

Creating a Digital Scholarly Edition with the Text Encoding Initiative

Edited by Marjorie Burghart

With contributions from:

Marjorie Burghart
James Cummings
Elena Pierazzo

DEMM - 2017

Creating a Digital Scholarly Edition
with the
Text Encoding Initiative

A Textbook for Digital Humanists

Introduction

This textbook was born from the experience of the Digital Editing of Medieval Manuscripts (DEMM) programme, an Erasmus SP+ project funded by the European Union.¹ From 2014 to 2017, the partners of this programme had the opportunity to teach digital methodologies and technologies to international groups of young scholars interested in producing various types of digital editions of medieval texts and documents.

This textbook is our way to share with the DH community the courses and material we prepared for this programme. It is an ideal companion to the online course *Scholarly Digital Editions: Manuscripts, Texts and TEI Encoding*, freely available on the #dariahTeach platform.² If the official *Guidelines* remain the main resource about the encoding of digital editions in TEI, both this book and the online course offer a different approach by targeting the particular needs of editors of medieval (and more generally manuscript) sources, and translating them into TEI terms.

We want to thank all the people who contributed to this experience, not least of all the DEMM students. Particular thanks go Martin Haltrich and the staff of the Stiftsbibliothek Klosterneuburg, who helped us gather relevant examples from their manuscript collections, and kindly contributed the images to this volume.

¹ See the DEMM website for more information: <https://www.digitalmanuscripts.eu>

² <https://teach.dariah.eu/course/view.php?id=32>

Structure and Layout

Marjorie Burghart

In this chapter you will learn:

- the general structure of a TEI document;
- how to represent the physical and logical structure of your document and/or text;
- a first hint at the resolution of concurrent hierarchies.

1. General structure of a TEI document

A TEI document always contains two mandatory parts: the header (`<teiHeader>`) which contains metadata about the document and its content, and the representation of the text itself (`<text>`). Those rules apply not only for editions, but for any kind of TEI document. In this chapter, we will only insist on the parts which are most essential for editing ancient texts and documents, without going over all the possibilities. For a general overview we recommend to read Burnard 2014, ch. [The structural organization of a TEI document](#) and [The TEI Header](#), and of course the TEI Guidelines. Note that, besides the mandatory header and text, the TEI root may also contain the `<facsimile>` and `<sourceDoc>` elements, which are mostly useful for documentary editions or a *dossier génétique*: they will be explained in chapter Digital facsimiles.

1.1. Header

The header (`<teiHeader>`) is particularly important as the ID card of your work, especially when it is distributed to others, or added to a corpus or digital library. It presents all the metadata in four sections.

The `<fileDesc>` (file description) “contains a full bibliographic description of an electronic file.”¹ This essential part of the header is subdivided into subsections, three of which are mandatory:

- `<titleStmt>` (title statement) “groups information about the title of a work and those responsible for its content”² It must contain at least one `<title>` element, used to indicate the title of the text edited here. It may be followed by an `<author>` element,

1. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-fileDesc.html>

2. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-titleStmt.html>

containing the name of the original author of the text you are editing, and an **<editor>** element to indicate the name of the scholarly editor of the text. A **<principal>** element is also available, to supply “the name of the principal researcher responsible for the creation” of this document, that the principal investigator who bears the scientific responsibility but may not be the actual editor.³ If you need to indicate more roles, for a collaborative effort for instance, you can use the **<respStmt>** element (statement of responsibility):⁴ each **<respStmt>** contains a **<resp>** and one or more **<name>** (or **<persName>**) elements, which supply respectively a description of the role played and the name of the people who played it. For instance, let us imagine that a fictional Dr. Anna Rossi created a digital scholarly edition of Cicero’s De Amicitia, a project for which she assigned tasks to her assistants John and Lenka. A basic **<titleStmt>** could be:

```

<titleStmt>
  <title>De Amicitia</title>
  <author>Marcus Tullius Cicero</author>
  <editor xml:id="arossi">Anna Rossi</editor>
  <respStmt xml:id="lnovakova">
    <resp>Identification of person names mentioned in the text,
    and redaction of short biographies of these persons</resp>
    <persName>Lenka Novákova</persName>
  </respStmt>
  <respStmt xml:id="jsmith">
    <resp>Proofreading</resp>
    <persName>John Smith</persName>
  </respStmt>
</titleStmt>

```

Note that we have given a unique **@xml:id** value to each of these roles. This will allow us to point to these roles (and people) later, when we want to give them responsibility for an action, interpretation or change in the course of the creation or revision of the edition.

- **<publicationStmt>** (publication statement) “groups information concerning the publication or distribution” of the text.⁵ In its simplest form, it may consist in a simple prose paragraph (**<p>**). Alternatively it can contain elements dedicated to supply the name of the publisher (**<publisher>**), place (**<pubPlace>**) and date (**<date>**) of publication, and most importantly the conditions of availability of the text (**<availability>**), where you can indicate in a prose paragraph under which licence, if any, the text is distributed.
- **<sourceDesc>** (source description) “describes the source from which an electronic text was derived or generated”⁶ For scholarly editions, there can be two cases: the TEI

3. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-principal.html>

4. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-respStmt.html>

5. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-publicationStmt.html>

6. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-sourceDesc.html>

document may be a “born digital” edition, and the **<sourceDesc>** will contain the description of the original document (or documents) on which the edition is based.⁷ Or the TEI document may be the digital version of a pre-existing edition, and the **<sourceDesc>** will contain a bibliographic description of the book or article used as the source for this version. In either case, the simplest form of description is a simple prose paragraph (**<p>**), but it can be refined into a very elaborate manuscript or bibliographic description.

The other, non mandatory parts of **<fileDesc>** are **<editionStmt>**, **<extent>** **<notesStmt>** and **<seriesStmt>**. We invite you to check their definitions in the Guidelines, but they are less crucial to scholarly editions than the elements described above.

The **<encodingDesc>** (encoding description) “documents the relationship between an electronic text and the source or sources from which it was derived.”⁸ In its simplest form, it can consist of a simple prose paragraph (**<p>**), but it can also be much more detailed, with several sections, one of which is the **<projectDesc>** (project description) which “describes in detail” (in simple prose paragraphs) “the aim or purpose for which an electronic file was encoded, together with any other relevant information concerning the process by which it was assembled or collected.”⁹ You can explain there the context of your edition (student assignment, part of a thesis, exercise, professional work, etc.), what you are willing to achieve, etc. Another optional section particularly relevant to critical edition is **<variantEncoding>** which “declares the method used to encode text-critical variants”¹⁰ (you will learn about the possible methods in chapter Textual Variants). For the other possible sections of **<encodingDesc>**, we refer you to the relevant sections of the Guidelines.

The **<profileDesc>** (text-profile description) “provides a detailed description of non-bibliographic aspects of a text, specifically the languages and sublanguages used, the situation in which it was produced, the participants and their setting.” When present, it contains a simple prose paragraph (**<p>**).¹¹

The **<revisionDesc>** (revision description) “summarizes the revision history for a file”¹² In this section, you have the possibility to record all the revisions of the current file. Each revision will be represented by a **<change>** element containing a prose description of the change made, and complete with a set of optional attributes: **@when** to indicate the date of the change, **@who** to indicate the person(s) responsible for this change typically by pointing to the element giving their full name, or **@target** which “points to one or more elements that belong to this change” (for instance when a change consisted in adding some elements).¹³ If desirable, related **<change>** elements can be grouped together in a **<listChange>** element.¹⁴ The recommended practice is

7. To learn more about the encoding of manuscript descriptions, see chapter Manuscript Description

8. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-encodingDesc.html>

9. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-projectDesc.html>

10. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-variantEncoding.html>

11. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-profileDesc.html>

12. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-revisionDesc.html>

13. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-change.html>

14. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-listChange.html>

to list the changes in reverse chronological order (most recent first). Going back to our example of the fictional edition of *De amicitia*, let us imagine that John Smith proofread the first version on June 5th 2017, then Lenka Nováková marked up the person names present in the text, without providing a full identification yet, on June 9th 2017. Here is how we could translate this information in the `<revisionDesc>`:

```
<revisionDesc>
  <change when="2017-06-09" who="#jnovakova">Added markup for each
person name mentioned in the text.</change>
  <change when="2017-06-05" who="#jsmith">Proofread the entire
text.</change>
</revisionDesc>
```

Note that, to point to an element from an attribute that allows this, like `@when` here, we use the value of the `@xml:id` of the attribute we are pointing to, but preceded with a # sign. This is a general rule when pointing to an attribute, whatever the target element or attribute used to point to it.

1.2. Text

Placed after the `<teiHeader>`, the `<text>` element is the other mandatory component of any TEI document. As the name suggests, it contains the text of the edition, which may be divided into three parts: `<front>`, `<body>`, and `<back>`. There can be different approaches regarding the distribution of the text among those parts:

- If you are preparing a digital facsimile of a book, you might choose to put all prefatory matter of the edition you are encoding (preface, dedication, preliminary letters, etc.) in the `<front>`. In this case, the text in the `<front>` and `<back>` as well as the `<body>` is the ancient text you are working on, and your own contribution (for instance to describe the project that led to the creation of this document) will be restricted to the `<teiHeader>`.
- If you are preparing a digital scholarly edition, you may consider that the TEI document represents not only the ancient text, but your edition as a whole, with its own front and back matter added by the critical editors. You could therefore choose to put your introduction, discussion of the choice of witnesses, etc. in the `<front>`, so that the `<body>` contains only the ancient text you have edited. At any rate, be careful not to mix together, without a clear distinction, the ancient text and your own prose as editor: they should not, for instance, be together in the `<body>` of the document. Keeping the edited text clearly separated from the editor's commentary and introduction will, for instance, allow for linguistic processing of the ancient text.

1.2.1. Front matter

The `<front>` (front matter) is an optional element which “contains any prefatory matter (headers, abstracts, title page, prefaces, dedications, etc.) found at the start of a document, before

the main body”¹⁵ The **<front>** may contain paragraphs of prose (**<p>**), possibly grouped in **<div>** elements. They may also contain other elements which are too specific to be presented here: we invite you to consult the Guidelines if you wish to know more about the possible contents of the **<front>**.

1.2.2. Body

The **<body>** is the only mandatory part of the **<text>**.

1.2.3. Back matter

The **<back>** (back matter) is, like **<front>**, an optional part of **<text>**. It “contains any appendixes, etc. following the main part of a text.”¹⁶ In the context of an edition, for instance, you could decide to put there your bibliography, or any type of index-like information.

To summarise, here is, on the following page, how we could structure the TEI edition, by a fictional Jane Doe, of The Saga of Haakon Haakonarson or Hákonar saga Hákonarsonar, an Old Norse saga written in the 1260s by the Icelandic historian Sturla Þórðarson. Note that we can use the **@xml:lang** on any relevant element to indicate in which language it is written. In this case, this allows us to indicate two versions of the title: one in the original language, Old Norse, which language code is “non,” and the English translation. Accordingly, most of the content of **<text>** is in Old Norse, but the front and back matter are added, in English, by the editor. In this case, we can indicate on **<text>** an **@xml:lang** with the value “non,” and this value will be inherited by all the descendants of **<text>**, unless they have their own, different value for **@xml:lang**.

Now that we are familiar with the basic structure of a TEI document, it is time to move to more important matters: how to represent, in a TEI document, the logical and physical structure of the text or document that we wish to edit.

15. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-front.html>

16. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-back.html>

```

<TEI @xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title xml:lang="non">Hákonar saga Hákonarsonar</title>
        <title xml:lang="en">The Saga of Haakon Haakonarson</title>
        <author xml:lang="non">Sturla Þórðarson</author>
        <editor xml:id="jdoe">Jane Doe</editor>
      </titleStmt>
      <publicationStmt>
        <p>Publication Information would go here.</p>
      </publicationStmt>
      <sourceDesc>
        <p>Information about the source would go here.</p>
      </sourceDesc>
    </fileDesc>
    <encodingDesc>
      <projectDesc>
        <p>Optional description of the project's aim, scope, etc. </p>
      </projectDesc>
      <variantEncoding method="parallel-segmentation"
location="internal"/>
        <!-- The variantEncoding information is relevant only for
critical editions recording the variants of several witnesses. See
chapter "Textual Variants" -->
    </encodingDesc>
    <profileDesc>
      <p>Optional description of the language of the text, or the
circumstances of its production, </p>
    </profileDesc>
  </teiHeader>
  <text xml:lang="non">
    <front xml:lang="en">
      <p>Optional front matter, such as an introduction, discussion of
the witnesses, etc. could go here.</p>
    </front>
    <body>
      <p>The main text of the edition would go here.</p>
    </body>
    <back xml:lang="en">
      <p>Optional back matter, such as a bibliography, or index-like
information could go here.</p>
    </back>
  </text>
</TEI>

```

2. Representing the structure of a text or document

Texts have a logical structure: they have different parts, which may be arranged sequentially and / or hierarchically. For instance, a treatise may have a prologue, then several main sections called “books,” containing each several chapters, containing each in their turn several paragraphs. But, as philologists, the texts we edit are not born-digital: they are transmitted to us through objects like charters, codices or early printed books, which have their own physical structure. For instance, a manuscript codex comprises several folios, each of which may contain columns, and lines.

Of course, those logical and physical structures rarely coincide in ancient documents. Depending on your type of edition, you might be interested mostly in one or the other, but the TEI lets you encode both in a seamless way.

2.1. The logical structure

2.1.1. Main subdivisions of the text

The `<div>` (text division) “contains a subdivision of the front, body, or back of a text.”¹⁷ It is a powerful and versatile tool to represent the structure of a text: a `<div>` is a neutral subdivision, which can correspond to any relevant unit: book, section, chapter, subdivision of a chapter... In each case the same element is used, but the `@type` attribute can indicate the nature of the `<div>`. A `<div>` may contain other `<div>` elements, or paragraphs (`<p>`) and paragraph-like elements like `<ab>`, `<list>`, etc.,¹⁸ but never directly text.

For instance, a treatise containing a prologue and two books, comprising respectively two and three chapters could be represented like this:

```
<text>
  <body>
    <div type="prologue" xml:id="pro"/>
    <div type="book" xml:id="boo1">
      <div type="chapter" xml:id="cha1.1"/>
      <div type="chapter" xml:id="cha1.2"/>
    </div>
    <div type="book" xml:id="boo2">
      <div type="chapter" xml:id="cha2.1"/>
      <div type="chapter" xml:id="cha2.2"/>
      <div type="chapter" xml:id="cha2.3"/>
    </div>
  </body>
</text>
```

17. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-div.html>

18. For a full list of such elements, see <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-model.pLike.html>

Note that we have added an `@xml:id` attribute to each `<div>`. Those attributes are optional, but may prove useful in certain situations. Each `@xml:id` must have a unique value (within the same document) and conform to certain rules (the value cannot start with a number, for instance). This attribute is used to identify without ambiguity the element. When an element has an `@xml:id`, other elements can point to it, for cross-references for instance. When processing and displaying the edition, the value of `@xml:id` can be used to link directly to an element, from the table of contents or from an index for instance.

Subdivisions of a text, like chapters, often have a title. In manuscripts, those titles are often represented as “rubrics,” which may be in a different colour from the main text. The title of a `<div>` may be encoded with a `<head>` (heading) element¹⁹ The `<head>` element, which is optional, must be the first to appear in the `<div>` (paragraphs or other `<div>` elements must come after the `<head>`). For instance, on this image, we can see the end of a section, a rubric in red, with slightly larger letters, giving the title of the new section (“Noua uestigalia institui non posse,” that is “New taxes cannot be established”), then the actual text of this new section begins with a drop capital:

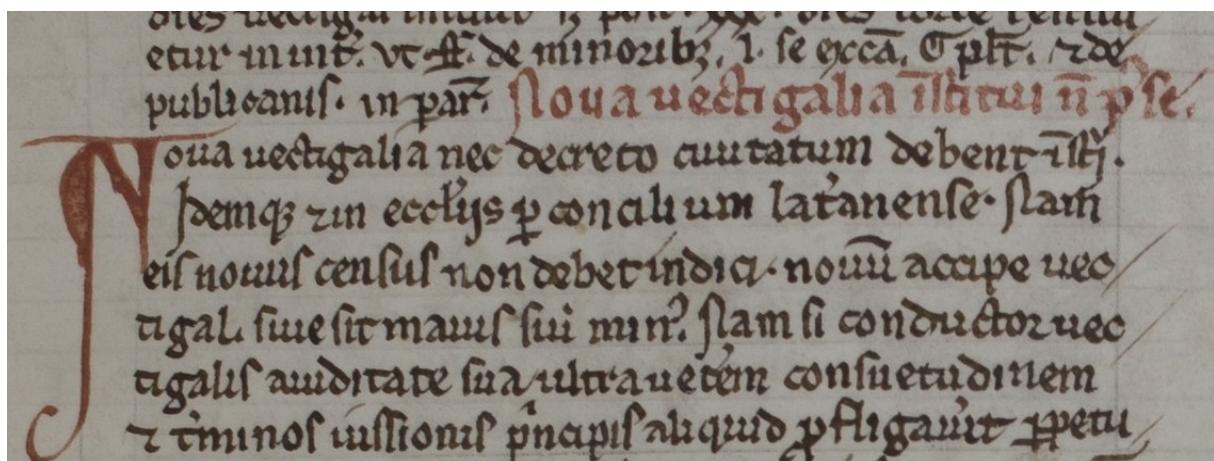


Figure 1: Klosterneuburg, Stiftsbibliothek, CCl 119, fol. 86r

We could represent this section, with its title, as follows:

```
<div>
  <head>Noua uestigalia institui non posse</head>
  <p>Noua uestigalia nec decreto ciuitatum debent institui...</p>
</div>
```

Sometimes, there may be more than a single heading, for instance “Distinctio XXIX. An in diuinis sit principium?” (that is, “Distinctio XXIX. Is there a beginning to divine matters?”). In such cases, you may use more than one `<head>`, and differentiate them with a `@type` attribute if necessary:

19. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-head.html>

```

<div>
  <head type="section_number">Distinctio XXIX.</head>
  <head type="section_title">An in diuinis sit principium?</head>
  <p/>
</div>

```

The opening formula starting with “Incipit...” in Latin (or the equivalent in vernacular languages) sometimes found at the beginning of a text or chapter may also be encoded as a **<head>** element. The closing formula, starting with or simply consisting in the word “Explicit,” may be encoded as a **<trailer>**, which “contains a closing title or footer appearing at the end of a division of a text”²⁰ A section containing an “incipit” and an “explicit” formula could be encoded as follows:

```

<div>
  <head>Incipit liber tetrius</head>
  <p/>
  <trailer>Explicit liber tertius</trailer>
</div>

```

2.1.2. Paragraphs and blocks of text

Paragraphs and similar elements may appear within **<div>**s, but if the text is not significantly subdivided they may also appear directly in **<body>**. The **<p>** (paragraph) element “marks paragraphs in prose,”²¹ which may sound like a fairly simple definition. Yet, as an editor, you will need to decide what defines a paragraph in your edition. Typically, when preparing a diplomatic edition, you will follow the punctuation and paragraph marks (the “pieds-de-mouche”) of the manuscript. When preparing a critical edition of a text, based on several witnesses, it is fairly common for editors of Latin texts to apply a modern punctuation and division in paragraphs to the text. Traditions and scholarly practices do vary with the type of literature.²²

In some cases, you might need to wrap some text and phrase-level markup into a paragraph-like element, but without using the semantic of a paragraph. The **<ab>** element²³ may be used instead of **<p>** in this case.

2.1.3. Lists

Lists are another very common feature in all sorts of texts: library catalogues, recipes, accounts, etc. They are very straightforward to encode: the whole list is wrapped in a **<list>** element,²⁴ which contains as many **<item>** elements as necessary, and possibly an optional **<head>** element. Note that no text is allowed to be directly in the **<list>** element: all the text must be

20. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-trailer.html>

21. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-p.html>

22. For a practical overview of various editing practices and strategies, see: The Arts of Editing Medieval Greek and Latin: A Casebook, E. Göransson, G. Iversen, B. Crostini (eds.), Turnhout: Brepols, 2016.

23. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-ab.html>

24. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-ab.html>

contained in the **<item>** elements (or **<head>**). For instance, if there are punctuation marks separating the items in the list, they must be enclosed in the **<item>** elements. Lists may appear within a paragraph, or instead of a paragraph. They may be nested: a **<list>** may appear within the **<item>** of a parent **<list>**.

Let us consider, for instance, this short extract of a medieval catalogue, listing the books held at the Klosterneuburg monastic library in 1330:²⁵ the document starts with an indication that this is the list of works by saint Jerome held by the library, which we can consider a header. Each entry describes the contents of a codex kept in the library. Some volumes contain a single work, but others contain several works, also listed: these entries can be encoded as nested lists.

```
<list>
  <head>Item libri Ieronimi</head>
  <item>Primo Ieronimus super Ysaiam prima pars</item>
  <item>Item Ieronimis super Ysaiam secunda pars</item>
  <item>Item expositio Ieronimi super Ezechielem libri XIIII</item>
  <item>
    <list>
      <item>Item epistule Ieronimi</item>
      <item>Ibidem Ieronimus super dicta quedam librorum et capitularum
          Pauli</item>
      <item>Ibidem epistula Augustini ad Ieronimum, in uno
          volumine</item>
      <!-- etc. -->
    </list>
  </item>
</list>
```

2.2. The physical structure: quires, pages, columns, lines

To represent the physical structure of the documents you encode, the TEI offers a series of “empty” ou “milestone” elements: instead of wrapping a whole passage of text, they simply mark the beginning of a new quire, page, etc. The advantage of empty elements is that they will not interfere with the markup representing the logical structure of the document. That way, a paragraph can start on one page and finish on another one without causing any overlapping of the markup.

The **<gb>** (gathering beginning) element can be used to mark up the limits of different quires or gatherings constituting a manuscripts. The **<pb>** (page beginning) element, marking the beginning of pages, is probably the most widely used of this series. Within the space of a page, **<cb>** (column beginning) and **<lb>** (line beginning) can be used to mark up the limits of columns and lines on the original document, respectively.

Note that all those milestones must be placed at the beginning of the phenomenon they are representing. Regarding lines, this goes against some traditional transcription practices, where editors are used to putting a marker at the end of lines. This may be the source of some errors

25. Klosterneuburg, 1251, f. 1r

when transitioning to digital editions: if you place `<1b>` milestones at the end of lines instead of the beginning, the count and display of the lines will be off.

Those milestones share similar attributes, especially; `@n`, `@ed` and `@edRef`, and `@break`.

The `@n` attribute may be used to record the number given to the element in the document. It is not mandatory, since it is possible to automatically count the relevant elements (quires, pages, columns or lines) with a script, allowing you for instance to automatically number the lines from 1 to x for each page of the document. But in some cases, you might want to record the precise number of one of those elements, as it is in the document. This is generally true of the page numbers which, in manuscripts, follow a particular naming pattern. For instance, the following paragraph is written across folio 4v and 5r in the manuscript we are transcribing. We could encode the beginning of f. 5r with a simple `<pb>` element in the text at the place where the page is changing, adding the `@n` attribute to indicate the folio number:

(...) `<p>` Comment messire Jehan le Boursier, gouverneur de Gand, durant les treves avoit avitaillié la ville de Gand. Et comment ungue maniere de gens qui s'appelloient `<pb n="5r"/>` les Comporselles faisoient moult de maulx. LIIII chapitre. `</p>` (...)

When working on an edition involving more than one witness, it is possible to record separately the boundaries and numbers of the pages, lines etc. in each witness, if this is useful. The `@ed` (edition) and `@edRef` (edition reference) attributes let us attach a milestone to a particular “edition” or witness in general. The difference between those two attributes is that `@ed` contains a sigil or other identifier of a witness, while `@edRef` contains a pointer to the description of a witness. Let us imagine that we are editing the paragraph above using two manuscripts, one from London and the other from Cambrai, which naturally have different page boundaries. We could represent them for each manuscript using the `@ed` attribute:

(...) `<p>` Comment messire Jehan le Boursier, `<pb ed="London" n="3v"/>`gouverneur de Gand, durant les treves avoit avitaillié la ville de Gand. Et comment ungue maniere de gens qui s'appelloient `<pb ed="Cambrai" n="5r"/>` les Comporselles faisoient moult de maulx. LIIII chapitre. `</p>` (...)

To use the `@edRef` attribute, we would need to have described the relevant manuscripts elsewhere in the document (on the description of witnesses, see chapter Manuscript Description), for instance something like this:

```
<listWit>
  <witness xml:id="C">[Description of the Cambrai
manuscript]</witness>
  <witness xml:id="L">[Description of the London
manuscript]</witness>
</listWit>
```

Then in the edition, we would point to those descriptions, using @edRef:

(...) <p>Comment messire Jehan le Boursier, <pb edRef="#L"
n="3v"/>gouverneur de Gand, durant les treves avoit avitaillié la
ville de Gand. Et comment ungue maniere de gens qui s'appelloient <pb
ed="#C" n="5r"/> les Comporselles faisoient moult de maulx. LIIII
chapitre. </p> (...)

The advantage of @edRef over @ed is that the link between the milestone and the witness is established in a more formal way, and is easier to process; the coherence of the value of the attribute can also be controlled more easily.

A change of page, line, etc. can of course occur right in the middle of a word. When this happens, the @break attribute lets you indicate that the milestone bearing this attribute with the value "no" does not separate the word in two like a white space would.

<p><lb/>Comment messire Jehan le Boursier, gouver<lb
break="no"/>neur de Gand, durant les treves avoit avitaillié <lb/>la
ville de Gand.</p>

Additionally, since all those milestones correspond to identifiable elements in the witnesses, it is possible to link them with their digital reproduction thanks to the @facs (facsimile) attribute. The value of @facs can be the relative or absolute path to an image, or the URL of this image:

<pb edRef="#C" n="5r" facets=".//images/f_005r.jpg"/>

This is a simple, basic way of linking an edited text to the facsimile of its witnesses. The TEI offers different and more refined ways of achieving this goal, which will be presented in Chapter Transcription.

3. Special structures

The particular structure of some texts, like poetry or drama, calls for special elements and attributes. We are just giving a quick overview here, and if you need to encode such texts we recommend that you read the relevant sections of the Guidelines for more details.²⁶

3.1. Verse

Each verse can be encoded with the <1> (verse line) element. This is the most basic form of encoding of verses, as illustrated with those opening lines of the Anglo-Irish poem entitled Fall and Passion:²⁷

26. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/CO.html#COVE> and <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/VE.html> for verse, or <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/CO.html#CODR> and <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/DR.html> for drama.

27. See the full text online in Anglo-Irish poems of the Middle Ages: The Kildare Poems, <http://celt.ucc.ie/published/E300000-001/>

```

<1>The grace of God ful of might </1>
<1>That is a king and euer was, </1>
<1>Mote amang vs alight</1>
<1>And yiue vs alle is swet grace,</1>

```

Coherent groups of lines, like stanza, can be encoded with the **<lg>** (line group) element. The full poem is also a group of lines of verse, and can be encoded as an particular type of **<lg>**, containing other types of **<lg>** elements, optionally distinguished thanks to the **@type** attribute.

Some special attributes availabe on **<1>** and **<lg>** elements will allow you to encode the rhyme scheme and the metrical analysis of the verse. The **@rhyme** attribute lets you represent the rhyme scheme of a group of lines, typically using letters to symbolise the rhymes, or any other system defined by the encoder. The **@met** (metrical structure, conventional) and **@real** (metrical structure, realised) attributes can be used to represent, through a user-defined formula, the metrical structure of an element, be it a full poem, stanza, a line... The former represents the conventional metrical structure, while the later indicates that a line or group of lines follows a different structure from the conventional one.

Let us encode, for instance, the first two stanza of the poem above, while representing their rhyme scheme. Note that we here we have chosen to encode the whole poem as an **<lg>** element, but could just as well have encoded it as a **<div>** containing **<lg>** elements.

```

<lg type="poem">
  <lg type="stanza" rhyme="abab">
    <1>The grace of God ful of might </1>
    <1>That is a king and euer was, </1>
    <1>Mote amang vs alight</1>
    <1>And yiue vs alle is swet grace,</1>
  </lg>
  <lg type="stanza" rhyme="cdcd">
    <1>Me to spek and you to lere </1>
    <1>That hit be worsip, lord, to the; </1>
    <1>Me to teche and you to bere </1>
    <1>That helplich to ure sowles be,</1>
  </lg>
  <!-- etc. -->
</lg>

```

3.2. Drama

The main divisions of drama texts, like acts and scenes, can be encoded with the all-purpose **<div>** element discussed earlier. But besides those divisions, a number of specific elements are available.

