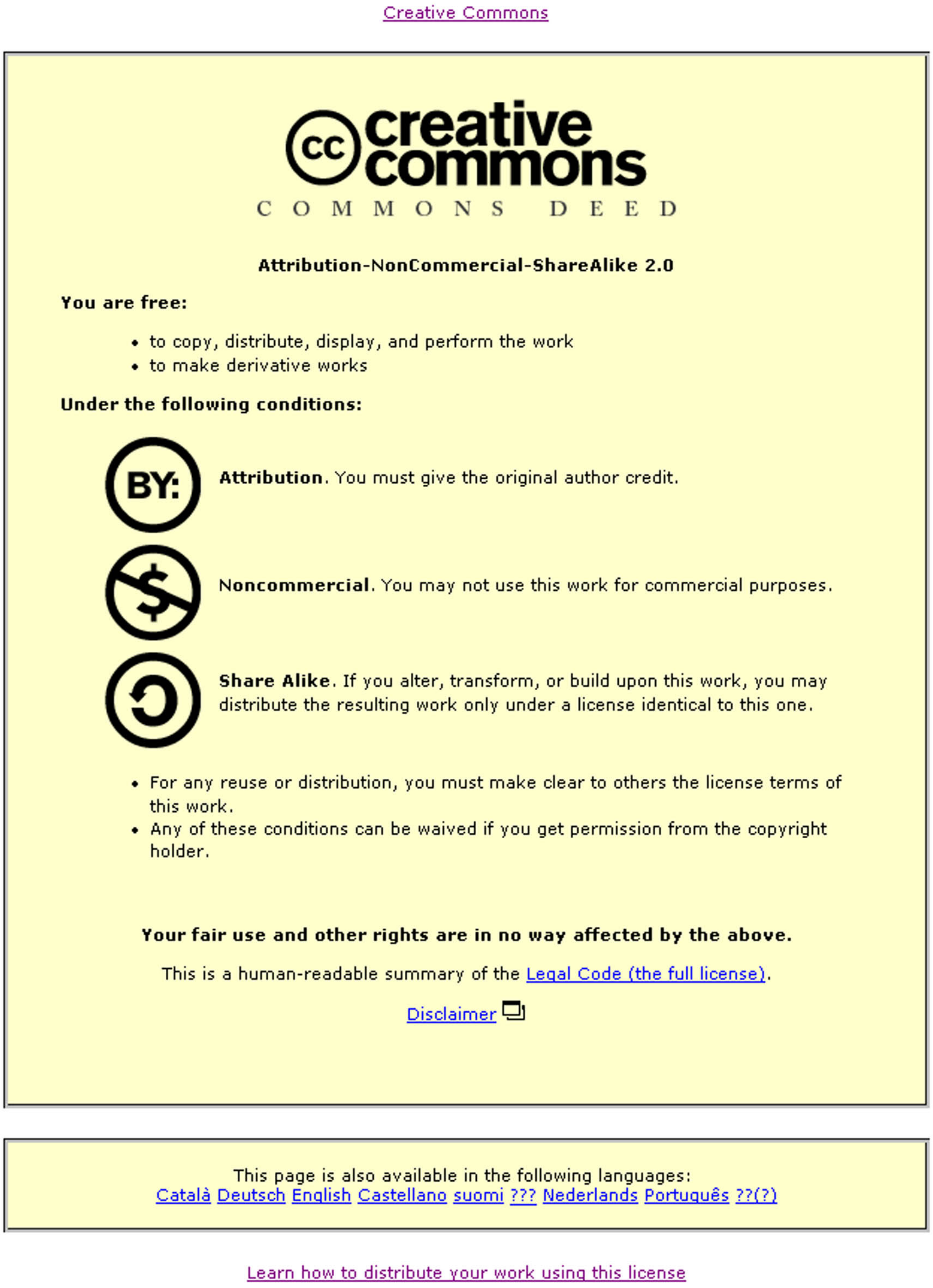
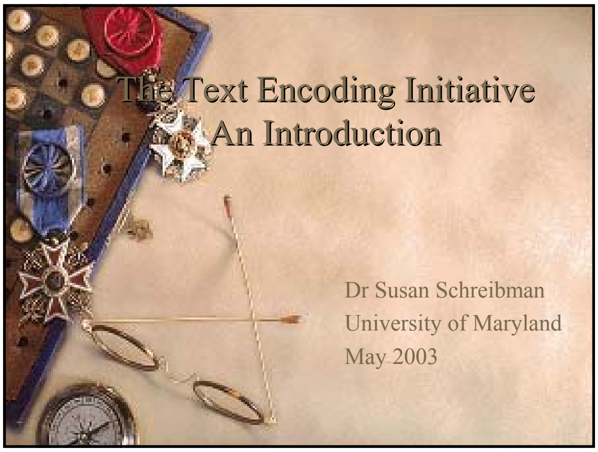
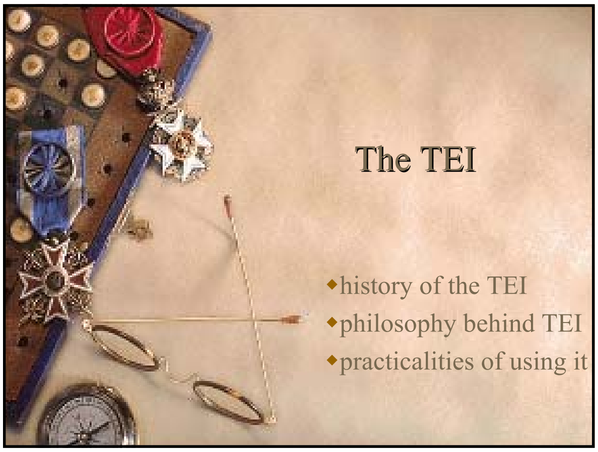




