ZENTRUM FÜR
INFORMATIONSMODELLIERUNG
AUSTRIAN CENTRE FOR
DIGITAL HUMANITIES

KARL-FRANZENS-UNIVERSITÄT GRAZ
UNIVERSITY OF GRAZ

Beyond TEI

Digital Editions with XPath & XSLT for the Web & in LaTeX

Sarah Lang

Harvard, April/May 2022

Overview

1. The workshop

The workshop

Goals

1. get to know XPath & XSLT (and learn how to use it)
2. understand the role of XML/TEI, XPath and XSLT in Digital Editing
3. be able to use XSLT to generate HTML and LaTeX output from TEI
4. Two days isn't enough for you to master XSLT!

Schedule

Day 1, morning	XML, TEI and Digital Editing → repetition of the basics, making sure we're all on the same page, understanding why we're even learning XSLT.
Day 1, afternoon	Navigating XML documents using XPath, introduction to HTML (& Bootstrap) and LaTeX (& `reledmac`)
Day 2	Transforming XML documents into HTML & LaTeX output formats using XSLT

Single point of entry for all workshop-related materials: LaTeX Ninja blogpost & Github Repository (‘additional resources’ directory)

Introductions

Please introduce yourselves!

1. Name, pronouns, field/topic of study
2. Why did you come to this workshop?
3. Previous experience with Digital Humanities (DH) or editing?

Contact

@SarahALang_
@latex_ninja

sarahalang.com
latex-ninja.com

sarah.lang@uni-graz.at

Sarah Lang (she/they)

- originally from Germany, now in Graz (Austria)
- Studied Latin, French & History (teacher's education) in Graz & Montpellier (France), then Archaeology Bachelor, Master in Religious Studies & Philosophy
- got a DH certificate & started working at Zentrum für Informationsmodellierung (ZIM) / Centre for Information Modelling in Graz
  - Moral Weeklies/Spectators → *gams.uni-graz.at/mws*
  - Graz Repository of Ancient Fables (GRaF) → *gams.uni-graz.at/graf*
  - *PhD thesis:* Decoding alchemical *Decknamen* digitally. A Polysemantic Annotation and Machine Reasoning Algorithm for the Corpus of Iatrochymist Michael Maier (1568–1622)
- *Now:* teaching in Graz, Passau & Vienna; PostDoc in Graz. *Research interests:* history of science (alchemy), Neo-Latin, text mining and computer vision

TEI for Digital Editing i

TEI can describe the structure of a text, e.g.

- speaker, verse line, stage directives
- greeting, signature
- Visual aspects of the script
- special characters, new lines

Simple layout markup

- beginning of a new **line**: *<lb/>*
- beginning of a new **page**: *<pb/>* *@n* for an explicit numbering
- beginning of a new **column**: *<cb/>*
- **highlighted** text: *<hi>*
  - Attribute *@rend* to describe the appearance
  - Alternative encoding: , ,
- graphical elements in the text: *<figure>*
- *<fw>* (forme work) contains a running head (e.g. a header, footer), **catchword**, or similar material appearing on the current page.

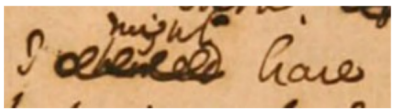
```
<fw place="top-centre" type="head">Poëms.</fw>
<fw place="top-right" type="page-no">29</fw>
```

TEI for Digital Editing ii

Documenting particularities of the writing surface

<damage>	*@agent, @degree, @unit, @quantity, @extent, @precision, @scope*
<unclear>	
<gap>	any ommission in the transcription – @reason, e.g. sampling, inaudible, irrelevant, cancelled

