LIMITING FACTORS IN CURRENT DATA MODELS

W. Kent

Abstract

We are too familiar with the data models that have evolved in current computer technology, to the point of being blind to their limitations. There are surprisingly many diffferences between the record structures of today and the configurations occurring naturally in information.

Introduction

All currently used and currently popular data models are based on the traditional technology of the record structure. This includes traditional access methods, current data base products (for example, IMS, TOTAL), the network model of CODASYL, the relational model, the aggregation model (Smith), and even the entity-relationship model in the form in which it is commonly presented. Even though some of these models impose additional structure over the records, they retain the record as their fundamental concept.

The structure of records naturally suits the context of automated data processing technology. It does not correspond very well with the general structure of real information. The more we are motivated to produce faithful models of real information, the more we will have difficulty with record based constructs.

The Characteristics of Records

- Disjoint types

  A record belongs to exactly one record type.

- A single-valued fields

  If we accept first normal form, i.e., disallow groups, of repeating fields.

- Horizontal homogeneity

  Every record has the same fields.

- Vertical homogeneity

  Every occurrence of a given field within a given record type contains the same kind of value.

- Name-based representation

  A field contains a simple, or simply structured, string of characters. Whether that string uniquely identifies anything, or whether several such strings refer to the same thing, is totally unknown.

- Lack of field name discipline

  The system regards field names simply as placeholders. They might correspond to entity types, relationship names, combinations, or neither. Fields with similar contents needn't have similar names.

- Multiple uses of multiple columns

  - Fact acout an entity.
  - Qualified or composite name.
  - Relationship (intersection record).

- Separation of records and descriptions

  Catalog or dictionary data is "essentially different" from data base data.

Representing Entities, Attributes and Relationships

While not the only possibility, we adopt the concepts of entity, attribute, and relationship as a natural framework for describing real information. We use the terms in an informal, real-world sense, and not as described in any formal data model.

How well do their characteristics correspond with the properties of records?

- Overlapping types (Figure 1).

  Employees, customers, stockholders, people, companies, government agencies, schools, legal entities, ...

- Non-uniform attributes within type (Figure 2).

  Married and unmarried employees have different attributes, and so do hourly vs. Salaried, and also temporary and permanent, etc. Some categories are even more diverse in their relevant attributes: clothing; tools; furniture; vehicles; animals; ....

- Multiple types as domains of relationships (Figure 3).

  Owners own property. Users authorised to resources.

- Multi-valued attributes and relationships.

- Relationships as entities in themselves (Figure 4).

- Non-unique names, multiple names, and no names (Figure 5).

- The data/description continuum (Figure 6).

Representation Mappings

- A record represents an entity -- sometimes (Figure 7).

- Many ways to represent relationships.

- Field names used for many purposes (Figure 8).

- Multiple fields are used for different things (Figure 9).

References

P.P.S. Chen, "The Entity-Relationship Model: Toward a Unified view of Data", ACM Transactions on Database Systems 1 (1), March 1976, pp. 9-36.

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

W. Kent, "Limitations of Record Based Information Models", ACM Transactions on Data Base Systems 4 (1), March 1979.

J.M. Smith and D.C.P. Smith, "Data Base Abstractions: Aggregation and Generalization", ACM Transactions on Data Base Systems 2 (2), June 1977.