Drama is characterised, most notably, by the alternance of utterances of speech by different characters, and by the presence of stage directions. Each utterance is encoded with the **<sp>**

(speech) element,²⁸ which typically contains a **<speaker>** element,²⁹ indicating which character is delivering this speech, and the speech itself, which can consist in a paragraph (**<p>**) or in lines of verse (**<l>**), possibly grouped in **<lg>** elements. Note that **<sp>** does not necessarily contain a **<speaker>** element. If the text does not explicitly give the speaker's name there, or if the editor want to create a more formal link between a speech and its performer, it is possible to use the **@who** attribute on **<sp>** to indicate who is speaking. This attribute may point to the **@xml:id** of a **<castItem>**, inside a **<castList>**, describing in more detail a character.³⁰ Stage directions, which may appear anywhere, are encoded with the **<stage>** element.³¹

Let us consider for instance a possible encoding for this passage from the 15th c. *Mystère des Actes des Apôtres* by Simon Gréban.³² We have here one example of stage direction, followed by two utterances in verse, for which the name of the speakers are given:

```

<stage>Il arrive à l'ostel d'Anne. </stage>
<sp>
  <speaker>Caïphas</speaker>
  <l>Dy, Agripart, que fait ton maistre ?</l>
  <l>Est il pas bien embesoignié ?</l>
</sp>
<sp>
  <speaker>Agripart</speaker>
  <l>Il est lēans plus rechigné </l>
  <l>que n'est ung renart pourbouilly.</l>
</sp>

```

4. Concurrent hierarchies

Overlapping or non-nesting information is an issue with all XML-based languages, and the TEI is not an exception. An XML document must be a tree structure, with each element under the root nesting in a parent element. This poses a problem when users want to encode “concurrent hierarchies,” that is layers of information that overlap. A typical example of such concurrent hierarchies is the logical and physical structure of a document, i.e., the division into logical parts vs. the division into pages. In this particular case, the potential conflict between these layers of information is solved as we have seen by the use of milestones for the physical structure, to avoid any risk of overlapping elements, but other cases of concurrent hierarchy may occur, requiring the use of different encoding strategies to overcome this difficulty.

The TEI Guidelines support four XML-based methods for handling overlapping information, which are exposed in details.³³ Since, with scholarly editions, the encoding of the critical

28. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-sp.html>

29. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-speaker.html>

30. On the encoding of cast lists, see section [7.1.4 - Cast Lists](#) in the Guidelines

31. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-stage.html>

32. See the full text online: Simon Gréban. *Le Mystère des Actes des Apôtres*. CNRS-Lamop (UMR 8589). [online]. <http://eserve.org.uk/anr/>

33. See chapter 20 of the Guidelines, [Non-hierarchical Structures](#)

apparatus frequently overlaps with the one of citations, we have chosen to use this example to fully develop the explanation of the different strategies available in TEI, at the end of the Citations and references chapter. We direct the readers interested in this topic to the section Handling overlaps of this chapter.

Bibliography

Lou Burnard, *What is the Text Encoding Initiative?*, Marseille, OpenEdition, 2014, DOI: 10.4000/books.oep.426.

Transcription or Diplomatic Edition

Marjorie Burghart

In this chapter you will learn:

- how to encode the transcription or diplomatic edition of a single manuscript document ;
- how to represent various features and interventions, either by the original scribe or by the editor ;
- how to encode a damaged document with poorly legible or illegible text.

When editing or transcribing a text from a single document, as it is often the case for archive documents like charters, ledgers, etc., it is generally desired to render all its features with the highest degree of accuracy, in what is sometimes called a diplomatic or documentary edition. Such editions clearly indicate the layout of the text, follow scrupulously the orthography of the source without trying to regularise it, mention all scribal interventions, etc. The TEI offers a range of solutions to encode such an edition, and also lets you produce more versatile editions, where the absolute respect for the original orthography can coexist with a different view of the document presenting the users with a more accessible, regularised version, for instance. We have already seen in chapter Structure how to represent the structure of a text as well as the layout of a document, so in this chapter we are going to focus on the encoding various features typical of manuscript texts (presence of abbreviations, various interventions, changes of hand, etc.), but also common editorial interventions like pointing out obvious errors in the document and suggesting a correct reading, offering a regularised orthography while keeping the original one, or supplying text that was obviously omitted in the document. Finally, we will deal with the particular case of damaged documents presenting areas with poorly legible or illegible text.

1. Scribal Features and Interventions

1.1. Abbreviations

Abbreviations are a frequent feature of manuscripts, especially in medieval documents. In most critical editions they are silently expanded, but in diplomatic editions however it is common to try and reproduce as accurately as possible all the features of the edited document, including its abbreviations.

Here is an example, containing one abbreviation:

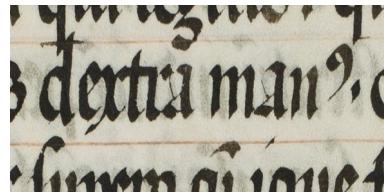


Figure 1: Klosterneuburg, Stiftsbibliothek, CCI 222, fol. 57v

Reproducing the aspect of the abbreviation can be achieved through typography, using a font allowing us to reproduce the aspect of the original writing.¹ This passage reads: “dextra man⁹,” which translates as “right hand.” The last sign at the end of the first word is an abbreviation, a tironian note standing for the suffix “-us” when it is at the end of a word. The passage should therefore be expanded as “dextra manus.” When relying only on typography, we have to choose between the former representation, faithful to the aspect of the original document but less easy to read, and the latter, an expanded version, easier for readers but lacking information about the aspect.

Using the encoding, it is possible to record alternative versions, allowing us to display different information in different contexts, allowing users to switch from a fully diplomatic version to a reading version with expanded abbreviations. To achieve this we can use the **<choice>** element, which “groups a number of alternative encodings for the same point in a text.”² For abbreviations and their expanded form, we are going to use those two children in **<choice>**:

- <abbr>** “(abbreviation) contains an abbreviation of any sort”;
- <expan>** “(expansion) contains the expansion of an abbreviation.”

Our example could be encoded as follows:

```
dextra <choice>
  <abbr>man⁹</abbr>
  <expan>manus</expan>
</choice>
```

This would let us display either “man⁹” or “manus.” But if we want a more precise encoding, showing which letters are the result of an expanded abbreviations, we can use **<ex>** tags within **<expan>**: **<ex>** (editorial expansion) “contains a sequence of letters added by an editor or transcriber when expanding an abbreviation.”

```
dextra <choice>
  <abbr>man⁹</abbr>
  <expan>man<ex>us</ex></expan>
</choice>
```

1. For medieval European texts, see the [Medieval Unicode Font Initiative: http://folk.uib.no/hnooh/mufi/](http://folk.uib.no/hnooh/mufi/), a project coordinating the encoding of non-Unicode characters from medieval texts in the Latin alphabet, including many abbreviation signs.

2. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-choice.html>

With this encoding, we could display not only “man” or “manus,” but also all sorts of variations highlighting the expanded letters: “manus,” “man(us),” etc. This means that our edition will be able to adapt to different needs: readers will be able to have several different views of the text.

1.2. Corrections: Deletions, Additions and Substitutions

Scribes were only humans after all and made mistakes while they were writing documents, and those mistakes may later have been corrected by the original scribe or a later reader deleting or adding text in the source document.

1.2.1. Deletions

According to the definition of the TEI Guidelines, **** (deletion) “contains a letter, word, or passage deleted, marked as deleted, or otherwise indicated as superfluous or spurious in the copy text by an author, scribe, or a previous annotator or corrector.”³ Let us consider the following example:



Figure 2: Klosterneuburg, Stiftsbibliothek, CCI 222, fol. 46v

This Latin line transcribes as follows: “Nulli irascenti ~~iraseenti~~ ira sua uidetur iniusta,” which means “No angry persons ~~angry persons~~ think their anger is unjust.” The scribe, seeing that he had repeated the word “irascenti,” deleted the superfluous occurrence. We could encode this scribal deletion as follows:

Nulli irascenti ****irascenti**** ira sua uidetur iniusta.

There are many ways to delete text from a document: the scribes could for instance strike a line through it as in the example above, scratch the ink, or expunctuate it (drawing dots under the letters or words that were to be deleted). If we are interested in categorising the way a deletion has been performed, we may use the **@rend** attribute on ****. The value of **@rend** is at the discretion of the editor, we therefore recommend that you establish your own list of values for the edition at hand. In our example, a more precise encoding would be:

Nulli irascenti **<del rend="strikethrough">**irascenti**** ira sua uidetur iniusta.

It may happen that a passage has been deleted so successfully that the words are barely legible anymore, or not legible at all. Here for instance, the scribe scraped the text so thoroughly that we cannot read it anymore, but we can estimate that a single word has been deleted:

3. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-del.html>



Figure 3: Klosterneuburg, Stiftsbibliothek, CCl 1195, f. 31v

To encode this deletion, we could proceed as above, but use the **<gap>** tag to represent the illegible deleted word. A **<gap>** “indicates a point where material has been omitted in a transcription, whether for editorial reasons described in the TEI header, as part of sampling practice, or because the material is illegible, invisible, or inaudible.”⁴ With the **@reason** attribute we can specify the reason why we omit material here, and optionally we can also use the dimension (**@unit**, **@quantity**, **@extent**, **@precision**, **@scope**) and ranging attributes (**@atLeast**, **@atMost**, **@min**, **@max**, **@confidence**). Here for instance, if we wanted to indicate that we cannot transcribe the deleted text because it is illegible, but estimate that is a single illegible word, we could have the following encoding:

```
... et non <del rend="scraped">
  <gap quantity="1" unit="words" reason="illegible"/>
</del> fuit qui...
```

It may happen that a deletion overlaps other hierarchies, running across several paragraphs for instance. Let us consider this example, where a deletion has been made across the text, covering the text before a quotation and the beginning of this quotation:

Et ~~ut~~ dicitur in euangelio Iohannis: "In principio erat Verbum, et Verbum erat apud Deum."

Figure 4: A deletion overlapping a quote

The easiest solution is to break down the deletion into two, so the individual part do not overlap any other element:

Et ~~ut~~ dicitur in euangelio Iohannis: <quote>In principio erat Verbum, et Verbum erat apud Deum.</quote>

This is perfectly satisfactory if we simply want to record the fact that some text was deleted. If, however, we are interested in representing the deletions, we may not want to break a single deletion phenomenon into two parts. In this case, we can use a complementary mechanism, allowing us to mark up the beginning and the end of the deleted passage with empty tags. The first, **<delSpan>** “(deleted span of text) marks the beginning of a longer sequence of text deleted, marked as deleted, or otherwise signaled as superfluous or spurious by an author, scribe,

4. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-gap.html>

annotator, or corrector.”⁵ **<delSpan>** has a **@spanTo** attribute pointing to the **@xml:id** of the tag marking up the end of the deleted span, an **<anchor>**:

Et **<delSpan spanTo="#endOfDel1"/>**ut dicitur in euangelio Iohannis:
<quote>In principio erat Verbum, et**<anchor xml:id="#endOfDel1"/>** Verbum
erat apud Deum.**</quote>**

Nota bene: this encoding, using empty tags, is intellectually satisfying but will be more difficult to process than a regular ****.

1.2.2. Additions

Additions, represented in TEI by **<add>**, occur when “letters, words, or phrases inserted in the source text by an author, scribe, or a previous annotator or corrector.”⁶ In the following example, the scribe added the abbreviated words “Tharsis, et filios” in the margin of the document, and used the sign // to indicate where in the text those words should be inserted, that is between the words “filios” and “Israhel.” The original sentence meant “... and he pillaged all the children of Israel,” with the added words it means “... and he pillaged all the children of Tharsis, and the children of Israel.”



Figure 5: Klosterneuburg, Stiftsbibliothek, CCI 3, fol. 20r

A simple encoding of this phenomenon would be the following, putting the added word where it was meant to be inserted and marking it up with **<add>**:

... predauitque omnes filios **<add>**Tharsis, et filios**</add>** Israhel ...

The **@place** attribute on **<add>** lets us indicate where the addition occurred. You can add your own values to the ones suggested by the TEI Guidelines ⁷ We could improve the precision of the encoding by adding the **@place** attribute:

... predauitque omnes filios**<add place="margin">**Tharsis, et filios**</add>** Israhel ...

Of course, if we wanted to also encode the abbreviations, we could combine the two:

... et gaudia **<add place="margin">** **<choice>**
 <abbr>scli**</abbr>**
 <expan>s**<ex>**e**</ex>**c**<ex>**u**</ex>**li**</expan>**
 </choice>
</add> transitoria ...

5. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-delSpan.html>

6. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-add.html>

7. The suggested values are: “*below*: below the line; *bottom*: at the foot of the page; *margin*: in the margin (left, right, or both); *top*: at the top of the page; *opposite*: on the opposite, i.e., facing, page; *overleaf*: on the other side of the leaf; *above*: above the line; *end*: at the end of, e.g., chapter or volume; *inline*: within the body of the text; *inspace*: in a predefined space, for example left by an earlier scribe.”

1.2.3. Substitutions

Sometimes a deletion and an addition have been concomitant phenomena: a scribe has deleted a word in the document to replace it with another, which is added in the same operation. This is what the TEI calls a “substitution,” encoded with the **<subst>** elements which wraps together a **** and an **<add>**. In the following example, the word “dominus” (the Lord) has been deleted by expunction, and the word “diabolus” (the Devil) has been added above the line to replace it.

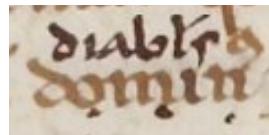


Figure 6: Paris, Bibliothèque nationale de France, Latin 588, f. 8r

The most basic encoding would be:

```
<subst>
  <del>dominus</del>
  <add>diabolus</add>
</subst>
```

We could also use the attributes we have already seen for **** and **<add>**:

```
<subst>
  <del rend="expunction">dominus</del>
  <add place="above">diabolus</add>
</subst>
```

1.2.4. Transpositions

In a transcription, a transposition “occurs when metemarks are found in a document indicating that passages should be moved to a different position.”⁸ This is commonly found in drafts, where full paragraphs may be moved. The words of a sentence are also sometimes marked to be reshuffled.⁹

1.3. Hands

The different scribes who have written a document are identified by their handwriting, or “hand.” Differentiating the various hands may be very important for the study of a document, and therefore its edition. If we want to encode the changes of hands in a document, we must begin with listing and describing the various hands we can identify. To this effect, we must create a

8. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/PH.html#transpo>

9. Note that “transposition” has a different meaning for philologists, who use it to describe a situation in which parts of a text (sections, paragraphs, sentences, words) are in a different order from one witness to another. See chapter Textual Variants

<handNotes> element in the TEI header of the edition, within which we are doing to describe each hand in a **<handNote>**. This element contains a prose description of the hand, and may have the following special attributes:

@scribe gives a name or other identifier for the scribe believed to be responsible for this hand.

@script characterizes the particular script or writing style used by this hand, for example secretary, copperplate, Chancery, Italian, etc.

@scribeRef points to a full description of the scribe concerned, typically supplied by a person element elsewhere in the description.

@scriptRef points to a full description of the script or writing style used by this hand, typically supplied by a scriptNote element elsewhere in the description.

@medium describes the tint or type of ink, e.g., brown, or other writing medium, e.g., pencil

@scope specifies how widely this hand is used in the manuscript.

For instance, if we had a document written by two different scribes in the twelfth century, where two different readers had then added annotations in fifteenth and seventeenth century script respectively, we could have the following declaration:

```
<handNotes>
  <handNote xml:id="scribe01" script="charter_hand" medium="brown-
ink">Charter hand, 12th c., with marked Gothic
characteristics.</handNote>
  <handNote xml:id="scribe02" script="charter_hand" medium="brown-
ink">Charter hand, 12th c., slightly less angular than the first, with
particularly lavish curls on letters a</handNote>
  <handNote xml:id="annotator01" script="informal_cursive"
medium="black-ink">Informal cursive, 15th c., used for adding
annotations in the margins</handNote>
</handNotes>
```

We have now declared the different hands, we can indicate which hand is writing what in two ways.

- The attribute **@hand** may be added to a number of elements, most notably those relating to scribal corrections like **<add>**, **** and **<subst>**.¹⁰ For instance, if an annotation was added by the hand identified above as “annotator01,” we could encode it as follows:

```
<add hand="#annotator01" place="margin">Episcopus
Lincolnensis</add>
```

- The empty tag **<handShift>** can be used to indicate where a change of hands occurs

10. For the complete list, please consult the “Members” section on this page: <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-att.written.html>

in the text. The link between a `<handShift>` and a hand is expressed by `@new`, which contains a link to a `<handNote>`. The attributes `@resp` (person responsible) and `@cert` (level of certainty) may optionally be used to indicate who is responsible for identifying that a handshift occurs, and with what level of certainty. Even in documents with a single hand, there might be some changes, like for instance a paler ink. To encode this, it is possible to use the `<handShift>` tag with the same attributes listed above for `<handNote>`. In the following example, for instance, there is a change of ink but the hand stays the same.

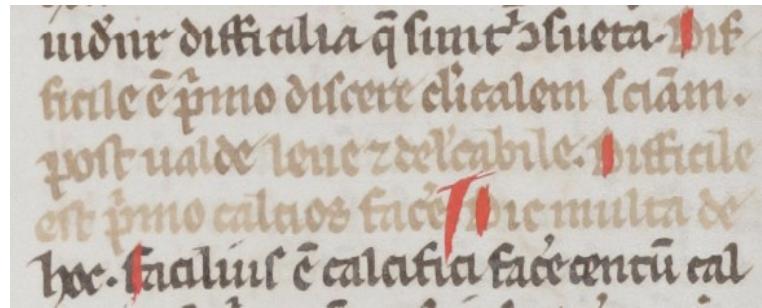


Figure 7: Fribourg, Couvent des Cordeliers / Franziskanerkloster, Ms. 117 I, f. 153r

- We could encode this phenomenon as follows, using the `@medium` attribute on `<handShift>` - note that, since this is the same hand, we do not use the `@new` attribute:

```
... que sunt inconsueta. <handShift medium="brown-ink"/>
Difficile est primo discere clericalem scientiam, post ualde leue
et delectabile. Difficile est primo calcios facere. Dic multa de
<handShift medium="black-ink"/>hoc. Facilius est...
```

1.4. Rendition

In some cases, you might want to encode the aspect or rendition of some elements of the text: words or letters written in a different colour, or underlined, or decorated, etc. To indicate how a textual element has been rendered in the original document we are encoding, the TEI offers the `<hi>` (highlighted) element, which “marks a word or phrase as graphically distinct from the surrounding text, for reasons concerning which no claim is made,”¹¹ and the `@rend` (rendition) attribute, which “indicates how the element in question was rendered or presented in the source text.”¹² The `@rend` attribute is global, which means that it can be used on any TEI element, and not only of `<hi>`. It means that if a highlighted textual element is already marked up (as a word, or sentence, or segment, etc.) there is no need to wrap it in a `<hi>` element: we can simply add a `@rend` attribute to the existing element. When, however, there is no relevant semantic or other markup, we can use the `<hi>` element.

11. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-hi.html>

12. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-att.global.rendition.html>

Let us consider the following example, the beginning of the tenth chapter of the Gospel according to Mark in an early 14th c. manuscript.

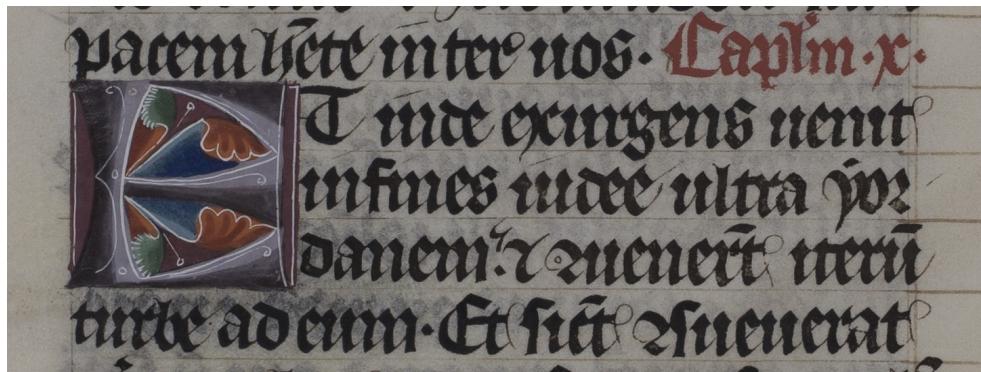


Figure 8: Klosterneuburg, Stiftsbibliothek, CCI 3, fol. 219v

```
<div>
  <head rend="red-ink">Capitulum X</head>
  <p><hi rend="decorated_initial"
    facs="../initial_E_f219v.png">E</hi>t inde exurgens uenit in fines
    Iudee ultra Yordanem, et conuenierunt iterum turbe ad eum, et sicut
    consueverat ... </p>
</div>
```

The whole chapter can be encoded as a `<div>`, and the rubric “Capitulum X” as `<head>`. This rubric is highlighted: it is written in red ink, while the rest of the text is written in black ink. To represent this, we can simply add a `@rend` attribute to the relevant `<head>` element. Note that the TEI does not give a fixed list of values, nor even recommendations for `@rend`: you can create your own list, and it is highly recommended to document carefully your own convention and practice.

The initial of this chapter, a decorated E, is also highlighted. But this time, there is no other reason for this letter to receive markup, other than its being highlighted. We can therefore use the `<hi>` element, combined of course with the `@rend` attribute. As a side note, since this is a graphical element, we can use the `@facs` attribute on `<hi>` to point to an image of the decorated initial.

It is important to notice that `<hi>` and `@rend` must be used to represent how an element is highlighted in the document we are encoding, and not to express how we would like the element to be visually rendered in the output of our edition.

2. Editorial Interventions

We have so far studied the encoding of phenomena and features linked to the work of the scribes who wrote the documents. Another important part of the encoding regards our own editorial interventions: most notably, a critical editor may need to normalise the text, modify it to make it

better intelligible, and foreign words from the rest of the text.

2.1. Normalisation

Editors have the possibility to point (and optionally offer a correction for) apparent errors in the text. They may also choose to offer a regularised version of the text alongside its original spelling. In each case, the the **<choice>** element, which “groups a number of alternative encodings for the same point in a text,”¹³ can be used to combined the original version with the corrected or regularised one.

2.1.1. *Sic*: Pointing Out (and ccorrecting) inaccurate or incorrect Text

The element **<sic>** in TEI “(Latin for ‘thus’ or ‘so’) contains text reproduced although apparently incorrect or inaccurate,”¹⁴ which corresponds to the definition used by philologists, who add the mention *sic* after an unexpected reading. If, for instance, you found the following sentence in a document:

I've been hear before...

As an editor, you should notice that there is a grammar / spelling mistake here: the scribe wrote “hear” instead of “here.” To convey this information to the users of the edition, we could encode this phenomenon as follows, simply pointing out that we deem “they’re” to be erroneous, but without explaining what we think would be the correct version, using only **<sic>**:

I've been **<sic>**hear**</sic>** before...

A more thorough encoding would let the readers know what we suggest should have been the correct version of the erroneous word, using a combination of **<sic>** and **<corr>** (correction)¹⁵ in a **<choice>** element:

I've been **<choice>**
 <sic>hear**</sic>**
 <corr>here**</corr>**
</choice> before...

This encoding could be used to generate a footnote in a print or online version of our edition, or to display different versions of the text.

2.1.2. Original and Regularised Spelling

Early Modern languages often use different spelling, grammar and punctuation rules from the ones to which we are accustomed. In some cases, it might be desirable for the editor to prepare a more easily readable version of a text, to make it more accessible to modern readers. An example could be the full title of this edition of Hamlet, printed in London by Valentine Simmes for

13. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-choice.html>

14. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-sic.html>

15. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-corr.html>

Nicholas Ling and Iohn Trundell, in 1603:

The tragicall historie of Hamlet Prince of Denmarke
by William Shake-speare

The modern reader might stumble upon the original spelling, and therefore it might be desirable to offer a possibility to access a version of the text offering regularised spelling.¹⁶ To this end, we are going to use, within a **<choice>** element, a combination of **<orig>** (original form) which “contains a reading which is marked as following the original, rather than being normalized or corrected,”¹⁷ and **<reg>** (regularisation) which “contains a reading which has been regularized or normalized in some sense.”¹⁸ The example above could be encoded like this:

```
The <choice>
  <orig>tragicall</orig>
  <reg>tronical</reg>
</choice>
<choice>
  <orig>historie</orig>
  <reg>history</reg>
</choice> of Hamlet Prince of <choice>
  <orig>Denmarke</orig>
  <reg>Denmark</reg></choice> by William <choice>
  <orig>Shake-speare</orig>
  <reg>Shakespeare</reg>
</choice>
```

2.2. Editors Adding or Skipping Text

It is paramount for scholars to respect the text they are editing, and not to start “rewriting it” for any reason. But in some cases, when a word appears to be missing in the text, or has been erroneously repeated, the editor can intervene, but always distinguishing their emendation from the original text.

2.2.1. Supplying an Omitted Word

When the scribe has omitted to write down a word or group of words, the text may become hard to understand. But if the editors can make an educated guess as to what this word or words might have been, they may supply the missing text while clearly signalling that the supplied text is their own suggestion. In the following example for instance, the sentence has no verb and it seems clear that the scribe forgot to write it down:

16. To see a working edition using this feature, see for instance François-Joseph Bérardier de Bataut, *Essai sur le récit, ou entretiens sur la manière de raconter* (Paris: Berton, 1776). Édition électronique sous la direction de Christof Schöch, URL: <http://www.berardier.org>, 2010 (version 0.6, 12/2010).

17. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-orig.html>

18. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-reg>

I only twelve when the war started.

An educated guess as to the missing verb would lead us to supply “was” just between “I” and “only.” To encode this, we could use the **<supplied>** element, which “signifies text supplied by the transcriber or editor for any reason.”¹⁹ Note that here we are considering the use of **<supplied>** for obvious omissions by the scribe, but it can also be used to encode a damaged or poorly legible document. For this type of use, see section [The Damaged Document](#) below.

The **<supplied>** element can be used with several useful elements:

- **@reason** can be used to indicate why a word had to be supplied here. In the case at hand, a value suggested by the Guidelines would be with “omitted-in-original.” Note that, in the value of this attribute, whitespaces are considered as separators, so the phrases describing the reason are hyphenated.
- **@source** “provides an attribute used by elements to point to an external source”²⁰ It can be useful when you guess the missing word based on another text. It happens typically when the omitted word occurs in a quotation of a standard text, like a biblical verse.
- The responsibility attributes, **@resp** (responsible party) and **@cert** (certainty), can be used to indicate, respectively, to indicate “the agency responsible for the intervention or interpretation, for example an editor or transcriber” and “the degree of certainty associated with the intervention or interpretation”²¹ You may use either of these elements alone, or both combined. The value of **@resp** must be one or more “data pointer” i.e., pointers to the **@xml:id** of an element describing each a person responsible for this suggestion. The value of **@cert** must be “high,” “medium,” or “low.”

In the light of this information, here is how we could encode the phenomenon described in the example above: the reason we have to supply a word if that it was omitted in the original, and given basic grammar rules the level of certainty of our hypothesis is very high. There is no external source supporting our hypothesis, and finally let us assume that we are not interested in recording the “responsible party” for this intervention since we are the only editor. The result would be the following encoding:

I **<supplied reason="omitted-in-original" cert="high">was</supplied>**
only twelve when the war started.

2.2.2. Cutting Out a Redundant Word

The opposite phenomenon occurs when a scribe has inadvertently added superfluous words to a sentence. This typically happens when a word is repeated by mistake, as in the following example:

I was was only twelve when the war started.

19. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-reg>

20. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-att.global.source.html>

21. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-att.global.responsibility.html>

To mark up the superfluous word (or words), we can use the **<surplus>** element, which “marks text present in the source which the editor believes to be superfluous or redundant.”²² The same attributes listed above for **<supplied>** are available for **<surplus>**. The example could be encoded as follows:

I was **<surplus reason="repeated" cert="high">was</surplus>** only twelve when the war started.

2.2.3. Words in a Different Language

It is common in editions to highlight, usually with italics, the words belonging to a different language from the surrounding text. To encode this phenomenon, we use the **<foreign>** element²³ combined with the **@xml:lang** attribute. The value of **@xml:lang** is a tag corresponding to the BCP 47 rules.²⁴ The language code should be taken from the list of codes registered by the Internet Assigned Numbers Authority (IANA).²⁵

For instance, in this liturgical treatise in Latin, the author has inserted a few Greek words in a sentence that means “Then follows *Kyrie Eleison*, which is sung by the choir”:

Deinde sequitur *Kύριε ἐλέησον*, quod a choro concinitur.

