King's College London
University of London

allc

CCH
Centre for Computing in the Humanities

Introduction to TEI
Why use it and how?

Workshop **Text Markup & Database Design**
University of Leipzig
13 December 2007

Dr Arianna Ciula

TEI

Consortium

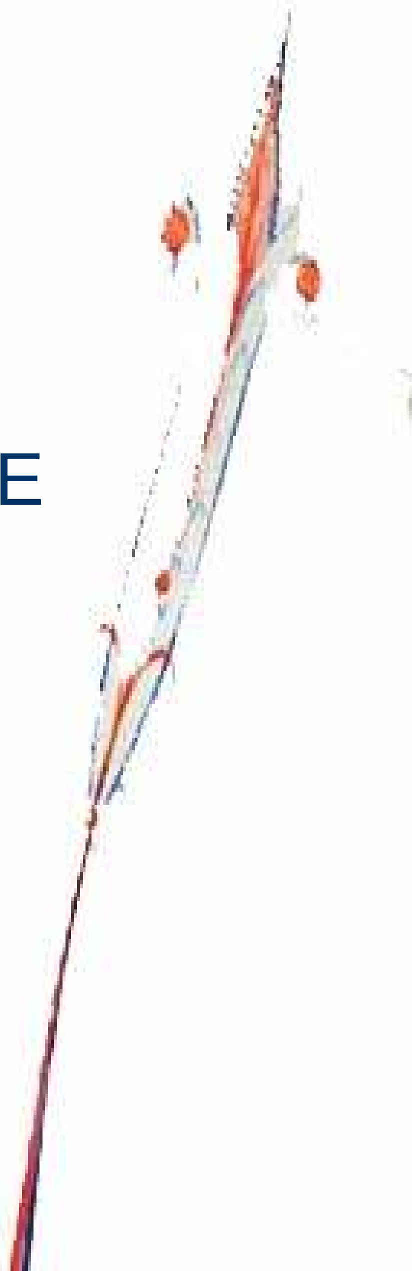
TEI sands for

Text Encoding Initiative

TEI is 20 years old!

- originally (1987) a research project within the humanities
  - Sponsored by three professional associations
  - Funded 1990-1994 by US NEH, EU LE Programme et al

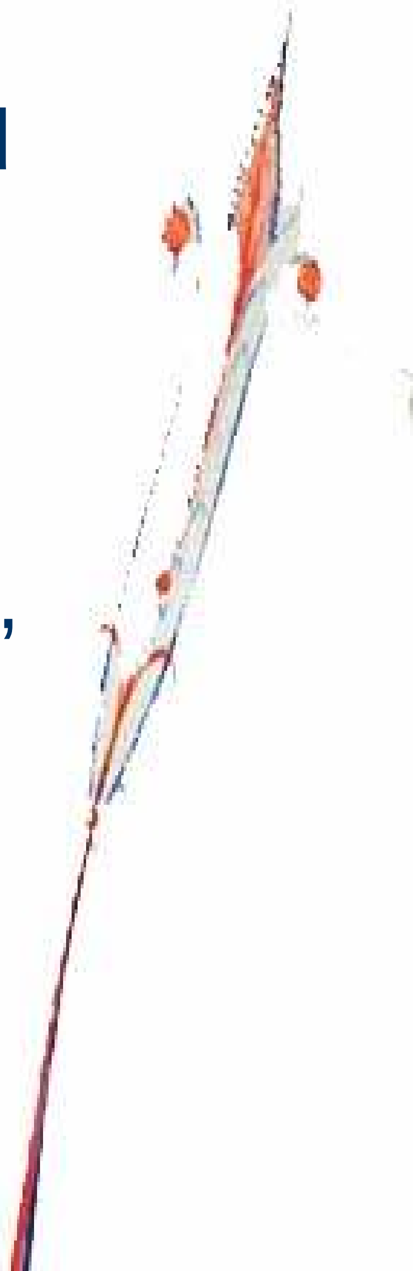
- Major influences
  - digital libraries and text collections
  - language corpora
  - scholarly datasets



TEI is 20 years old!

