Introduction to TEI for Research Centre for Japanese Language and Linguistics

TEI@Oxford

Hertford College, January 19th 2010

What are texts and markup?

JULY 1, 1914.] PUNCH, OR THE LONDON CHARIVARI. 1

CIVIL WAR

Bernard Partridge

Col. Calch.

PROGRESS.

["Giving evidence recently before a Select Committee of the House of Commons, Miss C. E. Collet, of the Home Office, said the commercial laundry was killing the small hand laundry."—*Evening News.*]

THE little crafts! How soon they die!
In cottage doors no shuttle clicks;
The hand-loom has been ousted by
A large concern with lots more sticks.

The throb of pistons beats around;
Great chimneys rise on Thames's banks;
The same phenomena are found
In Sheffield (Yorks) and Oldham (Lancs).

No longer now the housewife makes
Her rare preserves, for what 's the good?
The factory round the corner fakes

'Tis so with what we eat and wear,
Our bread, the boots wherein we splosh;
'Tis so with what I deemed most fair,
Most virginal of all—the Wash.

'Tis this that chiefly, when I chant,
Fulfils my breast with sighs of ruth,
To think that engines can supplant
The Amazons I loved in youth.

That not with tender care, as erst
By spinster females fancy-free,
These button-holes of mine get burst
Before the shift comes back to me;

That mere machines, and not a maid
With fingers fatuously plied,
The collars and the cuffs have frayed
That still excoriate my hide;

That steam reduces to such states
What once was marred by human skill;
That socks are sundered from their mates

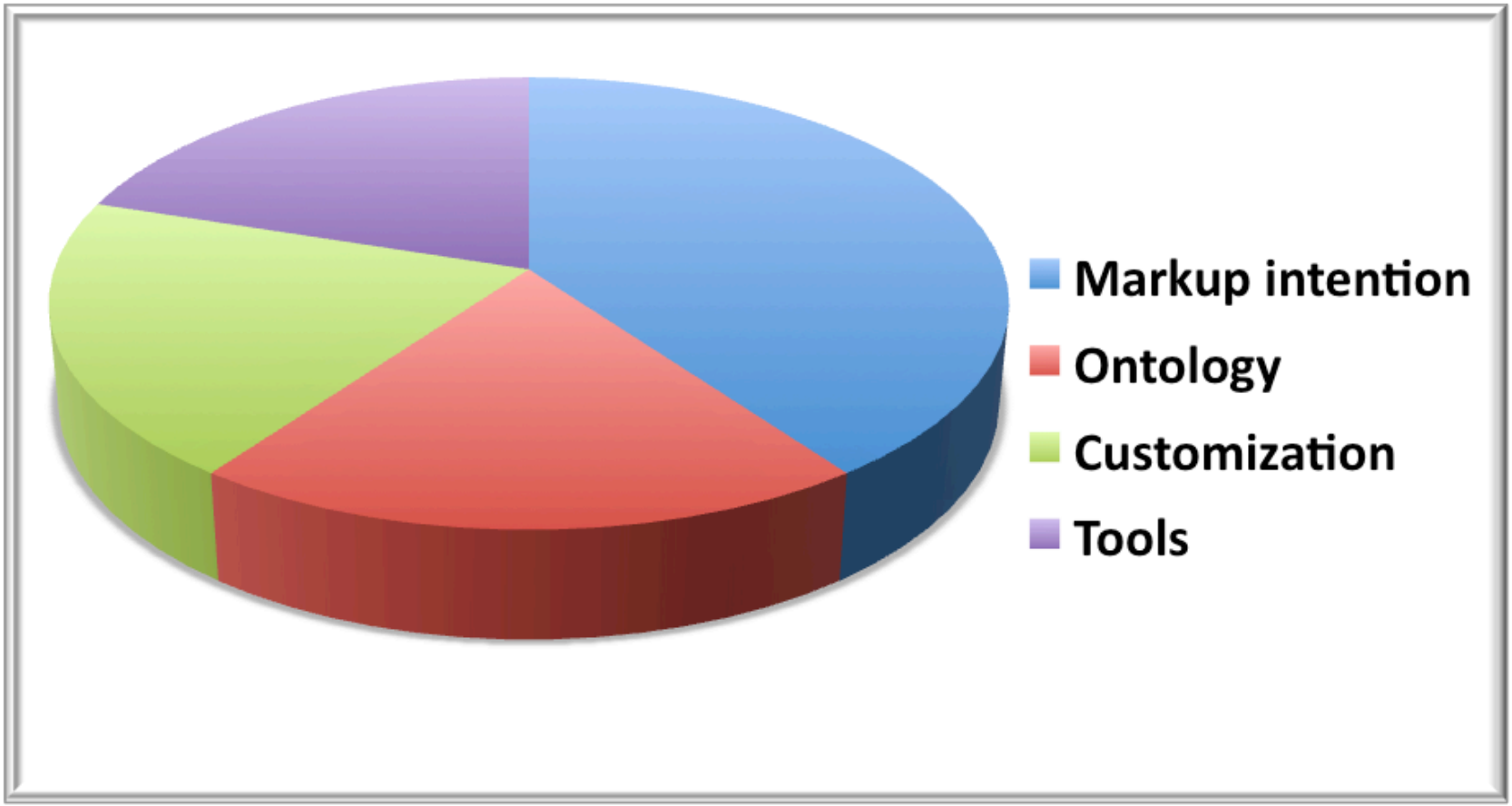
That not by Cupid's coy advance
(Some crone conniving at the fraud),
But simply by mechanic chance,
I get this handkerchief marked "Maud."

