1
Data Semantics on the Information Superhighway

D. Beech
Oracle Corporation
500 Oracle Parkway, Redwood Shores, CA 94065, USA
dbeech@us.oracle.com

Abstract

Major advances in data semantics are needed to keep pace with the generalized forms of data now becoming widely available on the information superhighway. We begin by describing a system that treats character string data as genuine text in a natural language, and adds the level of semantic understanding provided by thematic analysis. Extensions of conventional database type systems and query paradigms are needed for handling text and document and multimedia information, and are discussed especially in relation to SQL. Finally, seeking a theoretical foundation of sufficient breadth for future work in data semantics, we explore the semiotics of the philosopher Charles S. Peirce.

Keywords

Data semantics, information superhighway, thematic analysis, ConText, database systems, SQL, multimedia, signs, semiotics, Peirce.

1 INTRODUCTION

1.1 Historical perspective

While a title that juxtaposes Data Semantics and the Information Superhighway may seem to be straining after topicality, the intention of this paper is to explore some of the long-term implications of their interaction. The two phrases share a certain optimism about the possibility of achieving in the future what is currently beyond our grasp. While this optimism can degenerate into marketing slogans, or may be characterized as naïveté, it is essential to progress and is one of the more engaging characteristics of our profession.

R. Meersman et al. (eds.), *Database Applications Semantics*

We take the sense of *Data* in Data Semantics to imply a note of caution or pragmatism, a guarded optimism. Data Semantics might be interpreted here as 'Practical semantics in database systems'.

So where is the caution in speaking of a Superhighway? The answer must be that the term Highway has become debased, corresponding to a frustrating and dangerous means of travel, and that Superhighway is now a neutral term for a reasonably acceptable and safe model of what we should be aiming for in the future conveyance of information.

The thesis of this paper may be stated more fully as follows:

> The rapid increase in variety and accessibility of information on the Information Superhighway presents major opportunities, and corresponding challenges, to the field of Data Semantics.

Accessibility to an early form of the information superhighway is already available to millions of people via computers, and will expand many times more when new set-top units enable television sets to serve as the interface instead of full-fledged computers.

A single author cannot hope to be expert throughout so wide a field, and what follows will be an attempt to share some exploratory experiences in broadening from a conventional database view of the world to a more general treatment of information. It is in the nature of a personal odyssey or, to be less anachronistic, a hitchhiker's guide to the galaxy.

We will first offer a perspective on some technological achievements of the past, in order to suggest the magnitude of the historical opportunity that we have in this decade. Some of the specific challenges will then be identified.

Discussions of the social impact of science and technology often begin by considering gunpowder, the printing press, and the discovery of the circulation of the blood. However, we will focus our somewhat selective view in Figure 1 on the 19th through 21st centuries.

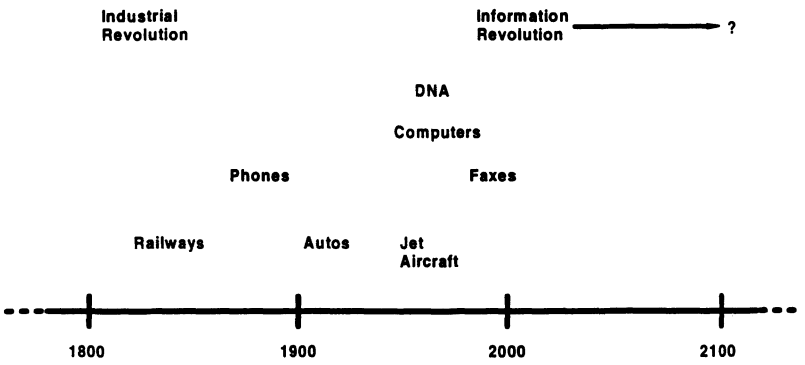

Figure 1 A historical perspective of technologies with widespread social impact

The effects of the Industrial Revolution have already spanned nearly two centuries, and may continue indefinitely. The Information Revolution now gathering speed is likely to be seen by future historians as a phenomenon on a similar scale.

Other technical breakthroughs in transport and communications have had enormous social consequences during the past two centuries. Although faxes may not be so vast a technical achievement as others that are shown, they have been included in the figure because of the speed with which they have been adopted and have changed people's lives, in a way that foreshadowed the no less dramatic growth in communication via the Internet and especially the World Wide Web.

The arrival of the computer and the revelation of DNA stand out as peaks in the mid-twentieth century. They have already had significant effects, but we are still only in the early stages of the way they will transform human life. In the case of information processing, we are living at a time of great opportunity, but of equally great responsibility and challenge. If we do not measure up to the challenge, there is the likelihood that the shortcomings of this decade will be perpetuated well into the next century, and our successors will not look kindly on us.

1.2 Data semantics challenges

Without attempting to be totally comprehensive, let us at least begin by indicating the breadth of the requirements for progress in data semantics. Here are some of the problems that need to be addressed:

Information overload

Communications have improved to the point where information overload is the problem of the age. On the Information Superhighway, users can all too easily be deluged with information they do not want. Even with traditional published materials, the quantities are now so great that a national conference of librarians is this summer debating 'The Crisis in Cataloging' – a crisis of overload without even passing beyond the title page.

As in Jorge Luis Borges' story 'The Library of Babel' (Borges, 1983), all knowledge is contained in the library (or on the Internet), but one can find nothing:

> 'When it was proclaimed that the Library contained all books, the first impression was one of extravagant happiness ... As was natural, this inordinate hope was followed by an excessive depression.'

Maybe the results today are no worse than they have been with clumsy old-fashioned access to information in the past, but the frustration level is greater:

> 'The certitude that some shelf in some hexagon held precious books, and that those precious books were inaccessible, seemed almost intolerable.'

This paper will discuss some of the semantic responses to the challenge of information overload, such as automated thematic analysis, filtering and ranking of documents, and the use of intelligent agents.

Heterogeneity of semantic models
Semantic heterogeneity is a major issue that must be mentioned here, although it will not be addressed in this paper, since it is the topic of another paper in this volume (Wiederhold, 1995).

Communication with non-technical consumers
The opening up of on-line information to large numbers of non-technical users raises its own challenges of matching the user's and the system's semantic models, and we shall attempt to at least lay some foundations for progress in this direction.

Object paradigm
The object paradigm in programming and database systems has, of course, considerable semantic utility. In principle, the type of an object specifies the operations that may be performed on the object, i.e. its behavior or semantics, although in practice this definition is hard to achieve at a convenient level of abstraction, and at present the semantics must mostly be prised out by reading the code that implements the operations. We shall mention some uses of object concepts in SQL, in document markup languages, and in multimedia scripting languages.

