Database Design for a Dynamic Dictionary

Frank Wm. Tompa
Darrell R. Raymond

Centre for the New Oxford English Dictionary
University of Waterloo
Waterloo, Ontario
N2L 3G1

1. A database researcher's perspective.

The following definitions from the Oxford English Dictionary capture the fundamental problem in designing a database for a dynamic dictionary:

> **DICTIONARY.**
> **1.** A book dealing with the individual words of a language (or certain specified classes of them), so as to set forth their orthography, pronunciation, signification, and use, their synonyms, derivation, and history, or at least some of these facts; for convenience of reference, the words are arranged in some order, now, in most languages, alphabetical; and in larger dictionaries the information given is illustrated by quotations from literature; a word-book, vocabulary, or lexicon.
> **2. b.** *fig.* A person or thing regarded as a repository of knowledge, convenient for consultation.

Dictionaries have typically been designed as alphabetically arranged books. As such they blend presentational technique and information content to produce external, stable, communicative artifacts.[1] A repository of knowledge, on the other hand, suggests an information resource similar to a computer database. Database design isolates presentational technique from information content to achieve a flexible, dynamic information resource. While we can separate certain components of dictionaries from their presentation (and hence provide the opportunity for dynamic changes of presentation and content), other components are tightly coupled to their presentation. This tension between book and database, between a specific message about the world and its representation in a form that is capable of generating many messages, is the central problem in database design for a dynamic dictionary.

Traditional databases are one step removed from reality. This step is captured in the conceptual model, which is a formal description of the mapping between elements of reality and elements of the database. Databases for dictionaries and other reference works, however, involve two mappings, because the dictionary is simultaneously a model of reality and a reality to be modelled. There is a fundamental difference between queries that address reality as modelled by the dictionary (e.g., ‘‘what is the meaning of the word *chthonic*?’’), and queries that address some model of the dictionary as an artifact itself (e.g., ‘‘how many quotes from Shakespeare are in the dictionary?’’), even though it may be possible to address both with string searching. Preserving the dictionary implies preserving the text as well as preserving its mapping to reality, but this mapping may not be formal enough for traditional database design, despite its utility to the user of the printed dictionary.

Many ideas expressed herein arose from consideration of a database for the *OED*, and many of our examples are drawn from the *OED*. However, this description is independent of any particular dictionary, and it is expected that the dictionary hypothesized is more extensive than any current dictionary. We do not imply that either the Oxford University Press or the University of Waterloo intends to develop the *OED* as described herein.

2. Elements for database design.

In conceptual modelling of an enterprise, it has become customary to identify the *entities* of the enterprise, their *attributes*, and the *relationships* among them.[2] Entities typically represent real or conceptual objects or events that can be distinctly identified. Classical examples revolve around institutional phenomena such as projects, parts, students, and courses, but dictionaries contain other types of entities. Consider the following text used as evidence in the *OED* entry for MACAQUE:

> **1698** FROGER *Voy.* 115 We observed two sorts of Monkeys there [*viz.* Brazil], which they distinguished by the Names of Sagovins and Macaques [Fr. orig. *Macaqs*]...The Macaques are..of a brown Colour.

Some entity types from this sample text include persons (e.g., Froger), published works (e.g., *Relation of a voyage made in 1695-97 on the coasts of Africa*, etc. tr. 1698), subjects (e.g., monkeys, sagovins, macaques), and places (e.g., Brazil). Each entity may have attributes that determine subclasses of that type: for example, persons have names and dates of birth and death, and published works have titles, editions, and publication dates. Finally, relationships are used to associate entities with each other: for example, persons may be related to published works by authorship, subjects may be related to other subjects by hyponymy, and subjects may be related to places by location.

