Complexity in named entity modelling, and linked open data

TEI @ Oxford

September 2014

Aims

- getting into some minutiae of namesdates module in TEI
- more use of <relation>
- thinking about places
- thinking about linked data

Remember the inner place example:

```
<place xml:id="leipzig">
  <placeName>Leipzig</placeName>
  <place xml:id="tkirch" type="church">
    <placeName xml:lang="de">Thomaskirche</placeName>
    <placeName xml:lang="en">St Thomas' Church</placeName>
    <event type="consecration"
      when="1496-04-10">
      <desc>The current building was consecrated on 10 April
            1496 by the<persName>
        <roleName>Bishop</roleName> of
        <placeName>Merseburg</placeName>
        </persName>.</desc>
    </event>
  </place>
</place>
```

We can also model that as an explicit relation

```
<relation name="partOf" active="#tkirch"
  passive="#leipzig"/>
```

<relation> is not just about relationships between people

Further relations

<relation> can also model RDF:

```
<relation  type="CRM"
  name="P87_is_identified_by"
  active="http://id.clarosnet.org/places/ashmol/placecode/22584"
  passive="http://id.clarosnet.org/places/ashmol/placename/orvieto"/>
```

This indicates that there is a relation, defined by CIDOC CRM, between two resources identified by URLs.

```
<relation  resp="http://viaf.org/viaf/44335536/"
  ref="http://purl.org/saws/ontology#isVariantOf"
  ac-
tive="http://www.ancientwisdoms.ac.uk/cts/urn:cts:greekLit:tlg3017.Syno298.sawsGrc01:divedition.divs
  passive="http://data.perseus.org/citations/urn:cts:greekLit:tlg0031.tlg002.perseus-
grc1:9.35"/>
```

This shows that a passage of text identified by a CTS URN, and a variant passage of text in the Perseus Digital Library, are related, and assigns the identification of the relationship to a particular editor

Time Periods and Relative Chronology

Time periods and relative chronology can be defined under <encodingDesc> and <classDecl>.

```
<taxonomy xml:id="periods">
 <category xml:id="hellenistic">
  <catDesc>
   <ref target="http://www.wikipedia.com/wiki/Hellenistic">
Hellenistic</ref>. Commonly treated as
<date notBefore="-0323" notAfter="-0031"/>. </catDesc>
 </category>
<!--…-->
</taxonomy>
<!--…-->
<p> The city was built near a marble quarry which was
extensively exploited in the <date period="#hellenistic">Hellenistic</date>
and <date period="#roman"> Roman</date> periods.</p>
```

Thinking about other calendars

Trivial example:

```
<profileDesc>
  <calendar xml:id="Stardate">
    <p>Fictional Stardate (from Star Trek series)</p>
  </calendar>
</profileDesc>
<!-- .... -->
<p>Captain's log <date calendar="#Stardate">stardate 23.9 </date>
Enterprise is entering the DiXiT system, and seeing strange alien
forms...</p>
```

Getting more complicated

Normalizing to an alternative calendar:

```
<p>Alhazen died in Cairo on the
<date when="1040-03-06"
   when-custom="431-06-12"> 12th day of Jumada t-Tania, 430 AH
 </date>.</p>
<p>The current world will end at the
<date when="2012-12-21"
   when-custom="13.0.0.0.0">end of B'ak'tun 13</date>.</p>
<p>The Battle of Meggidu
(<date when-custom="Thutmose_III:23">23rd year of reign of Thutmose
III</date>).</p>
<p>Esidorus bixit in pace annos LXX plus minus sub
<date when-custom="Ind:4-10-11">die XI mensis Octobris indictione
IIII</date>
</p>
```

```
<event xml:id="FIRE1"
 datingMethod="#julian" from-custom="1666-09-02"
 to-custom="1666-09-05">
 <head>The Great Fire of London</head>
</event>
```

It can get worse...

```
<p>Contayning the Originall, Antiquity, Increaſe, Moderne
eſtate, and deſcription of that Citie, written in the yeare
<date when-custom="1598"
   calendar="#julian" datingMethod="#julian">1598</date>. by Iohn Stow
Citizen of London.</p>
```

```
<p>Encaenia is usually in <date when-custom="T5" dur-iso="P7D"
   when-iso="2014-06-03" datingMethod="#oxford"
   calendar="#gregorian">Week 5 of
    Trinity term</date>
</p>
```

That <geo> element

```
<place xml:id="craiglockhart">
  <placeName>Craiglockhart War Hospital</placeName>
  <settlement>Edinburgh</settlement>
  <region>Scotland</region>
  <country key="UK">United Kingdom</country>
  <location>
    <geo>55.91812 -3.24019</geo>
  </location>
</place>
```

What do the numbers in <geo> refer to?

