Live long and prosper! Lessons from the TEI

TEI @ Oxford

September 2008
1986 was a long time ago…

- The first computer virus – Brain – appears, in the USA
- Construction of the channel tunnel begins
- The Soviet Union launches space station Mir
- Disaster at Chernobyl
- Olaf Palme assassinated
- Records of the year: *Raising Hell* (Run DMC)… *Graceland* (Paul Simon)… Группа крови (Виктор Цой)
1986 was a long time ago…

- The first computer virus – Brain – appears, in the USA
- Construction of the channel tunnel begins
- The Soviet Union launches space station Mir
- Disaster at Chernobyl
- Olaf Palme assassinated
- Records of the year: *Raising Hell* (Run DMC)… *Graceland* (Paul Simon)… Группа крови (Виктор Цой)
1986 was a long time ago…

- The first computer virus – Brain – appears, in the USA
- Construction of the channel tunnel begins
- The Soviet Union launches space station Mir
- Disaster at Chernobyl
- Olaf Palme assassinated
- Records of the year: *Raising Hell* (Run DMC)… *Graceland* (Paul Simon)… Группа крови (Виктор Цой)
1986 was a long time ago…

- The first computer virus – Brain – appears, in the USA
- Construction of the channel tunnel begins
- The Soviet Union launches space station Mir
- Disaster at Chernobyl
- Olaf Palme assassinated
- Records of the year: *Raising Hell* (Run DMC)… *Graceland* (Paul Simon)… Группа крови (Виктор Цой)
1986 was a long time ago…

- The first computer virus – Brain – appears, in the USA
- Construction of the channel tunnel begins
- The Soviet Union launches space station Mir
- Disaster at Chernobyl
- Olaf Palme assassinated

Records of the year: *Raising Hell* (Run DMC)… *Graceland* (Paul Simon)… Группа крови (Виктор Цой)
1986 was a long time ago…

- The first computer virus – Brain – appears, in the USA
- Construction of the channel tunnel begins
- The Soviet Union launches space station Mir
- Disaster at Chernobyl
- Olaf Palme assassinated
- Records of the year: *Raising Hell* (Run DMC)… *Graceland* (Paul Simon)… Группа крови (Виктор Цой)
...but we used computers then

- Corpus linguistics
- Databases on CD ROM
- Largescale lexical resources already existed (e.g. TLF, TLG, LASLA…)
- Digital lexicography (e.g. OED)
- Document management systems (e.g. TeX, Scribe, tRoff..)
  - some proprietary (and expensive), some research
- Text archives
- Hypertext theory

But there was no world wide web and not many desktop pcs…
Birth of the Text Encoding Initiative

- Spring 1987: European workshops on standardisation of historical data (J.P. Genet, M Thaller)
- Autumn 1987: NEH funds an exploratory international workshop on the feasibility of defining "text encoding guidelines"
Today's question:

- So the TEI is *very old*!
- It comes from a time before the Web, before the DVD, the mobile phone, cable tv, or Microsoft Excel
- Not much in computing survives 5 years, never mind 20
- What relevance can it possibly have today?
- Why is it still here, and how has it survived?
Is the TEI still relevant?

- With XML everyone can create their own markup system and still share data!
- In the Semantic Web, XML systems will all understand each other's data!

If we have

- historical data marked up with a Historical Markup Language
- linguistic data marked up with a Linguistic Markup Language
- metadata marked up with a Metadata Markup Language

how will we integrate resources or ask interesting questions?

Haven't we been here before?
The TEI provides

- a language-independent framework for defining markup languages
- a very simple consensus-based way of organizing and structuring textual (and other) 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
- a large and active open source style user community
The TEI provides

- a language-independent framework for defining markup languages
- a very simple consensus-based way of organizing and structuring textual (and other) 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
- a large and active open source style user community
The TEI provides

- a language-independent framework for defining markup languages
- a very simple consensus-based way of organizing and structuring textual (and other) 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
- a large and active open source style user community
Relevance 1

The TEI provides

- a language-independent framework for defining markup languages
- a very simple consensus-based way of organizing and structuring textual (and other) 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
- a large and active open source style user community
The TEI provides

- a language-independent framework for defining markup languages
- a very simple consensus-based way of organizing and structuring textual (and other) 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
- a large and active open source style user community
The TEI provides

- a language-independent framework for defining markup languages
- a very simple consensus-based way of organizing and structuring textual (and other) 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
- a large and active open source style user community
Why would you want those things?

- because we need to interchange resources
  - between people
  - (increasingly) between machines

- because we need to integrate resources
  - of different media types
  - from different technical contexts

- because we need to preserve resources
  - cryogenics is not the answer!
  - we need to preserve metadata as well as data
Why would you want those things?

- because we need to interchange resources
  - between people
  - (increasingly) between machines
- because we need to integrate resources
  - of different media types
  - from different technical contexts
- because we need to preserve resources
  - cryogenics is not the answer!
  - we need to preserve metadata as well as data
Why would you want those things?

- because we need to interchange resources
  - between people
  - (increasingly) between machines

- because we need to integrate resources
  - of different media types
  - from different technical contexts

- because we need to preserve resources
  - cryogenics is not the answer!
  - we need to preserve metadata as well as data
The virtuous circle of encoding

**Resources**

- **abstract model**
- **encoding**
- **digital resources**
- **analysis**
The scope of intelligent markup

Even within the original scope of the TEI we have

- basic structural and functional components
- diplomatic transcription, images, annotation
- links, correspondence, alignment
- data-like objects such as dates, times, places, persons, events (named entity recognition)
- meta-textual annotations (correction, deletion, etc)
- linguistic analysis at all levels
- contextual metadata of all kinds
- ... and so on and so forth

Is it possible to delimit encyclopaedically all possible kinds of markup?
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
Old Skool TEI

- A traditional (if large) research project with soft funding, driven by academic curiosity
- A codification of best practice, with no formal maintenance method
- Uncertain licencing and development practices
- Perceived as unmanageably complex except by the priesthood — or simultaneously as too simple for real scholarly work
- Lack of specific tools to do something with a TEI text
- Failure to market the advantages of rich markup
Proper open source licence, with visible development on Sourceforge

Architecture rethought to facilitate expansion and integration with other systems

Self documenting, each release fully validated, delivered using standard mechanisms

Publicly available processing tools managed together with the Guidelines

Active developer community, wiki, etc. Test files, exemplars, regular updates…

New governance structure, new tools, new modules…
Three important things about TEI P5

1. Being a good digital citizen:
   - Support for multiple schema languages and namespaces
   - Reliance on XML, and hence on Unicode
   - Validation of attributes and datatyping
   - Use of W3C pointers and paths

2. Making it flexible:
   - ODD: a single specification language for developers, users, and teachers, integrating schema and documentation;
   - Verifiable conformance

3. Old annoyances removed and some new topics added
A set of TEI documents is described by an ODD, which is itself a TEI document that combines:
- references to existing declarations
- formal declarations for elements and attributes
- documentation and usage notes

Underlying this:
- a conceptual model which abstracts from specific elements to generic classes
- a modular architecture for combining sets of definitions

Specifications are chainable; modifications are written in ODD with ODD as input and output

Roma is one interface to this: there will be others
For example

An ODD file is a valid TEI document, containing descriptive prose, and a `<schemaSpec>` element to define the schema it documents.

```xml
<div>
  <head>Our Project Manual</head>
  <p>In this project we use the basic TEI structures with a few minor modifications to exclude elements we do not need</p>
  <schemaSpec ident="TEI-minimal" start="TEI">
    <moduleRef key="tei"/>
    <moduleRef key="header"/>
    <moduleRef key="core"/>
    <moduleRef key="textstructure"/>
    <!-- We don't need these drama elements: -->
    <elementSpec ident="sp" mode="delete" module="core"/>
    <elementSpec ident="speaker" mode="delete" module="core"/>
  </schemaSpec>
</div>
```
Support for many schema languages

- TEI schemas can be generated for
  - XML DTD language
  - ISO RELAX NG language
  - W3C Schema Language

- Content models are defined using RELAX NG syntax
- Datatypes are defined in terms of W3C datatypes
- Some facilities (e.g. alternation, namespaces) cannot be expressed in DTD
- Additional constraints can be expressed in Schematron
Two reasons why standards fail

- The theory is not yet ripe
- The "not invented here" attitude: the community of users is too diverse
Coping with partially-baked ideas

In a TEI ODD, you can …

- constrain the domain of a value list
- enforce Schematron rules about e.g. co-dependency
- provide new elements in your own namespace
- remove (non-mandatory) child elements
New elements

A schema is a grammar. How can you add new terminals to an existing syntax?

- All content models are expressed indirectly, by reference to element classes rather than elements
- Hence adding a new element is simply a matter of saying which class/es it belongs to

The TEI schema is also enriched with semantics. How can you explain what a new element means?

- Class membership also conveys some semantics
- ODD includes detailed documentation
Coping with the NIH Syndrome

- TEI P5 has extensive I18N features for translation of …
  - schema objects
  - schema documentation
- TEI is hospitable to other namespaces
  - You can use SVG for graphics, MathML for math, Word Table markup if you like
  - (but note this doesn't solve the Other Overlap Problem)
- ODD also includes an `<equiv>` element for mapping to external ontologies
Embedding SVG within TEI:

```xml
<figure>
  <svg xmlns="http://www.w3.org/2000/svg"
       width="6cm" height="5cm" viewBox="6 3 6 5">
    <ellipse xmlns="http://www.w3.org/2000/svg"
             style="fill:#ffffff"
             cx="9.75" cy="6.35" rx="2.75" ry="2.75"/>
  </svg>
</figure>
```

A user-defined attribute:

```xml
<div xmlns:my="http://www.example.org/ns/nonTEI">
  <p n="12" my:topic="rabbits">Flopsy, Mopsy, Cottontail, and Peter…</p>
</div>
```

An NVDL processor can validate a document using multiple namespace schemas.
Conformance issues

A document is TEI Conformant if and only if it …

- is a well-formed XML document
- can be validated against a TEI Schema, that is, a schema derived from the TEI Guidelines
- conforms to the TEI Abstract Model
- uses the TEI Namespace (and other namespaces where relevant) correctly
- is documented by means of a TEI Conformant ODD file which refers to the TEI Guidelines

Or if it can be transformed automatically using some TEI-defined procedures into such a document (it is TEI-conformable)

Standardization should not mean ‘Do what I do’, but rather ‘Explain what you do in terms I can understand’
Evolution works!