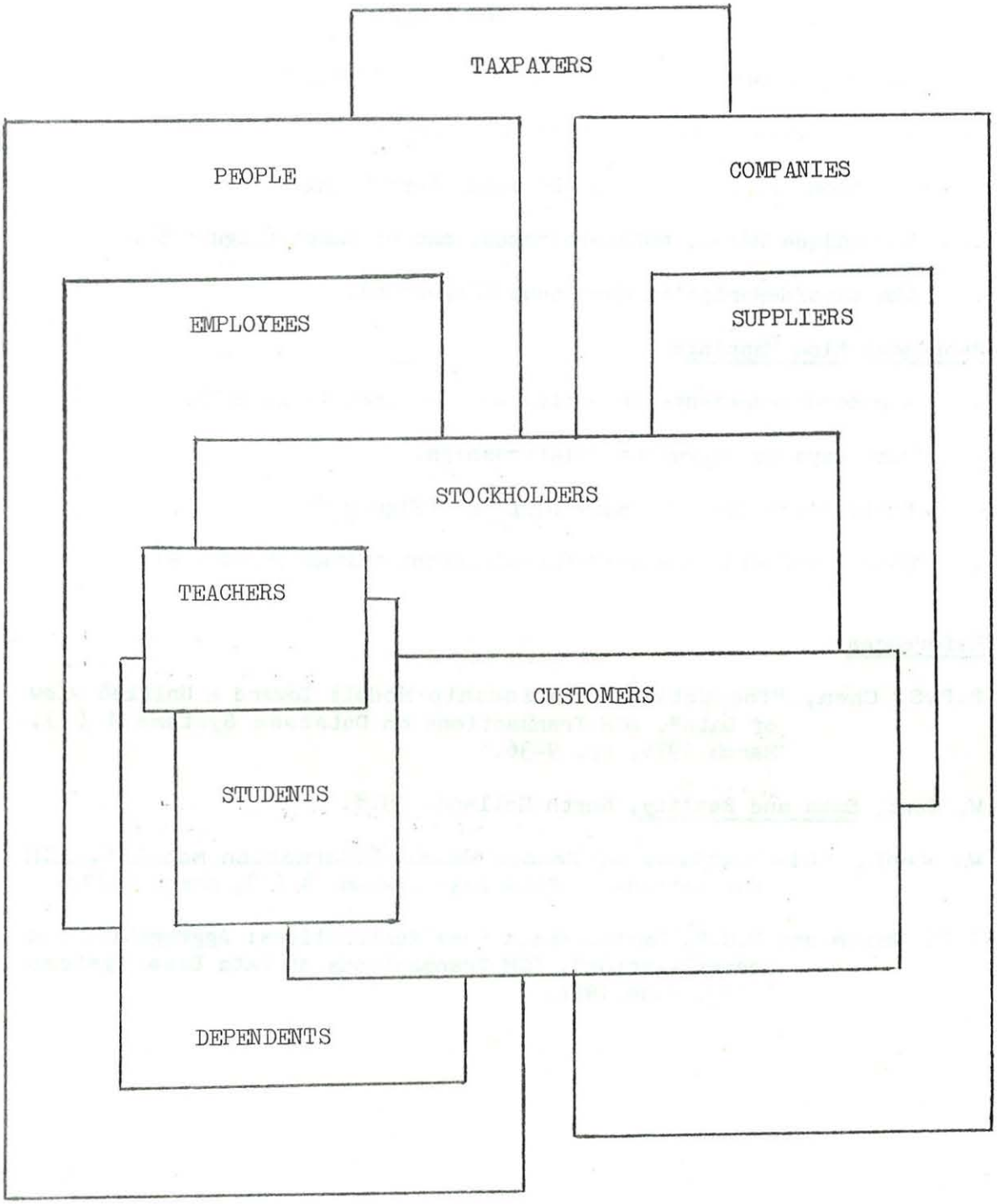

Figure 1 Overlapping Types

The "attributes" of clothing:

size, waist size, neck size, sleeve length, long or short sleeves, cup size, inseam length, button or zipper, sex, fabric type, heel size, width, color, pattern, pieces, season, number, collar style, cuffs, neckline, sleeve style, weight, flared, belt, waterproof, formal or casual, age, pockets, sport, washable . . . .