e.g. unclear passage



Other important attributes

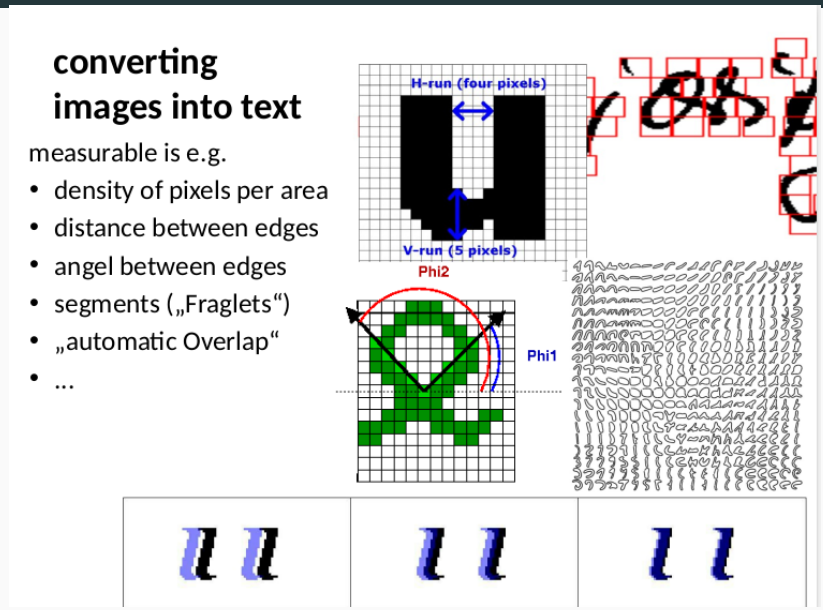
@cert(ainty)	how certain you are about the suggested transcription?
@resp(onsibility)	who did it?
@evidence	where you got the clues from (internal, external, conjecture)?

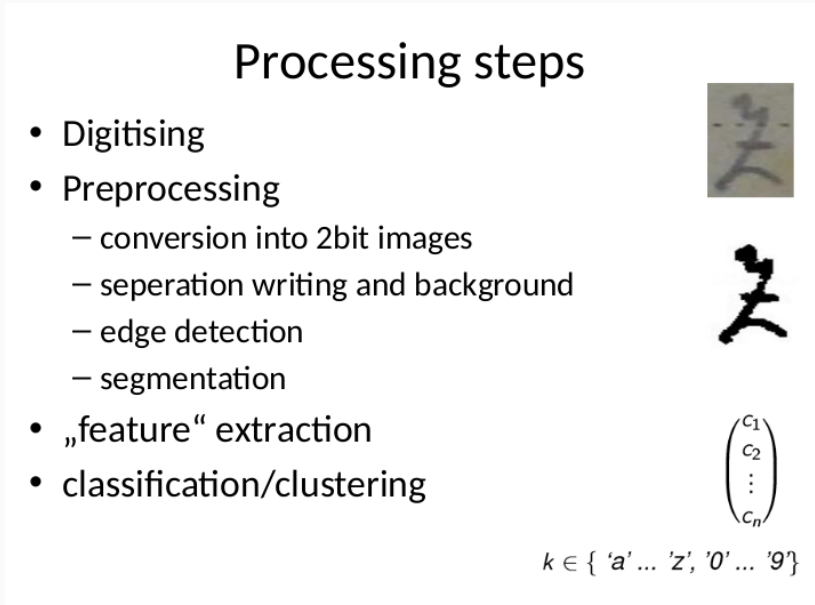
```
I <subst>
 <add place="above">might</add>
  <del>
   <unclear reason="overinking"
      cert="medium" resp="#LDB">
      should</unclear>
  </del> </subst> have
```

```
<gap reason="wormhole" quantity="5" unit="character"/>
<damage agent="coffee" quantity="3" unit="line"/>
```

Transcription i

- OCR (Optical Character Recognition) – e.g. Transkribus (transcription support)
- also: Transkribus Keyword Spotting
- also: fuzzy search which should also find the word if it's mistranscribed
- Writer identification





Transcription ii

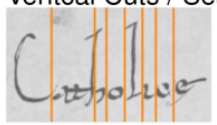
Typical phenomena

- “Special characters”
- Abbreviations
- damaged or unreadable text
- additions, deletions, substitutions, corrections
- editorial interventions (emendations and conjectures)
- editorial additions or omissions