The identification of entities, attributes, and relationships forms the basis of conceptual database design. Such identification is an art, requiring intensive interaction among database engineers, database users (including those who maintain the database as well as those who query it), and enterprise managers and directors. The design is deemed satisfactory if all three groups are satisfied with its ability to describe the data and provide operations on that data. However, we must be aware that user groups are unlikely to propose or initiate design elements for new or potential uses of the dictionary.[3]

In the next sections we will try to identify the major entities and relationships for a hypothetical dictionary.

3. Molecular data.

The first task is to identify the base elements contained in the dictionary. These base elements satisfy two main characteristics: they are not reducible to smaller elements (insofar as the dictionary is concerned) and they are possessed of a unique value that distinguishes them from other elements of the same type. This value is often called the *key* of the entity.

Dictionaries are constituted from text fragments. Unlike databases in which atomic values predominate (i.e., where the data is largely contained in numbers or other non-decomposable values), the basic elements of the dictionary are molecular: text forms are not necessarily sought nor specified by users as units. Instead, typical search requests will retrieve information based on exact match to a string or, more often perhaps, approximate match. Several classes of string represent typical self-contained units; these include lexemes, phonemes, syllables and embedded phrases. Retrieval software must provide access to these particular components, but also allow sophisticated users to define precisely what they mean by ‘‘successful match.’’† A list of molecular data elements appears in Figure 1.

cited text forms	lemmata
	transcribed foreign words
	transcribed pronunciations
	foreign script
	formulae
	special symbols
cited subjects	persons
	places
	events
	institutions
	products
	cited publications

Figure 1. Molecular data elements.

†One drawback in many text retrieval systems is that there is no control over the form of the word being sought: all searches are always for ‘‘near’’ matches using *system-defined* rules governing the forms of plurals, the equivalence of UK and US spellings, etc., which is often inappropriate for finding a particular form of a word.

The entities of a dictionary database must include ‘‘words,’’ as these are the (conceptual) objects for which the dictionary was created. Since ‘‘word’’ is a term that might easily cause confusion, we will call the principal entity in the dictionary design a *cited text form*, that is, a word form that is the object of discussion. Cited text forms are often printed in the *OED* in bold, italics, or small capital letters, or are enclosed in parentheses or quotation marks when the alphabet is roman. Cited text forms include the following:

- lemma (including headword, sub-ordinate lemma, variant, inflexion)
- transcribed foreign word or phrase
- transcribed pronunciation
- foreign script (e.g., Malay)
- mathematical, chemical, or other scientific formula
- special symbol (e.g., mordent)

As well as containing the text string itself, the attributes for a cited text form include the *domain* of the text form (e.g., French, music, chemistry) and its *alphabet*, i.e., the character class for the text form (e.g., Arabic, IPA, mathematical symbols).

It is also useful to subdivide cited text forms into ‘‘singleton words’’ *vs.* ‘‘multiword phrases.’’ This distinction is important for investigating collocation relationships. Such a classification is orthogonal to the distinction between lemmata and other text forms.

Murray comments in his preface to Volume I of the *New English Dictionary* (p. vi): ‘‘the Cyclopædia *describes things*; the Dictionary *explains words*.’’ While dictionaries may not attempt to describe things, they do contain many references to things. These *cited subjects* are typically represented by proper names, and their identification can greatly augment the usefulness of the database. Every cited subject has attributes indicating the *start date* and *end date* of its existence (e.g., birth and death dates). Cited subjects include the following:

- person (in any context, including author)
- place (in any context, including attributive uses)
- event (e.g., the signing of the Magna Carta)
- institution (in any context, including publisher’s or author’s affiliation)
- product (e.g., Dictaphone, Xerox)

Finally, the dictionary also contains *cited publications*. These include primarily printed works, but possibly electronic or oral works that are cited as part of the dictionary’s display of evidence and may be included in the dictionary’s bibliography. Attributes for an entity of type publication include *date*, *title*, and *form* (e.g., newspaper, sermon, novel, dictionary). Note that author is *not* included in the collection of attributes for a publication; rather ‘‘person’’ is a cited subject that can be related to a publication through an ‘‘author’’ relationship.