‘the assumption is that the content of each geo element will be a pair of numbers separated by whitespace, to be interpreted as latitude followed by longitude according to the World Geodetic System.’

Being explicit about geolocation

```
<encodingDesc>
  <geoDecl xml:id="WGS" datum="WGS84">World Geodetic System</geoDecl>
  <geoDecl xml:id="OS" datum="OSGB36">Ordnance
    Survey</geoDecl>
</encodingDesc>
<!-- .... -->
<location>
  <desc>A tombstone plus six lines of
    Anglo-Saxon text, built into the west tower (on the south side
    of the archway, at 8 ft. above the ground) of the
    Church of St. Mary-le-Wigford in Lincoln.</desc>
  <geo decls="#WGS">53.226658 -0.541254</geo>
  <geo decls="#OS">SK 97481 70947</geo>
</location>
```

Other schemes

- KML: `https://developers.google.com/kml/`
- GML: `http://www.opengeospatial.org/standards/gml`
- GPX: `http://www.topografix.com/gpx.asp`

Remember that places are not points, they are usually better regarded polygons

Remember that places have an historical dimension - they change!

Remember that places have a cultural/political dimension - we don't all agree on what the extent of Ukraine is :-{

An example of KML

```
<place  xml:id="Strat">
  <head>Stratfield Road, Oxford</head>
  <location>
    <geo>
      <Placemark xmlns="http://www.opengis.net/kml/2.2">
        <styleUrl>#lineStyle</styleUrl>
        <MultiGeometry>
          <LineString>
            <coordinates>-1.266010,51.776118,77.200000
                         -1.266015,51.776107,73.900000 -1.266010,51.776097,73.500000
                         -1.266013,51.776102,72.500000 -1.265967,51.776103,72.600000
                         -1.266020,51.776278,70.300000 -1.266052,51.776323,69.400000
                         -1.266083,51.776467,70.200000 -1.266090,51.776507,70.500000
                         -1.266272,51.776848,69.600000 -1.266295,51.776890,70.300000
                         -1.266325,51.776930,69.700000 -1.266352,51.776970,69.000000
                         -1.266533,51.777237,67.000000 -1.266548,51.777260,68.100000
                         -1.266605,51.777410,73.200000 -1.266693,51.777475,71.200000
                         -1.266842,51.777733,65.000000 -1.266865,51.777762,65.200000
                         -1.266942,51.777835,63.800000 -1.266962,51.777867,63.100000
                         -1.267075,51.778078,62.200000 -1.267097,51.778107,62.200000
                         -1.267168,51.778197,62.000000 -1.267180,51.778210,62.300000
                         -1.267203,51.778245,65.30000
            </coordinates></LineString></MultiGeometry></Placemark>
    </geo>
  </location>
</place>
```

An example of GML

```
<place xml:id="LYON3">
 <placeName notBefore="1400">Lyon</placeName>
 <placeName notAfter="-0056">Lugdunum</placeName>
 <location notAfter="-0056">
  <gml:Polygon>
   <gml:exterior>
    <gml:LinearRing> 45.1 -110.23 46.48 -99.08 31.74 -108.86 45.3 -78.2
42.25 -103.45
    </gml:LinearRing></gml:exterior></gml:Polygon>
 </location>
 <location notBefore="1950">
  <gml:Polygon>
   <gml:exterior>
    <gml:LinearRing> 45.256 -110.45 46.46 -109.48 43.84 -109.86 45.8
-109.2 45.256 -110.45
    </gml:LinearRing></gml:exterior></gml:Polygon>
 </location>
</place>
```

Extracting locations to make a map?