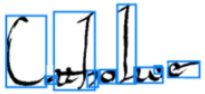
Preprocessing: e.g. different segmentation methods

- Grid

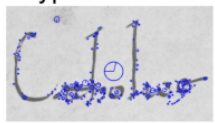
- Veritcal Cuts / Seam Cuts



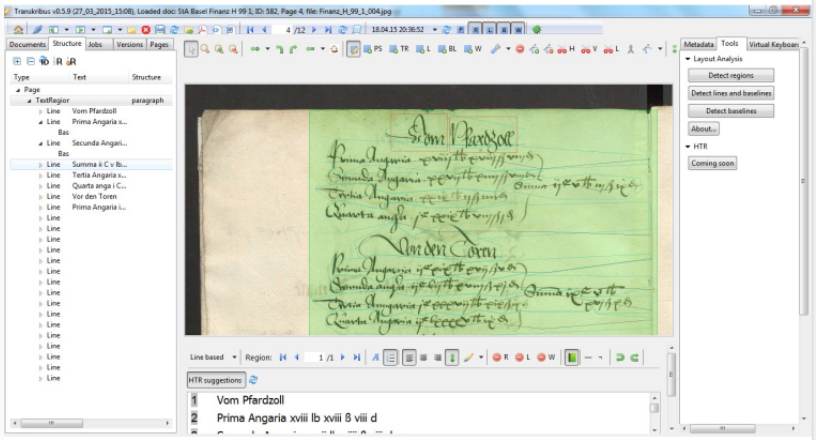
- Connected Components



- Keypoint-based



Transkribus



Transcription iii

Transcriptions contain:

- layout
- additions
- corrections
- modifications
- voids, space, holes, gaps ...
- alternative transcriptions
- editorial interventions
- Enhanced transcription

```
<pb>, <lb>, <cb>, <hi>, <g>,
<handShift/>
<add>, <addSpan>,
<corr>, <del>, <delSpan>, <sic>
<subst>
<gap>, <damage>
<choice>, <alt>
<unclear>, <supplied>, <reg>
<add>, <addSpan>, <corr>, <choice>,
<damage>, <del>, <delSpan>,
<restore>, <gap>, <sic>
```

Genetic Edition

Critique Génétique

Research interest: Reconstruct the writing process in the working manuscripts of the author

```
That's <del xml:id="del1">
superfluous
<restore>
 <redo target="del2" />
 <del xml:id="del2">deleted</del>
</restore> </del>
text.
```

```
<zone>
 <line>Alone
  <seg type="alternative"
   xml:id="alt1"> before </seg>
  <add place="above"
       type="alternative"
       xml:id="alt2">beside</add>
  his native river—
 </line>
 <alt targets="#alt1 #alt2"
      mode="excl" weights="01"/>
</zone>
```

```
He sat <seg type="transposition"
xml:id="trans1"> at
his table</seg>
<seg type="transposition"
xml:id="trans2"> head on hands
</seg>.

<listTranspose>
<transpose>
  <ptr target="#trans2"/>
  <ptr target="#trans1"/>
</transpose>
</listTranspose>
```

The script: palaeography

- soft hyphen: *@break="no"*
- *@rend | @rendition*
  - *@rend*: verbal description; each word describes a single facet (*rend="indented:5cm"*)
  - *@rendition*: reference to description of the rendition in the *teiHeader//encodingProfile*
- „special characters":
  - *<g>*
  - Does it exist in Unicode? (*http://www.unicode.org*). As an entity in XML:

```
&#x[hexadecimal code];
&#[decimal code];
```

Unicodefor historical texts

- *Combining Diacritical Marks* (0300–036F) and *Supplement* (1DC0–1DFF): Superscripts, Subscripts
- *Latin Extended Additional* (1E00–1EFF): characters with diacritics
- *Latin Extended-D* (A720–A7FF): Ligatures, abbreviations, ...
- *General Punctuation* (2000–206F) and *Supplement* (2E00–2E7F)
- *Ancient Symbols* (10190–101CF): roman measurements, coins...