4. Types of relationships.

Relationships collect molecular elements in certain predefined ways. While the base elements are shared by most dictionaries, it is the relationships between these elements that determines the character of individual dictionaries. Nevertheless, there are classes of relationships that occur in many dictionaries, some of which are prevalent enough to be considered for standardization.[4] A list of potential relationships appears in Figure 2.

aggregate pseudo-entities	dictionaries
	word occurrences
	entries
	artifacts
	quotations
sets	dictionary bodies
	supplementary lists
	aliases
sequences	headwords
	variant pronunciations
hierarchies	significations
	generalizations
	derivations
	bibliographies

Figure 2. Types of Relationships.

4.1. Aggregate pseudo-entities.

A basic relationship is that of the *aggregate pseudo-entity*, which draws together heterogeneous entities to form a virtual entity, often having its own attributes. Aggregates differ from the base entities in that they do not correspond to individual objects or events, but instead represent interactions among (possibly dissimilar) entities.

The simplest such aggregate is the *artifact*, which relates a creator to creation. The components of this aggregate include:

entity class	*role*
cited person	creator
cited publication	object

Artifact

Potential attributes of the artifact relationship include the *date* of creation and *responsibility* of the creator (e.g., authored, edited, translated, photographed, published, submitted, etc.).

A *quotation* is an aggregate of an artifact, relating a cited person to a cited publication, and a string of text used as evidence (i.e., a selection from the cited publication). Likely attributes for quotations include *location* (volume, book, part, chapter, page; act, scene, line; canto, stanza, line; etc.) and *mode* (e.g., heading, dialogue, narrative):

entity class	*role*
artifact	source
text	evidence

Quotation

The major pseudo-entity in a dictionary is the *word occurrence*, which captures all facets of a word related to a single use (commonly known as a *sense* of the word). In principle, each word occurrence comprises the following entities (although some will be omitted or unknown for some entries):

entity class	*role*	*comments*
cited text form	lemma	represents a written form
cited text form	pronunciation	represents a spoken form
text	sense	represents one meaning
text	etymology	includic specifics for this occurrence
set of illustrations	images	graphics depicting the meaning
set of quotations	exemplification	illustrative examples

Word Occurrence

Attributes for a relationship in this class describe the environment in which the particular word occurrence is found:

- *date* encodes the date span appropriate for the word occurrence; ante/circa and other ambiguous date forms are permitted;
- *grammatical class* typically captures the part of speech, but may also be ‘‘combinational form,’’ ‘‘prefix,’’ ‘‘suffix,’’ etc.;
- *usage indicators* includes as many labels as desirable chosen from area (e.g., Australian), currency (e.g., obsolete), grammar (e.g., transitive), register (e.g., colloquial), semantic (e.g., figurative), status (e.g., unnaturalized), and subject (e.g., nuclear physics). The domains of possible usage indicators could be greatly expanded and refined to maintain more semantic and pragmatic information.

A fourth aggregate pseudo-entity is *entry*, which serves to structure identification, morphology, signification, and evidence for a lemma. The major entities and sub-relationships in an entry for a comprehensive historical dictionary might include:

entity/relationship class	*role*
sequence of cited text forms	headword
sequence of cited text forms	pronunciation
set of cited text forms (with dates and usage labels)	variant list
editorial text	etymology
set of usage labels	usage indicators
hierarchy of senses (with usage labels) plus quotations	signification
hierarchy of other entries for sub-ordinate lemmata, etc.	derivatives

Entry

Each entry has as attributes the *date* of its creation, and its *skeleton*, which indicates the precise order and format for presentation.

