Handling primary sources in TEI XML

TEI @ Oxford

September 2008

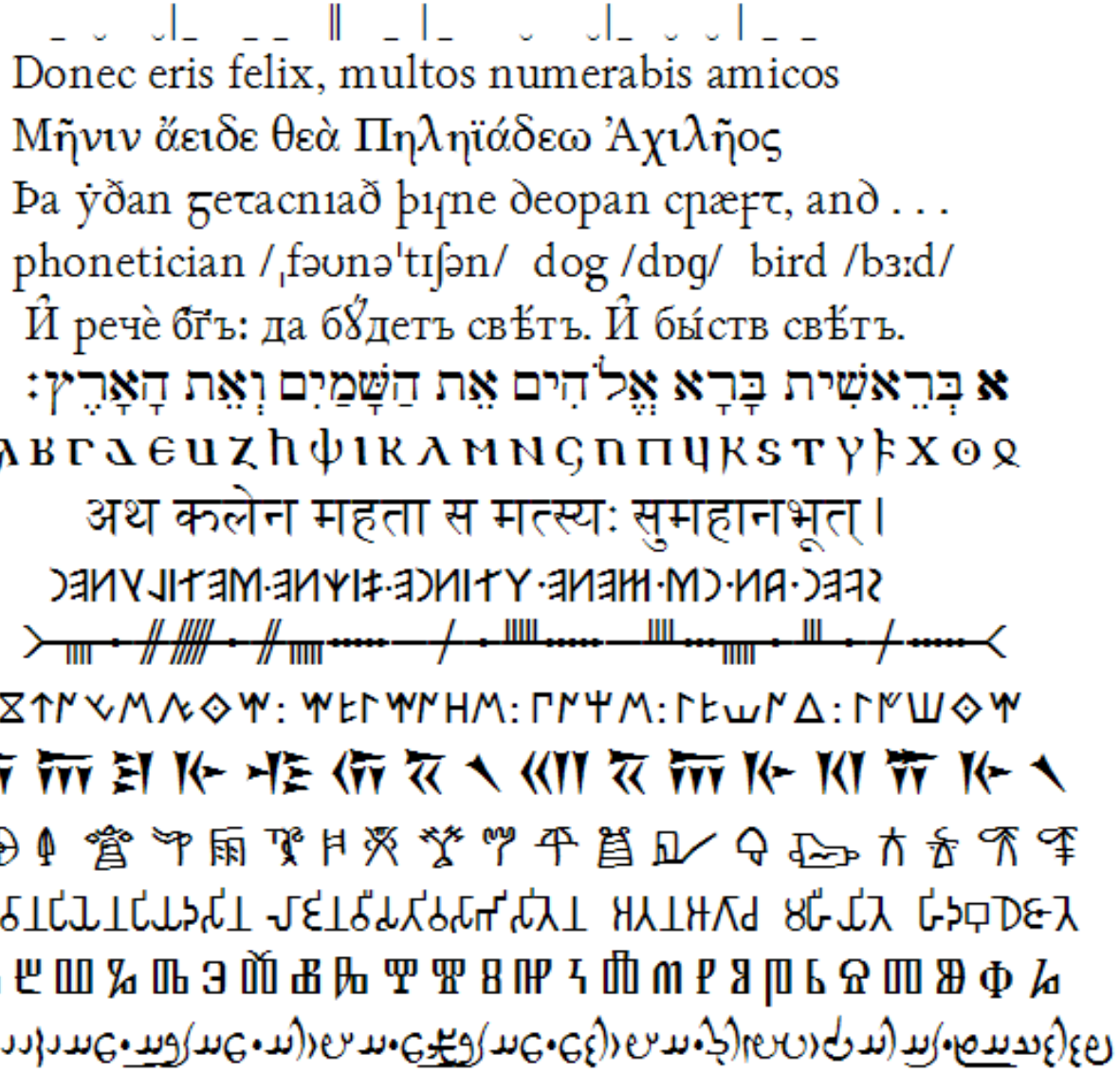
Transcribable features

Transcription is a special kind of encoding, in which the aim is to represent all the important features of a primary source without prejudging too much about it... hence the term diplomatic transcript.

