Encoding Texts

Markup, XML and oXygen

Dr Arianna Ciula
Dr Elena Pierazzo

London 1 May 2014
Encoding Texts

- Markup
  - Principles
  - Terminology
- XML
  - Concepts
  - Practicalities
  - Editing using oXygen
- Exercises
Text Structure

- Text is not an undifferentiated sequence of words

- Depending on the objectives, it can be divided in multiple units of different type and size
<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Dr</td>
</tr>
<tr>
<td>Forename</td>
<td>Ciula</td>
</tr>
<tr>
<td>Surname</td>
<td>Arianna</td>
</tr>
<tr>
<td>Affiliation</td>
<td>King’s College London</td>
</tr>
</tbody>
</table>
The SICK ROSE

O Rose thou art sick.
The invisible worm, 
That flies in the night 
In the howling storm: 
Has found out thy bed 
Of crimson joy: 
And his dark secret love 
Does thy life destroy.
Text Representation

- **Input**
  - digitization, imaging, OCR

- **Text analysis**
  - Concordance making, text retrieval, lemmatizing, collation, statistical analysis, stylometry, stemma (re)construction

- **Output**
  - typesetting, visualization, hypertext

Uniform system for text encoding and interchange

(Vanhoutte 2004)
Text Representation and Accessibility

- Workflow of editorial process
- Use and re-use of content
- Data recovery

Uniform system for text encoding and interchange

(Vanhoutte 2004)
What is Markup? (OED)

- 2. b. The process or result of marking corrections on copy or proofs in preparation for printing.

- c. Computing. The process of embedding tags in an electronic text so as to distinguish the text's logical, syntactic, or structural components; the tags so embedded.
No markup means problems!

The merchant of Venice by William Shakespeare

No markup means problems!
Markup Principles and Terminology

- Every writing is a kind of markup
- Make explicit (to a machine) what is implicit (to a person)
- Supply multiple annotations/interpretations

- 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 mark-up schemes, requirements for a rigorous and documented encoding
1. **Procedural**
   - Direct instructions to execute a visualization procedure (focus on formatting and blind to content)
   - Often system dependent, mainly proprietary (portability and preservation issues)
   - Useful in the right context

2. **Descriptive**
   - Encoding of content as well as structure (focus on interpretation, on the roles played by textual components – domain specific encoding)
   - Human readable
   - Software and platform independent
   - Hierarchical tree structure (possible to apply rules of validation)
   - Formatting applied later (interchange between applications and formats)
“Come here, Mr. Watson. I need you!”

“Come here, Mr. Watson. I need you!”

“Come here, Mr. Watson. I need you!”

“Come here, Mr. Watson. I need you!”
SGML, HTML, XML

- **SGML** *Standard Generalized Markup Language*
  - Possible to create new tagsets
  - ISO standard in 1986
  - Metalanguage (rules and procedures to define syntax and lexicon of languages)

- **HTML** *Hypertext Markup Language*
  - Fixed tagset
  - Markup scheme

- **XML** *eXtensible Markup Language*
  - *Simplified and improved* SGML (1998, W3C)
  - Well formed and valid documents
  - Metalanguage
- Main structure of the languages based on SGML

- Contain anything that can be viewed as a structural component of a text

“Structure” could mean here:
a paragraph, a section but also a word or a single character or even the space between two words
Elements are delimited by **tags** and enclose **content**

```
<element>content</element>
```
Relations between Elements

In SGML based languages relations among elements are:

- Hierarchical
- Ordinal

Relations among elements can be described as family relations (parents, ancestor, children, siblings...)

16
Semantic-hierarchical Markup

- Document parts identified by their function or meaning and by their context or position in a hierarchical structure

- Document seen as Tree format
  - One single root and multiple nodes
  - A node can be
    - A sub-tree
    - A single element (possibly with attributes)
    - A string of characters
Element Content

- String of characters
  `<para>Some text</para>`

- A sub-tree (other elements)

  `<bibl>`
  `<author>C. Collodi</author>`
  `<title>Le avventure di Pinocchio</title>`
  p. 23-34
  `</bibl>`

- Mixed content
Nesting of Elements

\[<b><i>Some text</i></b>\]
Attributes

- a type of a category
- a characteristic of something
- label or count something
- make explicit what is implicit

```xml
<city name="Montevideo"
    type="capital"
    country="Uruguay"/>
```
XML Focus on Meaning

### HTML

<i>Was the Novel a Spanish Invention?</i>

<i>mens sana in corpore sano</i>

### XML

<title>Was the Novel a Spanish Invention?</title>

<foreign>mens sana in corpore sano</foreign>
XML Focus on Explicit Meaning

<p><person name="Mendoza, Eduardo">E. Mendoza</person></p>

<p><person name="Mendoza, Eduardo">the author of ‘La ciudad de los prodigios’</person></p>

<p><person name="Mendoza, Enrique">E. Mendoza</person></p>
XML Separation Content/Design

- Marked up source data is independent of any given *rendition*
- Design applied later in separate process
- Same document to create different presentational formats for different media
XML for different Media and Formats

- Print
  - Body of text
  - Indices
  - ...

