TEI Guidelines for Electronic Text Encoding and Interchange

12: Print Dictionaries

ID: DI

Print Dictionaries

This chapter defines a base tag set for encoding human-oriented monolingual and polyglot dictionaries (as opposed to computational lexica, which are intended for use by language-processing software). Dictionaries are most familiar in their printed form; however, increasing numbers of dictionaries exist also in electronic forms which are independent of any particular printed form, but from which various displays can be produced --- e.g. CD-ROM dictionaries.

Both typographically and structurally, dictionaries are extremely complex. In addition, dictionaries interest many communities with different and sometimes conflicting goals. As a result, many general problems of text encoding are particularly pronounced here, and more compromises and alternatives within the encoding scheme may be required.
We refer the reader to previous and current discussions of a common format for encoding dictionaries. For example, Robert A. Amsler and Frank W. Tompa, *An SGML-Based Standard for English Monolingual Dictionaries,* in *Information in Text: Fourth Annual Conference of the U[niversity of] W[aterloo] Centre for the New Oxford English Dictionary* October 26-28, 1988, Waterloo, Canada, pp. 61-79; Nicoletta Calzolari et al., *Computational Model of the Dictionary Entry: Preliminary Report,* Acquilex: Esprit Basic Research Action No. 3030, Six-Month Deliverable, Pisa, April 1990; John Fought and Carol Van Ess-Dykema,*Toward an SGML Document Type Definition for Bilingual Dictionaries,* TEI working paper TEI AIW20 (available from the TEI); Nancy Ide and Jean Veronis, *Encoding Print Dictionaries*, *Computers and the Humanities* (special TEI issue --- to appear); Nancy Ide, Jacques Le Maitre, and Jean Veronis, *Outline of a Model for Lexical Databases,* (Information Processing and Management, 29, 2, 159-186, 1993); Nancy Ide, Jean Veronis, Susan Warwick- Armstrong, Nicoletta Calzolari, *Principles for Encoding machine readable dictionaries*,*Proceedings of the Fifth EURALEX International Congress, EURALEX'92* (to appear), University of Tempere, Finland; and The DANLEX Group, *Descriptive tools for electronic processing of dictionary data,* in *Lexicographica, Series Maior* (Tübingen: Niemeyer, 1987). Two problems are particularly prominent.

First, because the structure of dictionary entries varies widely both among and within dictionaries, the simplest way for an encoding scheme to accommodate the entire range of structures actually encountered is to allow virtually any element to appear virtually anywhere in a dictionary entry. It is clear, however, that strong and consistent structural principles do govern the vast majority of conventional dictionaries, as well as many or most entries even in more exotic dictionaries; ideally, a set of encoding guidelines should capture these structural principles. We therefore define two distinct elements for dictionary entries, one (<entry>) which captures the regularities of most conventional dictionary entries, and a second (<entryFree>) which uses the same elements, but allows them to combine much more freely. It is recommended that <entry> be used in preference to <entryFree> wherever the structure of the entry allows it. These elements and their contents are described in sections 12.2, The Structure of Dictionary Entries , 12.6, Unstructured Entries , and 12.4, Headword and Pronunciation References .

Second, since so much of the information in printed dictionaries is implicit or highly compressed, their encoding requires clear thought about whether it is to capture the precise typographic form of the source text or the

underlying structure of the information it presents. Since both of these views of the dictionary may be of interest, it proves necessary to develop methods of recording both, and of recording the interrelationship between them as well. Users interested mainly in the printed format of the dictionary will require an encoding to be faithful to an original printed version. However, other users will be interested primarily in capturing the lexical information in a dictionary in a form suitable for further processing, which may demand the expansion or rearrangement of the information contained in the printed form. Further, some users wish to encode both of these views of the data, and retain the links between related elements of the two encodings. Problems of recording these two different views of dictionary data are discussed in section 12.5, Typographic and Lexical Information in Dictionary , together with mechanisms for retaining both views when this is desired.