Here are some of the kinds of features concerned:

- letter forms
- page layout
- orthography
- word division
- punctuation
- abbreviations
- additions and deletions
- errors and omissions

Letter forms



Letter forms

- Unicode (ISO 10646) defines computer codepoints for most, though not all, of the abstract characters recognized by modern scholars when reading ancient sources.
- Different fonts realise those codepoints in different styles; however the underlying character remains the same.
- Data entry of Unicode characters can be
  - direct: some key combination or menu-selection generates the character æ for us
  - indirect, using a numeric character entity reference such as `&#xE6`
  - indirect using a mnemonic character entity reference such as `&aelig;` (this requires every document to carry a DTD with it)

Non-Unicode characters

Nevertheless, sometimes Unicode is not enough…

- … if your character doesn't exist
- … if you want to distinguish letter forms that Unicode regards as identical e.g. for statistical analysis.

The <g> (gaiji) element stands for any non-Unicode character. Its content can be a local approximation to the desired letter (or nothing); its *@ref* attribute points to a definition for the required character or glyph.

```
<!-- in text --><g ref="#x123"/>
```

or

```
<g ref="#x123">x</g>
```

in header:

```
<char xml:id="x123">
<!-- character definition here -->
</char>
```

Structure and layout

- As elsewhere we distinguish ‘structure’ (the way the intellectual content of a work is logically organized) from ‘layout’ (the physical arrangement of the text on the page).
- The structural view is generally privileged over the layout view in TEI documents. Common practice is to mark `<div><p>`, `<lg>`, `<l>` (etc) elements, elements, as in printed texts, and to use empty ‘milestone’ tags for significant points in the physical layout, for example `<pb/>`, `<cb/>`, and `<lb/>`, for page-, column- and line-boundaries respectively.
- (The opposite practice is also feasible: one could imagine marking up a structural hierarchy of `<gathering>`, `<leaf>`, etc. with milestone elements to mark the points at which ‘structural’ components begin and end.)

Abbreviation

Abbreviations are highly characteristic of manuscript materials of all kinds. Western MSS traditionally distinguish:

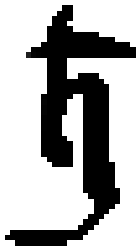
Suspensions	the first letter or letters of the word are written, generally followed by a point, or other marker: for example e.g. for exempla gratia
Contractions	both first and last letters are written, generally with some other mark of abbreviation such as a superscript stroke, or, less commonly, a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

Encoding abbreviations (1)

TEI proposes two levels of encoding:

- the whole of an abbreviated word and the whole of its expansion: <abbr> and <expan>
- abbreviatory signs or characters and the 'invisible' characters they imply: <am> and <ex>

The Old Icelandic word *hann* ('he') is usually written as a brevigraph, combining the letter h with a horizontal stroke representing nasalisation (Unicode character 0305, functionally similar to the modern tilde). It looks like this:



Encoding abbreviations (2)

Depending on editorial policy, we might represent this combination in any one of the following ways:

```
<abbr>h&#x305;</abbr>
```

```
<expan>hann</expan>
```

```
h<am>&#x305;</am>
```

```
h<ex>ann</ex>
```

```
<abbr>h<am>&#x305;</am>
</abbr>
```

```
<expan>h<ex>ann</ex>
</expan>
```

Encoding abbreviations (3)

We could also indicate multiple alternatives (at either level) by using the <choice> element

```
h<choice>
  <am>&#x305;</am>
  <ex>ann</ex>
</choice>
<choice>
  <abbr>h&#x305;</abbr>
  <expan>hann</expan>
</choice>
```

And much more besides...

