XML na Demografia Histórica: Anotação de Registos Paroquiais

Rafael Fernandes Félix[1]
WeDo Consulting, Braga, Portugal
rafael.felix@wedoconsulting.com

Fernanda Faria[2]
Núcleo de Estudos da População e Sociedade, Universidade do Minho, Campus de Azurém, Portugal
ffaria@neps.ics.uminho.pt

Maribel Yasmina Santos
Departamento de Sistemas de Informação, Universidade do Minho, Campus de Azurém, Portugal
maribel@dsi.uminho.pt

Pedro Rangel Henriques
Departamento de Informática, Universidade do Minho, Campus de Gualtar, Portugal
prh@di.uminho.pt

Resumo

O Método de Reconstituição de Paróquias, no qual os Historiadores Demógrafos do Núcleo de Estudos da População e Sociedade (NEPS) baseiam o seu trabalho para analisar o comportamento das populações ao longo dos quatro últimos séculos - através do estudo de factores como a natalidade, a fecundidade, a nupcialidade, a mortalidade e a mobilidade -, assenta na tratamento dos registos paroquiais de baptizados, casamentos e óbitos. Da leitura local e exaustiva desses documentos são extraídos os dados que permitem fazer a reconstituição das famílias. Uma vez estáveis, os dados são armazenados em Bases de Dados Paroquiais, que são posteriormente fundidas numa única Base de Dados Central sobre a qual actuam as ferramentas de análise.

A norma XML (*eXtensible Markup Language*) define um método de anotação de documentos, estabelecendo os princípios gerais de estruturação de textos e a sintaxe das marcas a usar. Documento original e marcas formam um ficheiro único de texto ASCII puro, independente de qualquer plataforma de *hardware/software*. A anotação estabelece a estrutura do documento e dá interpretação a determinados elementos nele contidos, sem qualquer informação de formatação ou transformação—é, portanto, também independente da aplicação final. Ao contrário de outros sistemas de anotação, a norma XML não estabelece um conjunto de anotações fixo, para um caso específico ou para todos os casos; ao invés é uma metalinguagem que indica como definir as anotações próprias de cada família de documentos.

Neste artigo é apresentada uma aplicação do XML ao domínio da Demografia, nomeadamente na Anotação dos Registos Paroquiais analisados pelos Historiadores Demógrafos. Concretamente, discute-se o desenvolvimento de um XML-Schema para definir uma instância de XML (a linguagem Schema-RP) para anotação dos Registos Paroquiais. Assim é possível construir uma Base de Dados Documental que contém, em formato electrónico, os registos originais devidamente marcados. O conteúdo desta Base de Dados apresenta-se mais rico quando comparado com as respectivas Bases de Dados Paroquiais, uma vez que permite: i) a reconstrução das fontes históricas originais, reproduzindo-as em diferentes meios, facilitando a sua disseminação; ii) a extracção dos dados necessários ao carregamento das referidas Bases de Dados Paroquiais; iii) a análise dos documentos por diversas áreas de estudo, como por exemplo, a linguística.

Este artigo evidencia ainda o processo de edição anotada dos documentos e exemplifica duas transformações possíveis dos mesmos: a visualização em HTML; e a geração automática de SQL para alimentação das Bases de Dados.

---

[1] Trabalho desenvolvido no âmbito da disciplina de projecto (Opção III), enquanto finalista da LESI.
[2] Bolseira da FCT ao abrigo do programa SAPIENS 99.

Palavras-chave: XML, XML-Schemas, XSL, Demografia, Registos Paroquiais, HTML, Base de Dados, SQL.

1. Introdução

O SEED (Sistema para Estudo da Evolução Demográfica) é um projecto de investigação aplicada no qual um grupo de docentes de informática, da Universidade do Minho, vem trabalhando de há cinco anos a esta parte, em estreita colaboração com o NEPS (Núcleo de Estudos da População e Sociedade) da mesma Universidade. Pretende-se com este projecto aplicar um conjunto de novas tecnologias, para análise de dados, extracção e gestão de conhecimento, ao domínio da Demografia.

Como ponto de partida para o SEED foi proposta uma arquitectura de cinco níveis [Rodrigues, 2000] [Rodrigues et al, 1999], que contempla desde o carregamento das Bases de Dados Paroquiais até ao Sistema de Gestão de Conhecimento, que integra as regras extraídas dos dados pelas várias ferramentas de análise [Rodrigues et al, 2001].

Para complementar o trabalho até ao momento desenvolvido tornou-se necessário estudar soluções que facilitem a aquisição dos dados, a partir das fontes históricas consultadas. Neste contexto idealizou-se a utilização de um computador portátil associado a um *scanner*, de modo a permitir a digitalização local (nos arquivos) dos documentos paroquiais. Esta solução requer por um lado a utilização de ferramentas de reconhecimento de caracteres (OCR) [Oliveira, 2000], capazes de lidar com manuscritos degradados e de diferentes épocas, e por outro, a escolha de um formato de representação digital desses documentos que possibilite o seu armazenamento e posterior tratamento.

A dificuldade em encontrar um sistema OCR, eficiente e a um custo acessível, adequado aos requisitos destes documentos, conduziu-nos a suspender temporariamente esta questão da digitalização directa dos mesmos. Optámos, então, por nos debruçarmos a fundo sobre o problema da estruturação e representação interna dos registos paroquiais transcritos manualmente.

A experiência adquirida no manuseamento dos documentos antigos dos arquivos distritais e na construção de bibliotecas digitais, sugeriu a utilização de um Sistema de Anotação Declarativa segundo a norma XML (*eXtensible Markup Language*) [W3C, 2001][Harold et al, 2001][Eckstein et al, 2001].