This is, indeed, a striking change;
I sometimes wonder if the world
Gets better as the skies grow strange
With coils of smoke about them curled.

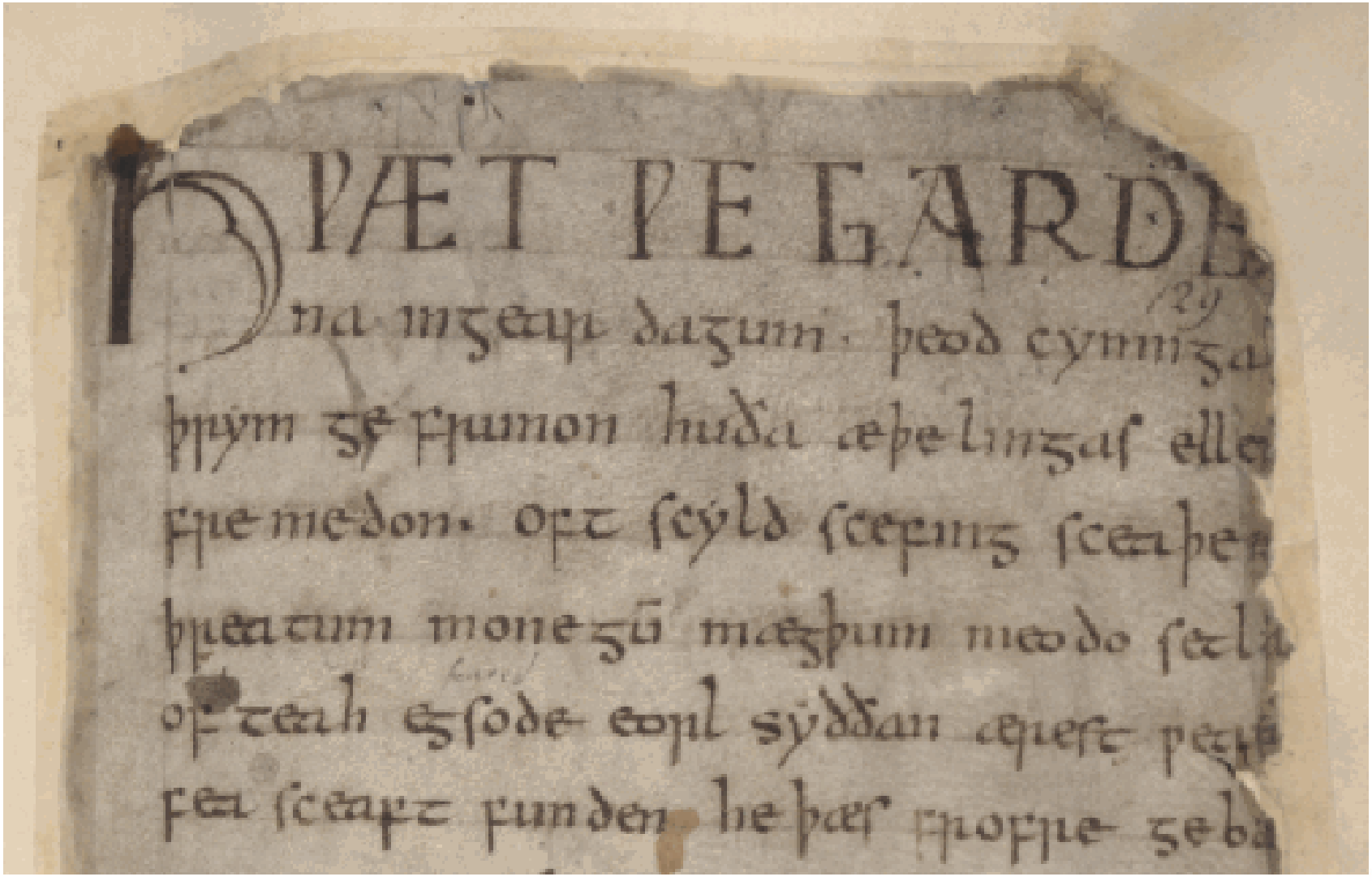
If the old days were not the best
Ere printed formulas conveyed
Sorrow about that silken vest
For all eternity mislaid;