Encoding abbreviations (3)

The *@type* attribute on <abbr> allows us to provide alternative renderings for the same markup in different contexts.

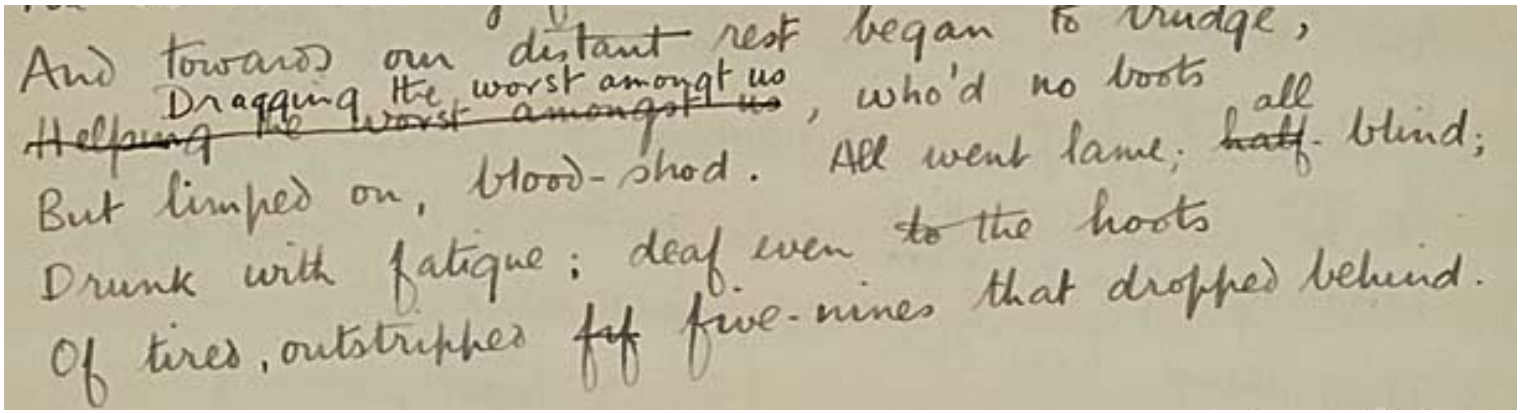
```
<choice>
 <abbr type="susp">k<am>&#x307;</am>
 </abbr>
 <expan>k<ex>onungr</ex>
 </expan>
</choice>
<choice>
 <abbr type="tittel">ml<am>&#x305;</am>i</abbr>
 <expan>m<ex>æl</ex>l<ex>t</ex>i</expan>
</choice>
```

*k(onungr) m**æll**t**i*

As elsewhere, the *@resp* and *@cert* attributes can also be used to indicate who is responsible for an expansion, and the degree of certainty attached to it.

Additions, deletions, and substitutions

- <add> (addition) or <del> (deletion) are used for evident alterations in the source
- a combined addition and deletion may be marked using <subst> (substitution)



Additions, deletions, and substitutions


```
<l>And towards our distant rest began to trudge,</l>
<l>
  <subst>
    <del>Helping the worst amongst us</del>
    <add>Dragging the worst amongt us</add>
  </subst>, who'd no boots
</l>
<l>But limped on, blood-shod.  All went lame;
<subst>
    <del>half-</del>
    <add>all</add>
  </subst> blind;</l>
<l>Drunk with fatigue ; deaf even to the hoots</l>
<l>Of tired, outstripped <del>fif</del> five-nines that dropped behind.</l>
```

Corrections and emendations

The <sic> element can be used to indicate that the reading of the manuscript is erroneous or nonsensical, while <corr> (correction) can be used to provide what in the editor's opinion is the correct reading:

```
<sic>giorit</sic>
```

```
<corr>giorir</corr>
```

Alternatively, they may be combined within a <choice> element, thus allowing the possibility of providing multiple corrections:

```
<choice>
 <sic>giorit</sic>
 <corr cert="high">giorir</corr>
 <corr cert="low">gioret</corr>
</choice>
```