- International consortium established June 1999 http://www.tei-c.org/
- Partners:

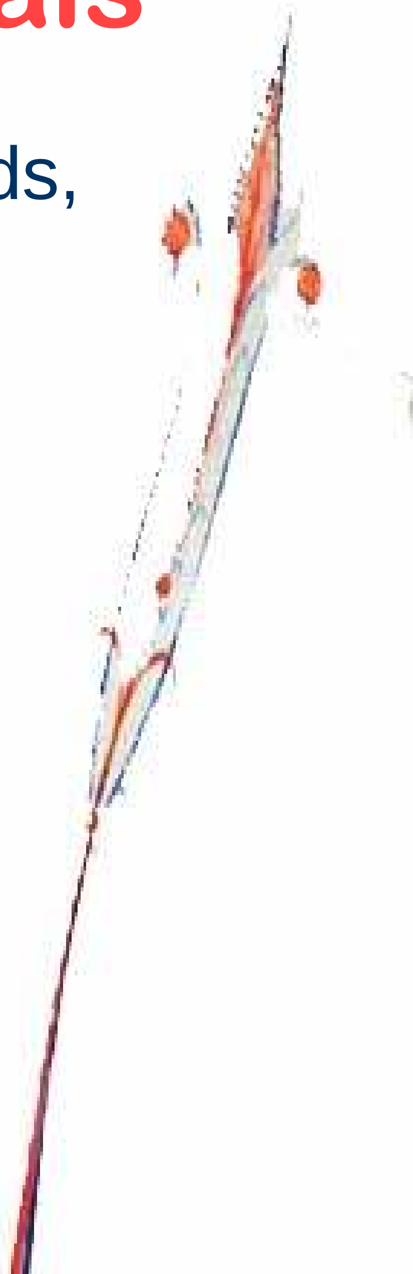
ACH, ACL, ALLC, US National Endowement for the Humanities, EU, Canadian Social Science Research Council, Andrew W. Mellon Foundation…



TEI Goals

Support the study of text

- for all texts, in all languages, from all periods, of any literary genre or textual type without restrictions of form or content
- guidance and assistance
  - for the perplexed: **what** to encode → hence, a user-driven codification of existing best practice
  - for the specialist: **how** to encode → hence, a loose framework into which unpredictable extensions can be fitted
- continuous material → textual flow
- discontinuous, more structured material → dictionaries and linguistic corpora



TEI Flexible and Modular

- standards for exchange and integration of humanities data
- guidelines for encoding texts in every format
- support on the encoding of every characteristic of every type of text of scholar interest
- platform independent approach

The TEI work has converged towards the development of **mark-up** schemes for texts based on **XML**

Markup

make explicit (to a machine) what is implicit (to a person)

supply multiple annotations/interpretations

Every writing is a kind of markup

- ***mark-up*** **language**
  - set of markup conventions to associate some categories to parts of text
  - called also encoding or annotation
- **meta-language or meta-markup**

  methodology, description of di murk-up schemes, requirements for a rigorous and documented encoding

XML *extendable mark-up language*

international standard

```
<div type="chap">

       <head>heading</head>

       <div type=“part">
           <p>paragrah</p>
           <p>an other paragraph</p>
       </div>

       <div type=“part">
            <p>paragrah</p>
       </div>

   </div>
```