Ere yet the unwieldy motor-van
Came clattering round the kerbstone's brink,
Its driver dreaming some new plan
To make my mauve pyjamas shrink

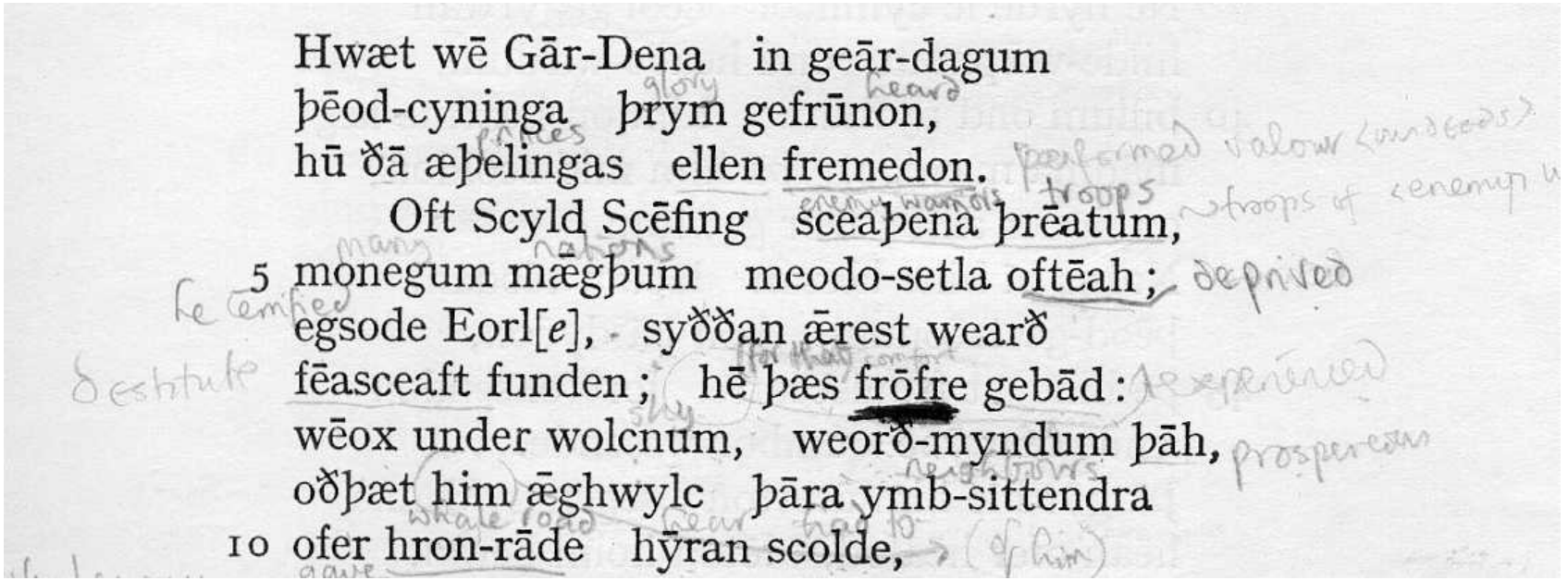
What do we need to make a textual resource?



What's in a text (2)?



What's in a text (3)?