We are going to wrap the Greek words in a **<foreign>** tag, and add an **@xml:lang** attribute with the IANA-registered code for Ancient Greek (pre-1453), “grc”:

Deinde sequitur **<foreign xml:lang="grc">***Kύριε ἐλέησον***</foreign>**, quod a choro concinitur.

Note that, if you have large parts of your text in different languages, like full paragraphs or sections, **<foreign>** is not the right solution. In this case, you should simply use the **@xml:lang** attribute on the **<p>** and **<div>** elements, and use **<foreign>** only for words or short phrases within those paragraphs. Clearly marking up the language of each part of the text is important if you consider processing your edition: lexical statistics and stylometry, for instance, can only be accurate if performed on parts written in the same language.

3. The Damaged Document

Working with ancient documents means that we often have to deal with the damage inflicted upon them through the course of their long history: tear and wear, vandalism, fire, water, mildew... Many aggressions may have taken their toll, and made the document impossible, or at least difficult to read in some parts. In this section, we will see how to encode such parts of the document.

When some part of a document bears text that is poorly legible or illegible because it the

22. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-surplus.html>

23. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-foreign.html>

24. <https://tools.ietf.org/html/bcp47>

25. <http://www.iana.org/assignments/language-subtag-registry/language-subtag-registry>

document has been damaged, or because the text has been deleted, we use a combination of the following elements:

- **<damage>** “contains an area of damage to the text witness”²⁶ We use it to mark up damaged areas of a document, whether or not it altered the legibility of the text. Besides many others, **<damage>** may have specific attributes,²⁷ among which **@agent** to indicate what caused the damage. Note that, just as for ****, there is a mechanism to allow the encoding of damage spanning over different hierarchies. If you need to use it, check the description of the **<damageSpan>** element in the Guidelines,²⁸ and see the explanation of the similar **<delSpan>** mechanism at the end of section [Deletions](#)
- **** (deletion), as we have seen above (section [Deletions](#)), “contains a letter, word, or passage deleted, marked as deleted, or otherwise indicated as superfluous or spurious in the copy text by an author, scribe, or a previous annotator or corrector.” Such deletion may result in some illegible or poorly legible text.²⁹
- **<gap>** “indicates a point where material has been omitted in a transcription”³⁰ We use it to mark the places where we are not able to provide the text that used to be borne by the document. With the **@reason** attribute we can specify the reason why we omit material here.
- **<unclear>** “contains a word, phrase, or passage which cannot be transcribed with certainty because it is illegible.”³¹ We use it to mark up text that has been damaged or deleted but remains partially legible, and for which we may propose a tentative reading.
- **<supplied>**, as we have seen above (section [Supplying an Omitted Word](#)), “signifies text supplied by the transcriber or editor for any reason,”³²

The following decision tree may help you to figure how to combine those tags:

26. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-damage.html>
27. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-att.damaged.html>
28. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-damageSpan.html>
29. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-del.html>
30. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-gap.html>
31. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-unclear.html>
32. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-supplied.html>



Figure 9: Decision Tree: How to Encode Deleted or Damaged Text

If we are interested in describing more precisely the extent of the gap, deletion or damage, we can use the we can also use on all the elements listed above the dimension (@**unit**, @**quantity**, @**extent**, @**precision**, @**scope**)³³ and ranging attributes (@**atLeast**, @**atMost**, @**min**, @**max**, @**confidence**).³⁴ For instance, if we wanted to indicate that we cannot transcribe some text because damage made it illegible, but estimate that there were between eight and ten words, we could have the following encoding:

```

<damage agent="rubbing">
  <gap atLeast="8" atMost="10" unit="words" reason="illegible"/>
</damage>
  
```

Or, if we wanted to indicate also that, for the same area, the extent of the damage is 3 inches, we could encode it like this:

```

<damage agent="rubbing" extent="3" unit="inch">
  <gap atLeast="8" atMost="10" unit="words" reason="illegible"/>
</damage>
  
```

As you can see, there is a great many possibilities. The important thing is to decide what information is important to you: are you interested in recording the extent of the damage, or just its presence? This will help you choose your own encoding rules for your edition. Let us now consider the four different possible situations in more detail.

33. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-att.dimensions.html>

34. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-att.ranging.html>

3.1. The Text is Illegible, No Text is Supplied

If the text is illegible and cannot be supplied, we simply use <gap>. The following attributes can be particularly useful with this element:

@reason which “gives the reason for omission. Sample values include sampling, inaudible, irrelevant, cancelled.”

@agent, which “in the case of text omitted because of damage, categorizes the cause of the damage, if it can be identified.”

Let us consider the following example: the ornate initial that was on the recto of this folio has been cut out. As a result, the text that was on the verso of this initial is missing and cannot be supplied:



Figure 10: Bern, Burgerbibliothek, Cod. A 9, f. 3v (www.e-codices.unifr.ch)

We could encode this transcription as follows (note that I have encoded the line breaks with <lb>, using @break with the value “no” when the line break occurs within a word):

```

<lb break="no"/>mini et replete terram, et subic<gap reason="cut-
out" extent="9 words"/>
<lb break="no"/>bus maris, et uolatilibus celi <gap reason="cut-out"
extent="4 words"/>
<lb/>mouentur super terram. Di<gap reason="cut-out" extent="6
words"/>
<lb break="no"/>bam afferentem semen super<gap reason="cut-out"
extent="6 words"/>
<lb break="no"/>bent in semetipsis sem<gap reason="cut-out"
extent="8 words"/>

```

3.2. The Text is Illegible, but the Text is Supplied

If the text is illegible but can be supplied, we are going to mark up the damaged or deleted area with **<damage>** or **** respectively, depending on the situation. Within **<damage>** or ****, we are going to put a **<supplied>** element containing the text we propose to supply. On **<supplied>**, we can also use the attributes we listed above (see section [Editors Adding or Skipping Text](#)).

Let us consider this example: this 11th c. Bible manuscript has been slightly damaged, so that the final words of this line have been lost:

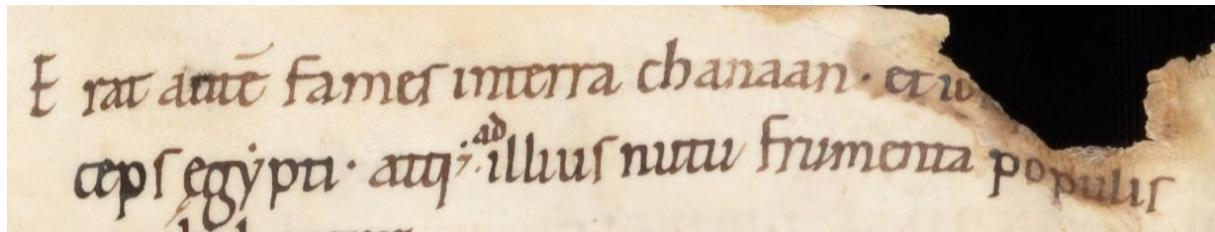


Figure 11: Sion/Sitten, Archives du Chapitre/Kapitelsarchiv, Ms. 15 – Giant Bible, f. 17r
www.e-codices.unifr.ch

Note that we do not have to use any attribute on **<damage>**. Remember to encode only the information you need and plan to use or process for your edition. In this case, we will assume that we are not interested in studying further the types of damage. Should you be interested in doing so, keep in mind that you can use on **<damage>** the **@reason** and **@agent** attributes we described above for **<gap>**, but also a group of attributes specific to **<damage>**,³⁵ and the whole range of attributes dedicated to describe the dimensions of the damaged area³⁶ Since it is a well-known text (here the book of Genesis), and only a very short passage was lost, we may choose to supply the missing word using the text of the Vulgate, the Latin Bible:

35. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-att.damaged.html>

36. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-att.dimensions.html>

```
<lb/>Erat autem fames in terra Chanaan et Io<damage>  
<supplied source="#Vulgate">seph erat prin</supplied>  
</damage><lb break="no"/>ceps Egypti...
```

3.3. The Text is Partially Legible

When, in a damaged or deleted area, the text is partially legible, we can use a combination of the following elements:

- We will use **<damage>** or **** to mark up the damaged or deleted parts.
- Within the damaged or deleted area, we will use **<gap>** to mark up the illegible part(s), or **<supplied>** if we wish to supply text.
- Also within the damaged or deleted area, we will use **<unclear>** to mark up the parts of the text which cannot be transcribed with perfect confidence.

Let us consider this deleted sentence, where the text has been blackened out. Most of the text is illegible, but we can venture a reading for the last word of the first line, and the first of the second line (I invite you to check for yourself on the high-definition image available online on the e-codices.unifr.ch website):



Figure 12: Fribourg, Couvent des Cordeliers/Franziskanerkloster, Ms. 117 I – Berthold of Regensburg, Sermones, f. 117v (www.e-codices.unifr.ch)

To encode this passage, we can combine within a **** the **<gap>** element for illegible text, and the **<unclear>** element for our tentative readings. On **<unclear>**, we can use the **@reason** and **@cert** attributes we saw above:

```
... peruenit. <del>  
  <gap reason="deletion" extent="4 or 5 words"/>  
  <unclear reason="blackened-out" cert="middle">sunt</unclear>  
  <lb/><unclear reason="blackened-out" cert="high">discreti</unclear>  
  <gap reason="deletion" extent="3 to 5 words"/>  
</del> Tertio ...
```

3.4. The Text is Fully Legible

When the damaged or deleted text is still perfectly legible, we can simply use the **<damage>** or **** elements and their attributes, without additional tags.

Manuscript Description

James Cummings

This chapter investigates the creation of manuscript descriptions for digital editions through looking at the recommendations of the *Guidelines of the Text Encoding Initiative* for manuscript description. By detailing the methodology of encoding a manuscript description, we examine the basic categories and level of detail necessary to produce a competent scholarly description of the object or objects that are the source for our editions. This chapter looks at when manuscript descriptions are provided and what forms these take. It looks at how to encode such descriptions (using the TEI) to identify a manuscript, record the manuscript contents, detail the physical description, document its history, and provide additional information. The chapter concludes with brief thoughts on publishing manuscript descriptions.

1. When to describe manuscripts

In the creation of digital editions you may question when it is appropriate to fully describe a manuscript instead of just having a witness description as mentioned in the chapter on Textual Variants. It is possible to create an edition from one or more variant witnesses and merely provide a <witness> element to supply a brief bibliographic reference, making it possible for readers to locate the manuscript if necessary. This is equivalent to providing a brief bibliographic reference for a printed work in that it provides readers with the basic information needed to look up the copy in other sources, but does not give details about its contents, physical manifestation, history or overall structure.

The primary reason for providing a manuscript description is that a basic bibliographic reference is not sufficient to contain the information or represent the object according to the needs and uses we have for the description. Modern printed books are usually adequately described by a formulaic bibliographic reference, the conventions of which are very familiar to most readers. The provision of information such as the title, authors, editors, publisher, publication place, date, and perhaps a cited range of pages for an individual contribution are sufficient to represent the part of the physical object to which we are referring, because we accept the illusion that one printed book is, for our needs, pretty much identical to the other copies of it. This is, of course, not actually true at very precise levels -- different printings, or even copies from the same printing may have quite striking physical differences. The reason we accept this comforting illusion is because these differences are not significant for our use of the bibliographic work.

A description is more suitable for a manuscript than a reference because manuscripts are inherently unique objects -- as much as one scribe might try to faithfully copy a source there will

always be differences, not only of content, but also of production of the physical object, which are interesting in themselves to scholars. This is true of manuscripts of all time periods, content, language, writing systems, and form. Moreover, it is also true for certain categories of printed objects, such as incunabula or modern book art, which are not adequately described by bibliographic conventions because of their structure, content, or even the sentiment attached to them. The recommendations made here, and indeed in the TEI Guidelines, apply equally to any other text-bearing objects that need a more detailed level of description, whatever their form or method of creation.

In an ideal world, manuscript descriptions would be provided for any manuscript worthy of study in itself. Large collections of manuscript descriptions in a comparable format such as TEI XML enable larger scale fields of study such as computational codicology, and it should therefore go without saying that, if a manuscript is important enough that a scholarly digital edition is being created, then a manuscript description should be crafted to describe it. This does not necessarily mean that the digital editor must always provide this manuscript description themselves; many libraries, archives and other resource-holding institutions are increasingly providing descriptions that a digital edition could reference. In some cases the underlying (hopefully TEI XML) data is freely available and thus could be referenced or included in a digital edition (with appropriate attribution) and potentially improved. If a description is not already available, then a digital editor should ideally create a manuscript description and make it available to both the resource-holding institution and readers of the edition.

2. Forms of manuscript description

Manuscript descriptions in themselves can take many forms, depending on both the tradition of description, the context of the description, and how it has been created. Some descriptions include lengthy discursive explanations of quite some size, distilling all that is currently known about the creation, history, and nature of the object and its context. Alternatively, others may consist of a mere summary record, providing little more than the location and identification information for the manuscript.

In many cases the amount of detail in a description is dependent on the context of its creation, for example manuscript descriptions created as part of a library finding aid may be less detailed than those crafted as part of an academic edition. How these descriptions are intended to be used will impact what details are recorded. For example a manuscript description used for a digital edition might be more or less detailed in certain aspects, depending on the concerns and priorities of that edition (for example it could be an edition focussed on the physical aspects of the textual object, or part of a dossier génétique).

One use of manuscript description is not as metadata for a digital edition but as part of a sustained discussion concerning one or more manuscripts in a catalogue raisonné, belle lettrist or other academic secondary study. However, in a digital edition a description can be used to document, locate, and describe various aspects of the edition and its physical source. The TEI aims to cope with these and other possibilities so is flexible enough to allow manuscript descriptions to appear not only as metadata but also inside and alongside paragraphs. Their use will differ depending on the aim, for example in a digital edition or a catalogue raisonné. In a

digital edition a manuscript description usually is provided as metadata for the source text of the edition. This enables encoders of manuscript descriptions using the TEI to both describe manuscripts in a manner which suits their needs and also do so in a systematic way suitable for further processing and analysis.

3. Manuscript description in the TEI

As described above, the methods of describing manuscripts provided by the TEI are designed to cope with the variation in the context of their production. It is quite common for institutions creating catalogues of manuscript descriptions to be retrospectively converting them from print, existing cataloguing systems, other markup formats (such as EAD XML), or even bespoke database systems. They may be converting them to benefit from the increased expressivity available through the TEI or to consolidate a range of catalogues all under the same processing system. It should be noted that the TEI recommendations for manuscript description, and indeed the rest of the TEI, are a moving target as the community continues to update, improve, and revise them. The advice given here reflects the state of the manuscript description module at the time of writing.

As the results of retrospective conversion are unpredictable as regards the granularity of the different aspects of a manuscript description, the TEI allows for a great deal of variation in the structure of manuscript descriptions. The overall structure is a `<msDesc>` element which itself is required to contain an `<msIdentifier>` usually containing its geographical and archival location, and manuscript identification information. While this identifying metadata usually contains full geopolitical and repository information, it could also contain only a manuscript name if that is all we know about that manuscript being described.

```
<msDesc xml:id="MySampleManuscript" xml:lang="en">
  <msIdentifier>
    <msName>My Manuscript</msName>
  </msIdentifier>
</msDesc>
```

This `<msIdentifier>` can be followed by one or more paragraphs, or more structured information, depending on the source of the data which itself could be more or less structured.

```
<msDesc xml:id="MySampleManuscript" xml:lang="en">
  <msIdentifier>
    <msName>My Manuscript</msName>
  </msIdentifier>
  <p>One or more paragraphs concerning various aspects of the manuscript</p>
</msDesc>
```

If the source of information can be fragmented along the categories of its intellectual contents, physical description and history, then it should be, and the TEI caters for this by allowing paragraphs inside each of these grouping elements.

```

<msDesc xml:id="MySampleManuscript" xml:lang="en">
  <msIdentifier>
    <msName>My Manuscript</msName>
  </msIdentifier>
  <msContents>
    <p>One or more paragraphs concerning the manuscript's contents</p>
  </msContents>
  <physDesc>
    <p>One or more paragraphs concerning the manuscript's physical description</p>
  </physDesc>
  <history>
    <p>One or more paragraphs concerning the manuscript's history</p>
  </history>
</msDesc>

```

In this way, manuscript descriptions generated from other sources with either more or less structure, can still be considered TEI. However, it is generally recommended that for a digital edition, each of these main sections inside of `<msDesc>` should be as fully complete as possible, using the elements described below. While it should go without saying that a manuscript description with greater structure and precision is preferable, there is of course a balance between the time cost in doing so, versus the benefit for a particular edition.

When the TEI manuscript description module is loaded in the particular TEI customization being used, additional elements are available. What this means is that any paragraph or phrase-level content is able to contain additional manuscript description elements such as: `<catchwords>`, `<dimensions>`, `<heraldry>`, `<locus>`, `<locusGrp>`, `<material>`, `<objectType>`, `<origDate>`, `<origPlace>`, `<secFol>`, `<signatures>`, `<stamp>`, `<watermark>`. Although these are allowed to be used in most phrase-level content, it is clear that some of them make more sense in particular sections, however, this enables greater flexibility in descriptions.

3.1. Identifying manuscripts

All manuscripts must be identified in some way. This could be a name, even one given to it locally, but more usually this is a particular repository with a standardised institutional shelfmark. In the TEI the provision of an `<msIdentifier>` is required, and this must have at least one form of identification information (such as an `<idno>` or `<msName>`) in it. It is more usual to provide the full geo-political location information (such as `<country>`, `<region>`, and `<settlement>`), then the repository information (such as `<institution>` and `<repository>`), and then finally the object identifier information (such as `<collection>`, `<idno>`, and `<altIdentifier>`). This is a traditional tri-partite full manuscript identification structure.

```

<msIdentifier>
  <country>United Kingdom</country>
  <region type="county">Oxfordshire</region>
  <settlement>Oxford</settlement>
  <institution>University of Oxford</institution>
  <repository>Bodleian Library</repository>
  <collection>Digby</collection>
  <idno type="shelfmark">MS. Digby 133</idno>
  <altIdentifier type="internal">
    <idno type="SCN">1734</idno>
  </altIdentifier>
</msIdentifier>

```

In the manuscript identifier above, we locate the manuscript as in the United Kingdom, in Oxfordshire, in Oxford, at the University of Oxford, and inside that at the Bodleian Library, as part of the Digby Collection. We provide a canonical shelfmark (MS. Digby 133) by which the manuscript is most commonly known, but also the alternative identifier of a number by which the manuscript was once known internally. It is possible for manuscript identifiers to contain as many alternative identifiers, or manuscript names, as needed.

```

<msIdentifier>
  <country>United Kingdom</country>
  <region type="county">Oxfordshire</region>
  <settlement>Oxford</settlement>
  <institution>University of Oxford</institution>
  <repository>Bodleian Library</repository>
  <idno type="shelfmark">MS. Auct. T. inf. 1. 10</idno>
  <altIdentifier type="internal">
    <idno type="SCN">28118</idno>
  </altIdentifier>
  <msName xml:lang="la">Codex Ebnerianus</msName>
</msIdentifier>

```

It is possible to use the `@xml:lang` attribute to indicate the language of any element's content but inside an `<msIdentifier>` this is most commonly used to identify the language of variant names of the manuscript. In the example below, where the manuscript identifier is written in Danish, the name of the manuscript is given in Latin and Icelandic as these are the two names by which the manuscript is usually known.

```

<msIdentifier xml:lang="da">
  <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>

```

In general as many details should be provided in the `<msIdentifier>` as are necessary to be able to easily locate the manuscript, however, there is no reason not to include additional alternative identifiers if available.

In some cases, additional information giving an overall summary of the manuscript description (the kind of thing that might have appeared in a `<witness>` element if a full manuscript description was not being created) follows the `<msIdentifier>` element. For historical reasons this uses the standard TEI `<head>` element (usually for headings) to provide a general heading for the entire description.

```

<msDesc>
  <msIdentifier>
    <country>United Kingdom</country>
    <region type="county">Oxfordshire</region>
    <settlement>Oxford</settlement>
    <institution>University of Oxford</institution>
    <repository>Bodleian Library</repository>
    <idno type="shelfmark">MS. Lat. th. e. 46</idno>
    <altIdentifier type="internal">
      <idno type="SCN">Not in SC (late accession)</idno>
    </altIdentifier>
  </msIdentifier>
  <head>Miscellaneous theological works; English, 13th century, first quarter</head>

```

In general it is preferable that this information be stored throughout the manuscript description and that the processing for display, analysis, or interchange extracts this as necessary. However, in systems that can make use of it, providing this summary `<head>` may be useful.

3.2. Recording the intellectual contents

One of the reasons for creating a manuscript description is to detail the intellectual contents of a manuscript. To do this the TEI uses the `<msContents>` element with either structured `<msItemStruct>` or more usually with the less rigorously structured `<msItem>` elements. These provide bibliographic and other information concerning each of the content items in the manuscript.

```

<msContents>
  <msItem n="1">
    <locus from="1r" to="4v">fol. 1r - fol. 4v</locus>
    <p>Description of this item</p>
  </msItem>
  <msItem n="2">
    <locus from="5r" to="55v">fol. 5r - fol. 55v</locus>
    <msItem n="2.1">
      <locus from="5r" to="15v">fol. 5r - fol. 15v</locus>
      <p>Description of this sub-item</p>
    </msItem>
    <msItem n="2.2">
      <locus from="16r" to="55v">fol. 16r - fol. 55v</locus>
      <p>Description of this sub-item</p>
    </msItem>
  </msItem>
  <msItem n="3">
    <locus from="56r" to="109r">fol. 56r - fol. 109r</locus>
    <p>Description of this item</p>
  </msItem>
</msContents>

```

In this hypothetical example, only paragraphs have been provided alongside a `<locus>` element whereas more specific elements could be provided instead. Here three `<msItem>` elements are shown, with the second consisting of two sub-items. Each manuscript item is given a `<locus>` element with both machine-processable `@from` and `@to` attributes as well as an optional human-readable version. This could be used by later processing to create a table of contents linking to the items, or surrogates of them based on their folio numbers. If a `<locus>` is given in the `<msItem>` it must be given before other elements. While it is possible to provide only a paragraph of information, it is more usual for `<msItem>` elements to contain more specific bibliographic information such as `<author>`, `<title>`, and `<textLang>`.

```

<msContents>
  <msItem n="1" xml:id="a1">
    <author ref="http://viaf.org/viaf/10428840">
      <persName>John Mirk</persName>
    </author>
    <title>Festival</title>
    <textLang mainLang="eng">English</textLang>
  </msItem>
  <msItem n="2" xml:id="a2">
    <author ref="http://viaf.org/viaf/10428840">
      <persName>John Mirk</persName>
    </author>
    <title>Instructions for parish priests</title>
    <textLang mainLang="eng">English</textLang>
  </msItem>
  <msItem n="3" xml:id="a3" class="#sermones">
    <title type="desc">Sermon</title>
    <textLang mainLang="eng">English</textLang>
  </msItem>
  <msItem n="4" xml:id="a4">
    <author ref="http://viaf.org/viaf/35717526">
      <persName>Richard Lavinham</persName>
    </author>
    <title>Treatise on the Seven Deadly Sins</title>
    <textLang mainLang="eng">English</textLang>
  </msItem>
  <msItem n="5" xml:id="a5" class="#sacramenta">
    <title type="desc">Treatise on confession</title>
    <textLang mainLang="eng">English</textLang>
  </msItem>
</msContents>

```

In `<msContents>` above, there are five `<msItem>` elements each providing an author, title, and language of the item, if known. In this case the manuscript items are numbered and have an `@xml:id` attribute for processing reasons. Where an author is known for the item, it has been linked to VIAF (the Virtual International Authority File).¹ Inside the `<author>` elements this catalogue has further encoded names with `<persName>` elements (this is strictly unnecessary, but provides consistency across a collection of up-converted descriptions). The `<title>` element sometimes uses its `@type` attribute to indicate whether this is a real title (no `@type` attribute provided) or an editor's description of the work (a value of 'desc'). Some of the `<msItem>` elements also have a `@class` attribute, which points to more information about the text type or other classifications. The `<msItem>` elements also contain `<textLang>` elements detailing the language(s) of that particular item.

¹ The Virtual International Authority File (<http://viaf.org/>) combines multiple authority files for names into a single resource hosted by the Library of Congress in the USA. By referencing this it disambiguates this name from any other forms.

```

<msItem n="2">
  <locus from="132r" to="134v">(fol. 132r–134v)</locus>
  <title>List of emperors</title>
  <note>with the lengths of their reigns, as far as Alexios III, after which is
    recorded the Latin capture of CP.</note>
  <incipit xml:lang="grc">τῇ ἀπὸ κτίσεως κόσμου ,εώκα'</incipit>
  <explicit xml:lang="grc">καὶ ἀφηρέθη ἐξ αὐτῶν παρὰ τῶν Λατίνων ἡ Κπολις</explicit>
  <note><hi rend="italic">pr. </hi>O. Lampsides, <title>op . cit.</title>, pp.
    353–5.</note>
  <textLang xml:lang="grc">Greek</textLang>
</msItem>

```

As in the earlier example, this **<msItem>** gives the location of the item in the manuscript using the **<locus>** element, in this case formatted with parentheses for a specific display output. In this case, additional notes, an incipit, and an explicit also have been supplied. The second note is really a bibliographic citation, and so would be better in a **<bibl>** element, but here is produced because of the migration of legacy data.

Many TEI manuscript descriptions are created from existing legacy records, or are destined for specific systems and so the TEI allows significant flexibility as seen above. For a digital scholarly edition however, one should always use the appropriate elements at the most reasonable level of granularity. The **<msContents>** element could be used by a front-end developer, to generate a table of contents for the manuscript, or to enable resource discovery, filtering by various aspects, searching, or browsing over a collection as a whole, a single edition or images of an individual manuscript.

3.3. Describing the physical object

After the intellectual contents, one of the most important aspects of a manuscript description is a record of the physical object. Indeed, for those interested in larger-scale computational codicology this may be vital. While a **<physDesc>** element may just have paragraphs, these could also be embedded at lower levels for the description of the object (and its support and layout), hands, type (if applicable), scripts, music, decoration, additions, binding, seals, and accompanying matter.

```

<physDesc>
  <objectDesc>
    <supportDesc>
      <p>...</p>
    </supportDesc>
    <layoutDesc>
      <p>...</p>
    </layoutDesc>
  </objectDesc>
  <handDesc>
    <handNote><p>...</p></handNote>
  </handDesc>
  <typeDesc>
    <typeNote><p>...</p></typeNote>
  </typeDesc>
  <scriptDesc>
    <scriptNote><p>...</p></scriptNote>
  </scriptDesc>
  <musicNotation><p>...</p></musicNotation>
  <decoDesc>
    <decoNote><p>...</p></decoNote>
  </decoDesc>
  <additions><p>...</p></additions>
  <bindingDesc>
    <binding>
      <p>...</p>
    </binding>
  </bindingDesc>
  <sealDesc>
    <seal>
      <p>...</p>
    </seal>
  </sealDesc>
  <accMat>
    <p>...</p>
  </accMat>
</physDesc>

```

Many of these sections of the physical description can incorporate more details or structure provided, some of which are discussed below.

3.3.1. Physical Description: the <objectDesc> element

In describing the physicality of the manuscript one of the most important sections is the general <**objectDesc**> element. This contains elements for the description of its support, that is the physical material or object which supports the written part of the manuscript -- other aspects of

support such as the binding or seals are handled separately. A description of the layout of the writing can also be provided inside the **<objectDesc>**.

```
<objectDesc>
  <supportDesc>
    <support><p>Description of the physical support, materials, etc.</p></support>
    <extent>The size of the manuscript using convenient dimensions</extent>
    <foliation><p>Description of the foliation or other numbering of
      surfaces</p></foliation>
    <collation><p>Description of the physical arrangement of folios, including
      collation formulas and discussion of catchwords</p></collation>
    <condition><p>Description of the physical condition</p></condition>
  </supportDesc>
  <layoutDesc>
    <summary><p>Summary description of the layout if desired</p></summary>
    <layout columns="1" ruledLines="20" writtenLines="20">
      <p>One or more descriptions of layout, including columns, ruling, pricking,
        written lines, etc.</p>
    </layout>
  </layoutDesc>
</objectDesc>
```

The hypothetical example uses a **<support>** element to describe the physical support for the writing of the manuscript. Inside this it is possible to have either phrase-level description mixing text and markup, or in the case of longer descriptions of the source, multiple paragraphs of description. Inside the **<supportDesc>** in addition to describing the support generally, there are specialised elements for describing the extent (the size of the manuscript support), the foliation (how the surfaces are numbered), the collation (the arrangement of folios), and the overall physical condition of the manuscript.