Editorial interventions

- expansion of abbreviations
- Conjectures
- Normalisations

- `<abbr>, <expan>` plus `<am>, <ex>`
- `<sic>, <corr>`
- `<orig>, <reg>`

All these can be paired:

- General for all editorial interventions: `<choice>`.
- Explicitly for substitutions: `<subst>`.
- `<supplied>` for additions by the editor
- `<unclear>` for unreadable text (`@reason, @agent, @hand`)

Abbreviations i

In Western MS, we usually distinguish:

- **Suspensions:** the first letter or letters of the word are written, generally followed by a point : for example ‘e.g.’ for ‘exempla gratia’
- **Contractions:** both first and last letters are written, generally with some mark of abbreviation such as superscript strokes, or points : e.g. ‘Mr.’ for ‘Mister’
- **Brevigraphs:** Special signs such as the Tironian nota used for ‘et’, the letter p with a barred tail used for ‘per’, the letter c with a circumflex used for ‘cum’/’con’ etc.
- **Superscripts:** Superscript letters (vowels or consonants) used to indicate various kinds of contraction: e.g. ‘w’ followed by superscript ‘ch’ for ‘which’.

- *<expan>* The element content is considered as the expansion of an abbreviation. In the text: USA → transcription:

```
<expan>United States of America</expan>
```

- *<abbr>* The element content is an abbreviation

```
<abbr>USA</abbr>
```

- *<ex>* (expansion) and *<am>* (abbreviation mark) for the omitted part of the abbreviation, e.g.

```
e<ex>xempla</ex> g<ex>ratia</ex>
e<am>.</am> g<am>.</am>
```

Abbreviations ii

Abbreviations can also be considered as alternatives: *<choice>*, e.g. ‘Zum Beispiel’ and ‘z.B.’:

```
<choice>
 <expan>Zum Beispiel</expan>
 <abbr>Z.B.</abbr>
</choice>
```

Or respectively:

```
Z<choice>
   <am>.</am>
   <ex>um</ex>
  </choice>

B<choice>
   <am>.</am>
   <ex>eispiel</ex>
 </choice>
```

Transcription = Interpretation

‘gemination dash’ – possible solutions:

- Uncommented expansion
- Unicode m with ”combining macron”
- Encoding as an XML-Entity
- `<g>` refering to `<charDecl>`
- Only `<am/>` for the stroke
- Only `<ex>` for the expansion
- As a `<choice>` with `<abbr>` and `<expan>`, the first incl. a abbreviation mark `<am>` and the second the expansion `<ex>`

Gemination dash

horizontal, bended or curved stroke above a nasal letter indicating the omission of a further instance of the same letter. (source)

Modifications

- addition, deletion, substitution, transpose, *or:*
- modification (represents any kind of general modification without interpretation)

Changing writer

<handShift />	`@new` : the hand which writes from this place onward
<handDesc>	(part of `msDesc`)
<handNotes>	(part of `profileDesc`)
<handNote>	for a particular description
@xml:id	an identifier for the hand

Text and images i

Images of a text are encoded in a *facsimile* – structure parallel to *teiHeader* and *text*:

```
<tei>
  <teiHeader>...</teiHeader>
  <facsimile> ...</facsimile>
  <text>...</text>
</tei>
```

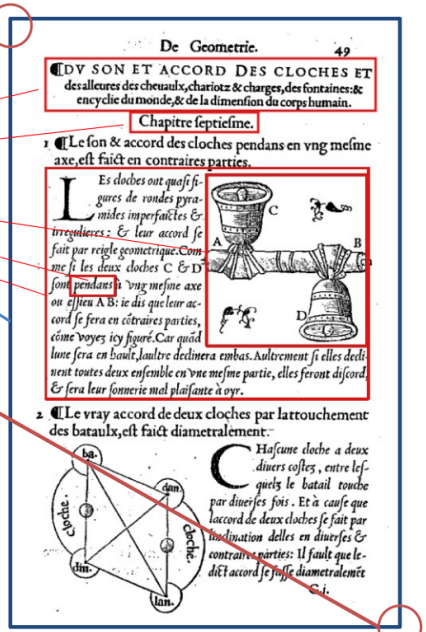
<facsimile>

- *<surface>* = something meant to be seen
  - *@uly, @ulx; @lrx, @lry* =upper left x/y- and lower right y/x coordinates
  - coordinates form a grid, which can be referred → *@ulx* and *@uly* are usually 0
- *<graphic>*: image, *@url* : image file
- *<zone>* = an area on the surface. Coordinates refer to the grid defined in *@uly*, *@ulx*; *@lrx, @lry* of the *<surface>*.