1. Make modifications in your own namespace
2. Document them in an ODD
3. Propose them to the TEI Council as amendments or feature requests
4. TEI P5 now has a 6 month release cycle…

Visit http://www.tei-c.org for more background info

Visit http://tei.sf.net to download
Using the basic TEI structural elements

TEI @ Oxford

September 2008
The TEI encoding scheme consists of a number of modules. These declare XML elements and their attributes. An element's declaration assigns it to one (or more) model classes. Another part declares its possible content and attributes with reference to these classes. This indirection allows strength and flexibility. It makes it easy to add/exclude new elements by referencing existing classes.
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 focus on elements drawn from a single module, which that chapter then defines
- A TEI Schema is created by selecting modules and add/removing elements from them as needed
<table>
<thead>
<tr>
<th>Module name</th>
<th>Chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>analysis</td>
<td>Simple Analytic Mechanisms</td>
</tr>
<tr>
<td>certainty</td>
<td>Certainty and Responsibility</td>
</tr>
<tr>
<td>core</td>
<td>Elements Available in All TEI Documents</td>
</tr>
<tr>
<td>corpus</td>
<td>Language Corpora</td>
</tr>
<tr>
<td>dictionaries</td>
<td>Dictionaries</td>
</tr>
<tr>
<td>drama</td>
<td>Performance Texts</td>
</tr>
<tr>
<td>figures</td>
<td>Tables, Formulae, and Graphics</td>
</tr>
<tr>
<td>gaiji</td>
<td>Representation of Non-standard Characters and Glyphs</td>
</tr>
<tr>
<td>header</td>
<td>The TEI Header</td>
</tr>
<tr>
<td>iso-fs</td>
<td>Feature Structures</td>
</tr>
<tr>
<td>linking</td>
<td>Linking, Segmentation, and Alignment</td>
</tr>
<tr>
<td>msdescription</td>
<td>Manuscript Description</td>
</tr>
<tr>
<td>namesdates</td>
<td>Names, Dates, People, and Places</td>
</tr>
<tr>
<td>nets</td>
<td>Graphs, Networks, and Trees</td>
</tr>
<tr>
<td>spoken</td>
<td>Transcriptions of Speech</td>
</tr>
<tr>
<td>tagdocs</td>
<td>Documentation Elements</td>
</tr>
<tr>
<td>tei</td>
<td>The TEI Infrastructure</td>
</tr>
<tr>
<td>textcrit</td>
<td>Critical Apparatus</td>
</tr>
<tr>
<td>textstructure</td>
<td>Default Text Structure</td>
</tr>
<tr>
<td>transcr</td>
<td>Representation of Primary Sources</td>
</tr>
<tr>
<td>verse</td>
<td>Verse</td>
</tr>
</tbody>
</table>
The Imaginary Punch Project

- **Punch** is a famous English humorous journal, published regularly between 1841 and 1992: see [http://www.punch.co.uk/historyofpunch.html](http://www.punch.co.uk/historyofpunch.html).

- A project plans to make available fully marked up texts of the journal, in conjunction with page images...
  - for social historians
  - for librarians
  - for linguists

- How will the TEI help? And which parts of the TEI will we use?
CHARIVARIA.

Two men carrying books were arrested last week on the outskirts of Paris, and are suspected of a plot against the French President. They alleged that the books were made for the use of Hessians, but the Trafalgar defence that gave the commission.

The town of Cracow, it is reported, has decided to give up gas in favour of electricity. This, of course, is not meant as a slight on its most illustrious resident.

Posted at Tenby, Isle of Wight, on July 14, 1901, a postcard has just been delivered at the Drapers Hotel in Dover. The recipient is said to have expressed the opinion that it would have been quicker, almost to have telephoned the message.

Miss Nellie Boyle, of the Women's Freedom League, has sent to the press a list of ladies on whom she considers the King ought to honour. Among the names there is one notable omission, and Miss Marie Connelly is said to be more of an anti-Suffragette than ever.

"NEW THEATRE FOR LONDON: ALL SEATS IN THE HOUSE TO BE BOOKED."

So the great difficulty has been solved at last! So may theatres fall because the seats are not taken.

A movement is on foot to induce Mr. Charles Garnett to change the name of his play, A Heritage of Hate, as so many patrons of melodrama have experienced difficulty in pronouncing this title as it stands at present.

In a struggle between a British sailor and a German policeman at Wilhelmshaven the other day honour seems to have been fairly even. The police, who used their sword, lost his head, and the sailor a piece of his hand.

Two men of good position were tried last week before the State Court of Berlin for refusing to address a policeman as "Mr." That will surprise no one who knows Prussia. It is the sequel which wakes our breasts away. The two men were acquitted!

Volume 10 of the Census of 1911 shows that in the preceding ten years the number of householders of the Established Church increased from 35,383 to 34,800. "The decrease is accounted for by the lack of young men taking orders." The wonder is that such orders were not at once snapped up by elder Germans.

Miss Evelyn Woodhouse, of Neuburg, known as "The Big Hat Girl," has not yet sailed from New York in the "Imperator" with a hat which measures 58 inches in diameter. These giant ladies are justifying themselves.

We are glad that the Postmaster-General has promised a Bill against foreign sweepers. Only the other day we received a circular headed "Submires SchoeplySchwepit!"

While we are ready to grant that the French Navy is about to try the experiment ofristling (black sailors) we should say that they will be found to make the most admirable sailors, not showing the dirt like the white men.

Describing a recent visit of a party of Congressmen and State officials to one of the naval battleships of the American Navy, a contemporary says: "The distinguished guests took water with what grace they could." Evidently they thought it was only worth saying once for.

The statement made last week in the course of a certain trial that "as a man grows old he becomes more con" has had a curious sequel. Order, not pouring in fees the Cambridge I. O. S. for consignments of contraceptives.

THE PROFESSIONAL ATTITUDE.

The person said, according to a daily paper, is not in to be not by law-making. She prefers a cheerful and amusing companion.

Dear, of old I wore devotion In the manner knights employed, Wrote epistles with emotion (Which I trust have been destroyed).

Now at last, a practised lover. Blowing conquests not a few.

I am told to put a cover On my sentiments for you.

Ophelia's chat is cut of fashion. Stumpy wools are never sold.

Voiles come a crack with passion. Shake with meritment instead.

Poets qualified to tackle.

Love notes when inspired Strego to make the ladies crack— Nothing further is required.

Ishboness one whose occupation Has a still and solemn trend Might enjoy, as relaxation, Jesting with a female friend; But, corrupted by the mason.

That my written letters bring, Here on earth can't be funny. For the pleasure of the thing?

The Daily Chronicle on the latest submarine:

"It will also be equipped with a quick-firing gun, which disappears when the vessel is submerged."

This is far the best arrangement; it would never do for it to be left floating where any passer-by could pick it up.

According to La France Militaire...
Great Economy Effected by Co-Operation in Advertisement.

The Militants' Tariff.

Mrs. Bungalow Snatcher, having entered into partnership with the Misses Burnham Blazer, as General Agents of Destruction, begs to inform the public that the firm will be prepared to examine communications of all kinds, at the shortest notice, on the very moderate terms given below:

For breaking windows, per window........ $1.00
For bowling, kicking, or hitting during service in church, per hour, if fine........ 0.25
For sitting on doorsteps of obnoxious persons, per hour, if soiled........ 0.15
For sitting on doorsteps of obnoxious persons, per hour, if wet........ 0.10
For damaging gull-gums, per gram........ 0.05
For throwing stones at magistrates in court, according to size and weight of stone, from 0.25 to 1.00
For beating officials connected with goss........ 0.50
For slashing and hacking valuable pictures, from........ 0.75 to 2.00
For books not intended to explode........ 0.50
For burning down a house, according to value and social position of owner, from........ 10.00 to 25.00
For insulting exalted Personages, per insult........ 10.00
For burning down a modern red-brick church........ 15.00
For burning down a splendidly valuable and interesting ancient one (eleventh and twelfth centuries extra)........ 25.00
For bombs warranted to destroy an ordinary church........ 30.00
For bombs suited to wreck really superior buildings, such as Westminster Abbey and St. Paul's........ 50.00

For disturbing public meetings and the general harvesting and annoyance of all peaceable and decent people........ No charge.

Bungalow Snatcher, Burnham Blazer & Co. beg to assure their patrons that all the choppers, hammers, bombs, stones, etc., employed in their business are of the very best quality, and only refined paraffin and wax matches will be used in burning down any building.

Being in a position to offer such exceptional advantages, they trust to receive a large measure of support in their elevating and enlightening work.

If none of the above is found suitable to the needs of intemperate clients, a further list of assorted outrages will be supplied on application.

Love's Logic.

My happiness is in another's keeping,
My heart delivered to a maiden's care.
And she can cast it down or set it ranging.
(The latter process is extremely rare).

Oh, would that love indeed had made me blind,
That I might put her image out of mind.

Yet if I looked at her with eyes unseeing,
Her voice and laughter would not pass unheeded;
I should not be a reasonable being,
I still would trouble at her lightest word.

How could I then gain freedom from the spell
Unless I turned completely deaf as well?

So, blind and deaf, I might perhaps recover
A partial peace of mind, but all in vain,
For memory pursues the helpless lover,
And only death can ease him of his pain.

Thus, having proved that I was better dead,
I think I'll go and talk to her instead.
### Example 3

**The World’s Workers.**

L. G. V.: “Please, Mrs. Murphy, never say, if it’s five o’clock, will ten o’clock with ‘er?”

**The Punch, or the London Charivari.**

<table>
<thead>
<tr>
<th>Date</th>
<th>Author/Reviewer</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>June 15, 1914</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**The Sunday Deliverer.** “... remarkable book.”

**The Simla Gazette.** “... verdict... profoundly enlivening work of fiction.”

**The Bury Post.** “... a brilliant first effort.”

**The Enquiry.** “... well written...”

**The Tatler Express.** “... amazing...”

**The Monthly Citizen.** “The characters have life and movement.”

---

*Before lunch each session held its annual meeting in private, and at two o’clock the respectable report, which was greatly disliked by the volunteer. After being posted upon by a photographer the party split.*

*We were too young, and afraid they had overdone it.*

*From a photographic catalogue...* “This is a most complete little Projector... It is quite self-contained and will project a thirty-inch picture anywhere at a moment’s notice.”