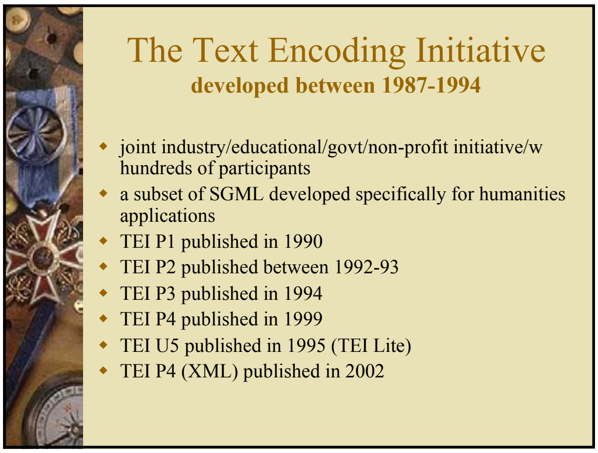
The Text Encoding Initiative
An Introduction

Dr Susan Schreibman
University of Maryland
May 2003



The TEI

- history of the TEI
- philosophy behind TEI
- practicalities of using it



The Text Encoding Initiative

developed between 1987-1994

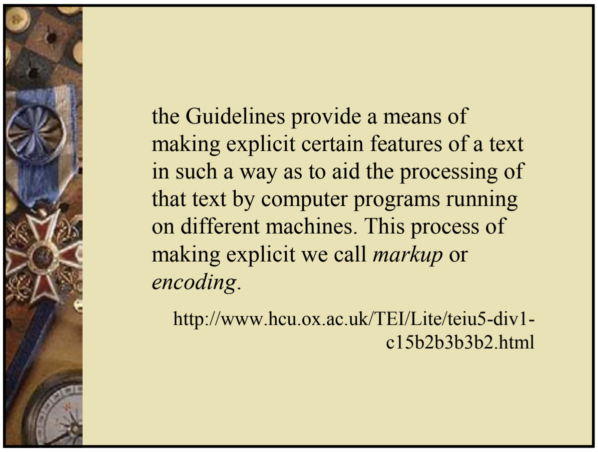
- joint industry/educational/govt/non-profit initiative/w hundreds of participants
- a subset of SGML developed specifically for humanities applications
- TEI P1 published in 1990
- TEI P2 published between 1992-93
- TEI P3 published in 1994
- TEI P4 published in 1999
- TEI U5 published in 1995 (TEI Lite)
- TEI P4 (XML) published in 2002

An International Consortium: old website http://www.uic.edu/orgs/tei/

- funders
  - the Natl Endowment of the Humanities
  - Commission of the EU
  - Andrew W Mellon Foundation
  - Social Science & Humanities Research Council of Canada
- sponsors
  - Assoc of Computers and the Humanities
  - Assoc for Computational Linguistics
  - Assoc for Literary and Linguistic Computing

New Consortium to maintain the TEI formed in 1999

- hosted by:
  - University of Virginia
  - Brown University
  - Oxford University
  - University of Bergen
- more info at:
  - http://www.tei-c.org/



the Guidelines provide a means of making explicit certain features of a text in such a way as to aid the processing of that text by computer programs running on different machines. This process of making explicit we call *markup* or *encoding*.

http://www.hcu.ox.ac.uk/TEI/Lite/teiu5-div1-c15b2b3b3b2.html

The guidelines create a standard and guide for "the preparation and interchange of electronic texts for scholarly research, and to satisfy a broad range of uses by the language industries more generally"

Intended Use

- guidance for individual or local practice in text creation and data capture
- support for data interchange
- support of application-independent processing

Goals of the TEI (1987):

- suffice to represent the textual features needed for research;
- be simple, clear, and concrete;
- be easy for researchers to use without special-purpose software;
- allow the rigorous definition and efficient processing of texts;
- provide for user-defined extensions;
- conform to existing and emergent standards.