New forms of structure and relationship
Turning to the treatment of information structure, there are several active areas. Later in this paper, we will briefly discuss markup languages, compound objects and documents, and hyperlinks and ostension.

New media
The widening of the scope of data beyond the alphanumeric world to a world that engages all the senses will cause us to consider sensory domains such as the audible, olfactory, and tactile. The visual domain is more thoroughly treated elsewhere in this volume (Foley, 1995).

Theoretical foundations
In searching for adequate theoretical foundations, we shall explore the potential of semiotics (the theory of signs).

1.3 Symphony in Three Movements

The substance of the paper will be contained in the next three sections. Rather than attempting a monotonic progression, we will offer a slightly more dramatic or symphonic structure. The two outer movements will contain the more taxing ideas, being less familiar to database researchers, while the middle movement will provide some earthbound light relief:

I. Textual semantics – *Theme and variations*
II. Generalized databases – *Allegro barbaro*
III. Generalized semantics – *Mysterioso*

2 TEXTUAL SEMANTICS – *Theme and variations*

Vast quantities of textual information reside in file systems, since database systems have traditionally been aimed at structured data in shorter segments. Now, however, we see a trend towards better integration of textual and other data in database systems, so that text can benefit from the wide range of management facilities of the database system and from the unification of query and update operations across text and other data types.

Before discussing the implications for the familiar database semantic model expressed in SQL, we will devote this section of the paper to some of the linguistic aspects of unstructured text. There is a spectrum of possible approaches, increasing in sophistication from the literal matching of individual words to the complete semantic understanding that would be exhibited by a successful artificial intelligence system. Existing text processing systems have progressed some way from the low end of the spectrum by supporting wildcards and stemming, proximity and phrase searches, and soundex and fuzzy matching. A more strongly semantic element is introduced by matches that can use a thesaurus to recognize related words. Yet the user of these systems still has to work hard to express appropriate criteria to find documents that are semantically relevant, and that are ordered according to their degree of relevance. Another kind of current information retrieval system converts semantics into statistics. If the system is trained by telling it which of a large number of sample documents satisfy which queries, it will recognize similar statistical patterns of word frequencies in new documents and categorize them accordingly.

The approach that we want to discuss here is somewhere in the middle of the spectrum. It is by no means a complete semantic approach for natural language, but it does represent a potentially significant step forward from the user-driven or statistical analysis of most current approaches. The key feature is that the system carries out thematic analysis, and attempts to grasp just enough of the semantics to say what various things a document *is about*.

In order to show that the thematic approach is now practicable, we will give an outline of how it is carried out in a running system known as ConText (Oracle,1995a).

2.1 Utility of thematic analysis

Quite apart from the direct use of thematic analysis to support queries for topical browsing ('select each document that *is about* this theme'), this macro level of semantic understanding can be used in a number of other applications.

Speed reading of documents can be aided by highlighting words and phrases related to the major themes, with some user control over how the highlighting is applied. For example, only the n most important themes might be highlighted, in the same or different colors. For a given theme, there could also be control of the thresholds of relevance that determine whether given words should be highlighted.

A variant of highlighting is the construction of digests of a document – compressions formed by omission of material rather than freshly written summaries. The thematic analysis has identified the most important themes that should be retained, and the sentences and parts of sentences that contribute most to them. Here again the user can be given some control, e.g. over the ratio of compression.

A more structured use of semantics is seen in the automatic generation of back-of-book indices. The thematic analysis helps to identify parts of the text that need to be indexed, even where different words are used, and the index itself, in conventional style, contains indented subentries a few levels deep that reflect semantic relationships to their containing entries.

In the routing of documents, statistical methods of categorization can be replaced by thematic methods where the latter prove to be more effective. The training of a system can take the form of teaching it to recognize the major themes and their relative importance for each category of document in the training sample. New documents can then be matched against the thematic patterns.

2.2 System architecture

A brief overview of the ConText architecture (see Figure 2) will help to clarify where syntactic issues shade into the semantic ones that are the focus of this paper.

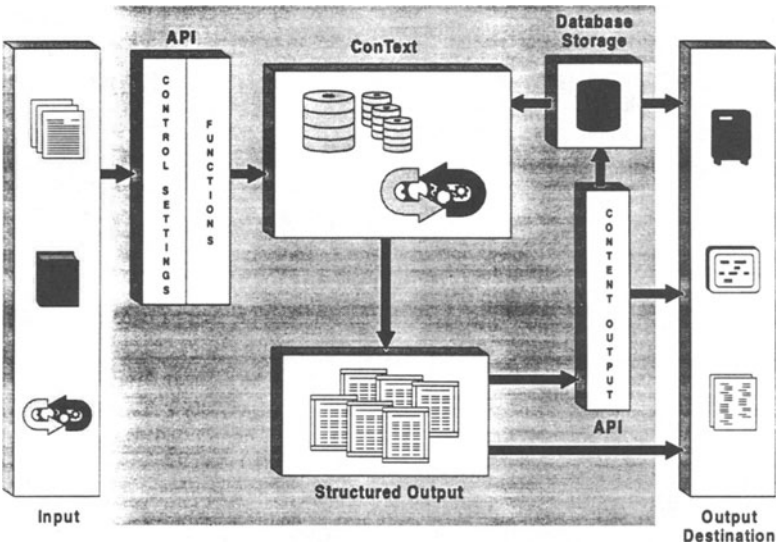

Figure 2 Architecture of the ConText system

The system is not intended as a self-contained product, but as an engine that can be used in various ways. It can be parameterized by front-end control settings, and produces large quantities of structured output to be massaged appropriately for various applications.

The linguistic engine itself (see Figure 3) begins with the syntactic parsing, and, even at this early stage of identifying parts of speech and sentence structure with the aid of the Lexicon, can lay some thematic foundations. The content extraction phases complete the thematic analysis by using the Knowledge Catalog, which contains the ontology, a classification hierarchy of concepts.