Hwæt wē Gār-Dena in geār-dagum
þēod-cyninga þrym gefrūnon,
hū ðā æþelingas ellen fremedon.
Oft Scyld Scēfing sceaþena þrēatum,
monegum mægþum meodo-setla oftēah;
egsode Eorl[e], syððan ærest wearð
fēasceaft funden; hē þæs frōfre gebād:
wēox under wolcnum, weorð-myndum þāh,
oðþæt him æghwylc þāra ymb-sittendra
ofer hron-rāde hȳran scolde,

The ontology of text

Where is the text?

- in the shape of letters and their layout?
- in the original from which this copy derives?
- in the stories we read into it? or in its author's intentions?

A "text" is an abstraction, created by or for a community of readers. Markup encodes and makes concrete such abstractions.

The ontology of text

Where is the text?

- in the shape of letters and their layout?
- in the original from which this copy derives?
- in the stories we read into it? or in its author's intentions?

A "text" is an abstraction, created by or for a community of readers. Markup encodes and makes concrete such abstractions.

The ontology of text

Where is the text?

- in the shape of letters and their layout?
- in the original from which this copy derives?
- in the stories we read into it? or in its author's intentions?

A "text" is an abstraction, created by or for a community of readers. Markup encodes and makes concrete such abstractions.

Encoding of texts

- Texts are more than sequences of encoded glyphs
  - They have structure and content
  - They also have multiple readings
- Encoding, or markup, is a way of making these things explicit

Only that which is explicit can be reliably processed

What's the point of markup?

- To make explicit (to a machine) what is implicit (to a person)
- To add value by supplying multiple annotations
- To facilitate re-use of the same material
  - in different formats
  - in different contexts
  - by different users

What's the point of markup?

- To make explicit (to a machine) what is implicit (to a person)
- To add value by supplying multiple annotations
- To facilitate re-use of the same material
  - in different formats
  - in different contexts
  - by different users

What's the point of markup?

- To make explicit (to a machine) what is implicit (to a person)
- To add value by supplying multiple annotations
- To facilitate re-use of the same material
  - in different formats
  - in different contexts
  - by different users

Markup is also a scholarly activity!

- Deciding what markup to apply, and how far this represents the original, is a kind of textual editing
- There is (almost) no such thing as neutral markup — all of it involves interpretation
- Useful markup is rarely as easy or as quick to produce as people would have you believe
- Markup can assist both in answering and in expressing fundamental research questions
- Deciding *what* to markup is at least as important as deciding *how*... we call that *document analysis*

Markup is also a scholarly activity!

- Deciding what markup to apply, and how far this represents the original, is a kind of textual editing
- There is (almost) no such thing as neutral markup — all of it involves interpretation
- Useful markup is rarely as easy or as quick to produce as people would have you believe
- Markup can assist both in answering and in expressing fundamental research questions
- Deciding *what* to markup is at least as important as deciding *how*... we call that *document analysis*

Markup is also a scholarly activity!

- Deciding what markup to apply, and how far this represents the original, is a kind of textual editing
- There is (almost) no such thing as neutral markup — all of it involves interpretation
- Useful markup is rarely as easy or as quick to produce as people would have you believe
- Markup can assist both in answering and in expressing fundamental research questions
- Deciding *what* to markup is at least as important as deciding *how*... we call that *document analysis*

Markup is also a scholarly activity!

- Deciding what markup to apply, and how far this represents the original, is a kind of textual editing
- There is (almost) no such thing as neutral markup — all of it involves interpretation
- Useful markup is rarely as easy or as quick to produce as people would have you believe
- Markup can assist both in answering and in expressing fundamental research questions
- Deciding *what* to markup is at least as important as deciding *how*... we call that *document analysis*

Markup is also a scholarly activity!

- Deciding what markup to apply, and how far this represents the original, is a kind of textual editing
- There is (almost) no such thing as neutral markup — all of it involves interpretation
- Useful markup is rarely as easy or as quick to produce as people would have you believe
- Markup can assist both in answering and in expressing fundamental research questions
- Deciding *what* to markup is at least as important as deciding *how*... we call that *document analysis*

What does markup capture?

