An Introduction to XML

Simon Mahony
From an original document by Susan Hockey

This document is part of a collection of presentations and exercises on XML. For full details of this and the rest of the collection see the cover sheet at: http://humbox.ac.uk/3110/

Aims and Outcomes

- Principles and role of structured generic markup
- Create well-formed and valid XML documents
- Write DTDs and Schemas
- Deliver XML documents over Web
- Apply style sheets

- Assess and evaluate role of XML for management and delivery of electronic information

What is XML?

- OED:
  - "Extensible Markup Language, a standard for the mark-up of electronic documents <remove>for display on the Web</remove>, which is based on SGML and allows users to customize their own tags."
- SGML:
  - Standard Generalized Markup Language
  - Describe the document rather than how it should be displayed

XML: Extensible Markup Language?

- Extensible – yes
- Markup – yes
- Language – not really
  - A framework for creating languages
  - Languages used to structure text files and describe their content
  - NOT a programming/scripting language
- Intended to be used by machines, but can be read (and understood) by humans

XML: Extensible Markup Language?

- Meta-language: a language used to describe other languages.
- International standard for the exchange of data

- Markup (encoding): adding a level of interpretation of text.
- Text already has markup (punctuation, spaces, position on the page)
- Encoding makes this explicit

Why is it important?

- Interoperable
  - Machine and software independent
  - ASCII or Unicode
  - Separate the data from the software
- Reusable
  - Not presentation dependant
  - Encode structure/content of the document not its appearance
- It saves you a lot of time and money

Markup?

- Nothing new: as we shall see
  - Proofreaders
  - Typesetters
  - Leiden convention (epigraphic texts)

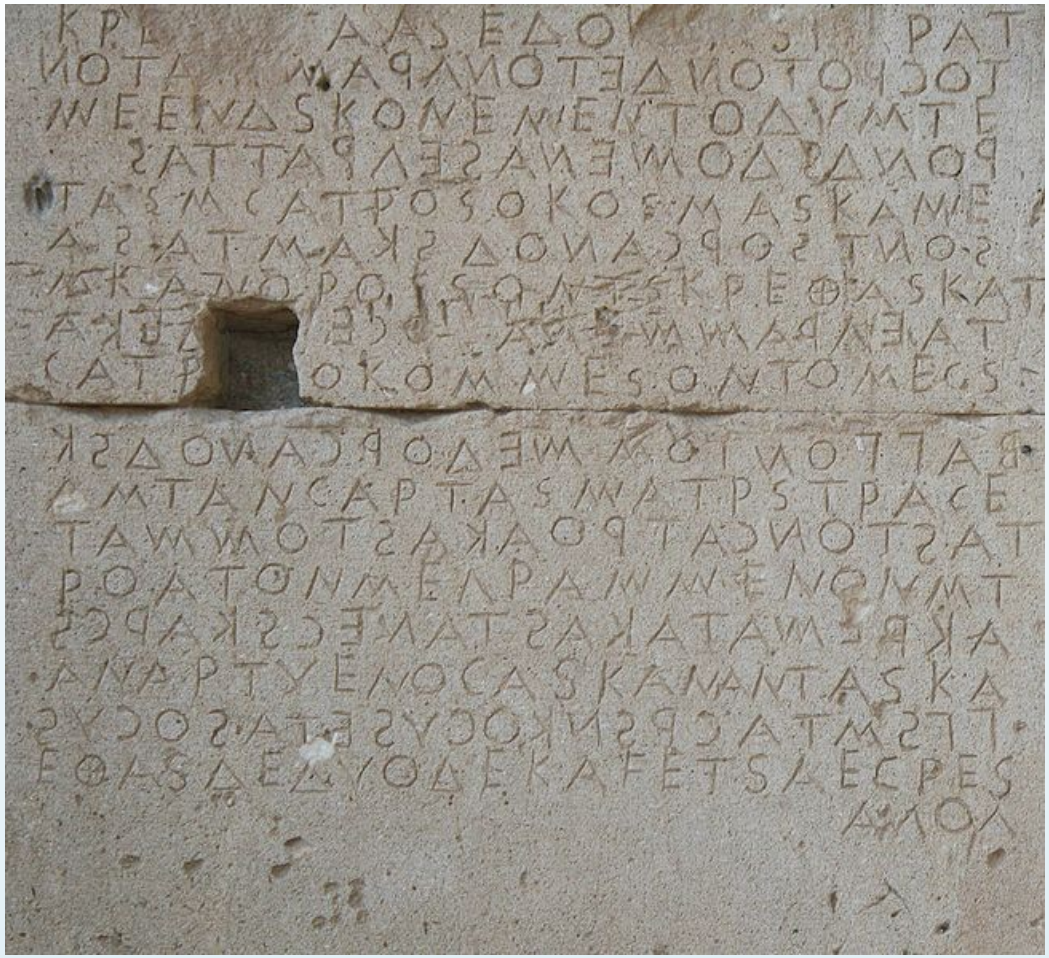
Humanities research is heavily TEXT orientated