It should be installed at the Royal Academy without delay.
Looking at Punch, what do we need to mark up?

- issue information and page number for reference purposes
- "chunks" or divisions of text, which may contain a picture, a poem, some prose, some drama, or a combination
- within the chunks, we can identify formal units such as
  - a picture, a caption
  - stanzas, lines
  - paragraphs
  - speeches and stage-directions
- and more...
We will treat each issue as a single `<text>` element, and each identifiable chunk within it as a `<div>` element of a particular type (e.g. cartoon, verse, prose).

For example, page 1 has two divisions,

```xml
<pb n="1"/>
<div type="cartoon">....</div>
<div type="poem">
  <head>Progress</head>....
</div>
```

Page 2 also has two, of different types:

```xml
<pb n="2"/>
<div type="prose">
  <head>The enchanted castle</head>....
</div>
<div type="snippet">
  <head>Correspondence</head>....
</div>
```
Why divisions rather than pages?

Because a division can start on one page (page 5 for example) and finish on another (page 6).
We use an empty element `<pb>` to mark the boundary between pages, rather than enclosing each page in a `<div type="page">`.

```xml
<pb n="5"/>
<div type="cartoon">...</div>
<div type="review">
    <head>Egypt in Venice</head>...
    <pb n="6"/>
</div>
...<div type="cartoon">...</div>
<div type="verse">
    <head>Enigma</head>...
</div>
<div type="snippets">...</div>
```

The sequence in which divisions appear is rather arbitrary.
Divisions can contain divisions...

TEI also provides division elements with names that indicate their degree of nesting (<div1>, <div2> etc.) which some people prefer.

Divisions must always tessellate: once "down" a level, you cannot pop "up" again within the same division.

```xml
<div type="snippets">
  <div type="snippet">Curiously....Chancellor</div>
  <div type="snippet">Men for the Antarctic... Canadians</div>
</div>
```
As mentioned above, `<div>`s must tessellate over the entire text

```
<div1>
  <p> ... </p>
<div2>
  <p> ... </p>
</div2>
<div2>
  <p> ... </p>
</div2>
</div1>
```

is valid but

```
<div1>
  <p> ... </p>
<div2>
  <p> ... </p>
</div2>
<div2>
  <p> ... </p>
</div2>
<div2>
  <p> ... </p>
</div2>
</div1>
```

is *not valid*.

A special `<floatingText>` element is available for "interruptions"
What are divisions made of?

(apart from other smaller divisions)

- `<head>` (heading)
- `<p>` (paragraph)
- `<sp>` (speech, contains any of the foregoing, also `<stage>` and `<speaker>`)
- `<list>` (contains `<head>`, `<label>`, `<item>`)  
- `<table>` (contains `<row>` containing `<cell>`)...
- `<l>` (verse line) optionally grouped into `<lg>` (line group) stanzas
- `<figure>` (contains `<graphic>`, `<figDesc>`, `<head>`)...
For example....

Page 3 contains a figure and a dialogue...

```
<div type="cartoon">
  <figure>
    <head>When the ships come home</head>
    <figDesc>A man in Turkish dress lounges on a sofa, smoking a cigarette and consulting a book labelled "Naval ledger". Another man, in traditional Greek costume, stands beside him, also reading a notebook.</figDesc>
    <graphic url="Punch/XML/Graphics/003.png"/>
  </figure>
  <sp>
    <speaker>Greece.</speaker>
    <p>Isn't it time we started fighting again?</p>
  </sp>
  <sp>
    <speaker>Turkey.</speaker>
    <p>Yes, I daresay. How soon could you begin?</p>
  </sp>
  <sp>
    <speaker>Greece.</speaker>
    <p>Oh, in a few weeks.</p>
  </sp>
  <sp>
    <speaker>Turkey.</speaker>
    <p>No good for me. Shan't be ready till the autumn.</p>
  </sp>
</div>
```
For example...

*The militants' tariff* (on Page 15) contains headings, paragraphs, and a table...

```xml
<div type="prose">
  <head>THE MILITANTS' TARIFF.</head>
  <head rend="right">Etna Lodge, W.</head>
  <p>Mrs. Bangham Smasher, having entered into partnership with the Misses Burnham Blazer, as General Agents of Destruction, begs to inform the public that the firm will be prepared to execute commissions of all kinds, at the shortest notice, on the very moderate terms given below: —

</p>
<table>
  <row role="label">
    <cell>£</cell>
    <cell>s.</cell>
    <cell>d.</cell>
  </row>
  <row>
    <cell>For breaking windows, per window ...</cell>
    <cell>0</cell>
    <cell>7</cell>
    <cell>6</cell>
  </row>
  <row>
    <cell>For howling, kicking, or biting during service in church, per howl, kick, or bite ...</cell>
    <cell>0</cell>
    <cell>10</cell>
  </row>
</table>
```
Global attributes

Some features (potentially) apply to everything:

- identity
- language
- rendition

TEI provides global attributes for these:

- `@xml:id` provides a unique identifier for any element;
- `@n` provides a name or number for any element;
- `@xml:lang` specifies the language of any element, using an ISO standard code;
- `@rend` and `@rendition` provide ways of specifying the visual appearance (rendition) of any element.
Egypt in Venice (on Page 05) begins with two headings, one in French....

<div type="prose" xml:lang="en" xml:id="I1914-07-01_05_02">
  <head>Egypt in Venice. </head>
  <head xml:lang="fr" rend="it">"La Légende de Joseph." </head>
  <p>Those who know the kind of attractions that the Russian ballet offers in so many of its themes .... </p>
</div>

Each stanza of the poem on page 10 has a last line which is significantly indented:

<lg>
  There were eight pretty walkers who went up a hill;
  They were Jessamine, Joseph and Japhet and Jill,
  And Allie and Sally and Tumbledown Bill,
  And Farnaby Fullerton Rigby.
</lg>
Macrostructure 1

All the issues of *Punch* for one year make up a volume. We could regard the volume as a single `<text>`, and each issue as a `<div>` within it. Or we could use the `<group>` element:

```xml
<text xml:id="v147">
  <front>
    <!-- introductory materials for volume 147 here -->
  </front>
  <group>
    <text xml:id="I1914-07-01">
      <body>
        <!-- first issue (1 July) -->
      </body>
    </text>
    <text xml:id="I1914-07-15">
      <body>
        <!-- second issue (15 July) -->
      </body>
    </text>
    <!-- etc... -->
  </group>
  <back>
    <!-- volume index, appendix etc. -->
  </back>
</text>
```
As well as the texts, we have detailed metadata about each volume, and images of its pages. These are the three parts of a canonical TEI document:

```xml
<TEI>
  <teiHeader>
    <!-- required; provides metadata -->
  </teiHeader>
  <facsimile>
    <!-- the text, represented in image form -->
  </facsimile>
  <text>
    <!-- the text, transcribed and marked up -->
  </text>
</TEI>
```
If many such documents are grouped together to form a corpus (rather than a collection), it may be useful to factor out the metadata they have in common:

```xml
<teiCorpus>
 <teiHeader>
 <!-- shared metadata -->
  </teiHeader>
 <TEI>
  <teiHeader>
   <!-- specific metadata -->
  </teiHeader>
  <text>
   <!-- ... -->
  </text>
 </TEI>
 <teiCorpus>
</teiCorpus>
```
What kinds of metadata?

For the *Punch Project* and for any other comparable project, we will need a place for such information as

- identification of the resource itself ("what is this thing?")
- statements of responsibility ("who did what when?")
- indication of source ("what was this derived from?")
- publication statement ("how is this item distributed and by whom?")
- declaration of encoding practice ("what do the codes we added mean?")

The TEI Header supports all these, and more.
The TEI Header

The TEI header was designed with two goals in mind

- needs of bibliographers and librarians trying to document ‘electronic books’
- needs of text analysts trying to document ‘coding practices’ within digital resources

On the one hand, the Librarian's header

- uses standard bibliographic concepts
- respects established mappings to other such records (e.g. MARC)
- has a preference for structured data over loose prose

On the other, Everyman's header

- Supports a (potentially) huge range of very miscellaneous information, organized in fairly ad hoc ways -
- Unpredictable combinations of narrowly encoded documentation systems and loose prose descriptions
The TEI header has four main components:

- `<fileDesc>` (file description) contains a full bibliographic description of an electronic file.
- `<encodingDesc>` (encoding description) documents the relationship between an electronic text and the source or sources from which it was derived.
- `<revisionDesc>` (revision description) summarizes the revision history for a file.
- `<profileDesc>` (text-profile description) provides a detailed description of non-bibliographic aspects of a text, specifically the languages and sublanguages used, the situation in which it was produced, the participants and their setting. (just about everything not covered in the other header elements)

Only `<fileDesc>` is required; the others are optional.
Simple TEI Header for Punch Project

<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>Punch, or the London Charivari, Vol. 147, July 1, 1914</title>
    </titleStmt>
    <publicationStmt>
      <idno type="gutenberg">24357</idno>
      <availability>
        <p>This text is freely available for re-use under US and UK law, consult your local legal restrictions if elsewhere.</p>
      </availability>
    </publicationStmt>
    <sourceDesc>
      <p>This text is a TEI version of a Project Gutenberg text originally located at <ptr target="http://www.gutenberg.org/dirs/2/4/3/5/24357/"/>. As per their license agreement we have removed all references to the PG trademark.</p>
    </sourceDesc>
  </fileDesc>
  <revisionDesc>
    <change when="2008-07-26T23:49:55.968+01:00"/>
  </revisionDesc>
</teiHeader>
Within the elements already introduced, TEI offers plenty of scope for mark-up of smaller components. For example:

- boundaries, such as page, column, or line breaks
- highlighting, emphasis and quotation
- editorial changes such as correction, normalization etc.
- names, numbers, dates, addresses...
- links and cross-references
- notes, annotation, indexing
- graphics
- bibliographic citations
- words and other analyses
Highlighting

By highlighting we mean any combination of typographic features (font, size, hue, etc.) which distinguishes the highlighted text from its surroundings. This may be for many reasons...