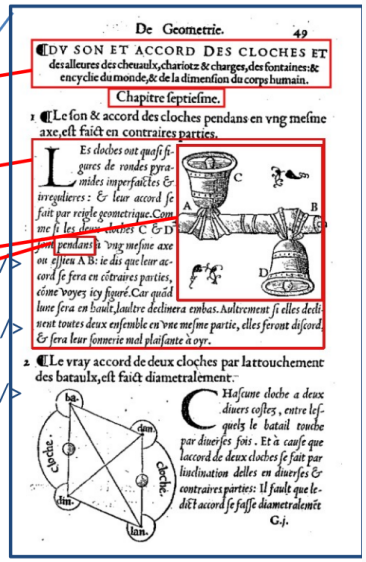
Text and images ii

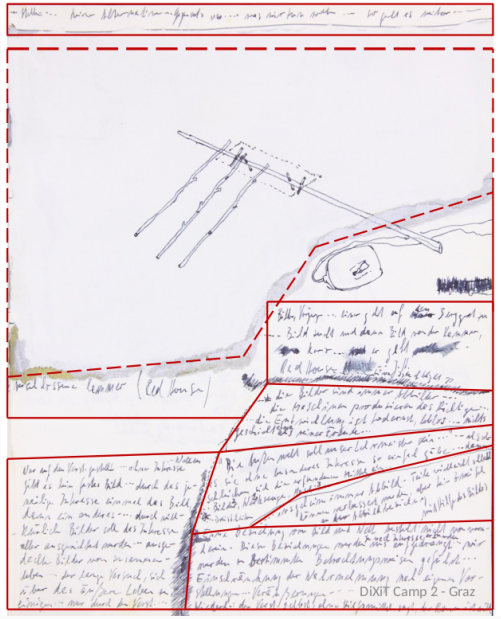
Example

surface
zone
@ulx, @uly
@lrx, @lry
graphic = http://www.tei-c.org/release/doc/tei-p5-doc/en/html/Images/facs-fig1.png



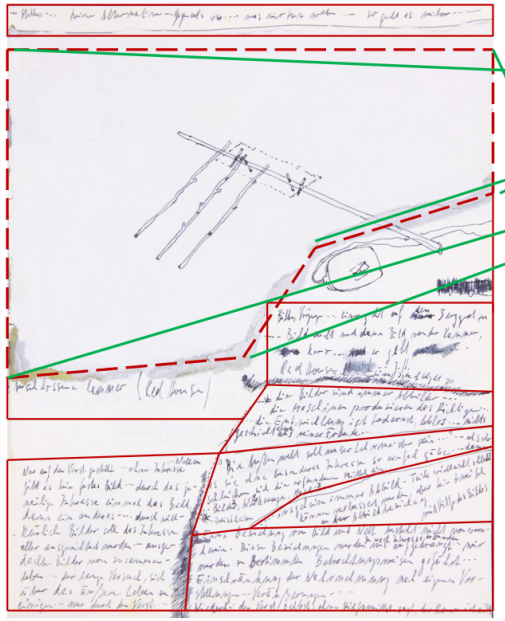

```
<facsimile>
 <surface
  ulx="0" uly="0" lrx="200" lry="300">
 <graphic url="Bovelles-49r.png"/>
 <zone
  ulx="25" uly="25" lrx="180" lry="60">
 </zone>
 <zone
   ulx="28" uly="75" lrx="175" lry="178"/>
 <zone
ulx="105" uly="76" lrx="175" lry="160" />
 <zone
   ulx="45" uly="125" lrx="60" lry="130"/>
</surface>
</facsimile>
```