Dos resultados desta decisão - aplicação do XML aos registos paroquiais para suporte a análise demográfica - se dá conta neste artigo. A secção 2 apresenta resumidamente o problema da demografia e o método MRP (Método de Reconstituição de Paróquias) [Amorim, 1991] adoptado pelo NEPS para identificar e caracterizar os registos paroquiais considerados. A secção seguinte, secção 3, apresenta a linguagem Schema-RP, uma instância XML, definida a partir de um XML-Schema [Duckett et al, 2001] para anotação dos referidos registos. A este propósito introduz-se o XML e os conceitos de DTD (*Document Type Definition*) [Ramalho, 2000] e XML-Schema, comparando-se, ainda, soluções alternativas para o desenho dessa linguagem. Prossegue-se, secção 4, com a discussão dos problemas de edição e de validação de documentos XML, apresentando-se a solução baseada no editor XML-Spy, entre outras. A secção 5 debruça-se sobre a problemática do processamento semi-automático dos registos paroquiais recorrendo a folhas XSL (eXtensible Stylesheet Language) [W3C, 2001][Bradley, 1999]. Por último, a secção 6 sistematiza o trabalho descrito, apresentando, ainda, algumas propostas de trabalho futuro.

2. Os Registos Paroquiais nos estudos demográficos

A demografia é a ciência da estatística da população. A demografia histórica estuda a estatística da população ao longo dos anos, ou seja, adiciona à demografia a componente histórica. A demografia

histórica estuda, geralmente, fenómenos sociais relacionados com a natalidade, a nupcialidade, a mobilidade e a mortalidade. Assim, o principal objecto de estudo desta ciência é o indivíduo em particular e as famílias constituídas por esse mesmo indivíduo (uma ou mais).

Na Europa e, consequentemente, em Portugal, as principais fontes de informação que permitem o estudo dos fenómenos sociais atrás referidos são os registos paroquiais, pois desde meados do séc. XVI há uma relativa continuidade da existência desses registos.

Existem três tipos de registos paroquiais: registos de baptismo, registos de casamento e registos de óbito. Os registos de baptismo contêm geralmente a seguinte informação: nome do baptizando (apenas o primeiro nome), nomes dos pais, data de baptismo, local de baptismo, nomes dos padrinhos. Os registos de casamento contêm a data e local de casamento, os nomes dos nubentes, as localidades de origem dos nubentes, os nomes dos pais e os nomes dos padrinhos. Os registos de óbito contêm o nome do indivíduo falecido, a data e local do óbito, entre outros dados.

Contudo, a maior parte das vezes, a informação que os três tipos de registo fornecem encontra-se ilegível, ou incompleta. Além disso, devido a erros por parte do pároco, há ainda a eventualidade dos dados inscritos nos registos não corresponderem exactamente à verdade. É também importante referir que um mesmo indivíduo pode ser identificado nos registos de várias formas, por exemplo num registo de baptismo é referido como Maria, num registo de casamento é referido como Maria Gonçalves, noutro registo de casamento é referido como Maria de Carvalho e no seu registo de óbito é referido como Maria de Carvalho Gonçalves.

Para efectuar a recolha dos dados existentes nos três tipos de registos paroquiais, tendo em consideração que um mesmo indivíduo se encontra uma vez nos registos de baptismo, zero ou várias vezes nos registos de casamento e uma vez nos registos de óbito (sendo necessário proceder à sua identificação inequívoca para evitar a duplicação de indivíduos e, em consequência, a duplicação de famílias), surgiu o Método de Reconstituição de Paróquias (MRP) desenvolvido por Norberta Amorim[3] [Amorim, 1973] [Amorim, 1991].

O MRP tem sido utilizado em algumas dezenas de estudos demográficos realizados em Portugal nas últimas décadas. Baseia-se na reconstituição de famílias por cruzamento dos registos de baptizados, casamentos e óbitos, procedendo à desagregação automática dos filhos de cada família num ficheiro de indivíduos. Este ficheiro, armazenando todos os indivíduos que nasceram, casaram ou faleceram numa determinada comunidade, juntamente com o de famílias, constitui uma Base de Dados Paroquial (BDP) ou Central [Costa et al, 2000].

Todo o trabalho de recolha de dados, preenchimento das fichas de família (incluindo o cruzamento da informação) e tratamento dos dados era feito manualmente. Para facilitar algumas tarefas do MRP, nomeadamente a realização dos diversos cálculos envolvidos na análise, foi criada uma aplicação informática em DBaseIII que permitiu a introdução e armazenamento dos dados em formato digital, e que tem sido bastante usada. Entretanto e a par com o projecto SEED [Lopes, 1999], acima referido, foi desenvolvida uma nova aplicação, com interface gráfica e suportada pelo *Microsoft ACCESS,* que permite a inserção e cruzamento dos dados numa base de dados relacional--o SRP (Sistema de Reconstituição de Paróquias) [Ferreira, 2001].

Contudo, a passagem da informação dos registos paroquiais para uma base de dados deixa "pelo caminho" informação que poderia ser importante para diversos outros estudos históricos complementares sobre a caligrafia, a evolução da língua portuguesa, a alfabetização, etc. Assim, o ideal seria armazenar o documento integral, em formato digital, guardando o seu conteúdo intacto no estilo e forma original, de modo a que pudesse ser explorado por ferramentas informáticas diversas para dele extrair toda a informação relevante a cada investigação particular. Para tornar

[3] Norberta Amorim - Professora catedrática do Instituto da Ciências Sociais da Universidade do Minho e coordenadora do NEPS - Núcleo de Estudos da População e Sociedade.

essa intenção uma realidade, viabilizando a exploração automática referida, tem de se recorrer à *Anotação declarativa de Documentos*, como se propõe e explica na secção seguinte.

3. Anotação de Registos Paroquiais em XML

A anotação de documentos "*consiste na identificação dos blocos estruturais e na 'marcação' de elementos no texto com um determinado significado, acrescentando assim, uma descrição ('etiqueta') da informação contida nos mesmos a que podemos chamar meta-informação"* [Félix, 2002]. Uma estratégia possível, hoje em dia muito em voga, para anotação de documentos é o XML, que se define como um método para lidar com informação semi-estruturada. O XML é uma forma de partilhar informação independentemente do meio e do modo como será visualizada por cada utilizador.

Nesta secção apresenta-se a linguagem XML, Schema-RP, concebida para anotar os registos paroquiais, discutindo-se o seu desenvolvimento (hipóteses e decisões tomadas em cada fase). Antes porém comparam-se as duas formas possíveis para descrever uma linguagem XML: o DTD ou o XML-Schema.

a) Especificação de Documentos XML: DTD vs XML-Schema

Uma vez que o XML permite a partilha de informação independentemente do tipo de aplicação usada para a visualizar, é necessário identificar a estrutura do documento XML em questão.