- to mark foreign, archaic, technical usages
- for emphasis when spoken
- to show something is not part of the text.. (e.g. cross references, titles, headings)
- or is attributed to some other agency inside or outside the text (e.g. direct speech, quotation)

TEI provides both a generic `<hi>` tag and a large number of specific ones...
A few highlighting examples

- `<hi>` (highlighted: reason unknown or unimportant)

  ```html
  <p>[The rest of this communication is omitted owing to considerations of space. <hi rend="sc">Ed</hi>.] </p>
  ```

- `<emph>` (emphasized)

  ```html
  <said>'E won't bite yer <emph>if you buy 'im</emph> guv'ner. </said>
  ```

- `<title>` and `<foreign>`:

  ```html
  <p>
    <foreign xml:lang="fr">À propos</foreign> of Oxford, it is a question whether that extremely amusing book <title>Verdant Green</title> is still much read by freshers.
  </p>
  ```

- `<distinct>` (linguistically marked)

  ```html
  But then I remind myself that the Russian ballet is nothing if not <distinct>bizarre</distinct>
  ```
Quotation marks can similarly be used to set off text for many reasons:

- `<q>` (used if the reason is unknown or unimportant)
- `<said>` (speech or thought)
- `<quote>` (attributed to an external source)
- `<mentioned>` and `<soCalled>` (nuances of narrative status)

```xml
<p>
  <said who="#Celia">I know a lovely tin of potted grouse, </said> said Celia, and she went off to cut some sandwiches.
</p>
```

```xml
<head>How to utilise the art of <soCalled>suggestion</soCalled></head>
<head>The Doctor, six down at the turn, <soCalled>suggests</soCalled> to his opponent that they are playing croquet, and wins by two and one. </head>
```
Note that these elements can nest within one another:

```xml
<p>The poet returned to his work. <said>
   <quote>In
       tooth and claw,</quote>
   </said> he muttered to himself,
<said>
   <quote>In tooth and claw.</quote>
</said>
</p>
```
Editorial intervention

As a simple example, consider: ‘Excuse me sir, but would you like to buy a nice little dawg?’ on page 6.

We can:

- use `<orig>` to show that "dawg" is what it says, even though this is a nonstandard spelling
- use `<reg>` to show that "dog" is an editorially-supplied regularisation of what it says
- or provide both within a `<choice>` element to say either is a valid encoding:

...a nice little

```
<choice>
  <orig>dawg</orig>
  <reg>dog</reg>
</choice>
```
Names of persons, places, things...

- `<name>` (a name in the text, contains a proper noun or noun phrase)
- `<rs>` (a general-purpose name or referencing string)
- `<title>` (any form of title)

The `@type` attribute is useful for categorizing these, and they both also have `@key`, `@ref`, and `@nymRef` attributes.
Examples of names

Using @type to distinguish personal from geographic names:

<p>The scene opens at a party given by
<name type="person">Potiphar</name> in
<name type="place">Venice</name>. </p>

Using @key and @ref to de-reference names:

<p>
<label>Business done.</label>—The Commons still harping on the Budget.
<name type="person"
    ref="http://en.wikipedia.org/wiki/Timothy_Michael_Healy">Tim Healy</name> enlivened proceedings by vigorous personal attack on <q>the most reckless and incapable <rs key="LLG">Chancellor of the Exchequer</rs> that ever sat on the Treasury Bench.</q>
<name key="LLG">Lloyd George's</name> retort courteous looked forward to with interest.
</p>
Dates

- `<date>` contains a date and time in any format
- For processing it is convenient to add a normalized version, using the `@when` attribute
- Uncertain dates and times, and ranges, can be indicated by other attributes: `@notBefore`, `@notAfter`, `@from@to`

```
<p>House of Commons, `<date when="1914-06-22">Monday, June 22, 1914</date>`. </p>
<p>`<date notAfter="1914-06-01" notBefore="1914-03-01">Sunday, a month ago, </date> was hot. </p>
```
Cross references

A cross reference is a link from one point in a text (the source) to another (the target). TEI provides generic elements `<ptr>` and `<ref>` for this purpose. If the linking text can be automatically generated use `<ptr>`; otherwise use `<ref>`. The source is the location of the `<ptr>` or `<ref>`; the target is specified by the `@target` attribute, in the form of a URI reference.

See `<ref target="#Section12">section 12 on page 34</ref>`.

See `<ptr target="#Section12"/>`. 
TEI provides special elements for bibliographic citations or references:

- `<bibl>` (loosely structured)
- `<biblStruct>` (standard bibliographic structure)
- `<listBibl>` (encloses a bibliography)

These are typically used in preparing bibliographies, or in footnotes. But even in Punch, there are examples.
Simple \texttt{<bibl>} Example

In Punch, bibliographic citations are usually associated with a quotation from another paper:
The \texttt{<cit>} element groups the two:

\begin{verbatim}
<cit>
  <quote>It was the time when Henry III. was batting with Simon de Montfort and his Barons.</quote>
  <bibl>
    <title>Straits Times.</title>
  </bibl>
</cit>
\end{verbatim}
Embedded notes

Notes, whether appearing in the original source, or added by an editor, can be marked using the `<note>` element. We might use this to add biographical details to the Punch transcriptions:

```
<p>By-the-by, it is denied that Sir <name rend="sc">Joseph Beecham</name> <note>Sir Joseph Beecham, 1st Baronet (8 June 1848 - 23 October 1916)...</note> was in any way responsible for the Government's "Pills for Earthquakes," by which it was hoped to avert the Irish crisis.</p>
```

