Gestão e Tratamento de Informação

Laboratório 7 – Exercícios

Nesta aula de laboratório iremos resolver alguns exercícios relacionados com problemas de integração e consolidação de dados, através da tecnologia XQuery.

A linguagem XQuery permite abordar **problemas de integração e consolidação de dados**, tais como os apresentados nos exercícios, através da **criação de vistas**, i.e. funções que acedem aos dados na sua representação original e procedem à sua conversão para um formato de representação comum. Estas vistas podem, por exemplo, ser usadas na construção de ***wrappers*** sobre as fontes de dados originais.

Um processo de integração e consolidação de dados envolve ainda uma outra vista, denominada ***mediador***, que consolida os dados provenientes dos vários *wrappers* e, opcionalmente, verifica a integridade dos dados. A vista mediadora pode usar operadores como *union, intersect* e *except*, os quais retornam os elementos na mesma ordem em que são fornecidos como entrada.

Segundo a especificação do W3C para a linguagem XQuery, os operadores de conjuntos *union*, *except* e *intersect* procedem à **eliminação dos elementos duplicados exactos** que se encontrem nos resultados, mas apenas com base na **identidade dos nós XML** e não com base nos valores associados aos próprios nós ou aos seus descendentes. No entanto, é relativamente simples definir funções utilizador em XQuery que apresentem uma semântica de eliminação de duplicados diferente (i.e., com base nos valores e na estrutura dos elementos). A Figura 1 apresenta o código XQuery para o caso de funções *dup-union* e *dup-intersect* que apresentem esta nova semântica.

```
declare function local:dup-union ($arg1 as node()*, $arg2 as node()*) as node()* {
 let $arg := $arg1 union $arg2
 for $a at $apos in $arg
 let $before_a := fn:subsequence($arg, 1, $apos - 1)
 where every $ba in $before_a satisfies not(deep-equal($ba,$a))
 return $a
};

declare function local:dup-intersect ($arg1 as node()*, $arg2 as node()*) as node()* {
 for $a at $apos in $arg1
 let $before_a := fn:subsequence($arg1, 1, $apos - 1)
 let $test1 := every $ba in $before_a satisfies not(deep-equal($ba,$a))
 let $test2 := some $bb in $arg2 satisfies deep-equal($bb,$a)
 where $test1 and $test2 return $a
};
```

Figura 1 – Funções XQuery correspondendo aos operadores union e intersect.

A Figura 2 representa o fluxo de dados normalmente envolvido num processo de integração e consolidação de dados, em que os wrappers e o mediador são implementados como vistas XQuery.

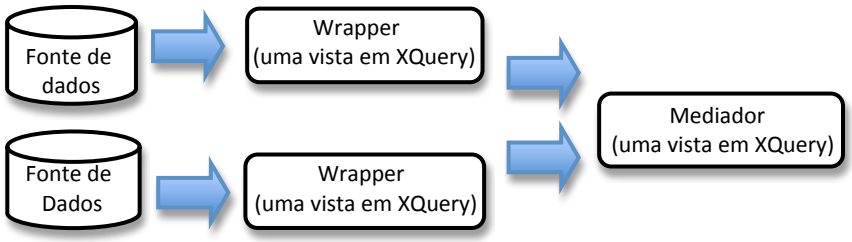

Figura 2 – Fluxo de dados num processo de integração e consolidação.

Exercício 1

No contexto de informação bibliográfica sobre artigos científicos, considere as 3 seguintes fontes de dados disponíveis na Web:

1. http://dbappl.cs.utwente.nl/pftijah/data/sigir.xml
2. http://liinwww.ira.uka.de/bibliography/rss?query=area.hci%20area%20hri
3. http://www.ismir.net/proceedings/index.php?table_name=papers&function=search&where_clause=&page=0&order=Authors&order_type=ASC&export_to_csv=1

A primeira fonte de dados representa, em XML, informação bibliográfica sobre artigos apresentados em edições anteriores da conferência "*ACM SIGIR Conference on Information Retrieval*". A segunda fonte de dados representa em XML, mais concretamente no formato RSS, informação bibliográfica sobre artigos relacionados com "*Human-Computer Interaction*". A terceira fonte de dados representa informação bibliográfica sobre artigos relacionados com "*Music Information Retrieval*," através de um ficheiro de texto com valores separados por vírgulas (formato CSV).