Figure 2 Non-uniform Attributes within Type

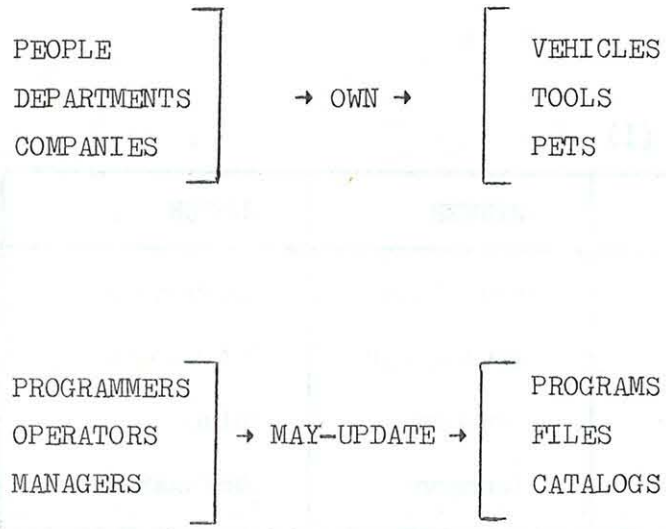

Figure 3 Multiple Types as Domains of Relationships

Harry owns Fido:

Since 1970.

Completely.

Longer than Fred owns Rover.

Under pet registry number 9999.

Figure 4 Relationships as Entities in themselves

Non-unique names: Harry

Multiple names: "W. Kent" = "William Kent" = "Kent, W." = "451999" = "199-99-9990" = ...

Complex names: the man who first ran the mile in under four minutes.

No names:

ELECTIONS (!)

YEAR	WINNER	LOSER
1952	Eisenhower	Stevenson
1956	Eisenhower	Stevenson
1960	Kennedy	Nixon
1964	Johnson	Goldwater

Figure 5 Non-unique names

Which of the following should be answerable from the data base, and which from the description (catalog, schema, dictionary)?
Why?

Does Fred Smith work in the Accounting department?

How is Fred Smith connected with the Accounting department?

How many employees are there in the Accounting department?

How many managers?

What is the maximum number of employees allowed in the Accounting department?

In any department?

What skills do the employees of the Accounting department have?

Which are required?

Is the Accounting department allowed to own vehicles?

Is any department?

Figure 6 The Data/Description Continuum

EMPLOYEES

E-NUM	NAME	BORN	SPOUSE	BORN
765	A. Bee	Boston	B. Bee	Rome

←(1)

↑ (2) ↑ (3)

(4) ↓

CUSTOMERS

C-NUM	NAME	LOC'N
432	A. Bee	Paris

↑ (2)

SALARY-HISTORY

E-NUM	DATE	SAL
765	1/78	700
765	6/78	750

(1) One record, many entities.

(2) One entity, many records.

(3) One entity, no records.

(4) One record, no entities.

Figure 7 How do Records represent Entities

SOCIAL-SECURITY-NUMBER: A domain (of identifiers).

DEPARTMENT: A domain of related entities. The nature of the relationship is not indicated.

HIRED-BY: The name of a relationship. The type of the related entities is not indicated.

ASSIGNED-PROJECT: A hybrid -- relationship name plus domain.

FIELDS: None of the above.

DEPT: Means the same as an earlier example. Can the data system detect that?

DEPARTMENT: In a furniture inventory record, it means the department currently using the furniture. Its meaning is different from the earlier example. Can the data system detect that?

Figure 8 Field Names

CITY-DATA (U.S.A.)

CITY	STATE	YEAR	POPULATION	MAYOR	PARTY
Dayton	Ohio	1950	59500	Jones	Demo.
Dayton	Ohio	1960	64250	Smith	Rep.
Dayton	Oregon	1950	11300	Brown	Demo.

CITY+STATE: Qualified city name. Wouldn't need STATE if cities had unique names.

CITY+POPULATION: A relationship involving the record's "subject".

CITY+MAYOR: Another such relationship.

CITY+YEAR: Identifying the complex subject of a fact. Not asserting a relationship.

MAYOR+PARTY: A relationship, but not involving the record's subject.

MAYOR+POPULATION: Accidental. No relationship, other than being facts about the same entity.

STATE+POPULATION: Accidental, misleading.

Figure 9 Multiple Fields