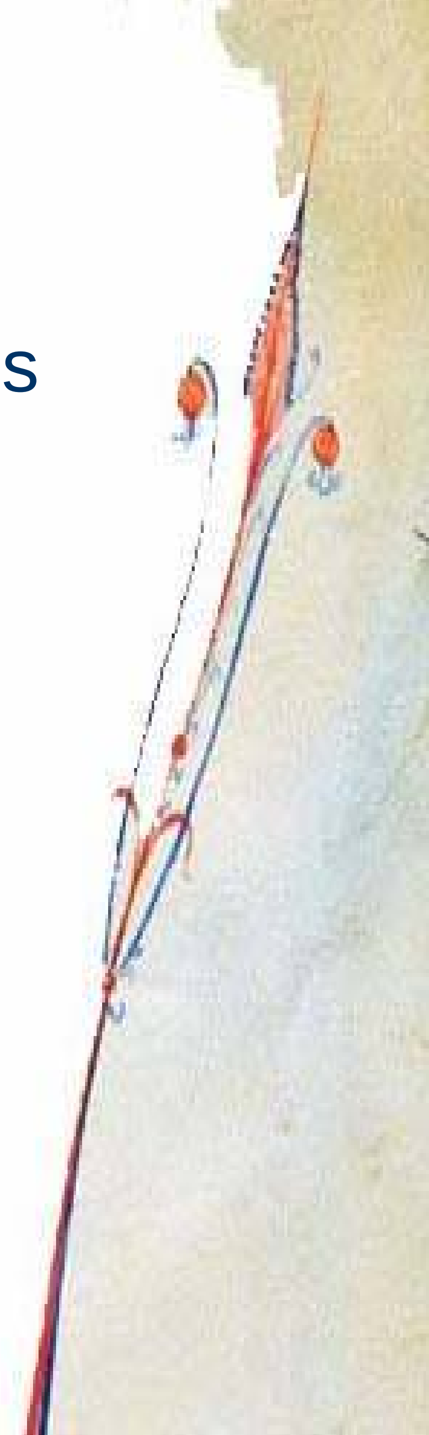
XML Schema

- 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



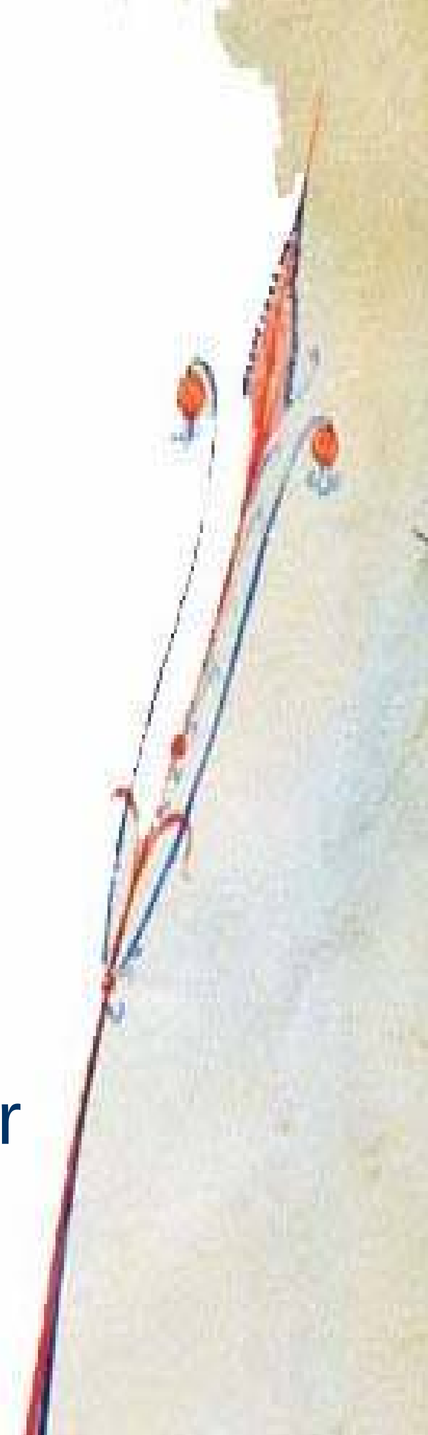
TEI Modular

- not one single schema, but framework for the definition of multiple schemas

- it defines and names several hundred useful textual distinctions

- it provides a set of modules that can be used to define schemas making those distinctions

- it provides a customization mechanism for modifying and combining those definitions with new ones using the same conceptual model



TEI Deliverables

- Set of recommendations for text encoding, covering both generic text structures and some highly specific areas based on (but not limited by) existing practice

- Very large collection of element definitions with associated declarations for various schema languages

- Modular system for creating personalized schemas or DTDs from the foregoing

TEI Legacy

- a way of looking at what ‘text’ really is

- a formalisation of current scholarly practice

- a set of shared assumptions and priorities about the digital agenda:
  - focus on content and function (rather than presentation)
  - identify generic solutions (rather than application-specific ones)

TEI How use it?

- rigorous and flexible at the same time: based on *best practice* → syntactic standard open customisations and use of new tools

  you use its modular system to build an encoding scheme appropriate to your needs, by selecting specific modules

- active discussion list to which everyone interested in TEI can subscribe and contribute to specific issues raising questions or providing expertise http://www.tei-c.org/Support/
- TEI *Special Interest Groups*

TEI How use it?

- The TEI is a *modular* system: each module defines a group of elements and attributes

- Elements are classified structurally and semantically
  - semantic classes group elements which have similar meanings — elements like names, or like editorial interventions for example
  - structural classes group elements which behave similarly in the structure — elements like paragraphs, or like phrases for example
  - there are also attribute classes: these group elements which all have the same attributes

TEI Modules

- **teistructure**
  - defines all named element classes and macros
  - basic “book-like” structure for prose, verse, drama
- **Core**
  - the TEI header
  - ‘core’ elements “common to all kinds of text”
- **Alternative structures**
  - eg transcribed speech, dictionaries ...
- **Specialist applications**
  - linking and alignment; analysis; non-standard characters and glyphs; feature structures; certainty; transcription; textual criticism; names and dates; language corpora; manuscript description. . .
  - the ODD system

TEI Basic structure(s)

In **P5** the root element is **`<TEI>`**

Every TEI-conformant document comprises a **`<teiHeader>`** followed by (at least one) **`<text>`**

- **`<teiHeader>`** contains:
  - – mandatory file description
  - – optional encoding, profile and revision descriptions
- **`<teiHeader>`** is essential for:
  - – bibliographic control and identification
  - – resource documentation and processing

TEI *header* Some components

`<fileDesc>` (file description) contains a full bibliographic description of an electronic file