Compare

```
<hi rend="dropcap">H</hi>&amp;WYN;ÆT WE GARDE <lb/>na in
gear-dagum þeod-cyninga <lb/>þrym gefrunon,
hu ða æþelingas <lb/>ellen fremedon. oft scyld scefing sceaþe
<add>na</add>
<lb/>þreatum, moneg<ex>um</ex> mægþum meodo-setl
<add>a</add>
<lb/>of<damage>
 <desc>blot</desc>
</damage>teah ...
```

and

```
<lg>
 <l>Hwæt! we Gar-dena in gear-dagum</l>
 <l>þeod-cyninga þrym gefrunon,</l>
 <l>hu ða æþelingas ellen fremedon,</l>
</lg>
<lg>
 <l>Oft Scyld Scefing sceaþena þreatum,</l>
 <l>monegum mægþum meodo-setla ofteah;</l>
 <l>egsode Eorle, syððan ærest wearþ</l>
 <l>feasceaft funden...</l>
</lg>
```

A useful mental exercise

Imagine you are going to markup several thousand pages of complex material....

- Which features are you going to markup?
- Why are you choosing to markup this feature?
- How reliably and consistently can you do this?

Now, imagine your budget has been halved. Repeat the exercise!

Working with markup

What is markup for?

- exchanging data
  1. between people
  2. between people and machines
  3. between machines
- preserving information
  1. independent of particular applications
  2. independent of medium or hardware

Schemas and namespaces

- A namespace is one way of specifying the meaning of the markup introduced in a document: like a dictionary
- A more powerful way is to use a *schema*: a kind of grammar for your markup
- A namespace tells you who defined or claims this element, but not much more about how to use it
- A schema tells you how you are supposed to use an element, and thus allows you to validate your documents against a set of rules

What can a schema do for you?

- ensure that your documents use only predefined elements, attributes, and entities
- enforce structural rules such as ‘every chapter must begin with a heading’ or ‘recipes must include an ingredient list’
- make sure that the same thing is always called by the same name

Schema languages vary in the amount of validation they support

Schemas

There are a variety of schema languages

- XML DTD language
- ISO RELAX NG
- W3C Schema
- ISO Schematron

All have different tool kits, different syntaxes, and different methods of doing things

The TEI aims to be independent of schema language

The TEI encoding scheme is a framework providing:

- definitions and names for several hundred useful textual distinctions
- a set of modules that can be used to generate schemas making those distinctions
- a customization mechanism for modifying and combining those definitions with new ones using the same conceptual model
- a very simple consensus-based way of organizing and structuring textual (and others) resources...
- ... which can be enriched and personalized in highly idiosyncratic or specialised ways
- a very rich library of existing specialised components
- an integrated suite of standard stylesheets for delivering schemas and documentation in various languages and formats

The TEI aims to be independent of schema language

The TEI encoding scheme is a framework providing:

- definitions and names for several hundred useful textual distinctions
- a set of modules that can be used to generate schemas making those distinctions
- a customization mechanism for modifying and combining those definitions with new ones using the same conceptual model
- a very simple consensus-based way of organizing and structuring textual (and others) resources...
- ... which can be enriched and personalized in highly idiosyncratic or specialised ways
- a very rich library of existing specialised components
- an integrated suite of standard stylesheets for delivering schemas and documentation in various languages and formats

The TEI aims to be independent of schema language

The TEI encoding scheme is a framework providing:

- definitions and names for several hundred useful textual distinctions
- a set of modules that can be used to generate schemas making those distinctions
- a customization mechanism for modifying and combining those definitions with new ones using the same conceptual model
- a very simple consensus-based way of organizing and structuring textual (and others) resources...
- ... which can be enriched and personalized in highly idiosyncratic or specialised ways
- a very rich library of existing specialised components
- an integrated suite of standard stylesheets for delivering schemas and documentation in various languages and formats

The TEI aims to be independent of schema language

The TEI encoding scheme is a framework providing:

- definitions and names for several hundred useful textual distinctions
- a set of modules that can be used to generate schemas making those distinctions
- a customization mechanism for modifying and combining those definitions with new ones using the same conceptual model
- a very simple consensus-based way of organizing and structuring textual (and others) resources...
- ... which can be enriched and personalized in highly idiosyncratic or specialised ways
- a very rich library of existing specialised components
- an integrated suite of standard stylesheets for delivering schemas and documentation in various languages and formats

The TEI aims to be independent of schema language

The TEI encoding scheme is a framework providing:

- definitions and names for several hundred useful textual distinctions
- a set of modules that can be used to generate schemas making those distinctions
- a customization mechanism for modifying and combining those definitions with new ones using the same conceptual model
- a very simple consensus-based way of organizing and structuring textual (and others) resources...
- ... which can be enriched and personalized in highly idiosyncratic or specialised ways
- a very rich library of existing specialised components
- an integrated suite of standard stylesheets for delivering schemas and documentation in various languages and formats