Inside the **<objectDesc>** the **<layoutDesc>** element can be used to group one or more **<layout>** elements, which describe how the text is laid out on the surface (e.g. whether the text is in columns, the number of ruled or written lines, description of surface pricking or any other aspects of the text layout). A real-world example might have more or less of this information:

```

<objectDesc form="codex">
  <supportDesc material="perg">
    <support><material>Parchment</material>, of mediocre quality, with numerous flaws,
      irregular edges, and pronounced hair-/flesh-side differences.</support>
    <extent>iii (modern card/paper) + ii (medieval parchment) + 168 + ii (medieval parchment).
      <dimensions type="leaves" unit="mm">
        <height max="305" min="300">300-5</height>
        <width>220</width>
      </dimensions>
      <dimensions type="ruled" unit="mm">
        <height min="234" max="239">234-9</height>
        <width min="163" max="146">163-4</width>
      </dimensions>
    </extent>
    <foliation>Foliated in modern pencil, i-v, 1-170
      (a former foliation left fols. i-iii blank, and numbered
      fols. iv, and v as 'i' and 'ii'; this was altered, 5 Aug. 1997);
      a 15th-/16th-century hand foliated fols. 1-4 in ink in the
      centre of the top margin, in arabic numerals.</foliation>
    <collation>
      <formula>I-XIV<hi rend="sup">12</hi> (fols. 1-168)</formula>;
      <catchwords>catchwords in all quires except the last, in the script of the
      main text, a little below the right-hand side of the right-hand column of text;</catchwords>
      <signatures>no leaf signatures visible</signatures>.
    </collation>
  </supportDesc>
  <layoutDesc>
    <layout ruledLines="45 46" columns="2">Ruled with 45 (Quires I-II, fols. 1-24) or 46
      (Quires III-XIV, fols. 25-168) lines of text in 2 columns, in brown ink (Quires I-II),
      and in grey/brown 'pencil'/'crayon' (Quires III-XIV), each column with single vertical
      bounding lines extending the full height of the page, the top and bottom horizontal
      lines extending the full width; fols. 162v-168r ruled with a considerably narrower
      inter-columnar space; each column 74-5 mm. wide.</layout>
  </layoutDesc>
</objectDesc>

```

In this example, the description provides a great deal of information. It has used both the **@material** attribute and **<material>** element to provide both an easy processable standard form ('perg') and a human-readable version ('parchment'). The **<extent>** element provides a mixture of textual description and dimensions for the leaves and ruled sections. There is a description of the foliation and the collation that includes not only a brief collation formula using the **<formula>** element, but also information about the **<catchwords>** and **<signatures>** of the text. There is only one major layout format for the manuscript, although this varies from 45 to 46 ruled lines, and the amount of inter-columnar space varies. It would also be possible to separate these into individual **<layout>** elements if it was needed to record a more distinct separation of layouts.

3.3.2. Physical Description: **<handDesc>**, **<typeDesc>**, **<scriptDesc>**

After the description of the object, the **<physDesc>** allows more specific descriptions for other aspects such as the hand, type, or script used. These can then be referred to from other parts of the manuscript description or transcription. For example, in the hypothetical example below, the **<handNote>** has an identification number, which can be pointed to from within a transcription to indicate where a hand changes. Similarly the note itself could point to more information about a scribe (with **@scribeRef**) or script (with **@scriptRef**). An arbitrary identifier could be provided for **@scribe** and **@script** attributes if it is not feasible to point to more information.

```

<handDesc>
  <handNote xml:id="handID" medium="inkType" scope="major" scribe="Scribel"
    scribeRef="#scribeID" script="scriptName" scriptRef="#script1">
    <p>One hand note for each identified hand in the manuscript with optional
      attributes for a hand id number, the medium of the hand, the scope of its
      writing stint, a scribe or reference to more information about one, a script
      or reference to one</p>
  </handNote>
</handDesc>
<typeDesc>
  <typeNote>
    <p>If describing a source with printed aspects, an optional typeNote for each
      typographic feature</p>
  </typeNote>
</typeDesc>
<scriptDesc>
  <scriptNote xml:id="script1">
    <p>Discussion of a particular script in the manuscript, its scriptorium, or
      usual use</p>
  </scriptNote>
</scriptDesc>

```

The `<typeNote>` element is used for either printed sources or manuscripts with printed aspects. It is important to note that the `<msDesc>`, although developed for manuscripts, can be used for any textual source whose description is not adequately covered by standard bibliographic metadata. Here, the identification number on the `<scriptNote>` is referenced from the earlier `<handNote>` to give a more detailed description of the script in which that hand is written.

```

<handDesc>
  <handNote>Written in dark brown ink in a gothic bookhand,
    in two sizes according to liturgical function;
    apparently by a single scribe throughout, except for the
    suffrage to Adrian (fols. 49r-50r) which is by a different but
    contemporary hand, perhaps in the same workshop.</handNote>
</handDesc>

```

Any of these more specific descriptive aspects can be more or less specific. In the example above two hands are recorded in a single `<handNote>` but where there are multiple hands these could be given as separate `<handNote>` elements.

```

<handDesc>
  <handNote xml:id="hand1">
    <locus from="1r" to="188r"/>
    Fols. 1r-188r written in gothic textualis, first half of the thirteenth
    century, with contemporary corrections pencilled in the margins (e.g. fols. 39v,
    44r-45r, 74r-77r). At least two scribes, one writing a smaller, more delicate
    script.</handNote>
  <handNote xml:id="hand2">
    <locus from="188v" to="192r"/>
    Fols. 188v-192r written in a charter hand, first half of the thirteenth
    century, at least two scribes.</handNote>
</handDesc>

```

3.3.3. Physical Description: <musicNotation>, <decoDesc>, <additions>

In addition to hands, physical type used, and scripts, there are also grouping elements for recording the presence and descriptions of musical notation, any forms of decoration, and any additions, marginalia, or annotations on the manuscript. In a hypothetical example, this might look like:

```

<musicNotation>
  <p>Description of the type of musical notation in the manuscript</p>
</musicNotation>
<decoDesc>
  <decoNote type="initials">
    <p>Description of the decoration of the manuscript or one homogenous class of
       decoration in it</p>
  </decoNote>
</decoDesc>
<additions>
  <p>Description of any additions, marginalia, or annotations on the manuscript</p>
</additions>

```

While <musicNotation> and <additions> have no specific structures in the content models -- that is they allow either text with phrase-level content or one or more paragraphs -- the <decoDesc> element contains one or more <decoNote> elements. These <decoNote> elements can each contain either text with phrase-level content or one or more paragraphs, allowing flexibility in the amount and level of description. Many collections of manuscript descriptions will prefer to use paragraphs where allowed, regardless of whether they need more than one, for consistency in processing the output. In other cases a <decoNote> element may be used for each separate form of decoration.

```

<decoDesc>
    <decoNote>Fine miniatures, borders, initials. Historiated penwork borders.
        (P&A ii. 73, pl. VI)
    </decoNote>
    <decoNote>Fine miniatures in colours and gold at the
        start of the five books (fols. 1r, 54r, 101r, 151v, 168r).</decoNote>
    <decoNote>Fine red and
        blue penwork decoration, often extending from two-line initials, includes
        birds, animals, dragons, human-headed hybrid creatures (see. e.g. fol. 113v),
        etc.</decoNote>
</decoDesc>

```

It is, of course, possible to also have `<locus>` elements inside `<decoNote>` to indicate where the decoration is situated. Similarly an `@xml:id` attribute can be provided, and referenced to from particular places in the transcribed text to localise the notes.

Physical Description: `<bindingDesc>`, `<sealDesc>`, `<accMat>`

The final sections of a physical description are those of external and adjunct aspects of the manuscript. In particular, information concerning the binding, seals, or other matter which accompanies the manuscript.

```

<bindingDesc>
    <binding notBefore="1700" contemporary="false">
        <p>Description of the binding, including boards and coverings, with optional
            dating attributes and indication of whether the binding is contemporary</p>
        <condition>
            <p>Optional description of the condition of the binding</p>
        </condition>
        <decoNote>
            <p>Optional description of the decoration on the binding</p>
        </decoNote>
    </binding>
</bindingDesc>
<sealDesc>
    <seal when="1518" contemporary="true">
        <p>Description one or more seals, with optional dating attributes and
            indication of whether the seal is contemporary</p>
    </seal>
</sealDesc>
<accMat>
    <p>Description of any significant additional material closely associated with the
        manuscript being described (e.g. papers not catalogued separately)</p>
</accMat>

```

Following the same pattern as many of the elements in `<physDesc>`, the `<bindingDesc>` and `<sealDesc>` elements may contain one or more `<binding>` or `<seal>` elements. Multiple `<binding>` elements usually are historical in nature, where details of earlier bindings are known because of existing records, or the manuscript has recently been rebound, or because of surviving artifacts of its presence. Multiple `<seal>` elements are used to indicate that multiple seals survive (or there is information about additional missing ones). The `<accMat>` element (for

accompanying material) is used, in this location, to describe the physicality of any accompanying objects that do not usually have their own descriptions.

```
<bindingDesc>
  <binding>
    <p>
      <dimensions type="binding">
        <height quantity="360">360</height>
        <width quantity="258">258</width>
        <depth min="54" max="83"/>
      </dimensions>
      Leather over boards, 18th or 19th century.
    </p>
    <condition>Binding is tight, putting unnecessary pressure on leaves.</condition>
  </binding>
</bindingDesc>
```

In this example the `<binding>` element has a paragraph inside it because it is mixing element content such as `<dimensions>` with prose text. The `<condition>` element is provided separately inside `<binding>` to record the condition of the binding and its effect on the manuscript as an object.

The `<seal>` element usually contains one or more paragraphs inside it and records information about the seals, their contents, and condition. It could also contain a description of other aspects of the document which might be used to validate its authenticity. A `@contemporary` attribute can be used (as with bindings) to record whether the seal is contemporary or not.

```
<sealDesc>
  <seal>
    <p>Remains of four seals in red wax, plaqué</p>
  </seal>
</sealDesc>
```

The physical description is a key aspect of any manuscript description. The layout, foliation, collation, hands, decoration, bindings, seals, and other physical aspects can provide more contextual information for an edition. This is not only useful for understanding the context of production of the physical object, but also how it may interact with the creation or copying of the text that is at the heart of the edition.

3.4. History: Origin, Provenance, and Acquisition

The history of the creation of a manuscript, and its provenance, are some of the important aspects of a manuscript description. The history of how this text-bearing object has come to survive to the modern day is of great interest for those using it in an edition, as that history may have affected the physical and textual make up of the document. This history is stored in a `<history>` element, all of whose children are optional. These include providing a summary in a `<summary>` element, information about the creation of the manuscript in an `<origin>` element, as many `<provenance>` elements as needed to record any episodes in its history, and an

<acquisition> with details of its current acquisition. As with other sections, these may contain phrase-level content or a series of paragraphs.

```
<history>
  <summary>
    <!-- Summary entry of history of the manuscript -->
  </summary>
  <origin>
    <!-- Information concerning the origin and creation of the manuscript -->
  </origin>
  <provenance>
    <!-- Information concerning any single identifiable episode during the history
        of the manuscript after its creation but before its acquisition
        -->
  </provenance>
  <acquisition>
    <!-- Information concerning the acquisition of the manuscript
        by the current resource-holding institution
        -->
  </acquisition>
</history>
```

Other than <summary> each of these is ‘datable’ -- by which we mean that they are members of the TEI att.datable class, and thus get a whole slew of dating attributes like @when, @notBefore, @notAfter, @from, and @to. This enables any of these steps in its history to be given a date on this container element. In manuscript description phrase-level content there are two elements which while available almost anywhere are best used inside <origin>, namely <origDate> (for providing an origin date) and <origPlace> (for providing an origin placename).

```
<history>
  <origin>
    Written in the
    <origDate calendar="Gregorian" notAfter="1500" notBefore="1400">15th century</origDate>
    in <origPlace><country>England</country></origPlace>
  </origin>
  <provenance>
    Previously owned by
    <persName type="person" role="formerOwner" ref="http://viaf.org/viaf/73979081">
      Sir Thomas Phillipps (1792-1872)</persName>, MS. 13443* and 13446
  </provenance>
  <provenance>
    Previously owned by
    <persName type="person" role="formerOwner" ref="http://viaf.org/viaf/11255508">
      Sir Robert Leicester Harmsworth</persName>
  </provenance>
  <acquisition>
    Harmsworth Trust sale at Sotheby's, <date when="1945-10-16">16 Oct. 1945</date>,
    lot 2051, bought by Quaritch for £38.
  </acquisition>
</history>
```

This example has the information that this manuscript was written in the fifteenth century ‘in

England'. (Although in some manuscript description traditions this might be expressed as 'English', in this case it is an indication of its nationality not its text language.) There are two episodes in its provenance that are known, that it had previous owners that are identified and referenced to VIAF. Finally, the purchase of this manuscript by the Bodleian Library in 1945 from an auction is recorded. While this information is minimal, often the confirmed provenance of a manuscript is very limited. It is very useful for any description, especially when attached to a surrogate such as a digital edition, to record and preserve any knowledge about the manuscript's history.

3.5. Recording additional metadata

The above sections clearly identify the manuscript, its intellectual contents, physical form, and history. However, there are other forms of metadata that are important to record, such as information about the sources of the manuscript description, events in its custodial history (such as photography and conservation), digital or print surrogates for the manuscript, and secondary works concerning the manuscript. These may be provided in an **<additional>** element which contains (optional) **<adminInfo>**, **<surrogates>**, and **<listBibl>** elements.

```

<additional>
  <adminInfo>
    <recordHist>
      <source>
        <!-- Information about the source of the manuscript description -->
      </source>
      <change>
        <!-- Descriptions of changes to the manuscript description -->
      </change>
    </recordHist>
    <custodialHist>
      <custEvent>
        <p> <!-- Custodial event history here --> </p>
      </custEvent>
    </custodialHist>
  </adminInfo>
  <surrogates>
    <p> <!-- Description of surrogates --></p>
    <listBibl>
      <bibl>
        <!-- Bibliographic entries of surrogates -->
      </bibl>
    </listBibl>
  </surrogates>
  <listBibl>
    <bibl>
      <!-- Bibliographic entries to works concerning the manuscript -->
    </bibl>
  </listBibl>
</additional>

```

Inside `<adminInfo>` it is possible to provide `<recordHist>`, `<availability>`, and `<custodialHist>`, elements. These help to provide information about the source and history of the manuscript description itself, the license under which that manuscript description is made available, and events in the custodial history of the manuscript.

```

<additional>
  <adminInfo>
    <recordHist>
      <source>Summary description by Elizabeth Solopova based on the Summary
      and Quarto Catalogues and supplementary sources.
      <listBibl>
        <bibl type="QUARTO" facs="abw0019.gif">Quarto Catalogue, cols. 19-20</bibl>
        <bibl type="QUARTO" facs="abw0192.gif">Quarto Catalogue (addenda), p. 15</bibl>
        <bibl type="QUARTO" facs="abw0193.gif">Quarto Catalogue (addenda), p. 16</bibl>
        <bibl type="SC" facs="aap0091.gif">Summary Catalogue, vol. 2, part 1, p. 70</bibl>
      </listBibl>
    </source>
  </recordHist>
</adminInfo>
<surrogates>
  <bibl facs="http://image.ox.ac.uk/list?collection=bodleian">
    Timaeus and Chanson de Roland: full digital facsimiles
  </bibl>
</surrogates>
</additional>

```

In the example above, the **<recordHist>** element provides information about the source of the description, listing previous catalogues that were used in composing it. Meanwhile the **<surrogates>** element gives a bibliographic reference to a digital facsimile of the manuscript.

3.6. Composite manuscripts and fragments

So far, the manuscript descriptions discussed have been for unitary manuscripts, which may contain multiple items but were produced originally as a single object. In many cases manuscripts are composite, i.e. now treated as a single physical object, but were originally distinct objects (or a part of a previous manuscript) before being grouped together. This is different from manuscript fragments, which are separate pieces of a single manuscript.

```

<msDesc xml:id="MyCompositeMS">
    <msIdentifier>
        <!-- MS Identification for composite object -->
        <msName>My Composite Manuscript</msName>
    </msIdentifier>

    <!-- MS Part elements for each manuscript in the composite -->
    <msPart xml:id="msPartA">
        <msIdentifier>
            <!-- Identification for this manuscript part -->
        </msIdentifier>
        <msContents>
            <!-- Intellectual contents of this part -->
        </msContents>
        <physDesc>
            <!-- Physical description of this part -->
        </physDesc>
        <history>
            <!-- History of this part -->
        </history>
        <additional>
            <!-- Additional metadata for this part -->
        </additional>
    </msPart>

    <!-- Additional parts -->
    <msPart xml:id="msPartB">
        <msIdentifier></msIdentifier>
    </msPart>

    <!-- Additional parts -->
    <msPart xml:id="msPartC">
        <msIdentifier></msIdentifier>
    </msPart>
</msDesc>

```

Inside each `<msPart>` you have the same elements as are available inside `<msDesc>` and thinking of them as nested manuscript descriptions is a useful approach. You may have the usual elements such as `<physDesc>` inside the parent `<msDesc>` which cover aspects that affect the composite object, and also a `<physDesc>` inside the `<msPart>` that describes the physicality of the part.

```

<msPart xml:id="bod-MS._Digby_133-1">
  <msIdentifier>
    <idno type="part">MS. Digby 133 - Part 1</idno>
  </msIdentifier>
  <msContents>
    <msItem n="1" xml:id="a1_1">
      <author ref="http://viaf.org/viaf/2470550"> Galileo Galilei </author>
      <title>Discorso del flusso e reflusso del mare</title>
      <textLang mainLang="ita">Italian</textLang>
    </msItem>
  </msContents>
  <physDesc>
    <objectDesc form="codex">
      <supportDesc material="chart">
        <support>paper</support>
      </supportDesc>
    </objectDesc>
  </physDesc>
  <history>
    <origin>
      <date calendar="Gregorian" notBefore="1600" notAfter="1700">17th century</date>
      <origPlace>Italian</origPlace>
    </origin>
  </history>
</msPart>

```

In this **<msPart>**, a brief description of its contents, physical form and history are provided, but these could also contain full and detailed descriptions, depending on what information is available.

In the event that a single manuscript has been fragmented into separate parts, and it is necessary to describe each of these fragments as part of a larger manuscript description, **<msFrag>** can be used. The main difference between **<msPart>** and **<msFrag>** is that the former is used for part of an existing object that was previously distinct, and the latter is used for a fragment of an original that is not now bound or attached. A manuscript description containing **<msFrag>** is not describing a single object, but a putative reconstructed or original object from which these fragments originate. On the other hand, a manuscript description containing an **<msPart>** is a single existing object that happens to be formed of parts which were originally distinct.

```

<msFrag xml:id="MS_Example-frag1">
  <msIdentifier>
    <!-- Manuscript identification of this fragment -->
  </msIdentifier>
  <msContents>
    <!-- Manuscript contents of this fragment -->
  </msContents>
  <physDesc>
    <!-- Physical description of this fragment -->
  </physDesc>
  <history>
    <!-- History of this fragment -->
  </history>
  <additional>
    <!-- Additional metadata of this fragment -->
  </additional>
</msFrag>

```

An **<msFrag>** is structurally identical to an **<msPart>** -- the only real difference between them is the semantics of their definition. An **<msDesc>** element containing either of these can be as detailed as the available information allows. The same elements that appear inside **<msDesc>** are also available inside **<msPart>** and **<msFrag>** because information relating to the part or fragment might concern its identification, contents, physical description, history, or other additional aspects. Usually, any information that applies to the manuscript as a whole is stored in the main sections, and only those aspects that apply to individual parts are recorded inside **<msPart>** or **<msFrag>**.

4. Publishing manuscript descriptions

A manuscript description created as part of a digital edition should form part of the published digital edition itself. For example, the information stored in the organised sections of the manuscript description can be extracted to form a description given in the introduction to the edition. Some of the publication tools for manuscript editions do make use of descriptions if they are present, but for the most part even these take a generalistic approach. More complicated and nuanced displays of manuscript descriptions are found where institutions present whole catalogues describing their holdings. In cases such as these you can usually browse and/or search the entire collection, and often filter by various facets. The software behind such systems range from native XML databases to bespoke systems developed for individual institutions. For example, in the recent redevelopment of all TEI manuscript description catalogues at the Bodleian Library, University of Oxford, the TEI P5 XML was converted to HTML to be ingested by Blacklight and indexed by Solr. In the Bodleian's case this was driven from TEI

manuscript descriptions stored in a publicly accessible GitHub account with the same underlying software used across several different catalogues, resulting in improved overall maintenance and support for the system.² However, systems such as these may be unnecessary when presenting a manuscript description from a digital edition, and the decision as to what software should be used to display the description is more dependent on how the edition itself is being published. Two basic rules should always be followed:

- Always offer the resource-holding institution a copy of your final manuscript description. They may not have need or use for it, or be willing to convert it to whatever system they are using, but since you've put work into increasing the detail of the description it is fair to offer it to them. However, if you've put significant intellectual effort into creating it, then do offer it on condition that you are cited as the creator of it (indeed you could store this in the `<recordHist>` element's `<source>` child element while listing sources). If you license the description to them as Creative Commons Attribution, using the `<availability>` element in the header, then they are duty bound to acknowledge you in their use of it, but this does not hamper them in using it in their systems otherwise.
- Always make the TEI XML of your description available. This is not in question if you are making the whole of your edition's XML available, but if the manuscript description is presented separately it is sometimes overlooked in the version of the XML that is released. Not only is this good practice, as it shows your underlying methodologies, but it enables others to make use of this data in ways you might not have anticipated. (e.g. the programmatic study of the markup itself or linguistic analysis of manuscript descriptions).

Conclusion

This chapter seeks to give an introduction to creating manuscript descriptions through the lens of encoding them according to the recommendations of the Guidelines of the Text Encoding Initiative, but the information given herein should still be useful in highlighting the categories of information and detail that is expected in a full manuscript description, whatever system is used. Thinking clearly and precisely about the physical object, its history, and describing the intellectual contents of that object will provide an intimate familiarity with a manuscript that can only benefit the creation of a sensitive and useful digital edition of its text.

² The TEI XML medieval manuscript descriptions from Bodleian Libraries are available from <https://github.com/bodleian/medieval-mss>.

Textual Variants

Marjorie Burghart

In this chapter you will learn:

- how to understand a traditional printed edition and its conventions;
- how to encode textual variants according to different methods and styles;
- how to solve several special cases common in critical editions.

When a text is transmitted through more than one witness, a critical edition will generally take a strong interest in recording the variant readings of some or all of those manuscripts or editions. There are many schools of textual criticism, more or less prevalent depending on the geographic area, field (literary studies, history, ...), and period of the original text. The history of (and relationships between) currents of textual criticism goes far beyond the scope of this chapter, and deserves a book of its own - if not a full library. The readers wishing to familiarise themselves with the various approaches could turn to Greetham 2013 for a short summary. But whatever the school or current of textual criticism, the TEI offers the same mechanism to record textual variance, described in the [Critical Apparatus chapter of the Guidelines](#).¹

For some readers, the concept of critical edition may be new. We will therefore open this chapter with a brief introduction to textual variants and the traditional way to record them, before moving to a survey of the TEI methods available for digital editions. We will give an overview of the different possible styles of apparatus, within the frame of the TEI. Finally, we will see some examples of particular cases commonly occurring in critical editions.

1. Understanding textual variants in a critical edition

The presence of variants and errors in the various witnesses of a text (manuscripts and editions) is particularly important in manuscript cultures, especially for classical and medieval texts. As texts were copied from manuscript to manuscript, the scribes introduced in their copies readings that differed from the model (or models) from which they were copying. A common distinction is to call “errors” the differences that are mere mistakes (typo, forgotten or repeated word, etc.), and “variants” the ones willingly introduced by a scribe, showing some creativity. However, it is sometimes difficult to distinguish between the two, and besides the concept of “error” implies the existence of a correct, reference text, which can be seen as a view too partial to particular currents of textual scholarship. We will therefore use only the more neutral term of “variant” or

1. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/TC.html>

“reading” here. For editions of texts published in the print era, the proportion of variants is naturally much less significant, but nevertheless variant readings exist between editions and must be taken into account.

Critical editors need to render account of their work on the tradition and the choices they made, in a way that would allow their readers to verify their work. In a traditional printed edition, scholars have to conciliate, in the two-dimensional space of the page, the legibility of the text they propose with the informativity of the annotations representing their exploration of the tradition, and their choices (if any) among the variant readings. More than a century of modern scholarship has shaped the traditional critical edition page: it is composed of a main text, which is the version of a text proposed by the editor, to which refer one or more groups or layers of footnotes (one for textual variants, one for source identifications, one for the reference of biblical citations, one for historical notes, etc.). To avoid the multiplication of footnote markers in the text, which impairs its legibility, it is common in printed editions to link the notes to the text by the means of line numbers: the lines of the text are numbered, and in the footnotes a reference is made to a line number, or the numbers of a range of lines.

The [following illustration](#) is an example of such traditional critical edition layout. It has three distinct layers of footnotes. The first two (textual variants, and biblical quotations) refer to line numbers in the main text, while the last (source identifications) has footnote markers in the text. The editor has proposed a text from the evidence found in four manuscripts, A, S, T and H. Using the textual variants footnotes, the reader must be able to deduct the readings of each witness. Each footnote about a textual variant has the same structure: following the line number is the “lemma,” that is the word or group of words from the main text for which there is a variant reading. The lemma is followed by a description of what changes (is this word omitted? is it replaced with another word? etc.), and in which manuscript(s). For instance, if we look at the footnotes referring to line 44 in the text, we can reconstruct the text of each manuscript:

- A has the same sentence as the one presented in the main text: “scilicet malis que obuenerunt ei,” which translates as “that is to say the misfortunes that hit him”;
- S has one omission, “scilicet” is missing: “malis que obuenerunt ei”;
- T has two omissions, “scilicet” is missing at the beginning, and “ei” at the end: “malis que obuenerunt”;
- H has one variant reading “euenerunt” instead of “obuenerunt”: “scilicet malis que euenerunt ei”

You will notice that many abbreviations are used in those footnotes, in order to save space: *om.* for an omission, *add.* for an addition to the text of the edition, etc.

omnibus hiis, scilicet malis que obuenerunt ei, non peccauit
 45 *Iob labiis suis, nec contra Deum, nec contra proximum, nec*
contra uxorem. Item sufferunt omnia patienter, confortati
exemplo eiusdem Iob iuxta illud Iacobi ultimo: Sufferentiam
Iob audistis. Et bene debet tale exemplum animare. Si enim
 50 *Iob, qui fuit quasi rex et diues ualde et quasi innocens, tam*
patienter portauit amissionem tantorum bonorum et propri-
orum filiorum, et tam grauem et tam uilem infirmitatem a plan-
ta pedis usque ad uerticem, quanto magis alii in quibus desunt
conditions predice debent esse patientes in minoribus!

Preterea notandum quod amor presentis uite occasio est
 55 omnium malorum sicut ostendit beatus Sebastianus in sermo-
 ne quem fecit Marco et Marcelliano, ubi, enumeratis multis
 et grauibus malis, concludit dicens: « Numquid non presentis
 uite amore ducti ista committunt? »¹. Sed huiusmodi infirmi-
 tias acerba et desperata et abhominabilis amaricat omnes io-
 60 cunditates huius uite, et ideo retrahere debet ab amore huius
 uite, sicut amaritudo circumposita uberibus retrahit puerum
 ab amore uberum. Item amaritudo huius uite debet nos mouere
 ad desiderandum aliam uitam beatam que est sine amaritudi-
 ne. Augustinus: « Cum molestie in huius uite fragilitate cre-
 65 bescunt, eternam requiem nos desiderare compellunt »². Et
 ideo in huiusmodi infirmitatis constituti amaritudine debent
 ad illam beatam dulcedinem uehementer suspirare. Et sic fe-
 cit quidam religiosus leprosus qui sanctissime se habens in

⁴⁴ scilicet *om.* S T obuenerunt / euenerunt H ei *om.* T ⁴⁶ sufferunt / su-
 ferant S T ⁴⁷ ultimo / 5 S T sufferentiam / sufficientiam S T ⁴⁹ quasi *om.*
 H S T tam / et *praem.* S T ⁵¹ filiorum *om.* A ⁵² desunt *om.* S
 T ⁵⁴ preterea / postea H ⁵⁶ enumeratis / et innumeratis S ⁵⁷ et graui-
 bus *om.* H numquid / numquam S T non *om.* S T ⁵⁸ committunt / alias di-
 mittunt *add.* H ⁶⁰ et *om.* A ⁶⁰⁻⁶¹ et ideo... uite *homom.* A ⁶¹ uberibus
 / ubique S T ⁶³ desiderandum/considerandum A ⁶⁴ in *om.* A fragilitate /
 flagelitate S T, fragilitates H ⁶⁵ requiem *om.* S T ⁶⁷ uehementer *om.*
 A ⁶⁸ se habens / sequens S T

⁴⁷ Iac 5, 11 ⁵² Iob 2, 7

¹ Cf. *Passio sancti Sebastiani*, PL 17, 1025D-1026A.