A estrutura de um documento XML pode ser definida associando-lhe um DTD ou um XML-Schema. Ambos definem *a estrutura do documento XML, as anotações permitidas, os atributos disponíveis...* [Félix 2002], contudo o DTD (o mais antigo, herdado do SGML [Herwijnen, 1994]) não segue a notação XML e não permite nem a definição de tipos nem os mecanismos de herança e modularidade introduzidos posteriormente em XML-Schema, como se explicará a seguir.

Um DTD [Kimber, 1997] é um conjunto de declarações que especificam um tipo de documento. O tipo caracteriza uma família, ou conjunto, de documentos (as instâncias) do mesmo género. Assim, um DTD:

-Define a estrutura de um documento, à custa de identificar os elementos (blocos que correspondem a componentes distintas, ou segmentos com significado especial) que o constituem;

-Especifica as anotações disponíveis para marcar cada um dos elementos do documento;

-Para cada elemento, especifica os atributos que lhe estão associados, o domínio de cada atributo e os seus valores por omissão;

-Para cada elemento, define ainda a estrutura do seu conteúdo: que subelementos tem; em que ordem; onde é que pode aparecer texto normal; onde é que pode aparecer dados que não sejam texto.

Um XML-Schema [Buck, 1999] é uma alternativa ao DTD, para a descrição da estrutura de um documento XML; no entanto, apresenta algumas vantagens sobre este, nomeadamente:

-a utilização da sintaxe do XML;

-o suporte de tipos de dados, simples e complexos, para especificar elementos e atributos;

-a extensibilidade, graças à hierarquia de tipos e à modularidade.

Como se disse acima, o suporte a tipos de dados, talvez a mais importante novidade do XML Schema, usa-se para especificar a forma e o conteúdo dum documento, servindo para validar os dados pois podem-se formular restrições aos valores concretos das instâncias.

Como o XML-Schema utiliza a sintaxe do XML, não é necessário aprender outra linguagem e qualquer editor de XML pode ser usado para o desenvolver; de igual forma, um *parser* de XML normal é depois usado para validar sintacticamente a especificação escrita. Além disso, o XML-Schema é extensível, o que permite a reutilização duma especificação através dos mecanismos de herança, segundo uma hierarquia de tipos, em tudo semelhante às abordagens orientadas a objectos.

Neste projecto optou-se pela utilização do XML-Schema para a descrição da estrutura da família de documentos XML em causa—os registos paroquiais--, dadas as suas vantagens em relação ao DTD.

b) O desenvolvimento do Schema-RP: alternativas e opções

A linguagem Schema-RP não surgiu imediatamente após a primeira análise dos registos paroquiais; aliás a princípio havia mesmo a ideia de definir um sistema de anotação diferente para cada um dos 3 tipos de registos. Ao contrário e à medida que a análise foi progredindo, a definição do XML-Schema foi evoluindo num processo iterativo, passando pelas fases abaixo descritas. Em cada alternativa foram-se anotando vários registos para se recolher o *feedback* necessário para refinar a solução da etapa seguinte:

1. Na primeira fase, o *schema* usa elementos (ex. noivo, noiva) que identificam a "personagem" a quem pertencem os dados (conteúdo anotado, tal como, nome, idade, ou profissão). De acordo com esta aposta, um registo anotado teria o seguinte aspecto:

> ... na minha presença compareceram os nubentes <**noivo**><**nome**>António Pereira Madruga</**nome**></**noivo**> e <**noiva**><**nome**>Roza da Conceição do Coração de Jesus</**nome**></**noiva**>, os quais sei serem os próprios, com todos os papeis de estylo correntes, e sem impedimento algum canonico ou civil para o casamento, elle <**noivo**>de edade de <**idade**>vinte e hum annos completos</**idade**>, <**estadocivil**>solteiro</**estadocivil**>; <**profissao**>agricultor</**profissao**>, ... </**noivo**>, e ella <**noiva**>de edade de <**idade**>vinte e quatro annos</**idade**>, <**estadocivil**>solteira</**estadocivil**>, de profissão <**profissao**>domestica </**profissao**> ... </**noiva**> ...

Embora seja a que mais naturalmente surge no início, esta abordagem obriga a que o mesmo elemento seja aberto e fechado várias vezes, pois os dados de determinado indivíduo não aparecem todos seguidos, como se pode verificar no exemplo acima.

2. Na segunda fase, procurou-se utilizar uma abordagem "orientada aos objectos" em que o nome dum elemento identifica a "personagem" e o respectivo dado pessoal (ex: noivo.idade, noiva.nome). De seguida, é apresentado um exemplo da anotação resultante da escolha desta fase:

> ... na minha presença compareceram os nubentes <**noivo.nome**>António Pereira Madruga</**noivo.nome**> e <**noiva.nome**>Roza da Conceição do Coração de Jesus</**noiva.nome**>, os quais sei serem os próprios, com todos os papeis de estylo correntes, e sem impedimento algum canonico ou civil para o casamento, elle de edade de <**noivo.idade**>vinte e hum annos completos</**noivo.idade**>, <**noivo.estadocivil**>solteiro</**noivo.estadocivil**>; <**noivo.profissao**>agricultor </**noivo.profissao**>, ..., e ella de edade de <**noiva.idade**>vinte e quatro annos </**noiva.idade**>, <**noiva.estadocivil**>solteira</**noiva.estadocivil**>, de profissão <**noiva.profissao**>domestica</**noiva.profissao**>, ...

Um inconveniente encontrado nesta abordagem é a explosão da quantidade de elementos usados, além da ortografia de cada elemento que é manifestamente pesada.

3. Por último, utilizaram-se elementos genéricos, que designam o dado em si, conjugados com atributos que, entre outras coisas, indicam a "personagem" a que dizem respeito. Como exemplo de elementos temos: nome, idade, data, local, profissão, etc. Os atributos relacionam os dados com a "personagem" como se mostra nos exemplos seguintes: nomedeQuem (pode ser = a "noivo" ou a "noiva"); localdeQue (pode ser = a "casamento" ou a "baptismo").