<zone>

@points:
List of coordinates (pairs of numbers), which combined by lines enclose a region on the surface.



```
<zone
points="0,29
534,20 536,215
334,282 259,376
0,409"/>
```

Text and images iii

Linking text and image

- @facs, content corresponding with a xml:id in the facsimile structure:

```
<surface xml:id="p49">
  <zone xml:id="p49z2" />
 <graphic url="test.png" />
</surface>
<text><body><div>
 <pb n="49" facs="#p49"/>...
 <head facs="#p49z2">
  Chapitre septiesme </head>
</div></body></text>
```


De Geometrie. 49

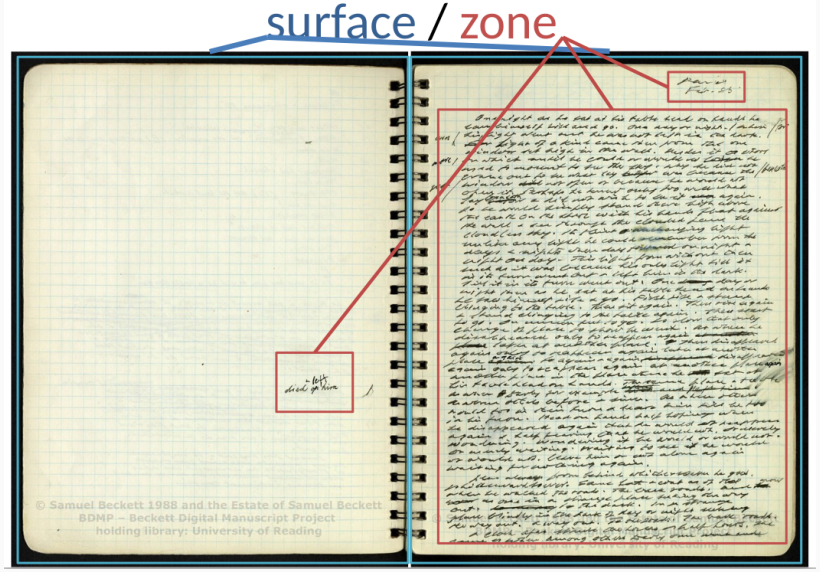
DV SON ET ACCORD DES CLOCHES ET des alleures des cheuaulx, chariotz & charges, des fontaines: & encyclie du monde, & de la dimenſion du corps humain.

Chapitre ſeptieſme.

1 Le ſon & accord des cloches pendans en vng meſme axe, ou faict en contraires parties.

Les cloches ont quaſi figures de rondes pyramides imperfaictes & irregulieres: & leur accord ſe fait par reigle geometrique. Comme ſi les deux cloches C & D ſont pendans à vng meſme axe ou eſſieu A B: ie dis que leur accord ſe fera en cõtraires parties, cõme voyez icy figuré. Car quãd lune ſera en hault, laultre declinera embas. Aultrement ſi elles deſcendent toutes deux enſemble en vne meſme partie, elles feront diſcord & ſera leur ſonnerie mal plaiſante à oyr.

2 Le vray accord de deux cloches par lattouchement des bataulx, eſt faict diametralement.



Tools for text-image linking

1. **Image markup tool** (Martin Holmes, *http://www.tapor.uvic.ca/~mholmes/image_markup/index.php*)
2. **TextGridLab:** http://www.textgridlab.de
3. **T-PEN** (*http://www.t-pen.org*)
4. *http://imagecoordinates.com*

Text and images iv

Embedded transcription

- "Embedded transcription": Text directly in *<surface>*
- *Relevant elements:* *<sourceDoc>*, *<surface>*, *<zone>*, *<line>*, *@rotate*

Embedded transcription

```
<sourceDoc ulx="0" uly="0"
  lrx="..." ...>
<surface>
  <zone ulx=".." ... >
    <line>Chapitre
      septiesme</line>
  </zone>
  <graphic url="test.png" />
</surface>
</facsimile>
```


De Geometrie. 49

DV SON ET ACCORD DES CLOCHES ET des alleures des cheuaulx, chariotz & charges, des fontaines: & encyclie du monde, & de la dimenſion du corps humain.