Supplied text

Sometimes, a transcript may need to include words not visibly present in the source:

- because the carrier has been damaged or is barely legible
- because of (assumed) scribal error

The `<supplied>` element is provided for use in either situations; the *@reason* attribute is used to distinguish them.

```
…Dragging the worst
among<supplied reason="omitted">s</supplied>t us…
```

Metadata for supplied text

Attributes *@resp* and *@cert* can be used here as elsewhere. A *@source* attribute is also available to indicate that another witness supports the reconstruction:

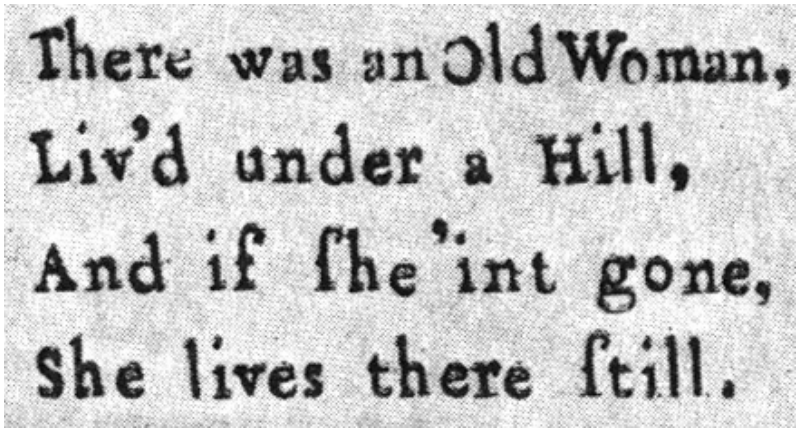
```
<p>ath þeir <supplied reason="omitted" source="AM02-152">mundu</supplied>
sundr ganga</p>
```

When missing text cannot be confidently reconstructed, the <gap> element should be used. Its *@reason* attribute explains the reason for the omission and its *@extent* attribute indicates its presumed size.

```
<gap reason="damage" extent="7cm"/>
```

Normalization

Source texts rarely use modern normalized orthography. For retrieval and other processing reasons, such information may be useful in a transcription. The <reg> (regularized) element is available used to mark a normalized form; the <orig> (original) element to indicate a non-standard spelling. These elements can optionally be grouped as alternatives using the <choice> element:



Normalization example

```
<lg>
 <l>There was an Old Woman,</l>
 <l>
  <choice>
   <orig>Liv'd</orig>
   <reg>Lived</reg>
  </choice> under a hill,</l>
 <l>And if she <orig>'int</orig> gone,</l>
 <l>She lives there still.</l>
</lg>
```

Why are manuscript descriptions special?

- Manuscripts are *unique objects*, sometimes (though not always) of great cultural or political value
- Books, by contrast, exist in multiple copies, and can be described adequately by well-established and formalized bibliographic conventions.
- For manuscripts, there are several traditions, often descriptive or belle lettriste, and little consensus.

Similar concerns apply to other text-bearing objects.

Objectives

The TEI <msDesc> element is intended for several different kinds of applications:

- standalone database of library records (finding aid)
- discursive text collecting many records (catalogue raisonné)
- metadata component within a digital surrogate (electronic edition)
- tool for ‘quantitative codicology’

Catalogue Raisonné

An <msDesc> can appear anywhere a <p> paragraph can

```
<div>
 <head>The Arnamagnæan Institute and its records</head>
 <p>Probably the finest collection of ……
 </p>
 <p>For example:  </p>
 <msDesc xml:id="AMI-1" xml:lang="en">
<!-- …-->
 </msDesc>
 <p>In the following manuscript….
 </p>
 <msDesc xml:id="AMI-2" xml:lang="en">
<!-- …-->
 </msDesc>
</div>
```

