The Management of a Large Data Base in IRIS†

PETER A. ALSBERG

Center for Advanced Computation,
University of Illinois at Urbana-Champaign, Urbana, Illinois 61801

Received November 15, 1974

IRIS (Illinois Resource Information System) is a retrieval and analysis system for land-use planners. The system stores the attributes of land parcels. Currently two major data bases are in use: a 78,000 parcel data base covering eight counties around Chicago and a 16,000 parcel data base covering six counties around Chicago.

The IRIS data structure is a two-level class-element hierarchy for each land parcel. The data classes are repeating groups with a fixed number of data elements/class. There are currently 18 classes with 4–30 elements/class. The data base is currently 20–25 million bytes of compressed data. The system can grow to accommodate over one million parcels with over a billion bytes of compressed data.

The user is a land-use planner with no computer science training. A simple query may scan the data in 10,000 parcels. Furthermore, those queries may use very complex data interpretations that require the analysis of large Boolean expressions for each parcel. To further compound the problem, the user needs exploratory access to this data base. This implies an interactive system which can scan and analyze a large number of records in a few seconds in order to meet a reasonable response time objective. Finally, the user has little money and is exceptionally cost conscious.

The development of a cheap, interactive, complex query, large data base system for use by noncomputer scientists requires some ingenuity. A very pragmatic approach had to be taken and coupled with a program to develop data management techniques with the required cost and performance parameters.

The major problems of storage costs and response time were solved by compressing the data base to a minimum number of bits. Since a data management system is traditionally heavily I/O bound, a reduction in the amount of data should produce a reduction in response time. For IRIS, the data compression is about 4:1. This reduces both response time and storage costs by about 4:1.

For each data element, the number of different possible values are calculated and used to determine a minimum field size (in bits) and an internal compressed data coding scheme. The element compression is semiautomatic. It requires that the data base builder indicate the range of all possible data element values, and that he specify whether collating sequences are to be preserved in the compression coding. The system automatically packs the various data fields and aligns them relative to word and byte boundaries to optimize retrieval. For example, data classes are aligned on a word boundary, the beginning of a new repetitive group is aligned on a byte boundary, and data elements are aligned so that they are not split by word boundaries (to reduce compare operations).

† This work was supported by the Ford Foundation, the Northeastern Illinois Planning Commission, and the Illinois Institute for Environmental Quality. This paper is a summary of a paper presented in the "Conference on Large Data Bases," sponsored by the NAS/NRC Committee on Chemical Information, National Academy of Sciences, Washington, D.C., May 22–23, 1974.

Frequently, one set of data element values predominates for a data class. For example, the forestry class frequently specifies that there are no trees on the parcel. In this case a default value of zero trees is stored for the forestry data class. The default value is stored once, in a separate file, and does not have to be repeated many times in the main file. This further compresses the data base.

The system keeps track of user and system usage parameters. These parameters are checked at the end of each session. If it is warranted, IRIS will automatically spawn a batch job that will change hash coding schemes and restructure some files to improve efficiency.

During retrievals, a query tuning algorithm detects patterns in the data records. The evaluation of Boolean expressions is altered on a microsecond to microsecond basis to take advantage of these patterns. IRIS attempts to do the minimum number of operations to either qualify or disqualify a record for query. The query tuning algorithm dramatically reduces the cost of evaluating complex queries. The technique will be described fully in a future paper.

A special retrieval and analysis language was designed for the land use planner. It uses his jargon and is organized to coincide with his approach to data analysis. The "region" is a user-defined index into the data base. These indexes are dynamically created and destroyed by the user. The regions scope the parcels of the data base in which the user is interested and eliminate I/O on parcels in which he has no interest. All logic is true, false, and maybe. The maybe allows for missing data. If parcels cannot be qualified or disqualified for inclusion in a study region, because of maybe's, they become part of an error region. The user is notified of this condition and can then examine the error region in more detail.

Many other facilities are made available to the users including:

1. abbreviations—a macro facility used to create interpretations and canned queries
2. weighting functions—an analysis technique commonly used by land use planners
3. multilevel libraries—individual, agency, and public libraries of regions, abbreviation, and functions
4. on-line help—diagnostics and complete data descriptions are available to define any data element or class and to retrieve and display the original command used to create any region, abbreviation, or function
5. the escape clause—a facility to prepare machine readable files, based on IRIS data, for use outside of the system.
6. data cleaning—extensive data cleaning facilities ensure the accuracy of the data base.