`<note>` has attributes `@place` and `@resp`
Since we have several references to the same person, it might be better to put the notes elsewhere and point to them from the names:

```xml
<div type="notes">
  <note xml:id="BEECHJO">Sir Joseph Beecham, 1st Baronet (8 June 1848 - 23 October 1916) the eldest son of Thomas Beecham (1820-1907) played a large part in the growth and expansion of his father's medicinal pill business which he joined in 1866....</note>
</div>

<!-- other notes -->

<div type="snippets">
  <p>... Both Earl <name rend="sc">Beauchamp</name> and <name>Sir <ref target="#BEECHJO">Joseph Beecham</ref></name> appear in the recent Honours List.</p>
  <p>By-the-by, it is denied that Sir <name rend="sc" ref="#BEECHJO">Joseph Beecham</name> was in any way responsible...</p>
</div>

Could also use specialised <person> element, in this case. "Elsewhere" can be anywhere on the Internet...
Using TEI in a Philological Context: Reports from work in progress

1. Producing a critical edition
2. Encoding a dictionary
3. Compiling translation corpora
The TextGrid project

1. Project
2. Motivation
3. Collaborative design
4. TextGridLab: user toolbox
5. TextGridRep: archiving data
TextGrid: Toolbox
Using TEI in a Philological Context: Reports from work in progress

1. Producing a critical edition
2. Encoding a dictionary
[3. Compiling translation corpora]
Producing a critical edition

Jean Paul Arbeitsstelle, University of Würzburg: The new critical edition of Jean Paul's works
Volume 1: Hesperus, edited by Barbara Hunfeld

[www.jean-paul-portal.de]
Producing a critical edition

Jean Paul edition: texts published by the author
Hesperus versions 1795 - 1798 - 1819
[altogether more than 20,000 pages]
Producing a critical edition

Jean Paul, Nachlass, pages used for 'Hesperus'
Staatsbibliothek zu Berlin - Preußischer Kulturbesitz
Producing a critical edition

Jean Paul, Nachlass (Detail: “Holbeins gemalte Beine”
Staatsbibliothek zu Berlin - Preußischer Kulturbesitz
Producing a critical edition

Jean Paul, Nachlass, Detail: “Holbeins gemalte Beine”
Staatsbibliothek zu Berlin - Preußischer Kulturbesitz
Producing a critical edition

Test of typesetting, using TEI encoding & \TeX
Encoding a dictionary

Joachim Heinrich Campe:

„Wörterbuch der deutschen Sprache“, 5 volumes, Braunschweig 1807 – 1811
+ „Wörterbuch zur Erklärung und Verdeutschung der unserer Sprache aufgedrungenen fremden Ausdrücke. Ein Ergänzungsband zu Adelungs und Campes Wörterbüchern“, Braunschweig 1813

6000 pages, about 140.000 entries

[Testbed for the TextGrid Project]
Encoding a dictionary

Workflow

Typographical analysis

Image digitization
- resolution
- dynamic range
- Type of scanner
- image format
- quality control
- metadata
- archiving

Text capture
- page format
- line structure
- font information
- font size
- indentation
- special characters
- quality control

DC 2008 Berlin
Encoding a dictionary

Workflow

Typographical analysis: page
Encoding a dictionary

Workflow
Typographical analysis: line & fonts
Encoding a dictionary

Workflow

Guidelines for typing

<table>
<thead>
<tr>
<th>FRAKTUR</th>
<th>ANTIQUA</th>
<th>ANTIQUA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Großbuchstabe</strong></td>
<td><strong>Großbuchstabe</strong></td>
<td><strong>Großbuchstabe</strong></td>
</tr>
<tr>
<td><strong>Kleinbuchstabe</strong></td>
<td><strong>Kleinbuchstabe</strong></td>
<td><strong>Kleinbuchstabe</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ansage,</th>
<th>Fackeln,</th>
<th>Adverbium</th>
<th>adj. u. adv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A·······(Ansage)</td>
<td>a········(Fackeln)</td>
<td>A······(Adverbium)</td>
<td>a······(adj. u. adv)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Überdächen,</th>
<th>Album</th>
<th>Admirable</th>
</tr>
</thead>
<tbody>
<tr>
<td>A·</td>
<td>a·······(Überdächen)</td>
<td>A······(Album)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accidenz</th>
<th>Admiral,</th>
</tr>
</thead>
<tbody>
<tr>
<td>A·······(Accidenz)</td>
<td>a······(Admiral)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ärmel</th>
<th>Baumsäge</th>
<th>Pädagog</th>
</tr>
</thead>
<tbody>
<tr>
<td>A······(Ärmel)</td>
<td>a·······(Baumsäge)</td>
<td>a······(Pädagog)</td>
</tr>
</tbody>
</table>
## Encoding a dictionary

### Workflow

<table>
<thead>
<tr>
<th>Kürzelzeichen</th>
<th>Verwendungssachen</th>
<th>mit ihren Unicode-Werten (hex):</th>
</tr>
</thead>
<tbody>
<tr>
<td>Der Felsenfall</td>
<td>Kreis</td>
<td>Kreis-im-Kreis</td>
</tr>
<tr>
<td>△ Das Felsenherz</td>
<td>leeres Dreieck</td>
<td>Dreieck-mit-Innenkreis</td>
</tr>
<tr>
<td>Die Felsenbrust</td>
<td>△ 9651 UP-POINTING TRIANGLE, geometric shapes</td>
<td>△ 9708 UP-POINTING TRIANGLE WITH DOT, geometric shapes</td>
</tr>
<tr>
<td>Die Überart</td>
<td>Kreuz</td>
<td>Doppelkreuz</td>
</tr>
<tr>
<td>Δ Evangelisch</td>
<td>Doppelhalbkreis, seitlich-öffnen</td>
<td>Doppelhalbkreis, durchgestrichen</td>
</tr>
<tr>
<td>Wältigen</td>
<td>Stern</td>
<td>** Wältig</td>
</tr>
<tr>
<td>△ Überallmächtig</td>
<td>umgekehrtes Venuszeichen</td>
<td>△ 9793 EARTH</td>
</tr>
</tbody>
</table>

### Guidelines for typing
Die Aalraupe, <W+1>Mz. die -- n</W+1>, ein in süßem Wasser lebender Fisch,
der eine glatte Haut wie der Aal, aber einen dicken Kopf und Bauch hat, (<A>Gadus lota L.</A>). Jn N. D. die <W+1>Aalquappe</W+1>, in O. D. <W+1>der Raubaal, der Ruppe, Rutte, Rutten, Trusch, Trusche</W+1> und <W+1>Trüsche</W+1>. Die Aalraupe, <W+1>Mz. die -- n</W+1>, eine Reuse zum Aalfangen.

Die Aalschlange, <W+1>Mz. die -- en</W+1>, oder <W+1>der Meeraal</W+1>, eine Aalart, größer als der gemeine Aal, mit aschgrauem Rücken, weißem Bau=che, weißlich getüpfelten Seitenlinien und silberfarbigen Ringen um die großen Augen. Er hat zwei kurze Bartfaden und die Kiefern sind mit kleinen spitzigen Zähnen besetzt (<A>Muraena conger L.</A>).
Der Aar, des -- es, oder -- en, Mz. die -- e, oder -- en, die alte
Benennung aller großer Raubvögel, besonders aber des Adlers, die
noch in N. D. üblich ist und bei Dichtern vorkommt.

Ein kühner Aar theilt mit gewalt'gen Schwingen
Die Lüfte, -- -- -- -- -- Schreiber.
Bald werdet ihr im Meer der Haien, am Gestade
Der Aaren Beute sein. -- Ramler.

Der Aaron, eine Pflanze, s. Arum.
Encoding a dictionary

Workflow

Lexicographical analysis

Entry structures

Example: Der Aar

* Der Aar, des — es, oder — en, M. die — e, oder — en, die alte
  Benennung aller grosser Haubögel, besonders aber des Ablers, die
  noch in R. D. üblich ist und bei Dichtern vorkommt.

  Ein führer Aar theilt mit gewalt'gen Schwingen
  Die Luste, — — — — Schreiber.
  Bald werdet ihr im Meer der Haisen, am Gestad
  Der Aaren Beute sein. — Kamler.
Encoding a dictionary

Workflow

Lexicographical analysis

Entry structures: marker + keyword

Example: Der Aar

Der Aar, des—es, oder—en, Ms. die—e; oder—en, die alte Benennung aller großer Haubvögel, besonders aber des Ablers, die noch in R. D. üblich ist und bei Dichtern vorkommt.

Ein füßer Aar theilt mit gewaltgen Schwingen
Die Küste, — — — — — Schreiber.
Bald werdet ihr im Meer der Hainen, am Gestade
Der Haren Beute sein. — Ramler.
Encoding a dictionary

Workflow

Lexicographical analysis

Entry structures: grammatical forms

Example: Der Aar
Encoding a dictionary

Workflow

Lexicographical analysis

Entry structures: definition

Example: Der Aar
Encoding a dictionary

Workflow

Lexicographical analysis

Entry structures: context examples
(literary quotations)

Example: Der Aar

* Der Aar, des – es, oder – en, Ms. die – e; oder – en, die alte

Benennung aller großer Naußvögel, besonders aber des Adler's, die
noch in N. D. üblich ist und bei Dichtern vorkommt.

Ein fähn Ein Aar theilt mit gewaltigen Schwings
Die Lust, — — — — Schreiber.
Bald werdet ihr im Meer der Hainen, am Gestade
Der Aaren Beute sein. — Ramler.
Encoding a dictionary

Workflow

TEI Encoding structures

<entry>
  <form>
    <form> ... </form>
  </form>
  <sense>
    <def> ... </def>
    <note> ... </note>
    <etym> ... </etym>
    <cit> ... </cit>
    <re> ... </re>
  </sense>
</entry>
Encoding a dictionary

Workflow

Guidelines for the project encoding

Zitate werden durch <quote>-Tags ausgezeichnet. Die zugehörige Quelle wird in <bibl> </bibl> angegeben. Die Einheit aus <quote> und <bibl> wird durch <cit type="quote">... </cit> geklammert.

Bsp: <cite>

<quote> Ein kühner Aar theilt mit gewaltigen Schwingen
Die Lüfte, — — — — — Schreiber. </quote>

</cit>

<cit>

<author n="#Schreiber"> Schreiber </author>

</cit>

<cit>

<author n="#Name"> Name </author>

</cit>

Quelle

<author n="#Name"> Name </author>

<title> Titel </title>

in <bibl> </bibl>

</cit>

<author> bezeichnet den Autor eines Zitates. Sein Name ist gesperrt gedruckt. Das Attribut n="#Name" wird vergeben, um immer eine eindeutige Zuordnung von Zitat und Autor zu gewährleisten, z.B. auch dann, wenn der Name abgekürzt ist oder der Verweis „derselbe“ auf das Vorausgegangene verweist.

DC 2008 Berlin
Encoding a dictionary

Workflow

Defining a TEI Relax NG Schema and customize it according to the Guidelines using Roma

Programming a converter to produce valid TEI files for every entry of the dictionary

Using the TEI header do keep track of the workflow
Encoding a dictionary
Converter output: valid TEI file

<?xml version="1.0" encoding="UTF-8" ?>
<?oxygen RNGSchema="../dictionary_schemas/campe_dictionary.rnc" type="compact"?>
- <TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="de">
- <teiHeader xmlns="">
- <fileDesc>
- <titleStmt>
<title>Joachim Heinrich Campe, Wörterbuch der Deutschen Sprache, Band 1 mit TEI-Auszeichnung</title>
<author>Kompetenzzentrum für EDV-Philologie, Universität Würzburg</author>
</titleStmt>
- <publicationStmt>
- <availability>
<p>für die Kodierungs- und Korrekturphase</p>
</availability>
<authority>Kompetenzzentrum für EDV-Philologie, Universität Würzburg</authority>
</publicationStmt>
- <sourceDesc>
<p>Wörterbucheintrag Aar</p>
- <biblStruct xml:lang="de" type="dictionary">
- <monogr>
<author>Joachim Heinrich Campe</author>
<title>Wörterbuch der Deutschen Sprache</title>
- <imprint>
<pubPlace>Braunschweig</pubPlace>
<publisher>In der Schulbuchhandlung.</publisher>
<date when="1807">1807</date>
</imprint>
Encoding a dictionary
Workflow
Converter output: valid TEI file 2

</fileDesc>
- <encodingDesc>
- <editorialDecl>
- <normalization>
  <p>Formatterfassung der Lemmata: Alle Lemmata (Stichwort ggf. mit Artikel) sind ausnahmlos in einem ca. 0,5 Punkt größeren Font gesetzt; dieses wird nur an dieser Stelle notiert. Der negative Einzug (Linksausrückung) wird nicht kodiert</p>
  <p>Typographische Besonderheiten: Alle übrigen typographischen Veränderungen sind mit dem Element <hi/> und spezifizierendem Attribut kodiert.</p>
</normalization>
</editorialDecl>
</encodingDesc>
- <revisionDesc>
  <change who="#CG">2008-09-22 - automatisch generiert durch CPM</change>
  <change who="#--">Überarbeitung</change>
  <change who="#--">Textkontrolle</change>
  <change who="#--">Übernahme in Bandkodierung</change>
</revisionDesc>
</teiHeader>
Encoding a dictionary

Workflow

Converter output: valid TEI file 3