Digital edition

- metadata in the header
- transcription in the body, with links to
- images in a `<facsimile>` element

```
<TEI>
 <teiHeader>
<!-- … metadata describing the manuscript -->
<!-- includes a msDesc within the sourceDesc -->
 </teiHeader>
 <facsimile>
<!-- … metadata describing the digital images -->
 </facsimile>
 <text>
<!-- (optional) transcription of the manuscript -->
 </text>
</TEI>
```

Example minimal structure


```
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>[Title of manuscript]</title>
  </titleStmt>
  <publicationStmt>
   <distributor>[name of data provider]</distributor>
   <idno>[project-specific identifier]</idno>
  </publicationStmt>
  <sourceDesc>
   <msDesc xml:id="ex1" xml:lang="en">
<!-- [full manuscript description ]-->
   </msDesc>
  </sourceDesc>
 </fileDesc>
 <revisionDesc>
  <change when="2008-01-01">[revision
information]</change>
 </revisionDesc>
</teiHeader>
```

Quantitative Codicology: is it possible?

Two conflicting desires:

- preserve (or perpetuate) existing descriptive prose
- reliable search, retrieval, and analysis of data

The <msDesc> tries, wherever possible, to have its cake and eat it.

Components of a manuscript description

We separate, and tag differently, aspects concerned with…

- identification
- intellectual content
- physical description
- history and curation
- … and other manuscript descriptions

msDesc structure


```
<msDesc xml:id="ex2" xml:lang="en">
 <msIdentifier>
<!-- Repository location, shelfmarks, etc. -->
 </msIdentifier>
 <msContents>
<!-- Structured description of MS contents -->
 </msContents>
 <physDesc>
<!-- Physical and codicological description -->
 </physDesc>
 <history>
<!-- Origin, provenance, acquisition, etc. -->
 </history>
 <additional>
<!-- Additional bibliographic and curatorial information,
and associated materials etc. -->
 </additional>
 <msPart>
<!--Composite manuscript details -->
 </msPart>
</msDesc>
```

<msIdentifier> is the only one that is required.

Simple example <msDesc>


```
<msDesc xml:id="ex3" xml:lang="en">
  <msIdentifier>
    <settlement>Oxford</settlement>
    <repository>Bodleian Library</repository>
    <idno>MS. Add. A. 61</idno>
    <altIdentifier type="other">
      <idno>28843</idno>
    </altIdentifier>
  </msIdentifier>
  <p>In Latin, on parchment: written in more than one hand of the 13th
    cent. in England: 7¼ x 5⅜ in., i + 55 leaves, in double columns: with
    a few coloured capitals.</p>
  <p>'Hic incipit Bruitus Anglie,' the De
    origine et gestis Regum Angliae of Geoffrey of Monmouth (Galfridus
    Monumetensis: beg. 'Cum mecum multa &amp;amp; de multis.'</p>
  <p>On fol. 54v very faint is 'Iste liber est fratris guillelmi de
    buria de … Roberti ordinis fratrum Pred[icatorum],' 14th cent. (?):
    'hanauilla' is written at the foot of the page (15th cent.). Bought
    from the rev. W. D. Macray on March 17, 1863, for £1 10s.</p>
</msDesc>
```

Structured form of <msDesc> (1)


```
<msDesc xml:id="ex4" xml:lang="en">
  <msIdentifier>
    <settlement>Oxford</settlement>
    <repository>Bodleian Library</repository>
    <idno>MS. Add. A. 61</idno>
    <altIdentifier type="internal">
      <idno>28843</idno>
    </altIdentifier>
  </msIdentifier>
  <msContents>
    <msItem>
      <author xml:lang="en">Geoffrey of Monmouth</author>
      <author xml:lang="la">Galfridus Monumetensis</author>
      <title type="uniform" xml:lang="la">De origine et gestis Regum
Angliae</title>
      <rubric xml:lang="la">Hic incipit Bruitus Anglie</rubric>
      <incipit xml:lang="la">Cum mecum multa & de multis</incipit>
      <textLang mainLang="la">Latin</textLang>
    </msItem>
  </msContents>
<!-- … -->
</msDesc>
```