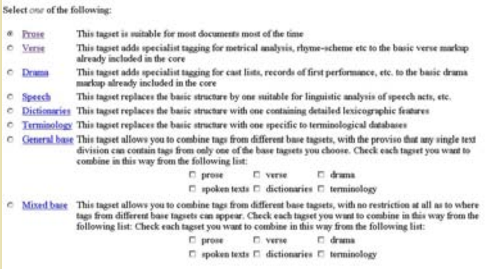
http://www.hcu.ox.ac.uk/TEI/Lite/teiu5-div1-c15b2b3b3b2.html

to facilitate the widest possible usership, it was important to ensure:

- the common core of textual features be easily shared;
- additional specialist features be easy to add to (or remove from) a text;
- multiple parallel encodings of the same feature should be possible;
- the richness of markup should be user-defined, with a very small minimal requirement;
- adequate documentation of the text and its encoding should be provided.

http://www.hcu.ox.ac.uk/TEI/Lite/teiu5-div1-c15b2b3b3b2.html

- the common core of textual features be easily shared;



Select *one* of the following:

● Prose This tagset is suitable for most documents most of the time

○ Verse This tagset adds specialist tagging for metrical analysis, rhyme-scheme etc to the basic verse markup already included in the core

○ Drama This tagset adds specialist tagging for cast lists, records of first performance, etc. to the basic drama markup already included in the core

○ Speech This tagset replaces the basic structure by one suitable for linguistic analysis of speech acts, etc.

○ Dictionaries This tagset replaces the basic structure with one containing detailed lexicographic features

○ Terminology This tagset replaces the basic structure with one specific to terminological databases

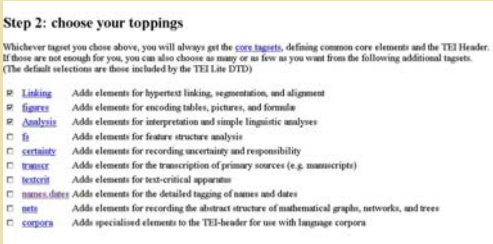
○ General base This tagset allows you to combine tags from different base tagsets, with the proviso that any single text division can contain tags from only one of the base tagsets you choose. Check each tagset you want to combine in this way from the following list:

□ prose □ verse □ drama
□ spoken texts □ dictionaries □ terminology

○ Mixed base This tagset allows you to combine tags from different base tagsets, with no restriction at all as to where tags from different base tagsets can appear. Check each tagset you want to combine in this way from the following list: Check each tagset you want to combine in this way from the following list:

□ prose □ verse □ drama
□ spoken texts □ dictionaries □ terminology

- additional specialist features be easy to add to (or remove from) a text;



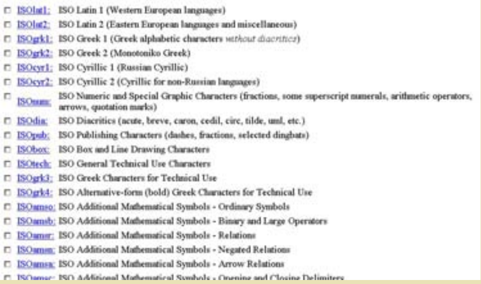
Step 2: choose your toppings

Whichever tagset you chose above, you will always get the core tagsets, defining common core elements and the TEI Header. If those are not enough for you, you can also choose as many or as few as you want from the following additional tagsets. (The default selections are those included by the TEI Lite DTD)

☑ Linking Adds elements for hypertext linking, segmentation, and alignment
☑ figures Adds elements for encoding tables, pictures, and formulæ
☑ Analysis Adds elements for interpretation and simple linguistic analyses
□ fs Adds elements for feature structure analysis
□ certainty Adds elements for recording uncertainty and responsibility
□ transcr Adds elements for the transcription of primary sources (e.g. manuscripts)
□ textcrit Adds elements for text-critical apparatus
□ names.dates Adds elements for the detailed tagging of names and dates
□ nets Adds elements for recording the abstract structure of mathematical graphs, networks, and trees
□ corpora Adds specialised elements to the TEI-header for use with language corpora

Entity Sets

•additional specialist features be easy to add to (or remove from) a text;



□ ISOlat1: ISO Latin 1 (Western European languages)
□ ISOlat2: ISO Latin 2 (Eastern European languages and miscellaneous)
□ ISOgrk1: ISO Greek 1 (Greek alphabetic characters *without diacritics*)
□ ISOgrk2: ISO Greek 2 (Monotoniko Greek)
□ ISOcyr1: ISO Cyrillic 1 (Russian Cyrillic)
□ ISOcyr2: ISO Cyrillic 2 (Cyrillic for non-Russian languages)
□ ISOnum: ISO Numeric and Special Graphic Characters (fractions, some superscript numerals, arithmetic operators, arrows, quotation marks)
□ ISOdia: ISO Diacritics (acute, breve, caron, cedil, circ, tilde, uml, etc.)
□ ISOpub: ISO Publishing Characters (dashes, fractions, selected dingbats)
□ ISObox: ISO Box and Line Drawing Characters
□ ISOtech: ISO General Technical Use Characters
□ ISOgrk3: ISO Greek Characters for Technical Use
□ ISOgrk4: ISO Alternative-form (bold) Greek Characters for Technical Use
□ ISOamso: ISO Additional Mathematical Symbols - Ordinary Symbols
□ ISOamsb: ISO Additional Mathematical Symbols - Binary and Large Operators
□ ISOamsr: ISO Additional Mathematical Symbols - Relations
□ ISOamsn: ISO Additional Mathematical Symbols - Negated Relations
□ ISOamsa: ISO Additional Mathematical Symbols - Arrow Relations
□ ISOamsc: ISO Additional Mathematical Symbols - Opening and Closing Delimiters

