
        <termMaterialsTech>preto e branco</termMaterialsTech>
    </indexingMaterialsTechSet>
</indexingMaterialsTechWrap>
```

9. *Style Wrapper* - Termo que identifica o período artístico ou histórico. Apesar de não termos esta informação e não ser um campo obrigatório, optámos por colocar "indefinida" porque no futuro este campo será útil.

```
<styleWrap>
    <style>indefinida</style>
</styleWrap>
```

10. *Display Creation Date* - Descrição da data associada à criação da obra (obrigatório).

```
<displayCreationDate>1967</displayCreationDate>
```

11. *Indexing Dates Wrapper* - Local para colocar as várias datas que estejam associadas com a obra (obrigatório).

```
<indexingDatesWrap>
    <indexingDatesSet>
        <dateQualifier>Impressa</dateQualifier>
        <earliestDate>2009</earliestDate>
    </indexingDatesSet>
    <indexingDatesSet>
        <dateQualifier>Incorporada no Museu da Emigração e das
Comunidades</dateQualifier>
        <earliestDate>2009</earliestDate>
    </indexingDatesSet>
    <indexingDatesSet>
        <dateQualifier>Tirada</dateQualifier>
            <earliestDate>1967</earliestDate>
    </indexingDatesSet>
</indexingDatesWrap>
```

12. *Location/Repository Wrapper* - As várias localizações geográficas que estão relacionadas com a obra (obrigatório).

```
<locationWrap>
    <locationSet>
        <locationName type="creationLocation">França</locationName>
    </locationSet>
    <locationSet>
        <locationName type="currentRepository">Museu da Emigração e das
Comunidades</locationName> </locationSet>
</locationWrap>
```

13. *Classification Wrapper* - Termo utilizado para classificar uma obra, agrupando-a juntamente com outras obras com base em características semelhantes, incluindo materiais, forma, região de origem, o contexto cultural, ou período histórico ou estilístico.

```
<classificationWrap>
    <classification>Fotografia</classification>
</classificationWrap>
```

14. *Inscriptions Wrapper* - Uma descrição ou transcrição de qualquer distinção ou identificação física, anotações, textos, marcações ou etiquetas que são afixadas. No caso das fotografias colocamos a legenda.

```
<inscriptionsWrap>
        <inscriptions>Emigrante português lendo um jornal do seu país num estaleiro de
obras nos subúrbios de Paris, 1967.</inscriptions>
</inscriptionsWrap>
```

Adminsitrativa and Resource Metadata:

1. *Record Wrapper* - Informações sobre o registo que contêm a informação da catalogação.

```
<recordWrap>
    <recordID>CGB001</recordID>
    <recordType>Inventário</recordType>
</recordWrap>
```

2. *Resource Wrapper* - Local para informações sobre as imagens ou outros recursos que servem como substitutos visuais da obra, incluindo imagens digitais, slides, fotografias, vídeos ou áudio.