Whichever view is adopted, a parameter entity TEI.dictionaries must be declared within the document type subset of any document using this base tag set. This should have the value INCLUDE, as further described in section 3.3, Invocation of the TEI DTD . A document using this base tag set and no other additional tag sets will thus begin as follows:

```
<!DOCTYPE TEI.2 system 'tei2.dtd' [
     <!ENTITY % TEI.dictionaries 'INCLUDE' >
]>
```

12.1: Dictionary Body and Overall Structure

Dictionary Body and Overall Structure

Overall, dictionaries have the same structure of front matter, body, and back matter familiar from other texts; the base tag set for dictionaries uses the same front-matter and back-matter elements as other TEI base tag sets; these are documented in chapter 7, Default Text Structure . In addition, dictionaries define the elements <entry>, <entryFree>, and <superEntry> as component-level elements which can occur directly within a text division or the text body.

The following tags should be used to mark the gross structure of a printed dictionary; the dictionary-specific tags are discussed further in the following section.

`<text>`
- contains a single text of any kind, whether unitary or composite, for example a poem or drama, a collection of essays, a novel, a dictionary, or a corpus sample.

`<front>`
- contains any prefatory matter (headers, title page, prefaces, dedications, etc.) found before the start of a text proper.

`<body>`
- contains the whole body of a single unitary text, excluding any front or back matter.

`<back>`
- contains any appendixes, etc. following the main part of a text.

`<div>`
- contains a subdivision of the front, body, or back of a text.

`<div0>`
- contains the largest possible subdivision of the body of a text.

`<div1>`

- contains a first-level subdivision of the front, body, or back of a text (the largest, if <div0> is not used, the second largest if it is).

`<entry>`
- contains a reasonably well-structured dictionary entry.

`<entryFree>`
- contains a dictionary entry which does not necessarily conform to the constraints imposed by the <entry> element.

`<superentry>`
- groups successive entries for a set of homographs.

The text-division elements <div2> through <div7> may also be used, as described in chapter 7, Default Text Structure .

As members of the class entries, <entry> and <entryFree> share the following attributes:

`type`
- indicates type of entry, in dictionaries with multiple types. Suggested values include:

`main`
- a main entry (default).

`hom`
- a homograph with a separate entry.

`xref`
- a reduced entry whose only function is to point to another main entry (e.g. for forms of an irregular verb or for variant spellings: was pointing to be, or esthete to aesthete).

`affix`
- an entry for a prefix, infix, or suffix.

`abbr`
- an entry for an abbreviation.

`supplemental`
- a supplemental entry (for use in dictionaries which issue supplements to their main work in which they include updated information about entries).

`foreign`
- an entry for a foreign word in a monolingual dictionary.

`key`
- contains a (sortable) character sequence reflecting the entry's alphabetical position in the printed dictionary.

The front and back matter of a dictionary may well contain specialized material like lists of common and proper nouns, grammatical tables, gazetteers, a guide to the use of the dictionary, etc. These may be tagged as elements defined in the core tag set (chapter 6, Elements Available in All TEI Documents) or as specialized dictionary elements as defined in this chapter.

The <body> element consists of a set of entries, optionally grouped into one or several , <div0>, or <div1> elements. These text divisions might correspond, for example, to sections for different languages in a bilingual

dictionaries, sections for different letters of the alphabet, etc.
It is unlikely that many conventional dictionaries will require smaller divisions, but all the usual division elements <div2> through <div7> may be used. In print dictionaries, entries are typically typographically distinct entities, each headed by some morphological form of the lexical item described (the headword), and sorted in alphabetical order or (for non-alphabetic scripts) in some other conventional sequence. Dictionary entries should be encoded as distinct successive items, each marked as an <entry> element. The type attribute may be used to distinguish different types of entries, for example main entries, related entries, run-on entries, or entries for cross-references, etc.

Some dictionaries provide distinct entries for homographs, on the basis of etymology, part-of-speech, or both, and typically provide a numeric superscript on the headword identifying the homograph number. In these cases each homograph should be encoded as a separate entry; the <superEntry> element may optionally be used to group such successive homograph entries. In addition to a series of <entry> elements, the <superEntry> may contain a preliminary <form> group (see section 12.3.1, Information on Written and Spoken Forms) when information about hyphenation, pronunciation, etc., is given only once for two or more homograph entries. If the homograph number is to be recorded, the global attribute n should be used for this purpose. In some dictionaries, homographs are treated in distinct parts of the same entry; in these cases, they may be separated by use of the <hom> element, for which see section 12.2.1, Hierarchical Levels .