Structured form of <msDesc> (2)

```
<physDesc>
  <objectDesc  form="codex">
    <supportDesc  material="perg">
      <support>
        <p>Parchment.</p>
      </support>
      <extent>i + 55 leaves <dimensions  scope="all"  type="leaf"  unit="in">
          <height>7 ¼</height>
          <width>5 ⅜</width>
        </dimensions>
      </extent>
    </supportDesc>
    <layoutDesc>
      <layout  columns="2">
        <p>In double columns.</p>
      </layout>
    </layoutDesc>
  </objectDesc>
  <handDesc>
    <p>Written in more than one hand.</p>
  </handDesc>
  <decoDesc>
    <p>With a few coloured capitals.</p>
  </decoDesc>
</physDesc>
```

Structured form of <msDesc> (2)

```
<history>
  <origin>
    <p>Written in <origPlace>England</origPlace> in the
<origDate notAfter="1300" notBefore="1200">13th cent.</origDate>
    </p>
  </origin>
  <provenance>
    <p>On fol. 54v very faint is <quote xml:lang="la">Iste liber est
          fratris guillelmi de buria de
      <gap reason="illegible"/> Roberti ordinis
          fratrum Pred<ex>icatorum</ex>
      </quote>, 14th cent. (?):
    <quote>hanauilla</quote> is written at the foot of
        the page (15th cent.).</p>
  </provenance>
  <acquisition>
    <p>Bought from the rev. <name type="person" key="MCRAYWD">W. D.
Macray</name> on
    <date when="1863-03-17">March 17,
          1863</date>, for £1 10s.</p>
  </acquisition>
</history>
```

Identification (1)

The <msIdentifier>

Traditional three part specification:

- place (<country>, <region>, <settlement>)
- repository (<institution>, <repository>)
- identifier (<collection>, <idno>)

```
<msIdentifier>
  <country>France</country>
  <settlement>Troyes</settlement>
  <repository>Bibliothèque Municipale</repository>
  <idno>50</idno>
</msIdentifier>
```

Identification (2)

Alternative or additional names can also be included:

```
<msIdentifier>
  <country>Danmark</country>
  <settlement>København</settlement>
  <repository> Det ArnamagnæanskeInstitut </repository>
  <idno>AM 45 fol.</idno>
  <msName  xml:lang="la">Codex Frisianus</msName>
  <msName  xml:lang="is">Fríssbók</msName>
</msIdentifier>
```

Intellectual Content

- May simply use paragraphs of text…
- … or a tree of <msItem> elements
- … optionally preceded by a prose summary

We can describe the content in general terms:

```
<msContents>
  <p>An extraordinary charivari of heroic deeds and
    improving tales, including an early version of
  <title>Guy of Warwick</title> and several hymns.
  </p>
</msContents>
```

or we can provide detail about each distinct item:

```
<msContents>
  <summary>An extraordinary charivari of heroic deeds,
    improving tales, and hymns</summary>
  <msItem>
<!-- details of Guy of Warwick here -->
  </msItem>
  <msItem>
<!-- other items here -->
  </msItem>
</msContents>
```

The <msItem> element

Manuscripts contain identifiable items, usually physically tied to a locus.

- <locus>, if present, must be given first
- then any of the following, in a specified order:
  - <author>, <respStmt>
  - <title>, <rubric>, <incipit>, <explicit>, <colophon>, <finalRubric>
  - <quote>, <textLang>, <decoNote>, <bibl>, <listBibl>, <note> …
  - … or nested <msItem>s