Just an XSLT transform away ...

```
<xsl:template match="/">
  <Document>
    <name>TEI Extract</name>
    <xsl:for-each select="//geo">
      <Placemark id="{generate-id()}">
        <name>
          <xsl:value-of select="ancestor::place/placeName"/>
        </name>
        <Point>
          <coordinates>
            <xsl:value-of select="(tokenize(.,' ')[2],tokenize(.,' ')[1])"
              separator=","/>
          </coordinates>
        </Point>
        <description>... something here...
        </description>
      </Placemark>
    </xsl:for-each>
  </Document>
</xsl:template>
```

What if you don't have the location yet?

1. go to the place in person with a GPS-enabled device, make a note of the coordinate
2. go to Google Maps, find your place, do a right-click, select "What's here", and read off the coordinate
3. go to Google Earth, put a placemark on your place(s), and save the result as KML (right click, "Save place as")
4. use one of the many other mapping apps and web sites (eg `http://www.mapmywalk.com/`), many of which which export an XML format you can read from

... but that is not actually the best way to proceed

Better to link your places to a standard gazetteer

- for the modern world, Geonames:
  `http://www.geonames.org/`
- for the classical world, Pleiades:
  `http://pleiades.stoa.org/`
- or the Getty Thesaurus of Geographic Names:
  `http://www.getty.edu/research/tools/vocabularies/tgn/`

How do we reference the gazetteers?

We want to talk about a place called Aphrodisias:

- `http://www.geonames.org/7733053/aphrodisias.html`
- `http://pleiades.stoa.org/places/638753/`
- `http://www.getty.edu/vow/TGNFullDisplay?find=Aphrodisias&place=&nation=&prev_page=1&english=Y&subjectid=7002357`

```
<place xml:id="Aph">
  <placeName>Aphrodisias</placeName>
  <idno type="geonames">7733053</idno>
  <idno type="pleiades">638753</idno>
  <idno type="tgn">7002357</idno>
</place>
```

Not entirely satisfactory, as we need magic to go from (eg) 638753 to `http://pleiades.stoa.org/places/638753/rdf`

Thinking about RDF

- What is the relationship of a TEI-encoded text to RDF?
- What is/are our target ontology or ontologies?
- How do get from TEI XML to RDF?

We want to inject our data into the world of the semantic web in a standardized way, and let other people find or assert links.

Linked Data

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs so that they can discover more things.

1. * make your stuff available on the Web under an open license
2. ** make it available as structured data (e.g., Excel instead of an image scan)
3. *** use non-proprietary formats (e.g., CSV instead of Excel)
4. **** use URIs to denote things, so that people can point at your stuff
5. ***** link your data to other data to provide context

Four possible relationships of TEI to RDF

Discovery	Extract enough bibliographical information to make RDF triples for insertion into a resource finding aid
Extraction	Comb the text, using the TEI markup, looking for assertions about the 'real world' which we can represent in RDF
Mapping	Map everything found in the text into RDF
Container	Decorate our existing TEI markup with extra attributes which map to RDF

Where do we find potential linked data in a TEI file?

1. In <person> or <place> elements, or manuscript descriptions
2. In the bibliographical metadata (authors, titles, dates, subjects
3. In bibliographies
4. In specific <relation> elements
5. Up to a point, in transcribed text

Target ontology

It depends on where we want to get:

- Dublin Core may be good for a general overview
- FRBRoo will make library communities happy
- FOAF may be good for mapping persons and dates
- http://schema.org/Book might suit some TEI work

but it has long been a target of the TEI's Ontology SIG to align the TEI with ISO 21127:2006, the CIDOC conceptual reference model (CRM): see `http://www.cidoc-crm.org/` and `http://www.tei-c.org/SIG/Ontologies/guidelines/guidelinesTeiMappableCrm.xml`

Whatever the chosen ontology, the approach remains the same.

Where to store mapping?

We can look at the TEI <person> element and say it corresponds to the CRM *E21_Person*.

This class comprises real persons who live or are assumed to have lived. Legendary figures that may have existed, such as Ulysses and King Arthur, fall into this class if the documentation refers to them as historical figures. In cases where doubt exists as to whether several persons are in fact identical, multiple instances can be created and linked to indicate their relationship. The CRM does not propose a specific form to support reasoning about possible identity. Examples: - Tut-Ankh-Amun - Nelson Mandela

The <equiv> element allows us to points from a TEI element to an external identifier, and say how to get there.

<equiv>??

```
<elementSpec ident="person" mode="change">
  <equiv filter="crm.xsl"
    mimeType="text/xsl" name="E21"
    uri="http://erlangen-crm.org/110404/E21_Person"/>
</elementSpec>
```

More on <elementSpec> tomorrow...