² AUGUSTIN, *Ep.* 145, 2, PL 33, 983.

Figure 1: An example of traditional critical edition in print (from Nicole Bériou et François-Olivier Touati, *Voluntate Dei leprosus: les lépreux entre conversion et exclusion aux XIIe et XIIIe siècles*, Spoleto: Centro Italiano di Studi sull'Alto Medioevo, 1991, 162)

Now that we have deciphered a print edition, we can work in the opposite direction, and see how we would build such an edition from manuscript evidence. Let us imagine a simple case, where our text is transmitted through three surviving manuscripts that we will call A, B and C. These witnesses of the text form its “tradition,” that is all manuscripts and editions which have born the text. In these three manuscripts, the text is almost the same, but in manuscript B, we have emphasised a word that differs from the readings of the two other manuscripts:

A

... ut dum spiritus reficitur
dolores corporis leniantur...

B

... ut dum spiritus reficitur
dolores corporis *laniantur*...

C

... ut dum spiritus reficitur
dolores corporis leniantur...

The Latin sentence born by manuscripts A and C, “... ut dum spiritus reficitur dolores corporis leniantur...,” means “so that, while the spirit is restored, the sufferings of the body are soothed.” It is grammatically correct and makes perfect sense. The sentence born by manuscript B, “... ut dum spiritus reficitur dolores corporis *laniantur*...,” has only one word that changes (actually, only one letter), but its meaning is totally different: it translates as “so that, while the spirit is restored, the sufferings of the body are *torn*.” This is not a correct and meaningful sentence, and a critical editors will deduce that the scribe copying manuscript B made a mistake while copying “leniantur,” or copied from a manuscript (now lost) already bearing “*laniantur*” instead of “leniantur.” If we were to produce a traditional printed edition, here is how we would build our critical apparatus note:

- **Text:**

ut dum spiritus reficitur dolores corporis leniantur^(a)

- **Footnote:**

^{a)} leniantur] *laniantur B*

quoting the lemma | rejected reading | witness

Figure 2: Structure of the information contained in a textual variant footnote

Digital editions, unlike printed ones, are not limited to “the two-dimensional space of the ‘page’ and to typographic means of information representation” (Sahle 2008). Yet, it is very important for digital Humanists to understand the traditional semiology of critical editions and the decades of scholarship that produced it. Conversely, it is capital for scholars familiar with this traditional semiology to be able to think beyond it, in order to create truly digital scholarly editions which are free from the constraints of the printed page and convey to the reader the same information, and even more, through different means. This is what we will learn in the following section, explaining how textual variants are represented in a TEI XML edition.

2. Encoding textual variants in TEI

To record textual variants, the TEI uses the **<app>** (apparatus entry) element, which “contains one entry in a critical apparatus, with an optional lemma and usually one or more readings or notes on the relevant passage.”² An apparatus entry can be critical, i.e., express a choice made by the editor to favour one reading over the others. In this case, it will contain two types of elements:

- a single **<lem>** element (lemma), which records the reading chosen by the editor. In a printed edition, this is the text we would find in the main text of the edition;
- one or more **<rdg>** (reading) elements, which record the rejected readings. In a printed edition, this is the information we would find in a critical footnote, saying that instead of the lemma which manuscript has which variant instead of the lemma.

Or the apparatus entry can be neutral and simply record how the various witnesses vary one from another at a given point of the text, without expressly favouring a reading. In this case it will only contain **<rdg>** elements.

The critical or neutral style of apparatus entries is chosen according to the school of textual criticism in which belongs the editor. But even editors who want critical apparatus entries in their final edition might want to use neutral entries, simply recording variants, while they are still collating manuscripts and have not yet decided which reading will be the lemma. It is perfectly possible to start working on an edition with neutral entries, then choose a lemma for each entry in the later stages of the work.

For **<lem>** as well as for **<rdg>**, the **@wit** (witness) attribute is used to materialise the link between the given text and the manuscript(s) bearing it: **@wit** may contain one or more reference(s) to the unique **@xml:id** of witnesses described in a list of witnesses (**<listWit>**), typically placed in the **<sourceDesc>** section of the header or in the **<front>** part of the **<text>**. This list contains several **<witness>** elements identified by their siglum (**@xml:id**). The first step must therefore be the creation of this list of witnesses, to be able to use the references to the witness sigla in the apparatus entries. A minimalist version of this list could be encoded like this:

```
<listWit>
  <witness xml:id="A">Short description of witness A (city, library,
shelfmark)</witness>
  <witness xml:id="B">Short description of witness B (city, library,
shelfmark)</witness>
  <witness xml:id="C">Short description of witness C (city, library,
shelfmark)</witness>
</listWit>
```

2. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-app.html>

To learn more about the encoding of witness descriptions, which you can make as complete and precise as you wish, please read the chapter Manuscript Description.

The **<lem>** and **<rdg>** elements wrapped in an **<app>** are the basic tools to represent textual variants, but the TEI offers three different methods to combine them, each with its pros and cons. Even though we generally recommend to use the Parallel segmentation method, the following review of the three methods might help you make well-informed choices. This is of course just an overview, and we refer you to the [Guidelines](#) for a full description of the methods and their options. Let us examine them individually, and see how our latest example, where we have one variant among three manuscripts, would be encoded with each of these methods.

A	B	C
... ut dum spiritus reficitur dolores corporis leniantur...	... ut dum spiritus reficitur dolores corporis <i>leniantur</i> ut dum spiritus reficitur dolores corporis leniantur...

2.1. Location-referenced method

In the Location-referenced method, we need to have a base text to attach apparatus entries to it. It means that it is not suitable if we want to make an edition without using the concept of base text. The apparatus entries are not linked to a precise spot in the text, but to a block of text: we indicate to our readers that there is an apparatus entry linked to a particular paragraph, or line of verse, etc. Human readers will be able to understand the meaning of the entry and its relation to the base text, but a script cannot process the information automatically beyond simply displaying an apparatus note linked to the relevant block of text. It means that this method is not suitable either if you wish to be able to virtually reconstruct the individual text of each witness, for instance.

One of its advantages though is that, since the apparatus entries are linked to the text by a simple system of references, they will not interfere with other hierarchies, i.e., the **<app>** elements will not overlap with other elements, like the ones used to encode citations, text structures, etc.³ The apparatus entries may be “internal” (stored in the same file as the edition) or “external” (stored in a different file).

This method can be useful when converting printed edition, if you simply want to display notes loosely attached to the text, without the need to process the result in elaborate fashions. Its compatibility with overlapping hierarchies is also an argument in its favour, but the other methods also have solutions. However, we would not generally recommend using the Location-referenced method, since it is just as long to encode as with other methods, but offer less possibilities for processing.

Here is how our example would be encoded with the Location-referenced method. In the base text we would have a block of text identifiable by some sort of reference (here, paragraph “1” in the division “serm1”):

3. For a discussion of the methods available, see section Handling overlap in chapter Citations and references.

```

<div n="sermol">
  <p n="1">...ut dum spiritus reficitur dolores corporis leniantur...</p>
</div>

```

Somewhere else, either in the same file as the edition (in the **<back>** for instance) or in another related file, we would have the apparatus entries. The **@loc** attribute of **<app>** is used to represent the reference to the relevant block of text in the base text:

```

<app loc="sermol 1">
  <lem>leniantur</lem>
  <rdg wit="#B">laniantur</rdg>
</app>

```

2.2. Double-end-point-attached method

The Double-end-point-attached method, just like the previous one, implies the existence of a base text - with the same consequences. But this time, the apparatus entries are linked to a precise span of the base text, using the empty tags **<anchor>** if no other element is already present: one to mark up the beginning of the lemma, and another to mark up the end. This means that, beyond generating notes at the relevant places, you will be able to process your edition in more elaborate ways than with the previous method. Since it uses empty tags, marking up just points in the text, this method is also convenient if you have to deal with overlapping hierarchies.

This method is intellectually satisfying: it allows us to link your apparatus precisely to the base text, leaving all the possible processing options, while allowing potentially overlapping hierarchies to coexist without any problem. Its major drawback is practical: first, you will have to place the elements or anchors marking the limits of the lemmata in the base text, giving each an **@xml:id**, and then you will need to be very careful when linking the **<app>** to these identifiers, since a mistake will mean that your apparatus entry will be linked to a wrong span of the base text. In the absence of a specialised XML editor that would take care of these delicate steps, it is a tedious and error-prone task. Second, and perhaps more importantly, at the time of writing there are no available tools offering support to display or process critical editions encoded following the Double-end-point-attached method. This means that you will have to develop your own scripts and applications to be able to display and/or process an edition encoded with this method.

Our example could be encoded as follows with the Double-end-point-attached method. Anchors would mark the beginning and end of the lemma in the base text:

```

<div n="1">
  <p n="1">... ut dum spiritus reficitur dolores corporis <anchor
  xml:id="app1begin"/>leniantur<anchor xml:id="app1end"/> ... </p>
</div>

```

Somewhere else, either in the same file as the edition (in the **<back>** for instance) we would have the apparatus entries. The **@from** and **@to** attributes of **<app>** contain a reference to the **@xml:id** of the tags marking up the beginning and end of the lemma in the base text:

```

<app from="#app1begin" to="#app1end">
  <rdg wit="#B">laniantur</rdg>
</app>

```

2.3. Parallel segmentation method

The parallel segmentation method does not necessarily imply a base text: we may have one, if we use `<lem>` in our apparatus entries, or we may not, if we only use `<rdg>`. With this method, the apparatus entries are placed “inline,” instead of being placed somewhere else in the edition file or in a related XML file. It means that the apparatus does not need to be linked to the text through references, since it occurs in the flow of the edition’s text. Each time witnesses have different readings, an `<app>` element is inserted, containing either a lemma and one or more readings, or a simple list of variant readings. These `<app>` are of course at risk of overlapping with other elements, but should it occur there are mechanisms allowing you to handle them.⁴

Relatively easy to encode manually, giving the option of critical or neutral apparatus entries, and supported by several third-party tools,⁵ the Parallel segmentation method is the most widely used. It is the method we recommend to use, unless you have a particular reason to prefer another. All the further examples of apparatus entries will be encoded using this method.

We might encode the example as follows - we could also add a `@wit` attribute to `<lem>`, but it is optional as, by default, all the witnesses not mentioned in a `<rdg>` are supposed to bear the text of the lemma:

```

<div n="1">
  <p n="1">... ut dum spiritus reficitur dolores corporis <app>
    <lem>laniantur</lem>
    <rdg wit="#B">laniantur</rdg>
  </app> ... </p>
</div>

```

3. Different styles of apparatus

Besides the encoding methods, editors must opt for a style of apparatus. The choice of a style of apparatus has nothing to do with technical considerations: it is a methodological issue, linked to the type of edition you intend to produce. Some schools of textual criticism will recommend the neutral recording of the variants, other will consider a lemma absolutely necessary. However, even for the latter type of editions, we recommend a mixed approach: a neutral recording of variants during the collation stage, then a conversion to critical apparatus entries by choosing a lemma.

4. See section *Handling overlap* in chapter Citations and References. The mechanisms are explained with the example of quotations, but they may be applied to any other element.

5. At the time of writing, the Parallel segmentation method is the only one supported by tools dedicated to TEI critical editions, like [Edition Visualization Technology](#), [Stemmaweb](#), the [TEI Critical Apparatus Toolbox](#) or the [Versioning Machine](#)

3.1. Neutral recording of the variants

When opting for a neutral recording of the variants, the editor lists the varying readings of the witnesses, without choosing a lemma. Editors who do not want to privilege the readings of a witness over the others may choose this style of apparatus.

It is worth mentioning that even for editors wishing to produce an edition using lemmata, this style of apparatus may be a valuable first step: when the editor has not yet collated all the witnesses, it may be difficult to decide which reading is the lemma. In this case, a common strategy is to encode first the apparatus in a neutral style, then, once the collation stage of the edition is over, make an editorial decision for each `<app>` and choose which `<rdg>` will become the `<lem>`.

With this style of apparatus, the same example would be encoded as follows:

```
ut dum spiritus reficitur dolores corporis <app>
  <rdg wit="#A #C">leniantur</rdg>
  <rdg wit="#B">laniantur</rdg>
</app> ...
```

3.2. Negative critical apparatus

When editors wish to produce an edition using lemmata, critical apparatus entries are required, but they may belong in two different styles: negative, or positive. In a negative critical apparatus entry, the witnesses are listed in `@wit` only for the variant readings (`<rdg>`). By default, all the witnesses which are not listed in a `@wit` attribute of a `<rdg>` of the current `<app>` are supposed to bear the text of the `<lem>`.

Practically, a negative critical apparatus is quicker to encode, especially when there is a long list of witnesses: we do not need to add a `@wit` attribute, and insert references to the witnesses in its value. On the other hand, processing it may be more complicated, depending on your goal, since the scripts will need to reconstruct for themselves the list of witnesses bearing the lemma. Another argument against a negative critical apparatus is that it is more difficult to verify: during the collation, it is only human to make mistakes - forgetting to add the reference to a new witness bearing the same variant as an existing `<rdg>` for instance, or making a typo which will result in the wrong manuscript being mentioned twice in the `@wit` value or a reading, etc. If this is a concern, we recommend choosing the positive style. It is worth noting that, in terms of display, it is perfectly possible to display or print in a negative style an apparatus that has been encoded as positive.

Example with a negative critical apparatus:

```
<app>
  <lem>leniantur</lem>
  <rdg wit="#B">laniantur</rdg>
</app>
```

3.3. Positive critical apparatus

In a positive critical apparatus entry, the witnesses are listed for the lemma as well as for the other readings. This is the method we generally recommend for editions using the concept of lemmata. This style is slightly longer to encode than a negative apparatus, but has some advantages: by forcing the encoder to explicitly state which witnesses bear the lemma and which have variants, a positive apparatus allows automatic verifications of the consistency of your encoding. Scripts will be able, for instance, to check whether each apparatus entry mentions a reference to all the witnesses - if one does not, it means that a mistake has been made and we must verify this apparatus entry. Another advantage is that scripts will generally have less difficulties processing your edition, since less operations are required to reconstruct the text of individual witnesses. Finally, as we noted above, encoding a positive apparatus does not imply you will have to display a positive apparatus: those are two different things, and the display scripts may perfectly be configured to display one style or the other, since in both cases the same information is recorded - implicitly in the first case, explicitly in the latter.

Example with a positive critical apparatus:

```
<app>
  <lem wit="#A #C">leniantur</lem>
  <rdg wit="#B">laniantur</rdg>
</app>
```

4. Going further with the Parallel segmentation method

We have seen the basic principle of the Parallel segmentation method, in various styles, and how they apply to one type of variant, the “substitutions” - when a word or phrase is replaced with another word or phrase. In this section we will first examine the other three main types of variants: “omissions,” “additions,” and “transposition” and suggest efficient ways to encode them.⁶ Then we will turn to some examples of the variety of particular that may arise.

4.1. Omissions, additions and transpositions

4.1.1. Omissions

There is an omission when a witness has nothing where the exemplar from which it was copied has a lemma. In this case, you can simply leave the **<rdg>** element empty where the witnesses have nothing. For instance, if we consider the following example, manuscript B has an omission: the word “corporis,” which is present in the other manuscripts and makes perfect sense in the sentence, is absent.

A

... ut dum spiritus reficitur
dolores corporis leniantur...

B

... ut dum spiritus reficitur
dolores _____ leniantur...

C

... ut dum spiritus reficitur
dolores corporis leniantur...

6. The use cases presented here are derived for the most part from Burghart 2011.

- *Text:*
ut dum spiritus reficitur dolores^(a) corporis leniantur
 - *Footnote:*
a) dolores om. *B*
-
- The diagram shows a superscript 'a)' followed by 'dolores om. *B*'. Two lines descend from this to the word 'lemma' in purple and 'editorial explanation' in green. A third line descends from 'om.' to the letter '*B*' in orange. Below these three items is a horizontal bar with vertical dividers, containing the words 'lemma | editorial explanation | witness'.

Figure 3: Representing this omission in a printed apparatus

We may encode this variant in the following way:

```
<app> ... ut dum spiritus reficitur dolores <lem wit="#A
#C">corporis</lem>
    <rdg wit="#B"/>
</app>
```

4.1.2. Additions

Conversely, there is an addition when one or more of the witnesses has some text where the exemplar from which it was copied has nothing. Additions may stem from innovation (a conscious intervention of the scribe) or mere errors (copying the same sentence twice, for instance), and if we wish to make this distinction we may use the @type attribute on **<rdg>** to categorise them.

In case of an addition, it is the **<lem>** element that we will leave empty, since there is no text at this place in the text of our edition. For instance, if we consider the following example, manuscript B has an addition: the word “corporis” has been written twice by the scribe - a type of scribal error known as “dittography.”

A	B	C
... ut dum spiritus reficitur dolores corporis leniantur...	... ut dum spiritus reficitur dolores corporis <i>corporis</i> leniantur...	... ut dum spiritus reficitur dolores corporis leniantur...

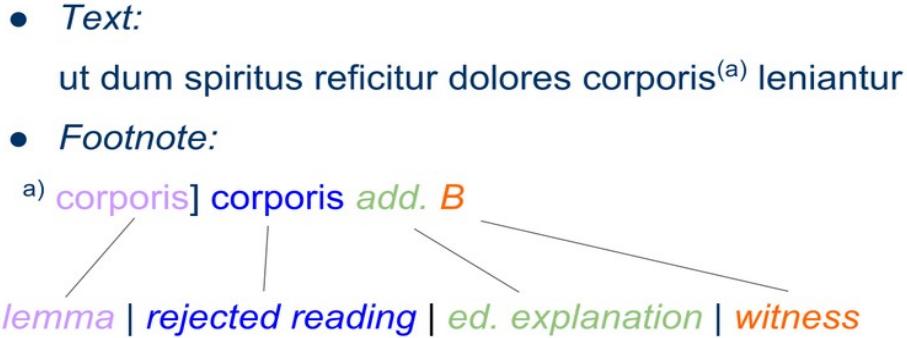


Figure 4: Representing this addition in a printed apparatus

We may encode this variant in the following way; here, since it is a particular type of addition, we have chosen to use the optional `@cause` attribute to indicate the type of phenomenon occurring:

```
... ut dum spiritus reficitur dolores corporis <app>
  <lem wit="#A #C"/>
  <rdg wit="#B" cause="dittography">corporis</rdg>
</app> leniantur...
```

Additions raise a particular issue when it comes to displaying your edition. Reconstructing the individual text of each witness is not a problem, but if we wish to display a main text with notes, in the fashion of traditional printed apparatus, we will need to display, as lemma, the word preceding the addition in the main text, so our readers may know where the addition occurs. Since the `<lem>` is empty, this word is the one immediately preceding the `<app>` in the main text, and it is not encoded as a lemma. There are various strategies: one consists in handling everything at the processing stage, typically by writing an elaborate XPath selecting this word in the XSLT - a perfectly reasonable solution, but requiring a good level of XPath knowledge. Another strategy is to record the desired lemma, as you wish it to appear in the future notes, in an unused attribute of `<lem>` - for instance, `@n`. Our encoding would in this case be the following, after documenting our particular usage of `@n`:

```
... ut dum spiritus reficitur dolores corporis <app>
  <lem wit="#A #C" n="corporis"/>
  <rdg wit="#B" cause="dittography">corporis</rdg>
</app> leniantur...
```

4.1.3. Transpositions

There is a transposition in a witness when the order of the words, or sentences, or paragraphs from the exemplar is changed. In this example, for instance, manuscript B has “corporis dolores” where the other manuscripts have “dolores corporis”:

A

... ut dum spiritus reficitur
dolores corporis leniantur...

B

... ut dum spiritus reficitur
corporis dolores leniantur...

C

... ut dum spiritus reficitur
dolores corporis leniantur...

In a traditional printed edition, transpositions are typically represented by giving the full phrase or sentence in the lemma, and only the first letter(s) of the transposed words in the reading part of the note, to save space on the page:

dolores corporis] c. d. *B*

Transpositions are the most difficult type of variants to encode, at least with the Parallel segmentation method. In simple cases, like the above, the encoding is rather simple:

```
... ut dum spiritus reficitur <app>
  <lem wit="#A #C">dolores corporis</lem>
  <rdg wit="#B" type="transposition">corporis dolores</rdg>
</app> leniantur...
```

When, however, a passage is moved to a different place in the text, for instance when the third paragraph of a text is moved to the first place in a witness, the encoding is a little more complex. To achieve this, we are going to use a linking mechanism available in the TEI, and more precisely the **@copyOf** attribute.

Let's consider two hypothetical witnesses of a nursery rhyme. In manuscript B, although the text is the same, the usual order of the verses has been changed: the third and fourth verses are swapped.

A

Hey diddle diddle, the cat and the fiddle
The cow jumped over the moon
The little dog laughed to see such sport
And the dish ran away with the spoon

B

Hey diddle diddle, the cat and the fiddle
The cow jumped over the moon
And the dish ran away with the spoon
The little dog laughed to see such sport

The first step is to give an **@xml:id** to each of the elements which are transposed. Here, we are working on lines of verse which can be encoded with the **<1>** element. The mechanism would be identical if we were dealing with words (**<w>** element), sentences (**<s>**), paragraphs (**<p>**) or even divisions (**<div>**). Of course it would be perfectly acceptable to give an **@xml:id** to each **<1>**, but it is not necessary for our purpose, so in this example we will do that only for the verses swapped in manuscript B:

```
<1>Hey diddle diddle, the cat and the fiddle</1>
<1>The cow jumped over the moon</1>
<1 xml:id="verse3">The little dog laughed to see such sport</1>
<1 xml:id="verse4">And the dish ran away with the spoon</1>
```

Now that the transposed elements have an **@xml:id**, we can add the markup for an apparatus entry. In the lemma, we simply put the two lines of verse swapped in manuscript B. In a **<rdg>**,

we represent the transposition by adding two `<1>` elements which are virtual copies of verse 3 and verse 4 respectively. This is done with the help of the `@copyOf` attribute, which is used to point to an element of which the current element is a copy.

```

<1>Hey diddle diddle, the cat and the fiddle</1>
<1>The cow jumped over the moon</1>
<app>
  <lem wit="#A">
    <1 xml:id="verse3">The little dog laughed to see such sport</1>
    <1 xml:id="verse4">And the dish ran away with the spoon</1>
  </lem>
  <rdg wit="#B">
    <1 copyOf="#verse4"/>
    <1 copyOf="#verse3"/>
  </rdg>
</app>

```

4.1.4. Nested apparatus entries

It is also worth mentioning that variants may be “nested.” Let us consider the following example: manuscript B has a substitution (“laniantur” instead of “leniantur”), but manuscript C has an omission spanning over the same passage (“dolores corporis leniantur” is omitted)

A	B	C
... ut dum spiritus reficitur dolores corporis leniantur...	... ut dum spiritus reficitur dolores corporis <i>laniantur</i> ut dum spiritus reficitur _____ ...

To encode this, we simply nest the shortest variant inside the lemma of the longest:

```

... ut dum spiritus reficitur <app>
  <lem wit="#A #B">dolores corporis <app>
    <lem wit="#A">leniantur</lem>
    <rdg wit="#B">laniantur</rdg></app>
  </lem>
  <rdg wit="#C"/>
</app>...

```

4.1.5. Overlap

As a final note, it may happen that overlaps occur between variants. The risk of having to deal with such phenomena increases greatly with the number of manuscripts used in an edition. Let us consider the following example: manuscript B has a substitution (“mentis laniantur” instead of “corporis leniantur”), but manuscript C has an omission overlapping the same passage (“reficitur dolores corporis” is omitted)

A	B	C
... ut dum spiritus reficitur dolores corporis leniantur...	... ut dum spiritus reficitur dolores <i>mentis lanian</i> tur...	... ut dum spiritus _____ leniantur ...

We will discuss in details the various methods available to handle overlaps at the end of the chapters dedicated to *Citation and References*. Here we will only consider encoding the example according to the Fragmentation and Reconstitution of Virtual Elements method, but you are invited to consider other methods. Here we may split in two the substitution in B, which will lead us to have a nested variant:

```
... ut dum spiritus <app>
  <lem wit="#A #B">reficitur dolores <app>
    <lem wit="#A">corporis</lem>
    <rdg wit="#B">mentis</rdg>
  </app>
</lem>
<rdg wit="#C"/></app>
<app>
  <lem wit="#A #C">leniantur</lem>
  <rdg wit="#B">laniantur</rdg>
</app> ...
```

4.2. Some special cases

In some cases, the apparatus entry may be complicated by other factors: the necessity to add precisions about how some text has been corrected (or at least changed) in a witness, where, and by which hand; or to point out (and maybe correct) inaccurate text; or to supply missing letters or words.

In those cases, the encoding of an apparatus entry must be combined with elements used to represent the more documentary aspects of a primary source. We give here an overview of some of the most common cases.

4.2.1. Word or phrases corrected by the scribe

Scribes make mistakes, which are sometimes corrected in the manuscript, either by the scribes themselves or by later readers. To encode this phenomenon, we must combine the apparatus entry with elements used in transcriptions. In this case, we can use the **<subst>** mechanism, used to encode a “substitution,” which in TEI “groups one or more deletions with one or more additions when the combination is to be regarded as a single intervention in the text”⁷ In other words, a substitution happens when the scribe or a medieval reader deletes a mistake (by scraping the parchment, crossing over the wrong word, or expunctuating it for instance), and in the same movement adds what he considered the correct word.

If we had the following situation:

7. See section Substitutions in the chapter dedicated to Transcription. Note that what the TEI calls a “substitution” is different from what philologists call a “substitution.”

A	B	C
... ut dum spiritus reficitur dolores corporis leniantur...	... ut dum spiritus reficitur corporis dolores corporum leniantur...	... ut dum spiritus reficitur dolores corporis leniantur...

In a printed edition, we would expect a footnote along those lines, expressing the fact that manuscript B bears the word “corporis,” but that it is a correction from the initial “corporum”:

corporis *corr. ex* corporum *B*

Here is how we could combine the substitution and the apparatus entry in our encoding. Note how we can use the value of **@rend** on **** to record how the deletion was effected in the manuscript (here an expunction), if this information is relevant to our edition:

```
<app>
<lem wit="#A #C">corporis</lem>
<rdg wit="#B">
<subst>
<del rend="expunctuated">corporum</del>
<add>corporis</add>
</subst>
</rdg>
</app>
```

If you wish to indicate which hand did what, you can add a **@hand** attribute to **<subst>**, to differentiate substitutions made by the first copyist from those made by later hands.⁸

4.2.2. Corrections or additions in special places (margin, interlinear...)

Corrections and additions may occur in various places: inline (typically when a scribe realises he has made a mistake just after having written a word, expunctuates or scrapes it and writes the right word), between the lines (usually above the place where a word must be added or corrected), or in a margin of the page. To represent this, we may use the **@place** attribute on **<add>**. For a list of the suggested values for **@place**, see section XXX

In the example above, for instance, the correct word “corporis” has been added above the line. We could therefore add this precision in the encoding:

```
<app>
<lem wit="#A #C">corporis</lem>
<rdg wit="#B">
<subst>
<del rend="expunctuated">corporum</del>
<add place="above">corporis</add>
</subst>
</rdg>
</app>
```

8. For a discussion of the mechanism allowing to record hands and handshifts in TEI, see

4.2.3. *Sic*: pointing out inaccurate or incorrect text

The element **<sic>** in TEI “(Latin for thus or so) contains text reproduced although apparently incorrect or inaccurate,”⁹ which corresponds to the definition used by philologists, who add the mention *sic* after an unexpected reading. In the example below, for instance, manuscript B has “*carporis*” instead of “*corporis*,” which is a spelling mistake.

A	B	C
... ut dum spiritus reficitur dolores corporis leniantur...	... ut dum spiritus reficitur dolores <i>carporis</i> leniantur...	... ut dum spiritus reficitur dolores corporis leniantur...

In a traditional printed edition, we could represent this information as follows:

corporis] carporis sic B

This is how we could encode this phenomenon:

```
... ut dum spiritus reficitur dolores <app>
  <lem wit="#A #C">corporis</lem>
  <rdg wit="#B"><sic>carporis</sic></rdg>
</app> leniantur...
```

In this case, it is obvious enough that the inaccurate word is a misspelt version of “*corporis*,” although in some cases it is worth mentioning explicitly what we think the correct form should be. In a traditional printed edition, we could represent this as follows:

corporis] carporis sic pro corporis B

The corresponding encoding would be:

```
... ut dum spiritus reficitur dolores <app>
  <lem wit="#A #C"/>
  <rdg wit="#B">
    <choice>
      <sic>carporis</sic>
      <corr>corporis</corr>
    </choice>
  </rdg>
</app> leniantur...
```

4.2.4. *Suppleamus*: when the editor supplies missing text

When the text of a witness is damaged or unclear, the editors may have to supply their own conjecture. Let us imagine for instance that, in yet another variant of our example, manuscript B has an extra word added after “*corporis*,” which has been made illegible by some later damage to the manuscript (for instance a stain):

9. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-sic.html>

A

... ut dum spiritus reficitur
dolores corporis leniantur...

B

... ut dum spiritus reficitur
dolores corporis
leniantur...

C

... ut dum spiritus reficitur
dolores corporis leniantur...

If conjecture as to the nature of this word is impossible, we may simply consider that it is a “locus desperatus” (or “hopeless passage” in Latin), and content ourselves with marking up the words as a damaged part of the manuscript, using the **<unclear>** element, which “contains a word, phrase, or passage which cannot be transcribed with certainty because it is illegible or inaudible in the source.”¹⁰ This element has various useful attributes, among which **@reason** that lets us specify what makes the passage unclear, and **@extent** that lets us give a full-text estimate of the number of letters or words of the unclear passage:

```
... ut dum spiritus reficitur dolores corporis <app>
<lem wit="#A #C"/>
<rdg wit="#B"/>
</app> leniantur...
```

When a conjecture is possible, the editors will indicate which word(s) should be in the text according to them, using the **<supplied>** element, which “signifies text supplied by the transcriber or editor for any reason; for example because the original cannot be read due to physical damage, or because of an obvious omission by the author or scribe.”¹¹ Among the attributes available for **<supplied>**, **@reason**, which may be used to explain why this word had to be supplied by the editors. Here, for instance, the editors think that the illegible word is “uestri,” meaning “your”

```
... ut dum spiritus reficitur dolores corporis <app>
<lem wit="#A #C"/>
<rdg wit="#B"><supplied reason="damage">uestri</supplied></rdg>
</app> leniantur...
```

In some cases, there might be various hypothesis regarding the elucidation of the illegible passage. Here for instance, the word added in manuscript B could be read as “uestri” (“your”), or maybe as “nostri” (“our”). To express those two possible readings of the illegible word, we could use the **<choice>** element to wrap several possible **<supplied>** elements. The optional **@cert** (certainty) attribute on **<supplied>** could be used to indicate our level of certainty regarding each supplied possibility. Here for instance, if we are fairly certain that the word is “uestri,” but still think there is a possibility it might be “nostri,” we could encode the phenomenon like this:

10. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-unclear.html>

11. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-supplied.html>

```

<app>
  <lem wit="#A #C"/>
  <rdg wit="#" B"><choice>
    <supplied cert="high">uestri</supplied>
    <supplied cert="low">nostri</supplied>
  </choice></rdg>
</app>

```

When different editors have different opinion on the right solution, we can use the `@resp` attribute to indicate who suggests to supply what:

```

<choice>
  <supplied cert="high" resp="#MB">uestri</supplied>
  <supplied cert="low" resp="#PJ">nostri</supplied>
</choice>

```

4.2.5. Displaying long lemmata

To finish up this long and rather dense chapter, let us look at a practical issue: we might have in our edition long lemmata, spanning over several lines of text, if not several paragraphs (for a long omission, for instance). This is not an issue in itself, but it may pose a problem if we wish to display our edition in a traditional fashion, in print or online, with footnotes giving some information to the readers. When a lemma is longer than a couple of words, the traditional convention is to simply give its first and last word separated by an ellipsis, instead of repeating the whole passage (the first and last few significant words may be given when the passage is long, or to avoid confusion).

Let us imagine for instance that a manuscript omits these two long sentences:

```

<app>
  <lem wit="#A #C">Amicitie primum, ut mihi uidetur, ipsa natura
humanis mentibus impressit affectum, deinde experientia auxit,
postremo legis auctoritas ordinauit. Deus enim summe potens, et summe
bonus, sibi est ipsi sufficiens, qui bonorum nostrorum non eget;
uoluit ut omnes creaturas suas pax componeret, et uniret
societas.</lem>
  <rdg wit="#" B"/>
</app>

```

In a traditional printed critical edition, the footnote about this phenomenon would have a shortened version of the lemma, like this:

amicitie primum ... uniret societas] *om. B*

If we wish to display the same sort of notes for some version of our digital edition, we have two solutions:

- Leave our encoding as above, and process the lemma (typically in XSLT) to transform its text into the shortened version before displaying it. It is a good solution for people who

are familiar enough with a suitable processing language, and also if the encoding is not too complicated by nested variants, which is likely to happen in a very long lemma.

- Add an element to our encoding, to record how we wish the lemma to be displayed in notes. There is no standard way of doing this, but we could use for instance the generic **<note>** element, with a specific **@type** attribute value (for instance "**altLem**" for "alternative to the lemma"):

```
<app>
  <lem wit="#A #C">Amicitie primum, ut mihi uidetur, ipsa natura
  humanis mentibus impressit affectum, deinde experientia auxit,
  postremo legis auctoritas ordinavit. Deus enim summe potens, et
  summe bonus, sibi est ipsi sufficiens, qui bonorum nostrorum non
  eget; uoluit ut omnes creaturas suas pax componeret, et uniret
  societas.</lem>
  <rdg wit="#B"/>
  <note type="altLem">amicitie primum ... uniret societas</note>
</app>
```

With this encoding, it is much easier to display a shortened version of the lemma in notes when we need to, as an alternative to the full text.

Bibliography

- Marjorie Burghart, TEI: Critical Apparatus Cheatsheet, 2011,
<http://marjorie.burghart.online.fr/?q=en/content/tei-critical-apparatus-cheatsheet>
- David C. Greetham, “A History of Textual Scholarship,” in Neil Fraistat and Julia Flanders (eds), *The Cambridge Companion to Textual Scholarship*, Cambridge:Cambridge University Press, 2013, 16–41.
- Patrick Sahle, *A Catalog of Digital Scholarly Editions*, 2008, <http://www.digitale-edition.de/vlet-about.html>

Facsimile and Document-Centric Editing

Elena Pierazzo

Digital facsimiles have become ubiquitous in editorial practices, whether the outcome is to be in digital form only or in print or both. This chapter explores the ways the TEI handles the use of facsimile within the production of digital editions, and connects them to different theoretical editorial approaches and conceptions. Readers will learn how to establish a relation between a portion of a text and a digital facsimile or a portion of it; they will learn how to *zone* an image and also how to embed a transcription within an image; they will lastly learn how to document and time the process of producing a text.

1. Facsimile and (Digital) Editing

The combination of the affordability of digital photography and the rise of digital methods in editing has brought the images of manuscripts and documents in the scholarly spotlight. The high quality digital reproductions of primary sources has represented a major improvement with respect to microfilms, and a much cheaper option with respect to photography, and it has been saluted with great enthusiasm by scholars around the world. The improvement has been so remarkable that some scholars have even been carried away by their enthusiasm, declaring that once you have a digital image “we need never see the document itself” (Twycross 2008, p. 23) and that the facsimile provide an “unmediated” access to the source (Nell Smith, 2004, p. 309), or that a facsimile edition subsumes the role of diplomatic editions (Kiernan 2006, p. 266).¹

However exciting the new phase might be, the use of facsimile and the inclusion of reproduction of documents in editions is not a new fact in textual scholarship; editors have in fact tried to provide their readers with some sort of access to the documentary sources for centuries; since the sixteenth century typesetters tried to reproduce in print the look of ancient scripts,² while engravings reproducing charters and manuscripts were firstly used by Jean Mabillon in his *Re Diplomatica* (1st ed. 1681) (Stokes 2010, 14). Such publications were nevertheless very expensive and therefore extremely rare; even when photographs replaced engravings, making the publication of representations marginally more affordable, publications including facsimile remained scarce and expensive. The easy availability and reduced cost of digital photographs have determined a more pervasive presence of reproductions within the editorial practices, not to mention the centrality they have gained in any digital edition to the point that producing a digital edition without providing access to the digitised sources has become extremely rare. Such change, which is a direct consequence of the major technological innovation that has brought us digital photography, is also coincidental with important changes in textual scholarship itself,

1. See Pierazzo 2015, pp. 93–98

2. See for instance *A testimonie of antiquity*, collected by John Joscelyn and Matthew Parker, published in 1566 by Iohn Day.

namely with a renewed attention given to documentary evidence brought in by different scholars and scholarly approaches (Pierazzo 2104).

These combinations of technological innovation and changes in scholarly theorisation have on the one hand had a major impact on editorial workflow, and on the other hand have created new expectations on the part of the readers (and users) of digital scholarly editions (Sutherland and Pierazzo, 2012, p. 202). But there is more: in fact the combined changes brought in by the availability of reproductions of primary sources and by the new textual scholarship attention to documents are determining changes on the perceptions of textuality itself, on what text is and what it is supposed to convey to its readers (Pierazzo 2016). These new perspectives and new attention to documents have been even considered dangerous by certain scholars (Robinson, 2013c, p. 127), but while calling it dangerous is probably an overstatement, it is without a doubt that this change is unsettling and has manifold implications for scholarship and practices (Treharne 2013).

From a practical point of view, the inclusion of facsimile within a digital edition can take different forms and role, from being an accessory to illustrate the scholarly argument brought forward by the edition, to taking centre stage and becoming the focal point of the edition. This difference in weight is the result of different workflows which in turns reflects different conceptions and editorial purposes, and while a digital edition can be produced using different tools and methods, it is without a doubt that for the past thirty years the Text Encoding Initiative has provided (and still does) a very solid, scholarly-based approach to the handling of facsimile within the editorial work.

2. The TEI and the facsimiles

The TEI has been part of this pivotal change from the very beginning, but particularly in the last few years with a provision of new facilities for the handling of representations of source documents. Admittedly, the earlier version of the TEI (the so-called P3 and P4 releases) provided limited support for facsimile images, the inclusion of which was mostly mandated to the imaginative use of a handful of generic elements. However, since the release of the P5 version in 2007, the TEI has been on the front line of documentary editing. In fact this version has introduced an entire new module for the handling of facsimile and for the connection of images, or portions of images, to transcriptions or editions.³ Furthermore, since 2012 the TEI has provided a further support to documentary editing in the form of new specialised elements for the inclusion of transcription within, so-to-speak, the facsimile itself, therefore allowing for the production of hyper-diplomatic editions, able not only to reproduce the text in documents to a great extent of details, but also its layout, including also non-textual features such as blurs and doodles.

This chapter will present the different ways the TEI offers to the editor for the inclusion of facsimile into digital editions, and will be divided in two main sections discussing, respectively, the connection of an edition to an image through the use of the **<facsimile>** element, and the

3. The entire following discussion subsume the reading of chapter 11 of the TEI Guidelines (see <http://www.tei-c.org/Vault/P5/current/doc/tei-p5-doc/en/html/PH.html>)

embedding of an edition within an image, through the use of the `<sourceDoc>` element.

3. Side-by-side: editions and `<facsimile>`

There are several ways one can put in relation an image to a portion of a text within the TEI framework, most of them are connected to the use of the `<facsimile>`⁴ element which is meant to contain information about a facsimile image (or group of images) that is related in some way to a text or, more commonly, to a portion of a text (the content of a page, of a document, of a side of a stone, etc.); once that the presence of a facsimile has been declared, the actual connection between the information contained in the `<facsimile>` and the relevant portion of the text itself is then performed via the global attribute `@facs`.⁵ The purpose of this attribute is in fact to contain a reference (an URI) to an image, by the means of a reference to an identifier (established by an `@xml:id` within the `<facsimile>` element) or thanks to a link to the actual file containing the image of the facsimile itself. In fact, it is even possible to simply use the `@facs` attribute without having to use the element `<facsimile>` at all, which represents the easiest way to connect a text to an image, as in the following example which typically uses the `@facs` attribute on the `<pb>` (i.e., *page-beginning*)⁶ element; the example then connects then an image of a page of a manuscript to the element that signal the beginning of a new page in the transcribed source.

```
<body>
  <pb n="001r" facs="../images/SBK-CC1-126_001r.jpg"/>
  <div>
    <argument rend="rubric">
      <p>Incip<ex>it</ex> lib<ex>er</ex> prim<ex>us</ex> de
        inge<ex>n</ex>iosa<ex>n</ex>itatis ...</p>
    </argument>
    <p>...</p>
  </div>
</body>
```

This method presents the double advantage of being very straightforward and lightweight; however it only allows for a connection of a text (or a page of a text) to one image at the time, which might prove insufficiently detailed or nuanced in some circumstances. For instance, it might be useful to be able to connect a portion of a text to only a portion of an image; or it could be desirable for an edition to offer to the users several versions of the image containing a page of the source, for instance: a full-sized, high-resolution image for storage and/or for more advanced or registered user, a thumbnail image for browsing, a low resolution image for the side-by-side display, an infrared version for reading the underlying text of a palimpsest, a PDF version for download, and so on. For all this cases, the use of the `@facs` attribute is not sufficient and it is necessary to use the `<facsimile>` element.

4. <http://www.tei-c.org/Vault/P5/3.2.0/doc/tei-p5-doc/en/html/ref-facsimile.html>

5. <http://www.tei-c.org/Vault/P5/3.2.0/doc/tei-p5-doc/en/html/ref-att.global.facs.html>

6. <http://www.tei-c.org/Vault/P5/3.2.0/doc/tei-p5-doc/en/html/ref-lb.html>

3.1. One surface, many images

The **<facsimile>** element is a direct child of the root element **<TEI>**, and has to be used immediately after the **<teiHeader>** and before the **<text>** element.

```
<TEI>
  <teiHeader>
    <!-- Here metadata -->
  </teiHeader>
  <facsimile>
    <!-- Here the list of all facsimile -->
  </facsimile>
  <text>
    <!-- Here the transcription of the text -->
  </text>
</TEI>
```

The **<facsimile>** element typically contains one or many **<surface>** elements,⁷ corresponding to a page or any type of surface over which writing is found. Each of them is to contain the references of the image file(s) representing in some way the said surface. It is also good practice to give this element an **@xml:id** attribute to which then one can point to.

The link to the actual files storing the images is normally performed via a **<graphic>** element,⁸ or a sequence of them if the edition is to provide several versions of the same surface. The **<graphic>** element is normally an empty one, only providing a link to the image file via a **@url** attribute, but it can also contain a **<desc>** element to provide useful information about the usage of the image, its content or any other information that might be relevant.

In the following example, the **<facsimile>** element has been used to declare the availability of several versions of the same image for different purposes. The **<surface>** element has been given an **@xml:id** attribute, consisting in an arbitrarily chosen sequence of characters providing a unique identifier of the surface, and an **@n** attribute, providing the folio number of the surface itself. Once all the images of that surfaces have been declared and described, it is possible to refer to the surface from the transcribed and/or edited text using the **@facs** attribute that this time does not provide a direct link to an image file, but a reference to the **@xml:id** of the **<surface>** element. When the edition is ready to be published, it will be then the role of some types of scripts (typically XSLT scripts) or of an API to choose the appropriate image among the ones available for the surface and to display it according to the situation, as decided by the editors.

7. <http://www.tei-c.org/Vault/P5/3.2.0/doc/tei-p5-doc/en/html/ref-surface.html>

8. <http://www.tei-c.org/Vault/P5/3.2.0/doc/tei-p5-doc/en/html/ref-graphic.html>

```

<facsimile>
  <surface n="001r" xml:id="c001r">
    <graphic url="../images-low-res/SBK-CCL-126_001r.jpg">
      <desc>Low Resolution</desc>
    </graphic>
    <graphic url="../images-archival/SBK-CCL-126_001r.tiff">
      <desc>High Resolution</desc>
    </graphic>
    <graphic url="../images-thumb/SBK-CCL-126_small_001r.jpg">
      <desc>Thumbnail</desc>
    </graphic>
    <graphic url="../infrareds/SBK-CCL-126_infra_001r.tiff">
      <desc>Infrared version</desc>
    </graphic>
  </surface>
</facsimile>
<text>
  <body>
    <pb facs="#c001r"/>
    <div>
      <argument rend="rubric">
        <p>Incip<ex>it</ex> lib<ex>er</ex> prim<ex>us</ex> de
        inge<ex>n</ex>iosa<ex>n</ex>itatis ...</p>
      </argument>
      <p>...</p>
    </div>
  </body>
</text>

```

3.2. One surface, many zones

The **<facsimile>** element can also be conveniently used to connect only portions of images to portions of texts. This could be useful if the edition is to provide detailed commentary on decoration or to relate marginal additions or comments to their actual position, or for any other circumstance when an editor wishes to discuss a particular detail of a surface. For that purpose, the **<surface>** element may contain one or more **<zone>** elements,⁹ each of them representing an arbitrarily (or better: editorially) defined area of an image. The area itself is delimited thanks to a series of attributes on the **<zone>** element containing spatial coordinates that allow to draw either a rectangular or polygonal area.

Let us suppose that we want to draw a rectangle to encase the left hand-side of the folio 1r of Cod. 126 of the Klosterneuburg, Augustiner-Chorherrenstift Library in Austria¹⁰ in order to be able to comment on the iconography of the decoration, and in particular on the archer below the drop capital. In this case we can use a series of four attributes (or better: two couples of

9. <http://www.tei-c.org/Vault/P5/3.2.0/doc/tei-p5-doc/en/html/ref-zone.html>

10. A description of the manuscript can be found at <http://manuscripta.at/m1/lib.php?libcode=AT5000>.

attributes) to provide the coordinates corresponding to a rectangle-shaped zone over the image. This operation is performed thanks to a familiar geometric mechanism, namely by considering the image as a bi-dimensional Cartesian space, where any point can be defined by two values, i.e., the “abscissa” (or “x” value) and the “ordinate” (or “y” value) coordinates of that particular point. It is important to note, however, that with respect to the way Cartesian coordinates normally work (i.e., with origin positioned on the lower-left corner of a bi-dimensional space), coordinates on screens are calculated from the upper-left corner.

Now, in a Cartesian space a rectangle can be ideally drawn by providing the coordinates of only two diagonally opposite corners of that rectangle, and then letting the computer to infer the coordinates of the other two remaining corners based on the ones provided. For the drawing of a rectangle, the TEI has chosen to provide the coordinates of the top left corner and the bottom-right corner, therefore the attributes offered by the TEI are:

Coordinates for the upper-left corner

- **@ulx**: provides the upper-left X value
- **@uly**: provides the upper-left Y value

Coordinates for the lower-right corner

- **@lrx**: provides the lower-right X value
- **@lry**: provides the lower-right Y value

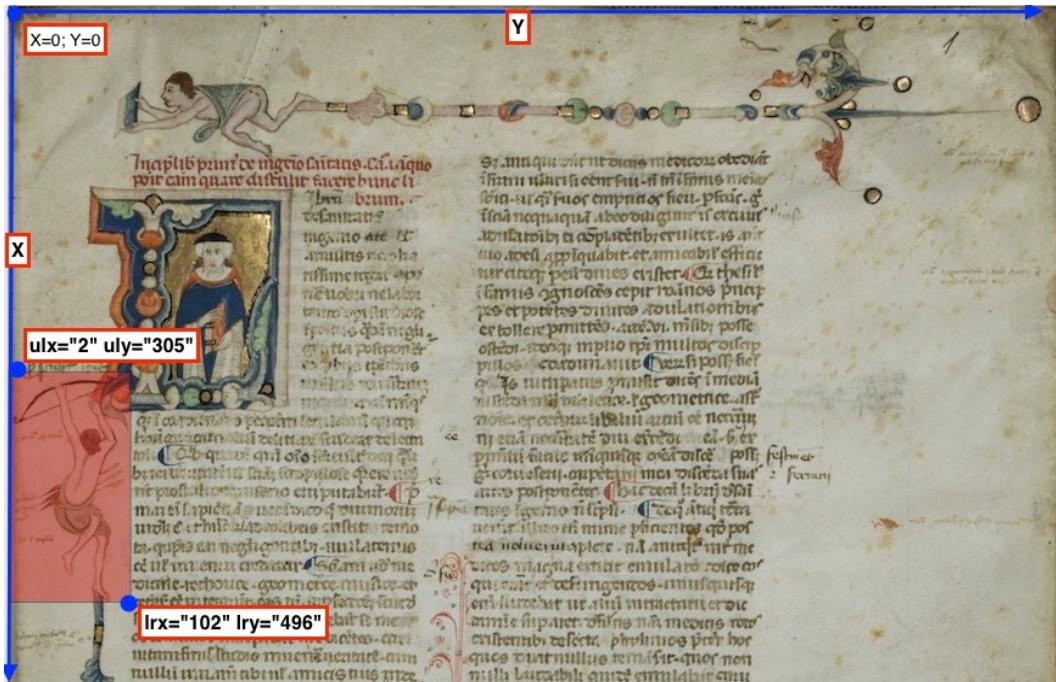


Figure 1: Rectangle drawn upon f. 1r of Cod. 126 of the Klosterneuburg, Augustiner-Chorherrenstift Library

The **<zone>** element in the snippet below is then provided with the attributes containing the coordinates for the upper-left and lower-right corner of a rectangle;¹¹ it also has an **@xml:id** attribute which is then being referenced by the **@faccs** attribute in a **<note>** (but it could be a paragraph or any other appropriate element used for annotation) describing the content of the zone.

```
<facsimile>
  <surface n="001r" xml:id="c001r">
    <graphic url="SBK-CC1-126_001r.jpg" width="1481px"
height="953px"/>
    <zone ulx="2" uly="305" lrx="102" lry="496" xml:id="c001r-
archer"/>
  </surface>
</facsimile>
<text>
  <body>
    <pb faccs="#c001r"/>
    <div>
      <argument rend="rubric">
        <p>Incipit liber primus de  

ingenio sanitatis ...</p>
      </argument>
      <p>...</p>
      <note type="decoration" faccs="#c001r-archer">Archer on the left
margin</note>
    </div>
  </body>
</text>
```

A similar approach can be adopted in all those cases in which it is considered appropriate to connect a portion of a text with a portion of an image, for annotating or documenting or providing support to the editorial arguments. It is also possible to use the **@faccs** attribute from within the **<msDesc>** element; in this latter case it is particularly appropriate to use it from the **<locus>** element. For instance:

11. The coordinates have been determined thanks to the TEI Zoner, a web-based tool developed by Chris Sparks and available from the TEI Critical Apparatus Toolkit, <http://ciham-digital.huma-num.fr/teitoolbox/zoner.php>.

```

<msDesc>
  <msIdentifier>...</msIdentifier>
  <msContents>
    <msItem>
      <incipit><locus from="1r" facs="#c001r"/>Incipit Liber primus de
ingenio sanitatis</incipit>
    </msItem>
  </msContents>
  <physDesc>
    <decoDesc>
      <decoNote facs="#c001r-archer">Archer on the left
margin.</decoNote>
    </decoDesc>
  </physDesc>
</msDesc>

```

Rectangles are at times not the best way to encase decorations or jots. In these case the four attributes holding the coordinates can be substituted by a single attribute called `@points` which syntax allows to supply any number of couples of coordinates, each of them representing the X and the Y values of a point, parted by a comma; each couple is then separated from the next one by a space. The following example represents a polygon encasing a complex decoration at folio 7r of Cod. 253 of the Klosterneuburg, Augustiner-Chorherrenstift Library.



Figure 2: Polygon drawn upon f. 7r of Cod. 253 of the Klosterneuburg, Augustiner-Chorherrenstift Library

```

<surface xml:id="c007r">
  <graphic url="../images/SBK-CC1-253_7r.jpeg" width="1055px"
height="1652px"/>
  <zone points=" 352,33 474,75 546,288 580,319 574,355 616,414
739,523 762,565 738,640 747,666 736,693 757,729 757,762 696,940
777,1089 732,1191 682,1227 468,1243 141,1273 57,1209 42,993 28,714
37,514 148,355 138,213 195,96 351,33 ">
</surface>

```

The zoning mechanism could also be used to record the exact position of a textual addition or a variant. In the following example taken from folio 1r of Cod. 110 of the Klosterneuburg, Augustiner-Chorherrenstift Library¹² the word *divine* has been added on the right margin and connected to the main text thanks to a caret.

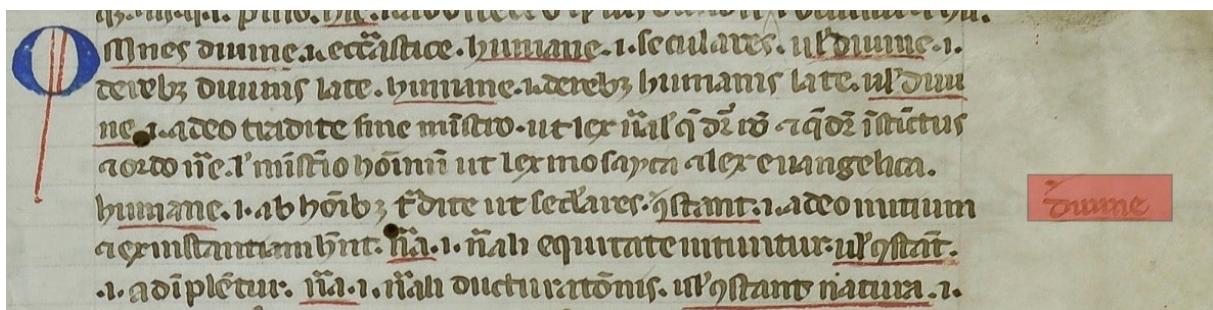


Figure 3: Addition to right margin of Cod. 110, f. 1r, of the Klosterneuburg, Augustiner-Chorherrenstift Library.

The encoding snippet below features the `<zone>` element surrounding the word *divine* within the `<facsimile>` section, as well as the `@facs` attribute used on the `<add>` element pointing at the `<zone>`.

```

<facsimile>
  <surface n="001r" xml:id="c001r">
    <graphic url="../images/SBK-CCL-110_001r.jpg" width="1312px"
height="312px"/>
    <zone ulx="1002" uly="170" lrx="1143" lry="217" xml:id="add-01"
type="marginal-addition"/>
  </surface>
</facsimile>
<text>
<body>
  <pb facs="#c001r"/>
  <div>
    <p>...humane id est ab hominibus tradite ut sec<ex>u</ex>lares.
<metamark function="add" type="caret"/> <add facs="#add-01"
place="margin">divine</add> constant id est a deo initium </p>

```

12. A description of the manuscript can be found at http://manuscripta.at/m1/hs_detail.php?ID=131.

```
</div>
</body>
</text>
```

4. Embedded: editions and <sourceDoc>

From an editorial perspective, the use of the **<facsimile>** element as a sibling of the **<text>** element responds to a vision of the editorial work that aims at reconstructing the text of a document or of a work, or that seeks to produce a diplomatic edition of one or more witnesses of a work alongside a critical edition. All of these editorial purposes assume however that there is a text, and that the text can be determined editorially from the extant documents and that determining it is the purpose of the editions. However, there are cases when this seemingly obvious purpose is not so simple to achieve, and these are cases where the physicality of the document is overwhelmingly present (as for draft manuscripts or highly decorated ones) or where the visual component of the document has a major importance from a cultural, historical or editorial point of view.

In cases like the ones described by Pierazzo and Stokes 2010 the extraction of texts from documents determines a loss of information that could be deemed unacceptable by editors and readers. In some other cases, the purpose of editing itself may be to explore the materiality of the document or the process of creating texts and documents, studying the stratification of the ink traces left by the author or by readers. In these cases fitting the textual component of a document in a linear way, organising them in paragraphs or verses, does not seem to be the best way to represent either these scholarly purposes or the peculiarity of the documents. In these cases, the TEI offers the possibility of transcribing the words on the page following a documentary approach as opposed to a textual approach, i.e., page by page and line by line instead that paragraph by paragraph, which consists in migrating the transcription within the **facsimile** instead of keeping it alongside it. In these cases, instead of using the **<facsimile>** as a parent element, **<surface>** and **<zone>** are used as children of a **<sourceDoc>** element.¹³ In this latter case the element **<zone>** becomes a text holder and not only a way to draw virtual shapes over a facsimile. The **<sourceDoc>** element assumes all the functions of the **<facsimile>** element which becomes then redundant.