- What is a text?
  - A construct created by the reader?
  - It is more than just the words on the page.

- Book culture
- We know the rules
- Punctuation, space have meaning (to us)
- How would we render this electronically?

Clay tablet inscribed with 'Linear B'. Minoan c.1400BC Knossos. British Museum (image: Simon Mahony)

Phaistos Disc (side A) poss 1700BC: Heraklion Archaeological Museum (image from Wikipedia – Wikimedia commons).

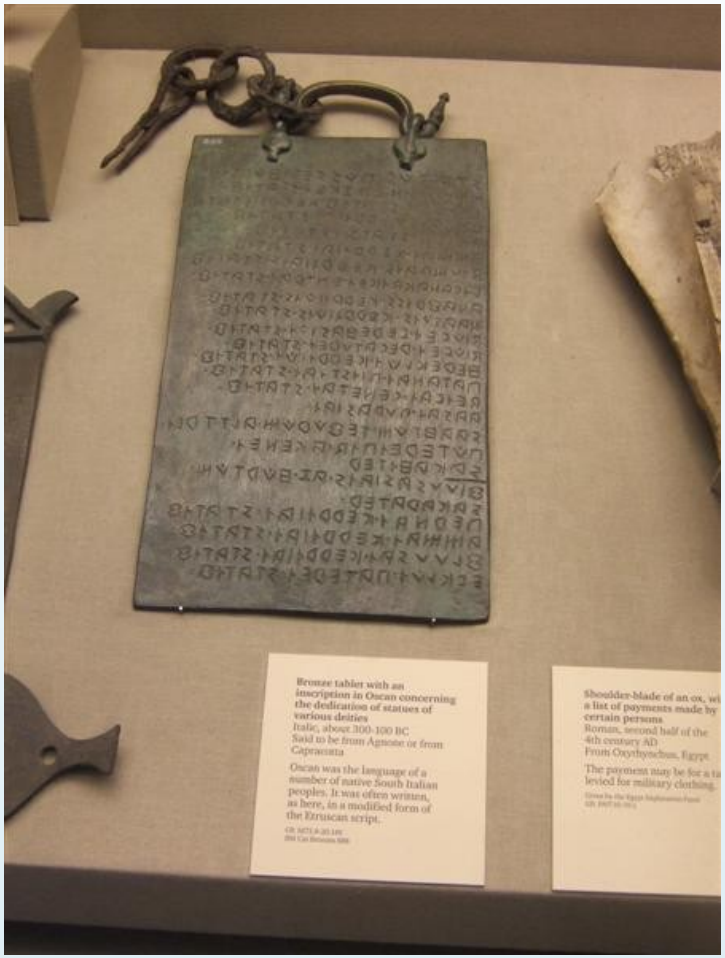


***Boustrophedon*: (like an ox ploughing a field) the direction of each line is reversed. Gortyn law code inscription. Crete 5BC. Image: Wikimedia Commons**

Bronze tablet with an Oscan dedication.