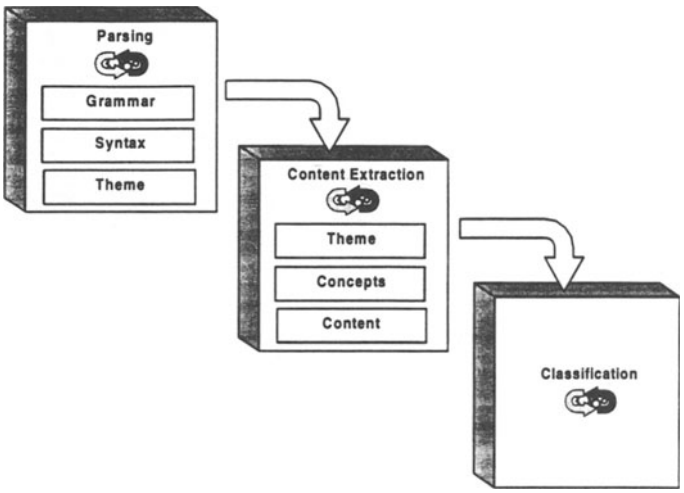

Figure 3 Linguistic engine

The Lexicon and Knowledge Catalog between them describe the morphology of a particular language, currently English (see Figure 4). For present purposes, we are mainly concerned with the concepts described in the Knowledge Catalog, and in showing how the notion of an ontology, familiar in data semantics already in its application to structured data, is employed in the thematic analysis of unstructured natural language text.

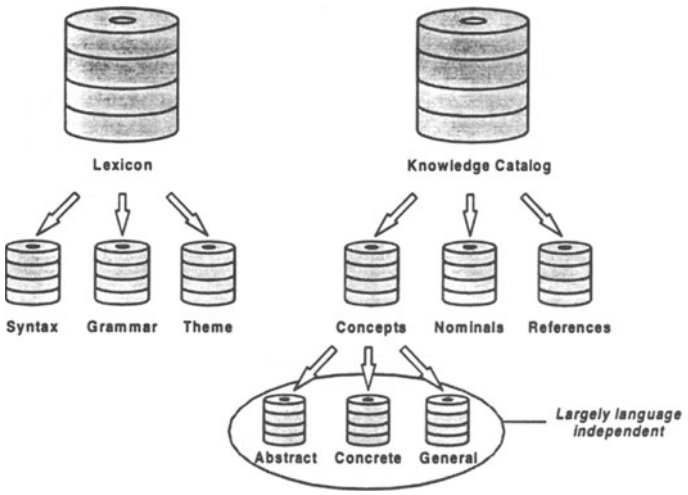

Figure 4 Morphology

ConText is implemented by a variety of complex and eclectic linguistic techniques, rather than by relying on having found some simplifying philosopher's stone. The system is rule-driven, and the rules are iterated over until (usually) a best-fit point of equilibrium is found. The rules are effectively a set of constraints that must be satisfied by a successful parse, and hence this form of grammatical specification may be termed a *constraint grammar*.

There is an analogy here with those chess-playing programs that succeed by pragmatism rather than purity, and indeed the English language has now been explicitly shown to be even more grammatically complex and impure than had been feared. In a project led by Professor (later Sir Randolph, and now Lord) Quirk of University College London over the course of nearly 20 years, the language was scrutinized so thoroughly as to produce *A Comprehensive Grammar of the English Language* (Quirk *et al*, 1985) that runs to 1779 pages. Thus in order to handle on-line documents ranging from the formal to the colloquial, a parser is well-advised to embody all of the general principles and special cases contained in this definitive work. It turns out that such a parser needs a set of about 20,000 rules, as compared to perhaps 2,000 in a less complete research system.

To give some idea of the scale of the accompanying lexicon, the ConText Lexicon for the English language currently contains about 600,000 entries, with over 700 pieces of grammatical or thematic information for each entry.

In the Knowledge Catalog, each concept is nominalized (i.e. represented by a noun form). There are currently about 1,500 categories, which may be composite general themes such as *genetic engineering* or *haircare industry*.

While the grammatical rules of a natural language are relatively stable, the vocabulary and the senses of words and phrases are continually being extended. Hence the Lexicon and the Knowledge Catalog need to be readily extensible. The system itself can be learning while it is processing text, adding new words to the Lexicon, and their apparent conceptual roles to the Catalog where these are clear enough. Beyond this, provision is made for explicit user inputs to both Lexicon and Catalog.

2.3 Theme extraction

The process of theme extraction is founded on identifying the following (when present) for each sentence:

- topic
- focus
- action.

A clear and entertaining exposition of this approach may be found in Williams (1990).

The *topic* is the psychological subject of a sentence (often being also the grammatical subject, and coming first in the sentence). The topic is often carried through from one sentence to another, and the *focus* is then the new information that a sentence conveys about the topic. However, when a topic is newly introduced, topic and focus coincide. The *action*

serves to relate the topic and focus. In general, topic, focus, and action are taken respectively as the primary, secondary, and tertiary themes of a sentence.

Theme grading in ConText is applied to sentences in isolation, with information being collected per word for up to 16 different views of the sentence. *Theme strength* is then computed per word, with adjustable weights being given to the different views.

This leads to the production of *theme vectors* per sentence, paragraph, and document. A theme vector retains the top 16 themes, with their relative strengths normalized as fractions of 255. Each theme is tied to a category from the Knowledge Catalog, as shown in Figure 5, where the categories are in angle brackets.

Treasury bill yields dropped substantially Friday morning in anticipation of further reductions in the federal funds rate by the Fed, market watchers said. The three–month bill fell 8 basis points to a discount equivalent rate of 7.70 percent, while the 1–year bill was down 12 basis points to 7.30 percent.

Paragraph theme vector

	Strength	Theme
1:	43	banking <finance and investment>
2:	25	basis points <stocks, bonds, and commodities>
3:	24	treasury bill yields <banking>
4:	22	stocks, bonds, and commodities <finance and investment>
5:	22	points <stocks, bonds, and commodities>
6:	21	yields <banking>
7:	17	bills <bills>
8:	12	federal funds rates <banking>
9:	11	reductions <banking>
10:	10	rates <banking>
11:	9	discount equivalent rates <commerce and trade>
12:	9	three–month <three–month>
13:	8	1–year <1–year>
14:	8	rates <commerce and trade>
15:	7	discounts <commerce and trade>
16:	7	equivalents <equivalencies>

Figure 5 Example of paragraph themes

Note that some of the categories acquire enough weight to be shown as themes themselves, and others do not. For example, <banking> becomes the top theme, above specific themes belonging to that category, while <stocks, bonds, and commodities> comes below the theme 'basis points' belonging to that category, and <finance and investment> is too general to be shown as a theme itself.

With theme vectors at the document level, there is a natural tendency for the more general themes to acquire cumulative weight through recurrence, corresponding to intuitive notions of what one would say the document as a whole was about.

Figure 6 illustrates the top of the category hierarchy in the Knowledge Catalog.

Science, Technology, and Education
- communications
- education
- hard science and technology
- social sciences
- transportation

Business and Economics
- business and industry
- economics

Government and Military
- government and law
- military

Social Environment
- belief systems
- clothing and appearance
- family
- food and agriculture
- home
- leisure and recreation

Geography
- cartography
- political geography
- physical geography

Abstract Ideas and Concepts
- abstract relations
- animate actions and states
- metrics
- physical properties and relations

Generic Classification

Figure 6 Top of Knowledge Catalog

3 GENERALIZED DATABASES – *Allegro barbaro*

In order to show more clearly the relevance and feasibility of the longer-term advances in data semantics that we are discussing, we will come down to earth in this section and focus on some generalizations that are currently finding their way into commercially available database systems.

We will begin by discussing SQL extensions to handle full-text as a meaningful data type, including the use of the thematic analysis discussed above. We shall consider in some detail the challenge presented by the need for ranked approximate searching. This will lead to considerations of multi-media document structure and queriability, expressed for example in the worlds of HTML, SGML, and HyTime. Finally, we will mention the integration of audio, image, and video information into database systems.

3.1 SQL/MultiMedia Full-Text

The ANSI/ISO standard for SQL is being extended with facilities for definition of abstract data types. This allows not only for individual users to define types, but for standard libraries of such types to be defined, and there is now a subsidiary ISO project known as SQL/MultiMedia (SQL/MM) defining a library of multimedia types, including a FullText type.

Consider a Documents table that stores the documents themselves in a FullText column, together with descriptive information in other columns (see Figure 7).

TABLE documents

doc # INTEGER	title CHAR(80)	author CHAR(40)	text Full Text

Figure 7 Documents table

FullText as a standard SQL/MM library type has a Contains function in its interface, which can thus be applied to instances of FullText data:

```
CREATE VALUE TYPE FullText
  (PUBLIC FUNCTION
        Contains (text FullText, pattern FT_Pattern)
                  RETURNS BOOLEAN;
    ....);

SELECT *
FROM Documents d
WHERE Contains(text,'soundex(''Peirce'') OR
                        is_about(''pragmatism'')' );
```

SQL3 also allows the definition of infix operators as syntactic sugar for function calls, so that it could be made possible to express the condition as

```
WHERE text CONTAINS 'soundex(''Peirce'') OR
                        is_about(''pragmatism'')'.
```

The pattern argument of Contains offers Boolean combinations of a variety of string matching functions. In particular, is_about could be implemented by a thematic searching engine as described above. Just as indices may be maintained on columns of conventional data types to aid in retrieval, so can indices be maintained on FullText columns to support the Contains function and its pattern subfunctions, including an index of theme vectors to support is_about.

3.2 Ranked Approximate Searching

Two aspects of searching for information become of much greater significance when full text or multimedia data are involved – the potentially large volume of responses due to imprecise search criteria, and the need to rank those responses in some order of relevance.

Searching for relevant textual information is already a problem on the Internet, and future applications such as home shopping will call for improved visual searching, e.g. for similar colors or shapes or textures in clothing catalogs, as in Query By Image Content (IBM, 1995). There are highly interesting semantic questions as to what are intuitively satisfactory metrics for similarity in textual and multimedia searches, but they would take us too far afield at this point. Assuming that there exist various state-of-the-art functions that can determine whether there is a match, and can quantify how close that match is, we shall focus on the implications for familiar query mechanisms using predicates as in an SQL SELECT statement.

Conventional queries over numeric and string data types have a limited set of operators that test for strict equality, or use comparisons based on the total ordering of a numeric type or a collating sequence for a string type. In standard SQL, at least when using a cursor, the ORDER BY clause on columns of the result can be used to simulate the closeness of the matches in comparisons:

```
DECLARE C CURSOR FOR
   SELECT name, salary
   FROM Emp
   WHERE salary > 50000 AND name >= 'Jones'
   ORDER BY salary DESC, name ASC;
```

Although the ordering is on the result values rather than the matches in the comparisons themselves, the ordering is the same with a simple linearly ordered data type, and the choice between ascending and descending order can be used to decide whether the closest comparison or the most distant one is given the greatest weight. However, this will not work in general for arbitrary matching functions, such as Contains on FullText data.

In fact, as shown above, Contains is specified to return a Boolean value, and provides no ranking information about the closeness of the match. Similarly, the functions Soundex and Is_About in the search pattern are conceived as Booleans that may be combined by operators such as OR. (Another semantic question that we do not have space to pursue here concerns the ways in which an overall ranking should be computed from a combination of component rankings in a composite pattern like this).

There seem to be two main alternative ways in which weighting can be added to search functions like Contains. The first of these, currently favored by the ISO SQL committee, is to define the functions to return numeric values instead of Booleans, so that the manipulation of the weights is explicit in the query. For each row of the Documents table, the WHERE clause is evaluated, and then (conceptually) the Contains function is reevaluated in the SELECT clause to form the column named w in the result, so that ordering can be on this column:

```
DECLARE C CURSOR FOR
   SELECT d.*, Contains(text, 'soundex(''Peirce'') OR
                              is_about(''pragmatism'')' AS w
   FROM Documents d
   WHERE Contains(text, 'soundex(''Peirce'') OR
                         is_about(''pragmatism'')' ) > 0
   ORDER BY w DESC;
```

To spare the user from having to write the Contains function twice, it is now possible in SQL-92 to write a derived table in the FROM clause:

```
DECLARE C CURSOR FOR
   SELECT *
   FROM (SELECT d.*, Contains(text, 'soundex(''Peirce'') OR
                                    is_about(''pragmatism'')' ) AS w
         FROM Documents d)
   WHERE w > 0
   ORDER BY w DESC;
```

This is easy to understand, as first computing an additional column w of weights, and then referring to those in the WHERE clause and (as part of the * expansion) SELECT clause. However, it represents a considerable complication and reconstruction of the original form of the query in order to capture the weighting and use it for ordering.

The comparison '>0' would not be required if there were implicit casting from a non-zero number to true, and zero to false. Comparisons would then only need to be written for non-zero thresholds. To achieve this, the language could define a multi-valued Boolean type so that search functions could return values of this type, combining the arithmetic and Boolean properties.

The other major alternative would be to retain functions like Contains as Booleans, but to have the ranking returned as a side-effect in some way, e.g. by adding an optional OUT parameter to Contains, or by some more implicit language feature. In this way, the main Boolean outline of a query can be preserved, with the ranking being a minor decoration if desired. The challenge is that the recipient of the side-effect is not a single variable, but has to be a column of values, one per row of the table being queried, e.g.

```
DECLARE C CURSOR FOR
   SELECT *
   FROM Documents d OUT w
   WHERE Contains(text, 'soundex(''Peirce'') OR
                         is_about(''pragmatism'')', w )
   ORDER BY w DESC;
```

Values in an OUT column would only be referenceable as OUT parameters, and could be integers by default. Although arbitrary side-effects in queries are undesirable, this local assignment to a transient column appears harmless. However, the weighting is still intrusive, and the reader is invited to improve on this solution.

3.3 HTML, SGML, DSSSL, and HyTime

Moving up from linear text to completed structured documents, it is difficult to be unaware of some of the developments that are taking place. Our local newspaper recently devoted a whole page to describing in detail how its readers could use the HyperText Markup Language HTML to set up their own home pages on the World Wide Web. Admittedly the San Jose Mercury News is not quite a typical provincial newpaper, catering as it does to the residents

of Silicon Valley, but this kind of visibility for the intricacies of a document markup language is probably unprecedented.

HTML (IETF, 1995; Netscape, 1995) describes hierarchical document structure and, of course, hypermedia links. The semantic significance is that these are meaningful properties of hyperdocuments, that could for example be useful in queries. Space permits only the briefest of introductions here for those who wish to explore the implications of this world for data semantics.

HTML has grown up as part of the World Wide Web, and is syntactically a subset of the formal ISO standard SGML (ISO, 1986; Goldfarb, 1990), although it adds some presentation semantics of its own, whereas the SGML philosophy is to have presentation separately described in the style language DSSSL (ISO/IEC, 1995), which is approximately a purely functional subset of the Lisp-like language Scheme.

HyTime is another formal international standard (ISO/IEC, 1992; DeRose and Durand, 1994), ahead of its time (not always a recommendation for a standard), and well worthy of study. HyTime specifies what it calls *architectural forms*, which are effectively abstract superclasses of objects whose attributes can be inherited by the actual object types that the HyTime user specifies. It uses SGML syntax, and defines richer forms of hypermedia linkage than HTML yet contains. For example, the anchors (ends of links) can be 'rectangles' in n-dimensional space, such as time sequences of parts of video frames. (HyTime evolved from a music description language!) This control of granularity helps in dealing with the problem of ostension – what exactly is being pointed at?

To draw the threads together for our present semantic purposes, DSSSL and HyTime both define a standard parse tree representation of the components of an SGML document, and a common set of functions for querying such trees. (A corrigendum to HyTime has just been approved to reflect the convergence with the newly finalized DSSSL.) This means that one could treat an SGML document as being an instance of an SGML abstract data type in SQL3 terminology, with the query functions providing the interface to the type, and defining the behavior, the semantics, of such objects. This would be exactly analogous to the treatment described above of FullText as an abstract data type, with the Contains function in its interface.

3.4 Audio, Image, and Video

Other information in media such as audio, image, and video lends itself to the abstract data type treatment, with appropriate functions being defined in the interfaces, and subtypes defining more specialized forms. SQL/MultiMedia plans to add Audio, Image, and Video types, although no detailed proposals have yet been made. The actual binary representation of the objects could be encapsulated in a private attribute of the type Binary Large Object, which has already been added to SQL3. As soon as such types are defined (and indeed an SQL3 user could define their own without waiting for an SQL/MultiMedia standard), the semantics of these types, as exhibited in their interfaces, become usable in queries and other SQL statements.

Several MultiMedia scripting languages for graphical user interface applications, such as Visual Basic (Microsoft, 1993) and Media Talk (Oracle, 1995b), already provide class

libraries for multimedia types. Indeed, the term 'scripting language' seems to mean a (usually) simple interpreted programming language, with a built-in class library of this kind.

Although we have dwelt on the conceptual simplicity of the introduction of new data types and functions as a semantic model, the implementation challenges will be considerable. Query optimizers will be baffled by the frequent occurrence of functions that they do not recognize, and would rather not have to evaluate at run-time. Hence there will be the need for techniques such as maintaining indices on the results of frequently-used functions, by analogy with the indices and theme vectors that are built to support the use of Contains on FullText data.

In case anyone doubts the potential value of automating indexing of audio-visual information, we will give a contemporary illustration. After directing the film 'Schindler's List', Steven Spielberg established a project to record interviews with Survivors of the Shoah (the holocaust). It is expected that 40,000 interviews will be carried out over the next three years, and these will be catalogued by subject, key phrases, and location. There will be 64 interviews scheduled per day, averaging about two and a half hours each. Most of the information to be abstracted is from the audio, although interviewees may also show photographs and possessions of historical relevance. In the absence of adequate technology, 64 human cataloguers will be employed to keep pace with the incoming videos.

4. GENERALIZED SEMANTICS – *Mysterioso*

We will begin this section with some consideration of the generality of signs. Then follows an outline of semiotics, especially the work of C.S. Peirce, which is suggested as the basis of a new paradigm for data semantics. Finally, there is brief discussion of computer systems as agents, designed to observe and generate signs and act on behalf of human users.

4.1 Generality of Signs

We take a sign to be anything that may be interpreted as conveying information. Signs may be generated by humans (language, gestures), or by other living things (tail-wagging), or they may even be inanimate natural phenomena (rainbows).

In the case of signs such as linguistic utterances, hand waves, or traffic lights, the conveyance of information may be intentional (although the intention may be remote from the sign situation – we are not implying that traffic lights themselves have intentions). But there seems to be no advantage to limiting signs to intentional use of this kind, and there are good reasons not to – we can adopt a uniform approach to conveyance of information, we can be spared the difficult determination of whether intention is present, and we avoid doing violence to ordinary English usage where 'sign' is used independently of intention ('a tell-tale sign', 'a sign of nervousness', 'a sign of rain').

Signs can employ a variety of media. Audible and/or visual signs are the most obvious, with language being a special case of a system of audio-visual signs. Tactile signs are also already commonplace in communication with computers, in the use of a keyboard or a mouse.

A handshake is a sign that conveys information through the sense of touch, but with a subtlety not yet to be found in human-computer interfaces. Ten years ago, we used to suggest a simulated hand as a futuristic addition to the videoconference desktop, since the handshake

is an important part of the business culture of many societies. The simulated hand was to be endowed with the dual capability of (i) conveying the characteristics (shape, pressure, temperature, humidity) of the handshake of a remote party using a similar device, and (ii) sensing one's own handshake characteristics to be conveyed to the other party, with the characteristics dynamically responding to each other in the simulation. This may have seemed ludicrously unrealistic at the time, and yet technology has now reached the point where one reads of prosthetic devices that can restore some forms of tactile sensation as though they came from the missing limbs – possibly a harder problem.

The olfactory senses of taste and smell are also important media of information-bearing signs in our daily lives, whether to give pleasure or to warn us that all is not well.

Finally, as we shall see later, even our own thinking may be considered as a medium for sign-activity, a process of generating signs to be further interpreted.