A sort key, given in the key attribute, is often required for superentries and entries, especially in cases where the order of entries does not follow the local character-set collating sequence (as, for example, when an entry for ``3D'' appears at the place where ``three-D'' would appear).

The body of a bilingual dictionary with two parts will thus have an overall structure resembling the following:

```
<body>
  <div0 type='dictionary'>
    <!-- English-French -->
    <entry>...</entry>
    <entry>...</entry>
    <entry>...</entry>
  <!-- ... -->
  </div0>
  <div0>
    <!-- French-English -->
    <entry>...</entry>
    <entry>...</entry>
    <entry>...</entry>
  <!-- ... -->
  </div0>
</body>
```

A dictionary with no internal divisions might have a structure like the following; a <superEntry> is shown grouping two homograph entries.

```
<body>
  <entry>...</entry>
  <entry>...</entry>
  <!-- ... -->
  <superEntry>
    <entry type=hom n='1'>...</entry>
    <entry type=hom n='2'>...</entry>
  </superEntry>
  <!-- ... -->
</body>
```

The base tag set for dictionaries is contained in the files teidict2.ent and teidict2.dtd. The first of these defines the class comp.dictionaries, so that the generic text-division elements , <div0>, <div1>, etc. can contain <entry> elements:

```
<!-- 12.1:  Element classes for dictionary base               -->
<!-- Text Encoding Initiative: Guidelines for Electronic      -->
<!-- Text Encoding and Interchange. Document TEI P3, 1994.    -->

<!-- Copyright (c) 1994 ACH, ACL, ALLC. Permission to copy    -->
<!-- in any form is granted, provided this notice is          -->
<!-- included in all copies.                                  -->

<!-- These materials may not be altered; modifications to     -->
<!-- these DTDs should be performed as specified in the       -->
<!-- Guidelines in chapter "Modifying the TEI DTD."           -->

<!-- These materials subject to revision. Current versions    -->
<!-- are available from the Text Encoding Initiative.         -->


<!-- First we define attributes available on all the          -->
<!-- elements in this tag set.                                -->

<!-- ... declarations from section 12.5.4                     -->
<!--     (Attributes for dictionary work)                     -->
<!--     go here ...                                          -->

<!-- Next we define comp.dictionaries, which will be used in  -->
<!-- the declaration of component, within file TEI2.DTD.      -->

<!ENTITY % x.comp.dictionaries ''                              >
<!ENTITY % m.comp.dictionaries '%x.comp.dictionaries entry |
          entryFree | superentry'                              >
<!ENTITY % mix.dictionaries '| %m.comp.dictionaries'            >

<!-- Next, we declare some specialized element classes, used  -->
<!-- in various content models in the dictionary tag set.     -->

<!ENTITY % a.entries '
          type              CDATA              "main"
          key               CDATA              #IMPLIED'      >

<!-- ... declarations from section 12.2.2                     -->
<!--  (Class for top-level structure of dictionary entries)   -->
<!--     go here ...                                          -->
<!-- ... declarations from section 12.3.1                     -->
<!--     (Classes for morphological and form information)     -->
<!--     go here ...                                          -->
<!-- ... declarations from section 12.3.2                     -->
<!--     (Elements for grammatical information)               -->
<!--     go here ...                                          -->
<!-- ... declarations from section 12.4                       -->
<!--     (Classes for headword references)                    -->
<!--     go here ...                                          -->
<!-- ... declarations from section 12.6                       -->
<!--     (Model class for unstructured dictionary entries)    -->
<!--     go here ...                                          -->
```