◆multiple parallel encodings of the same feature should be possible;

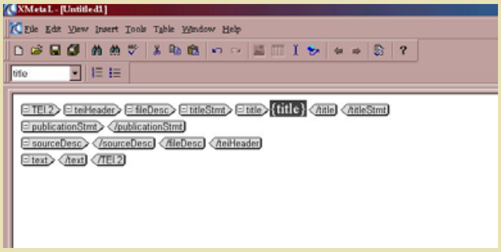
<P>However, few have been so scathing in their condemnation of <PERSNAME

REG="WB Yeats">Yeats</PERSNAME> as <PERSNAME><CORR

REND="Brian Coffey" CERT="100%">Brian Cofey</CORR></PERSNAME>, a friend and contemporary of <PERSNAME

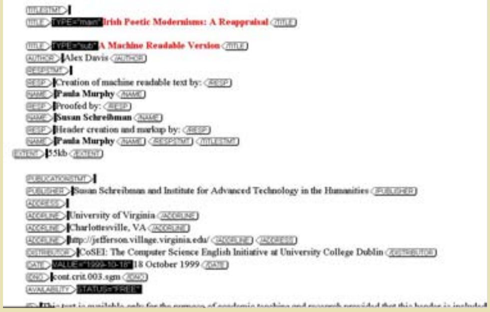
REG="Samuel Beckett">Beckett</PERSNAME>, who described Yeats as 'A power-hungry seducer who gathered a right gang of praisers around him, and who blocked off the kind of talent he didn't like'.

◆the richness of markup should be user-defined, with a very small minimal requirement;



```
TEI.2 teiHeader fileDesc titleStmt title {title} /title /titleStmt
publicationStmt /publicationStmt
sourceDesc /sourceDesc /fileDesc /teiHeader
text /text /TEI.2
```

◆ adequate documentation of the text and its encoding should be provided

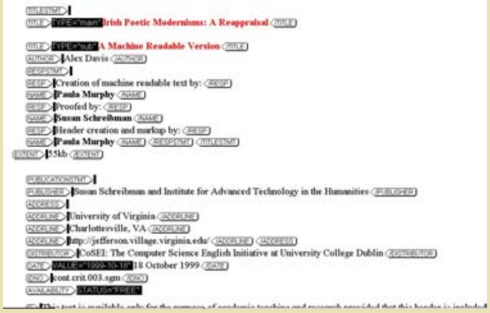


```
TITLESTMT
TITLE TYPE="main" Irish Poetic Modernisms: A Reappraisal /TITLE

TITLE TYPE="sub" A Machine Readable Version /TITLE
AUTHOR Alex Davis /AUTHOR
RESPSTMT
RESP Creation of machine readable text by: /RESP
NAME Paula Murphy /NAME
RESP Proofed by: /RESP
NAME Susan Schreibman /NAME
RESP Header creation and markup by: /RESP
NAME Paula Murphy /NAME /RESPSTMT /TITLESTMT
EXTENT 55kb /EXTENT

PUBLICATIONSTMT
PUBLISHER Susan Schreibman and Institute for Advanced Technology in the Humanities /PUBLISHER
ADDRESS
ADDRLINE University of Virginia /ADDRLINE
ADDRLINE Charlottesville, VA /ADDRLINE
ADDRLINE http://jefferson.village.virginia.edu/ /ADDRLINE /ADDRESS
DISTRIBUTOR CoSEI: The Computer Science English Initiative at University College Dublin /DISTRIBUTOR
DATE VALUE="1999-10-18" 18 October 1999 /DATE
IDNO cont.crit.003.sgm /IDNO
AVAILABILITY STATUS="FREE"
This text is available only for the purpose of academic teaching and research provided that this header is included
```

◆ adequate documentation of the text and its encoding should be provided



TITLESTMT
TITLE TYPE="main" Irish Poetic Modernisms: A Reappraisal /TITLE

TITLE TYPE="sub" A Machine Readable Version /TITLE
AUTHOR Alex Davis /AUTHOR
RESPSTMT
RESP Creation of machine readable text by: /RESP
NAME Paula Murphy /NAME
RESP Proofed by: /RESP
NAME Susan Schreibman /NAME
RESP Header creation and markup by: /RESP
NAME Paula Murphy /NAME /RESPSTMT /TITLESTMT
EXTENT 55kb /EXTENT

PUBLICATIONSTMT
PUBLISHER Susan Schreibman and Institute for Advanced Technology in the Humanities /PUBLISHER
ADDRESS
ADDRLINE University of Virginia /ADDRLINE
ADDRLINE Charlottesville, VA /ADDRLINE
ADDRLINE http://jefferson.village.virginia.edu/ /ADDRLINE /ADDRESS
DISTRIBUTOR CoSEI: The Computer Science English Initiative at University College Dublin /DISTRIBUTOR
DATE VALUE="1999-10-18" 18 October 1999 /DATE
IDNO cont.crit.003.sgm /IDNO
AVAILABILITY STATUS="FREE"
This text is available only for the purpose of academic teaching and research provided that this header is included

a DTD flexible enough to encompass a wide range of texts, periods & purposes

- history
- literature
- art history
- linguistics
- ancient texts
- medieval texts
- modern texts

has become the de facto standard

- most major humanities computing projects utilise it
- in theory, allows for the exchange of texts across projects and archives
- helps to ensure uniform encoding of text: extremely important for both humans and parsers

What does TEI facilitate?