- Digital
  - Linear text in HTML
  - Linear text in PDF
  - Decomposed text
  - Indices
  - ...

From same XML source
XML Document Type
Definition or Schema

- A good thing!
- Vocabulary and grammar of a specific XML language
- Defines which elements, attributes and entities are allowed in a specific XML document
- Defines how to use elements and attributes (required, optional, repeatable, types...)
Ensure that your documents use only predefined elements, attributes, and entities. Enforce structural rules such as 'recipes must include an ingredient list' or 'every chapter must begin with a heading'. Make sure that the same thing is always called by the same name. Schema languages vary in the amount of validation they support.
XML Levels of Correctness

- Documents must be *well-formed*
- Documents must be *valid*
- Documents must be *consistent*

Encoding is not mechanical!
XML Editing

- You can edit XML in any plain text editor (e.g. Notepad)
- But an XML editor is XML aware
  - helps in editing (e.g. by completing tags)
  - is context-aware
  - checks whether a file is valid and well-formed

  - one of the best
  - relatively easy to set up and to use
  - very powerful
  - cross-platform
  - relatively cheap
  - developers keen in digital humanities needs
oXygen Opening Window
<line>Nobie, simple, all conditions:</line>
<line>Worth beneath a threadbare cover,</line>
<line>Villainy bedaubed all over.</line>

</stanza>

<stanza>
<line>Women black, red, fair and grey,</line>
<line>Prudes and such as never pray,</line>
<line>Handsome, ugly, noisy, still,</line>
<line>Some that will not, some that will.</line>
</stanza>

</stanza>
oXygen Attributes and Elements Panel

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/css" href="rec.css"?>
<!DOCTYPE recipe-collection SYSTEM "rec.dtd">
<recipe-collection>
  <recipe author="pjs" id="num1">
    <title>Hummus</title>
    <ingreds>
      <ing>2 cups chick peas</ing>
      <ing>1/2 cup lemon juice</ing>
      <ing>1/2 cup water</ing>
      <ing>12 cloves of garlic</ing>
      <ing>1 teaspoon cumin</ing>
      <ing>Dash of salt</ing>
      <ing>1 cup tahini</ing>
      <ing>Olive oil</ing>
      <ing>Dried parsley</ing>
      <ing>Paprika</ing>
    </ingreds>
    <steps>
      <p>Put the chick peas, lemon juice, water, garlic, cumin, salt, tahini, olive oil, parsley, and paprika in a blender or food processor and grind them up. Then add the tahini and run the blender again. This will probably need to add some water (typically not more than a few tablespoons) to get it smooth. Take a taste and adjust to taste.</p>
  </recipe>
</recipe-collection>
```
Exercise 1

Open a file and save it with another name

1. Click or select File → Open
2. Go to the mmsda_examples/xml/exercises folder saved in your Desktop
3. Choose the sample XML file in the poem folder
4. Save the file with another name within the same folder
5. Open the file within a browser (Firefox, Mozilla, Explorer or Safari) and look at the rendered output
6. Take a few minutes to get to know each of the editing panels.
<poems>

<poem>
<title>A Description of London</title>
<author>John Bancks</author>

<stanza>[5 lines]</stanza>

<stanza>
<line>Gaudy things enough to tempt ye,</line>
<line>Showy outsides, insides empty;</line>
<line>Bubbles, trades, mechanic arts,</line>
<line>Coaches, wheelbarrows and carts.</line>
</stanza>
Exercise 2

Use the basic editing features for elements

1. Fold the content of the first 'poem' element and then reveal it again
2. Delete the second poem
3. Create a new poem after the poem by Walt Whitman and add your own text!
Centre for Computing in the
Centre for Computing in the <hi rend=""">development of print original or surrogates.

automating the capturing of print originals or surrogates.

Element
oXygen Tagging a string of text

She made her initial London appearance in important notices were for the role of <name type="fictional">Jew Suss</name> (1929).

Her appearance in <name>Victoria</name> established <name type="fictional">Cleopatra</name>, <name type="fictional">Mary Stuart</name>, and <name type="fictional">Elizabeth</name>. Her portrayal of <name type="real">Shakespeare</name>...
oXygen Validate document function

An XML Schema annotation is not available for the current selection.
Exercise 3

Copying and Pasting

1. Open the 'recipes' example at mmsda_examples/xml/exercises/recipe in oXygen
2. Open the same folder in Windows Explorer and open 'recipes-to-add.txt': this is the file with the text that you will be pasting in
3. Add the new recipes to the XML file by adding the required elements
4. Validate the document
   - Select menu Document → XML Document → Validate document (Ctrl +Shift+V) or click the toolbar button “Validate document”
1. Associate a CSS

2. Open the Author View

Put the chick peas, lemon juice, water, garlic, grind them up. Then add the tahini and run the...
oXygen Using the Author view
Exercise 4
The Author view

1. Open the recipe.xml file in Author view and change the display option, first as Full Tags, and to Partial tags.
2. Add a new Ingredient and a new step to your recipe.
Exercise 5
Editing Attributes

1. Add attributes to each recipe. There are three different attribute types here. How are they different?
2. Add imaginary links to each of the recipes (after 'steps') What else do you have to add?
   e.g.
   
   `<link url="http://www.google.com">Google</link>`

3. Validate the document
References


TEI: Text Encoding Initiative

Dr Elena Pierazzo
TEI is 27 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....
What is the TEI for?

- Interchange of data
- Platform independence
- Global digital library (hundreds of thousands of texts all over the world)
- Open to new tools
- Preservation
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
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 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)
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 to 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
Roma: generating validators for the TEI

- You chose your template
- You choose your modules
- You create new elements
- You save your customization
- You download your schema
TEI Modules: basics

- textstructure
  - defines all named element classes and macros
  - basic “book-like” structure for prose, verse, drama
- core
  - ‘core’ elements “common to all kinds of text”
- header
  - teiHeader
Basic text structure

- `<TEI>` the root element
- Every TEI document contains:
  - `<teiHeader>`
  - `<text>`
Medieval/Modern Manuscript Studies in the Digital Age
Medieval/Modern Manuscript Studies in the Digital Age
Medieval/Modern Manuscript Studies in the Digital Age
<teiHeader>
<fileDesc>
<titleStmt>
<title>XML training test</title>
<author>Author's name</author>
</titleStmt>
<publStmt>
<distributor>Centre for Computing in the Humanities</distributor>
<address>Strand, London WC2R 2LS</address>
</publStmt>
<sourceDesc>
<p>No source: created in electronic format</p>
</sourceDesc>
</fileDesc>
<revisionDesc>
<change>
<date>2007-05-10</date>
<name>EP</name>
<desc>File created</desc>
</change>
</revisionDesc>
</teiHeader>
Medieval/Modern Manuscript Studies in the Digital Age
To our three selves

Author's Note

All the characters in this book are purely imaginary, and if the author has used names that may suggest a reference to living persons she has done so inadvertently.

...
TEI block level elements (start in a new line)

- Divisions
- Headings
- Paragraphs
- Lists
- Tables
- Verses
Divisions and Heading

<div type="section">
  <head>heading</head>
  <div type="subsection">
    <head>heading</head>
    <!-- TEXT HERE -->
  </div>
</div>

<div type="subsection">
  <head>heading</head>
  <!-- TEXT HERE -->
</div>
Paragraphs

<p>Some text</p>
<list type="ordered">
    <item>Apples</item>
    <item>Oranges</item>
    <item>Lemons</item>
</list>
<table>
<thead>
<tr>
<th>Name</th>
<th>Lucia</th>
</tr>
</thead>
</table>

Tables 1 (i)
### Tables 1 (ii)

<table>
<thead>
<tr>
<th>Name</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lucia</td>
<td>Spain</td>
</tr>
</tbody>
</table>
The SICK ROSE

O Rose thou art sick.
The invisible worm,
That flies in the night
In the howling storm:
Has found out thy bed
Of crimson joy:
And his dark secret love
Does thy life destroy.
<head>The SICK ROSE</head>

<lg>
  <l>O Rose thou art sick. </l>
  <l>The invisible worm, </l>
  <l>That flies in the night </l>
  <l>In the howling storm: </l>
  <l>Has found out thy bed </l>
  <l>Of crimson joy: </l>
  <l>And his dark secret love </l>
  <l>Does thy life destroy.</l>
</lg>
Exercise

1. Go to the `mmsda_examples/tei/exercises` folder saved in your Desktop and open `Ex1.doc` within MS Word
2. Take a few minutes to examine the structure
3. Go to the `mmsda_examples/tei/exercises` folder and open the `ex1.xml` file within oXygen
4. Copy and paste the text from the Word document and add the relevant elements (`<div>`, `<head>`, `<p>...`)
Phrase-level elements

1. Names
2. Titles
3. Dates
4. Alternatives to italics
The script of Pembroke 25 is a form of Standard Late English Caroline Minuscule, also known as Style IV Anglo-Caroline minuscule. Caroline minuscule was invented around the late eighth century in France, combining elements from Cursive Half-Uncial and local minuscules to produce a clear, elegant script with fewer ligatures and variant letter forms. Its association with the court of Charlemagne gives it its modern name. Typical features include the a like a modern printed a, with an arched top stroke over the round body; g like a formal printed g, with a round top bowl above a tail (which may have been open or closed); r like a modern r; and s much like a modern f but without the cross bar and not descending below the line. This style of script usually has a neat, restrained aspect.
Names

- <name>
- <persName>
- <placeName>
- <orgName>
- <geogName>
- ...

Medieval/Modern Manuscript Studies in the Digital Age
Bibliographies

<bibl>
  <author>B. Bischoff</author>
  <title level="m">Latin Palaeography: Antiquity and the Middle Ages</title>
  <pubPlace>Cambridge</pubPlace>
  <date>1990</date>
</bibl>

<bibl>
  <author>T.J. Brown</author>
  <title level="a">Latin Palaeography since Traube</title>
  <title level="j">Transactions of the Cambridge Bibliographical Society</title>
  <date>1959–63</date>
  <biblScope>361–81</biblScope>
</bibl>
Semantics and not presentation!!! Well, most of the time!

particularly in the `<hi rend="italic">ihc</hi>` abbreviation for `<foreign xml:lang="lat">Iesus</foreign>`
Exercise

- Use the same file used for the previous exercise and add phrase-level elements
Manuscript Description

Identifier and Intellectual content

Dr Elena Pierazzo
The `<msDesc>` element

- Can occur within the `<sourceDesc>`
- Can occur within `<witness>`
- Can occur within `<body>`
The `<msDesc>` element: content

- `<msIdentifier>`: mandatory
- `<head>`
- `<msContent>`
- `<physDesc>`
- `<history>`
- `<additional>`
- `<msPart>`
The Identifier

- Parker Library, Ms 174, *Chronicle of England*
- Bodleian Library, MS. Don. e. 7, *Volume the First*
The Identifier

- A very simple description

```xml
<msDesc>
  <msIdentifier>
    <repository>Parker Library</repository>
    <idno>174</idno>
  </msIdentifier>
</msDesc>

<msDesc>
  <msIdentifier>
    <repository>Bodelian Library</repository>
    <idno>Ms. Don. e. 7</idno>
  </msIdentifier>
</msDesc>
```
The Identifier

And a more detailed description:

```xml
<msDesc>
  <msIdentifier>
    <country>United Kingdom</country>
    <district>Cambridgeshire</district>
    <settlement>Cambridge</settlement>
    <institution>Corpus Chirsti College</institution>
    <repository>Parker Library</repository>
    <idno>174</idno>
  </msIdentifier>
  <head>Chronicle of England</head>
  <altIdentifier>
    <idno>Chronicle of England</idno>
  </altIdentifier>
</msDesc>
```
<msDesc>
  <msIdentifier>
    <country>United Kingdom</country>
    <district>Oxfordshire</district>
    <settlement>Oxford</settlement>
    <institution>University of Oxford</institution>
    <repository>Bodleian Library</repository>
    <idno>Ms Don. e. 7</idno>
    <altIdentifier>
      <idno>Volume the First</idno>
    </altIdentifier>
  </msIdentifier>
</msDesc>
Exercise

1. Go to the mmsda_examples/msDesc/exercises folder saved in your Desktop
2. Open msIdentifier.xml
3. Open on a web browser the link contained in the exercise (copy the link and paste in the browser address field)
4. use some of the entries for filling the <msIdentifier>
The Content

<msContents>
<msItem>
<locus from="1r" to="198v"/>
<title>Chronicle of England</title>
<title type="xiv">William Caxton's Fructus Temporum.</title>
<incipit><locus from="1r"/>Here may a man hure (how) Engelande was fferst callede albyon and þoruʒ wham hit had þe name</incipit>
<lb/> In the noble land of Syrrie.</incipit>
<explicit><locus from="198v"/>he deide att Shene and is beried worshipfully at Westmynster on whos soule god haue mercy. Amen</explicit>
<textLang mainLang="enm">Middle Eghlish/Latin</textLang>
</msItem>
</msContents>
The Content

- Cambridge, University Library, li. 1. 33.
<msItem>
  <locus from="2r" to="28v/8"/>
  <title type="supplied">Ælfric's preface to <title>Genesis</title> and his translation of Chapters I-XXIV.22.</title>
  <incipit>Ælfric munuc gret æþpelpeard ealdorman eadmod <lb/>lice</incipit>
  <explicit>sealde hyre Žarpreo <lb/>nas eallgildene sona; gildene biagas god herlENDE</explicit>
  <textLang mainLang="eng">English</textLang>
  <filiation>Almost identical to the versions in Claudius B iv and Laud Misc. 509. However, according to Crawford, chapters iv, v, xi, xxiii and xxiv offer an independent translation of the Vulgate and vary from the versions in these two manuscripts. Crawford edited these versions in parallel columns (See Crawford 1922, p. 5).</filiation>
  <listBibl>
    <bibl>Crawford 1922, pp. 76-149.</bibl>
    <bibl>Ker 1957, Art. 1.</bibl>
  </listBibl>
</msItem>
The fifth Item

<msItem>
  <locus from="31r" to="12">right margin</locus>
  <quote>'sunu . Spilche heo o?<gap/>
    <lb/> ne siððan bere<gap/>
    <lb/> ac forðl þe c<ex>ri</ex>st <gap/>
    <lb/> frumcenned of <lb/> manegu<ex>m</ex> gast[...] <lb/>
    cu<ex>m</ex> broðru<ex>m</ex>..</quote>
  <textLang mainLang="eng">English</textLang>
  <note>This Marginalia is in fact a correction which was added in the right margin by scribe 1 to Ælfric's Catholic Homilies, First Series, II, Christmas Day to supply a few lines from the text which were probably omitted at some stage due to eye-skip (See Clemoes 1997, p. 193, ll. 88-90).</note>
</msItem>
1. Go to the mmsda_examples/msDesc/exercises folder saved in your Desktop
2. Open msContent.xml or msContent2.xml (modern)
3. Open on a web browser the link contained in the exercise (copy the link and paste in the browser address field)
4. use some of the entries for filling <msItem>s
<msDesc>
  <msIdentifier>
    <settlement>Cambridge</settlement>
    <repository>Corpus Christi College</repository>
    <idno>422</idno>
  </msIdentifier>
  <msPart>
    <altIdentifier>
      <idno>Vol. I</idno>
    </altIdentifier>
    <msContents></msContents>
  </msPart>
  <msPart>
    <altIdentifier>
      <idno>Vol. II</idno>
    </altIdentifier>
    <msContents></msContents>
  </msPart>
</msDesc>
Digital Cataloguing

TEI Manuscript Description: Physical Description, History and Additional Information

Dr Arianna Ciula – Dr Elena Pierazzo

mmsda@sas.ac.uk
Digital Cataloguing

- Physical description
  - Carrier
  - Text (writing, illumination, music, additions)
  - Bindings, Seals, and Additional Material

- History
  - Origin
  - Provenance
  - Acquisition

- Additional Information
  - Administrative information
  - Surrogates
  - Bibliography
Grouping of many distinct items

1. You can simply supply paragraphs of prose, covering such topics directly within `<physDesc>`

2. Or, group your discussion within the following specific elements
   - `<objectDesc>`: the physical carrier
   - `<handDesc>`: what is carried on it
   - `<musicNotation>, <decoDesc>, <additions>`
   - `<bindingDesc>` and `<sealDesc>`
   - `<accMat>`: accompanying material

   Similarly, within the specific elements, you can supply paragraphs of prose, or further specific elements
1. The `<objectDesc>` can contain just paragraphs
2. Or `<supportDesc>` and `<layoutDesc>`
   - `<supportDesc>` can contain a more complex substructure with specific elements for
     - `<support>`: description of the materials etc. which make up the physical support
     - `<extent>`: approximate size specified in any convenient units
     - `<foliation>`: numbering system or systems used to count the leaves or pages
   - `<collation>`: description of how the leaves or bifolia are physically arranged
   - `<condition>`: contains a description of the physical condition of the manuscript
   - `<layoutDesc>` can include multiple
     - `<layout>` elements
1. `<objectDesc>` can contain just paragraphs

```xml
<msDesc>
  <msIdentifier>
    <settlement>Cambridge</settlement> <repository>Parker Library</repository>
    <idno>204</idno>
  </msIdentifier>
  <physDesc>
    <objectDesc>
      <p>parchment</p>
      <p>two columns of 46 lines</p>
    </objectDesc>
  </physDesc>
</msDesc>
```
2. `<objectDesc>` can contain `<supportDesc>` and `<layoutDesc>`

```
<physDesc>
  <objectDesc>
    <supportDesc>
      <p><material>parchment</material></p>
    </supportDesc>
    <layoutDesc>
      <p>two columns of 46 lines</p>
    </layoutDesc>
  </objectDesc>
</physDesc>
```
The Carrier

Structured version
The Carrier
Collation

<objectDesc>
  <supportDesc>
    <collation> The manuscript offers an irregular make up, with gaps and codicological incongruities. It was possibly assembled over time and the present sequence of the quires may not have been the original one (see also <ref>Schipper 1983</ref> and <ref>Ker 1957, p. 23</ref>). One scribe (<name>Scribe 1</name>) copied fols 2-36v and 120v-227v using a mixed quiring technique: catchwords and signatures in quires 1-5, but used catchwords only in quires 16-27; another scribe prepared quires 6 and 7, fols. 37-53 with no indication about quire boundaries; and the same scribe (<name>Scribe 2</name>) scribe signed regularly their quires from i. to viii. and these correspond to quires 8 to 15 (fols 53-120v). It is likely that those quires that now make up the beginning of the manuscript were not the initial ones before the sixteenth century binding. Although it is not possible to say which quire was the first one, the possible candidate is quire 5. The xiii/xiv century note on the top margin of fol. 29r, commenting on the nature of the language of the texts in the manuscript, may suggest that once the manuscript started with this folio.</collation>
  </supportDesc>
</objectDesc>
<condition>

The manuscript is in good condition. The parchment has holes on fols 117, 118, 171, 173, 225 and 226 (repaired with a rounded vellum patch) and sewing on fols 8, 16, 38, 80, 204, 205.

</condition>
Multiple layouts may also be specified:

```xml
<layoutDesc>
  <layout>
    <locus from="1" to="26"/> 24 lines </layout>
  <layout>
    <locus from="27" to="586"/> 20 lines </layout>
</layoutDesc>
```
The second group of elements within a structured physical description concerns aspects of the writing, illumination, or other notation (notably, music) found in a manuscript, including additions made in later hands

- `<handDesc>` contains a description of all the different kinds of writing used in a manuscript
- `<handNote>` describes a particular style or hand distinguished within a manuscript
- `<decoDesc>` contains a description of the decoration of a manuscript
- `<musicNotation>` contains description of type of musical notation
- `<additions>` contains a description of any significant additions found within a manuscript, such as marginalia or other annotations
About f. 24r the hand begins to get larger and settles into a script which continues to the end.
The scribe was a Robert of one of the houses called Vallis Dei. There were several of these in Eastern France and in Flanders.
<handDesc hands="2">
  <handNote xml:id="SC1" scope="major" script="English Vernacular minuscule">
    This scribe copied <locus from="2" to="36v">fols 2-36v</locus>, <locus from="120v" to="227v">120v-227v</locus>. The scribe uses a type of Anglo-Saxon Insular minuscule characteristic of the mid to second half of the twelfth century. The aspect is slightly backward-leaning; the ascenders are slightly wedged. The descenders are either tapered or curve on the left. The hand does not extend too much either below or above the ruled line. The nib of the pen is cut in a squarish proportion to give to the script a squarish aspect. The scribe mixes both caroline and insular features.
  </handNote>

  <handNote xml:id="SC2" scope="major" script="English Vernacular minuscule">
    This scribe copied <locus from="53r" to="120v">fols 53r-120v</locus>. The scribe uses a type of Anglo-Saxon Insular minuscule characteristic of the mid second half of the twelfth century. The aspect is upright and fractured; the ascenders are slightly wedged. The descenders occasionally taper to the left. The hand extends below or above the ruled line in a proportionate way. The scribe mixes caroline and insular letter forms.
  </handNote>
</handDesc>
<decoDesc> (decoration description) contains a description of the decoration of a manuscript, either as

1. A sequence of paragraphs
2. A sequence of topically organised <decoNote> elements

<decoDesc>

<p>The catchwords are enclosed in rather elaborate ornaments.</p>
</decoDesc>
<decoDesc>
  <decoNote type="initials">
    <p>Decorated initials and rubrics in red; litterae notabiliiores highlighted in red. Green is used only on <locus>fol. 227v</locus>.
    </p>
    <list>
      <item><locus>fol. 2r</locus>, three-line initial in red</item>
      <item><locus>fol. 3r</locus>, nine-line initials in red</item>
      <item><locus>fol. 24v</locus>, five-line initial in red</item>
      <item><locus>fol. 29r</locus>, ten-line initial in red</item>
    </list>
  </decoNote>
</decoDesc>
<musicNotation>
  <p>
    <locus>52v</locus> Below the miniature, the text of the preface is continued in minuscule, with neumes.
  </p>
</musicNotation>
The Text
Additions

- `<additions>` contains a description of any significant additions found within a manuscript, such as marginalia or other annotations element.
- It can be used to list or describe any other additions to the manuscript, such as scribbles, doodles, etc., which are considered to be of interest or importance.

```xml
<additions>
  <p>Note by Parker: `<q xml:lang="lat">hic desunt usque ad 7 h. quinti</q>`. This refers to no. 182. A later note: `<q>imperfect. he wrot to almost the end of Edw. 4th</q>.

  </p>
</additions>
```
Physical Description

Exercise 1a

1. In oXygen open the msPhysDesc.xml within the folder mmsda_examples/msDesc/exercises
2. Open Parker on the Web CCCC 54 file within your browser
   http://parkerweb.stanford.edu/parker/actions/manuscript_description_long_display.do?ms_no=54
3. Reflect on the content of this catalogue entry
4. Create and fill in the `<physDesc>` element in the XML file
5. Validate and save
1. In oXygen open the msPhysDesc.xml within the folder mmsda_examples/msDesc/exercises
2. Open the description of Volume the Second within your browser http://www.janeausten.ac.uk/edition/ms/VolSecondHeadNote.html Reflect on the content of this catalogue entry
3. Create and fill in the <physDesc> element in the XML file
4. Validate and save
Bindings, Seals, and Additional Material

The third major component of the physical description relates to supporting but distinct physical components, such as bindings, seals and accompanying material.

- `<bindingDesc>` present and former bindings of a manuscript, either as a series of paragraphs or as a series of distinct `<binding>` elements
- `<binding>` description of one binding, i.e. type of covering, boards, etc.
- `<sealDesc>` seals or other external items attached to a manuscript, either as a series of paragraphs or as a series of distinct `<seal>` elements, possibly with additional `<decoNote>`s
- `<seal>` description of one seal or similar attachment applied to a manuscript
- `<accMat>` details of any significant additional material
<bindingDesc>
  <p>Bound, s. XVII in</p>
  <material>brown suede</material>.
</bindingDesc>

<bindingDesc>
  <binding>
    <p>Nineteenth-century binding on five</p>
    <material>medieval sewing-stations</material>.
  </binding>
  <condition>It is in need of repair.</condition>
</bindingDesc>
<accMat>

Contains details of any significant additional material which may be closely associated with the manuscript being described (e.g. non-contemporaneous documents or fragments bound in with the manuscript at some earlier historical period)

Added flyleaves of a printed edition of a Latin text at the beginning and the end of the manuscript.

</accMat>
History

- **<history>** groups elements describing the full history of a manuscript or manuscript part
  - **<origin>**: where it all began
    - *datable* element → attributes: @notBefore and @notAfter
  - **<provenance>**: everything in between (after its creation but before its acquisition)
  - **<acquisition>**: how it was acquired by the holding institution
<history>
  <provenance notBefore="1400" notAfter="1600">
    <p>On <locus>f. 1r</locus> is a pressmark (older than Parker's time, but seemingly of cent. xv-xvi) <q>C. 23</q>, possibly a late Norwich mark, and a note in another hand (xvi cent.): <q>Autor huius operis Egidius archiep. bituri<ex>c</ex>ensis tritemius</q>.</p>
  </provenance>
</history>
<history>

<origin>Trehane suggests a south-easter origin probably <origPlace>Christ Church, Canterbury of Rochester</origPlace> (<ref>Trehane 1998, pp. 240-244</ref>). Schipper argued that part of the manuscript was copied in the <origPlace>cathedral-priory of St Aethelthryth and St Peter in Ely</origPlace>. This argument is based on the presence in the manuscript of this homily, which could have constituted the central starting point for the process of copying (<ref>Schipper 1983, pp. 291-292</ref>). Traxel does not offer a conclusive argument on the history of the manuscript. However, he suggests a possible compilation originally from <origPlace>St Augustine's Canterbury</origPlace> and then <origPlace>Ely</origPlace> (<ref>Traxel 2004, p. 207</ref>).</origin>

<provenance notBefore="1541"> It is possible that the manuscript was at <placeName>Ely</placeName> due to the evidence on <locus>fol. 25r</locus>. The manuscript was amongst the collection of <persName>Matthew Parker</persName> (<ref>James 1912, p. 388, n. 17</ref>), see also <ref>Ker 1957, p. 23</ref>). It is not clear how the manuscript came into the possession of Archbishop Parker. However, it may be possible that Parker acquired it. 33 at Ely during his residence there. On 10 September 1541, Ely was re-founded with a new chapter, a dean and eight canons ('Ely Introduction') and Parker was 'among the first canons of the new foundation'. Other manuscripts now in Corpus Christi College Cambridge came into Parker's possession while he was at Ely (CCCC 44, 335, 393 and 416, see <ref>Owen 1973, p. 4</ref>). </provenance>

<acquisition when="1574"> Donated to <orgName>Cambridge University Library</orgName> in 1574 by <persName>Matthew Parker</persName> together with other manuscripts containing Old English texts: li.2.4; li.2.11; li.4.6; Kk.3.18 and Hh.1.10. At a similar time Parker gave also Ff.1.23 and Gg.3.28 (Oates 1986, pp. 97-101). Possibly under the influence of <persName>Andrew Perne</persName>, master of <placeName>Peterhouse</placeName> in 1554 and dean of Ely in 1557, who was restoring the University Library in 1574 (see also, <ref>Wright 1949-53, p. 224</ref>). </acquisition>

</history>
Additional Information

- `<additional>` groups additional information, combining bibliographic information about a manuscript, or surrogate copies of it with curatorial or administrative information
  - `<adminInfo>`: administrative information
    - `<recordHist>`
      - `<source>`
      - `<change>`
    - `<availability>`
    - `<custodialHist>`
      - `<custEvent>`
  - `<surrogates>`: information about other representations
  - `<listBibl>`: bibliographic descriptions
Additional Information

History
Exercise 2a

1. In oXygen open the msHistory.xml within the folder mmsda_examples/msDesc/exercises
2. Keep open Parker on the Web CCCC 54 file within your browser http://parkerweb.stanford.edu/parker/actions/manuscript_description_long_display.do?ms_no=54
3. Create and fill in the <history> element in the XML file
4. Validate and save
1. In oXygen open the `msHistory.xml` within the folder `mmsda_examples/msDesc/exercises`
2. Keep open Volume the Second file within your browser `http://www.janeausten.ac.uk/edition/ms/VolSecondHeadNote.html`
3. Create and fill in the `<history>` element in the XML file
4. Validate and save


Transcription

Documentary level

Dr Elena Pierazzo
Synopsis

- Layout
- Hands
- Abbreviations
- Deletions, additions
- Marginalia
- Decorations
- Damages, omissions, uncertainty
... haben<nb rend="hyphen" type="inWord"/>tes claro...

<nb>: line break
<nb page number>: page break
<nb column break>: column break
Hands: description

<physDesc>
  ... <handDesc hands="2">
    <handNote xml:id="Eirsp-1" scope="minor">
      <p>English Vernacular minuscule, fairly wide with forked ascenders and prominent foot-serifs.</p>
    </handNote>
    <handNote xml:id="Eirsp-2" scope="major">
      <p>English Vernacular minuscule. Not always consistent, more laterally compressed, and descenders turn to the left</p>
    </handNote>
  </handDesc>
</physDesc>
Hands: encoding

... [and] s[an]c[t]e petre. for his moder saple. <handShift new="#Eirsp-2">Her sputelab on ðyssere cristes béc ðæþ sapí hagg. ðæþ pidecumæ hæþ gedon út</handShift>
Abbreviations

Qui sua<ex>m</ex> amica<ex>m</ex>

Qui <choice>
  <abbr>sua</abbr><g ref="#titulus"/></g><abbr/>
  <expan>suam</expan>
</choice>

Qui <choice>
  <abbr>sa</abbr><g ref="#titulus"/></g><abbr/>
  <expan>sua</expan>
</choice>
<encodingDesc>
  <charDecl>
    <char xml:id="titulus"/>
    <charName>Abbreviation Super Stroke</charName>
    <charProp>
      <unicodeName>COMBINING MACRON</unicodeName>
      <value>Above attached</value>
    </charProp>
    <mapping type="Unicode">U+0304</mapping>
  </char>
</charDecl>
</encodingDesc>

→ UNICODE
Exercise

1. Click  or select File → Open
2. Go to the mmsda_examples/transcription/exercises folder saved in your Desktop
3. Open transcription.xml
4. Open on a web browser the link contained in the exercise (copy the link and paste in the browser address field)
5. Add to the transcribed text the markup for marking:
   a) Line breaks
   b) Page breaks
   c) Abbreviations
   d) Hands
<p>servi<del type="erasure">ens</del> s&amp; et plebs</p>

<choice><abbr>N</abbr><expan>NOMEN</expan></choice>
<del type="overstrike" resp="#hand2">papa nostro et</del>
Deletions/Additions

- V
Otto rex unguitur in imperatorem ad Iohanne papa. Cenobium in Domnania, quod...
Decorations

Or a more semantic approach…