4.2. Sets.

A *set*, in database terminology, is an arbitrarily large, unordered, homogeneous collection of objects. For example, a usage indicator (e.g., ‘‘military, colloquial, chiefly US’’) is a set because its elements are chosen from an unbounded domain, it has little or no internal structure, and it can be ordered as deemed convenient. Sets are manipulated and queried using operations derived from mathematical set theory, including set union, intersection, and difference (e.g., ‘‘Find word occurrences that are used chiefly in the US military and do not relate to ships or shipping’’). Although the print medium imposes order on every collection of objects in a traditional dictionary, for some collections the order is immaterial or derivable from the values of the objects themselves (e.g., alphabetical ordering).

A set of entries constitutes a *dictionary body*. Other parts of the dictionary are also sets: the *abbreviations*, *external consultants*, *bibliography*, and *forms of address* are just a few examples of sets, ordered by value for user convenience.

A class of set that is important for retrieval is that of *aliases*.[5] Cited text forms, subjects, and publications can each be referenced under multiple names, which make up a set of aliases for the entity. The most important set of aliases for cited text forms is the set of variant forms, including abbreviated forms and transcriptions. Cited subjects also require alias sets for maintaining alternative names (e.g., *Bruxelles*/*Brussels* or *S.Clements*/*M.Twain*), name revisions, or abbreviated forms. Finally, cited publications may also have aliases to capture alternative editions, translations, name revisions (e.g., *Busy man's Magazine/Maclean's magazine*), and abbreviated forms. An attribute *mode* can be associated with each member to indicate the type of aliasing.

4.3. Sequences.

Some collections of elements are more properly represented by *sequences*, i.e., sets in which a defined ordering is preserved. For example, the collection of headwords for a single entry, representing multiple acceptable spellings, are often ordered by editorial preference, which (unlike alphabetical ordering) is not derivable from the lemmata themselves. Similarly the pronunciations associated with the headwords of an entry might be better represented by a sequence of forms than by an unordered set.

When modelling existing dictionaries, it is often difficult to distinguish between intended and unintended sequence. Unintended sequence is an ordering of information which is purely a result of the print-induced sequential presentation of text. If such a collection of text is represented by a set, an important, but as yet undetected piece of information captured by the ordering would be lost. If a set is represented by a sequence, convenient reordering will be ruled out and there is a possibility that an incorrect inference will be made about the ordering rule.

4.4. Hierarchies.

Several dictionary structures are based on nesting. A *hierarchy* structures arbitrarily large, homogeneous collections of elements such that each member is classified according to some ‘‘containment’’ property defined among the members. In many hierarchies, each component is also positioned in some order relative to the other members of its partition. Hierarchies share with sequences the characteristic that the relative position of members is not derivable from the members themselves, but must be specified.

The major example of a hierarchy in a dictionary is the *signification* of entries. A signification is an organized collection of senses, typically arranged historically or by frequency of use, each sense containing arbitrarily many sub-senses and perhaps augmented with graphics or illustrative quotations. Often the senses of an entry are numbered to reflect their positions; such numbering, however, is a result of the order, not an inherent value from which order can be deduced.

Two other hierarchical structures are important to the sense structure of a dynamic dictionary. The first is *generalization*, which relates word occurrences using hyponymy as the basis of containment.† For example, ‘‘cat’’ and ‘‘dog’’ are within ‘‘domestic pet’’ which, together with ‘‘barnyard animal’’ *et al.*, are within ‘‘animal.’’ The second hierarchical structure is *derivation*, classifying word occurrences according to morphology (e.g., ‘‘mace-bearer’’ derives from ‘‘mace’’ and ‘‘bearer’’ which, in turn, derives from ‘‘bear’’). This structure illustrates an additional complexity possible in a hierarchy: a word may be derived from more than one other word occurrence, thus there is more than one ‘‘parent’’ in this containment relation.

Other hierarchies may also be incorporated in a dictionary. For example, the bibliography of cited publications may be better represented as a hierarchy of works (e.g., showing that *The Merry Wives of Windsor* is contained in Shakespeare’s *Comedies, Histories, and Tragedies*) than as an unordered set. The choice of hierarchy *vs.* set must be made depending on the relevance of the information conveyed by the structure and not recoverable from the information associated with each individual member.

