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 oviously 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:

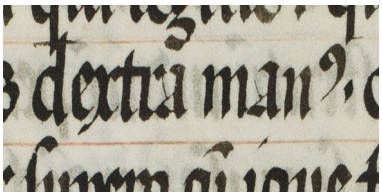

Figure 1: Klosterneuburg, Stiftsbibliothek, CCl 222, fol. 57v

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 man^{9}," which translates as "right hand." The last sign at the end of the first word is and 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 "man□" 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/, 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 “man⁹” or “manus,” but also all sorts of variations highlighting the expanded letters: “man*us*,” “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 orginal 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, **<del>** (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:

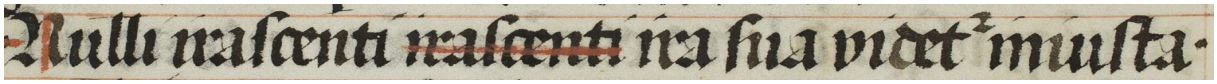

Figure 2: Klosterneuburg, Stiftsbibliothek, CCl 222, fol. 46v

This Latin line transcribes as follows: “Nulli irascenti ~~irascenti~~ 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 <del>irascenti</del> 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 expuctuate 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 **<del>**. 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</del> 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

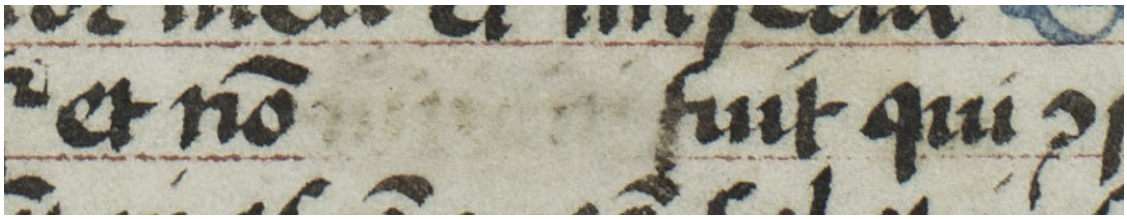

Figure 3: Klosterneuburg, Stiftsbibliothek, CCl 1195, f. 31v

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 evangelio 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 <del>ut dicitur in euangelioIohannis:</del> <quote><del>In
principio erat Verbum, et</del> 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 **<del>**.

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."

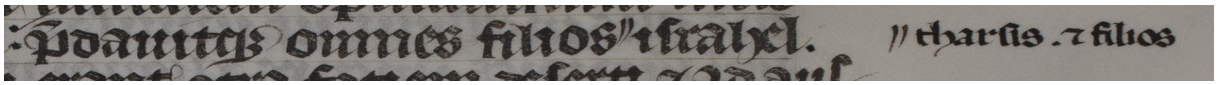

Figure 5: Klosterneuburg, Stiftsbibliothek, CCl 3, fol. 20r

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 addiction 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 concommitent 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 **<del>** and an **<add>**. In the following example, the word "dominus" (the Lord) has been deleted by expunctuation, and the word "diabolus" (the Devil) has been added above the line to replace it.

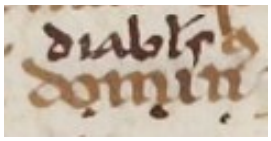

Figure 6: Paris, Bibliothèque nationale de France, Latin 588, f. 8r

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 **<del>** and **<add>**:

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

1.2.4. Transpositions

In a transcription, a transposition "occurs when metamarks 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 twelth 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>**, **<del>** 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 attributres **@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.

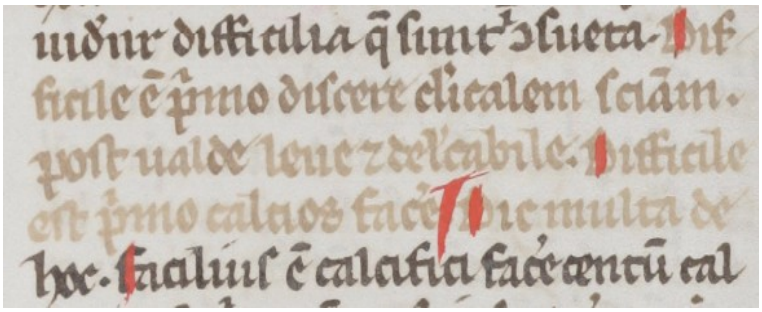

Figure 7: Fribourg, Couvent des Cordeliers / Franziskanerkloster, Ms. 117 I, f. 153r

- 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.

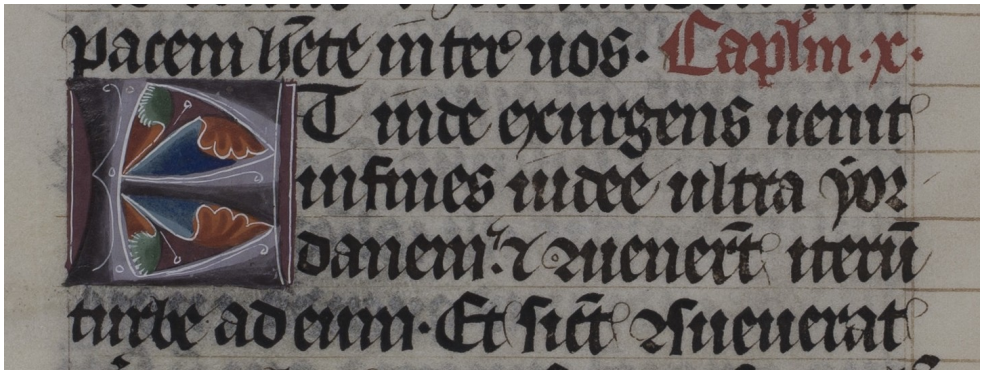

Figure 8: Klosterneuburg, Stiftsbibliothek, CCl 3, fol. 219v

```
  <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 conuenerunt iterum turbe ad eum, et sicut
consueverat ... </p>
  </div>
```