British Museum

(image: Simon Mahony)



Shoulder blade of an ox,

with a list of payments.

British Museum

(image: Simon Mahony)

Roman writing instruments and materials.
British Museum (image: Simon Mahony)

Texts or documents

- Is the object a text or document?
- What is the difference?

- Text: the letters or the ideas therein?

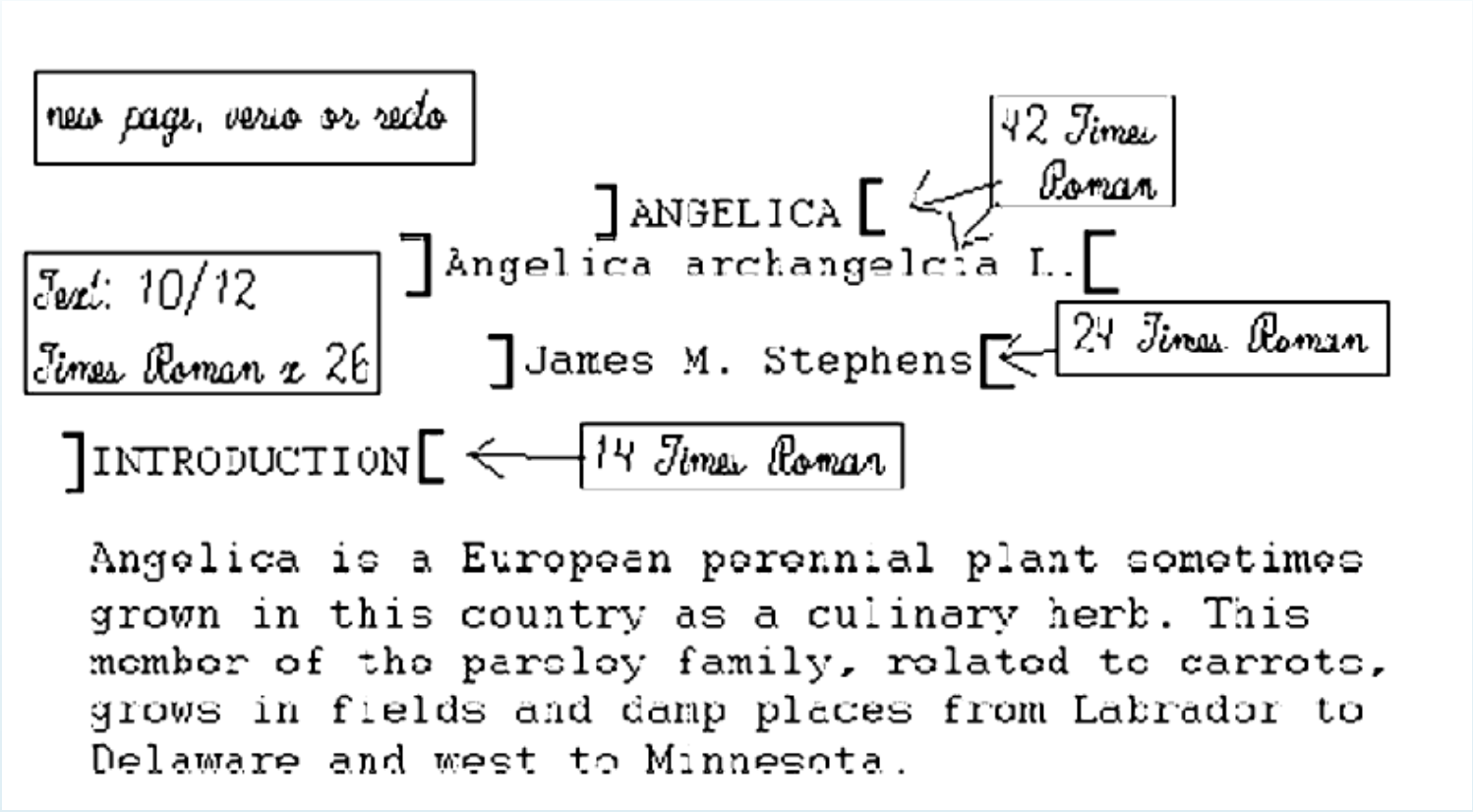
- Markup makes explicit things that we understand implicitly
- Once made explicit they can be processed