- **structural divisions within a text**
  - **title-page, chapter, scene, stanza, line, etc**
- **typographical elements**
  - **changes in typeface, special characters, etc**
- **other textual features**
  - **grammatical structure, location of illustrations, variant forms, etc**

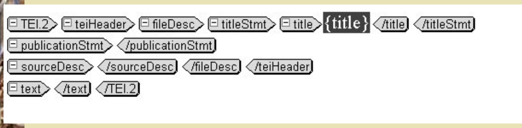
facilitates the long-term preservation of electronic texts

- repositories have one basic DTD to preserve and understand
  - though that DTD may have many "flavors"

all TEI documents follow the same essential format

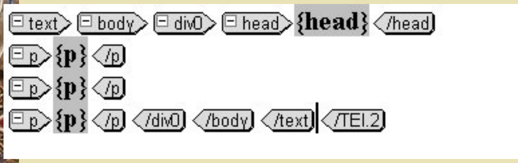
- TEI *header*
  - documents the electronic edition being created
- TEI *body*
  - contains the content being created

TEI Header



```
TEI.2 teiHeader fileDesc titleStmt title {title} /title /titleStmt
publicationStmt /publicationStmt
sourceDesc /sourceDesc /fileDesc /teiHeader
text /text /TEI.2
```

TEI Body



```
text body div0 head {head} /head
p {p} /p
p {p} /p
p {p} /p /div0 /body /text /TEI.2
```

TEI flavors: The TEI Pizza Site

The Full TEI Tag Set

- developed to meet a range of encoding needs
  - across disciplines
  - across language needs
    - literature vs linguistics
  - developed in a pre-internet environment

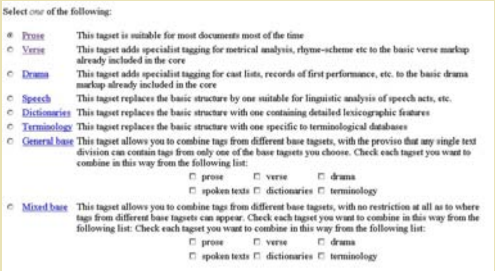
Concept behind Pizza

“The TEI Guidelines define several hundred SGML elements and associated attributes, which can be combined to make many different DTDs, suitable for many different purposes, either simple or complex. With the aid of the Pizza Chef, you can build a DTD that contains just the elements you want, suitable for use with any SGML or XML processing system.”

Pizza concept

- there are various base tag sets (including)
  - prose
  - verse
  - drama
  - dictionaries
  - speech
  - mixed base

- the common core of textual features be easily shared;



Select *one* of the following:

● Prose This tagset is suitable for most documents most of the time

○ Verse This tagset adds specialist tagging for metrical analysis, rhyme-scheme etc to the basic verse markup already included in the core

○ Drama This tagset adds specialist tagging for cast lists, records of first performance, etc. to the basic drama markup already included in the core

○ Speech This tagset replaces the basic structure by one suitable for linguistic analysis of speech acts, etc.

○ Dictionaries This tagset replaces the basic structure with one containing detailed lexicographic features

○ Terminology This tagset replaces the basic structure with one specific to terminological databases

○ General base This tagset allows you to combine tags from different base tagsets, with the proviso that any single text division can contain tags from only one of the base tagsets you choose. Check each tagset you want to combine in this way from the following list:

□ prose □ verse □ drama
□ spoken texts □ dictionaries □ terminology

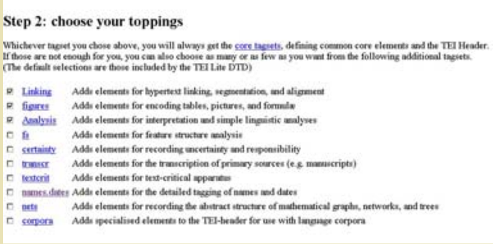
○ Mixed base This tagset allows you to combine tags from different base tagsets, with no restriction at all as to where tags from different base tagsets can appear. Check each tagset you want to combine in this way from the following list: Check each tagset you want to combine in this way from the following list:

□ prose □ verse □ drama
□ spoken texts □ dictionaries □ terminology

Pizza concept

- individuals choose various "toppings", ie additional tag sets based on encoding needs (including)
  - linking
  - figures
  - analysis (linguistic analysis)
  - transcription
  - textcrit
  - names.dates
  - certainty

- additional specialist features be easy to add to (or remove from) a text;



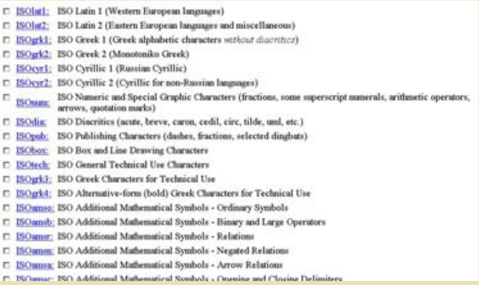
Step 2: choose your toppings

Whichever tagset you chose above, you will always get the core tagsets, defining common core elements and the TEI Header. If those are not enough for you, you can also choose as many or as few as you want from the following additional tagsets. (The default selections are those included by the TEI Lite DTD)