A anotação resultante desta abordagem (final) é a ilustrada no seguinte exemplo:

> Aos **<data** deQue="casamento">vinte e quatro dias do mez de Fevereiro do anno de mil oitocentos setenta e seis**</data>**, n'esta **<local** deQue="casamento">Igreja paroquial Matriz da Santíssima Trindade da Villa Lages do Pico, Diocese de Angra**</local>**, na minha presença compareceram os nubentes **<nome** deQuem="noivo">António Pereira Madruga**</nome>** e **<nome** deQuem="noiva">Roza da Conceição do Coração de Jesus**</nome>**, os quais sei serem os próprios, com todos os papeis de estylo correntes, e sem impedimento algum canonico ou civil para o casamento, elle de edade de **<idade** deQuem="noivo"**>**vinte e hum annos completos**</idade>**, **<estadocivil** deQuem="noivo"**>**solteiro**</estadocivil>**; **<profissao** deQuem="noivo"**>**agricultor **</profissao>**, natural **<local** deQue"naturalidade" deQuem="noivo">d'esta freguesia**</local>**, **<local** deQue"morada" deQuem="noivo">n'ella morador **</local>** e **<local** deQue"baptismo" deQuem="noivo">na mesma baptisado**</local>**, filho **<filiacao** deQuem="noivo">legítimo**</filiacao>** de **<nome** deQuem="paiNoivo">Manuel Vicente**</nome>** e de **<nome** deQuem="maeNoivo"> Maria Perpetua**</nome>**, naturais **<local** deQue="naturalidade" deQuem="paiNoivo maeNoivo">d'esta freguesia**</local>**, e ella de edade de **<idade** deQuem="noiva">vinte e quatro annos**</idade>**, ...

Definida a estrutura do *schema* para o caso dos registos de baptismo e dado que existem três tipos de registos paroquiais--baptismo, casamento e óbito--colocava-se um novo problema: criar um *schema* para cada tipo de registo, ou criar um "*super-schema*" capaz de suportar os três tipos de registos (uma vez que a sua estrutura é muito semelhante).

A utilização de um *schema* para cada tipo de registo obriga o utilizador a um maior esforço de aprendizagem e de anotação (na medida que tem de sobreviver com as pequenas diferenças associadas a cada caso). Por outro lado, a utilização de um só *schema* para os três tipos de registos implica um grande conjunto de valores possíveis para os atributos de cada elemento. Por exemplo, num registo de baptismo, para o elemento nome, o atributo deQuem poderia tomar um dos valores da lista {indivíduo, pai, mãe, avô, avó, padrinho, madrinha, assistente, padre}; no entanto, ao utilizar um "*super-schema*", os valores desse mesmo atributo de nome teriam de incluir, para além dos que constam da lista anterior: {noivo, noiva}.

Uma vez que uma das principais preocupações deste projecto é a facilidade de utilização da linguagem de anotação (para que os demógrafos não se sintam intimidados e por isso recusem o sistema), a opção recaiu sobre a última alternativa: o recurso a um único *schema* para os três tipos de registos.

Na Figura 1, mostra-se, esquematicamente, o *XML-schema* obtido. Esse diagrama, de leitura muito directa, poupa-nos aos detalhes da especificação que estão escritos ao pormenor na versão textual e deixa-nos aperceber rapidamente dos elementos que constituem um documento anotado em Schema-RP e da forma como se combinam entre si. Assim, um livro de registos tem uma abertura, um ou mais registo e um fecho. Por sua vez, um registo tem uma notaMargem, um texto, uma ou mais assinatura e, opcionalmente, uma declaração. O texto contém, por qualquer ordem, zero ou mais elementos do conjunto {data, local, sexo, nome, hora, filiacao, profissao, estadoCivil, idade}.

4. Edição e Validação dos Registos Paroquiais em XML

Nesta secção é apresentada a problemática associada à edição e validação dos registos paroquiais em XML. São ainda descritas as características do editor XML que deve ser utilizado.

a) Edição e Validação de documentos Schema-RP

A anotação XML dos registos paroquiais, de acordo com a linguagem Schema-RP definida na secção anterior, é uma operação que, embora apoiada por uma ferramenta informática apropriada—um editor de XML—, requer a intervenção directa dum operador (neste contexto, um demógrafo).

Recorde-se que essa tarefa consiste na inserção das marcas (identificação do início e fim de cada bloco de texto que corresponde a um elemento) ao longo do texto simples que constitui a versão inicial do documento.

Idealmente essa operação deveria seguir-se à digitalização e reconhecimento dos registos, o que permitiria obtê-los em formato digital automaticamente (isto é, sem ninguém ter de o escrever). Contudo e pelas razões apontadas na Introdução, essa tarefa ainda não é exequível e portanto o texto base tem de ser introduzido manualmente pelo operador; na prática, tal tarefa pode ser feita previamente ou em simultâneo com a anotação, pelo que muitas vezes se confundem estas duas operações—introdução e marcação do texto—com a operação de *edição* .

Conceptualmente, a anotação passa por duas etapas:

1 - anotação dos elementos, isto é, inserção das marcas de início e fim de cada elemento;

2 - anotação dos atributos e valores, que corresponde a acrescentar a cada marca os atributos que lhe estão associados, inicializando cada um com o valor que lhe pertence em cada contexto.

Na prática é vulgar confundir estas duas etapas e fazê-las em simultâneo, embora tal seja deixado ao critério da experiência e vontade do operador.

