1 Introduction

This working paper reviews the requirements for document structuring identified by Memoria Deliverable 1.1. *Reading Activity: objects, functionalities for reading process, and impact on the holding structure* and proposes a set of SGML elements which might be used to represent them. The set proposed constitutes a particular application of the Text Encoding Initiative's document type definition, and will form the basis for production of Memoria Deliverable D8, the major output of Work Package 3. Some discussion of general encoding issues and principles is also included.

The present document also includes, as an annex, a copy of the full documentation for the TEI Lite subset, which has effectively defined the standards applied during the rest of this work package

Section 4 of D1.1 lists a number of "Types of Structure Taken into Consideration". Our discussion here follows the order in which topics are presented in that document, with cross references to it for the convenience of the reader.

2 Ontological Issues

Under the heading "Volume/author/work/book", section 4.1 of D1.1 raises a number of important ontological issues concerning the status of the various components making up *La Pleiade*. Like any other collection, this may be analysed along two dimensions:

- as a series of physical volumes (or *tomes*)
- as a series of hierarchical groupings of bibliographic items, for example. the works of a given author (*oeuvre*); a collection of related items (e.g. the dramatic works of a given author, or set of authors); a single work.

We propose to restrict the term *text* to this last sense, and introduce the term *composite* to refer to any combination of several texts treated for some purpose as a single item.

Complications arise for three largely unrelated reasons:

- The mapping of bibliographic items to physical volumes is to a large extent arbitrary. An *oeuvre* may span several *tomes*; one *tome* may contain several unrelated *texts* or a mixture of other bibliographic items. Items making up a composite may be spread across several *tomes*, with other unrelated items intervening
- Some composite items have a bibliographic identiy of their own, and thus need to be regarded both as unitary and composite. For example, in certain circumstances *A la recherche du temps perdu* is both a composite (comprising individual texts *Du cote de chez Swann*, etc.) and an individual work, itself forming part of the oeuvre of Marcel Proust.

- The individual components of a given composite may also form part of other composites. For example, the plays of Sade appear both as a subcomponent of the composite “18th century drama” and as a subcomponent of the oeuvre of Sade.

The TEI proposes SGML elements for the representation of both unitary and composite texts: for the former, the tag `<text>` is used, and for the latter the tag `<group>`. A `<text>` may contain either a `<body>` element or a `<group>`; a `<group>` may contain further nested `<group>` or `<text>` elements. Thus, Proust’s work might be tagged as follows:

```
  <text><!--A la recherche du temps perdu-->
    <group>
       <text>
       <body><!-- Du cote  de chez Swann--></body>
       </text>
       <text>
       <body><!-- A l'ombre des jeunes filles en fleur--></body>
       </text>
<!-- ... etc -->
    </group>
 </text>
```

Several other more complex examples are given in the Guidelines, (section 7.3).

This allows us to treat both the whole work and its components as individual texts, but it does not represent all the other composites in which (say) *Swann* might appear. However, it is our view that the purpose of any encoding must be not to show all possible groupings – since a moment’s reflection shows that the number of such possible groupings is infinite. – but only the particular grouping applicable in a given case. In encoding a work such as *La Pleiade*, one would normally choose to respect the structural view of the original editors (this may even be the main object of the analysis); in creating new collections, however, editorial decisions will have to be made and sometimes, the decision as to which view is appropriate will be an arbitrary one.

3 “Registers” of the main text

Section 4.2 of D1.1 identifies a number of “Registers” which should be distinguished within the main body of a text, specifically

- *metatext* (that is, parts of a text which describe it, such as contents pages, indexes, and running titles)
- *peritext* (that is, more extensive parts of a text which comment on, introduce,or complement a text, such as introductions, annexes, bibliographies etc.)
- *appareil critique* (that is, critical or interpretative commentary of any kind, other than that relating to the state of the text, which is classed as paratext; note that in English the term *critical apparatus*may be used to describe either or both).

