These textual structures overlap with each other in complex and unpredictable ways. Particularly when dealing with texts as instantiated by paper technology, the reader needs to be aware of both the physical organization of the book and the logical structure of the work it contains. Many great works (Sterne's *Tristram Shandy* for example) cannot be fully appreciated without an awareness of the interplay between narrative units (such as chapters or paragraphs) and page divisions. For many types of research, it is the interplay between different levels of analysis which is crucial: the extent to which syntactic structure and narrative structure mesh, or fail to mesh, for example, or the extent to which phonological structures reflect morphology.

2.3 SGML Structures

This section describes the simple and consistent mechanism for the markup or identification of structural textual units which is provided by SGML. It also describes the methods SGML provides for the expression of rules defining how combinations of such units can meaningfully occur in any text.

2.3.1 Elements

The technical term used in the SGML standard for a textual unit, viewed as a structural component, is *element*. Different types of elements are given different names, but SGML provides no way of expressing the meaning of a particular type of element, other than its relationship to other element types. That is, all one can say about an element called (for instance) <blort> is that instances of it may (or may not) occur within elements of type <farble>, and that it may (or may not) be decomposed into elements of type <blortette>. It should be stressed that the SGML standard is entirely unconcerned with the semantics of textual elements: these are application dependent.

Work is currently going on in the standards community to create (using SGML syntax) a definition of a standard "document style semantics and specification language" or DSSSL.

It is up to the creators of SGML conformant tag sets (such as these Guidelines) to choose intelligible names for the elements they identify and to document their proper use in text markup. That is one purpose of this document. From the need to choose element names indicative of function comes the technical term for the name of an element type, which is *generic identifier*, or GI.

Within a marked up text (a *document instance*), each element must be explicitly marked or tagged in some way. The standard provides for a variety of different ways of doing this, the most commonly used being to insert a tag at the beginning of the element (a *start-tag*) and another at its end (an *end-tag*). The start- and end-tag pair are used to bracket off the element occurrences within the running text, in rather the same way as different types of parentheses or quotation marks are used in conventional punctuation. For example, a quotation element in a text might be tagged as follows:

```
    ...  Rosalind's remarks <quote>This is the silliest stuff
  that ere I heard of!</quote> clearly indicate ...
```

As this example shows, a start-tag takes the form <name>, where the opening angle bracket indicates the start of the start-tag, "name" is the generic identifier of the element which is being delimited, and the closing angle bracket indicates the end of a tag. An end-tag takes an identical form, except that the opening angle bracket is followed by a solidus (slash) character, so that the corresponding end-tag would be </name>.

The actual characters used for the delimiting characters (the angle brackets, exclamation mark and solidus) may be redefined, but it is conventional to use the characters used in this description.

2.3.2 Content Models: An Example

An element may be *empty*, that is, it may have no content at all, or it may contain simple text. More usually, however, elements of one type will be *embedded* (contained entirely) within elements of a different type.

To illustrate this, we will consider a very simple structural model. Let us assume that we wish to identify within an anthology only poems, their titles, and the stanzas and lines of which they are composed. In SGML terms, our document type is the *anthology*, and it consists of a series of *poem*s. Each poem has embedded within it one element, a title, and several occurrences of another, a stanza, each stanza having embedded within it a number of line elements. Fully marked up, a text conforming to this model might appear as follows:

The example is taken from William Blake's *Songs of innocence and experience* (1794). The markup is designed for illustrative purposes and is not TEI-conformant.

```
<anthology>
  <poem><title>The SICK ROSE</title>
     <stanza>
```

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

            <!- more poems go here    ->

</anthology>
```

It should be stressed that this example does *not* use the same names as are proposed for corresponding elements elsewhere in these Guidelines: the above is *not* a valid TEI document. It will however serve as an introduction to the basic notions of SGML. White space and line breaks have been added to the example for the sake of visual clarity only; they have no particular significance in the SGML encoding itself. Also, the line

```
            <!- more poems go here    ->
```

is an SGML *comment* and is not treated as part of the text.

This example makes no assumptions about the rules governing, for example, whether or not a title can appear in places other than preceding the first stanza, or whether lines can appear which are not included in a stanza: that is why its markup appears so verbose. In such cases, the beginning and end of every element must be explicitly marked, because there are no identifiable rules about which elements can appear where. In practice, however, rules can usually be formulated to reduce the need for so much tagging. For example, considering our greatly over-simplified model of a poem, we could state the following rules:

1. An anthology contains a number of poems and nothing else.
2. A poem always has a single title element which precedes the first stanza and contains no other elements.
3. Apart from the title, a poem consists only of stanzas.
4. Stanzas consist only of lines and every line is contained by a stanza.
5. Nothing can follow a stanza except another stanza or the end of a poem.
6. Nothing can follow a line except another line or the start of a new stanza.

From these rules, it may be inferred that we do not need to mark the ends of stanzas or lines explicitly. From rule 2 it follows that we do not need to mark the end of the title—it is implied by the start of the first stanza. Similarly, from rules 3 and 1 it follows that we need not mark the end of the poem: since poems cannot occur within poems but must occur within anthologies, the end of a poem is implied by the start of the next poem, or by the end of the anthology. Applying these simplifications (applicable only in SGML), we could mark up the same poem as follows:

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


<poem>
            <!- more poems go here    ->
```