K. Navickas

Week 4: xml and big data

What is metadata?

To quote Jeff Good, a library scientist, metadata is:

Metadata is a new word based on an old concept. Any summary of the contents of a library or archive, like a card catalogue, contains metadata. It is the preferred term of the technical community to refer to ``card-catalogue'' data, and it will, therefore, become increasingly used as more technical tools are developed for linguistic research. ... There is little conceptually new about metadata. The shortest definition for the term is ``data about data,'' and, while many of us may not be accustomed to thinking about metadata very much, we create it and make use of it all the time.

http://www.language-archives.org/documents/gentle-intro.html

A citation like the one below, is a form of metadata:

Bloomfield, Leonard. 1933. *Language*. New York: Holt, Rinehart & Winston.
or
Leonard Bloomfield, *Language* (New York, 1933)

The reference above is information about a book--that is, data about data.

Another way of representing the metadata in the above reference would be as follows:

Document type	Book
Last name of author	Bloomfield
First name of auther	Leonard
Year of publication	1933
Title	Language

City	New York
Publisher	Holt, Rinehart & Winston

What is an XML Schema?

http://www.w3schools.com/schema/schema_intro.asp

The purpose of an XML Schema is to define the legal building blocks of an XML document.

An XML Schema:

- defines elements that can appear in a document
- defines attributes that can appear in a document
- defines which elements are child elements
- defines the order of child elements
- defines the number of child elements
- defines whether an element is empty or can include text
- defines data types for elements and attributes
- defines default and fixed values for elements and attributes

An example:

Just to give you a sense of what this looks like as raw XML, the following is a bit of text drawn from a short trial from 1720, at:

http://www.oldbaileyonline.org/browse.jsp?foo=bar&path=sessionsPapers/17200115.xml&div=t17200115-20

It looks like:

<!-- © 2003-2008 Old Bailey Proceedings Online -->
<TEI.2>
<text>
<body>
<div0 id="17200115" type="sessionsPaper" fragment="yes">
<div1 type="trialAccount" id="t17200115-20">
<xptr type="pageFacsimile" doc="171901150002"/>
<interp inst="t17200115-20" type="collection" value="BAILEY"/>
<interp inst="t17200115-20" type="year" value="1720"/>
<interp inst="t17200115-20" type="uri" value="sessionsPapers/17200115"/>
<interp inst="t17200115-20" type="date" value="17200115"/>
<join result="criminalCharge" id="t17200115-20-off105-c75" targOrder="Y" targets="t17200115-20-defend120 t17200115-20-off105 t17200115-20-verdict109"/>
<p>
<persName id="t17200115-20-defend120" type="defendantName"> Charles Cross , alias
<rs id="t17200115-20-alias-1" type="alias">
<join result="nameAlias" targOrder="Y" targets="t17200115-20-defend120 t17200115-20-alias-1"/>Williams</rs>
<interp inst="t17200115-20-defend120" type="surname" value="Cross"/>
<interp inst="t17200115-20-defend120" type="given" value="Charles"/>
<interp inst="t17200115-20-defend120" type="gender" value="male"/> </persName> , of
<placeName id="t17200115-20-defloc104">St. Andrews Holbourn</placeName>
<join result="persNamePlace" targOrder="Y" targets="t17200115-20-defend120 t17200115-20-defloc104"/> was indicted for
<rs id="t17200115-20-off105" type="offenceDescription">
<interp inst="t17200115-20-off105" type="offenceCategory" value="theft"/>
<interp inst="t17200115-20-off105" type="offenceSubcategory" value="grandLarceny"/> feloniously stealing a Portmanteau Trunk, a Cloth Coat, Wastcoat and Breeches, a pair of Scarlet Breeches laced with Gold Lace, and 7 Holland Shirts, in all to the value of 13 l. </rs> the Goods of
<persName id="t17200115-20-victim122" type="victimName"> Joseph Neal
<interp inst="t17200115-20-victim122" type="surname" value="Neal"/>
<interp inst="t17200115-20-victim122" type="given" value="Joseph"/>
<interp inst="t17200115-20-victim122" type="gender" value="male"/>
<join result="offenceVictim" targOrder="Y" targets="t17200115-20-off105 t17200115-20-victim122"/> </persName> , on the
<rs id="t17200115-20-cd106" type="crimeDate">19th of October</rs>
<join result="offenceCrimeDate" targOrder="Y" targets="t17200115-20-off105 t17200115-20-cd106"/> last. It appeared that
<rs id="t17200115-20-viclabel107" type="occupation">Captain</rs>
<join result="persNameOccupation" targOrder="Y" targets="t17200115-20-victim122 t17200115-20-viclabel107"/> Neal and his Servant came to Town that Night, and that his Man carried the Horses and Portmanteau to
<placeName id="t17200115-20-crimeloc108">Warwicks Stables</placeName>
<join result="offencePlace" targOrder="Y" targets="t17200115-20-off105 t17200115-20-crimeloc108"/>, and while he and two or three more were busy about the Horses in the Stables, the Prisoner, who was Lurking about the Yard, took the Portmanteau out of the Stable and carryed it off. That afterwards the Captain's Servant seeing the Prisoner in the Bear and Harrow Tavern in Butcher Row with his Master's Clothes on his back, apprehended him. The Prisoner in his Defence pleaded