<msContents> with multiple <msItem>s

```
<msContents>
  <msItem n="1">
    <locus>fols. 5r-7v</locus>
    <title>An ABC</title>
    <bibl>
      <title>IMEV</title>
      <biblScope type="pages">239</biblScope>
    </bibl>
  </msItem>
  <msItem n="2">
    <locus>fols. 7v-8v</locus>
    <title xml:lang="fr">Lenvoy de Chaucer a Scogan</title>
    <bibl>
      <title>IMEV</title>
      <biblScope type="pages">3747</biblScope>
    </bibl>
  </msItem>
<!-- … -->
  <msItem n="6">
    <locus>fols. 14r-126v</locus>
    <title>Troilus and Criseyde</title>
    <note>Bk. 1: 71-Bk. 5: 1701, with additional losses due to mutilation
        throughout</note>
  </msItem>
</msContents>
```

Physical Description

An artificial (but helpful) grouping of many distinct items.
You can simply supply paragraphs of prose, covering such topics as

- <objectDesc>: the physical carrier
- <handDesc>: what is carried on it
- <musicNotation>, <decoDesc>, <additions>
- <bindingDesc> and <sealDesc>
- <accMat>: accompanying material

Or, group your discussion within the specific elements mentioned above.
Similarly, within the specific elements, you can supply paragraphs of prose, or further specific elements.

The carrier 1

The <objectDesc> contains just paragraphs, or <supportDesc> and <layoutDesc>

```
<objectDesc form="codex">
  <supportDesc material="mixed">
    <p>Early modern <material>parchment</material> and
    <material>paper</material>. </p>
  </supportDesc>
  <layoutDesc>
    <layout columns="1" ruledLines="25 32"/>
  </layoutDesc>
</objectDesc>
```

The carrier 2

A more complex substructure with specific elements for <support>, <extent>, <foliation>, <collation>, <condition>.

Multiple layouts may also be specified:

```
<layoutDesc>
  <layout ruledLines="25" columns="1">
    <p>
      <locus from="1r-202v"/>
      <locus from="210r-212v"/>
        Between 25 and 32 ruled lines.</p>
  </layout>
  <layout ruledLines="34 50" columns="1">
    <p>
      <locus from="203r-209v"/>Between 34 and 50 ruled lines.</p>
  </layout>
</layoutDesc>
```

<handDesc> and <decoDesc>

- <handNote> (note on hand) describes a particular style or hand distinguished within a manuscript.
- <decoNote> contains a note describing either a decorative component of a manuscript, or a fairly homogenous class of such components.

<handDesc> examples


```
<handDesc hands="2">
  <p>The manuscript is written in two contemporary hands, otherwise
     unknown, but clearly those of practised scribes.  Hand I writes
     ff.  1r-22v and hand II ff.  23 and 24.  Some scholars, notably Verner
     Dahlerup and Hreinn Benediktsson, have argued for a third hand on
     f.  24, but the evidence for this is insubstantial.</p>
</handDesc>
```

```
<handDesc hands="3">
  <handNote xml:id="Eirsp-1" scope="minor" script="other">
    <p>The first part of the manuscript, <locus from="1v" to="72v:4">fols
1v-72v:4</locus>, is written in a
       practised Icelandic Gothic bookhand.  This hand is not
       found elsewhere.</p>
  </handNote>
  <handNote xml:id="Eirsp-2" scope="major" script="other">
    <p>The second part of the manuscript,
<locus from="72v:4" to="194v">fols 72v:4-194</locus>, is
       written in a hand contemporary with the first; it can
       also be found in a fragment of <title>Knýtlinga
          saga</title>, <ref>AM 20b II fol.</ref>.</p>
  </handNote>
  <handNote xml:id="Eirsp-3" scope="minor" script="other">
    <p>The third hand has written the majority of the
       chapter headings.  This hand has been identified as the
       one also found in <ref>AM 221 fol.</ref>.</p>
  </handNote>
</handDesc>
```