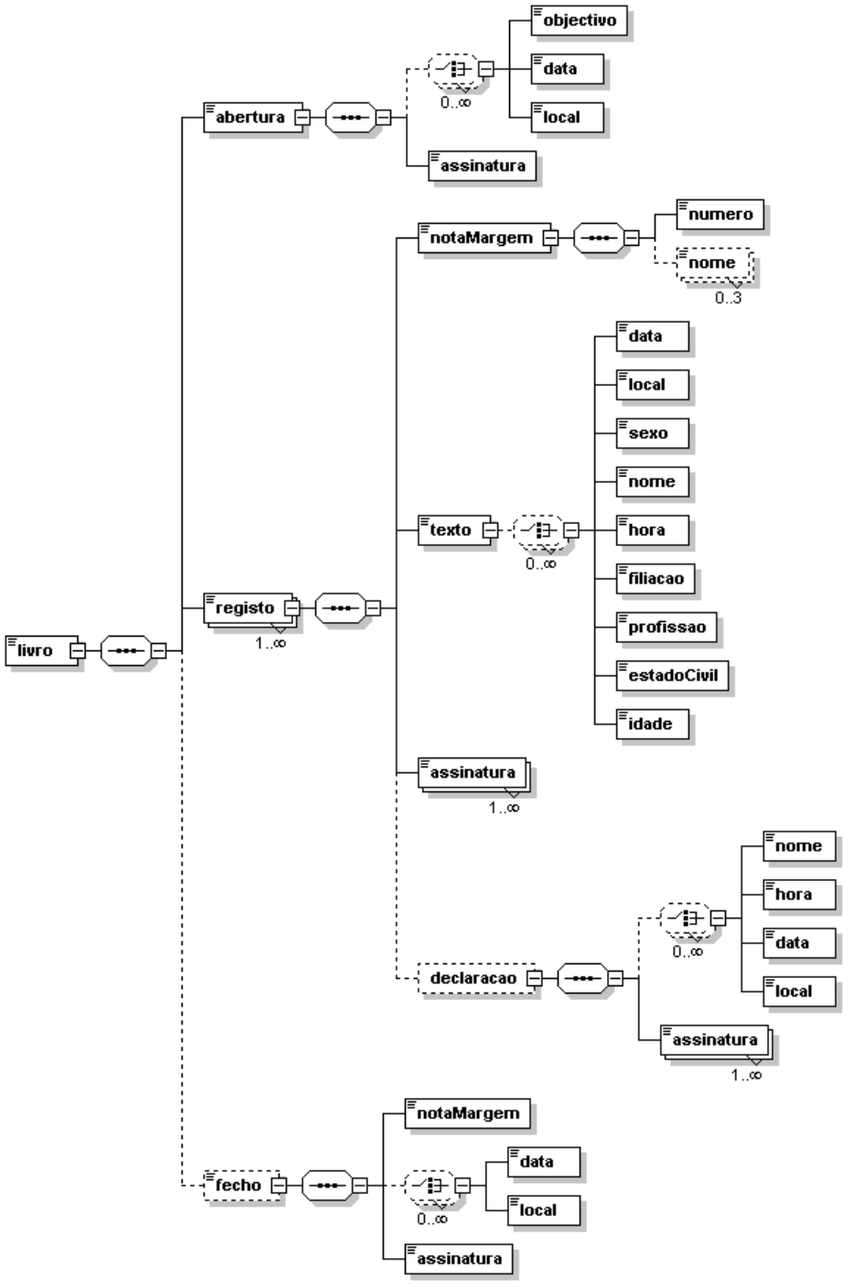

Figura 1 - Estrutura do Schema-RP

Terminada a escrita do texto base e a sua anotação, e antes de se proceder ao seu armazenamento em ficheiro e subsequente processamento, é necessário verificar se o documento está correctamente anotado, isto é, validar: se o nome de todas as marcas é conhecido e pode ser usado no respectivo contexto; se todas as marcas que se abrem, fecham; se a ordem de abertura e fecho, o aninhamento,

está bem feito; etc. Por outras palavras, é necessário validar a boa formação do documento face ao XML Schema que define a linguagem em causa.

A realização desta tarefa é executada de forma totalmente automática por um analisador sintáctico (um *parser*) de XML, o qual é independente do editor, mas deve integrar com ele.

b) Características desejáveis num Editor de XML

Na medida em que actualmente existem no mercado muitos editores de XML, parametrizáveis pelo DTD ou XML Schema que se quer usar, não é de todo necessário desenvolver um específico para a família de documentos que se quer processar. Porém a selecção do editor é muito crítica para o sucesso do processo em geral; como é óbvio, essa escolha deve ser realizada com a preocupação de que as características do editor/validador sejam as mais apropriadas para auxiliar o utilizador no processo de anotação dos documentos. Nomeadamente, o editor a adoptar deve:

-permitir indicar facilmente o DTD, ou XML Schema, a usar;

-permitir escrever com destreza o texto base, ou abrir um que já exista (seja qual for a sua proveniência);

-apresentar os elementos e atributos que podem ser inseridos em cada momento, tendo em conta o contexto e posição do cursor no documento;

-apresentar o conjunto de valores possíveis para determinado atributo;

-fazer a validação do documento face ao respectivo DTD/*Schema*, automaticamente (incrementalmente ou não) ou sob pedido, integrando o *parser* de forma transparente para o utilizador.

c) A solução em XML-Spy

O XML-Spy, da *Altova GmbH & Altova, Inc.,* por satisfazer plenamente e com elevada qualidade os requisitos acima (bem como muitos outros), foi o editor utilizado, neste projecto. Reconhece-se contudo que a sua interface é ligeiramente mais fácil de utilizar por um informático, do que por um utilizador não informático.

Como se pode verificar na Figura 2, o lado superior esquerdo apresenta uma árvore do projecto, e o lado inferior esquerdo apresenta informação sobre determinado elemento extraído do *schema*. Do lado superior direito, é apresentada a lista de todos os elementos definidos no *schema* e na parte inferior direita são apresentados os atributos existentes para cada elemento. Na parte central do editor visualiza-se o próprio registo. Neste exemplo vê-se uma pequena janela sobre a janela central, correspondente à localização actual do cursor, na qual surge uma lista de valores possíveis para um determinado atributo do elemento seleccionado. O operador apenas tem de seleccionar, da lista de opções que lhe vai aparecendo na janela central, o elemento, atributo ou valor que pretende inserir em cada ponto do texto base.