To show that these more general kinds of signs are urgently relevant to the information superhighway, it is perhaps sufficient to cite a couple of application areas.

First, for home shopping to be an effective alternative to first-hand inspection of the goods, it will often be important to offer more than audio-visual evidence. The shopper for food, say, needs to be convinced by simulation of the feel, the taste, the smell – imagine the French giving up their daily trips to the food stores and markets without this! Or the cautious shopper for vacation travel can be won over by simulation of the total experience – 'Reading about Patagonia can include the sensory experience of going there' (Negroponte, 1995).

Second, medical applications clearly benefit from a wide sensory range. Indeed, the word 'semiotics' has a technical sense in medicine as the study of symptoms – signs that convey, usually unintentionally, information about bodily and mental health. Diagnostic aids to physicians can include querying of multi-sensory databases to match against static images, or movements in video, or even distinctive odors. Some applications are already far advanced in this field – the Video Dissector, for example, simulates a corpse on which medical students can hone their skills, even having the advantage that 'unlike the actual cadaver, the program can be rewound'.

4.2 Semiotics

Historical background

In order to establish an appropriate framework for the generalization of semantics, we turn to semiotics, the theory of signs. This subject was given its modern impetus by the American logician and philosopher C.S. Peirce (1839–1914), who may come to be recognized paradoxically as the 'next' Turing – although Peirce was born 73 years earlier, his influence in computer science can only come after that of Turing. The two men shared the same penetrating intellectual vision founded in mathematical logic, and even appear to have had physiognomical similarities, with large direct eyes and high cheekbones.

A rather tenuous thread of semiotics has been maintained in this century by Saussure (1916), Ogden and Richards (1923) – 'The Meaning of Meaning' is highly readable and still in print – and Morris (1946, 1971). A rare early perception of the applicability of semiotics in computing is found in Zemanek (1966).

Another historical note that may be of interest is that Umberto Eco, before becoming celebrated as a novelist, was already professor of semiotics in Bologna, and had published a

book (1976) in which he frequently cites Peirce, and asks 'How then can one represent this type of semantic universe (which happens to be the one in which human beings live)?'. This leads him to find promise in the semantic network approach of Quillian (1968) in the computing world.

The influence of Peirce is seen also in the work of Sowa (1984), and in the ongoing studies of IFIP Working Group 8.1 on Information Systems Concepts.

Peirce's theories

In order to suggest some of the power and relevance of Peirce's thinking, we can give here only the briefest summary of his semiotics. An annotated bibliography may be found at the end of the paper to guide any readers interested in investigating further.

One contribution of Peirce was his extensive classification of signs, which eventually ran to 10 trichotomies and 66 classes of signs. Most instructive as an introduction is his trichotomy of Icons, Indices, and Symbols. An Icon serves as a sign due to a likeness to its referent, such as the floor plan of a house. It exemplifies Peirce's category of Firstness – in playing this role, the inherent likeness is independent of anything else. An Index, such as a weather vane or a bullet hole, serves as a sign in the category of Secondness, being directly, *compulsively*, related to the direction of the wind or the passage of the bullet that is its referent. A Symbol, however, serves as a sign only by convention – mediation is required between the sign and its referent, so that this exhibits Thirdness.

With his general treatment of the categories of Firstness, Secondness, and Thirdness, Peirce made one of the major contributions to metaphysics since Aristotle. He claimed that all of these categories were needed, but no more than these. A special case of this argument that is close to the field of data semantics is his assertion that there are relations that are *irreducibly triadic*, although all higher order relations are reducible. For example, *A gives B to C* cannot semantically be reduced to combinations of dyadic relations such as *A parts with B* and *C receives B* without loss of information. Adequate decompositions into dyadic relations require a unifying triadic relation at some point. But the issues are subtle and not uncontroversial, and Peirce's several expositions make fascinating reading.

The foundation of Peirce's semiotics is then his analysis of the observation of a sign ('sign activity') as involving a triadic relationship between:

- the Sign itself;
- its Object;
- and an Interpretant, *which is another sign.*

Typically, the Interpretant is a cognition produced in a human mind. Mental activity is sign activity – thought is a succession of interpretations of signs producing other signs, potentially infinite (logically, if not physically). However, a particular sequence of interpretants can come to an end at various points, depending on external factors such as loss of curiosity, or, to use a more Peircean phrase, cessation of doubt. This may seem a rather anti-climactic outcome of the sign activity, but we want to suggest that all is not lost. For a start, some of the interpretants may remain in the memory; but there is a much stronger positive case to be made, that this theory of an often desultory and inconclusive sign activity is a correct description of human cognitive behavior.

Let us narrow down our scope for a moment to consider signs that are single English words, and compare Peirce's theory with the way we use a dictionary to clarify semantics. We proceed from one dictionary definition often to several others as interpretants, until eventually we cease this activity. We have usually not arrived at a mathematical definition of concepts grounded in primitive axioms, but rather have satisfied ourselves that we have a sufficient understanding of the relationships between the concepts for our purposes (or at other times we give up and remain confused). Sometimes the definitions are almost circular, and yet we understand the way they hang together well enough to use them so that what we want to convey will be correctly understood by other proficient users of the language.

This imprecise, social notion of semantics (as later in Wittgenstein: 'The meaning of a word is its use') is echoed convincingly in Peirce's account of scientific activity, especially in the physical sciences. Science proceeds by progressive consensus, and scientific theories become accepted by the community of those proficient in a field of study, not by the sheer accumulation of positive evidence so much as by the ability to withstand attempts at refutation. If a crucial experiment shows a failure in a theory, this is not a failure for science, but rather a step towards a more successful theory. This is the philosophy of scientific progress known as Fallibilism, and espoused in this century by Karl Popper. Peirce also contributed to the theory of probability, and understood the indeterministic implications of its use in science, which culminated in quantum theory a few years after his death.

Thus meaning and truth are in general not absolute, but represent a consensus that is subject to revision. This does not mean that we are forced into a hopeless scepticism, but rather that we honestly accept in all sign activity a degree of approximation and the possibility of error. Often an acknowledged approximation is good enough – we still use Newton's laws of motion when the relativistic correction is so small as to be irrelevant. And when a consensus has been often relied upon and yet remained unchallenged for years or centuries, we have no need to doubt it.