- **`<titleStmt>`** provides a title for the resource and any associated statements of responsibility
- **`<sourceDesc>`** documents the sources from which the encoded text derives (if any)
- **`<publicationStmt>`** documents how the encoded text is published or distributed

- **`<encodingDesc>`** (encoding description) specifies the methods and editorial principles which governed the transcription or encoding of the text
- **`<revisionDesc>`** (revision description) summarizes the revision history for a file

TEI *Book like* Structure of `<text>`

- A text may be unitary or composite
- A unitary text contains
  - optional front matter **`<front>`**
  - a body **`<body>`**
    - usually constituted by a set of nested divisions **`<div>`**
  - optional back matter **`<back>`**

- In a composite text, the body is replaced by a group of texts (or nested groups)
- A corpus is a collection of text and header pairs, which also has its own header

TEI Text Divisions

A text usually has divisions

- generic, hierarchic subdivisions
- the `@type` attribute is used to label a particular level e.g. as "part" or "chapter"
- vanilla or numbered tags may be used to identify level explicitly
- the `@n` attribute gives a particular division a name or number
- the `@xml:id` attribute gives a particular division a unique identifier
- associated `<head>` and other elements may also be supplied

TEI
Text Structure

```
<text>
 <front> <!-- titlepage, etc here --> </front>
 <body>
    <head>Realism in Language and Music: Kurt Weill's <title>Street
    Scene</title></head>
    <div>
        <p>Kurt Weill (1900-1950) began his musical career at an early age, working
        for the theater in Dessau, and then studied under Ferrucio Busoni at the
        Akademie der Künste in Berlin; by the early 1920's he was already well-
        established and respected within the German musical and dramatic
        community. […] </p>
        <p>Weill's popularity in Germany was cut short, however, by the rise of the
        Nazi party, and he fled Berlin on March 21, 1933 for Paris. By January of
        1935 he had moved on to England, and finally arrived in New York on
        September 10, 1935. […] </p>
    </div>
        <!– more text here -->
    <div>
        <p>For his foray into the new musical theater, Weill chose, after extended
        deliberation, what he considered to be the prime material: the Pulitzer Prize
        winning play Street Scene (1929) by Elmer Rice. […]</p>
        <!– more text here -->
    </div>
 </body>
</text>
```

TEI Global Attributes

Defined in the core module, available for all elements:

- `@xml:id` supplies a unique identifier
- `@n` supplies a (non-unique) name or number
- `@rend` gives a suggestion about rendition (appearance)
- `@xml:lang` identifies the language using an ISO standard code

Defined in the linking module

- `@corresp`, `@synch`, `@ana` for specific association types
- `@next`, `@prev` for aggregating fragmented elements

TEI Text Components

What are divisions composed of?

- prose is mostly paragraphs **`<p>`**
- verse is mostly lines **`<l>`,** sometimes in hierarchic groups **`<lg>`**
- drama is mostly speeches **`<sp>`** containing **`<p>`** or **`<l>`** elements interspersed with stage directions **`<stage>`**

These may be mixed, and may also appear directly within undivided texts.

TEI Text Components

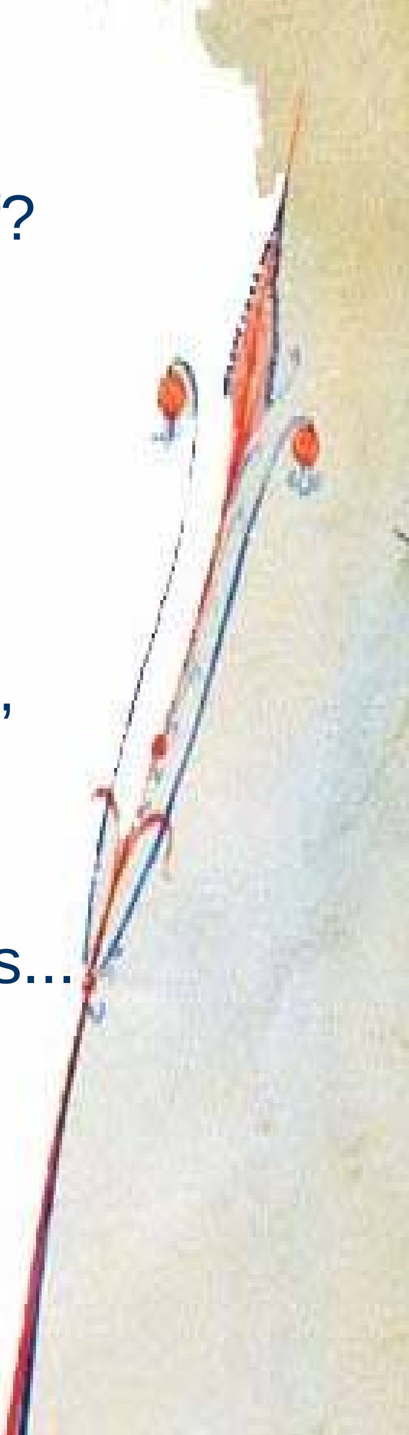
```
 <div>
     <head><title>The Journey</title></head>
     <lg>
          <l>Suevia, my mother, happy land!</l>
          <l>You also are like your more shining
          sister</l>
          <l>Lombardy over there</l>
          <l>Flowed through by a hundred streams</l>
          <l>And trees in plenty, white with blossom or
          reddish</l>
          <l>And the darker, deep, full green, the wild
          trees</l>….
     </lg>….
</div>
```