Chapitre ſeptieſme.

1 Le ſon & accord des cloches pendans en vng meſme axe, eſt faict en contraires parties.

Les cloches ont quaſi figures de rondes pyramides imperfaictes & irregulieres: & leur accord ſe fait par reigle geometrique. Comme ſi les deux cloches C & D ſont pendans à vng meſme axe ou eſſieu A B: ie dis que leur accord ſe fera en cõtraires parties, cõme voyes icy figuré. Car quãd lune ſera en hault, laultre deſcendra embas. Autrement ſi elles deſcendent toutes deux enſemble en vne meſme partie, elles feront diſcord, & ſera leur ſonnerie mal plaiſante à oyr.

2 Le vray accord de deux cloches par lattouchement des batаulx, eſt faict diametralement.

Haſcune cloche a deux diuers coſtez, entre leſquelz le batail touche par diuerſes fois. Et à cauſe que laccord de deux cloches ſe fait par

```
<sourceDoc>
 <surfaceGrp n="leaf1">
  <surface facs="page1.png"> <zone>All the writing on page 1</zone> </surface>
  <surface>
   <graphic url="page2-highRes.png"/>
   <zone> <line>A line of writing on page 2</line> </zone>
  </surface>
 </surfaceGrp>
</sourceDoc>
```

Critical Apparatus i

> ...aims at documenting the variants of a text in the witnesses of textual transmission.

A critical apparatus is encoded with: *<app>*, *<rdg>* / *<rdgGrp>* and *<lem>*. *<lem>* can contain text not documented in any textual witness.

```
<app>
  <rdg wit="#Sh1">Then</rdg>
  <lem wit="#Sh2">Than</rdg>
</app> is my deede to my most
<app>
  <lem wit="#Sh1">painted</rdg>
  <rdg wit="#Sh2">pained</rdg>
</app>word:

<app>
    <lem>deed</lem>
    <rdg wit="#Sh1 #Sh2">deede</rdg>
</app>
```

```
<app>
  <lem wit="#Sh2">Than</rdg>
  <rdg wit="#Sh1">Then</rdg>
</app> is my deede to my most
<app>
  <lem wit="#Sh1">painted</rdg>
  <rdg wit="#Sh2">pained</rdg>
</app>word:

<app>
  <rdg wit="#Sh1">Then</rdg>
  <rdg wit="#Sh2">Than</rdg>
</app> is my deede to my most
<app>
  <rdg wit="#Sh1">painted</rdg>
  <rdg wit="#Sh2">pained</rdg>
</app>word:
```

Critical Apparatus ii

The apparatus can be located anywhere as a *<listApp>*:

- in the *<body>* of the document
- in the *<back>* in other documents
- is referenced by *@loc*

→ *<rdgGrp>* aggregates several readings of a common type.

Witness list

@wit refers to descriptions in the header: *<listWit>* = list of *<witness>*-elements identified by *@xml:id* each describing a textual witness e.g. by *<bibl>*- or *<msDesc>* elements in *teiHeader*/*sourceDesc*

```
<app>
    <rdgGrp type="orthographic">
        <rdg wit="#Sh1">giue</rdg>
        <rdg wit="#Sh2">give</rdg>
    </rdgGrp>
    <rdg wit="#AS1">have</rdg>
</app>

<listWit>
    <witness xml:id="Sh1">
  <bibl>Folger STC 22276</bibl>
    </witness>
    <witness xml:id="Sh2">
  <bibl>Huntington 69304</bibl>
    </witness>
</listWit>
```

Critical Apparatus iii

There are different options:

Location referenced + external

```
<p><pb n="f13"/><lb n="f13-z1" />
Dis ouentúrlich buoch bewiset wye
von einer Frowen ge<lb/>nannt
Melusina ...</p>
<!-- ... -->
<listApp><app loc="f13-z1">
    <lem wit="#BR1">ouentúrlich</lem>
    <rdg wit="#SK1">ouentuorlich</rdg>
    <rdg wit="#AS1">abenteürlich</rdg>
</app></listApp>
```

