FUTURE REQUIREMENTS ON DATA MODELLING

W. Kent

Rapporteur: Dr. B.N. Rossiter

Abstract

The trend in data usage is toward increased sharing and distribution, and increasingly direct involvement of end users. Data is becoming recognised as a resource of the business enterprise. The consequent demand on data structures is to be semantically richer, and to reflect more of the enterprise's information structure and less of computer technology. An emerging class of data models addresses these needs.

Prologue

My appearance as the first speaker here should not be construed as establishing a keynote theme. Even before we've surveyed data base technology as it is, I'm going to talk about what it ought to become. My work is not mainstream; it is critical of the mainstream, and should perhaps come after the mainstream as a critique and a caveat. Nevertheless, this sequence may work well after all, if it prompts you to listen to all that follows a bit more critically.

I am probably one of those "data base thinkers" of which Tsichritzis says an infinitesimal number are needed. To oversimplify, my view is that the current technology of data base is not the best approach, and we ought therefore to be careful of how and why we teach it. Of course, the truth of that depends on what we mean by "best". To make an analogy, it's somewhat like saying that factory mass production is not the best way to make things. Hand craftsmanship is generally superior in terms of the quality of the product. But to be realistic, "best" usually has to mean economical and profitable, and being of the most use to the most people. Innate quality of the product is often legitimately compromised for business reasons. Not enough people want to pay much extra for quality.

But hand craftsmanship always shows us the ideal, a target to be attained. As more and more people need or want the qualities of hand crafted products, mass producers look to the hand crafters for the standards of quality and for ideas about techniques. The hoped for result is a marriage of the two, with the mass producers adapting some of the objectives and methods of the hand crafters to achieve better results for all concerned.

I tend to be involved in searching for ideal ways to represent information in data base systems, stressing the criteria of naturalness and completeness of representation, suitability to the

views and needs of all users, independence from the idiosyncrasies of technology, and long term stability. I subordinate other concerns, such as performance, size, integrity, recovery, etc. which are addressed more or less by current technology.

Most workers, I think, focus on the technology, just barely looking at reality out of the corner of the eye. Their mode is to nudge technology from where it is in increments which are useful and profitable.

But they lack a larger sense of purpose. The purpose of this technology is to represent information -- real information about real things in the real world -- for use by people. It wouldn't hurt occasionally to look squarely at that information, at that reality, and attempt to understand the nature of that beast -- and only then consider what kind of technology would be useful to deal with it.

What Data Modelling Is

Data model (or information model): system of concepts and constructs for describing data or information. A vocabulary. Could be, but not limited to, a data structure. For example:

- Records and fields.
- Columns and rows of tables.
- Entities, relationships, attributes.
- Sets, as in CODASYL.
- Sets, as in set theory.
- Business forms and line items.
- Processes and events.
- Predicates, axioms, inference rules.
- Sentences, grammar.
- Functions and variables.

Where it occurs

Data description occurs, in various forms, at various stages of the planning, design, and use of a data base.

- In the planning stages of a data base, there may be very gross descriptions of the data to be maintained, for example, "10,000 employee records", "one megabyte of tax records", etc.

- In the DB process, more detail is defined in terms of all the individual items of data required.

- Description of physical data, as required by data management software. For example, catalogues, IMS DBD's.

- Description of data as used by programs: data declarations.

- Description of available data for new applications, for example data dictionaries.

- Description of the enterprise's information which is represented by the data in the database. Conceptual schema, ANSI/SPARC framework.

- What a data administrator mentally keeps constant while he recognises the physical data.

A data model is implicitly present wherever language is defined for the access and manipulation of data, for example data management functions, query languages, end user interfaces.

Changing Needs

Past: data file was used by a few trained programmers.
Future: data base used by many people, some with no programmer training.

- Many people using the same data base:
  - Integrated data base.
  - Distributed systems.
  - End user interfaces.
- Many people have to understand the data base:
  - Application planners and programmers.
  - Non-programmer end users. Query systems, "user friendly" interfaces, automated offices.
  - Management: awareness and use of information as a corporate resource.

Increasing requirements on systems themselves to exhibit some "understanding" of the data.

- Integrity: enforcing validity and plausibility criteria. Not just passive data repository.

- Translation capability in various conversion situations. For very disparate forms, need to comprehend "deep structure".

  - Multiple application views, differing from stored view.

  - Migration to new DBMS.

  - Data export for use elsewhere.

  - Heterogeneous distributed systems.

End user facilities: man-machine interface shifts toward human dialog. For example, in query systems, office automation.

- Constructs and syntax like natural language.

- Inference.

- General understanding of what's relevant and plausible.

- Real-world common sense. Knowledge base.

- Error forgiveness.

Consequence Requirements

First, a better understanding of the various contexts (a framework), and the different requirements for each. For some contexts:

- Explicit documentation of conventions and assumptions. No private formats, unexplained usages of techniques, unwritten assumptons.

- Constructs derived from natural information, not computer technology.