- *paratext*(that is editorial commentary or variant information relating specifically to the state of the text)
- *quasi-text* (we propose this term for the items mentioned in D1.1 under the heading "Full-page plates"; it covers such items as free-standing illustrations with their own associated text, appendixes of unrelated material such as publishers advertisements, ex-libris plates etc.)

For brevity, we treat some of these distinctions together, while recognising the virtue in making them.

3.1 Peritext and metatext

We find it hard in the general case to distinguish peritextual and metatextual materials. Moreover, these terms seem to us to relate rather more to the perceived qualities of particular textual units than to their essential natuire. Thus, a division of a text which has a peritextual function (such as the preface to a scholarly publication) is not structurally distinct from a division which does not (such as a chapter of the work in question). Indeed, there is a long-standing literary tradition (from Rabelais to Queneau, including writers as distinct as Browne, Melville and Pope) in which scholarly or pseudo-scholarly peritextual units form an integral part of the work.

Nevertheless, the TEI scheme allows for a distinction between front or back matter and the body of a text. The normal assumption is that `<front>` and `<back>` elements will contain peritextual divisions, but these will themselves be composed of the same elements as the `<body>` element. To distinguish amongst the role of these various divisions, the `type` attribute must be used. See further the discussion of the `<div>` element below (section 4.6 *Architectural units* on page 7); for more details of the TEI recommendations specific to front and back matter see U5 section 19, and P3 chapter 7). These include elements appropriate to all the features listed above as having a metatextual or peritextual function.

Full meta-textual information for a given electronic text is also encoded in the TEI header, which is not discussed further here, except to note the important distinction between the header, which contains information added to an electronic text during or after its creation, and metatextual information already present or pre-existing in a text, which will generally appear as its front and back matter.

3.2 Paratext and *appareil critique*

Paratextual and text-critical information may appear in front or back matter, or it may be embedded within a text. The elements `<note>` and `<interp>` (further discussed below) may be used to record it in either case.

For the specific case of textual variation, the TEI scheme contains a wealth of encoding mechanisms. The elements `<add>` and `<del>` may be used to mark additions or deletions; correction and regularisation may also be marked using elements `<corr>` and `<reg>`; lacunae in an encoding may be marked by the `<gap>` element, and so forth. These "editorial interventions" are described further in section 9 of U5 and in more detail in section 6.5 of P3.

Three additional TEI element sets provide for more elaborate paratextual requirements. The textual criticism tagset described in chapter 19 of P3 may be used for the encoding of the traditional textual apparatus; as noted below, the tagset described in chapter 18 of P3 provides specialised elements for the transcription of primary textual sources; additional elements for the encoding of certainty and responsibility is described in chapter 17.

3.3 Quasi-text

The TEI scheme permits a `<text>` element to be nested arbitrarily at various points within another `<text>`, with the implied semantics that the nested text has this quasi-textual function. This facility should be carefully distinguished from the notion of composite texts discussed above, and may be appropriate for such cases as highly structured and freely floating components of this type. The `<figure>` element used for any form of graphic for example may contain an embedded `<text>`. Here as elsewhere, serious ontological questions may need to be resolved before an appropriate tagging can be chosen.

4 Specific Textual Units

Under this heading, D1.1 groups nine categories of "textual unit". Again, for ease of comparison, we follow here the presentation of that deliverable, although it does not correspond very naturally with the organization of the TEI scheme.

4.1 Language

The TEI scheme recognizes the importance of specifying the language for every component of a text by defining a global attribute `lang`as well defining a specific element `<foreign>`. It will often be the case that the textual unit whose language is to be recorded will itself be tagged for reasons independent of its language. For example, a quotation will be marked as such, whatever language it uses. In other cases, the `<foreign>` element may be used to introduce a textual unit whose only interest is the language used. In all cases, language is specified by an attribute.