The dictionary-specific elements are all declared in the file teidict2.dtd, which has the following overall structure.

```
<!-- 12.1:  Base tag set for printed dictionaries             -->
<!-- Text Encoding Initiative: Guidelines for Electronic      -->
<!-- Text Encoding and Interchange. Document TEI P3, 1994.    -->

<!-- Copyright (c) 1994 ACH, ACL, ALLC. Permission to copy    -->
<!-- in any form is granted, provided this notice is          -->
<!-- included in all copies.                                  -->
```

```
<!-- These materials may not be altered; modifications to     -->
<!-- these DTDs should be performed as specified in the       -->
<!-- Guidelines in chapter "Modifying the TEI DTD."           -->

<!-- These materials subject to revision. Current versions    -->
<!-- are available from the Text Encoding Initiative.         -->

<!-- First we embed the default text structure.               -->

<![ %TEI.singleBase [
<!ENTITY % TEI.structure.dtd system 'teistr2.dtd'              >
%TEI.structure.dtd;
]]&nil;>


<!-- Now we define the dictionary-specific material.          -->

<!-- ... declarations from section 12.2.1                     -->
<!--     (Dictionary entries and their structure)             -->
<!--     go here ...                                          -->
<!-- ... declarations from section 12.3.1                     -->
<!--     (The form group)                                     -->
<!--     go here ...                                          -->
<!-- ... declarations from section 12.3.2                     -->
<!--     (The gram group)                                     -->
<!--     go here ...                                          -->
<!-- ... declarations from section 12.3.3.1                   -->
<!--     (Definition text)                                    -->
<!--     go here ...                                          -->
<!-- ... declarations from section 12.3.3.2                   -->
<!--     (Translation information)                            -->
<!--     go here ...                                          -->
<!-- ... declarations from section 12.3.4                     -->
<!--     (Etymologies)                                        -->
<!--     go here ...                                          -->
<!-- ... declarations from section 12.3.5.1                   -->
<!--     (Examples and citations)                             -->
<!--     go here ...                                          -->
<!-- ... declarations from section 12.3.5.2                   -->
<!--     (Usage information)                                  -->
<!--     go here ...                                          -->
<!-- ... declarations from section 12.3.5.3                   -->
<!--     (Cross References)                                   -->
<!--     go here ...                                          -->
<!-- ... declarations from section 12.3.6                     -->
<!--     (Related entries)                                    -->
<!--     go here ...                                          -->
<!-- ... declarations from section 12.4                       -->
<!--     (Headword references)                                -->
<!--     go here ...                                          -->
```

12.2: The Structure of Dictionary Entries

The Structure of Dictionary Entries

A simple dictionary entry may contain information about the form of the word treated, its grammatical characterization, its definition, synonyms, or translation equivalents, its etymology, cross-references to other entries, usage information, and examples. These we refer to as the constituent parts or constituents of the entry; some dictionary constituents possess no internal structure, while others are most naturally viewed as groups of smaller elements, which may be marked in their own right. In some styles of markup, tags will be applied only to the low-level items, leaving the constituent groups which contain them untagged. We distinguish the class of

top-level constituents of dictionary entries, which can occur directly within entries, from the class of phrase-level constituents, which can normally occur only within top-level constituents. The top-level constituents of dictionary entries are described in section 12.2.2, Groups and Constituents , and documented more fully, together with their phrase-level sub-constituents, in section 12.3, Top-level Constituents of Entries .

In addition, however, dictionary entries often have a complex hierarchical structure. For example, an entry may consist of two or more sub-parts, each corresponding to information for a different part-of-speech homograph of the headword. The entry (or part-of-speech homographs, if the entry is split this way) may also consist of senses, each of which may in turn be composed of two or more sub-senses, etc. Each sub-part, homograph entry, sense, or sub-sense we call a level; at any level in an entry, any or all of the constituent parts of dictionary entries may appear. The hierarchical levels of dictionary entries are documented in section 12.2.1, Hierarchical Levels .

12.2.1: Hierarchical Levels

Hierarchical Levels

The outermost structural level of an entry is marked with the elements <entry> or <entryFree>. The <hom> element marks the subdivision of entries into part-of-speech homographs. The <sense> element marks the subdivision of entries and part-of-speech homographs into senses; this element nests recursively in order to provide for a hierarchy of sub-senses of any depth. All of these levels may each contain any of the constituent parts of an entry. A special case of hierarchical structure is represented by the <re> (related entry) element, which is discussed in section 12.3.6, Related Entries .

`<entry>`
- contains a reasonably well-structured dictionary entry.

`<entryFree>`
- contains a dictionary entry which does not necessarily conform to the constraints imposed by the <entry> element.