- Resiliency: usable in multiple ways, minimally impacted by changing usages and requirements.

- Supplementary information, beyond simply the form and content of the data itself. Constraints, implications, general knowledge. Relationships implied by the data.

Current Indications

Considerable interest in the literature.

ANSI/SPARC framework, conceptual schema.

ISO working group on conceptuall schema.

Work in progress: Schank, Sowa, Ladder (Hendrix), Rendezvous (Codd 78), Bari (Dell'orco).

University of Stuttgart, University of Malan, University of Brussels, IRIA (France), Siemens, CCA.

Semantic enhancements of relational models (RM/T (Codd 79), entity join (Kent 79c), others).

NDB (Sharman), functional models (Shipman), some dictionary interfaces.

Status

Many issues outstanding, concerning requirements, models, frameworks (Olle, Nijssen), and even the nature of reality.

ANSI, ISO.

Emerging models.

- Relational
- Binary
- Entity, Attribute, Relationship
- Sentence
- Predicate Calculus
- Aggregations
- Case models
- Functional

(Each in many variants.)

Discussion

Professor Tsichritzis reported that his group had attempted to develop a general natural language interface using a semantic network to a relational system, but the vast amount of natural language available and its complexity made it impossible to accommodate the system even on an IBM 370/168. Only in specific areas, such as lunar rocks, did this approach appear to be feasible. Mr. Kent agreed that it would be many generations of computers before we came close to achieving anything like the sophistication of the human brain in semantics. Dr. King in addition observed that after they have been set-up, semantic data bases are difficult to change.

References

ANSI — The ANSI/X3/SPARC DBMS Framework, Report of the Study Group on Database Management Systems, (D. Tsichritzis and A. Klug, editors), AFIPS Press, 1977.

Codd 78 — E.F. Codd, R.S. Arnold, J-M. Cadiou, C.L. Chang, N. Rossopoulos, "Rendezvous Version 1: An Experimental English-Language Query Formulation System for Casual Users of Relational Databases", IBM Research Report RJ2144, Jan. 1978.

Codd 78 — E.F. Codd, "Extending the Database Relational Model to Capture More Meaning", ACM SIGMOD International Conference on Management of Data, May 31 - June 1, 1979, Boston, Mass.

Dell'orco — P. Dell'orco, V.N. Spadavecchia, and M. King, "using Knowledge of a Data Base world in Interpreting Natural Language Queries", Inf. Proc. 77, Proc. 6th IFIP Congress, Aug. 8 - 12, 1977, Toronto, Canada, North Holland Publ. Co., 1977.

Galtieri — C.A. Galtieri, "Framework for an Information System", IBM Research Report RJ2394, Nov. 1978.

Hammer — M.M. Hammer and D.J. McLeod, "On Database Management System Architecture", Technical Report TR79-4, Computer Science Dept., University of Southern California, April 1979.

Hendrix — G.G. Hendrix, E.D. Sacerdoti, D. Sagalowicz, and J. Slocum, "Developing a Natural Language Interface to Complex Data", ACM Transactions of Database Systems 3 (2), June 1978.

Kent 78 — W. Kent, Data and Reality, North Holland, 1978.

Kent 79a — W. Kent, "Limitations of Record Based Information Models", ACM Transactions on Database Systems 4(1), March 1979.

Kent 79b — W. Kent, "Data and Reality: the Next Five Years", Infotech Conference (Database - the Next Five Years), San Francisco, May 1979.

Kent 79c — W. Kent, "The Entity Join", Proc. Fifth Intl. Conf. on Very Large Data Bases, Oct. 3 - 5, 1979, Rio de Janeiro, Brazil.

Nijssen — G.M. Nijssen, "Modelling in Database Management Systems", (EURO-IFIP 79, London. Proceedings: North Holland.)

Olle — T.W. Olle, "Multistage Data Definition in a Multi-component DBMS Architecture", in B. Shnidermann, Databases: Improving Usability and Responsiveness, Academic Press, 1978.

Schank — R.C. Schank and R. Abelson, Scripts, Plans, Goals and Understanding, Lawrence Erlbaum Accoc., Hillsdale N.J., 1977.

Sharman 77 — G.C.H. Sharman, "Update-by-Dialogue: An Interactive Approach to Database Modification", ACM SIGMOD International Conference on Management of Data, Aug. 3 - 5, 1977, Toronto, Canada, D.C.P. Smith (ed.).

Sharman 78 — G.C.H. Sharman and N. Winterbottom, "The Data Dictionary Facilities of NDB", Proc. Fourth Intl. Conf. on Very Large Databases, Sept. 13 - 15, 1978, Berlin, Germany.

Shipman — D.W. Shipman, "The Functional Data Model and the Data Language DAPLEX", ACM SIGMOD International Conference on Management of Data, May 31 - June 1, 1979, Boston, Mass.

Sowa — J.F. Sowa, "Conceptual Graphs for a Data Base Interface", IBM J. Res. & Dev. 20 (4), July 1976, pp. 336-357.