This new approach privileges the document over the text, therefore it does not allow to use elements that are normally used to enrich and annotate semantically a text; this is the case of elements to mark names of people or places, and also elements to mark dates and events, or any internal textual structure such a verses, paragraphs, speeches, and so on. This situation is specular to the one that only allows to mark the beginning of new lines and pages (elements **<lb>** and **<pb>**, respectively) as empty elements in the textual approach of the TEI. The reason behind the decision of limiting the content of these elements is to prevent the risk of overlapping hierarchies: in fact, when adopting a textual approach and therefore transcribing the text on a source paragraph by paragraph (for instance) a page break may happen within a paragraph, therefore, if the elements for a page was to be non-empty, it would cause an overlap between

13. <http://www.tei-c.org/Vault/P5/3.2.0/doc/tei-p5-doc/en/html/ref-sourceDoc.html>

paragraph and page. In the same way, if adopting a documentary approach, the name of a person, for instance, could be split between two lines, and so on. The TEI therefore only allow strictly *editorial* elements, such as, for instance, **<add>**, ****, **<gap>**, **<ex>**, **<sic>** and few others, which are essential transcriptional phenomena, and also for theses it warns about the fact that some overlapping issues will be inevitable.

The use of the **<sourceDoc>** approach determines also a different use of the **<zone>** than can have three main types of contents:

- Free text, possibly segmented by **<1b>** (line-beginning) elements .
- A sequence of **<line>** elements, each of them representing a typographical written line.
- One or more embedded **<zone>** elements.

The choice between this different contents is to some extent arbitrary, but there are situations in which one choice could be more appropriate than another. The use of the element **<line>** (which is the non-empty counterpart of the **<1b>** element)¹⁴ is recommended only in those cases where the page layout clearly shows lines of text. For instance, in case the text in a zone is not clearly structured in lines, a free text approach could be more appropriate, or if a zone shows signs of having internal structures, sub-zones could be more suitable.

The following example presents the zoning and transcription of some of the marginalia of folio 80r of the Cod. 126 seen above. The text of the first two zones is organised in clear sets of lines, while the third zone contains only one word and the forth zone is wrapped around the decoration and has a more untidy line structure.

14. <http://www.tei-c.org/Vault/P5/3.2.0/doc/tei-p5-doc/en/html/ref-line.html>

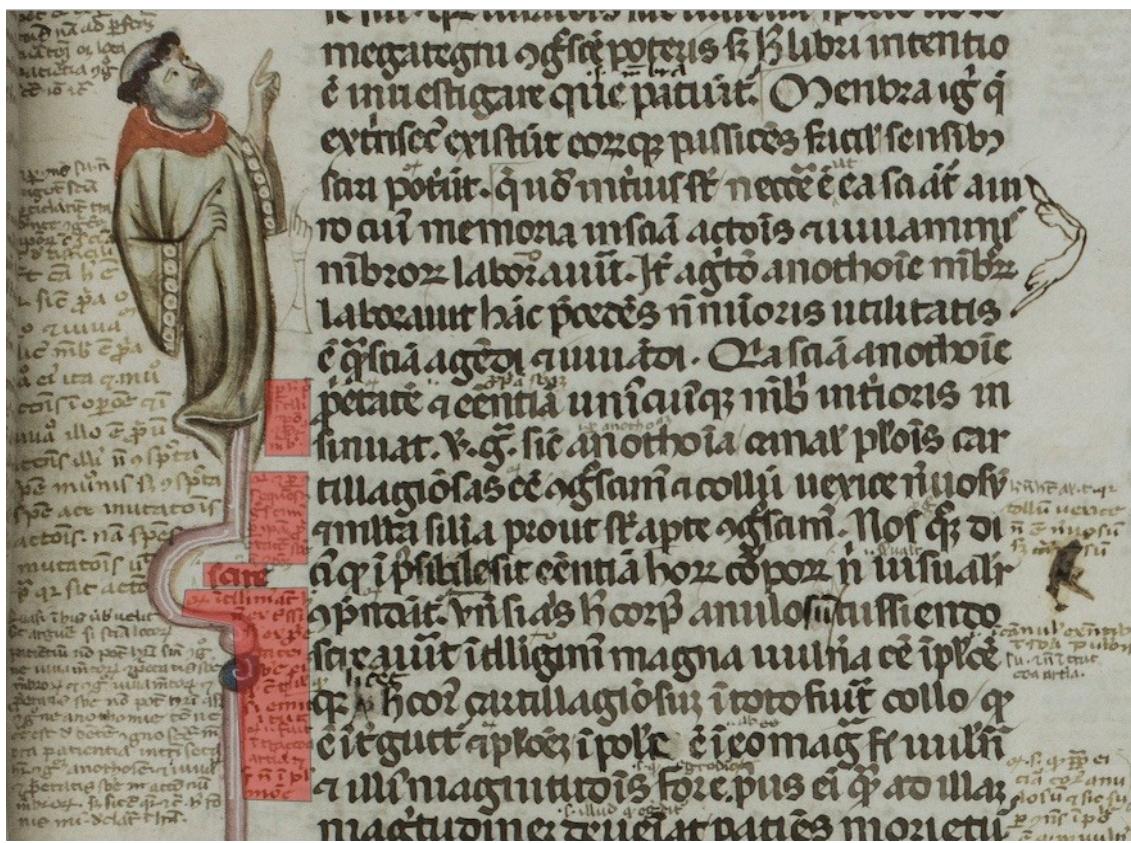


Figure 4: Zoning of cod.126 Klosterneuburg, Augustiner-Chorherrenstift Library

Because of the particularity of each zone, it may not be advisable to adopt the same approach for each of them, so in the example below the first two zones will propose a line by line transcription, while the latter two will have a free text approach, in one case because the use of `<line>` seems unnecessary, given the fact that the zone contains only one word, and in the other because of the unplanned and untidy appearance of the text for which a rigid encasing of text into `<line>` elements seems inappropriate.

```

<sourceDoc>
  <surface xml:id="c007r">
    <graphic url="../images/marginalia.png" width="1761px"
height="1349px"/>
    <zone ulx="405" uly="598" lrx="480" lry="721">
      <line>per hoc potest</line>
      <line>intelligi</line>
      <line>compositio</line>
      <line>et forma</line>
      <line>membri</line>
    </zone>
    <zone ulx="369" uly="753" lrx="477" lry="900"><line><metamark
  
```

```

function="reference"/>et per</line>
<line>consequens</line>
<line>ergo scimus</line>
<line>per ipsam propri-</line>
<line>-etatem substantie</line>
<line>et autem compositionem</line></zone>
<zone ulx="303" uly="895" lrx="421" lry="939">scire</zone>
<zone points="271,946 351,939 484,939 472,1287 369,1285 369,1111
394,1078 387,1024 354,1015 334,987 271,979"><lb/><metamark
function="reference"/>intelligimus autem hoc <lb/>non ex crassi
<lb/>sed ex proprie<lb/>tate <lb/>substantie eius ... </zone>
</surface>
</sourceDoc>

```

Given the fact that zones are independent from one another (they are like self-contained boxes), their graphical plotting on the surface can overlap without problems, as in the example below, where the zones containing the blocks of texts overlap with the zones containing the decoration (see next page).

```

<surface>
  <graphic url="..../images/SBK-CCL-685_074r.jpg" width="686px"
height="1019px"/>
  <zone points="36,236 288,236 258,833 27,825">
    <note>Soldier facing right</note>
  </zone>
  <zone ulx="19" uly="171" lrx="460" lry="264">
    <line>Ex geminis...</line>
  </zone>
  <zone ulx="168" uly="276" lrx="307" lry="378">
    <line>Ysidorus dixit..</line>
  </zone>
</surface>

```



Figure 5: Zoning of cod.685, f. 74r, Klosterneuburg, Augustiner-Chorherrenstift Library

Sometimes manuscripts present portion of writing that are rotated with respect to the orientation of the page; to capture this characteristic, the `<zone>` element is provided with a `@rotate` attribute which value expresses the amount of clockwise rotation that the document had to undergo in order to allow for the rotated writing to be supplied; the value of the attribute is expressed in arc degrees. In the following example there are two sets of zones on the right of the main block writing: four blocks of writing close to the spine rotated by 270 arc degrees with respect to the page orientation, and a smaller addition constituted of only two lines rotated by 90 arc degrees.

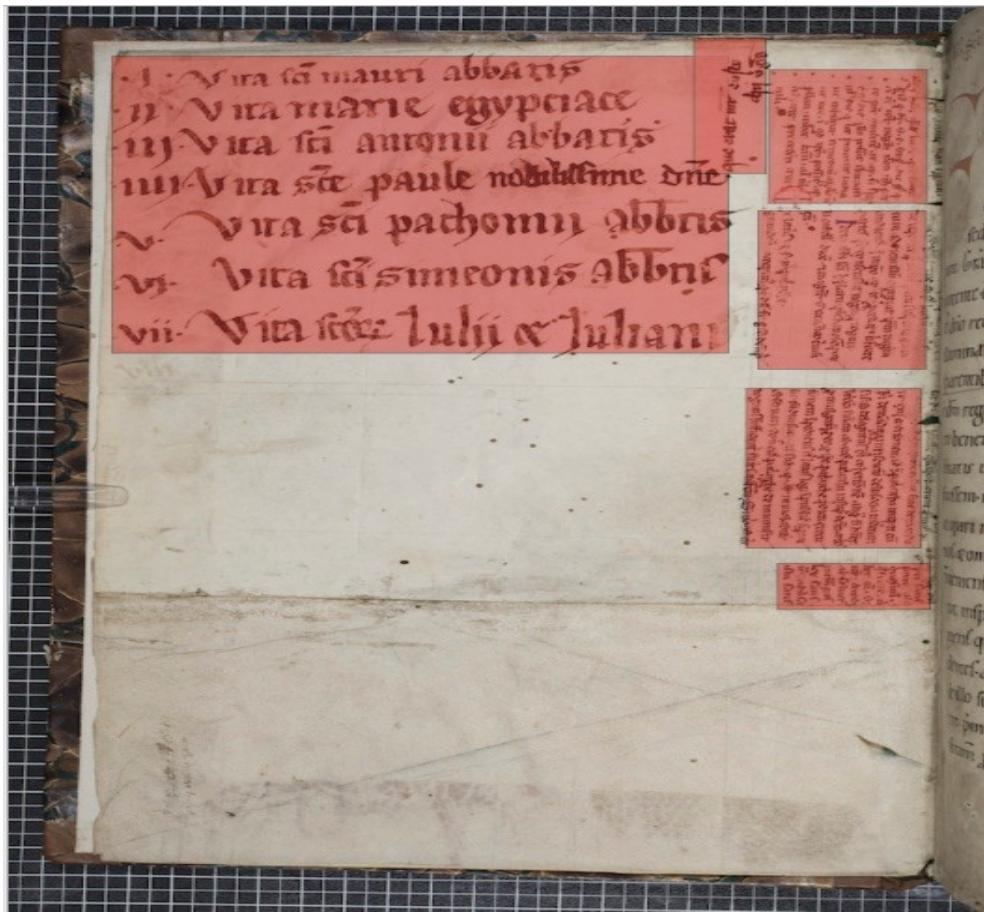


Figure 6: Zoning of cod. 705, f. Iv, Klosterneuburg, Augustiner-Chorherrenstift Library

```

<surface>
  <zone ulx="72" uly="57" lrx="484" lry="366"/>
  <zone ulx="461" uly="38" lrx="509" lry="179" rotate="90"/>
  <zone ulx="511" uly="71" lrx="617" lry="210" rotate="270"/>
  <zone ulx="504" uly="218" lrx="616" lry="383" rotate="270"/>
  <zone ulx="496" uly="403" lrx="613" lry="569" rotate="270"/>

```

```
<zone ulx="517" uly="585" lrx="619" lry="633" rotate="270"/>  
</surface>
```

It is clear how this type of encoding and these preoccupations are only appropriate in very special cases, since the capture of these type of details in the encoding (and not, for instance, within an editorial note) implies that they will be used to process or represent the transcription in a meaningful way. As a rule of thumb, an editor should refrain from applying unnecessary markup, where unnecessary means “not used computationally in any way.” This is because applying markup that is not used in any way distracts from the main purpose of the edition and clutters the encoding which should be otherwise kept as tidy and consistent as possible. On the other hand, only what is explicitly marked up can be later processed, either for display, or statistical analysis or any other purpose, therefore editors should always be very careful with their encoding to allow only the markup that is required by the purpose of their edition, but the one that is required should be consistently there (Pierazzo 2011).

4.1. Composite surfaces

At times it may be helpful to consider a level of analysis that goes beyond the page; for instance, if analysing the decoration of a manuscript, it could be necessary consider units such as openings; the codicology could be better studied at gathering or at bifolium level; damages or ink bleeding through the page may requires to studio the leaf, and so on. The TEI proposes one element to that purpose, **<surfaceGrp>**¹⁵ which requires one or more **<surface>**s as its children; a **@type** attribute can be used to specify the kind of grouping. In the example below two images relatives to f. 71v and f. 72r of ms. 685 of the Klosterneuburg, Augustiner-Chorherrenstift Library that presents a complex and rich iconography.

15. <http://www.tei-c.org/Vault/P5/3.2.0/doc/tei-p5-doc/en/html/ref-surfaceGrp.html>



Figure 7: Opening of f. 71v and f. 72r of ms. 685 of the Klosterneuburg, Augustiner-Chorherrenstift Library

The encoding of the grouping could be looking as in the example below:

```

<surfaceGrp type="opening" n="71v-72r">
  <surface n="71v">
    <graphic url="AT5000-685_71v-d.jpg"/>
    <!-- Here zones -->
  </surface>
  <surface n="72r">
    <graphic url="AT5000-685_72r-d.jpg"/>
    <!-- Here zones -->
  </surface>
</surfaceGrp>
  
```

This type of encoding will later allow to display the edition page by page or opening by opening, according to the preferences of the reader or of the choices of the editors. However, the mechanism does not allow for the creation of <**zone**>s trespassing the border of one surface; if the particular characteristics of a document or the purpose of an edition requires such trespassing zone to be established, the editors may consider to consider the entire opening as a <**surface**> in its own right, nesting two surfaces, as in the example below; in fact it could be arguable that if a decoration or some writing acts spread over the two faces of the opening, the opening has been considered as a single surface by the scribe or the artist.

```
<surface type="opening" n="71v-72r">
  <graphic url="opening.jpg"/>
  <!-- Here zones at opening level -->
  <surface n="71v">
    <graphic url="AT5000-685_71v-d.jpg"/>
    <!-- here zones relative to the page -->
  </surface>
  <surface n="72r">
    <graphic url="AT5000-685_72r-d.jpg"/>
    <!-- here zones relative to the page -->
  </surface>
</surface>
```

The <**sourceDoc**> approach is also helpful in cases where the writing support (i.e., the <**surface**>) is fragmentary or made from different materials, for instance when patches have been glued or stitched to the main surface. The image below (taken by a miscellaneous liturgical manuscript of the Capitular Library of Vercelli, Italy), shows a page where a cartouche has been stitched to the main page.

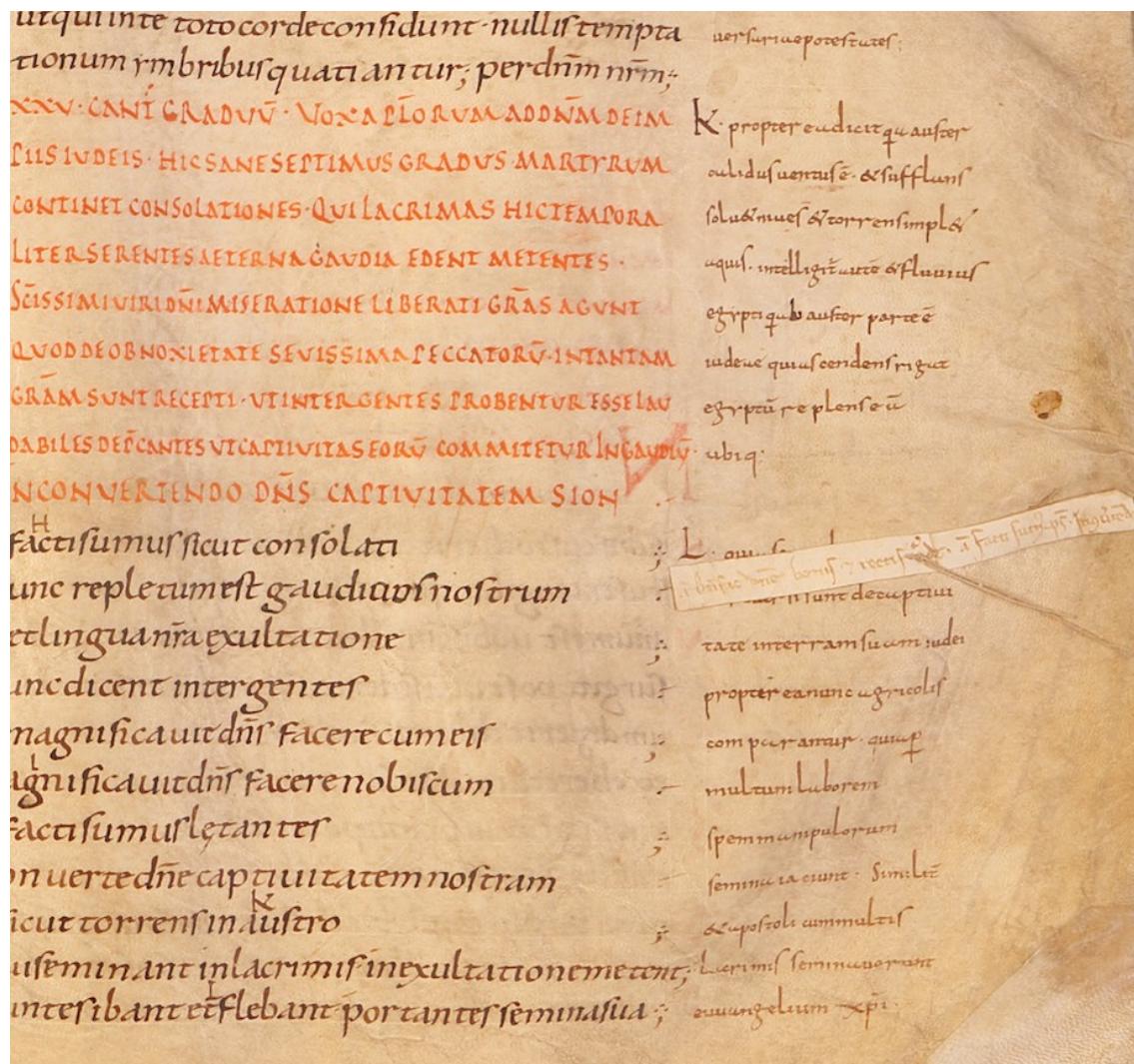


Figure 8: Zoning of cod.LXII Vercelli, Capitular Library

In these cases, since the patch (a sort of a medieval post-it note) represents a new material object, it should be encoded within a **<surface>** of its own, which can be embedded within the main one. This patch-surface carries a series of specialised attributes that allow to specify that the embedded **<surface>** is indeed a patch, that it is attached to the main one via a stitch and that the patch is flipping, i.e., its verso can be seen as well as its recto.

```

<surface type="page">
  <!-- many <zones> here -->
  <surface attachment="stitch" flipping="true" type="patch">
    <!-- reference to the image of the patch, zones, etc. here -->
  </surface>
</surface>

```

In case the editors wish to embed a transcription within the patch, a **<zone>** can be included as well; however, it should be noted that the coordinates of the embedded zone are to be given with respect to the **<surface>** relative to the patch, and not to the one relative to the page. In fact each **<surface>** element “resets” the coordinates to zero, since each **<surface>** represents an independent material object which could be (at least theoretically) be completely separate from any other.

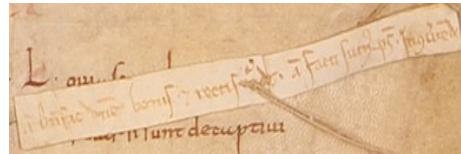


Figure 9: Detail with the patch Vercelli, Capitular Library, cod. LXII

```

<surface type="page">
  <graphic url="../images/BCVCmsLXII-f137r-crop.png" width="918px"
height="850px"/>
  <surface attachment="stitch" flipping="true" type="patch">
    <graphic url="../images/patch.png" width="385px" height="100px"/>
    <zone points="7,65 231,30 369,2 384,34 235,69 13,99">Antiphona:
      Benefac, Domine, bonis et rectis corde. Antiphona: Facti sumus.
      Psalmus: In convertendo</zone>
    </surface>
  </surface>

```

However, the encoding above does not allow to establish where exactly the patch is located within the main **<surface>**; in order to do that, it will be necessary to wrap the patch-surface within a **<zone>** belonging to the parent **<surface>** and which coordinates establishes the exact location of that patch.

```

<surface type="page">
  <graphic url="../images/BCVCmsLXII-f137r-crop.png" width="918px"
height="850px"/>
  <zone type="patch" ulx="529" uly="381" lrx="908" lry="478">
    <surface attachment="stitch" flipping="true" type="patch">
      <graphic url="../images/patch.png" width="385px" height="100px"/>
      <zone points="7,65 231,30 369,2 384,34 235,69 13,99">Antiphona:
        Benefac, Domine, bonis et rectis corde. Antiphona: Facti sumus.
        Psalmus: In convertendo</zone>
    </surface>
  </zone>
</surface>

```

The complexity of this encoding reflects the complexity of unpacking a seemingly trivial phenomenon, namely the insertion of a piece of writing support (paper, parchment or similar) somewhere within another one. This operation, however, needs to be considered from different points of view: the insertion of new text (which could be an addition or a comment to existing text, or even a totally unrelated new material), the act of connecting such new surface to the existing one, the materiality of the added surface, as well as its exact location with respect to the “host” surface.

4.2. Encoding the time

As complex as the previous encoding may seem, an extra layer can be added if the editor wishes to capture not only the location, size and contents of the patch, but also the timing of its addition to the document. In fact, one of the purposes of an edition could also be to describe the process of producing the document or of writing the text, distinguishing also relative timing (i.e., something was written after or before something else) from absolute timing. This kind of preoccupations is mostly characterising the so-called genetic editing scholarly approach, which usually focusses on modern authorial draft manuscripts,¹⁶ but that can be usefully applied also to medieval manuscripts, particularly when studying readership traces, ownerships and provenance, palimpsests and so on. Editors might also want to group together certain phenomena such as a group of interlinear annotation, corrections of various sort, and so on. To do this it is necessary firstly to define and describe the criteria according to which such phenomena are being grouped and then a way to connect the single phenomenon on the surface to its description. The TEI calls such phenomena “changes.”

The first operation, namely the definition and description of changes or groups of changes, is done within the **<teiHeader>**, and more specifically within the element **<creation>**,¹⁷ a child of **<profileDesc>**. The element may contain one or more **<listChange>** element,¹⁸ which in turns contains one or more **<change>** element.¹⁹

Let us consider again the example seen in [Fig. 6](#). The image represents the verso of a flyleaf of ms. 705 of the Klosterneuburg, Augustiner-Chorherrenstift Library, which is a twelfth century manuscripts containing Lives of Saints. The flyleaf has clearly been taken from a pre-existing codex (probably a heavily glossed legal manuscript), which has been dismantled and its parts reused as it was usual for unwanted codices. Traces of the previous life of the flyleaf are well visible on the image: they are namely represented by the writing on the four blocks close to the spine, rotated by 270 arc degrees clockwise with respect to the main orientation of the page. After the insertion of the page as flyleaf, the empty space has been used to provide the newly formed codex with a table of content; to conclude the tormented life of this page, a note has been added on the right hand side of the table of content, rotated by 90 degrees clockwise. This complex history could be formally described as follow:

16. For a comprehensive introduction to *Critique Génétique*, see Grésillon 1994

17. <http://www.tei-c.org/Vault/P5/3.2.0/doc/tei-p5-doc/en/html/ref-creation.html>

18. <http://www.tei-c.org/Vault/P5/3.2.0/doc/tei-p5-doc/en/html/ref-listChange.html>

19. <http://www.tei-c.org/Vault/P5/3.2.0/doc/tei-p5-doc/en/html/ref-change.html>

```

<profileDesc>
  <creation>
    <listChange>
      <change xml:id="ch01" notBefore="0900"
notAfter="1100">Preparation and writing of the manuscript containing a
legal text with glosses.</change>
      <change xml:id="ch02" notBefore="1100"
notAfter="1200">Dismantling of the original manuscript and reuse of
parchment as flyleaves of a codex containing lives of saints; the
reused pages are rotated by 270 arc degrees clockwise and cut into
shapes to fit the format of the new codex</change>
      <change xml:id="ch03" notBefore="1200">Writing of a table of
content</change>
      <change xml:id="ch04">Addition of a note rotated by 90 arc
degrees</change>
    </listChange>
  </creation>
</profileDesc>

```

Here the elements **<change>** are provided with an **@xml:id** attribute, which will then allow to make references to each particular description. The termini ante and post quem are respectively encoded thanks to the **@notAfter** and **@notBefore** attributes.

Once the changes have been declared and described in the **<teiHeader>**, it is possible then to connect the transcription of the zones (as declared in the example below Fig. 6); this can be achieved thanks to the global attribute **@change²⁰** which function is to make a reference to an **@xml:id** declared by a **<change>** element and therefore connecting the data to their metadata. The example below updates the code seen above with the references to the changes declared in the previous example.

```

<surface>
  <zone ulx="72" uly="57" lrx="484" lry="366" change="#ch03"/>
  <zone ulx="461" uly="38" lrx="509" lry="179" rotate="90"
change="#ch04"/>
  <zone ulx="511" uly="71" lrx="617" lry="210" rotate="270"
change="#ch01"/>
  <zone ulx="504" uly="218" lrx="616" lry="383" rotate="270"
change="#ch01"/>
  <zone ulx="496" uly="403" lrx="613" lry="569" rotate="270"
change="#ch01"/>
  <zone ulx="517" uly="585" lrx="619" lry="633" rotate="270"
change="#ch01"/>
</surface>

```

20. <http://www.tei-c.org/Vault/P5/3.2.0/doc/tei-p5-doc/en/html/ref-att.global.change.html>

The order of the **<change>** elements does not need to reflect the order of the **<zone>** (or any other element) carrying the **@change** attribute; furthermore, referring many times to the same **<change>** has the result of grouping together a set of zones or other elements.

The elements **<listChange>** can also nest or be repeated to present sub-ordering of **<change>**s or different types of grouping; in case the **<listChange>** is presenting a non-chronological list, it is recommended to use it with the attribute **@ordered** carrying the value of “false.”

The use of facsimiles in editing can assume different forms and different roles, which reflects different editorial theories and practices. It is certain that, whatever the reasons and the theoretical motivations behind the use of a digital facsimile within an editions, images of primary sources have had (and are likely to have in future) a major impact in the way we edit and access texts from the past. The TEI offers a solid way to handle facsimile and the different ways they can be used within an edition, but it is not by any means the only way nor it is perfect. It is nevertheless clear that the level of theoretical reflection and the modelisation that lies behind the handling of facsimiles in TEI is among the most advanced and in-depth analysis available to editors and textual scholars.²¹ Document-centric editing has taken a central role in textual scholarship, an evolution that is the result of a major technological improvement and of a modification in editorial theories; whether this is matter of a coincidence or that the change in textual scholarship is determined by the development of digital photography is not completely clear, but it seems reasonable to assume that this change will not be of short duration, and that new scholarly achievements will follow as a consequences.

Bibliography

- Grésillon, Almuth 1994. *Eléments de critique génétique: lire les manuscrits modernes*, Paris: Presses universitaires de France
- Kiernan, Kevin 2006. “Digital Facsimile in Editing.” In *Electronic textual editing*, Lou Burnard, Katherine O’Brien O’Keeffe and John Unsworth 262–268 New York: Modern Language Association of America http://www.tei-c.org/About/Archive_new/ETE/Preview/kiernan.xml
- Nell Smith, Martha 2004. “Electronic Scholarly Editing.” In *A Companion to Digital Humanities*, Susan Schreibman, Ray Siemens and John Unsworth ?? Oxford: Blackwell <http://www.digitalhumanities.org/companion/>
- Pierazzo, Elena 2011. “A rationale of digital documentary editions.” In *Literary and Linguistic Computing* 463–477 4 4. Doi: 10.1093/lrc/fqr033

21. The TEI is updated twice a year by a large community of scholars, and therefore one can reasonably hope that possible drawbacks will be fixed in future (especially if the editors nothing them will take care to send an appropriate feedback to the TEI Council).