☑	Linking	Adds elements for hypertext linking, segmentation, and aligment
☑	figures	Adds elements for encoding tables, pictures, and formulæ
☑	Analysis	Adds elements for interpretation and simple linguistic analyses
□	fs	Adds elements for feature structure analysis
□	certainty	Adds elements for recording uncertainty and responsibility
□	transcr	Adds elements for the transcription of primary sources (e.g. manuscripts)
□	textcrit	Adds elements for text-critical apparatus
□	names.dates	Adds elements for the detailed tagging of names and dates
□	nets	Adds elements for recording the abstract structure of mathematical graphs, networks, and trees
□	corpora	Adds specialised elements to the TEI-header for use with language corpora

Entity Sets

•additional specialist features be easy to add to (or remove from) a text;



□	ISOlat1:	ISO Latin 1 (Western European languages)
□	ISOlat2:	ISO Latin 2 (Eastern European languages and miscellaneous)
□	ISOgrk1:	ISO Greek 1 (Greek alphabetic characters *without diacritics*)
□	ISOgrk2:	ISO Greek 2 (Monotoniko Greek)
□	ISOcyr1:	ISO Cyrillic 1 (Russian Cyrillic)
□	ISOcyr2:	ISO Cyrillic 2 (Cyrillic for non-Russian languages)
□	ISOnum:	ISO Numeric and Special Graphic Characters (fractions, some superscript numerals, arithmetic operators, arrows, quotation marks)
□	ISOdia:	ISO Diacritics (acute, breve, caron, cedil, circ, tilde, uml, etc.)
□	ISOpub:	ISO Publishing Characters (dashes, fractions, selected dingbats)
□	ISObox:	ISO Box and Line Drawing Characters
□	ISOtech:	ISO General Technical Use Characters
□	ISOgrk3:	ISO Greek Characters for Technical Use
□	ISOgrk4:	ISO Alternative-form (bold) Greek Characters for Technical Use
□	ISOamso:	ISO Additional Mathematical Symbols - Ordinary Symbols
□	ISOamsb:	ISO Additional Mathematical Symbols - Binary and Large Operators
□	ISOamsr:	ISO Additional Mathematical Symbols - Relations
□	ISOamsn:	ISO Additional Mathematical Symbols - Negated Relations
□	ISOamsa:	ISO Additional Mathematical Symbols - Arrow Relations
□	ISOamsc:	ISO Additional Mathematical Symbols - Opening and Closing Delimiters

TEI Lite

- a subset of of TEI designed to meet average encoding needs
- no choosing of bases and toppings
- no choosing of entity sets – all is chosen for you
- "conceived of as a simple demonstration of how the TEI encoding scheme might be adopted to meet 90% of the needs of 90% of the TEI user community."

http://www.hcu.ox.ac.uk/TEI/Lite/teiu5-div-c15b2b3b1b3.html

TEI Lite is most suited to

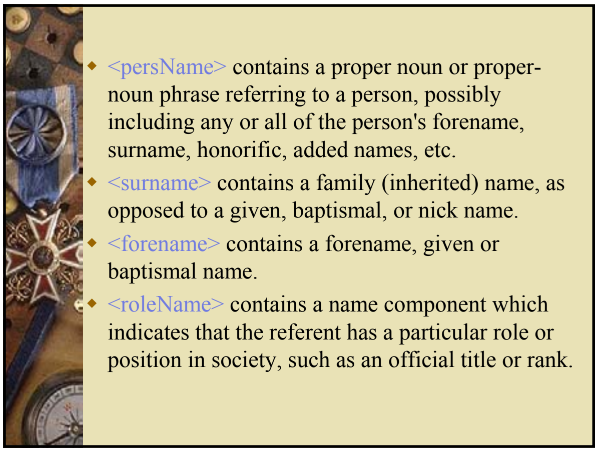
- printed texts
- structural encoding
- light content encoding
- electronic repositories which want to make lightly encoded texts available to scholars

Goals of TEI Lite

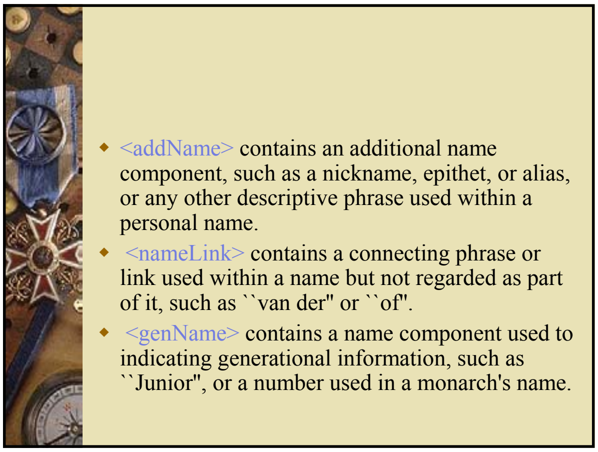
- it should include most of the TEI "core" tag set, since this contains elements relevant to virtually all text types and all kinds of text-processing work;
- it should be able to handle adequately a reasonably wide variety of texts, at the level of detail found in existing practice (for ex the Oxford Text Archive);
- it should be useful for the production of new documents as well as encoding of existing ones;

how does it fulfill its goals?