5. Amorphous components.

Molecular elements have precise values and are organized in specific, definable relationships. However, dictionaries also contain components which do not satisfy these conditions. Such components as discursive text, illustrations, tables, cross-references, and implicit aggregates are *amorphous* in nature rather than molecular. Amorphous components typically exhibit a weak distinction between structure and value (alternatively, a strong connection between presentation and information content). As a result, we cannot break amorphous components into entities and relationships.

---

† Such structures are often called ‘‘IS-A hierarchies’’ in database modeling.

The value of an amorphous component is usually not confined to a number or a short, precise text string. Instead, its value is derived from (but not reducible to) values of molecular elements and a subjective value that arises from a particular arrangement of those elements. In particular, it is often the case that the component remains the same if some of the elements are changed or left out (though its worth may be somewhat degraded). For example, definitions can withstand a significant amount of rewording or abbreviation before the meaning is lost. Similarly, tables are still useful even though some of their entries might be incorrect.

Amorphous components are drawn from an infinite variety of expression, and editors are not likely to have constrained themselves in their use. Indeed, it is the ability to employ variety in an artistic fashion that has prompted the use of the amorphous element. This suggests that users should not be expected to enter the value of an amorphous component in order to retrieve it. For example, users are unlikely to specify the content of a definition in order to find the word that it denotes, or to draw an illustration in order to retrieve it.

Yet while the value of an amorphous element is variable, it is not random; like intended sequence, it was chosen by an editor because it was the most suitable, perhaps the optimum, value for that particular component. Users who are interested in the dictionary in its own right will certainly want such subjective judgements to be preserved, and editors will need access to the presentation of the amorphous element during updating or revision. As a result, we must confront the fact that amorphous components contain an irreducible element of presentational, scholarly, or artistic merit, which can neither be ignored nor flexibly replaced with other presentational forms.[6]

In a dictionary, the primary class of amorphous data is *discourse*. This class includes text that is composed of unconstrained phrases and sentences. Subclasses of discourse include:

- *editorial text* including ‘‘etymology’’ and ‘‘sense’’ text;
- *evidence*, i.e., segments of text used to illustrate the senses of words, either extracted from external publications or generated by editors specifically for the dictionary.

The second class of amorphous data is *illustration*. This class includes non-textual material used to elaborate entries in the dictionary. There are three subclasses:

- *artwork* (e.g., line drawing, sample of type style) typically represented by discrete components;

- *image* (e.g., photograph, audio signal) typically represented by an approximation to a continuous image from the ‘‘real world’’.
- *tables*

Tables are particularly interesting because they are closer than other illustrations to text, yet are certainly neither sequences nor sets nor hierarchies. Other writers have commented on the special status of tables in document design.[1, 7, 8]

A third class of amorphous data is *implicit aggregates*. These are aggregates which may be defined in the real world but are not explicitly captured in the data model implicit in the dictionary. Typically they are pseudo-entities maintained not by the dictionary but by users who have invested some effort in defining them. Consider for example the set of all passages from the Bible that are cited in the *OED*. A typical citation looks like the following:

> **1611** BIBLE *Gen* xxvii. 40 Thou shalt breake his yoke from off thy necke.

OED citations contain a date, an author, a work, and the text of the citation. Here the ‘‘author’’ is given as Bible (implying the King James Version), largely so that the work field can be employed to indicate the relevant book of the Bible.

If the value ‘‘Bible’’ for the field ‘‘author’’ were sufficient, we might refer to the collection as a molecular set. There are some 5700 occurrences of the string ‘‘Bible’’ within the author field, but these are not all the Biblical citations. For example, in 340 cases ‘‘Bible’’ is the value of the cited publication instead of author. The problem is exacerbated by distinctions among various translations and publications, such as Hyll, Bagster, and Bishops’. Furthermore, many of these variants do not contain the string ‘‘Bible’’ anywhere in the citation: some estimates of variants not explicitly indicated as ‘‘Bible’’ are given in Figure 3.

Variant	Number of Occurrences
Wyclif	8000
Coverdale	4200
Tindale	2000
Nisbet	100
Paues	55
Geneva	32
Cranmer	28
Revised	17
Rheims	14
Great	11
Matthew	4
Tomson	3

Figure 3. Varieties of Biblical citations.

Further compounding our uncertainty is the variety of spellings (Tindale/Tyndale) and abbreviations (the Nisbet is variously given as *New Test. in Scots, N. Test. in Scots, N.T. in Scots, N.T. Scots, N. Test.*, and *N.T.*).

The notion of a cited publication is modelled adequately by most dictionaries. But the notion of a *Biblical* citation is not explicitly captured, though it seems like a valid real-world entity and can be approximated. Many of the translations given above were located because one or two citations had both the translation name and the qualifying word ''Bible'' present; searches were then performed for citations that used the translation name but did not contain the string ''Bible.'' These sets were reduced by examining the work field to see if the citation was likely to be from a book of the Bible. Such transitive relationships involve probabilistic inferences and hence are difficult to model.

In general, determining implicit aggregates requires specialist knowledge of the subject area for comprehensiveness and specialist knowledge of the dictionary to anticipate the many variations in spelling, abbreviation, and markup. While good approximations can be achieved by persistent effort, definitive answers are not generally possible. The amorphous nature of the aggregate is a direct result of the information loss in mapping from real-world to dictionary to database. This transitive mapping of implicit aggregates is common not only to dictionaries but also to most other reference texts.

A final amorphous component is the *cross-reference*. Consider the examples drawn from the *OED* shown in Figure 4. Cross-references seem as though they might be molecular relationships, connecting two entries, senses, or quotations. Nevertheless we have classified cross-references as amorphous components primarily because they are expressed in discursive text.

The free-form nature of cross-references leads to two amorphous aspects. First, it is often difficult to determine the scope and range of the cross-reference, that is, the extent of the source and target that are involved in the relationship.[6] As a result, the entities involved in the cross-reference are themselves amorphous. Second, it is difficult to accurately identify whether the source, the target, or the relationship is the most durable element of the cross-reference.[9] For example, when the dictionary is being updated, it is not sufficient to simply modify all sources to point to the new locations of their targets. It may be that the target has been completely removed, or that the relationship is no longer considered valid (perhaps because of developments in lexicography). As is the case for implicit aggregates, cross-references involve real-world entities which are not well modelled by the dictionary, and hence they are amorphous components.

Cross-references become even more problematic when they interact with each other. The etymology in the *OED* entry for PRIMROSE contains a cross-reference to its own trailing note, which has a cross-reference to PRIMULA's note

STRIATAL see STRIATUM
STREUTH var 'STREWTH
WARRE obs. f WAR and WARE (sb., a., and v.)
STRIATION (2b.) *Electr*. = STRIA 2d (in sing.)
STRETTO B sb. b.*stretto maestrale* [cf. MAESTRALE] (see quot. 1946).
MIN sb. 3 shortened form of MINUTE sb.
STRIDE 7. Ellipt. for *stride piano*
STRETCH-OUT [f. STRETCH v. + OUT adv.]
KNOWLEDGE v. see also the sb.
WILD see also Special Collocations (16), wild cat, fowl, goose in the main series.
RECRAY see also next
DAY 6 b. see also the various qualifying words
WARE 2. see quots. and s.v. the first element ... and others mentioned in 3
PUDDING 6 c. Also *Christmas pudding (Christmas 4), Sussex pudding, Yorkshire pudding* (See also these words.)
FLY-BY-NIGHT 1. See also quot. 1796
STROKE Perh. a misprint for *noke, nook sb.* (where see senses **3 d 3**).

Figure 4. Varieties of cross-reference forms.

(and PRIMULA has a cross-reference in its own etymology to its own note). In dynamic situations it might seem easier just to rewrite the entries than to attempt to modify the structure to maintain the cross-reference.

6. Implementing the model.

Once entities, attributes, and relationships have been defined, the next step is selection of a data structure to represent these components. In traditional databases, the choice of data structure is usually confined to one of network, hierarchical, or relational data base, but it is the existence of amorphous components that makes such structures inadequate for text. Two current approaches to structuring text are SGML-style markup[10] and hypertext.[11] Neither of these approaches are sufficiently developed to support all the needs of a dynamic dictionary. SGML, for example, fails to distinguish sets from sequences, is unable to represent multiple fields which are not hierarchical (e.g., the etymology for PORE† contains the overlapping derviational segments ‘‘Sp., It. *poro*, ad. L. *porus*’’ and ‘‘L. *porus*, a. Gr. πόροζ ’’), and is unable to mark aggregates of elements that do not occur contiguously in a text (e.g., the sets of rhyming lines in a poem).[12] Hypertext addresses some of these problems but shares with SGML the fundamental

† [a.F. *pore* (*porre*, 1312 in Hatz.-Darm.) = Sp., It. *poro*, ad. L. *porus*, a. Gr. πόροζ passage, pore.]

characteristic of explicit fragmentation of text. This fragmentation is expensive and of questionable utility;[6] more importantly, fragmentation always risks the destruction of amorphous components.

The choice of data structure to represent the dynamic dictionary is still an open problem. Currently we structure the text of the *OED* with SGML-style markup, despite the fact that it is formally inadequate to handle amorphous components (nor are all of the molecular components and relationships identified herein represented in our text). Nevertheless, we often deal with amorphous components, typically through a combination of powerful string searching capabilities, ingenuity, and knowledge of the text. Although these techniques have often provided satisfactory answers, we always experience a vague uneasiness about the process. It is unreliable, because one can never claim to know all the variants. Furthermore, ingenuity and dictionary-specific knowledge are hard to transmit to other users. Finally, without an explicit statement of what data is actually modelled, users often become confused about what can be legitimately queried.

Once the data structure has been chosen, a final step in database design is development of a set of operations for accessing and manipulating the data. Possible operations for the dynamic dictionary include:

Query and extraction functions:

- *browse*: pose queries against the dictionary iteratively, chaining through the relationships recorded in the database; interactively search for strings or patterns of interest throughout the dictionary; users might require displays highlighting structure as well as content and might require various degrees of formatting and typesetting codes respected; some users need access to ‘‘published’’ materials only, while others require access to ‘‘superfluous files,’’ materials in preparation, editorial notes, etc.
- *extract materials*: produce formatted reports summarizing some or all data related to particular entities for extensive study; requests can be based on specific content (e.g., medical terms), on missing information (e.g., no supporting evidence after 1880), or on stylistic forms (e.g., inclusion of ‘‘in recent usage...’’); reports can be used to service special-purpose requests from end-users, to obtain hardcopy form of working materials, or to check the accuracy of materials and the attractiveness of presentation.
- *interface*: access other databases to support browsing or extraction efforts.
- *consult*: query other users, including lexicographers and expert consultants, and reply to queries from other users.

Composition and maintenance functions:

- *assemble evidence*: sift through available materials, categorizing, ordering, and re-ordering evidence.
- *compose new entries*: create and edit text; select supporting evidence; insert cross-references to other entries and relationships to other entities.
- *edit materials*: update contents and structure of existing entries; augment materials to include new information about existing entities (e.g., adding hyphenation or synonym information to the *OED*).
- *annotate*: make private notes on entries, evidence, etc. for later review; leave written messages for others; record the editorial status of materials.
- *standardize*: impose consistency constraints on materials (e.g., check for house style, ensure symmetry in synonym sets).

7. Conclusions.

Data modelling inevitably involves philosophical questions about the nature of entities and their relationships. Such questions are always open-ended and their solutions subject to criticism. However, in traditional applications it is more important to resolve the questions than to defend their resolution. In a dynamic dictionary (and, we suspect, for many other scholarly works), the criteria by which the questions are resolved is at least as important as the actual resolution; furthermore, such criteria are likely to become database components themselves.

A meticulously crafted book is evidence of the value of interweaving presentation and representation in a single, inseparable whole. A meticulously crafted database is evidence of the value of separating presentation and representation, achieving flexibility in both. Addressing the tension created by these two incompatible forms is the key step in designing the dynamic dictionary of the future.

Acknowledgements

The research for this paper began during an extended visit to the Oxford University Press. The ideas became focussed primarily through extensive discussions with Edmund Weiner and John Simpson, co-editors of the Second Edition of the *OED*, as well as through continual fruitful exchanges with members of the Data Structuring Group at the University of Waterloo. Nevertheless, the contents of this paper are solely the responsibility of the authors and not of the Oxford University Press nor the University of Waterloo. Financial support from the University of Waterloo, the Office of the Canadian Secretary of State (through the Centres of Specialization Fund), and the Natural Sciences and Engineering Research Council of Canada (through grant A9292 and 0039063) is gratefully acknowledged.

8. References

1. Levy, David M., ‘‘Topics in Document Research,’’ *Proceedings of the ACM Conference on Document Processing Systems,* pp. 187-193 (December 5-9, 1988).
2. Kent, William, *Data and Reality: Basic Assumptions in Data Processing Reconsidered,* North-Holland Publishing Co., New York (1978).
3. Raymond, Darrell R. and Frank Wm. Tompa, ‘‘The Limits of User Consultation in Database Design,’’ *Canadian Journal of Information Science,* **12**(3/4) pp. 98-106 (1987).
4. Amsler, Robert A. and Frank Wm. Tompa, ‘‘An SGML-Based Standard for English Monolingual Dictionaries,’’ *Proceedings of the Fourth Annual Conference of the UW Centre for the New Oxford English Dictionary,* pp. 61-80 (October 26-28, 1988).
5. Lesk, Michael, ‘‘‘‘They Said True Things, But Called Them By Wrong Names’’ — Vocabulary Problems Over Time in Retrieval,’’ *Proceedings of the 4th Annual Conference of the UW Centre for the New Oxford English Dictionary,* pp. 1-10 (October 26-28, 1988).
6. Raymond, Darrell R. and Frank Wm. Tompa, ‘‘Hypertext and the Oxford English Dictionary,’’ *Communications of the ACM,* **31**(7) pp. 871-879 (July 1988).
7. Beach, Richard J., ‘‘Setting Tables and Illustrations with Style,’’ CS-85-45, Department of Computer Science, University of Waterloo, Waterloo, Ontario (May 1985).
8. Levy, David M., ‘‘Multiple Decomposition and Description of Documents: The Dependence of Document Structure on Use,’’ unpublished technical report, Xerox Palo Alto Research Centre, Palo Alto, California (February 27, 1989).
9. Warburton, Yvonne L. and Darrell R. Raymond, ‘‘Resolving Cross-References,’’ unpublished technical report, Centre for the New Oxford English Dictionary, University of Waterloo (March 1989).
10. ISO, ‘‘Information processing — text and office systems — Standard Generalized Markup Language (SGML),’’ ISO 8879-1986, International Organization for Standardization (1986).
11. Conklin, Jeff, ‘‘Hypertext: An Introduction and Survey,’’ *IEEE Computer,* **20**(9) pp. 17-41 (September 1987).
12. Raymond, Darrell R., ‘‘Reading Between the Tags: An Appraisal of Descriptive Markup,’’ unpublished technical report, Centre for the New Oxford English Dictionary, University of Waterloo (June 1989).