The TEI aims to be independent of schema language

The TEI encoding scheme is a framework providing:

- definitions and names for several hundred useful textual distinctions
- a set of modules that can be used to generate schemas making those distinctions
- a customization mechanism for modifying and combining those definitions with new ones using the same conceptual model
- a very simple consensus-based way of organizing and structuring textual (and others) resources...
- ... which can be enriched and personalized in highly idiosyncratic or specialised ways
- a very rich library of existing specialised components
- an integrated suite of standard stylesheets for delivering schemas and documentation in various languages and formats

The TEI aims to be independent of schema language

The TEI encoding scheme is a framework providing:

- definitions and names for several hundred useful textual distinctions
- a set of modules that can be used to generate schemas making those distinctions
- a customization mechanism for modifying and combining those definitions with new ones using the same conceptual model
- a very simple consensus-based way of organizing and structuring textual (and others) resources...
- ... which can be enriched and personalized in highly idiosyncratic or specialised ways
- a very rich library of existing specialised components
- an integrated suite of standard stylesheets for delivering schemas and documentation in various languages and formats

The TEI framework

Reasons for attempting to define a common framework:

- re-usability and repurposing of resources
- modular software development
- lower training costs
- 'frequently answered questions' — common technical solutions for different application areas

The TEI was *designed* to support multiple views of the same resource

Some terminology

- The TEI encoding scheme consists of a number of *modules*
- Each module contains a number of *element specifications* (marked up in TEI using the `<elementSpec>` element)
- Each element specification contains:
  - a canonical name (`<gi>`) for the element, and optionally other names in other languages
  - a canonical description (also possibly translated) of its function
  - a declaration of the *classes* to which it belongs
  - a definition for each of its *attributes*
  - a definition of its *content model*
  - usage examples and notes
- a TEI *schema* specification (`<schemaSpec>`) is made by selecting modules and (optionally) modifying their contents
- a TEI document containing a schema specification is called an *ODD* (One Document Does it all)

What is a module?

- A convenient way of grouping together a number of element declarations
- These are usually on a related topic or specific application
- Most chapters of P5 focus on elements drawn from a single module, which that chapter then defines
- A TEI Schema is created by selecting modules and adding or removing elements from them as needed

Which modules exist?

Module name	Chapter
analysis	Simple Analytic Mechanisms
certainty	Certainty and Responsibility
core	Elements Available in All TEI Documents
corpus	Language Corpora
dictionaries	Dictionaries
drama	Performance Texts
figures	Tables, Formulae, and Graphics
gaiji	Representation of Non-standard Characters and Glyphs
header	The TEI Header
iso-fs	Feature Structures
linking	Linking, Segmentation, and Alignment
msdescription	Manuscript Description
namesdates	Names, Dates, People, and Places
nets	Graphs, Networks, and Trees
spoken	Transcriptions of Speech
tagdocs	Documentation Elements
tei	The TEI Infrastructure
textcrit	Critical Apparatus
textstructure	Default Text Structure
transcr	Representation of Primary Sources
verse	Verse

The TEI Class System

- The TEI distinguishes over 500 elements,
- Having these organised into classes aids comprehension, modularity, and modification.
- *Attribute class*: the members share common attributes
- *Model class*: they can appear in the same locations (and are often semantically related)
- Classes may contain other classes
- An element can be a member of any number of classes, irrespective of the module it belongs to.

Attribute Classes

- Attribute classes are given (usually adjectival) names beginning with att.; e.g. *att.naming*, *att.typed*
- all members of `att.naming` inherit from it attributes *@key* and *@ref*; all members of `att.typed` inherit from it *@type* and *@subtype*
- If we want an element to carry the *@type* attribute, therefore, we add the element to the `att.typed` class, rather than define those attributes explicitly.

What now?

What tools do we need?

- Appropriately expressive vocabularies (eg TEI XML)
- Syntax-checking document creation tools (ie editors)
- Document transformation tools
- Document delivery tools
- Document storage and management tools
- Programming interfaces
- Specialized applications

Problems and solutions

- A methodology: markup
- An ontology: Text Encoding Initiative guidelines
- Modules: for different subject areas
- * ODD: Customizable schema(s)
- * Tools: to help build the schemas
- Tools: to creation and editing text

Now we've got a resource…

* we have not addressed these today.