The role of doubt in motivating sign activity was important to Peirce. He regarded doubt as something that was compulsive and could not be simulated. Hence one of his earliest papers (*Questions concerning certain faculties claimed for man*, in 1868) was a critique of Descartes, whose 'Cogito, ergo sum' was in response to a supposed ability to force himself to doubt absolutely everything before seeking a first indubitable proposition. For Peirce, sign activity, as in the example above illustrating the pursuit of semantics, is driven by doubt to the point where the doubt is quiesced, possibly due to our having reached an adequate approximation to an understanding (e.g. a competence to use or interpret the word in question) – or possibly due to the activity being interrupted and superseded by another doubt! Another way that Peirce often describes the removal of doubt and the acquisition of competence is as the formation of habits.

With this preparation, we are ready to digest Peirce's celebrated pragmatic maxim (in *How to make our ideas clear*, 1878) for signs that are intellectual concepts:

> 'Consider what effects, which might conceivably have practical bearings, we conceive the object of our conception to have. Then our conception of these effects is the whole of our conception of the object.'

Beyond its original roles of stripping away mystery, e.g. in metaphysics, or of showing whether two concepts were distinguishable, e.g. in science, the maxim summarizes Peirce's

theory of semantic understanding as a competence to conceive practical effects. Semantic communication with others is achieved insofar as they share that understanding.

4.3 Implications for data semantics

The semiotic theory outlined above deserves to be applied as a new paradigm for data semantics. First, it has the generality to handle information conveyed in any medium. The subsequent interpretation of the information may involve linguistic expression, or may itself involve signs in the multi-media imagination of a human being or in the multi-media processing of a computing system. Second, the theory is a better approximation to the reality of semantics in human discourse than one which assumes a simpler mathematical model of neat hierarchical ontologies of well-defined concepts.

This is not to say that existing and other paradigms should not continue to be explored. Computing systems are artefacts, and artefacts are not necessarily most successful when they imitate their animate counterparts – cars use wheels rather than legs, and aircraft do not flap their wings like birds. So we should not ignore the possibility that, for some time to come, information processing systems will succeed by using artificial abstractions.

Yet the complaints about the inhuman interfaces of computing systems are so loud and well justified that we need to explore other alternatives. Now is the crucial opportunity and the time to step up to the challenge, as computers are entering the lives of millions of users who have little relish for artificial abstractions. Computers are often replacing humans in performing many information processing tasks, but are viewed with distaste because they lack the human touch and are such frustrating substitutes.

In fact, the more general semiotic paradigm does not conflict with the use of conventional data semantics where the latter are effective. Some of our sign-activity is of a relatively precise and formal character, and Peirce's whole theory was inspired by his insight into this kind of activity in mathematics and science, so that it is certainly embraced by his semiotics. Yet even in these domains he shies away from assuming indubitable foundations and certain knowledge, replacing them with hypotheses and inferences that have so far withstood attempts at rebuttal.

Thus the main implication of this philosophy for data semantics is the need to deal with the fact that precision and certainty will be relatively rare in the semantics of the information that will pour down the information superhighway. Theories and systems that aim to correspond to human sign-activity must deal with vagueness and uncertainty and revisable truth as the norm, although they can certainly find a place for the more formal human means of communication where these occur. In our daily lives, we mostly communicate by natural language without reference to a precise ontology, and we use intonations and gestures that are difficult to quantify. More occasionally, we try to make our intended semantics more precise in natural language, or use artificial languages, for example in science, technology, or mathematics.

So how might this philosophy be realized in computational theories and systems? Peirce's notion of a kind of local equilibrium of a sign-processor, in which doubt has been replaced by amended dispositions or habits, fits well with the model of a system with a memory whose current state defines those habits or dispositions. As Eco saw 20 years ago, a semantic network of some kind might be appropriate, or one might want to regard a neural network as a more general primitive foundation, well adapted to the pattern recognition aspects of

imprecise and multimedia semantics (the semantics of 'similarity'). A semantic network could still be a valuable higher-level model specialized for linguistic semantics, related to the neural net model as presumably our introspective model of linguistic memory must be related to our underlying physical neurological state.

This suggests that the memory part of the model would not need to be very novel, although it might be called a *semiotic network* to emphasize that it was a collection of interrelated signs (only some of which were linguistic), and to avoid unintended assumptions about other properties that it might have.

It is in the area of the sign-activity of the accompanying *semiotic interpreter* that the breakthrough is needed. Peirce provided little guidance as to how the successive cycles of interpretation made their choices among the many possible interpretants, or as to how the removal of doubt could be a purposive part of the activity rather than a fortuitous outcome. Of course, a solution of this problem might be a major step towards a solution of the artificial intelligence problem, although not necessarily the whole solution. If intelligence is closely related to successful sign-activity, then a good theory of sign-activity is fundamental, but still leaves open the question of what leads to successful outcomes – why should the points at which doubt ceases be the appropriate ones?

Concentrating on the basic model of sign-activity, should the semiotic interpreter have an algorithm controlling the sign-activity, exerting will power or possessing free will, or are the algorithms part of the sign-activity itself? Peirce considered not only simple terms, but also propositions and arguments as signs, and thus as possible interpretants of other signs. Could algorithms also be modeled as signs or parts of signs, to participate in the generation of the next interpretant?

Models of this kind should be treated as hypotheses to be subjected to rigorous testing, so a good non-trivial first step would be to propose a feasible experiment involving some kind of simple sign-activity – something much simpler than Turing's test of machine intelligence.

4.4 Agents

Our final observation is that once a computing system has a strong enough implementation of sign-activity to be flexibly responsive to human queries, it will also be in a position to play a more symmetrical role in the interaction. For example, it may on its own initiative generate appropriate signs to communicate with a human user, possibly as a result of observing for itself some sign activity from another source.

A potential commercial application of a sign-generating system was reported recently under the heading 'Recliner that is Music to your Ears'. The article was illustrated by a photograph of a lady, fully wired for soothing sound and vision, stretched out in a special reclining chair. The system replaced a human masseur, controlling both the audio-visual entertainment, and the force and duration of the tactile signs – the massage – that the chair was capable of transmitting. This control was dependent on the agent's observation of the readings of sensors that monitored the beneficiary's vital signs.

This ideal of an observant *agent* acting on behalf of a user, much like a human assistant, has been pioneered for some while in research environments such as the MIT Media Lab. As Nicholas Negroponte has expressed it (1995): 'The challenge for the next decade ... is to

make computers that know you, learn about your needs, and understand verbal and non-verbal languages.'

This succinct statement brings us back to the starting-point of this paper. Even if the timeframe of a solution of the agent problem within a decade is over-optimistic, research in data semantics urgently needs to step up to the understanding of verbal and non-verbal languages that constitutes the semiotic challenge.

5. CONCLUSION

The information superhighway is making large quantities of linguistic and multimedia information readily available. In order to make this information truly useful, systems must become more powerful semantically.

