

Introduction to TEI and XML Encoding:

What are we doing and
why are we doing it?

Why do we encode texts?

- Preservation
- Publication
- Analysis
- Synthesis
- ???

Good and bad encoding systems

生

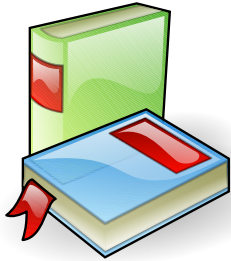
Meanings (pronunciations):

- *life* (sei, shō)
- *live* (i[kiru])
- *be born* (u[mu], u[mareru])
- *yield* (u[su])
- *grow* (ha[eru], ha[yasu])
- *bear fruit* (na[ru])
- *raw, uncooked* (nama)
- ...and more...

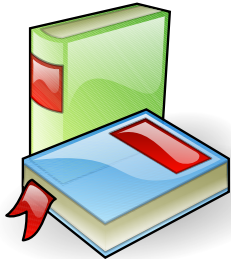
Good and bad encoding systems



buku



buku-buku



buku²

Computer encoding

Well-known quote:

```
010101000110111100100000011000100110010100
101100001000000110111101110010001000000110
1110011011110111010000100000011101000110111
1001000000110001001100101001011100001000000
111010001101000011000010111010000100000011
010010111001100100000011101000110100001100
101001000000111000101110101011001010111001
101110100011010010110111101101110
```

ASCII encoding

- 01010100 = 84 = T
- 01101111 = 111 = o
- 00100000 = 32 = [space]
- 01100010 = 98 = b
- 01100101 = 101 = e

Encoding through punctuation and style

Bradley, J. "Thinking about interpretation: Pliny and scholarship in the humanities." *Literary and Linguistic Computing* 23:3 (2008), 263–279.

- We know stuff that is not explicit; we can decode this encoding.
- Find encoded pieces of extratextual information.
- What other meanings do these signals have?

Problems with style-based encoding

- ambiguity: difficult for novices or machines to interpret.
- limitations: how many features can you encode?
- conflict: actual style vs. style-as-encoding. (What does a handwriting font look like in italics?)

Problems of encoding style itself

```
{\rtf1\ansi\deff3\adeflang1025
{\fonttbl{\f0\froman\prq2\charset0 Times New Roman;}{\f1\froman\prq2\charset2 Symbol;}{\f2\fswiss\prq2\charset0 Arial;}{\f3\froman\prq2\charset128 Times
New Roman;}{\f4\fswiss\prq2\charset128 Arial;}{\f5\fnil\prq2\charset128 Droid Sans;}{\f6\fnil\prq2\charset128 Lohit Hindi;}{\f7\fnil\prq0\charset128 Lohit
Hindi;}}
{\colortbl;\red0\green0\blue0;\red128\green128\blue128;}
{\stylesheet{\s0\snext0\nowidctlpar{\f1\hyphen2\hyphlead2\hyphtrail2\hyphmax0}\cf0\kerning1\hich\af5\langfe2052\dbch\af6\afs24\alang1081\loch\af3\fs24\lang1033
Normal;}}
{\s15\sbasedon0\snext16\sb240\sa120\keepn\hich\af5\dbch\af6\afs28\loch\af4\fs28 Heading;}
{\s16\sbasedon0\snext16\sb0\sa120 Text body;}
{\s17\sbasedon16\snext17\sb0\sa120\dbch\af7 List;}
{\s18\sbasedon0\snext18\sb120\sa120\noline\ldbch\af7\afs24\ai\fs24 Caption;}
{\s19\sbasedon0\snext19\noline\ldbch\af7 Index;}
}{\info{\author mholmes }{\creatim\yr2012\mo10\dy2\hr13\min58}{\revtim\yr0\mo0\dy0\hr0\min0}{\printim\yr0\mo0\dy0\hr0\min0}{\comment LibreOffice}
{\vern3500}}\def\ab709

{\*\pgdscctl
{\pgdsc0\pgdscuse195\pgwsxn12240\pghsxn15840\marglsxn1134\margrsxn1134\margtsxn1134\margbsxn1134\pgdscnxt0 Default;}}
\
formshade\paperh15840\paperw12240\margl1134\margr1134\margt1134\margb1134\sectd\sbknone\sectunlocked1\pgndec\pgwsxn12240\pghsxn15840\marglsxn1
134\margrsxn1134\margtsxn1134\margbsxn1134\ftnbj\ftnstart1\ftnrstcont\ftnna\enddoc\aftnrstcont\aftnstart1\aftnrlc
\pgndec\pard\plain
\so\nowidctlpar{\f1\hyphen2\hyphlead2\hyphtrail2\hyphmax0}\cf0\kerning1\hich\af5\langfe2052\dbch\af6\afs24\alang1081\loch\af3\fs24\lang1033{\rtlch \ltrch\loch
The Latin word }{\ai\rtlch \ltrch\loch
mundus}{\i0\ai0\rtlch \ltrch\loch
means \uc3 \u8220\’e2\’80\’9cworld\u8221\’e2\’80\’9d.\uc1 }
\par }
```

- The Latin word *mundus* means “world”.

What makes a good encoding system?

- intelligible
- precise
- nice (fine distinctions)
- concise
- comprehensive
- unambiguous
- systematic
- open, standardized, documented, maintained...

And so...

<s>

The

<lang>Latin</lang>

word

<term xml:id="mundus" xml:lang="lat">

mundus

</term>

means

<gloss target="#mundus">

world

</gloss>.

</s>

Encoding texts: where TEI fits