The whole chapter can be encoded as a ** **, 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>tragical</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 an other 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 attributed 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>** element[23] 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 Κύριε 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">Κύριε 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 ** ** 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 though the course of their long history: tear and wear, vandalism, fire, water, mildew... Many agressions 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 **<del>**, there is a mechanism to allow the encoding of damage spanning over different hierarchies. I you need to use it, check the description of the **<damageSpan>** element in the Guidelines,[28] ans see the explanation of the similar **<delSpan>** mechanism at the end of section Deletions

- **<del>** (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 delection 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

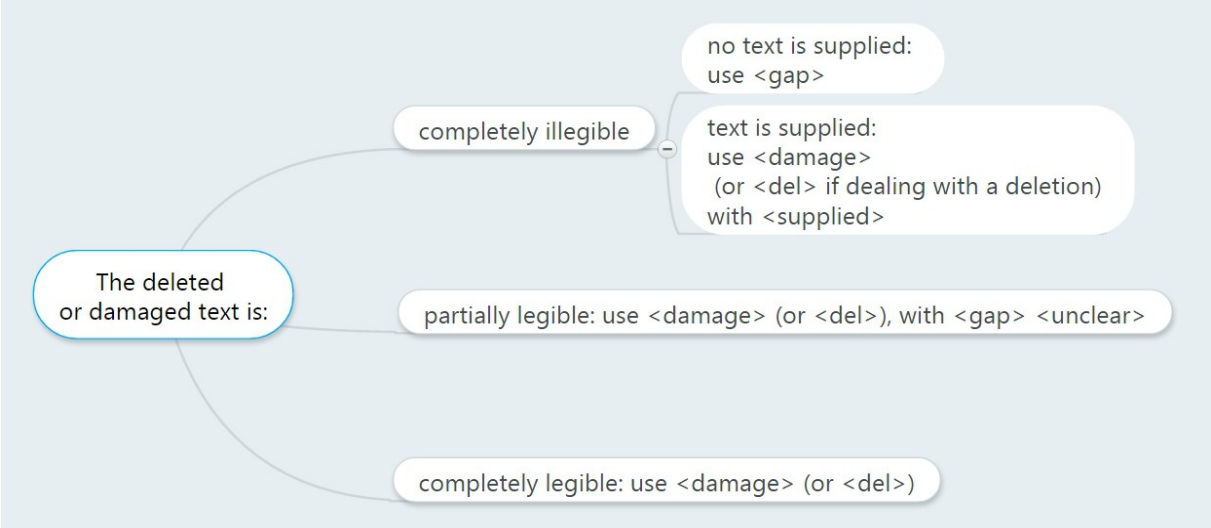

Figure 9: Decision Tree: How to Encode Deleted or Damaged Text

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 **<del>** respectively, depending on the situation. Within **<damage>** or **<del>**, 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:

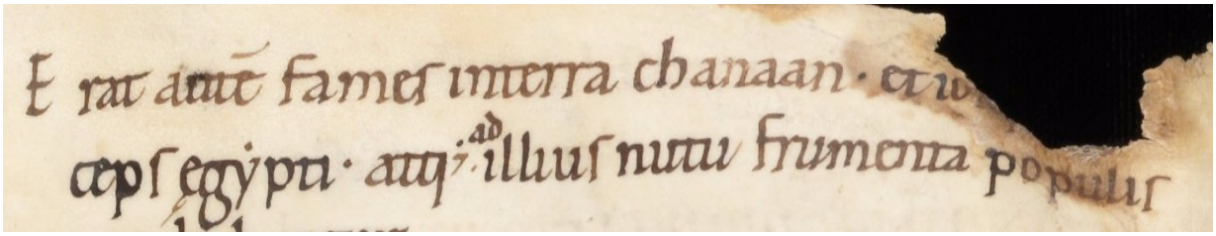

Figure 11: Sion/Sitten, Archives du Chapitre/Kapitelsarchiv, Ms. 15 – Giant Bible, f. 17r (www.e-codices.unifr.ch)

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 attributed dedicated to describe the dimensions of the damaged area[36] Since it is a well-know 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 **<del>** 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):

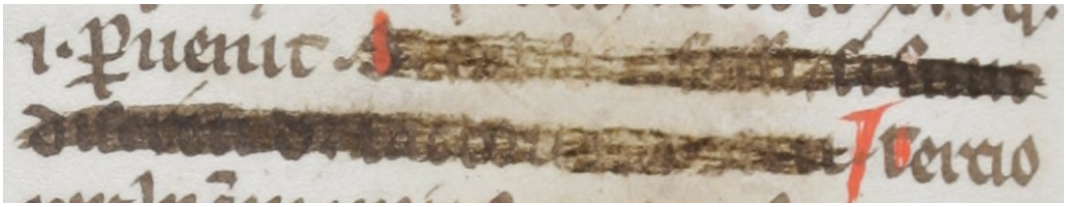

Figure 12: Fribourg, Couvent des Cordeliers/Franziskanerkloster, Ms. 117 I – Berthold of Regensburg, Sermones, f. 117v (www.e-codices.unifr.ch)

To encode this passage, we can combine within a **<del>** 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 **<del>** elements and their attributes, without additional tags.