We have indicated some encouraging current directions, including the thematic analysis of natural language text, and the use of object concepts in database languages, document markup languages, and scripting languages.

For the longer term, a breakthrough is still needed to establish a sufficiently general semantic model to encompass linguistic and multimedia sign-activity. Although a breakthrough cannot be scheduled, one may be imminent, given the stimulus provided by the gathering momentum of the information revolution. An approach based on Peirce's semiotics is advocated as the most likely to succeed.

6. ACKNOWLEDGEMENTS

I am grateful to Kelly Wical, Mike Kaposec, and Yitzik Brenman for their invaluable help with the description of ConText (to Yitzik also for the preparation of the illustrations); and to Daisy Miller for discussion of the Survivors of the Shoah project..

7. REFERENCES

Borges, J.L. (1983) *Labyrinths.* Modern Library, New York.

DeRose, S.J. and Durand, D.G. (1994) *Making hypermedia work : a user's guide to HyTime.* Kluwer Academic Publishers, Boston/Dordrecht/London.

Eco, U. (1976) *A theory of semiotics.* Indiana University Press, Bloomington.

Foley, J. (1995) Information visualization: the need for a data base approach. [This volume]

Goldfarb, C.F. (1990) *The SGML handbook.* Oxford University Press.

IBM Corporation (1995) Query by image content (QBIC). Home page on World Wide Web is http://wwwqbic.almaden.ibm.com/%7eqbic/qbic.html .

IETF (1995) Access official specifications and latest information on HTML via the IETF home page: http://www.ietf.org .

ISO/IEC (1992) ISO/IEC 10744:1992 *Hypermedia/time-based structuring language (HyTime).*

ISO/IEC (1995) ISO/IEC 10179:1995 *Document style semantics and specifications language (DSSSL).*
Microsoft Corporation (1993) *Microsoft Excel: Visual Basic user's guide.*
Morris, C.W. (1946) *Signs, language and behavior.* Prentice-Hall, New York.
Morris, C.W. (1971) *Writings on the general theory of signs.* Mouton, The Hague.
Negroponte, N.P. (1995) *Being digital.* Knopf, New York.
Netscape Corporation (1995) For introductory information on HTML, connect to http://home.netscape.com/home/how-to-create-web-services.html, or in Netscape Navigator, select How to Create Web Services from the Help menu.
Ogden, C.K. and Richards, I.A. (1923) *The meaning of meaning.* Routledge, London.
Oracle Corporation (1995a) *Oracle ConText linguistics toolkit guide and reference.*
Oracle Corporation (1995b) *Oracle Media Objects user's guide.*
Quillian, R.M. (1968) Semantic memory. *In* Minsky, M. (ed.): *Semantic information processing.* MIT Press, Cambridge.
Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985) *A comprehensive grammar of the English language.* Longman, London and New York.
Saussure, F. de (1916) *Cours de linguistique générale.* Payot, Paris.
Sowa, J.F. (1984) *Conceptual structures: information processing in mind and machine.* Addison-Wesley, Reading MA.
Wiederhold, G. (1995) Value-added mediation in large-scale information systems. [This volume]
Williams, J. (1990) *Style: toward clarity and grace.* University of Chicago Press, Chicago and London.
Zemanek, H. (1966) Semiotics and programming languages. *Communications of ACM,* **9:3,** 139–143.

8. BIBLIOGRAPHY

The following select bibliography is intended to help readers interested in Peirce. There is no substitute for reading his work in the original, either in the Collected Papers, or in shorter selections. Various commentaries are helpful, but none quite conveys the essence or force of Peirce's thinking.

Some of the books below are out of print, but are commonly to be found in libraries.

Collected Works

Peirce, C.S. (1931–58) *Collected papers of Charles Sanders Peirce. Vol 1–6* (ed: Hartshorne, C. and Weiss, P.), *Vol 7–8* (ed: Burks, A.W.). Belknap Press, Harvard.
The standard reference, to which citations are always given in the form v.p, for volume and paragraph number.
Peirce, C.S. (1982–93) *Writings of Charles Sanders Peirce: a chronological edition. Vol 1–6* (ed: Fisch, M. *et al*). Indiana University Press, Bloomington.
The start of an intended 30-volume edition, a major industry in Indiana. For specialists!

Selections

The following selections are quite similar, differing mainly in their editorial commentaries.

Buchler, J.. (ed.) (1939) *Philosophical writings of Peirce.* Dover, New York.
A good selection, and excellent value.
Cohen, M.R. (ed.) (1949) *Chance, love, and logic.* Peter Smith, New York.
Also contains an essay by John Dewey.
Housner, M. and Kloesel, C.J.. (ed.) (1992) *Essential Peirce. Vol 1. (1867–1893)* Indiana University Press, Bloomington.
Hoopes, J. (ed.) (1991) *Peirce on signs.* University of North Carolina Press, Chapel Hill.
The usual selection, despite the title.
Wiener, P.P. (ed.) (1958) *Values in a universe of chance.* Doubleday, New York.
Contains 'Concerning certain faculties ...', omitted by Buchler.

Commentaries

Buchler, J. (1939) *Charles Peirce's empiricism.* Harcourt Brace, New York.
Deely, J. (1990) *Basics of semiotics.* Indiana University Press, Bloomington.
Fisch, M. (1986) *Peirce, semeiotic and pragmatism.* Indiana University Press, Bloomington.
Gallie, W.B. (1952) *Peirce and pragmatism.* Penguin, Harmondsworth.
See especially chapter 5 on the doctrine of thought-signs.
Goudge, T.A. (1950) *The thought of C.S. Peirce.* Dover, New York.
See especially chapter V section B on general semiotic.
Hookway, C. (1985) *Peirce.* Routledge, London and New York.
See especially chapter IV on the theory of signs.
Knight, T.S. (1965) *Charles Peirce.* Washington Square Press, New York.
Well written and accurate. 60 cents when new, probably hard to find now.
Wiener, P.P. and Young, F.H. (ed.) (1952) *Studies in the philosophy of Charles Sanders Peirce.* Harvard University Press.
A collection sponsored by the Peirce Society of America.

9. BIOGRAPHY

David Beech is currently Director of Media Standards for Oracle Corporation. Since 1988, he has been pioneering the introduction of object database technology at Oracle. Prior to that, he had initiated in 1982 the work in object databases at Hewlett Packard Laboratories that led to the Iris system, and he has been closely involved with the design of object extensions to SQL. His earlier work with IBM was mainly concerned with the design and specification of the PL/I language, and with the prehistory of the DB2 relational database system. He holds an M.A. in mathematics from Cambridge University.