Em conclusão, a ferramenta em causa auxilia efectivamente o operador na sua tarefa de percorrer o documento e ir inserindo as marcas necessárias. Naturalmente que a validação é feita dentro do editor sem dificuldades para o utilizador.

d) Outras soluções

Além do XML-Spy, foram analisados outros dois editores, o XMetal da *SoftQuad Software, Inc.* e o Document Editor da *Altova GmbH & Altova, Inc.*

O XMetal apresenta uma interface bastante amigável, o que o torna talvez ainda mais adequado para utilizadores não técnicos. No entanto, a versão experimentada (versão 2.1) não permite a

"parametrização" com XML Schemas, mas apenas com DTDs, o que impede o seu uso no presente projecto (ver acima decisão tomada sobre a especificação da linguagem Schema-RP).

O Document Editor tem características muito semelhantes às do XMetal, com a vantagem de estar integrado no ambiente do XML-Spy. Além disso, pode ser inserido (por ser um componente ActiveX) numa página Web e pode ser usado de forma distribuída por toda a organização, desde que os computadores possuam instalado o Internet Explorer 5.5 ou superior. Embora seja um produto ainda recente será, concerteza, a escolha correcta no futuro próximo.

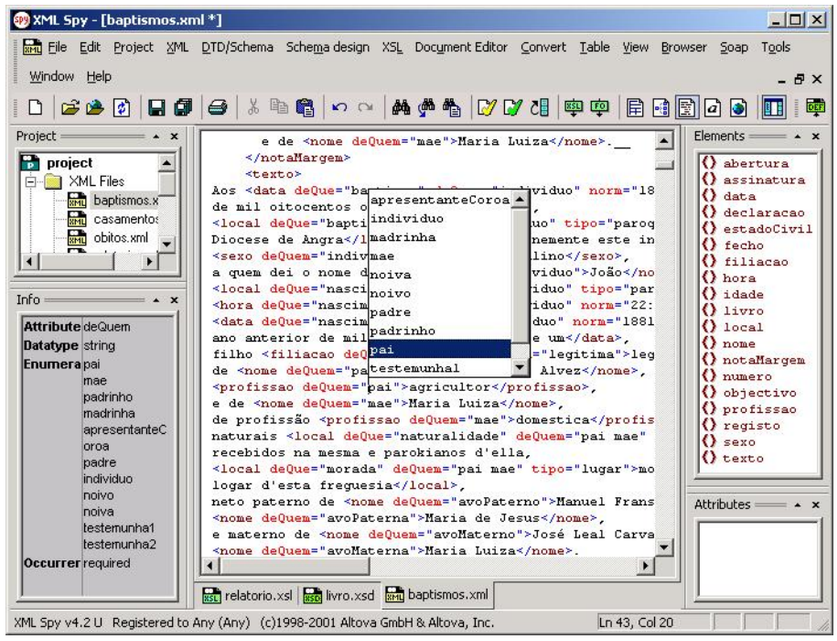

Figura 2 - Aspecto do ambiente de trabalho do XML-Spy

5. Processamento de Registos Paroquiais em XML

Como se disse no início, a maior vantagem em transcrever para o computador todo o registo paroquial procedendo, depois, à sua anotação num linguagem declarativa, baseada na norma XML, é poder dar múltiplas utilizações a esse documento, que quando anotado é independente do uso final. O documento é guardado em ficheiro de texto ASCII puro (texto original mais marcas) tendo depois de ser processado de modo a transformá-lo no formato pretendido, ou a extrair dele a informação requerida. Para realizar essa tarefa existem já várias abordagens bem estabelecidas, estando também disponíveis os programas que realizam a transformação segundo a abordagem escolhida. Essas abordagens, entre as quais se encontra a norma XSL [W3C, 2001], consistem basicamente no uso de uma linguagem que especifica a transformação pretendida para cada elemento do DTD, ou XML Schema, em causa e no recurso a um processador que abre o documento anotado e lhe aplica as regras de transformação prescritas, produzindo o resultado desejado.

Nesta secção, após uma muito breve introdução ao XSL, serão apresentadas duas transformações realizadas com os Registos Paroquias anotados em Schema-RP, a fim de ilustrar as ideias expostas.

a) XSL

XSL, eXtensible Stylesheet Language, é uma norma que permite escrever *especificações* (designadas por *folhas de estilo*) que determinam como *formatar* ou *transformar* documentos XML. Por formatação entende-se qualquer reorganização dos elementos contidos num documento anotado e embelezamento da sua apresentação, realizadas com a finalidade de tornar esses documentos mais compreensíveis e esteticamente agradáveis quando lidos por utilizadores humanos. Por transformação designa-se qualquer operação que analise um documento XML e proceda à sua manipulação estrutural e de conteúdo (sintáctica e semântica) para produzir um novo documento XML, estruturalmente diferente do original, cujo conteúdo informativo seja derivado da informação constante do original. Enquanto que a formatação se destina sempre a apresentar os dados a utilizadores humanos, as transformações têm normalmente por finalidade colocar a informação em formatos adequados ao seu posterior processamento automático através de outros programas.

Para mostrar precisamente estas duas formas possíveis de tratar os registos paroquiais usando XSL, apresentam-se a seguir: um caso de formatação—a conversão dos registos paroquiais de Schema-RP para HTML, de modo a ser possível vê-los através de um *browser da Web*; e um caso de transformação—extracção dos dados relevantes e geração de SQL para carregamento automático das Bases de Dados Paroquiais.

b) Caso 1: Visualização em HTML

Aplicando, através de um processador tipo Saxon, uma *folha de estilo,* especificamente escrita para o efeito em XSL [Robie et al, 1998], aos registos em XML é possível convertê-los para HTML (*HyperText Markup Language*).

Na Figura 3 encontra-se um exemplo da visualização dos Registos Paroquiais, através do *browser* da MicroSoft, Internet Explorer, de acordo com o formato experimental descrito pela folha de estilo (cuja listagem se omite aqui por questões de espaço) construída no contexto do projecto em causa. De acordo com a especificação desenvolvida, todos os nomes são apresentados a negrito, os locais estão sublinhados e as assinaturas estão em itálico.

c) Caso 2: Geração de SQL

Um dos principais objectivos do trabalho aqui apresentado era o carregamento das tabelas de indivíduos e de famílias, da Base de Dados Paroquial, automaticamente a partir da versão electrónica dos registos. Daí que uma das tarefas em foco desde o início fosse o desenvolvimento de um conjunto de regras XSL (uma *folha de estilo*) para especificar o processamento necessário a retirar dos documentos os dados relevantes e transformá-los num conjunto de instruções SQL para os inserir nas respectivas tabelas. Essa tarefa é, à partida, uma pretensão natural e trivial de programar em XSL. Porém, e no contexto específico deste projecto, a situação complicou-se e não foi possível encontrar a solução totalmente pensada, tal como se explicará a seguir.