- 2014. “Digital Documentary Editions and the Others.” In *Scholarly Editing*, 35
<http://www.scholarlyediting.org/2014/essays/essay.pierazzo.html>
- 2015. *Digital Scholarly Editing: Theories, Models and Methods*, Aldershot: Ashgate
- 2016. “Modelling Digital Scholarly Editing: From Plato to Heraclitus.” In *Digital Scholarly Editing: Theories and Practices*, Matthew James Driscoll and Elena Pierazzo, 41–58 Cambridge: Open Book Publisher
<https://www.openbookpublishers.com/product.php/483/digital-scholarly-editing--theories-and-practices>
- Pierazzo, Elena and Stokes, Peter A. 2010. “Putting the Text Back into Context: a Codicological Approach to Manuscript Transcription.” In *Kodikologie und Paläographie im Digitalen Zeitalter 2 - Codicology and palaeography in the digital age* 2, Franz Fischer, Christiane Fritze and Georg Voegler, Norderstedt: Books on Demand
- Robinson, Peter M.W. 2013. “Toward a Theory of Digital Editions.” In *Variants* 105–132 10
- Stokes, Peter A. 2010, “Scripts.” In *Handbook of Medieval Studies: Terms, Methods, Trends*, A. Classen, 1217–33 2 Berlin: A. Classen.
- Sutherland, Kathryn and Pierazzo, Elena 2012. “The Author’s Hand: from Page to Screen.” In *Collaborative Research in the Digital Humanities*, Marilyn Deegan and Willard McCarty, 191–212 Aldershot: Ashgate
- TEI Consortium 2017. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 3.3.0, last updated on 10th July 2017 <http://www.tei-c.org/Guidelines/P5/>
- Treharne, Eleine 2013. “Fleshing out the TEXT: the Transcendent Manuscript in the Digital Age.” In *Postmedieval*, 1–16 4 4
- Twycross, Meg 2008. “Virtual Restoration and Manuscript Archeology.” In *The Virtual Representation of the Past*, Mark Greengrass and Lorna Hughes, 23–47 Aldershot: Ashgate

Citations and References

Marjorie Burghart

In this chapter you will learn:

- to encode the sources used or referred to in a text, and link them to bibliographic citations;
- to encode the biblical references, and represent them as canonical references;
- to handle overlapping hierarchies if they occur.

Identifying the sources of a text is a key part of the editing process. Sources, in this context, are passages of the edited text which are quoted explicitly or implicitly from other works, or simple allusions to or reminiscences of other works. Their identification by the editor is very important to reconstruct the intellectual environment of the work at hand, the influences, and the methods used by the author, leaning on the authority of famous authors or compiling unattributed quotations. These sources can belong in various categories or styles: sometimes they are quoted silently, i.e., without any attribution, sometimes the author gives their reference (attributing them to an author and/or work), which may or may not be correct. Other categories, defined by the editor, may also be useful in order to categorise the sources: classical, patristic and contemporary sources for instance. Canonical references, like biblical quotations usually form a distinct group and need to be displayed and processed in a specific fashion.

The TEI offers a general frame for the encoding of quotations, within which you are free to develop your own strategies. The solutions suggested here are only possibilities, among many, that are offered to you with the needs of scholarly editors in mind. You are invited to consult the relevant sections of the Guidelines¹ to further your knowledge of the various possibilities.

1. Defining your needs

The first step consists in a reflection upon your needs and what you want to achieve with your edition.

If you want to display or process the quoted bits according to their category, you will need to attach a different type to each quoted source. This will be useful, for instance, if you want to be able make a separate index or group of critical notes for the sources quoted with an attribution and those quoted silently, or if you wish to display biblical quotations in italics and other sources

1. See in particular chapter [3.3.3, Quotations](#)

between quotes. Attributing a type to the sources will also let you process the edition in a more detailed way. You could for instance count the words in order to determine what proportion of your edition each type of quoted source represents, or you could also map the use of a certain category of sources across a corpus.

If you plan to have a lot of source identifications, it may be useful to consider using a mechanism that lets you encode a link between a citation and a full bibliography (in the case of a literary source), or formally represent a canonical reference (in the case of biblical quotations). It will prevent the addition of redundant information.

The general mechanism to encode a quotation together with its reference relies upon the **<cit>** element (cited quotation), which “contains a quotation from some other document, together with a bibliographic reference to its source” ([Guidelines](#)). The element **<quote>** contains the quotation, and the element **<bibl>** the bibliographic reference.

2. A basic example

Let us start with a basic case: imagine you find this source quoted in your edition, ascribed to “Bernardus”:

Bernardus : Auferatur malus ne generet malos. Non potest arbor mala fructus nisi malos facere.²

2.1. Minimal encoding

If you were interested only in distinguishing quotations from the rest of the text (for statistical purpose for instance), without attempting to give a bibliographic reference or identification of the source, you could simply use the **<quote>** element:

Bernardus : **<quote>**Auferatur malus ne generet malos. Non potest arbor mala fructus nisi malos facere.**</quote>**

But in most cases, the point of marking up quotations in an edition is precisely to give some form of identification for each of them. To this purpose, we will use the **<bibl>** element to give the bibliographic reference, and wrap together the **<quote>** and **<bibl>** in a **<cit>** element. You could encode this quotation as follows, using the **<bibl>** element to encode the ascription given in the text (which may or may not be accurate):

```
<cit>
  <bibl>Bernardus :</bibl>
  <quote>Auferatur malus ne generet malos. Non potest arbor mala
fructus nisi
    malos facere.</quote>
</cit>
```

2. This Latin quotation translates as follows “Bernard : He who is bad must be removed so he does not generate other bad ones. The bad tree cannot bear fruits, except bad ones.”

This is a perfectly valid encoding, but in the following examples, we will use **<bibl>** to encode the bibliographic reference that we, the critical editor, have identified as the source. In this case, the data contained by **<bibl>** is not part of the text of the edition, but of its critical apparatus, and should be displayed and processed accordingly. In any case, documenting your usage of the element in the header is useful.³

You have identified the source as a letter by Bernard of Clairvaux, numbered 102 in the reference edition of his letters. You may encode the quoted text together with the reference you have found like this:

```
Bernardus : <cit>
  <quote>Auferatur malus ne generet malos. Non potest arbor mala
fructus nisi malos facere.</quote>
  <bibl>Bernardus Claraevallensis, Epistolae, 102</bibl>
</cit>
```

If you want a more precise encoding, you can enhance the **<bibl>** element from the previous example:

```
<bibl><author>Bernardus Claraevallensis</author>,
<title>Epistolae</title>, 102</bibl>
```

This will allow you to process the different parts of the bibliographic reference according to your needs, for instance to display the author's name in small capitals and the title in italics.

2.2. Distinguishing between different categories of citations

You might want to give a type to this source. Since the author of the text has quoted it with an ascription, you might decide to add a **@type** attribute to **<cit>**, with the value "**ascribed**", for instance. The TEI does not have specific recommendations for the possible value of **@type** here, you are free to use your own categories according to your needs.

```
Bernardus : <cit type="ascribed"> ... </cit>
```

If you wish to specify different values in **@type**, just separate them with a simple space. For instance, to express that this source is quoted with an attribution and also is a literal quotation (as opposed to a rephrasing or a mere allusion), you could use the following:

```
Bernardus : <cit type="ascribed literal"> ... </cit>
```

3. Linking sources to a bibliography

In the previous example, all the bibliographic information was contained in the **<bibl>** element. This is inconvenient when the same source is quoted multiple times in an edition,

3. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/HD.html#HD57-2>

because it forces you to repeat redundant information, at the risk of being inconsistent in your bibliographic citations of the same work.

Let's consider the example we used above. So far we have just given a short, incomplete reference, but we would need to link to the full reference edition of the letters of Bernard of Clairvaux, which is the following:

Bernardus Claraevallensis, *Epistolae*, in *Sancti Bernardi Opera*, Jean Leclercq and Henri Rochais eds., Rome, 1957-1977, vol. VII and VIII.

First, anywhere in the document, let us create a detailed description for each bibliographic citation we want to link to. You could put this bibliography in a **<front>** or **<back>** element, for instance, to clearly separate it from the **<body>** of your edition.

We offer below an example of bibliographic citation encoding. You could opt for less or more detailed options: the TEI offers many options to encode a bibliography, which go beyond the scope of this chapter. If you are interested in this topic, we recommend that you check the relevant sections of the Guidelines.

The important point here is that each bibliographic citation we want to be able to link to must have an @xml:id attribute, with a unique value.

```
<listBibl>
  <bibl xml:id="bernEpist"><author>Bernardus
Claraevallensis</author>, <title>Epistolae</title>, in <title>Sancti
Bernardi Opera</title>, <editor>Jean Leclercq</editor> and
<editor>Henri Rochais</editor> eds., <pubPlace>Rome</pubPlace>,
<date>1957-1977</date>, <biblScope unit="volume">vol. VII and
VIII</biblScope>.</bibl>
  <bibl xml:id="gregMor"><author>Gregorius Magnus</author>,
<title>Moralia in Iob</title> (etc.) </bibl>
</listBibl>
```

Within the body of the edition, when we want to link a source to a bibliographic citation, we can proceed as follows, using the **<ref>** element with its **@target** attribute to wrap the **<bibl>** description:

```
Bernardus : <cit>
  <quote>Auferatur malus ne generet malos. Non potest arbor mala
fructus nisi malos facere.</quote>
  <ref target="#bernEpist"><bibl><author>Bern.</author>,
<title>Epist.</title>, 102 (VIII, 257-8)</bibl></ref>
</cit>
```

This encoding is very similar to the basic example above, except that thanks to the **<ref>** element we now have encoded a link to a full description of the bibliographic citation where the edition of this letter appears, while the **<bibl>** let's us give the details for this particular quotation (which volume, and which pages).

When displaying or processing your edition, you will be able, for instance, to create an index of the works cited, to count or highlight the references to a particular bibliographic citation, or to display bibliographic notes in full or short version.

4. Canonical references: the example of biblical quotations

Canonical references are “any means of pointing into documents, specific to a community or corpus”⁴ They are identified not by bibliographic citation, like most literary sources, but by a short reference following rules defined by a scholarly community. That is the case of many religious texts (the Mishnah, the Bible, the Quran), but not only: some classical works have well-established systems of canonical references (Aristotle, for instance, is often quoted by a “Bekker number” in modern literature).⁵ A very common usage of canonical references in critical editions is the identification of quotations from or references to religious texts. It is common to treat those quotations slightly differently from other source quotations: they are generally displayed differently from other sources, have their own searchable index, and in print editions they usually have their own series of footnotes. In this section, we are going to use the example of biblical quotations to illustrate canonical references: with the adaptation of the reference scheme, these principles are directly applicable to other types of canonical references.

4.1. Particularities of biblical quotations

Biblical quotations are identified by canonical references, which take the form of:

- a book name or abbreviation thereof;
- a chapter number (optional);
- a verse number, or a range or series of verses (optional).

It might seem very simple: if I give the reference “Gn 1:1,” my readers will understand that this means Genesis, chapter 1, verse 1. But there are many potential difficulties, which may also apply to other canonical references. Some are rather trivial and pertain to the formatting of the references:

- there is no universal list of book names or abbreviations: they will differ from one language to another, and also from one editor to another, etc. For Genesis, instead of “Gn” I could have used “Gen,” “Genes,” the French name “Genèse,” the Italian name “Genesi,” etc. The names or abbreviations you are using should therefore be stated explicitly in your edition.
- the separators between the data may vary: some will use a colon to separate the chapter and verse numbers, others a comma followed by a space, etc.

Another difficulty is linked with the version of the Bible that is quoted: despite the canonical

4. <http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/SA.html#SACR>

5. For an introduction to canonical references, their issues and the solutions to handle them, see Kalvesmaki 2014.

character of the Bible, there are several translations, even for a single language. In a single edition, the Bible may be quoted after different translations, for instance the Vulgate and the Vetus Latina (two different Latin versions), or the Vulgate (Latin) and the Douay-Rheims version (English translation from the Vulgate).

4.2. Basic example

The encoding of biblical quotations bears many similarities with the one of literary sources. Let's consider this example:

Unde in Genesim: In principio creavit Deus celum et terram.

A basic encoding could be the same as the one used for literary sources in our first example:

```
Unde in Genesim: <cit>
  <quote>In principio creavit Deus celum et terram.</quote>
  <bibl>Gn 1:1</bibl>
</cit>
```

This is simple to implement, but also has limited possibilities: it will only let you display a note with the reference of the quotation.

Another drawback is that, as it is, it does not distinguish biblical quotations from other quotations. You can achieve that simply by adding a `@type` attribute to `<cit>`, with the value `"bible"` for instance:

```
Unde in Genesim: <cit type="bible"> ... </cit>
```

If you need more categories, you may add different values to `@type`, separated with a space, as we saw above.

But the identification of a biblical quotation is not really a bibliographic citation. More accurately, we could encode it as a reference, using the `<ref>` element as an alternative to `<bibl>`. According to the Guidelines, “`<ref>` (reference) defines a reference to another location, possibly modified by additional text or comment.” A minimal example would be the following:

```
Unde in Genesim: <cit type="bible">
  <quote>In principio creavit Deus celum et terram.</quote>
  <ref>Gn 1:1</ref>
</cit>
```

4.3. Going further with `<ref>`

The `<ref>` element may bear two optional and mutually exclusive attributes: the `@target`, which “specifies the destination of the reference by supplying one or more URI References,” and `@cRef` (canonical reference) attribute, which “specifies the destination of the pointer by supplying a canonical reference expressed using the scheme defined in a `<refsDecl>` element

in the TEI header.” For biblical quotations, `@cRef` is the best suited. To use it fully, we need to proceed in two stages.

First, within the edition, we are going to encode the biblical quotations as above, but we will now use the `@cRef` attribute to contain the canonical reference, which you must always construct following the same rules. Here I have opted for a book abbreviation followed by a space, then the chapter and verse numbers separated with a colon:

```
Unde in Genesim: <cit type="bible">
  <quote>In principio creavit Deus celum et terram.</quote>
  <ref cRef="Gn 1:1"/>
</cit>
```

Should you want to add precisions or comments about this reference, you are free to add them:

```
<ref cRef="Gn 1:1">This is a literal quotation of the first verse of
the Bible</ref>
```

Then, we must declare in the `<teiHeader>` how the canonical references in `@cRef` are constructed. To do so, we must add a `<refsDecl>` (references declaration) to the `<encodingDesc>` part of the header.

For our purpose, there are two way to explain your referencing scheme in the `<refsDecl>` element:

- either you simply explain in prose how you construct your references, giving all necessary information in plain text, as in the following example:

```
<refsDecl xml:id="biblicalCitations">
  <p>The <att>cRef</att> attribute of <gi>ref</gi> elements
  citing the Bible contain a canonical reference in one to three
  parts. The first part is an abbreviation for the name of the
  relevant book of the Bible, and it is the only mandatory one. The
  second part, preceded by a space, is a chapter number. The third,
  preceded by a colon, is either a single verse number, or a series
  of consecutive verses with the number of the first verse followed
  by a hyphen, followed by the number of the last verse. </p>
  <p>The following list of abbreviations has been used for the
  books of the Bible: Gn: Genesis; Exodus: Ex; Leviticus: Lv;
  [etc.]</p>
</refsDecl>
```

This information will be very useful to future users (or even current collaborators) working on your edition, but it cannot be processed automatically.

- or you define the pattern(s) that your canonical references must match and how they must be transformed into a valid URI, using regular expressions.⁶ You can define several

6. Regular expressions are an extremely useful tool, available in many computing languages. They allow the

patterns, each in a `<cRefPattern>` element, using the two attributes `@matchPattern` (to define how canonical references are constructed), and `@replacementPattern` (to define how the different parts of the matching references must be transformed).

In the following example, we define a single rule: it matches canonical references (i.e., the values of `@cRef` attributes) composed of a string of any number of characters followed by a space, followed by another string, followed by a colon, and followed by a string. Then it transforms the parts identified into a URL pointing to an online edition of the Vulgate:

```
<refsDecl xml:id="biblicalCitations">
  <cRefPattern matchPattern="(.) (.+):(.+)"
    replacementPattern="http://vulsearch.sourceforge.net/html/
    $1.html#${$2}_{$3}">
    <p>This pointer pattern extracts and references the
    <q>book,</q> <q>chapter,</q> and <q>verse</q> parts of a biblical
    reference pointing to a single verse, like "Gn 1:1", and
    reconstructs a link to an online version of the biblical
    text.</p>
    <p>The following list of abbreviations has been used for the
    books of the Bible: Gn: Genesis; Exodus: Ex; Leviticus: Lv;
    [etc.]</p>
  </cRefPattern>
</refsDecl>
```

Here is how the regular expression works: in `@matchPattern`, each expression between parenthesis is a match. If the `@cRef` value "`Gn 1:1`" is processed, for instance, the first expression will retrieve the value "`Gn`", because it is a string of characters at the beginning of the reference and before the first separator, the space; the second will retrieve the value "`1`", as well as the third, because they are strings of characters respectively between the space and colon separators, and after the colon separator.

In `@replacementPattern`, each expression composed of a dollar sign and a number means it must be replaced with the value of the matched expression bearing this number. With the same example of "`Gn 1:1`", `$1` should therefore be replaced with the first match "`Gn`", `$2` with the second match "`1`" and `$3` with the third and last match "`1`", and the following URL would be constructed:

`http://vulsearch.sourceforge.net/html/Gn.html#x1_1`

You could add more `<cRefPattern>` elements defining more rules, for instance when a canonical reference points to a full chapter (Gn 1), or a full book (Gn).

definition of patterns for search and replace operations. A full introduction to regular expressions is beyond the scope of this chapter, but we highly recommend scholars interested in advanced search-and-replace operations to read further on this topic.

Note that there might be more than one `<refsDecl>` in a single file, so it may be useful to give it an `@xml:id` (unique identifier). If there are several `<refsDecl>`, you can specify in each `<ref>` which one applies with the attribute `@decls`.

```
<ref cRef="Gn 1:1" decls="#biblicalCitations">Gen 1, 1</ref>
```

Whatever solution you opt for, well-constructed canonical references will allow you to process your edition in a meaningful and useful way: extracting all quotations from a particular book or chapter of the Bible for instance, or building an index of biblical citations, etc.

5. Allusions without actual quotation

For literary as well as biblical sources, it may happen that a reference is made to a text without actually quoting it. For instance, here we have a reference to an evangelical parable, which appears in Luke, chapter 10, verses 25–37, but there is no quotation from the parable:

(...) as it is demonstrated by the Parable
of the Good Samaritan (...)

Similarly, here we have an allusion to a text by saint Augustine, where he laments on his youthful sins, among which the theft of pears in vain. This comes from the Confessions, book II, chapter 4, but again there is only a reference and no quotation:

(...) Augustine says he stole pears in a
garden when he was a young man (...)

Since the source is not actually quoted, we cannot use the `<quote>` element, but the `<bibl>` or `<ref>` elements are still relevant.

... as it is demonstrated by the Parable
of the Good Samaritan`<ref cRef="Lc 10:25-37"/>`...

Augustine says he stole pears in a
garden when he was a young man`<ref target="#AugConf"><bibl type="source">Aug., <title>Confessiones</title>, II, 4</bibl></ref>`

You could also choose, perhaps for considerations linked to the processing of the edition, to still wrap those `<ref>` and `<bibl>` elements in a `<cit>`, although it is not necessary.

6. Handling overlaps

It may happen that quotations, which are often numerous, overlap with other TEI elements in an edition - typically `<app>` elements dedicated to recording the textual variants. Overlapping or non-nesting information is an issue with all XML-based languages, and the TEI is not an exception. The impossibility for elements to overlap certainly poses a problem, but the TEI has various mechanisms to overcome this difficulty.

Let us consider the following example: we see there are two biblical quotations, the first from Ephesians, 5, 30 and the second from Luke, 24, 39:

And as the Apostle says in Eph. V: We are members of his body, of his flesh, and of his bones. And also in Luke XXIIII: A spirit hath not flesh and bones, as you see me to have.

But we have already encoded the variants from three witnesses A, B and C. As it happens, witness C presents a particular type of omission, caused by homeoteleuton: as two lines had similar endings, (here, “flesh”), the scribe’s eye slipped and he missed several words, right across the two citations:

And as the Apostle says in Eph. V: We are members of his body, of his
<app>
 <lem>flesh, and of his bones. And also in Luke XXIIII: A spirit
hath not**</lem>**
 <rdg wit="#C" cause="homeoteleuton"/>
</app> flesh and bones, as you see me to have.

The TEI Guidelines support four XML-based methods for handling overlapping information, which are exposed in details.⁷ Each method has its pros and cons, but for our purpose we recommend using in most cases the Fragmentation and Reconstitution of Virtual Elements method.

6.1. Fragmentation and Reconstitution of Virtual Elements method

This method consists in choosing a privileged hierarchy, and breaking up overlapping elements from the other hierarchy in smaller elements, connected between them but not overlapping anymore.

With our example, here is how we could solve the overlapping problem with this method, while privileging the textual variance hierarchy (**<app>** etc.): we would split each overlapping **<cit>** element into two smaller elements (or more if it was necessary) that do not overlap. To help with the virtual reconstruction of the full **<cit>** elements, we have given each split part of the original **<cit>** a unique identifier (**@xml:id**), and used the attributes **@prev** and **@next** to point to the previous and following part of the citation, respectively (this is called a virtual join, or a chain). We put the **<ref>** element, which does not overlap, only in the final part of the citation:

And as the Apostle says in Eph. V:

<cit xml:id="cit01a" next="#cit01b" type="bible"><quote>We are
members of his body, of his **</quote></cit><app>**
 <lem><cit xml:id="cit01b" prev="#cit01a" type="bible"><quote>flesh,
and of his bones**</quote><ref cRef="Eph 5:30"/></cit>**. And also in Luke
XXIIII:

```

<cit xml:id="cit02a" next="#cit02b" type="bible"><quote>A spirit
hath not</quote></cit></lem>
<rdg wit="#C" cause="homeoteleuton"/>
</app>
<cit xml:id="cit02b" prev="#cit02a" type="bible"><quote>flesh and
bones, as you see me to have</quote><ref cRef="Lc 24:39"/></cit>.

```

This method is a bit verbose, but easy enough to encode and process.

One thing you must remember when processing a document created with this method is that there are more **<cit>** elements than there are actual citations. Therefore, if you want to know, for instance, how many citations there are in your edition, counting the **<cit>** elements will be misleading. With this example, you would find a result of 4, while there are only 2 biblical citations, each split in two.

This is easily solved by preparing a more precise query, for instance you could add up the number of **<cit>** elements that do not have a **@prev** nor a **@next** attribute, and the number of **<cit>** elements with a **@next** but no **@prev** attribute (i.e., the first **<cit>** elements of a chain).

Nota bene: in the example above, we have chosen to privilege the textual variants hierarchy of the sources hierarchy, but we could do the opposite if it was more suited to our edition for instance if processing the sources were our top priority, and textual variance only accessory. The exact same principle would apply: we would break up the apparatus entries into smaller elements nesting in the **<cit>** hierarchy, and virtually join the split elements.

6.2. Alternative methods

The TEI also supports three other XML methods to handle overlapping hierarchies. Depending on your data, your goal and your familiarity with computer processing of the result, they may also be a good choice for your edition.

6.2.1. Multiple Encodings of the Same Information

Multiple Encodings of the Same Information is technically the simplest method, which consists in re-encoding the same information as many times as you have overlapping hierarchies. In practical terms, it means that we would need to have one text of the edition with the variants encoded, and another text of the same edition with the sources encoded. This is therefore very redundant, and can be difficult to maintain if your text is not stable: should you spot a typo or an error in your transcription at some point, you will have to correct it consistently across all the different encodings of this text.

With our example, it would result in the following: a first text with the textual variance encoded, without information on the quotations:

And as the Apostle says in Eph. V: We are members of his body, of his
<app>
<lem>flesh, and of his bones. And also in Luke XXIIII: A spirit

```

hath not</lem>
  <rdg wit="#C" cause="homeoteleuton"/>
</app> flesh and bones, as you see me to have.

```

And a second text with the quotations encoded, without information on the textual variance, which is not without problems in the case of a critical edition. Here we have a clearly identified lemma vs. an erratic manuscript, which makes things easier. But if we had more subtle variants, it would be much more difficult to determine what is the text:

And as the Apostle says in Eph. V:

```

<cit type="bible">
  <quote>We are members of his body, of his flesh, and of his
bones.</quote>
  <ref cRef="Eph 5:30"/>
</cit> And also in Luke XXIIII: <cit type="bible">
  <quote>A spirit hath not flesh and bones, as you see me to
have.</quote>
  <ref cRef="Lc 24:39"/>
</cit>

```

6.2.2. Boundary Marking with Empty Elements

With this method, the start and end points of the problematic (non-nesting) elements are marked with empty tags, like **<anchor>**. Although efficient, this method poses various problems. It can get confusing for the encoders and lead them to introduce errors that will be difficult to screen. The processing stage will also be difficult: no off-the-shelf application will be able to process your data satisfactorily, and you will need to write complex transformation programs capable of handling this type of encoding.

Using again our example, here is what the encoding would look like with this method. Let's say that we are going to privilege the textual variance hierarchy over the identification of sources. It means that we are going to keep classic elements for textual variance (**<app>** etc.), and represent only the start and end points of overlapping source identifications (**<cit>** and **<quote>**) with **<anchor>**. The result is cumbersome and poorly readable, on top of being difficult to encode and process:

And as the Apostle says in Eph. V:

```

<anchor type="delimiter" subtype="citBibleStart"/>We are members of
his body, of his <app>
  <lem>flesh, and of his bones <ref cRef="Eph 5:30"/> <anchor
type="delimiter" subtype="citBibleEnd"/>. And also in Luke
XXIIII: <anchor type="delimiter" subtype="citBibleStart"/>A spirit
hath not</lem>
  <rdg wit="#C" cause="homeoteleuton"/>
</app> flesh and bones, as you see me to have <ref cRef="Lc
24:39"/><anchor type="delimiter" subtype="citBibleEnd"/>.

```

6.2.3. Stand-off Markup

Stand-off Markup offers many possibilities, but it is the most complicated to encode and process. With this method, it is possible to annotate a file without writing in it: for instance, you could annotate a read-only edition made available online by a library, even if it is only a plain-text file. When you are annotating your own file, you can choose to put the stand-off markup in the same file, or in another. It is a very good method, offering many advantages for collaboration (many people can add annotations at the same time, without the risk of overlapping), and for texts with multiple, complex and overlapping layers of annotations. But it is very difficult to process, and often to encode. This is why we recommend using this method only when this is the only solution to represent satisfactorily your data.

Using this method, we could encode our example like this: first, we would need elements with a unique identifier to mark the boundaries of the bits of texts we want to annotate. We do not always need to add new elements for that, but in this case there is nothing at this place in our encoding, so we are going to use empty elements (`<anchor>`), at the beginning and end of each citation.

And as the Apostle says in Eph. V: `<anchor xml:id="cit01"/>`We are members of his body, of his `<app>`
`<lem>`flesh, and of his bones`<anchor xml:id="cit02"/>`. And also in Luke XXIIII: `<anchor xml:id="cit03"/>`A spirit hath not`</lem>`
`<rdg wit="#C" cause="homeoteleuton"/>`
`</app>` flesh and bones, as you see me to have`<anchor xml:id="cit04"/>`.

Somewhere else in the same document (or even in a separate file), we would reconstruct the desired encoding for the annotations:

```
<cit type="bible">
  <quote>
    <xi:include xpointer="range(element(cit01),element(cit02))"/>
  </quote>
  <ref cRef="Eph 5:30"/>
</cit>
<cit type="bible">
  <quote>
    <xi:include xpointer="range(element(cit03),element(cit04))"/>
  </quote>
  <ref cRef="Lc 24:39"/>
</cit>
```

Bibliography

Joel Kalvesmaki, “Canonical References in Electronic Texts: Rationale and Best Practices,” in *Digital Humanities Quarterly*, Volume 8, Number 2, 2014
[<http://www.digitalhumanities.org/dhq/vol/8/2/000181/000181.html>]