1.1 - Defina um formato de representação comum para as 3 fontes de dados listadas anteriormente. O formato comum deverá representar a informação por forma a preservar ao máximo os diferentes campos definidos em cada uma das fontes de dados originais.

1.2 - Escreva 3 *vistas* em XQuery que actuem como *wrappers* sobre os dados originais. Deverá ser construída uma vista para cada uma das fontes de dados, acedendo aos dados no formato de representação original e convertendo os dados para o formato de representação comum, definido na alínea anterior.

1.3 – Escreva uma *vista* em XQuery que actue como um *mediador*, a qual permita consolidar a informação das 3 vistas desenvolvidas na alínea anterior. Esta vista deverá retornar a união dos resultados das 3 vistas definidas no exercício 1.2. O resultado deverá ser apresentado segundo o formato de representação da alínea 1.1, eliminando-se ainda os duplicados exactos.

1.4 – Classificaria a abordagem desenvolvida nas perguntas anteriores como integração de dados virtual ou materializada? No caso de integração de dados virtual, a abordagem seguida segue a filosofia "*global-as-view*" ou "*local-as-view*"? Justifique.

1.5 - Indique os passos genéricos de uma abordagem alternativa aquela que foi seguida na resolução das perguntas anteriores (e.g., considerar uma integração de dados materializada, caso na pergunta anterior tenha respondido integração de dados virtual), implementada através dos mecanismos que conhece da linguagem XQuery e do Qizx. Apresente quais as vantagens e desvantagens associadas a cada uma das abordagens.

Exercício 2

2.1 – Escreva uma vista em XQuery que actue como um *wrapper* sobre a fonte de dados que se encontra online em http://people.ischool.berkeley.edu/~hearst/irbook/biblio.html (um ficheiro HTML contendo informação bibliográfica sobre artigos relacionados com "*Information Retrieval*"), convertendo os dados para o formato de representação comum desenvolvido no contexto do exercício anterior.

2.2 – Altere a vista XQuery desenvolvida no exercício 1.3 por forma a considerar também a informação resultante do *wrapper* desenvolvido no exercício 2.1. A vista deverá ainda efectuar algumas operações sobre os dados resultantes do processo de integração, nomeadamente:

a) Verificar que todas as referências bibliográficas contêm informação nos campos título, autor e data de publicação.

b) Verificar que a data de publicação de cada referência é superior a 2006.

c) Substituir os acrónimos associados a nomes de associações, presentes nas referencias bibliográficas, pelo nome completo, usando para isso um dicionário (e.g., o acrónimo ACM deve ser substituído por *Association for Computing Machinery*). O dicionário a ser utilizado é apresentado na Figura 3:

```
let $accr = ( 'ACM', 'IEEE', 'ACL' , 'WIC', 'BCS')
let $repl = ( 'Association for Computing Machinery',
              'Institute of Electrical and Electronics Engineers',
              'Association for Computational Linguistics',
              'Web Intelligence Consortium',
              'British Computer Society')
```

Figura 3 – Um dicionário de acrónimos e nomes completos correspondentes.

Exercício 3

Os dois URL que se apresentam abaixo correspondem a documentos XML com informação de perfil sobre empresas Norte-Americanas cotadas em bolsa.

- https://dspace.ist.utl.pt/bitstream/2295/230761/1/profiles1.xml
- https://dspace.ist.utl.pt/bitstream/2295/230762/1/profiles2.xml

3.1 - Defina um **formato de representação comum** para as 2 fontes de dados listadas acima. Escreva ainda dois ***wrappers*** (i.e., duas funções) em XQuery que permitam converter as fontes de dados para o formato de representação comum.

3.2 - Escreva uma *vista* em XQuery que actue como um ***mediador***, permitindo **consolidar a informação dos 2 *wrappers*** desenvolvidos na alínea anterior. O mediador deverá retornar a **união dos resultados** dos 2 wrappers, eliminando os **duplicados** exactos e seguindo o formato de representação comum.