Double endpoint + external

The apparatus is integrated into the *<body>* linked to identifiers for its beginning and its end (e.g. with *<anchor>*) but encoded anywhere (e.g. in the place where it was located in the printed source); referenced by *@from* & *@to*.

```
<p>Dis <anchor xml:id="A1"/>
ouentúrlich<anchor xml:id="A2"/> buoch
    bewiset wye von einer
    Frowen ge<lb/>nannt Melusina
    ...</p>
<!-- ... -->
<app from="#A1" to="#A2">
    <rdg wit="#SK1">ouentuorlich</rdg>
    <rdg wit="#AS1">abenteürlich</rdg>
</app>
```

Critical Apparatus iv

Location referenced + inline

```
<p n="p1">
    Dis ouentúrlich
    <app loc="p1">
        <rdg wit="#SK1">
           ouentuorlich</rdg>
        <rdg wit="#AS1">
           abenteürlich</rdg>
    </app>
    buoch bewiset wye
    von einer Frowen
    ge<lb/>nannt Melusina ...
</p>
```

Double endpoint + internal

The apparatus is integrated into the *<body>* after the referenced passage and linked an identifiers for its beginning; referenced by *@from*.

```
<p n="1">
    Dis <anchor xml:id="a"/>
    ouentúrlich<app from="#a">
        <rdg wit="#SK1">
           ouentuorlich</rdg>
        <rdg wit="#AS1">
           abenteürlich</rdg>
    </app>
    buoch bewiset wye von einer
    Frowen ge<lb/>nannt Melusina
    ...
</p>
```

Critical Apparatus v

Last but not least...

Parallel segmentation

Encoding the „base text“ as the *lem* in the *app*-element. Can be done only inline. Possibility to nest variants.

```
<p n="1">Dis
    <app>
        <lem wit="#BR1">
           ouentúrlich</lem>
        <rdg wit="#SK1">
           ouentuorlich</rdg>
        <rdg wit="#AS1">
           abenteürlich</rdg>
    </app>
    buoch bewiset wye von einer
    Frowen ge<lb/>nannt Melusina
    ...
</p>
```

Which one to choose?

1. **Referenced** imitates the classical print version, is relatively fast to create but can be imprecise in referencing
2. **Double-Endpoint** is relatively complex to encode and to process, but exact and the only form to handle overlapping structures
3. **Parallel Segmentation** can be easily processed with XSLT but not very flexible in documenting complex changes and overlapping structures

Further suggestions

Suggestions for typical problems in editing

Missing text (om.) *<rdg>* remains empty; *@cause* can contain a controlled term to describe the situation (e.g. omisit)

Additions (add.) *<lem>* remains empty

Corrections (corr. ex ...) *<rdg>* constains the complete encoding

```
<subst><del>…</del><add>…</add></subst>
```

Tools for collation

Juxta Commons Texts are reduced to flat text. Variants are encoded in the parallel segmentation method.

CollateX Creates a graph. Compares every version with the existing graph and searches for gaps.

Stemmatology in TEI

<eTree> each part of the tree which can have descendants

<eLeaf> each part of the tree, which has only ancestors

@type e.g. hypothetical, extant, lost ...

<label> for the short names („Sigla“)

<ptr> for „contaminations“ i.e. texts influenced by other manuscript traditions

TEI Critical Apparatus Toolbox

- *http://teicat.huma-num.fr/*
- by Marjorie Burghart
- **Check encoding:** consistency etc.
- **Display parallel versions.**
- **Print an edition of a TEI XML edition,** with a TEI-to-LaTeX and PDF transformation (*reledmac*! → XSL is here).
- **Annotate images:** lets you easily trace zones on an image to prepare a documentary edition (sometimes kind of buggy) → create your *<facsimile>*.
- **Get statistics** on the XML tags used in different parts of your edition plus word counts.