TEI Further down the hierarchy

What are speeches, paragraphs, and lines made of?

- phrases that are conventionally typographically distinct
- “data-like” (names, numbers, dates, times, addresses)
- editorial interventions (corrections, regularizations, additions, omissions ...)
- cross references and links
- lists, notes, graphics, tables, bibliographic citations...
- all kinds of annotations!

What you tag will depend on your research agenda



TEI Further down the hierarchy

```
<p>
    <q>I praise my <name>Leipzig</name>! It
    is a small <name>Paris</name> and
    educates its people.</q>
</p>
```

TEI Foreign language phrases

- The @xml:lang attribute may be attached to any element
- Use <foreign> if nothing else is available
- Use ISO 639-2 code to identify language

Have you read <title xml:lang="de">Die Dreigroschenoper</title>?

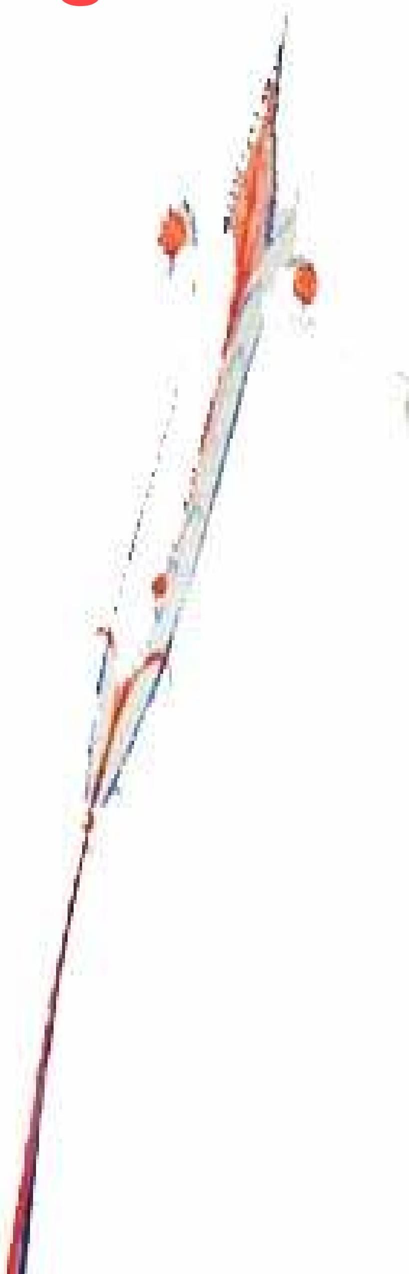
<mentioned xml:lang="fr">Savoir-faire</mentioned> is French for know-how.

John has real <foreign xml:lang="fr">savoir-faire</foreign>.

TEI Names and other referring strings

The **<rs>** (referring string) element is used for any kind of name or reference

```
<q>My dear <rs type="person" key="BENM1">Mr.
Bennet</rs>,</q> said <rs type="person"
key="BENM2">
his lady</rs> to him one day,<q>have you heard
that
<rs type="place" key="NETP1">Netherfield
Park</rs>
is let at last?</q>
```



TEI Correction and Regularization

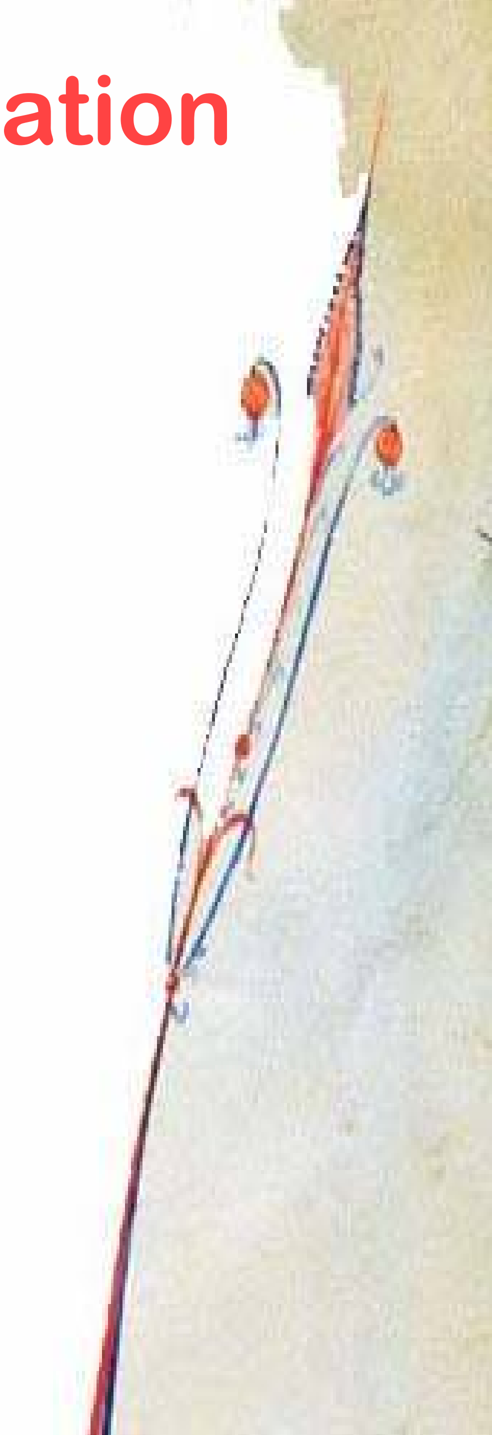
- `<corr>` marks a correction
- `<sic>` marks a (deliberate) non-correction
- `<reg>` marks a regularization
- `<orig>` marks something deliberately un-normalized
- Use `<choice>` to indicate a combination of possible encodings

TEI Correction and Regularization

```
<choice>
 <orig>a</orig>
 <reg>he</reg>
</choice>

<choice>
 <sic>table</sic>
 <corr resp="Theobald">babbl’d</corr>
</choice>

<choice>
 <orig>feelds</orig>
 <reg>fields</reg>
</choice>
```



TEI ‘Inter’ elements

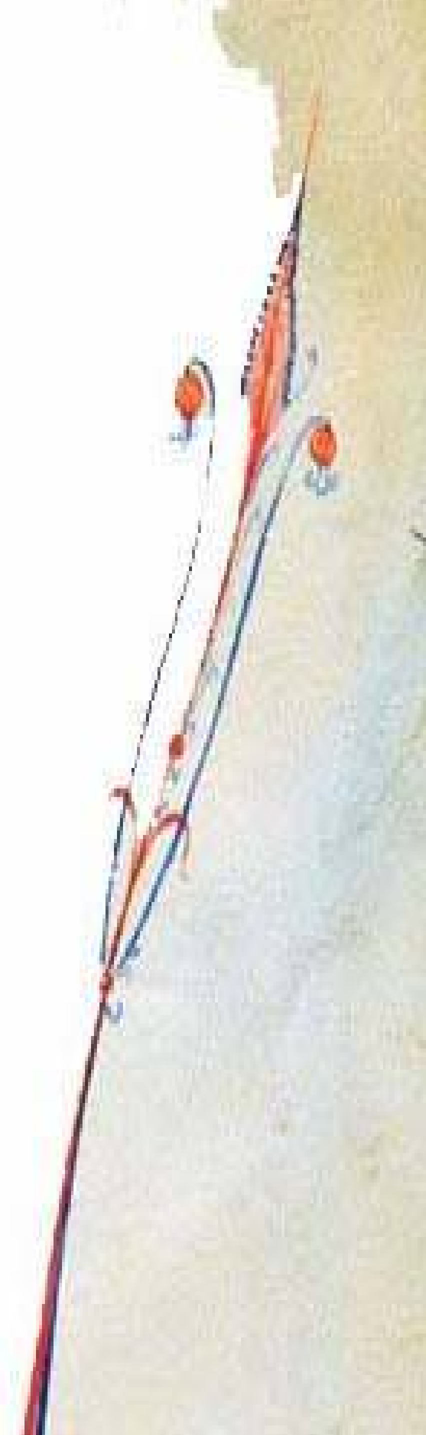
<list> lists of all kinds

<note> notes (authorial or editorial)