`<hom>`
- groups information relating to one homograph within an <entry>.

`<sense>`
- groups together all information relating to one word sense in a dictionary <entry> (definitions, examples, translation equivalents, etc.) Attributes include:

`level`
- gives the nesting depth of this sense.

For example, an entry with two senses will have the following structure:

```
<entry>
  <!-- ... information common to both senses -->
  <sense n='1'>
     <!-- ... sense number 1 -->
  </sense>
  <sense n='2'>
     <!-- ... sense number 2 -->
  </sense>
</entry>
```

An entry with two homographs, the first with two senses and the second with three (one of which has two sub-senses), will have a structure like this:

```
<entry>
  <!-- ... information common to both homographs, if any ... -->
  <hom n='1'>
    <sense n='1'>...</sense>
    <sense n='2'>...</sense>
  </hom>
  <hom n='2'>
    <sense n='1'>
      <sense n='a'>...</sense>
      <sense n='b'>...</sense>
    </sense>
    <sense n='2'>...</sense>
    <sense n='3'>...</sense>
  </hom>
</entry>
```

In some dictionaries, homographs typically receive separate entries; in such a case, as noted in section 12.1, Dictionary Body and Overall Structure , the two homographs may be treated as entries, optionally grouped by a superentry:

```
<superEntry>
<!-- ... form information common to both
     homographs, if any ... -->
<entry n='1'>
  <sense n='1'>...</sense>
  <sense n='2'>...</sense>
</entry>
<entry n='2'>
  <sense n='1'>
    <sense n='a'>...</sense>
    <sense n='b'>...</sense>
  </sense>
  <sense n='2'>...</sense>
  <sense n='3'>...</sense>
</entry>
</superEntry>
```

The hierarchical levels of dictionary entries are declared as shown in the following DTD fragment. As may be seen, the content model for <entry> specifies that entries do not nest, that homographs nest within entries, and that senses nest within entries, homographs, or senses, and may be nested to any depth to reflect the embedding of sub-senses. Any of the top-level constituents (<def>, <usg>, <form>, etc.) can appear at any level (i.e., within entries, homographs, or senses).

```
<!-- 12.2.1:  Dictionary entries and their structure          -->
<!ELEMENT superentry    - O  (form?, entry+)                    >
<!ATTLIST superentry         %a.global;
                             %a.entries;                        >
<!ELEMENT entry         - O  (hom | sense |
                             %m.dictionaryTopLevel)+
                                          +(anchor)     >
<!ATTLIST entry              %a.global;
                             %a.entries;                        >
<!ELEMENT entryFree     - O  (#PCDATA)

                                         +(%m.dictionaryParts
                                         | %m.phrase |
                                         %m.inter)      >
<!ATTLIST entryFree          %a.global;
                             %a.dictionaries;
                             %a.entries;                        >
<!ELEMENT hom           - O  (sense | %m.dictionaryTopLevel)*
                                            -(entry)    >
<!ATTLIST hom                %a.global;
```

```
                          %a.dictionaries;                  >
<!ELEMENT sense         - -  (sense | %m.dictionaryTopLevel |
                          %m.phrase | #PCDATA)*             >
<!ATTLIST sense           %a.global;
                          %a.dictionaries;
          level           NUMBER             #IMPLIED       >
<!-- This fragment is used in sec. 12.1                     -->
```

12.2.2: Groups and Constituents

Groups and Constituents

As noted above, dictionary entries, and subordinate levels within dictionary entries, may comprise several constituent parts, each providing a different type of information about the word treated. The top-level constituents of dictionary entries are:

- information about the form of the word treated (orthography, pronunciation, hyphenation, etc.)
- grammatical information (part of speech, grammatical sub-categorization, etc.)
- definitions or translations into another language
- etymology
- examples
- usage information
- cross-references to other entries
- notes
- entries (often of reduced form) for related words, typically called related entries

Any of the hierarchical levels (<entry>, <entryFree>, <hom>, <sense>) may contain any of these top-level constituents, since information about word form, particular grammatical information, special pronunciation, usage information, etc., may apply to an entire entry, or to only one homograph, or only to a particular sense. The examples below illustrate this point.

The following elements are used to encode these top-level constituents: