

From: R. A. Amsler and F. W. Tompa, 1988. "An SGML-based Standard for English Monolingual Dictionaries," *Information in Text: Proc. 4th Conf. of Univ. of Waterloo Centre for the New OED (October 26-28, 1988)*, pp. 61-80.

## Appendix — Proposal for Dictionary Encoding

### 1. Presentation of proposed tags and attributes

This appendix contains a preliminary list of tags and attributes that could be used to represent the information included in several dictionaries.

For each proposed tag, we first present a tag identifier and a definition or some usage notes to clarify the intent of the tag. We then identify how the proposed content of the tag is encoded in current representations:

- For the *Oxford English Dictionary*, we give the corresponding tag name in the encoding currently used by the Oxford University Press followed by the tag name in the encoding currently used by the University of Waterloo. For example,

[OED]:hwlem/HW means that Oxford has named its tag "hwlem" and Waterloo has chosen the identifier to be "HW".

- Similarly, for the *Oxford Advanced Learner's Dictionary*, we give the tag name used in the tagged version available from the Oxford Text Archive followed by the tag name used in a trial version being distributed by the Oxford University Press.
- For *Webster's Seventh*, we give the record identifier for the corresponding information and we give the field number in cases where the complete record contains other data as well. This information is represented for the existing encoding used by many researchers and for the encoding proposed by Slocum, as in the following example:

[W7]: text definition D:6 / text M:3

which shows that the existing encoding of W7 includes the information in record D field 6 (entitled "text definition") and Slocum has proposed it to be in record M field 3, which he calls "text".

- Similarly, for *Longman's Dictionary*, we give the record number and field number used to encode the information. For example,

[LDOCE]: 03,04:3,10:4

shows that the information corresponding to the proposed tag is represented in the LDOCE data in record 03, in record 04 field 3, and in record 10 field 4.

The last part of each explanation contains our proposed list of attributes, and shows for each its name, the intent of the attribute, the domain from which values are to be taken, and (where appropriate) a list of corresponding encodings in current machine-readable dictionaries. For example,

**dial** regional or social dialectic variant = CDATA

[OALD] / lab = CDATA

[W7] / dial (L:5)

shows that the attribute **dial** will take arbitrary strings of characters (CDATA) as values, that the OALD available from the Oxford University Press uses an attribute "lab" for the same purpose, and that W7 uses field 5 in record L (named "dial") for encoding the same information. If a proposed attribute is represented by a *tag* in either the OED or the OALD, the name of the tag is given in angle brackets; thus,

[OALD] <pos> / ps

indicates that the OALD available from the Oxford Text Archive uses the tag identifier "pos" whereas the one available from the Oxford University Press uses an attribute name "ps".

## 2. Layout and Content Hierarchies

Physical units (page, line, column, etc.) page

*attributes:*

**left-guideword**

**right-guide word**

num

illustration

column

Logical units (entry, headword, etc.) — See below.

## 3. Universal attributes

*definition:* attributes that can show up on any tag

Note: In many cases, attributes shown as single units might appear instead as lists; this has to be accommodated somehow.

**ed** edition/printing

*current usage:*

[W7]: ver# (H:7)

**by** author/scholar

**id** identification number = CDATA

*current usage:*

[OED]: id = NUMBER / —

[LDOCE]: Serial No on 01 records = CHAR NUMBER

## 4. Logical Units

### 4.1. factotum

*definition:* [W3] an ornamental oversized capital letter used in printing

*current usage:*

[LDOCE]: 20

### 4.2. ME — main entry

*definition:*

[W7] *vocabulary entry* — a word (as the *noun book*), hyphenated or open compound (as the verb *book-match* or the noun *book review*), word element (as the affix *pro-*\ abbreviation (as *agt*)<sub>9</sub> verbalized symbol (as *No*), or term (*as man in the street*) entered alphabetically in a dictionary for the purpose of definition or identification or expressly included as an inflectional form (as the noun *godless-ness* or the adverb *globally*) or related phrase (as *one for the book*) run on at its base word and usu. set in a type (as boldface) readily distinguishable from that of the lightface running text which defines, explains, or identifies the entry.

*current usage:*

[OED]: entry / E

[OALD]: entry / ent

[W7]: F / H

[LDOCE]: 01

*attributes:*

**id** identification number = CDATA

*current usage:*

[OED] id = NUMBER / —

	[LDOCE] Serial No = CHAR NUMBER
<b>key</b>	sort key to determine entry's placement = CDATA <i>current usage:</i> [OALD] — / h = CDATA [W7] — / hdwrđ (H:2)
<b>type</b>	entry type = (main   xref   affix   abbr   suppl) <i>current usage:</i> [OED] xref = (t) and use of tags for suppl/del/com/etc. [OALD] type = (main   xref) / — [W7] Prefix/Suffix/Infix (F:4) / [LDOCE] entry form codes {S,A,R,N,Z} (02:3.1)
<b>status</b>	word status = status-NAMES <i>current usage:</i> [OED] status = (obs   ali   spu   err)
<b>hom</b>	homonym/homograph number = NUMBER <i>current usage:</i> [OED] hom = NUMBER [OALD] <hom> / hn = NUMBER [W7] homograph number (F:3) / hom# (H:3) [LDOCE] Homograph (02:2)
<b>pos</b>	entry part of speech = pos-NAMES <i>current usage:</i> [OED] <ps> / <PS> [OALD] <pos> / — [W7] part of sp.. joiner, secondary part of speech (F:6-8) / cats (H:6) [LDOCE] POS (05:2)
<b>geo</b>	geographic region (e.g., Australia)
<b>dom</b>	subject domain (e.g., nuclear physics)
<b>regis</b>	register (e.g., colloquial)
<b>time</b>	currency or frequency (e.g., obsolete, rare)
<b>sem</b>	semantic (e.g., figurative)
<b>gram</b>	grammatical code (e.g., transitive)
— The next two features are included for compatibility with [LDOCE] —	
<b>posf</b>	[LDOCE] part of speech of first element of open compound (02:4.1)
<b>posl</b>	[LDOCE] part of speech of last element of open compound (02:4.2)

### 4.3. F — Forms (written/spoken)

*definition:*

set of associated written and/or spoken forms of lexical items

note: existing dictionaries keep this implicitly by ordering the parts of entries.

Comment: Must this be elaborated to handle OED's entry for "be v." which includes <orth> and <pron>?

*current usage:*

[LDOCE]: 04 for variants

*attributes:*

**infl** inflectional use (e.g., pi, pt, pp, comp) = infl-NAMES

**dial** regional or social dialectal variant = CDATA

**hist** temporally restricted variant = CDATA

#### 4.3.1. orth — orthography

*current usage:*

[OED]: distributed among hwlem, vf, blem, ilem, etc.

[W7]: in F, V, R records / hdwrđ, form in H, D, and F records

[LDOCE]: in 01, 04, 05, 10

*attributes:*

<b>cap</b>	capitalization convention (usu, sometimes, etc.) = freq-NAME [OALD] /cap = (t   f)
<b>dial</b>	regional or social dialectal variant = CDATA [O ALD] / lab = CDATA [W7] / dial (L:5)
<b>hist</b>	temporally restricted variant = CDATA [OED] variant date (<vdat>)
<b>pref</b>	preference level (implicit, or "usu" vs. "also" etc.) = CDATA [W7] level (V:4) / level (V:5)
<b>type</b>	(word   affix   phrase   alleged)

*note* regarding word division: Some dictionaries (e.g. [LDOCE], [RH2]) indicate syllable boundaries, whereas others (e.g., [W7]) show only points where words can be broken at ends of lines (i.e., hyphenation points). An easy test is to check words which begin or end with single-letter syllables (e.g., aback, awash, any).

<b>syl</b>	syllabification distance encoding (DIGIT   CHAR)* [LDOCE] embedded codes in headword
<b>hyph</b>	hyphenation distance encoding (DIGIT   CHAR)* [OALD] embedded codes in headword [W7] hyph (F:5,R:3) / hyph (H:4,D:3,F:3)

— This feature is included for compatibility with [LDOCE] —

<b>form</b>	headword form (no. of words, hyphens, etc.) (DIGIT   CHAR) [LDOCE] entry form codes {1-9,D,T,Q,C,Y,X,W,Z} (02:3.1)
-------------	---

#### 4.3.2. P — pronunciation

*note:* IPA vs. other encoding may be recorded using ed or by attributes.

*current usage:*

[OED]: pr / PR

[W7]: P / P

[LDOCE]: 03, 04:3, 10:4

*attributes:*

<b>type</b>	amount of pronunciation given - (whole, prefix, infix, suffix)
<b>dial</b>	regional or social dialectal variant = CDATA [OALD] lab = CDATA
<b>hist</b>	temporally restricted variant = CDATA
<b>pref</b>	preference level (implicit, or "USU <sup>H</sup> " vs. "also" etc.) = CDATA
<b>stress</b>	stress syllable distance encoding = (P   S   U)* P=Primary Stress, S=Secondary Stress, U=Unstressed
<b>syl</b>	syllable distance encoding (DIGIT   CHAR)* [LDOCE] embedded codes in headword

#### 4.3.3. hwd — headword

*note:* The headword is an artifact of existing machine-readable data records. It delimits that set of one or more (e.g. 'A,a') orthographies placed together as the alphabetic basis of an entry. A headword tag is only needed if form and orthography tags do not convey the same information.

*current usage:*

[OED]: hwlem / HW  
 [OALD]: hwlem /hw  
 [W7]: main entry (F:2) / orth (H:5)  
 [LDOCE]: Headword (01:3)

**4.4. M — meaning**

*definition:* Aggregate of all senses. The M tag delimits the block of senses from the other types of information in the entry.

*current usage:*

[OED]: signification (no longer tagged?)

**4.4.1. S — sense**

*definition:* definition (including the text, examples, and related matter)

*current usage:*

[OED]: sen0..sen7 / S0..S7  
 [OALD]: sen / hsn  
 [W7]: D / M  
 [LDOCE]: 08, 18

*attributes:*

**sn** sense number = NMTOKENS  
*current usage:*  
 [OED] num (repeated for each level descended) / <#>  
 [OALD] lab / sn  
 [W7] sense number, s. letter, s. subnumber (D:2-4) / sns (M:2)  
 [LDOCE] definition no (07:2) — doesn't identify subsenses

**pos** sense part of speech = pos-NAMES  
*current usage:*  
 [OED] <ps> / <PS>  
 [OALD] <pos> / ps  
 [W7] part of speech (D:5) / cats (M:3)

**geo** geographic region (e.g., Australia)  
**dom** subject domain (e.g., nuclear physics)  
**regis** register (e.g., colloquial)  
**time** currency or frequency (e.g., obsolete, rare)  
**sem** semantic (e.g., figurative)  
**gram** grammatical code (e.g., transitive)

— The following attributes are used to encode the semantic restrictions encoded in LDOCE's box codes documented to be #7, 9, 10 (but occurring as #5, 9 and 10) —

**nclass** noun class (e.g., slipstream can be used as "Gas")  
**aclass** adjective class (e.g., lambent modifies nouns in nclass "Gas")  
**vclass** verb class — value is a list giving semantic restrictions on the nclass of a verb's subject, direct object, and indirect object

**4.4.2. deftext — definition text**

*usage note:* typically in roman font

*current usage:*

[OED]: —  
 [OALD]: —

[W7]: text definition D:6 / text M:3

[LDOCE]: definition text 08:2, definition text continuation 18:2

*attributes:*

**type** (also | esp | specif | broadly)

#### 4.4.3. ex — examples of form or usage

*current usage:*

[OED]: quot / Q

[OALD]: ex / ex, gx

[W7]: — / X

[LDOCE]: not tagged

*attributes:*

**type** classification of the example = (cited | invented)

*current usage:*

[not present in any MRD]

**geo** geographic region (e.g., Australia)

**dom** subject domain (e.g., nuclear physics)

**regis** register (e.g., colloquial)

**time** currency or frequency (e.g., obsolete, rare)

**sem** semantic (e.g., figurative)

**gram** grammatical code (e.g., transitive)

#### 4.4.4. utext — usage text

#### 4.4.5. xr — cross-reference

*current usage:*

[OED]: xra / XR

[OALD]: xra / xr

[W7]: cross reference X / K

[LDOCE]: usage note text (09:3)

*attributes:*

**hom** xref homonym/homograph number = NUMBER

*current usage:*

[OED] hom = NUMBER

[OALD] <hom> / hn - NUMBER

[W7] homograph number (X:3) / H-sns (K:3)

[LDOCE] —

**sn** xref sense number = NMTOKENS *current*

*usage:*

[OED] num (repeated for each level descended) / <SN>

[OALD] lab / sn

[W7] subscript (X:4) / H-sns (K:3)

[LDOCE] —

**pos** xref part of speech = pos-NAMES

*current usage:*

[OED] <ps> / <PS>

[OALD] <pos> / ps

[W7] unneeded — info carried in hom

[LDOCE] —

**type** (illus | see | also | ... | external)

*current usage:*

[OED] —

[OALD] — / illus=y or type  
 [W7] (X:5) / type (K:1)  
 [UDOCE] X-Ref (09:2)

**sub** sub-part or associated section referenced = CDATA  
*current usage:*  
 [OED] —  
 [OALD] — / <xrc>  
 [W7] secondary word (X:6) / secform (K:4)  
 [LDOCE] —

#### 4.5. RE — related entry

*usage note:* typically in bold appearing within an entry (including run-in, run-on, compounds, idioms, derivatives)

An <RE> can contain any tag defined for <ME>

*current usage:*

[OED]: sube / E  
 [OALD]: entry type=sub / cd or ip  
 [W7]: R / D or F  
 [LDOCE]: 10, or phrases as untagged boldface in 08

*attributes:*

**key** sort key to determine entry's placement = CDATA  
*current usage:*  
 [W7] — / form (D:2,F:2)

**style** placement style = (run-on | run-in)  
*current usage:*  
 unused

**type** relation = ( root | deriv | idiom | compound | phrase)  
*current usage:*  
 [OED] — / —  
 [OALD] type = (main | xref) / <cd> <ip>  
 [W7] / D or F  
 [LDOCE] —

**ref** sense(s) to which this RE is related = NMTOKENS  
*current usage:*  
 unused

**pos** entry part of speech = pos-NAMES *current usage:*  
 [OED] <ps> / <PS>  
 [OALD] <pos> / ps  
 [W7] pt of sp.joiner, 2nd pt of sp. (R:4-6) / cats (D or F:4)  
 [LDOCE] POS (10:5)

**geo** geographic region (e.g., Australia)  
**dom** subject domain (e.g., nuclear physics)  
**regis** register (e.g., colloquial)  
**time** currency or frequency (e.g., obsolete, rare)  
**sem** semantic (e.g., figurative)  
**gram** grammatical code (e.g., transitive)

## 4.6. Synonyms & antonyms

### 4.7. E — etymology

*current usage:*

[OED]: etym / ET  
 [OALD]: not applicable  
 [W7]: E / E  
 [LDOCE]: not applicable

*attributes:*

**type** class of word formation = NAME  
*note:* (reflecting W7's ISV (= Intern. Scientific Vocab.), acronym, biographical, geographical, borrowed word, affixation, unknown origin, etc.)  
*current usage:*  
 [—] not tagged in any MRD

#### 4.7.1. epart — etymology of one variant form of the entry

*note:* (as in W7's among/amongst)

#### 4.7.2. es — etymological segment

*definition:* a unit of derivation in an etymology

#### 4.7.3. eu — etymon unit

*definition:* a package uniting an etymon with its lang, deftext, etc.

*note:* The following are lowest-level tags found particularly in etymologies. To date these have not been tagged in any MRD (except for cf in the OE D)

#### 4.7.4. etymon — word, morpheme, or phrase cited in an etymology

*usage note:* nearly always printed in italics

*attributes:*

**lang** language of the etymon = lang-NAME  
*usage note:* lang attribute inherits its value from previous <etymon>.  
**type** (word | affix | phrase | alleged)  
**gender** (M | F | N)

#### 4.7.5. lang — language = lang-NAME

#### 4.7.6. rel —relation = rel-NAME

*examples:* a., fr., ad., <, :-

Note:

Previous tags <deftext>, <ex> also typically appear in etymologies. The following additional tags might also play a useful role:

#### 4.7.7. cert — certainty

*examples:* prob., ?

#### **4.7.8. basis — basis for etymologist's belief**

*examples:* 'by folk etymology', 'assumed', 'according to'

### **4.8. Additional text tags**

#### **4.8.1. taxon — taxonomic name**

*usage note:* typically in italics.

*attributes:*

**lev** level = lev-NAMES e.g. (K|P|C|O|F|G|S|V)

#### **4.8.2. wd — anaphoric word**

*usage note:* typically represented by swung dash

*current usage:*

[OALD] — / h

*attributes:*

**h** referenced word