The value of the `lang` attribute on an element is inherited by all elements contained within it. Thus, if a text is primarily in French, with occasional paragraphs in Latin, one need specify the latter only, as in the following example:

```
<text lang=FRA>
<p>Cette paragraphe est en fran&ccedill;ais.
<p>Celle-ci &eacute;galement.
<p lang=LAT>hic lingua latina loquitur
<p>Pas besoin de le redire: c'est toujours du bon fran&ccedill;ais.
</text>
```

A section of the TEI Header defines the codes used for `lang` attribute values.For languages using different character sets (e.g. Greek, Russian), this definition may also specify a *writing system declaration* (see further TEI P3, chapter 25).

4.2 Authorship

Authorship is a rather more contentious issue than language, and no specific elements or attributes analogous to those for language are provided for it within the TEI scheme. A number of general purpose mechanisms are however provided to associate interpretative judgments such as this with particular stretches of text. One could use the general purpose <interp> element to mark a possibly disputed authorship attribution; or alternatively, where authorship is an uncomplicated feature-value pair, the <milestone> tag might be used to mark points at which it changes.

In the simplest case, of course, authorship is specified within the header of a text. If stated explicitly in the text itself, it will be tagged within the <docAuthor> element in the front matter.

4.3 Registers

A number of cases of distinct levels of discourse are mentioned in section 4.3.3. of D1.1. They are listed below, together with the corresponding TEI element

stage directions etc. The <stage> element marks stage directions; it has a `type` attribute with an open-ended set of values for finer discrimination. Sub-parts of a dramatic text such as interludes, epilogues etc. should be marked using the <div> or <div1> elements and their `type` attribute, as discussed below.

direct or indirect speech Several elements are available: <q> marks quoted or direct speech; <sp> marks a spoken paragraph; <soCalled>, <mentioned> and <term> distinguish various discourse levels. (see further TEI U5, section 6; P3 chapter 6.3)

annotations and comments The <note> element may be used to specify any kind of annotation, either authorial or editorial. It bears a number of attributes which can be used to distinguish the kind of annotation,and also to link it with others

4.4 Intertextuality

Notes may be linked with their object by means of a `target` attribute, or by other means discussed in the next section. Cross-references which have important content may be represented by means of the <ref> element; those which are empty, by means of the <ptr> element, as discussed in TEI P3 chapter 14, and in TEI U5 section 8. For linkage outside the current document, the TEI scheme provides for a simple subset of HyTime compliant addressing methods, using the corresponding <xptr> and <xref> elements.

4.5 Reflected structures

Under this heading, D1 discusses the fact that the structure of parts of a given text may deliberately echo the structure of parts as when, for example, a division composed of notes is deliberately organized to reflect the structure of the object annotated.

This is a particular (and striking) instance of the general need to represent linkage and alignment between arbitrary sections of the work. The general purpose alignment mechanisms defined by the TEI scheme are certainly adequate to this task, though their exact application may require some careful attention. For the case cited, and many others, either of the following approaches would be appropriate.

4.5.1 In-place linking

This is most simply achieved by what we will term the *implicit* linkage effected by the <note>element, given in line. When notes are collected together in a distinct division, they may be linked to their object by means of the <note>element'starget attribute. Alternatively, given a section S1 composed of a series of analytic notes, each tagged as a <note>, <interp> or other appropriate element, each of which is to be linked to a section (or other element) of a text, the linkage can be made in either or both directions by using the ana (analysis) and/or inst (instance) global linking attributes, as discussed in chapter 14 of TEI P3, or section 8.3 of TEI U5.

For example, if the note with identifier n1 is to be associated with the textual division with identifier TD9 , a two way link might be effected by tagging the two elements thus:

```
<div1 type=notes>
    <note id=n1 inst=TD9> .... </note>
     <!-- other notes here -->
</div1>
<div1 id=TD9 ana=n1> ... </div1>
<!-- possibly other divisions here -->
```

The use of the inst attribute allows one to associate the same note with several distinct elements, where this is appropriate, by supplying all the relevant identifiers as the value of the attribute.