```html
<p><hi rend="rubric">Dccclxxxiii</hi> xxv</p>

<p><num rend="rubric" value="983">Dccclxxxiii</num> <num value="25">xxv</num></p>
```
More complex decorations

```xml
<p>
  <figure>
    <graphic facs="#capital"/>
  </figure>
  <hi rend="uppercase">
    <c rend="drop">I</c>am dudum...</hi>
</p>
```
Damages, uncertainty omissions

<p>
  <damage agent="water" degree="3"><gap extent="4" unit="chars"/></damage>
  eode una haet ocmund</p>
Exercise

1. Using the same file, add the markup to mark:
   a) Deletions/additions
   b) Marginalia
   c) Decorations
   d) Damages, uncertainty, omissions
Transcription

Editorial level

Dr Elena Pierazzo
Synopsis

- Errors and corrections
- Spelling regularisation
  - Word spacing
- Supplied text
- Critical apparatus
Errors and Corrections

\[\text{<choice>}
\text{<sic>Qria</sic>}
\text{<corr>Tria</corr>}
\text{</choice>}\text{ sunt homini necessaria ad salutem}
<p><choice>
  <orig>&amp;</orig>
  <reg>et</reg>
</choice> ut filios <choice>
  <orig>k</orig>
  <reg>c</reg>
</choice><ex>arissi</ex>mos</p>
Regularisations: punctuation

\[\text{\textless p\textgreater } \textit{passionis} \textit{nec} \textit{non} \texttt{&amp;} \textit{ab} \textit{inferis} \textit{resurrectionis} \textless \text{choice} \textgreater
\textless \text{orig} \textgreater . \textless \text{orig} \textgreater
\textless \text{reg} \textgreater , \textless \text{reg} \textgreater
\textless \text{choice} \textgreater \texttt{sed} \texttt{amp;} \texttt{in} \texttt{celos} \texttt{glorioso} \textless \text{p} \textgreater\]
Regularisations: capital letters

<p><choice>
<orig>V</orig>
<reg>U</reg>
</choice>nde et memore <choice>
<orig>d</orig>
<reg>D</reg>
</choice>omine, <reg>nos tui</reg></p>
White spaces: missing spaces

p<ex>ri</ex>ncipes &amp;<reg> </reg>tiranni

p<ex>ri</ex>ncipes &amp;<choice><orig/></choice><reg> </reg>/</choice>tiranni

p<ex>ri</ex>ncipes &amp;<g ref="#edspace"/>tiranni
White spaces:
extra spaces

<p>pe sot<space/>lice</p>
Supplied text

<supplied>AD M</supplied> ENSAM
Critical Apparatus

\(\text{eius in omnibus]} \text{ in omnibus } H, \text{ omnibus eius } L\)

\(<p>\text{eius in omnibus]} \text{ in omnibus } H, \text{ omnibus eius } L</p>\)

\(<\text{app}>\)
\(<\text{rdg wit=""H"">}\text{in omnibus}</\text{rdg}>
\(<\text{rdg wit=""L"">}\text{omnibus eius}</\text{rdg}>
\)</app>\)

\(<\text{app}>\)
\(<\text{lem wit=""C3"">}\text{eius in omnibus]} \text{ in omnibus } H, \text{ omnibus eius } L</\text{lem}>\)
\(<\text{rdg wit=""H"">}\text{in omnibus}</\text{rdg}>
\(<\text{rdg wit=""L"">}\text{omnibus eius}</\text{rdg}>
\)</app>\)
<sourceDesc>
  <listWit>
    <witness xml:id="C3"><msDesc>...<msDesc></witness>
    <witness xml:id="H"/>
    <witness xml:id="L">Manuscript L...<witness>
  </listWit>
</sourceDesc>
Exercise 4a

- Using the same file, add markup to encode:
  - Regularisations of capital letters and whitespaces
  - Corrections of errors
Exercise 4b

- Using the images provided in the package (John Donne, *The Curse*), try to produce a collation and a critical apparatus.
- You may find easier to take Ms Add. 8467 as base text