```xml
<text xmlns="">
  <body>
    <entry xml:id="cwds1_00045_aar">
      <form>
        <form type="lemma">
          <usg type="time"> <g>*</g> </usg>
          <gramGrp> <pos value="noun" /> <gen value="m" /> </gramGrp>
          <form type="determiner"> <orth>Der</orth> </form>
          <form type="headword"> <orth>Aar</orth> <c type="$,">,</c> </form>
        </form>
        <form type="inflected">
          <gramGrp> <case value="genitive" /> <number value="singular" /> </gramGrp>
          <form type="determiner"> <orth>des</orth> </form>
          <form type="headword"> <orth> <oVar> <oRef /> -- es </oVar> </orth> </form>
        </form>
        <form type="inflected">
          <gramGrp> <case value="nominative" /> <number value="plural" >Mz.</number> </gramGrp>
          <form type="determiner"> <orth>die</orth> </form>
          <form type="headword"> <orth> <oVar> <oRef /> -- e </oVar> </orth> </form>
        </form>
        <sense> oder <c type="$,">--</c> <hi rend="large">en</hi> <c type="$,">,</c>
        die alte <lb n="0002.036" >Benennung aller großer Raubvögel</lb> <c type="$,">,</c>
        besonders aber des Adlers <c type="$,">,</c> die <lb n="0002.037" >noch</lb> <usg type="geo"> in N <c type="$,">.</c> </usg>
        üblich ist und <usg type="style">bei Dichtern</usg> vorkommt <c type="$,">.</c>
      </form>
    </entry>
  </body>
</text>
```
Encoding a dictionary

Workflow

Correcting the encoding manually

Final control of the encoding

Compiling larger units

Transform into baseline encoding for efficient search
Encoding a dictionary
Workflow
Baseline encoding

<revisionDesc>
<change who="#CG">Neuerstellung</change>
<change who="#MB">20080430 Textkontrolle</change>
</revisionDesc> </teiHeader>

<text> <body> <div n="a" type="letter" xml:id="letter-a"> <lb/>

<entry xml:id="cwds1_000045_Aar">
<form type="lemma"> <gramGrp> <pos value="noun"/> <gen value="m"/> </gramGrp>
<usg type="time">*</usg> Der <form type="headword">Aar</form>, des -es, oder -en,
Mz. die -e, oder -en,<form type="headword">Aar</form>, des -es, oder -en,
</entry>

<sense>
<def>die alte<lb/>Benennung aller großer Raubvögel, besonders aber des
Adlers,</def> die<lb/>noch <usg type="geo">in N. D.</usg> üblich ist und <usg
type="style">bei Dichtern</usg> vorkommt.<lb/>
<cit type="example"> <cit type="verse">
<quote>Ein kühner Aar theilt mit gewalt'gen Schwingen<lb/>Die Lüfte, - - - -  -</quote>
</cit>
<bibl><author n="#Schreiber">Schreiber.</author></bibl> <cit type="verse">
<quote>Bald werdet ihr im Meer der Haien, am Gestade<lb/>Der Aaren
Beute sein. -</quote> <bibl><author n="#Ramlr">Ramlr.</author></bibl> <cit type="verse">
</cite>
</body>
</text>
Baseline Encoding for the compilation of dictionaries in TEI P5

Mirjam Blüm, Stefan Büdelbender, Christian Graiger, Werner Wegstein
June 7* 2008

Abstract

This paper is concerned with the structuring of dictionary entries and gives an overview of the XML elements and element classes of the modules core, dictionaries, textstructure and gajj of TEI P5, which were specified for the baseline encoding.

Open Issues

- Treatment of obvious errors
- Ontology for attribute values (e.g. for quotations)
- Placement of meta information within an element <entry>

Contents

1 Baseline elements of TEI P5
   1.1 XML elements ............................................ 2
   1.2 XML-Attribute ............................................. 2
   1.3 Examples .................................................. 4
       1.3.1 Dictionary entry “Der Aar” .......................... 5
       1.3.2 Dictionary entry “Abtinden” ......................... 7

2 Hierarchy

A Description of the schema

1.1 XML elements

Table 1 shows those XML elements of the module dictionaries, which are included in the baseline encoding. The marker optional refers to elements, which are used only if demanded by the structure of the dictionary entry. If an entry contains a definition, for example, the appropriate tag <def> has to be used. If not, it is omitted.

<table>
<thead>
<tr>
<th>Element</th>
<th>Function</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;entry&gt;</td>
<td>contains a well-structured dictionary entry</td>
<td>required</td>
</tr>
<tr>
<td>&lt;form&gt;</td>
<td>groups all the information on the written and spoken forms of one headword</td>
<td>required</td>
</tr>
<tr>
<td>&lt;sense&gt;</td>
<td>groups together all information relating to one word sense in a dictionary entry, for example definitions, examples, translation equivalents, etymology ...</td>
<td>required</td>
</tr>
<tr>
<td>&lt;gramInfo&gt;</td>
<td>groups morpho-syntactic information about a lexical item, groups the elements &lt;pos&gt;, &lt;gen&gt; and &lt;sub&gt;</td>
<td>required</td>
</tr>
<tr>
<td>&lt;pos&gt;</td>
<td>indicates the part of speech</td>
<td>required</td>
</tr>
<tr>
<td>&lt;sub&gt;</td>
<td>contains subclassification information (“transitive”/“intransitive”)</td>
<td>optional</td>
</tr>
<tr>
<td>&lt;gen&gt;</td>
<td>identifies the morphological gender</td>
<td>optional</td>
</tr>
<tr>
<td>&lt;def&gt;</td>
<td>contains definition text in a dictionary entry</td>
<td>optional</td>
</tr>
<tr>
<td>&lt;etym&gt;</td>
<td>encloses etymological information</td>
<td>optional</td>
</tr>
<tr>
<td>&lt;usage&gt;</td>
<td>contains usage information</td>
<td>optional</td>
</tr>
<tr>
<td>&lt;ref&gt;</td>
<td>contains a dictionary entry for a lexical item related to the headword, such as a compound phrase or derived form, embedded inside a larger entry</td>
<td>optional</td>
</tr>
<tr>
<td>&lt;superEntry&gt;</td>
<td>groups successive entries for a set of homographs</td>
<td>optional</td>
</tr>
<tr>
<td>&lt;xref&gt;</td>
<td>cross-reference entries: contains a phrase, sentence, or icon referring the reader to some other location in this or another text</td>
<td>optional</td>
</tr>
</tbody>
</table>

The elements <entry>, <form> and <sense> are necessary preconditions for search enquiries and thus represent the obligatory components of the baseline encoding. The element <form> groups all the information on the lemma, the element <sense> all information relating to the word sense. This applies also to entries, which consist of
Encoding a dictionary

Wer ein Wörterbuch schreibt, baut recht eigentlich am Wege und setzt sein Werk dem Tadel aus.

Matthias von Lexer
Encoding a dictionary

Thank you.
Overview — what should be remembered in this section

- Concepts related to the TEI architecture
  - Modules, classes
  - ODD, Roma

- Bibliographical elements in the TEI
  - <bibl>
  - <biblStruct>
Following the TEI spirit

**Conformance** to the TEI means:

- Sharing a common text encoding culture
- Sharing the same vocabulary (when applicable)
- Allowing user autonomy in defining modifications (extensions, customization), but sharing the mechanisms to do so

The TEI gives you a lot of help in following these rules.
Colourful TEI: the bibliographic elements

- `<bibl>` freely annotated in-line element
- `<biblStruct>` highly structured bibliographical entry
- `<biblFull>` complex structure mimicking the TEI header
<bibl> - examples

<bibl>Blain, Clements and Grundy: Feminist Companion to Literature in English (Yale, 1990)</bibl>

<bibl><title>Dictionnaire des difficultés de la langue française</title> (<author>V. Thomas</author>, <publisher>Larousse</publisher>)</bibl>. </item>

<bibl><idno type="ISO">ISO/IEC 2382 (all parts)</idno>, <title type="introductory">Information technology</title><title type="main">Vocabulary</title></bibl>
<biblStruct> - example

[Export from WALS online - http://wals.info]

<biblStruct type="incollection">
  <analytic>
    <author>
      <forename>Pliny Earle</forename>
      <surname>Goddard</surname>
    </author>
    <title type="main" level="a">Athapascan (Hupa)</title>
  </analytic>
  <monogr>
    <editor>Boas, Franz</editor>
    <title type="main" level="m">Handbook of American Indian Languages</title>
    <imprint>
      <pubPlace>Washington, D. C.</pubPlace>
      <publisher>Government Printing Office</publisher>
      <date>1911</date>
      <biblScope type="pp">85-158</biblScope>
    </imprint>
  </monogr>
</biblStruct>
Defining a TEI Schema

- A schema helps you know a document is valid in addition to being well-formed
- A TEI schema is a combination of TEI modules, optionally including customizations of the elements/attributes/classes that they contain
- This schema is defined in an application-independent manner with a TEI ODD (One Document Does it all) file which allows for:
  - creation of schemas using DTD, RELAX NG or W3C Schema languages
  - internationalized documentation which reflects your customization of the TEI
Important ODD concepts

The TEI's literate programming with ODD (One Document Does it all) provides:

- Schema specification
- User oriented documentation
- Modularity: all specifications pertaining to a coherent sub-domain of the TEI
- Classes: identifying shared behaviours or semantics
- Extensibility: a consequence of the above mechanisms
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 often are structurally or semantically related)

- Classes may contain other classes
  - Elements inherit the properties from any classes of which they are members
Attribute Classes

- Attribute classes are given (usually adjectival) names beginning with att.;
  - e.g. members of the `att.canonical` class get a `key` attribute rather than have them define it individually
  - If another element needs a `key` attribute then the easiest way to provide it is to add it to the `att.naming` class

- Classes can be grouped together into a super classes
att.global

- All elements are members of att.global; this provides, among others:
  - xml:id
    - a unique identifier
  - xml:lang
    - the language of the element content
  - n
    - a number or name for an element
  - rend
    - how the element in question was rendered or presented in the source text.
Attribute classes — E.g. `<biblStruct>`/att.typed

---

**<biblStruct>** (structured bibliographic citation) contains a structured bibliographic citation, in which only bibliographic sub-elements appear and in a specified order. 3.11.1 Elements of Bibliographic References 2.2.7 The Source Description 15.3.2 Declarable Elements

<table>
<thead>
<tr>
<th>Module</th>
<th>core — 3 Elements Available in All TEI Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attributes</td>
<td>att.declarable (@default) att.typed (@type, @subtype)</td>
</tr>
<tr>
<td>Used by</td>
<td>model.bibLike</td>
</tr>
<tr>
<td>May contain</td>
<td>core: analytic monogr note relatedItem series</td>
</tr>
<tr>
<td></td>
<td>header: idno</td>
</tr>
<tr>
<td></td>
<td>textcrit: witDetail</td>
</tr>
</tbody>
</table>

**Declaration**

```xml
element biblStruct
{
    att.global.attributes,
    att.declarable.attributes,
    att.typed.attributes,

    { analytic*,
      ( monogr, series* )*,
      ( model.noteLike | idno | relatedItem )*  
    }
}
```
att.typed — definition

att.typed provides attributes which can be used to classify or subclassify elements in any way. 1.3.1 Attribute Classes

Module | tei — 1 The TEI Infrastructure
Attributes | In addition to global attributes
@type | characterizes the element in some sense, using any convenient classification scheme or typology.
@subtype | provides additional information

```xml
<biblStruct type="incollection">
   <analytic>
      <author>…</author>
      <title type="main" level="a">Athapascan (Hupa)</title>
   </analytic>
   <monogr>
      <editor>…</editor>
      <title type="main" level="m">Handbook of …</title>
      <imprint>
         <biblScope type="pp">85-158</biblScope>
      </imprint>
   </monogr>
</biblStruct>
```
Model Classes

- Model classes contain groups of elements allowed in the same place.
  - e.g. if you are adding an element which is wanted wherever the `<bibl>` is allowed, add it to the model.biblLike class
- Model classes are usually named with a Like or Part suffix:
  - model.divLike: structural class grouping elements for divisions
  - model.divPart: structural class grouping elements used inside divisions
  - model.nameLike: semantic class grouping name elements
  - model.persNamePart: semantic sub-class grouping elements that are part of a personal name
Model classes — E.g. `<bibl>/model.biblPart`