<figure> pictures or figures

<table> tables

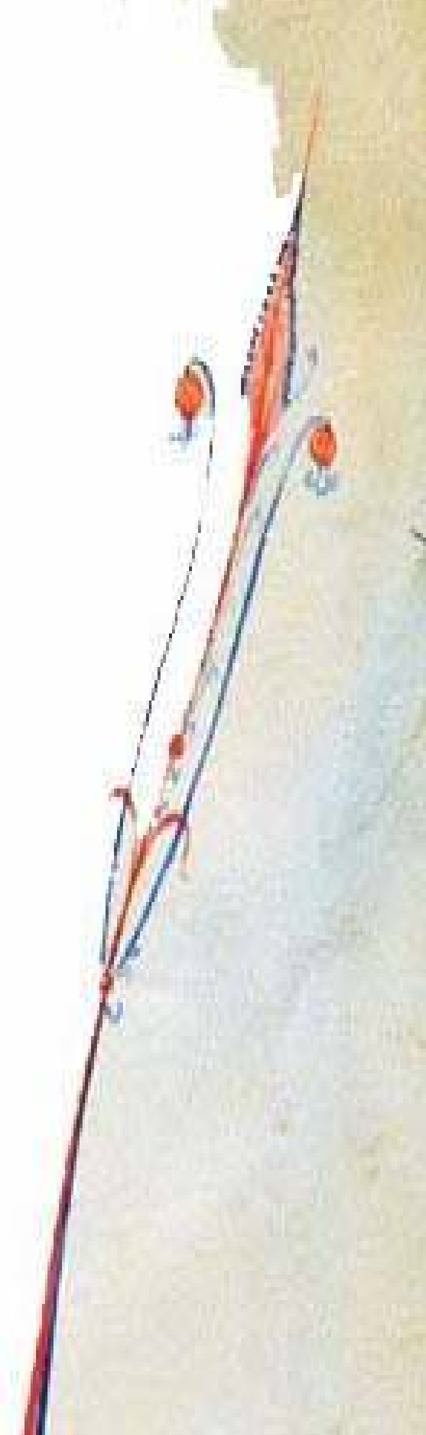
<bibl> bibliographic descriptions



TEI Lists

- use `<list>` for lists of any kind
  - – `@type` attribute to distinguish
- use `<label>` as alternative to `@n` attribute
- may be nested as necessary

```
<list type="xmas">
    <label>For my true love</label>
    <item>
        <list type="bullets">
            <item>three calling birds></item>
            <item>two french hens</item>
            <item>a partridge in a pear tree</item>
        </list>
    </item>
    <label>For Uncle Joe</label>
    <item>socks as usual</item>
</list>
```



TEI Figures and graphics

The presence of a graphic is indicated by the `<graphic>` element, usually contained within a `<figure>` element which groups together:

- The title of the graphic `<head>`
- A description of the graphic `<figDesc>` for use by software unable to render the graphic
- The graphic resource itself is pointed to by an `@url` attribute on the `<graphic>` element, and may also have attributes `@scale, @height, @width`
- `<figure>`s may self-nest, and may also contain other display class items such as `<formula>`s

TEI Figures and graphics

```
<figure>
    <head>Zimmerman's Coffeehouse</head>
    <figDesc>Zimmerman’s Coffeehouse on Catherine Street, just off the main market square, in Leipzig, where Bach's Collegium Musicum gave regular concerts</figDesc>
    <graphic url="http://en.wikipedia.org/wiki/Image:Zimmermannsches_Caffeehaus.jpg"/>
</figure>
```

TEI Tables

- a `<table>` element contains `<row>`s of `<cell>`s
- spanning is indicated by `@rows` and `@cols` attributes
- `@role` attribute indicates whether row or column holds `“data”` or a `“label”`
- embedded tables are permitted

TEI Tables

Bach’s Works		
Bach Works Catalogue Number	1–224	525–748
Type of Music	cantatas	organ works

```
<table>
      <row>
            <cell cols="3" role="label">Bach’s Works</cell>
      </row>
      <row>
            <cell role="label">Bach Works Catalogue Number</cell>
            <cell>1–224</cell><cell>525–748</cell>
      </row>
      <row>
            <cell role="label">Type of Music</cell>
            <cell>cantatas</cell><cell>organ works</cell>
      </row>
</table>
```

TEI Bibliography

The `<listBibl>` element lists bibliographic citations

- Individual citations may be represented loosely as
  - `<bibl>` elements
  - or `<biblStruct>` elements → more structured way
- In either case, they contain elements from the `model.biblPart` class, e.g.
  - `<author>`, `<editor>`, (generic) `<respStmt>` etc.
  - `<title>` with optional `@level` attribute to distinguish `“monographic”`, `“analytic”` etc.
  - `<imprint>` groups publication info (`<publisher>`, `<date>` etc.)
  - `<biblScope>` adds page references etc.
- Individual citations may be linked to the bibliography