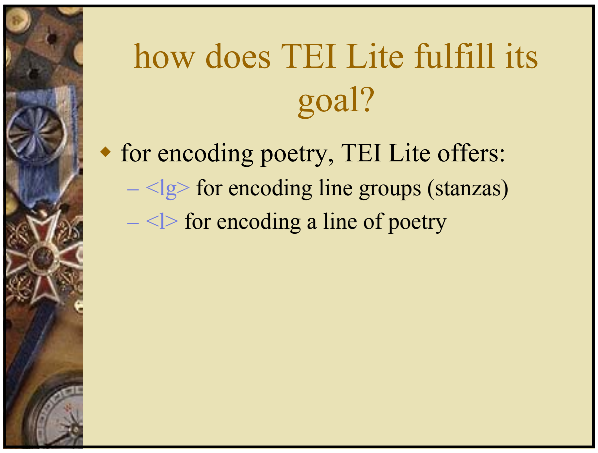
- it includes most features of the base prose tag set;
- but skips more specialised usages, for ex:
- TEI Lite includes <name>
- if one chooses the "names and dates" additional tag set one gets:



- <persName> contains a proper noun or proper-noun phrase referring to a person, possibly including any or all of the person's forename, surname, honorific, added names, etc.
- <surname> contains a family (inherited) name, as opposed to a given, baptismal, or nick name.
- <forename> contains a forename, given or baptismal name.
- <roleName> contains a name component which indicates that the referent has a particular role or position in society, such as an official title or rank.

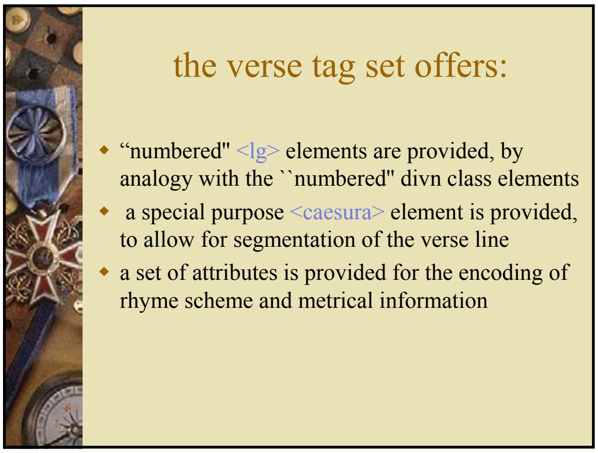


- <addName> contains an additional name component, such as a nickname, epithet, or alias, or any other descriptive phrase used within a personal name.
- <nameLink> contains a connecting phrase or link used within a name but not regarded as part of it, such as ``van der'' or ``of''.
- <genName> contains a name component used to indicating generational information, such as ``Junior'', or a number used in a monarch's name.



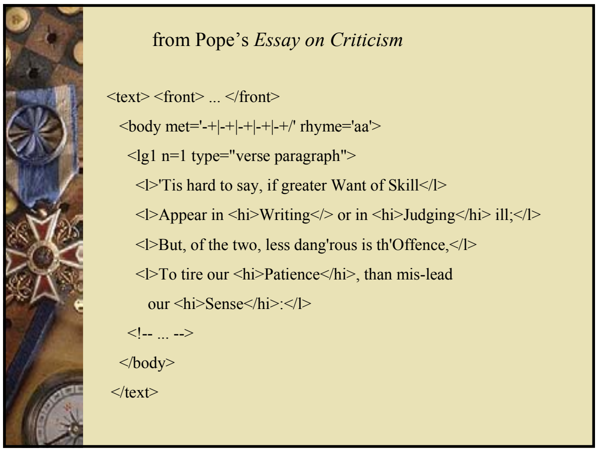
how does TEI Lite fulfill its goal?

- for encoding poetry, TEI Lite offers:
  - <lg> for encoding line groups (stanzas)
  - <l> for encoding a line of poetry



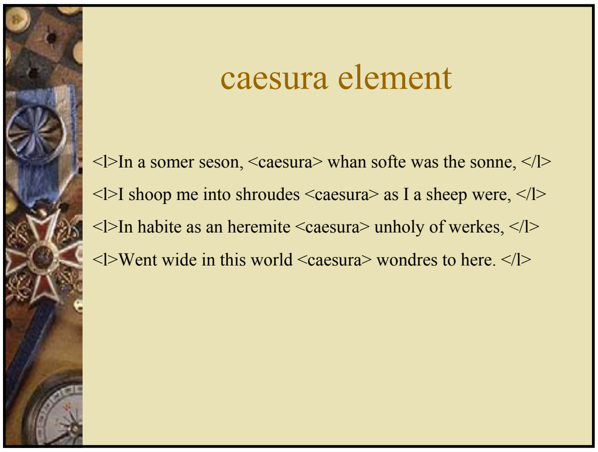
the verse tag set offers:

- “numbered'' <lg> elements are provided, by analogy with the ``numbered'' divn class elements
- a special purpose <caesura> element is provided, to allow for segmentation of the verse line
- a set of attributes is provided for the encoding of rhyme scheme and metrical information



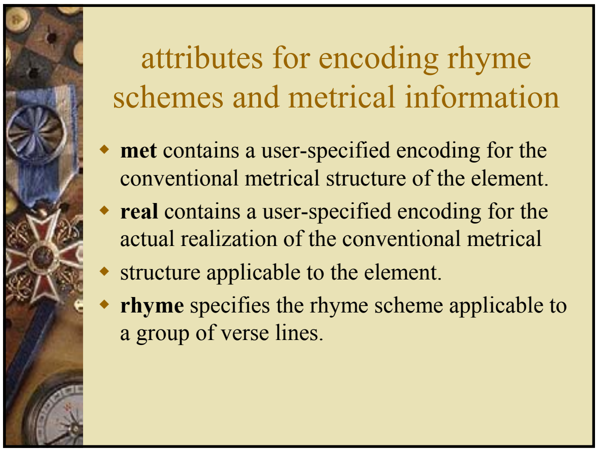
from Pope’s *Essay on Criticism*

```
<text> <front> ... </front>
  <body met='-+|-+|-+|-+|-+/' rhyme='aa'>
    <lg1 n=1 type="verse paragraph">
      <l>'Tis hard to say, if greater Want of Skill</l>
      <l>Appear in <hi>Writing</> or in <hi>Judging</hi> ill;</l>
      <l>But, of the two, less dang'rous is th'Offence,</l>
      <l>To tire our <hi>Patience</hi>, than mis-lead
         our <hi>Sense</hi>:</l>
    <!-- ... -->
  </body>
 </text>
```



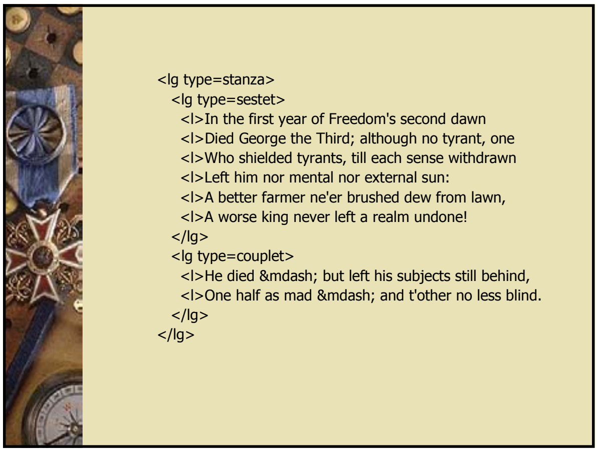
caesura element

```
<l>In a somer seson, <caesura> whan softe was the sonne, </l>
<l>I shoop me into shroudes <caesura> as I a sheep were, </l>
<l>In habite as an heremite <caesura> unholy of werkes, </l>
<l>Went wide in this world <caesura> wondres to here. </l>
```

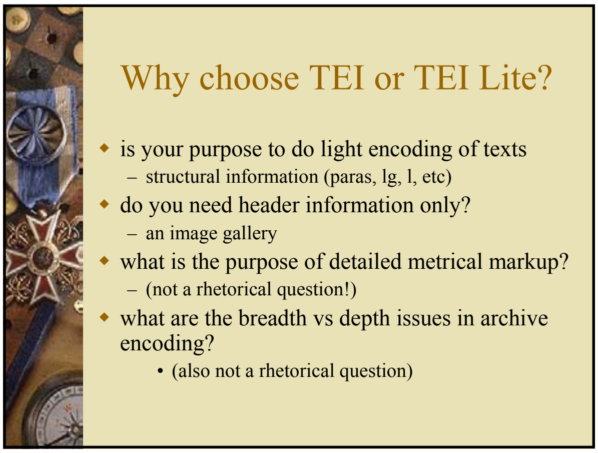


attributes for encoding rhyme schemes and metrical information

- **met** contains a user-specified encoding for the conventional metrical structure of the element.
- **real** contains a user-specified encoding for the actual realization of the conventional metrical
- structure applicable to the element.
- **rhyme** specifies the rhyme scheme applicable to a group of verse lines.



```
<lg type=stanza>
  <lg type=sestet>
    <l>In the first year of Freedom's second dawn
    <l>Died George the Third; although no tyrant, one
    <l>Who shielded tyrants, till each sense withdrawn
    <l>Left him nor mental nor external sun:
    <l>A better farmer ne'er brushed dew from lawn,
    <l>A worse king never left a realm undone!
  </lg>
  <lg type=couplet>
    <l>He died &mdash; but left his subjects still behind,
    <l>One half as mad &mdash; and t'other no less blind.
  </lg>
</lg>
```



Why choose TEI or TEI Lite?

- is your purpose to do light encoding of texts
  - structural information (paras, lg, l, etc)
- do you need header information only?
  - an image gallery
- what is the purpose of detailed metrical markup?
  - (not a rhetorical question!)
- what are the breadth vs depth issues in archive encoding?
    - (also not a rhetorical question)