```
<resourceWrap>
    <resourceSet>
        <linkResource>fotos/CGB001.JPG</linkResource>
        <resourceViewDescription>
            Emigrante português lendo um jornal do seu país
            num estaleiro de obras nos subúrbios de Paris, 1967.
        </resourceViewDescription>
    </resourceSet>
</resourceWrap>
```

Apesar de nos exemplos colocados em XML ser metainformação de uma fotografia, as zincogravuras possuem também uma estrutura praticamente igual a esta. Como é óbvio, os campos obrigatórios existem em ambas as descrições. Nos campos opcionais, algumas obras podem ter mais informação do que outras e isso faz com que haja alguma diferença da informação disponível para cada uma.

6 Concepção/desenho do Sistema

6.1 Ambiente de Trabalho

6.1.1 Linguagens de Programação e Tecnologias

As linguagens que mais vamos usar são o Perl e o PHP. O Perl usamos mais para as scripts e transformação dos documentos para outros formatos, enquanto que o PHP vai ser mais usado para construir o WebSite que permita gerir todo o Museu.

Tendo em conta os conhecimentos e experiência que tivemos noutros projectos, achamos melhor adoptar a utilização de uma Framework que nos facilitasse algumas tarefas na criação/gestão do WebSite. Depois de uma grande análise do que nos podia ser útil, optámos pela Framework em PHP Yii.

> *Yii é um framework de alta performance em PHP que utiliza componentes para o desenvolvimento de grandes aplicações Web. Permite máxima reutilização de códigos na programação Web e pode acelerar significativamente o processo de desenvolvimento. O nome Yii (pronunciado i) representa as palavras fácil (easy), eficiente (efficient) e extensível (extensible).*

Como o projecto tem uma grande parte de desenvolvimento Web, para além do HTML, vamos ainda usar CSS, JavaScript e outras tecnologias associadas ao Ajax.

Usámos ainda a linguagem de marcação XML e várias tecnologias a ela associadas, como: XPath, XML Schema, XML Stylesheet.

6.1.2 Suporte

Linux é o sistema operativo usado para suportar o sistema desenvolvido. Para a Base de Dados adoptámos o MySQL que noutros projectos se revelou-se mais do que suficiente para o que será exigido.

Instalámos nas nossas máquinas o xampp, ferramenta que incluiu um servidor PHP e MySQL para conseguirmos testar o que íamos fazendo.

6.1.3 Editor de Texto/IDE

Para desenvolver um projecto com esta dimensão e com as tecnologias que serão usadas, é essencial recorrer a ferramentas que nos facilitassem algumas tarefas.

O Vim foi muito usado para a criação de scripts em Perl, enquanto que para a escrita da linguagem PHP e HTML vamos usar os IDE NetBeans e Ecplise. Para a criação de ficheiros XML, Stylesheets (XSLT) e XML Schema recorremos ao IDE Oxygen XML.

6.2 Repositório para os Dados

O nosso suporte para a informação será uma Base de Dados Relacional em MySQL, no entanto, será ainda possível utilizar como suporte um ficheiro XML, mas este será usado, principalmente, para a partilhar da metainformação das obras com outros museus que também sigam as normas do CDWA-Lite.

O nosso desenho para a base de dados à primeira vista pode ser um pouco assustador, no entanto o facto de ter tantas tabelas deve-se à estrutura do CDWA-Lite, isto porque o que nós fizemos baseou-se em mapear o CDWA-Lite para uma base de dados relacional na 3ª forma normal. Como muitos dos campos do schema são listas, isso resulta em relações de muitos para muitos, logo aumenta mais um pouco o número de tabelas existente.

Apesar de a nossa base de dados cobrir praticamente tudo o que podem estar num museu, nós apenas vamos usar as tabelas necessárias para o nosso caso de estudo. Fica no entanto a camada de dados pronta, para no futuro poder ser possível expandir o museu sem grande dificuldade.

6.2.1 Descrição da Base de Dados Relacional

Aqui iremos explicar as tabelas presentes na base de dados e a informação que cada uma irá conter sobre uma obra.

6.3 Povoamento da Base de Dados a partir dos documentos fornecidos

Para inserir a informação que nos foi disponibilizada na Base de Dados, o conteúdo dos documentos fornecidos sofreu uma série de transformações até estar pronto a inserir no repositório.

A ideia era criar um mecanismo automático para transformar todos os registos das obras, para isso aplicámos os nossos conhecimentos adquiridos em de SPLN (Scripting no Processamento de Linguagem Natural) para criar scripts que fizessem o trabalho todo de uma forma eficaz e rápida.

Apesar de o resultado final ser equivalente nas fotografias e nas zincogravuras, o processo de transformação não foi exactamente o mesmo devido à estrutura dos documentos disponibilizados.

6.3.1 Zincogravuras

No caso dos registos das zincogravuras, a informação passou por duas fases.

Inicialmente os documentos estavam no formato do Word (doc), por isso começamos por passar essa informação para um ficheiro XML de acordo com o CDWA-Lite. Esta primeira transformação foi feita, tal como todas as outras, em Perl e recorremos à utilização do módulo `Text::Extract::Word` para extrair o texto do documento, depois através de expressões regulares "apanhamos" o texto que queríamos e associámos às respectivas tags e atributos do Schema do CDWA-Lite. Depois de ter a informação num formato XML, foi mais simples de passar os registos da zincogravuras para SQL de forma a que se pudessem a inserir na base de dados.

6.3.2 Fotografias

Os documentos em Word que tinham os registos das fotografias, até chegar a SQL sofreu mais processos de transformação devido à informação no doc estar em várias tabelas.
O primeiro passo foi ficar apenas com o texto "limpo". Para isso guardamos o documento em PDF, em seguida utilizamos o comando em Linux `pdftohtml ficheiro.pdf` para passar a informação para HTML e quando estava neste formato, apenas executamos o comando `lynx -dump ficheiro.html` para ficar com o texto limpo.

Depois de termos os registos em apenas texto, passamos a informação para um ficheiro XML com uma estrutura criada por nós:

```
<peca>
    <inventario>CGB001</inventario>
    <classificacao_generica>Fotografia</classificacao_generica>
    <objecto>Fotografia P/B</objecto>
    <titulo>1111/4</titulo>
    <autor>Gérald Bloncourt</autor>
    <suporte>Papel</suporte>
    <tecnica>Fotografia P/B - Impressão digital</tecnica>
    <dimensoes>42 x 59,5 cm</dimensoes>
    <legenda>
        Emigrante português lendo um jornal do seu país num estaleiro de
        obras nos subúrbios de Paris, 1967.
    </legenda>
    <datacao>Impressão em 2009</datacao>
    <lugar_producao>França</lugar_producao>
    <historia>Sem moldura. Não foi exposto neste Museu.</historia>
    <incorporacao>Doação do autor ao Museu das Migrações no ano de 2009.</incorporacao>
</peca>
```

A partir daqui foi mais simples criar outro documento XML de acordo com o Schema do CDWA-Lite e deste para SQL até ser inserido na base de dados.

7 Alternativas, Decisões e Problemas de Implementação

8 Conclusões

A Scripts em Perl