```plaintext
element bibl{
    __att.global.attributes,
    __att.declarable.attributes,
    __att.typed.attributes,
    (text
        |    model.gLike
        |    model.highlighted
        |    model.pPart.data
        |    model.pPart.edit
        |    model.segLike
        |    model.ptrLike
        |    model.biblPart
        |    model.global
    )*
}
```
### model.biblPart

The `model.biblPart` element groups elements which represent components of a bibliographic description. It is used in the context of the bibliographic citation and reference module.

#### 3.11 Bibliographic Citations and References

<table>
<thead>
<tr>
<th>Module</th>
<th>tei —</th>
<th>1 The TEI Infrastructure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Used by</td>
<td>bibl</td>
<td></td>
</tr>
<tr>
<td>Members</td>
<td></td>
<td>model.imprintPart, biblScope, distributor, pubPlace, publisher, model.respLike, author, editor, respStmt, edition, extent, idno, meeting, msIdentifier, relatedItem, series</td>
</tr>
</tbody>
</table>

[English] [Deutsch] [Español] [Italiano] [Français] [日本語] [中文]
The TEI ODD in practice

The TEI Guidelines, its schema, and its schema fragments, are all produced from a single XML resource containing:

1. Descriptive prose (lots of it)
2. Examples of usage (plenty)
3. Formal declarations for components of the TEI Abstract Model:
   - elements and attributes
   - Modules
   - classes and macros
Possibilities of customizing the TEI

The TEI has over 20 modules. A working project will:

- Choose the modules they need
- Probably narrow the set of elements within a module
- Probably add local datatype constraints
- Possibly add new elements
- Possibly localize the names of elements
Real life TEI customization

We aim to support a range of interactions with the TEI:

**Easy TEI**
- Simple access to the TEI through Roma

**Subsetting the TEI**
- Making the TEI even easier to use

**Enlarging the application profile**
- Using modules

**Modifying the TEI objects**
- First insights into extensibility

**Behind the scene - ODD**
- Starting to use the actual specification language
Roma

- The TEI knows you don't want to necessarily have to write TEI code in order to customize the TEI. So it has provided Roma, which is a command-line script, and corresponding web front-end to help you do this.

- The people behind Roma are:
  - Arno Mittelbach
    - Initial programming
  - Sebastian Rahtz
    - Maintenance and frequent improvements
  - Ioan Bernevig
    - A 'Sanity Checker' addition
Quick and simple access to the TEI

- Go to Roma at http://tei.oucs.ox.ac.uk/Roma/
- Generate a schema [Schema]
- Make a trial with the editor, creating a simple document

- That‘s all!
Scientific bibliographies

- additional needs
  - precise identification of persons and their affiliations
  - e.g. organisation, location, addresses
Roma: Start

TEI Roma: generating validators for the TEI

These pages will help you design your own TEI validator, as a DTD, RELAXNG or W3C Schema.

Create a new or upload existing customization

- Build schema (Create a new customisation by adding elements and modules to the smallest recommended schema)
- Reduce schema (Create a new customization by removing elements and modules from the largest possible schema)

- Create customization from template
- Open existing customization

Submit

Search TEI Guide

- TEI Absolutely Bare
- TEI Absolutely Bare
- TEI Lite
- TEI for Linguistic Corpora
- TEI for Manuscript Description
- TEI with Drama
- TEI for Speech Representation
- TEI for authoring ODD
- TEI with SVG
- TEI with MathML
- TEI with XInclude (experimental)
- TEI for Dictionaries (experimental)
Roma: Schema Select

Time to give you a schema

Creating a schema

Which format do you prefer?

- Relax NG schema (compact syntax)
- Relax NG schema (compact syntax)
- Relax NG schema (XML syntax)
- W3C schema
- DTD

Search TEI database

Roma was written by Arno Mittelbach and is maintained by Sebastian Rahtz. Please direct queries to the TEI at Oxford project.
This is Roma version 2.7, last updated 2006-10-20.
Roma: Generate Doc

Roma: generating validators for the TEI

Documentation?

Getting some nice documentation

Which output would you prefer?

- [ ] html
- [ ] pdf
- [ ] TEI Lite
- [ ] Tei

Search TEI database

Roma was written by Ann Mittelbach and is maintained by Sebastian Rohr. Please direct queries to the TEI@Oxford project.
This is Roma version 2.7, last updated 2006-10-29.
Subsetting the TEI

Suppose you now feel you want to use some more of the TEI, but not all of it

- Go to Roma...
- Look at [Modules]
- Explore default modules by pointing to main elements (by order of interest). You can throw away most things, but
  - In textstructure, you should really keep <TEI>, <text>, <body> and <div>
  - In core, most people need <p>, <q>, <list>, <pb/> and <head>
  - From header, keep everything unless you really understand the details
- Start checking out elements
- Make editorial choices (numbered vs. unnumbered divs)
Roma: Customize

Roma: generating validators for the TEI

Set your parameters

<table>
<thead>
<tr>
<th>Set your parameters</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>My TEI Extension</td>
</tr>
<tr>
<td>Filename</td>
<td>myTei</td>
</tr>
<tr>
<td>Prefix for TEI</td>
<td></td>
</tr>
<tr>
<td>pattern names</td>
<td></td>
</tr>
<tr>
<td>in schema</td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td>English, Deutsch, Français, Russian, Svenska, 日本語, 中文</td>
</tr>
<tr>
<td>Author name</td>
<td>generated by Roma 2.7</td>
</tr>
<tr>
<td>Description</td>
<td>My TEI Customization starts with modules tei, core, header, and textstructure</td>
</tr>
</tbody>
</table>
Roma: Modules

Roma: generating validators for the TEI

Modules

<table>
<thead>
<tr>
<th>Module name</th>
<th>A short description</th>
<th>Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>add analysis</td>
<td>Simple analytic mechanisms</td>
<td></td>
</tr>
<tr>
<td>add certainty</td>
<td>Certainty and uncertainty</td>
<td></td>
</tr>
<tr>
<td>add core</td>
<td>Elements common to all TEI documents</td>
<td></td>
</tr>
<tr>
<td>add corpus</td>
<td>Header extensions for corpus texts</td>
<td></td>
</tr>
<tr>
<td>add declarefs</td>
<td>Feature system declarations</td>
<td></td>
</tr>
<tr>
<td>add dictionaries</td>
<td>Printed dictionaries</td>
<td></td>
</tr>
<tr>
<td>add drama</td>
<td>Performance texts</td>
<td></td>
</tr>
<tr>
<td>add figures</td>
<td>Tables, formulae, and figures</td>
<td></td>
</tr>
<tr>
<td>add gaij</td>
<td>Character and glyph documentation</td>
<td></td>
</tr>
<tr>
<td>add header</td>
<td>The TEI Header</td>
<td></td>
</tr>
<tr>
<td>add iso-fs</td>
<td>Feature structures</td>
<td></td>
</tr>
<tr>
<td>add linking</td>
<td>Linking, segmentation and alignment</td>
<td></td>
</tr>
<tr>
<td>add msdescription</td>
<td>Manuscript Description</td>
<td></td>
</tr>
<tr>
<td>add namesdates</td>
<td>Names and dates</td>
<td></td>
</tr>
<tr>
<td>add nets</td>
<td>Graphs, networks and trees</td>
<td></td>
</tr>
</tbody>
</table>

List of selected Modules

- remove core
- remove tei
- remove header
- remove htmlstructure

26.09.2008
# Roma: Change Module

Roma: generating validators for the TEI

## Change module

<table>
<thead>
<tr>
<th>Include</th>
<th>Exclude</th>
<th>Tag name</th>
<th>Description</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>abbr</td>
<td></td>
<td>abbr</td>
<td>contains an abbreviation of any sort.</td>
<td>Change attributes</td>
</tr>
<tr>
<td>add</td>
<td></td>
<td>add</td>
<td>contains letters, words, or phrases inserted in the text by an author, scribe, annotator, or corrector.</td>
<td>Change attributes</td>
</tr>
<tr>
<td>addrLine</td>
<td></td>
<td>addrLine</td>
<td>contains one line of a postal or other address.</td>
<td>Change attributes</td>
</tr>
<tr>
<td>address</td>
<td></td>
<td>address</td>
<td>contains a postal or other address, for example of a publisher, an organization, or an individual.</td>
<td>Change attributes</td>
</tr>
<tr>
<td>altident</td>
<td></td>
<td>altident</td>
<td>supplies the recommended XML name for an element, class, attribute, etc. in some language.</td>
<td>Change attributes</td>
</tr>
<tr>
<td>analytic</td>
<td></td>
<td>analytic</td>
<td>contains bibliographic elements describing an item (e.g. an article or poem) published within a monograph or journal and not as an independent publication.</td>
<td>Change attributes</td>
</tr>
<tr>
<td>author</td>
<td></td>
<td>author</td>
<td>in a bibliographic reference, contains the name of the author(s), personal or corporate, of a work; the primary statement of responsibility for any bibliographic item.</td>
<td>Change attributes</td>
</tr>
<tr>
<td>bibl</td>
<td></td>
<td>bibl</td>
<td>contains a loosely-structured bibliographic citation of</td>
<td>Change attributes</td>
</tr>
</tbody>
</table>
## Roma: Show Class

### Class `att.divLike`

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>type</code></td>
<td>specifies a name conventionally used for this level of subdivision, e.g. act, volume, book, section, canto, etc.</td>
</tr>
<tr>
<td><code>Status</code></td>
<td>Recommended</td>
</tr>
<tr>
<td><code>Datatype</code></td>
<td><code>data.enumerated</code></td>
</tr>
<tr>
<td><code>Values</code></td>
<td>any string of characters</td>
</tr>
<tr>
<td><code>org</code></td>
<td>specifies how the content of the division is organized.</td>
</tr>
<tr>
<td><code>Status</code></td>
<td>Optional</td>
</tr>
<tr>
<td><code>Default</code></td>
<td>uniform</td>
</tr>
<tr>
<td>Legal values are:</td>
<td></td>
</tr>
<tr>
<td><code>composite</code></td>
<td>(composite content: i.e. no claim is made about the sequence in which the immediate contents of this division are to be processed, or their inter-relationships.)</td>
</tr>
<tr>
<td><code>uniform</code></td>
<td>(uniform content: i.e. the immediate contents of this element are regarded as forming a logical unit, to be processed in sequence.)</td>
</tr>
<tr>
<td>Sample indicates whether this division is a sample of the original source and if so, from which part.</td>
<td></td>
</tr>
<tr>
<td><code>Status</code></td>
<td>Optional</td>
</tr>
<tr>
<td><code>Default</code></td>
<td>complete</td>
</tr>
<tr>
<td>Legal values are:</td>
<td></td>
</tr>
<tr>
<td><code>initial</code></td>
<td>(division lacks material present at end in source.)</td>
</tr>
<tr>
<td><code>medial</code></td>
<td>(division lacks material at start and end.)</td>
</tr>
</tbody>
</table>
Roma: Change Class Attributes

Roma: generating validators for the TEI

Added Attributes

List of attributes

<table>
<thead>
<tr>
<th>Change attribute</th>
<th>Include</th>
<th>Exclude</th>
<th>Tag name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>type</td>
<td></td>
<td></td>
<td>type</td>
<td>specifies a name conventionally used for this level of subdivision, e.g. act, volume, book, section, canto, etc.</td>
</tr>
<tr>
<td>org</td>
<td></td>
<td></td>
<td>org</td>
<td>specifies how the content of the division is organized.</td>
</tr>
<tr>
<td>sample</td>
<td></td>
<td></td>
<td>sample</td>
<td>indicates whether this division is a sample of the original source and if so, from which part.</td>
</tr>
<tr>
<td>part</td>
<td></td>
<td></td>
<td>part</td>
<td>specifies whether or not the division is fragmented by some other structural element, for example a speech which is divided between two or more verse stanzas.</td>
</tr>
</tbody>
</table>

Submit Query

Search TEI database

Roma was written by Arno Mittelbach and is maintained by Sebastian Rahtz. Please direct queries to the TEI at Oxford project. This is Roma version 2.7, last updated 2006-10-29
Adding TEI objects

You can add your own elements and attributes. But

- make very sure you are not just making something which is syntactic sugar for an existing TEI concept
- do not rename existing elements - you can do that directly in ODD
- if you want facilities from a very different field of discourse, such as maths or vector graphics, use the existing standards in that area
- consider interoperability
Roma: Add Element

Roma: generating validators for the TEI

Add Element

Defining a new element

Name

Description

Model classes

- model.addPart
- model.bibIPart
- model.catDescPart
- model.complexVal
- model.divPart
- model.divPart.verse
- model.editorialDecidPart
- model.EntryParts
- model.formPart
- model.gLike
- model.global.meta
- model.graphicLike
- model.imprintPart
- model.assertLike
- model.blockLike
- model.choicePart
- model.dateLike
- model.divPart.spoken
- model.divPartWrapper
- model.encodingPart
- model.entryParts.top
- model.frontPart
- model.global
- model.glossLike
- model.headerPart
- model.inter
- model.bibILike
- model.castIItemPart
- model.common
- model.datePart
- model.divPart.stage
- model.divPartWrapper.bottom
- model.entryLike
- model.featureVal
- model.frontPart.drama
- model.global.edit
- model.gramPart
- model.hLike
- model.iLike
- model.iLike
Roma: language Selection

Roma: generating validators for the TEI

Choose a different language

Do you want the output schema to use a different language?

Language for element and attribute names
- English
- Deutsch
- Español
- Français
- 日本語
- 中文

Language for documentation
- English
- Deutsch
- Español
- Français
- 日本語
- 中文

Submit Query

Search TEI database

Roma was written by Arno Mittelbach and is maintained by Sebastian Rahtz. Please direct queries to the TEI@Oxford project. This is Roma version 2.7, last updated 2006-10-29.
Roma: In French (sorry…)

TEI Roma: generating validators for the TEI

Change module

<table>
<thead>
<tr>
<th>Inclure</th>
<th>Exclure</th>
<th>Nom de balise</th>
<th>Description</th>
<th>Attributs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>authority</td>
<td>L’élément Responsable de la mise à disposition authority fournit le nom d’une personne ou d’une institution responsable de la mise à disposition d’un fichier électronique, autre qu’un éditeur ou un distributeur.</td>
<td>Changer les attributs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>availability</td>
<td>L’élément Disponibilité availability fournit des informations sur la disponibilité du texte, par exemple toutes restrictions sur son usage ou sa diffusion, son statut de copyright, etc.</td>
<td>Changer les attributs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>broadcast</td>
<td>L’élément Diffusion broadcast décrit une émission utilisée comme source de la parole.</td>
<td>Changer les attributs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cRefPattern</td>
<td>L’élément Adressage de fragments fragmentPattern spécifie un modèle d’expression et des règles de remplacement pour transformer une référence canonique en URI.</td>
<td>Changer les attributs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>catDesc</td>
<td>L’élément Description de catégorie catDesc décrit une certaine catégorie à l'intérieur d'une taxinomie ou d'une typologie de texte, sous forme d'une brève description en prose suivie ou en termes de paramètres contextuels.</td>
<td>Changer les attributs</td>
</tr>
</tbody>
</table>
Under the hood

TEI customizations are themselves expressed in TEI XML, using elements from the *tagdocs* module.

For example:

```xml
<schemaSpec ident="myTEIlite">
  <desc>This is TEI Lite with simplified heads</desc>
  <moduleRef key="tei"/>
  <moduleRef key="core"/>
  <moduleRef key="textstructure"/>
  <moduleRef key="header"/>
  <moduleRef key="linking"/>
  <elementSpec ident="head" mode="change">
    <content>
      <rng:text/>
    </content>
  </elementSpec>
</schemaSpec>
```

produces something like TEI Lite, with a slight change
ODD processors

- The TEI maintains a library of XSLT scripts that can generate
  - The TEI Guidelines in canonical TEI XML format
  - The Guidelines in HTML or PDF
  - RELAXNG, DTD, or W3C schema fragments
- The same library is used by the customization layer to generate
  - project-specific documentation
  - project-specific schemas
  - translations into other (human) languages
The TEI abstract model

- The TEI abstract model sees a markup scheme (a schema) as consisting of a number of discrete modules, which can be combined more or less as required.
- A *schema* is made by combining references to *modules* and optional element over-rides or additions.
- Each *element* declares the module it belongs to: elements cannot appear in more than one module.
- Each module extends the range of elements and attributes available by adding new members to existing *classes of elements*, or by defining new classes.
Expression of TEI content models

TEI elements have to be defined using some language notation; choices include:

1. using XML DTD language (as in older versions of the TEI)
2. using W3C schema language
3. using the RELAXNG schema language
4. inventing an entirely new abstract language for later transformation to specific schema language

We chose a combination of 3 and 4 - using our abstract language, but switching to RELAXNG for content modelling.
What does an ODD look like?

```
<elementSpec module="spoken" ident="pause">
  <classes>
    <memberOf key="model.divPart.spoken"/>
    <memberOf key="att.timed"/>
    <memberOf key="att.typed"/>
  </classes>
  <content>
    <rng:empty/>
  </content>
  <attList>
    <attDef ident="who" usage="opt">
      <gloss>A unique identifier</gloss>
      <desc>supplies the identifier of the person or group pausing. Its value is the identifier of a
person</gi> or <gi>persGrp</gi> element in the TEI header.
    </desc>
    <datatype>
      <rng:ref name="data.pointer"/>
    </datatype>
  </attList>
  <desc>a pause either between or within utterances.</desc>
</elementSpec>
```
... from which we generate

element pause { pause.content, pause.attributes }
  pause.content = empty
  pause.attributes =
  att.global.attributes,
  att.timed.attributes,
  att.typed.attributes,
  att.ascribed.attributes,
  model.divPart.spoken |= pause
  att.timed |= pause
  att.typed |= pause
  att.ascribed |= pause
.. or

<!ELEMENT %n.pause; %om.RR; EMPTY>
<!ATTLIST %n.pause;
%att.global.attributes;
%att.timed.attributes;
%att.typed.attributes;
%att.ascribed.attributes;>
<!ENTITY % model.divPart.spoken
"%x.model.divPart.spoken; %n.event; | %n.kinesic;
| %n.pause; | %n.shift; | %n.u;
| %n.vocal; | %n.writing;">
... and, indeed

<pause/>

A pause either between or within utterances. **8.3.2 Pausing**

**Module**

spoken — **8 Transcriptions of Speech**

**Attributes**

[att.timed att.typed att.ascribed ]

**Declaration**

```xml
<pause dur="PT42S" type="pregnant"/>
```

**Example**

**Contained by**

`model.global.spoken`

**May contain**

Empty element
Adding a new element

```xml
<schemaSpec>
  <moduleRef key="header"/>
  <moduleRef key="core"/>
  <moduleRef key="text"/>
  <moduleRef key="textstructure"/>
  <elementSpec ident="soundClip" mode="add">
    <classes>
      <memberOf key="model.graphicLike"/>
    </classes>
    <attList>
      <attDef ident="location">
        <desc>supplies the location of the clip</desc>
        <datatype>
          <rng:ref name="data.pointer"/>
        </datatype>
      </attDef>
    </attList>
    <desc>includes an audio object in a document.</desc>
  </elementSpec>
</schemaSpec>```
Uniformity of description

- modules, elements, attributes, value-lists are treated uniformly
  - each has an identifier, a gloss, a description, and one or more equivalents
  - each can be added, changed, replaced, deleted within a given context

- Example
  - membership in the att.typed class gives you a generic type attribute, which can be over-ridden for specific class members
Overriding a value-list

```
<elementSpec ident="list" module="core">
  <classes>
    <memberOf key="att.typed"/>
  </classes>
  <attDef ident="type" mode="replace">
    <valList type="closed">
      <valItem ident="ordered">
        <gloss>Items are ordered</gloss>
      </valItem>
      <valItem ident="bulleted">
        <gloss>Items are bulleted</gloss>
      </valItem>
      <valItem ident="frabjous">
        <gloss>Items are frabjous</gloss>
      </valItem>
    </valList>
  </attDef>
</elementSpec>
```