AN INFORMATION ALGEBRA

R. Bosak

System Development Corporation

I BASIC CONCEPTS

The Algebra is concerned with three undefined concepts; entity, property, and value. Although these concepts are formally undefined, certain intuitive statements can be made.

Data processing is the activity of maintaining and processing data to accomplish certain objectives. The data are collections of values of certain selected properties of certain selected entities. e.g. A payroll manager has many objectives of which the primary one is the payment of his employees (the entities). He selects certain properties such as employee number, name, sex, hourly payrate, and he maintains a file in which he records values of these properties for each entity.

II BASIC POSTULATES

Postulate 1: Each property has one and only one set of values assigned to it. The set of values so assigned will be called the property value set (V).

For each selected property the payroll manager has chosen he selects a property value set adequate to contain any anticipated value. Thus, the value set he selects for employee number may be the integers from 1 to 1,000,000.

Postulate 2: Every property value set contains at least two values

$\not{u}$ (Undefined, not relevant)

$\not{\theta}$ (Missing, relevant but not known)

If the payroll file carried additional personnel data, the property "draft status" would be undefined and not relevant for female employees, while if its value were missing for a male employee, it would be relevant but not known. In practice, recognizable symbols, such as 99999 or 88888, are assigned to represent these possibilities in the value set.

Postulate 3: Every entity has one and only one value assigned to it from each property value set.

For example, a certain employee may have the values (721, male, Roy, $15.25) i.e. employee number is 721, sex is male, etc. Note that (721, male, Roy, $15.25) is not a property value set as defined. Rather, it is a set of values of properties representing an entity.

III PROPERTY SPACES

Definition: A coordinate set (Q) is a finite, non-empty, ordered set of distinct properties.

Definition: The property space (P) of a coordinate set (Q) is the cartesian product set

$$P = V_1 \times V_2 \times V_3 \ldots \times V_n$$

where V_i is the property value set assigned to the ith property of Q. Each point (p) of the property space will be represented by an n-tuple

$$p = (a_1, a_2, a_3, \ldots a_n)$$

where a_i is some value from V_i.

The property space of the coordinate set (employee number., hourly payrate) contains all points obtained by taking one coordinate as an employee number from the value set of employee number and one coordinate as an hourly payrate from its value set. If the first value set contains 1,000,000 values and there are 15 hourly payrates in its value set, then P contains 15,000,000 points.

IV AREAS

Definition: An area (A) is any subset of P.

The property space P represents the totality of all possible information in a data processing system. An area A represents only a portion or subset of this information. Having defined sub-sets of P we can now apply the operations of set theory.

Definition: A function of areas is a mapping that assigns one and only one value to each area.

This definition allows us, among other useful things, to name areas. We could also sum all employee numbers in the previous example modulo 1,000,000.

V LINES

Definition: A line (L) is an ordered set of points,

$L = (p_1, p_2, p_3, \cdots, p_n)$ chosen from P.

n is the span of the line.

Definition: A function of lines is a mapping that assigns exactly one value to each line in P. The set of distinct values assigned by a function is the value set of the function.

The term line is introduced to provide a generic term for a set of points which are related. In a payroll application the datum points might be individual records for each person for each day of the week. These records might contain employee number, number of hours worked, rate of pay, and the date. For each employee we can define a line such that the first point is the record for Monday, the second Tuesday, etc. The span of the line is five. A function of such a line would be the value of the product (hours worked x rate of pay) summed over the 5 days of the week (points of the line).

VI GLUMPS

Definition: A glump G(g,A) of an area A for a function of lines g is a partition of A by g such that each element of this partition consists of all points of A that have identical values for g.

Definition: A function of glumps is a mapping that assigns an area to each glump such that there is a one-to-one correspondence between the elements of the glump and the points of the area, and such that each property of a point in the assigned area is a function of areas of the elements of the glump.

As an example, if g were hourly pay rate, one element of the glump would consist of all points of A for which pay rate was equal to $1.25, another element would be all points for which pay rate was equal to $1.50, etc. A function of glumps provides the capability to develop summary information over these groupings. This summary information might be such things as, the number of persons having each pay rate or the highest employee number in each pay rate group.

VII BUNDLES

Definition: An area set of order n is an ordered n-tuple of areas

$$(A_1, A_2, \ldots A_n) = \mathbb{A}$$

Definition: The bundle $B=B(b,\mathbb{A})$ of an area set $\mathbb{A}$ for a function of lines b is the set of all lines L such that if

$$\mathbb{A} = (A_1, A_2, \ldots A_n) \text{ and } L = (p_1, p_2, \ldots p_n) \text{ then}$$

(a) p_i is a point of A_i for i = 1, 2...n

(b) b(L) has the value "True".

Definition: A function of bundles is a mapping that assigns an area to each bundle such that

(a) There is a one-to-one correspondence between the lines in the bundle and the points in the area.

(b) Each property of each point in the area is a function of the corresponding line in the bundle.

The notions of a bundle and a function of bundles is a generalization and formalization of the process of matching files on some key and then producing a new file whose records are derived from information in the matching records.