Input

```
<person xml:id="ArnMag01" sex="1"
  role="scholar">
  <persName xml:lang="is">Árni Magnússon</persName>
  <persName xml:lang="la">Arnas Magnæus</persName>
  <persName xml:lang="da">Arne Magnusson</persName>
  <birth when="1663-11-13">13 November 1663</birth>
  <death when="1730-01-07">7 January 1730</death>
  <residence>
    <date from="1663" to="1680">1663-1680</date>
    <placeName>
      <settlement type="farm">Hvammur</settlement>
      <region type="county">Dalasýsla</region>
      <region type="compass">Western</region>
      <country key="IS">Iceland</country>
    </placeName>
  </residence>
  <residence>
    <date from="1680" to="1683">1680-1683</date>
    <placeName>
      <settlement type="institution">Skálholt</settlement>
      <region type="county">Árnessýsla</region>
      <region type="compass">Southern</region>
      <country key="IS">Iceland</country>
    </placeName>
  </residence>
</person>
```

Result

```
<RDF>
  <E21_Person  rdf:about="http://www.example.com/idArnMag01">
    <P131_is_identified_by  xml:lang="is">
      <E82_Actor_Appellation  rdf:about="http://www.example.com/persname/ArnMag01">
        <value>Árni Magnússon</value>
      </E82_Actor_Appellation>
    </P131_is_identified_by>
    <P98i_was_born>
      <E67_Birth>
        <P4_has_time-span>
          <E52_Time-Span>
            <P82_at_some_time_within>
              <E61_Time_Primitive>
                <value>1663-11-13</value>
              </E61_Time_Primitive>
            </P82_at_some_time_within>
          </E52_Time-Span>
        </P4_has_time-span>
      </E67_Birth>
    </P98i_was_born>
  </E21_Person>
</RDF>
```

Result (continued)

```
<RDF>
  <E21_Person rdf:about="http://www.example.com/person/ArnMag01">
    <P74_has_current_or_former_residence>
      <E53_Place rdf:about="http://www.example.com/place/hvammur">
        <P2_has_type rdf:resource="http://www.tei-c.org/type/place/settlement"/>
        <P87_is_identified_by>
          <E48_Place_Name rdf:about="http://www.example.com/placename/hvammur">
            <value>Hvammur</value>
          </E48_Place_Name>
        </P87_is_identified_by>
        <P89_falls_within rdf:resource="http://www.example.com/place/dalassla"/>
      </E53_Place>
    </P74_has_current_or_former_residence>
  </E21_Person>
  <E53_Place rdf:about="http://www.example.com/place/dalassla">
    <P2_has_type rdf:resource="http://www.tei-c.org/type/place/region"/>
    <P87_is_identified_by>
      <E48_Place_Name rdf:about="http://www.example.com/placename/dalassla">
        <value>Dalasýsla</value>
      </E48_Place_Name>
    </P87_is_identified_by>
    <P89_falls_within rdf:resource="http://tei.it.ox.ac.uk/Talks/2011-10-
teimm/test.xml#IS"/>
  </E53_Place>
</RDF>
```

The problem of identifying things

RDF objects need to be *identified* to be at all useable, preferably with a real, stable, URI.

Among the ways we can generate an identifier:

- if a TEI element has a *@ref* attribute, that is perfect
- if we have an *@xml:id* and meaningful *@xml:base*, we can generate a good URL
- we may be follow a private *@key* if we know the scheme for the document
- we might generate new identifiers based on using <xsl:number> and an *@xml:base*

remembering that really *everything* needs an identifier. When CRM distinguishes Place from Place_Name, they both have to be identified.

The problem of ambiguity of context

Contrast

```
<place>
  <placeName>Bristol</placeName>
</place>
```

with

```
<p>He was born in <placeName>Bristol</placeName>
</p>
```

Problem of ambiguity: <name>

```
<name type="place">Zadar</name>
<name type="person">Oyvind</name>
<rs type="person">Piotr</rs>
```

A pre-processing stage may be order to resolve all such cases to a canonical format

The usual problems

- how to record location in TEI text of source claim
- date of claim
- how to actually express dates
- representing uncertainty and precision
- chronological periods
- how to actually express spatial coordinates