Markup

- Adding some additional information
- Disambiguate (cf interactive concordance software)
- Needs to be able to be *read* by computer AND humans

- WYSIWYG
  - word processor (hidden formatting)
  - Endnote

Typesetters markup



HTML vs XHTML

- HTML – displays your data
- XHTML – describes your data (HTML + XML)
  - Subset of XML family

- XHTML
  - separates style from content
  - structural and semantic markup
  - stricter syntax (limited elements)

- CSS – used to style XHTML pages

Example of the difference

I *really* liked the characterisation of Ajax in Homer's *Iliad.*

- HTML does not allow us to distinguish between the different uses of the italics
- With XHTML we can mark these up differently to differentiate between emphasis and the book title.
- Using XML we can also add more information if we wish

HTML:

```
<p>
I <i><ul>really</ul></i> liked the characterisation of Ajax
in Homer's <i><ul>Iliad</ul></li>.
</p>
```

XHTML:

```
I <em>really</em> liked the characterisation of Ajax in
Homer's<span class="title">Iliad</span>.
```

Generic Markup

- Used by early text formatting programs
- Markup identifies the content, not the format
  - Heading: not 14point, bold Times
- Waterloo Script – old formatting program
- LaTex – used for mathematics and science materials

SGML – Standard Generalized Markup Language

- International standard in 1986
- ISO 8879:1986
- Not a markup scheme in itself
- A syntax for defining markup schemes
- Assumes (mostly) that a document is a nested or hierarchic structure

- A descendant of IBM's Generalized Markup Language(GML)

SGML

- Separation of content and design
- Same document can be used for many different purposes
- Archival form of the material (simple text file)
- Separate the document from the processing
- Content-based markup
- Functionality is in the processing software

Development of XML

- Simplification of SGML
- Developed by a small group led by Jon Bosak of Sun Microsystems
- Became a World Wide Web Consortium recommendation in February 1998
- Now many associated activities in W3C and elsewhere

Not just the Web

- Allows transformation to multiple outputs
- Print publication
- Printable view on Web
- Create indices
- Table of contents
- Checking pages
- PDF
- Text

Where is XML used?

- Word processors (eg MS Word 2007)
- Google Maps
- ATM
- Banks exchanging data
- Petrol station
- Anywhere data needs to be transmitted

What is XML?

- A simple syntax for defining a markup scheme
- Elements
- Attributes
- Values
- Entities

- Document structures

Document Structures

- XML documents are tree structures
- Composed of nested structures of elements
- Some elements may also have attributes

(Image: Simon Mahony)

Made up of:

- Elements
- Attributes
  - Values

- Entities

Document Analysis

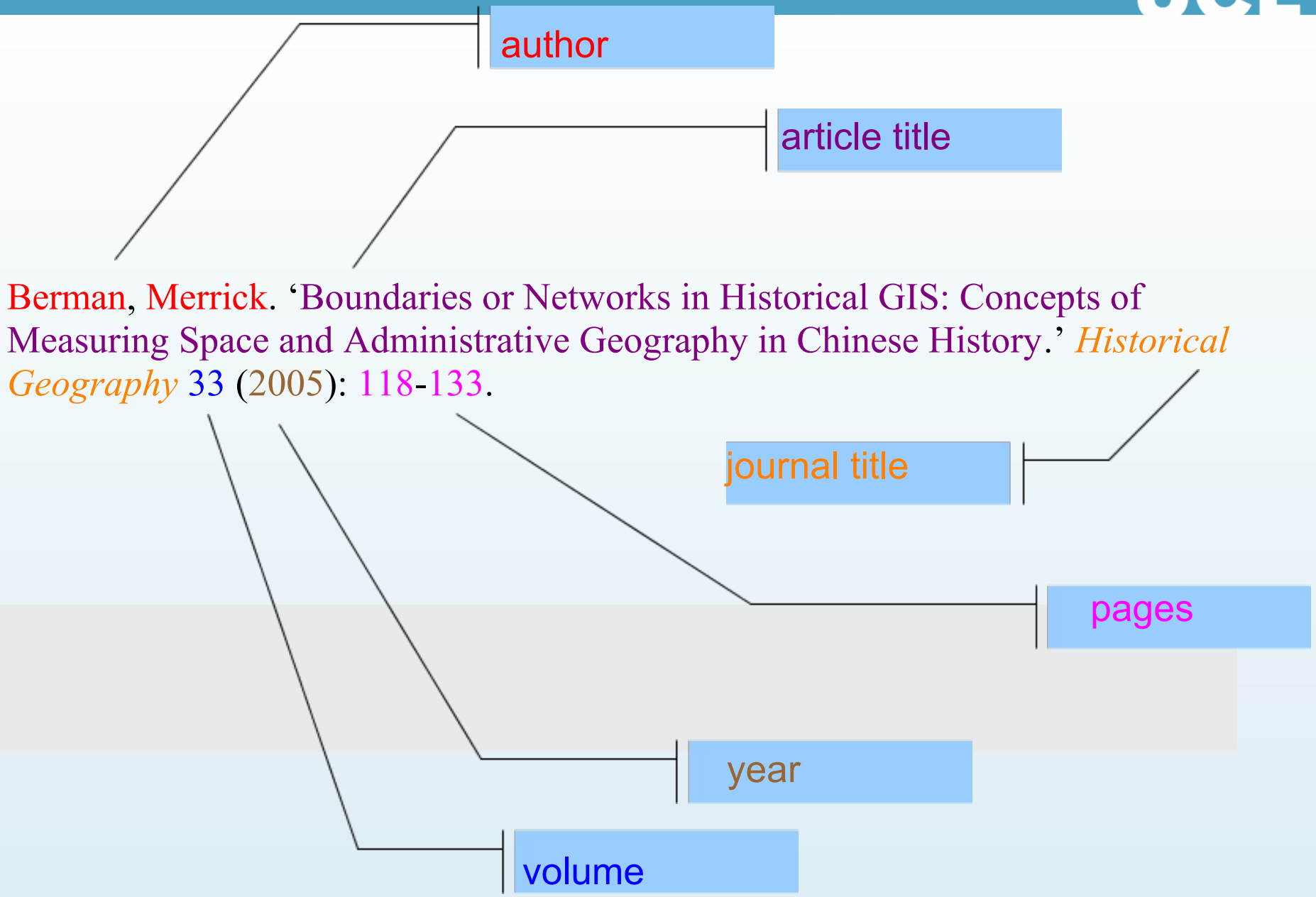
- First stage of an XML (and SGML) project
- Determine what are the important features within the document(s)
  - This will depend to some extent on the nature of the document
  - What is it you (or others) are interested in?
- Determine the relationships between the features
- Produce a tree structure with names for the elements

E.g. Bibliographic entry

Berman, Merrick. 'Boundaries or Networks in Historical GIS: Concepts of Measuring Space and Administrative Geography in Chinese History.' *Historical Geography* 33 (2005): 118-133.

This example is adapted from an original by Tom Elliott (NYU) and acknowledged with thanks.

(https://docs.google.com/present/view?id=drn6nzs_30d9vm77dt)


author
article title
Berman, Merrick. 'Boundaries or Networks in Historical GIS: Concepts of Measuring Space and Administrative Geography in Chinese History.' Historical Geography 33 (2005): 118-133.
journal title
pages
year
volume

Start marking up with XML

<bibl>

Berman, Merrick. 'Boundaries or Networks in Historical GIS: Concepts of Measuring Space and Administrative Geography in Chinese History.' Historical Geography 33 (2005): 118-133.