that he bought the Clothes in Newtoners Lane, and that the Prosecutor searched a House in Leather Lane, where he found some of his Goods; to prove the former he called one
<persName id="t17200115-20-person123"> William Beggar
<interp inst="t17200115-20-person123" type="surname" value="Beggar"/>
<interp inst="t17200115-20-person123" type="given" value="William"/>
<interp inst="t17200115-20-person123" type="gender" value="male"/> </persName> , who deposed that he lodged in the same House with the Prisoner, and saw their Landlord lend the Prisoner a Guinea to buy the Clothes, and saw him buy them; but being ask'd what Colour they were, and to describe them, said, they were a light Colour, lined with a light Silk, and trim'd with a Silver; which Description happened to be wrong, for the Clothes being produc'd and swore to by the Prosecutor to be his, and by his Man to be the same that were taken from off the Prisoner, proved not to be very light, but lined with a Blue Silk and no Trimming at all. The Jury found him
<rs id="t17200115-20-verdict109" type="verdictDescription">
<interp inst="t17200115-20-verdict109" type="verdictCategory" value="guilty"/> Guilty </rs>.
<rs id="t17200115-20-punish110" type="punishmentDescription">
<interp inst="t17200115-20-punish110" type="punishmentCategory" value="transport"/>
<join result="defendantPunishment" targOrder="Y" targets="t17200115-20-defend120 t17200115-20-punish110"/> Transportation </rs>.</p> </div1></div0>
</body>
</text>
</TEI.2>

Online tutorial on how to form an xml schema – probably best to start here:

http://www.w3schools.com/schema/schema_simple.asp

The text we are working with today is:

The Sydney Morning Herald (NSW : 1842 - 1954), Friday, 24 January 1890, p. 7. :

http://trove.nla.gov.au/ndp/del/article/13757043

INQUEST.

An inquest was held yesterday morning by Mr. J. C. Woore, City Coroner, at the court, Chancery-square, on the body of James Ward, a labourer by occupation, and, 22 years of age. Deceased was single, and resided at Murray-street, Pyrmont. Thos. Quinn, quarryman, deposed that Ward and himself were working together at a quarry at the Darling Harbour railway, on Wednesday, and were in the act of raising a stone with an iron crowbar, when suddenly a large lump of stone, weighing about 841b.,fell from the top, striking deceased, who bled freely from the ears; he was conscious, and was removed to the hospital, where he died at 7 o'clock. Other evidence was of a corroborative character. Dr. Wade stated that death was due to a fracture of the skull. The jury returned a verdict of accidental death.

Create a table reflecting the 'metadata', and create a list of XML tags. These should take the form of, for instance: <article> text <article/>