Devido à complexa interligação entre indivíduos, implícita na noção de família, acrescida pela dificuldade associada à identificação objectiva dos mesmos não é possível gerar numa só passagem as instruções SQL requeridas para alimentar a base de dados. Esta impossibilidade deve-se fundamentalmente às seguintes razões:

-existem códigos internos (as chaves da tabelas de indivíduos e de famílias) que são necessários para a inserção, mas que não vêm descritos nas fontes documentais e que o utilizador não conhece nem quer conhecer;

-não é possível, em alguns casos, identificar inequivocamente um indivíduo sem a ajuda do utilizador (é necessário que este tome a decisão de o considerar como um indivíduo já existente—e nesse caso terá de dizer precisamente qual é—ou como um novo indivíduo).

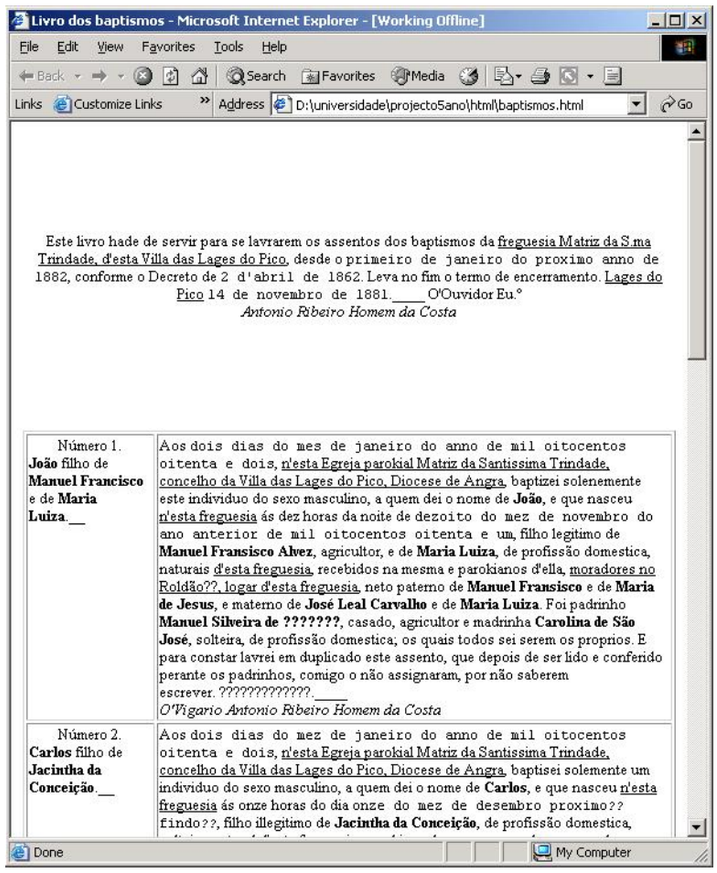

Figura 3 - Exemplo de visualização em HTML de registos XML

Como a linguagem XSL e seus processadores não permitem interactividade com o utilizador nem interrogações à base de dados (a meio do processamento), foi necessário desenvolver uma interface gráfica em Visual Basic para resolver estes dois problemas da interacção que têm de ser ultrapassados para se atingir o fim em vista.

Analisando criteriosamente o processo, verificou-se que a inserção dos dados extraídos dos registos XML terá de ser realizada em duas fases, que são ilustradas na Figura 4.

Na primeira fase, a aplicação em Visual Basic verifica se cada indivíduo reconhecido num documento existe ou não na BDP. Caso exista, apresenta ao utilizador os registos encontrados semelhantes na BDP semelhante ao indivíduo em análise; o utilizador deve então tomar uma de três decisões possíveis: o novo indivíduo é um dos encontrados; o novo indivíduo não se encontra na BDP; o novo indivíduo não se encontra na BDP mas já foi referido anteriormente nesse documento. O primeiro caso implica a actualização do registo existente na BDP, o segundo caso implica a

atribuição de uma nova chave para se proceder à sua inserção na BDP e o terceiro caso implica uma referência à chave anteriormente atribuída, quando esse indivíduo foi encontrado no documento pela primeira vez.

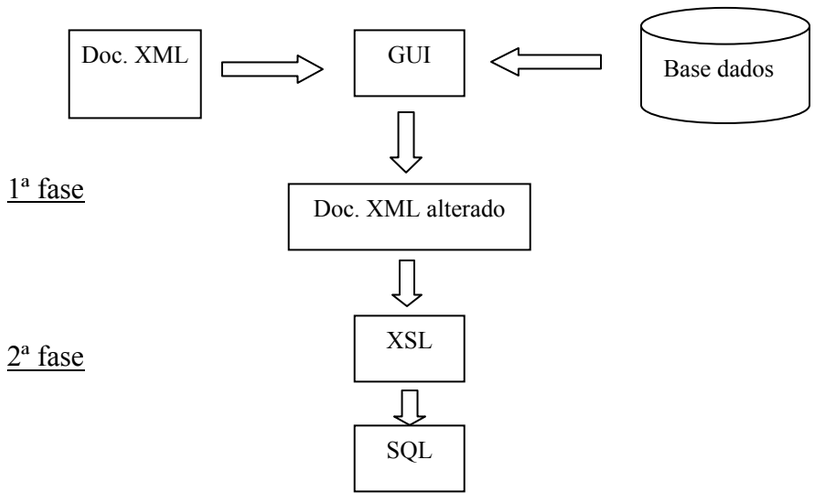

Figura 4 - Fases da geração de SQL para alimentação da base de dados

Além de ser verificada a existência de um indivíduo ou uma família na BDP, nesta fase também é feita a associação entre o valor textual de determinados elementos existentes no documento XML e o respectivo código interno da BDP. Por exemplo, o local "Trindade das Lages" será registado na BDP com o código "102030". O resultado desta fase é um documento XML com toda a informação necessária à segunda fase.