</bibl>

Adding element: <author></author>

<bibl>

<author>Berman, Merrick</author>. 'Boundaries or Networks in Historical GIS: Concepts of Measuring Space and Administrative Geography in Chinese History.'

Historical Geography 33 (2005): 118-133.

</bibl>

Adding <title></title> but more than one title

<bibl>

<author>Berman, Merrick</author>. '<title>Boundaries or Networks in Historical GIS: Concepts of Measuring Space and Administrative Geography in Chinese History</title>.' Historical Geography 33 (2005): 118-133.

</bibl>

<bibl>
<author>Berman, Merrick</author>. '<title level="a">Boundaries or Networks in Historical GIS: Concepts of Measuring Space and Administrative Geography in Chinese History</title>.' Historical Geography 33 (2005): 118-133.
</bibl>

title levels:

a = *analytic* title (article, poem, or other item published as part of a larger item)

j = *journal* title

m = *monographic* title (book, collection, or other item published as a distinct item,
including single volumes of multi-volume works)

s = *series* title

u = title of *unpublished* material (including theses and dissertations unless published by a commercial press)

XML: element > attribute > value

<bibl>
<author>Berman, Merrick</author>. '<title level="a">Boundaries or Networks in Historical GIS: Concepts of Measuring Space and Administrative Geography in Chinese History</title>.' <title level="j">Historical Geography</title> 33 (2005): 118-133.
</bibl>

Element to define volume number

```
<bibl>
<author>Berman, Merrick</author>. '<title level="a">Boundaries or Networks in Historical GIS: Concepts of Measuring Space and Administrative Geography in Chinese History</title>.' <title level="j">Historical Geography</title>
<biblScope type="vol">33</biblScope> (2005): 118-133
</bibl>
```

Date element

```
<bibl>
<author>Berman, Merrick</author>. '<title level="a">Boundaries or Networks in Historical GIS: Concepts of Measuring Space and Administrative Geography in Chinese History</title>.' <title level="j">Historical Geography</title>
<biblScope type="vol">33</biblScope> (<date>2005</date>): 118-133
</bibl>
```

Page numbers

```
<bibl>
<author>Berman, Merrick</author>. '<title level="a">Boundaries or Networks in Historical GIS: Concepts of Measuring Space and Administrative Geography in Chinese History</title>.' <title level="j">Historical Geography</title>
<biblScope type="vol">33</biblScope> (<date>2005</date>): <biblScope type="pp">118-133</biblScope>.
</bibl>
```

Punctuation?

```
<bibl>
<author>Berman, Merrick</author>. '<title level="a">Boundaries or Networks in Historical GIS: Concepts of Measuring Space and Administrative Geography in Chinese History</title>.' <title level="j">Historical Geography</title>
<biblScope type="vol">33</biblScope> (<date>2005</date>): <biblScope type="pp">118-133</biblScope>.
</bibl>
```

```
<bibl>
<author>Berman, Merrick</author>
<title level="a">Boundaries or Networks in Historical GIS: Concepts of Measuring Space and Administrative Geography in Chinese History</title>
<title level="j">Historical Geography</title>
<biblScope type="vol">33</biblScope>
<date>2005</date>
<biblScope type="pp">118-133</biblScope>
</bibl>
```

Author?

```
<bibl>
<author>Berman, Merrick</author>
<title level="a">Boundaries or Networks in Historical GIS: Concepts of Measuring Space and Administrative Geography in Chinese History</title>
<title level="j">Historical Geography</title>
<biblScope type="vol">33</biblScope>
<date>2005</date>
<biblScope type="pp">118-133</biblScope>
</bibl>
```

Deconstruct name into: surname / forename

```
<bibl>
<author>
<surname>Berman</surname>
<forename>Merrick</forename>
</author>
<title level="a">Boundaries or Networks in Historical GIS: Concepts of Measuring Space and Administrative Geography in Chinese History</title>
<title level="j">Historical Geography</title>
<biblScope type="vol">33</biblScope>
<date>2005</date>
<biblScope type="pp">118-133</biblScope>
</bibl>
```

<listBibl> **(Add unique ID and wrapper)**

….

