Processamento Estruturado de Documentos

LMCC & LESI, Universidade do Minho

Ano lectivo 2000/2001

Ficha Teórico-Prática Nº1

José Carlos Ramalho

13 de Outubro de 2000

1 Anotação de Documentos

1.1 *Partindo da forma impressa do "Soneto Já Antigo"de Álvaro de Campos, construa uma versão XML do documento.*

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<poema tipo="soneto">
  <titulo>Soneto Já Antigo</titulo>
  <autor>Álvaro de Campos</autor>
  <data>1922</data>
  <corpo>
    <quadra>
      <verso>Olha, <nome>Daisy</nome>: quando eu morrer tu hás-de</verso>
      <verso>dizer aos meus amigos aí de <lugar>Londres</lugar>,</verso>
      <verso>embora não o sintas, que tu escondes</verso>
      <verso>a grande dor da minha morte. Irás de</verso>
    </quadra>
    <quadra>
      <verso><lugar>Londres</lugar> p'ra </lugar>Iorque</lugar>, onde nasceste (dizes ..
      <verso>que eu nada que tu digas acredito),</verso>
      <verso>contar àquele pobre rapazito</verso>
      <verso>que me deu tantas horas felizes,</verso>
    </quadra>
    <terno>
      <verso>embora não o saibas, que morri ...</verso>
      <verso>Mesmo ele, a quem eu tanto julguei amar,</verso>
```

```
      <verso>nada se importará... Depois vai dar</verso>
    </terno>
    <terno>
      <verso>a notícia a essa estranha <nome>Cecily</nome></verso>
      <verso>que acreditava que eu seria grande...</verso>
      <verso>Raios partam a vida e quem lá ande!</verso>
    </terno>
  </corpo>
</poema>
```

2 Transformação usando o módulo Perl XML::DT

2.1 *Utilize a função mkdtskel pertencente ao módulo para gerar a script poema.pl.*

Solução

```
$perl -MXML::DT -e 'mkdtskel "poema.xml"' > poema.pl
$more poema.pl

#!/usr/bin/perl
use XML::DT ;
my $filename = shift;

%handler=(
#    '-outputenc' => 'ISO-8859-1',
#    '-default'   => sub"<$q>$c</$q>",
     'titulo' => sub"$q:$c",
     'poema' => sub"$q:$c",# remember $vtipo
     'corpo' => sub"$q:$c",
     'terno' => sub"$q:$c",
     'data' => sub"$q:$c",
     'autor' => sub"$q:$c",
     'lugar' => sub"$q:$c",
     'quadra' => sub"$q:$c",
     'nome' => sub"$q:$c",
     'verso' => sub"$q:$c",
);
print dt($filename,%handler);
```

2.2 *Acrescente o código necessário para gerar uma versão HTML do poema.*

Solução

```
#!/usr/bin/perl
use XML::DT ;
my $filename = shift;

%handler=(
     '-outputenc' => 'ISO-8859-1',
     '-default'   => sub"<$q>$c</$q>",
     'titulo' => sub"<H2>$c</H2>",
     'poema' => sub"$c",# remember $vtipo
     'corpo' => sub"$c",
     'terno' => sub"<P>$c<P>",
     'data' => sub"<P>($c)<P>",
     'autor' => sub"<I>($c)</I><P>",
     'lugar' => sub"<B>$c</B>",
     'quadra' => sub"<P>$c<P>",
     'nome' => sub"<B>$c</B>",
     'verso' => sub"$c<BR>",
);
print dt($filename,%handler);
```

2.3 *Use a nova script para gerar a versão HTML do poema.*

Solução

```
$chmod 755 poema.pl
$./poema.pl poema.xml > poema.html
$more poema.html
  <H2>Soneto Já Antigo</H2>
  <I>(Álvaro de Campos)</I><P>
  <P>(1922)<P>
    <P>
      Olha, <B>Daisy</B>: quando eu morrer tu hás-de<BR>
      dizer aos meus amigos aí de <B>Londres</B>,<BR>
      embora não o sintas, que tu escondes<BR>
      a grande dor da minha morte. Irás de<BR>
    <P>
    <P>
      <B>Londres</B> p'ra <B>Iorque</B>, onde nasceste (dizes ...<BR>
      que eu nada que tu digas acredito),<BR>
      contar àquele pobre rapazito<BR>
      que me deu tantas horas felizes,<BR>
    <P>
    <P>
      embora não o saibas, que morri ...<BR>
      Mesmo ele, a quem eu tanto julguei amar,<BR>
      nada se importará... Depois vai dar<BR>
    <P>
    <P>
      a notícia a essa estranha <B>Cecily</B><BR>
      que acreditava que eu seria grande...<BR>
      Raios partam a vida e quem lá ande!<BR>
    <P>
```

2.4 *Imagine que o seu poema iria integrar uma base de dados documental e que para se criar os índices necessários era condição obrigatória que todos os elementos no documento tivessem um atributo de nome ident com um valor único dentro do documento. Crie a script de transformação (atribid.pl) para realizar a tarefa com o XML::DT (utilize a função toxml).*

Solução