4.5.2 External linking

A more general alternative to the above procedure is to define the linkage between analysis and its target by means of the free-standing <link>element defined by TEI P3 (see section xx). Assuming, as above, that note n1 and textual division TD9 are to be associated, a <link>element is used to represent the association directly. This element carries a targets attribute which specifies the identifiers of the elements to be associated, as well as a number of others specifying more information about the nature of the association.

The above example might be encoded as follows:

```
<div1 type=notes>
    <note id=n1> .... </note>
       <!-- other notes here -->
</div1>
<div1  type=play id=TD9> ... </div1>
  <!-- possibly other div1s here -->

<linkGrp type=notes>
  <link targets='TD9 N1'>
   <!-- other links here -->
</linkGrp>
```

The position of the <link> elements and their grouping <linkGrp> element is not defined by the TEI; as a kind of electronic peritext they might most naturally appear in either the front or back matter of the text. Their use of the SGML identifier mechanism implies that they must appear within the same SGML document. If they do not, an intermediate <xptr> element may be used to specify their location.

4.6 Architectural units

The major structural components of texts are generally given highly culture-specific names. As discussed in P3 (section 7.1), the TEI approach is to use neutrally named <div> or <div1> (etc) elements for all such architectural units, optionally using a type attribute to distinguish amongst them. This attribute must be supplied for the first such unit at a given level, and is understood to apply to all others at the same level thereafter.

For example, a play divided hierarchically into acts and scenes might be encoded as follows:

```
<text><body>
<div1 type=act n=1>
  <div2 type=scene n=1.1> ... </div2>
  <div2 type=scene n=1.2> ... </div2>
</div1>
<div1 type=act n=2>
  <div2 type=scene n=2.1> ... </div2>
<div2 type=scene n=2.2> ... </div2>
</div1>
```

Note that the text type of a division (e.g. prose or verse) is not specified at this level, reflecting the heterogeneity commented on in D1 section 4.3.6. Instead , it is implied by the elements making up thedivision (for example, <p> for prose paragraph, <speech> for speech in a drama, <l>for verse line, and so forth). There are many such elements, described and exemplified in TEI U5. We may mention here only those specifically listed in D1: <formula> for mathematical theorems; <term> and <gloss>; <head> for section heading or subheading; <title> for an embedded title; <hi> for a typographically highlighted phrase.

4.7 Linguistically defined units

Segmentation of a linguistic nature is essentially hierarchic, although not all linguistic units are well-behaved with respect to a single hierarchy (as such phenomena as anaphora attest). The TEI scheme proposes a neutral <seg> element for all such segmentation in the general case, although for particular purposes it may be preferable to use “flavoured” segments such as <w> for word, <phr> for phrase, <m> for morpheme, etc. may also be used. See further P3, section 15.1. Not all of these elements are defined in TEI Lite however.

4.8 Lexically defined units

D1.1 calls attention to the potentially very large set of such units, whose lexically different nature suggests that they should be marked off from the rest of the text in some way. There is a large number of such elements, many of which are discussed and exemplified in TEI U5.

Again, we list here only those such elements mentioned in D1. <abbr> for abbreviations; <rs> for logos and other referring strings such as proper names of persons and places; <date> for dates; <num> for numbers; <formula> for chemical or mathematical formulas. The <index> and <term>elements are also defined specifically for units which are to be included in an index. Details of all these and others are to be found in TEI U5 and in chapter 6 of TEI P3.

4.9 Media defined units

As noted above, the structure defined by the units making up the physical media supporting a given text very rarely corresponds with the logical structure of a text. Paragraphs and pages do not form part of the same hierarchy. For this reason, the TEI proposes the use of empty "milestone" tags to mark points at which changes in the physical structure occur: <pb> for page breaks, <lb> for line breaks, <cb> for column breaks. A generic <milestone> element is also provided for other such boundary markers, which may be used to indicate e.g. foliation, changes in authorship (as noted above) etc.

Where close attention to the physical organization of a text is the primary object of analysis, more detailed and specific elements will be needed. Some of these are discussed in chapter 18 of TEI P3.