A segunda fase consiste em gerar automaticamente SQL a partir do documento XML obtido na primeira fase. O SQL é gerado aplicando uma *stylesheet* XSL sobre o documento XML, o qual pode ser utilizado para alimentar a BDP.

6. Conclusão

Neste artigo apresentou-se uma aplicação da norma de anotação declarativa de documentos XML ao domínio da história, no sentido de completar um sistema de análise de dados demográficos, SEED, melhorando o patamar de recolha de informação a partir dos Registos Paroquiais, a fonte histórica que suporta o método MRP usado no projecto em causa. Foi discutido com particular ênfase o desenvolvimento de uma instância XML, a linguagem Schema-RP, para a anotação dos registos de baptizados, casamentos e óbitos.

A solução apresentada permite ainda a utilização de diversas *folhas de estilo* XSL, as quais determinam a forma como processar os registos paroquiais. Numa primeira aplicação foi definida uma *folha de estilo* XSL para a conversão dos registos em XML em documentos HTML, permitindo a visualização dos mesmos através do vulgares *browsers da Web*. Posteriormente, uma outra *stylesheet* XSL, auxiliada por uma aplicação em Visual Basic, foi utilizada para alimentar as BDP.

Em termos de trabalho futuro refere-se que está a ser equacionada a utilização de técnicas de *Text Mining*, para auxiliar os Historiadores Demógrafos na exploração da informação recolhida, nomeadamente, na extracção de relacionamentos implícitos existentes nos dados armazenados nos documentos originais em vez, ou em complemento, à exploração que já é feita sobre as bases de dados.

Agradecimentos

Ao concluir este artigo os nossos sinceros agradecimentos vão para a Fátima Rodrigues, mentora da arquitectura SEED e parceira de projecto deste o instante zero, bem como para os muitos alunos finalistas da LMCC e da LESI que através dos projectos de Opção III e dos estágios contribuíram decisivamente para a prototipagem de todas estas ideias. Também agradecemos aos vários colaboradores do NEPS cujo diálogo tem sido fundamental para o avanço do SEED, com especial destaque para a Dr.ª Norberta Amorim e para o Dr. Antero Ferreira.

Referências

Amorim, M. Norberta, "*Rebordãos e a sua População nos séculos XVII e XVIII. Estudo Demográfico"*, Imprensa Nacional-Casa da Moeda, Lisboa, 1973

Amorim, M. Norberta, "*Uma Metodologia de Reconstituição de Paróquias"*, Universidade do Minho, 1991

Bradley, N., "*XSL Companion"*, Addison-Wesley, 1999

Buck, Lee., "*Data models as an XML Schema development method"*, in XML'99 proceedings, Phyladelphia, Dec. 1999

Costa, Américo, Jorge Freitas e Sandra C. Lopes, "*Interface para Aquisição de Dados de um Sistema para Estudo da Evolução Demográfica*", Relatório de Projecto (Opção III), Dep. de Informática, Universidade do Minho, Fev. 2000

Duckett, J., O. Griffin, S. Mohr et al, "*Professional XML Schemas"*, Wrox Press, 2001

Eckstein, Robert, Michel Casabianca, "*XML Pocket Reference"*, O'Reilly, 2.nd Ed., Apr. 2001

Félix, Rafael, "*Sistemas de Digitalização e Anotação de Documentos*", Relatório de Projecto (Opção III), Dep. de Informática, Universidade do Minho, Fev. 2002

Ferreira, João Antero, "*Sistema de Aquisição de Dados para a Reconstituição de Paróquias: a Reprodução Social em S. João das Caldas*", Tese de Mestrado, Universidade do Minho, Set.. 2001

Harold, E.R., W.S. Means, "*XML in a Nutshell"*, O'Reilly, 2001

Herwijnen, E., "*Practical SGML"*, Kluwer Academic Publishers, 1994

Kimber, E., "*Designing a DTD: Element or Attributes"*, acessível em "http://www.open-open.org/cover/attrKimber9711", 1997

Lopes, Sandra Cristina, "*Interface para Aquisição de Dados de um Sistema para Estudo da Evolução Demográfica*", Relatório de Estágio, Dep. de Informática, Universidade do Minho, Dez.. 1999

Oliveira, Luiz Soares, "*Processamento Automático de Dígitos Manuscritos aplicado ao contexto de Cheques Bancários Brasileiros*", Relatório de Qualificação, Programa de Pós-Graduação em Informática Aplicada, Pontifícia Universidade Católica do Paraná, Curitiba, Mai. 2000

Ramalho, José Carlos, "*Anotação Estrutural de Documentos e sua Semântica*", Tese de Doutoramento, Universidade do Minho, Jul. 2000

Robie, J., J. Lapp e D. Sach, "*XSL Transformations (XSLT) – version 1.0"*, acessível em "http://www.w3.org/TR/1998/WD-xsl-19980818", 1998

Rodrigues, Fátima, Carlos Ramos e Pedro Rangel Henriques, "*An Intelligent System to Study Demographic Evolution*", in the American Society SPIE Int. Congress "Data Mining \& Knowledge Discovery: Theory, Tools and Technology", Orlando -- USA, Abr. 1999

Rodrigues, Fátima, "*Arquitectura Heterogénea para Extracção de Conhecimento a partir de Dados"*, Tese de Doutoramento, Universidade do Minho, Dez. 2000

Rodrigues, Fátima, Maribel Yasmina Santos e Pedro Rangel Henriques, "*Realimentação em Sistemas para Extracção de Conhecimento em Bases de Dados*", Actas da II CAPSI - Conferência da Associação Portuguesa de Sistemas de Informação, Edição em CD-ROM, Évora, Portugal, Nov. 2001 (ISBN 972-97869-7-6)

W3C, World Wide Web Consortium, "*Extensible Markup Language (XML)"*, acessível em "http://www.w3.org/XML", Oct. 2001

W3C, World Wide Web Consortium, "*Extensible Stylesheet Language (XSL) – version 1.0"*, acessível em "http://www.w3.org/TR/xsl", Oct. 2001