```
$perl -MXML::DT -e 'mkdtskel "poema.xml"' > atribid.pl
$vi atribid.pl
$more atribid.pl

#!/usr/bin/perl
use XML::DT ;
my $filename = shift;

# Para gerar os vários identificadores vamos usar
# a variável $n como contador de elementos

%handler=(
     '-begin'     => sub$n=1;,
     '-outputenc' => 'ISO-8859-1',
     '-default'   => sub$vident="I$n"; $n++; toxml;,
);
print dt($filename,%handler);
```

2.5 *Use a script gerada para transformar o documento.*

Solução

```
$chmod 755 atribid.pl
$./atribid.pl poema.xml > poema-id.xml
$more poema-id.xml

<poema ident="I28" tipo="soneto">
  <titulo ident="I1">Soneto Já Antigo</titulo>
  <autor ident="I2">Álvaro de Campos</autor>
  <data ident="I3">1922</data>
  <corpo ident="I27">
    <quadra ident="I10">
      <verso ident="I5">Olha, <nome ident="I4">Daisy</nome>: quando eu morrer tu hás-de
      <verso ident="I7">dizer aos meus amigos aí de <lugar ident="I6">Londres</lugar>,<
      <verso ident="I8">embora não o sintas, que tu escondes</verso>
      <verso ident="I9">a grande dor da minha morte. Irás de</verso>
    </quadra>
    <quadra ident="I17">
      <verso ident="I13"><lugar ident="I11">Londres</lugar> p'ra <lugar ident="I12">Ior
      <verso ident="I14">que eu nada que tu digas acredito),</verso>
      <verso ident="I15">contar àquele pobre rapazito</verso>
      <verso ident="I16">que me deu tantas horas felizes,</verso>
    </quadra>
    <terno ident="I21">
      <verso ident="I18">embora não o saibas, que morri ...</verso>
      <verso ident="I19">Mesmo ele, a quem eu tanto julguei amar,</verso>
      <verso ident="I20">nada se importará... Depois vai dar</verso>
    </terno>
    <terno ident="I26">
      <verso ident="I23">a notícia a essa estranha <nome ident="I22">Cecily</nome></ver
      <verso ident="I24">que acreditava que eu seria grande...</verso>
      <verso ident="I25">Raios partam a vida e quem lá ande!</verso>
    </terno>
  </corpo>
</poema>
```

2.6 *Neste momento, você tem que calcular estatísticas sobre a utilização das anotações nos poemas. Crie uma script que ao invés de produzir uma versão transformada do documento produz um novo documento constituído por uma lista de nomes de elementos e respectivo número de ocorrências.*

Solução

```
$perl -MXML::DT -e 'mkdtskel "poema.xml"' > stat-elem.pl
$ ...
$more stat-elem.pl

#!/usr/bin/perl
use XML::DT ;
my $filename = shift;
my %elem = ();

%handler=(
     '-default'   => sub$elem$q++;"",
);
dt($filename,%handler);

print "<TABLE> <TR><TH>Elemento<TH>Nº de ocorrências";
for $i (keys %elem)

  print " <TR><TD>$i <TD>$elem$i";

print "</TABLE>";
```

2.7 *Utilize a script criada para gerar as estatísticas.*

Solução

```
$chmod 755 stat-elem.pl
$./stat-elem.pl poema.xml > stats.html
$more stats.html

<TABLE>
<TR><TH>Elemento<TH>Nº de ocorrências
<TR><TD>titulo <TD>1
<TR><TD>poema <TD>1
<TR><TD>corpo <TD>1
<TR><TD>terno <TD>2
<TR><TD>data <TD>1
<TR><TD>autor <TD>1
<TR><TD>lugar <TD>3
<TR><TD>quadra <TD>2
<TR><TD>nome <TD>2
<TR><TD>verso <TD>14
</TABLE>
```