<additions>

The <additions> element can be used to list or describe any additions to the manuscript, such as marginalia, scribblings, doodles, etc., which are considered to be of interest or importance.

```
<additions>
 <p>The text of this manuscript is not interpolated with
   sentences from Royal decrees promulgated in 1294, 1305
   and 1314.  In the margins, however, another somewhat
   later scribe has added the relevant paragraphs of
   these decrees, see pp.  8, 24, 44, 47 etc.</p>
 <p>As a humorous gesture the scribe in one opening of
   the manuscript, pp.  36 and 37, has prolonged the lower
   stems of one letter f and five letters þ and has them
   drizzle down the margin.</p>
</additions>
```

<accMat>

<accMat> (accompanying material) contains details of any significant additional material which may be closely associated with the manuscript being described, such as non-contemporaneous documents or fragments bound in with the manuscript at some earlier historical period.

```
<accMat> A copy of a tax form from 1947 is included in
the envelope with the letter.  It
is not catalogued separately.  </accMat>
```

History

- <origin>: where it all began
- <provenance>: everything in between
- <acquisition>: how you acquired it

<origin> is datable element and thus has attributes notBefore and notAfter

Example


```
<history>
 <origin>
  <p>Written in <origPlace>England</origPlace> in the
  <origDate notAfter="1300" notBefore="1200">13th
        cent.  </origDate>
  </p>
 </origin>
 <provenance>
  <p>On fol.  54v very faint is <q>Iste liber
        est fratris guillelmi de buria de
<gap reason="illegible"/>
        Roberti ordinis fratrum
Pred<expan>icatorum</expan>
   </q>,
     14th cent.  (?):  <q>hanauilla</q> is written at the
     foot of the page (15th cent.).</p>
 </provenance>
 <acquisition>
  <p>Bought from the rev.  <name type="person">W.  D.
Macray</name>on <date when="1863-03-17"> March 17,
1863</date>,
      for 1pound 10s.</p>
 </acquisition>
</history>
```

Additional information

- `<adminInfo>`: administrative information
- `<surrogates>`: information about other surrogates eg pictures
- `<accMat>`: accompanying material
- `<listBibl>`: bibliography

Administrative information

- record history
- availability
- custodial history
- miscellaneous remarks

```
<adminInfo>
  <recordHist>
    <source>
      <p>Information transcribed from <ref  target="IMEV123">IMEV 123</ref>
      </p>
    </source>
  </recordHist>
  <custodialHist>
    <custEvent  type="conservation"  notBefore="1961-03"  notAfter="1963-
02">
      <p>Conserved between March 1961 and February 1963 at Birgitte Dalls
           Konserveringsværksted.</p>
    </custEvent>
    <custEvent  type="photography"  notBefore="1988-05-01"  notAfter="1988-
05-30">
      <p>Photographed in May 1988 by AMI/FA.</p>
    </custEvent>
    <custEvent  type="other"  notBefore="1989-11-13"  notAfter="1989-11-13">
      <p>Dispatched to Iceland 13 November 1989.</p>
    </custEvent>
```

And finally

A <msDesc> can contain a nested <msDesc>, <msPart>, catering for a combination of two MSS, formerly distinct.

```
<msDesc xml:id="ex5" xml:lang="en">
  <msIdentifier>
    <msName xml:lang="la">Codex Suprasliensis</msName>
  </msIdentifier>
  <msPart>
    <altIdentifier type="partial">
      <settlement>Ljubljana</settlement>
      <repository>Narodna in univerzitetna knjiznica</repository>
      <idno>MS Kopitar 2</idno>
      <note>Contains ff. 10 to 42 only</note>
    </altIdentifier>
  </msPart>
  <msPart>
    <altIdentifier type="partial">
      <settlement>Warszawa</settlement>
      <repository>Biblioteka Narodowa</repository>
      <idno>BO 3. 201</idno>
    </altIdentifier>
  </msPart>
  <msPart>
    <altIdentifier type="partial">
      <settlement>Sankt-Peterburg</settlement>
      <repository>Rossiiskaia natsional'naia biblioteka</repository>
      <idno>Q. p. I. 72</idno>
    </altIdentifier>
  </msPart>
</msDesc>
```