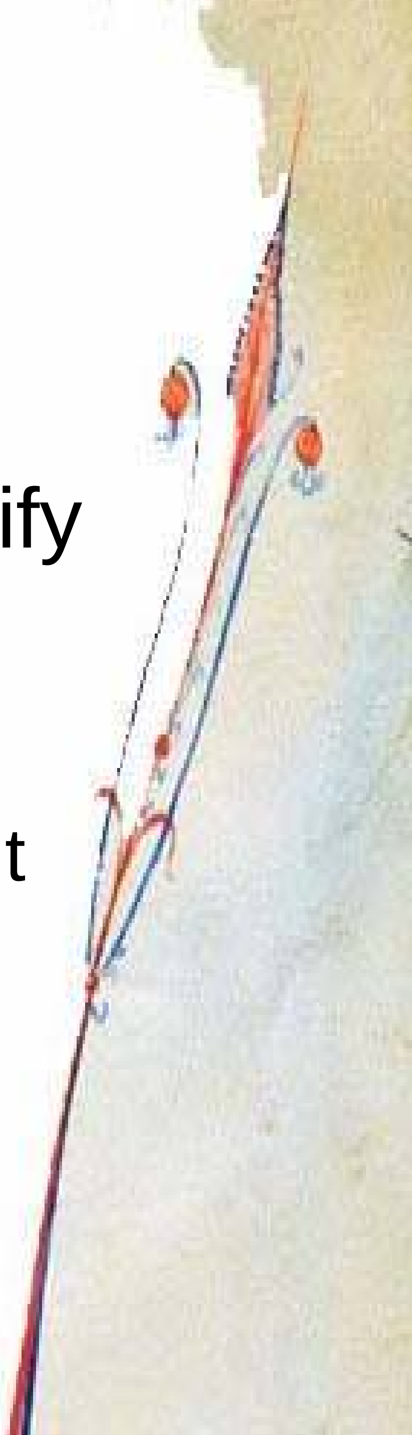
TEI Bibliography

```
<p>See for example <ref target="#Vanh2004">Vanhouette
(2004)</ref>...

<div><head>Bibliography</head>
    <listBibl>
        <bibl xml:id="Vanh2004">
            <author>Edward Vanhouette</author>
            <title level=“a">An Introduction to the TEI and the
            TEI Consortium</title>
            <title level=“j">Literary and Linguistic Computing
            </title>
            <date>2004</date>
            <biblScope type=“issue”>19.1</biblScope>
            <biblScope type=“pp”>pp. 9-16</biblScope>
        </bibl>
    </listBibl>
</div>
```

TEI Notes

- Use <note> for notes of any kind (editorial or authorial)
- if in-line, use @place attribute to specify location
- if out of line, either use
  - @target attribute to specify attachment point
  - or mark attachment point as a <ref>



TEI Notes

```
<lg>
    <l>The self-same moment I could pray></l>
    <l>And from my neck so free</l>
    <l>The albatross fell off, and sank</l>
    <l>Like lead into the sea.
    <note type="auth" place="margin">
    The spell begins to break.</note></l>
</lg>
```

OR

```
...
    <l>The albatross fell off, and sank</l>
    <l xml:id="L213">Like lead into the sea. </l>
</lg>
...
    <note type="auth" place="margin" target="#L213">
    The spell begins to break.</note>
```

TEI Other Modules

Your choice from:

- Transcription of spoken texts
- Dictionaries and lexica
- Varieties of linguistic annotation
- Nonstandard characters and glyphs
- Linking, alignment, non-hierarchic structures
- Detailed metadata (the TEI Header)
- Manuscript Description
- Text-critical apparatus
- Physical description
- Names, Dates, People, and Places
- The ODD system

References

- Association for Computers and the Humanities (ACH): http://www.ach.org
- Association for Computational Linguistics (ACL): http://www1.cs.columbia.edu/acl/home.html
- Association for Literary and Linguistic Computing (ALLC): http://www.allc.org
- TEI Consortium (TEI): http://www.tei-c.org
- World Wide Web Consortium (W3C): http://www.w3c.org

References

Lou Burnard, Matthew Driscoll, and Sebastian Rahtz *Digital Texts, XML, and TEI,* TEI Training, Sofia, October 2005: http://www.tei-c.org/Talks/2005/Sofia/talk-intro.pdf

Lou Burnard *TEI contents: an overview of the modules making up TEI P5,* TEI Training, Sofia, October 2005: http://www.tei-c.org/Talks/2005/Sofia/overview.pdf

Sperberg-McQueen, C. M., *Text in the Electronic Age: Textual Study and Text Encoding with Examples From Medieval Text*, « Literary and Linguistic Computing », 6.1 (1991), pp. 32-46.

Vanhouette, Edward, *An Introduction to the TEI and the TEI Consortium*, « Literary and Linguistic Computing », 19.1 (2004), pp. 9-16.