<bibl xml:id="berman2005">
<author>
<surname>Berman</surname>
<forename>Merrick</forename>
</author>
<title level="a">Boundaries or Networks in Historical GIS: Concepts of Measuring Space and Administrative Geography in Chinese History</title>
<title level="j">Historical Geography</title>
<biblScope type="vol">33</biblScope>
<date>2005</date>
<biblScope type="pp">118-133</biblScope>
</bibl>

…

</listBibl>

Anything else?

- Editorial decisions
- How much detail is required?
- How much detail can you afford?
  - Time = money and funding is limited

XML

- Texts are already encoded (book culture)
- For markup, texts need to be de-coded (by us)
- Then
  - Re-encoded in an unambiguous way
  - Read by both computers and humans

Transformation (via XSLT)

- For format (HTML, PDF etc)
- For editions (critical, diplomatic, etc)
- For collations (indices, TOCs, etc)
- Checking pages

Successful standard

- Data standard for many formats
- Underlying data: MS Word 2007 (.zip file)
- Platform independent
  - Plain text with .xml file extension
- Store information
  - Future-resistant
- Importantly: widely supported scholarly community (TEI)
  - Fosters interchange and collaboration
  - Open Source

Document analysis

- Study documents
- Construct an abstract model
- Define objectives
- Produce an encoded representation

To recap

What is XML?

- A simple syntax for defining a markup scheme
- Elements
- Attributes
  - Values
- Entities
- Document structures

Document Structures

- XML documents are tree structures
- Composed of nested structures of elements
- Some elements may also have attributes

Image source: Simon Mahony

Element / Attribute (value) / Entity

- Element: <title> <author> <date>
  - Syntax <title> … </title>
- Attribute: modify the elements
  - Syntax attribute-name="attribute-value"
  - <element attribute-name="attribute-value">some text</element>
- Entities: Non ASCII characters
  - Special characters in XHTML
  - eg & é etc
  - Text to be expanded (eg &UCL;)

Elements and Document Structures

- Elements can be repeated
- Elements can be optional
- Elements can contain other elements
- Elements can contain only text (the leaves of the tree)
- Elements can have mixed content – text and/or other elements

Elements

- Normally, elements have some content
- Start and end tags

<title> Pride and Prejudice</title>

- End tags MUST be present in XML
- rest of the file is PCDATA
  - ie Parsed Character data = untagged text
- File is a simple text file

Empty Elements

- Elements without any content

- <image filename="image.jpg" />
- <br/ >

- Mark a position in a document, rather than surrounding some text (cf. XHTML)
  - e.g. a page break

Attributes

- Further modify elements
- Attributes are always in quotes
  <name type="place">London</name>
  <name type="personal">Simon</name>
- Elements take more than one attribute type
  - eg <name language="english">
  - <name ID="26">

- This could also be expressed as
  <name><person>Simon</person></name>

Must have a nested Structure

- An XML document is a tree structure of nested elements
- Elements can repeat
- The tree can be any depth
- The document must have an outer (root) element

Nested structure

```
<body>

    <p>
<strong><em>Some text</em></strong>
```

```
<strong><em> Some text</strong></em>
    </p>

</body>
```

Nested structure: example from bibliography

```
<bibl>
    <author>
    <surname>Berman</surname>
    <forename>Merrick</forename>
    </author>
    <title level="a">Boundaries or Networks in Historical GIS: Concepts of Measuring Space and Administrative Geography in Chinese History</title>
    <title level="j">Historical Geography</title>
    <biblScope type="vol">33</biblScope>
    <date>2005</date>
    <biblScope type="pp">118-133</biblScope>
</bibl>
```

Elements and Document Structures

- Elements can be repeated
- Elements can be optional
- Elements can contain other elements
- Elements can contain only text (the leaves of the tree)
- Elements can have mixed content – text and/or other elements