A THEORETICAL ANALYSIS ON DATA DEFINITION AND TRANSLATION

FINAL SCIENTIFIC REPORT

AFOSR 72-2219

ABSTRACT

Over the past four years of research for AFOSR, considerable progress has been made toward development of a data translation methodology. A model for implementing data translators has been formulated and verified through a series of increasingly more general data translators. Mechanisms for prescribing stored-data transformations and descriptions, a Stored-Data Definition Language, and Translation Definition Language to direct the data translator have developed.

1.0 INTRODUCTION

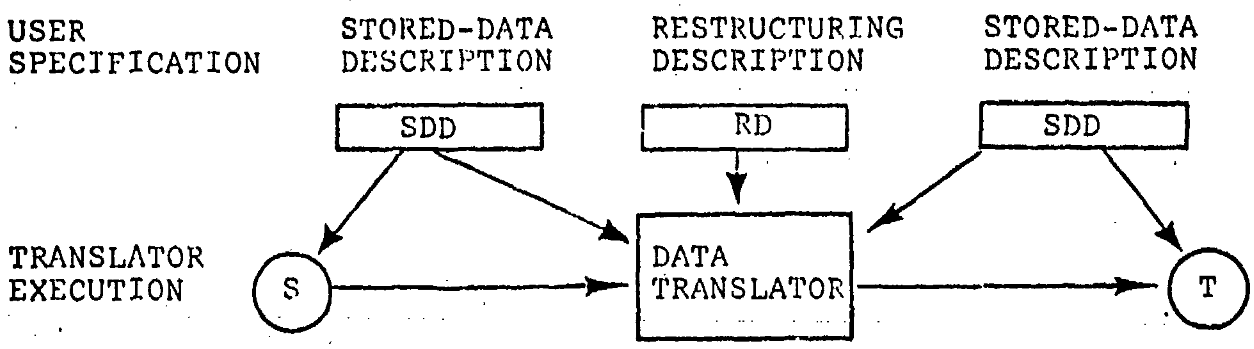
The outcome of the past ten years of accelerated growth in the computing industry has been the proliferation of data formats making it difficult to transfer data from one system to another. The state-of-the-art approach to this problem, termed data conversion is to develop a specific conversion program for each transfer of data from a source to a target system. This approach has the inherent disadvantage of requiring a different program to be written for each pair of source and target system. Hence for M (different) source systems and N (different) target systems, the number of programs required to translate data between differ t source and target system grows as the product of M and N increases.

Over the past four years of AFOSR funding, a substantive attack on the data conversion problem has been underway at The University of Michigan. Much progress has been made towards our overall goal of developing a data translation methodology to address the data conversion problem.

The descriptive approach to data base translation is based on a two-step process (Figure 1.1):

(i) the user specification of the necessary data descriptions

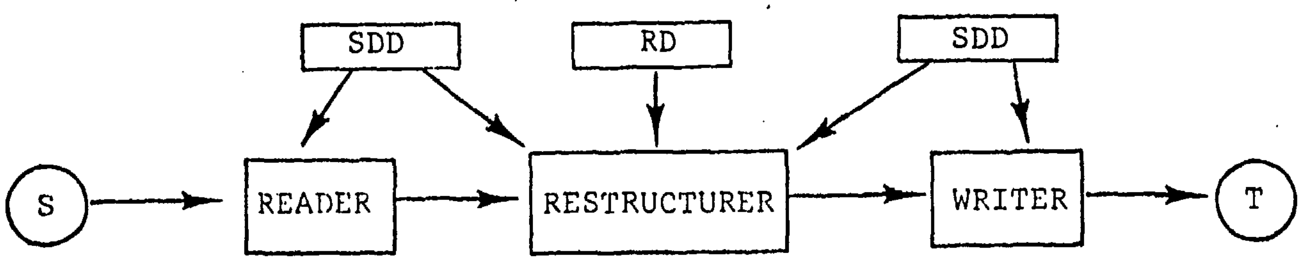
(ii) the execution of a data translator based on these descriptions



Data Description Approach
Figure 1.1

The data descriptions are the specification of the logical and physical attributes of the source and target data, along with the specification of the restructuring necessary to transform the source data into the target data.

The translation process consists of transforming the source data into the target data. This process is entirely driven by the stored-data descriptions prepared in Step i, and uses three components; a Reader, Writer, and Restructurer (Figure 1.2).



Components in the Translations Process

Figure 1.2

The Reader accesses the source data base, the Restructurer reorganizes the source data into a form suitable for the target, and the Writer outputs the target data base.

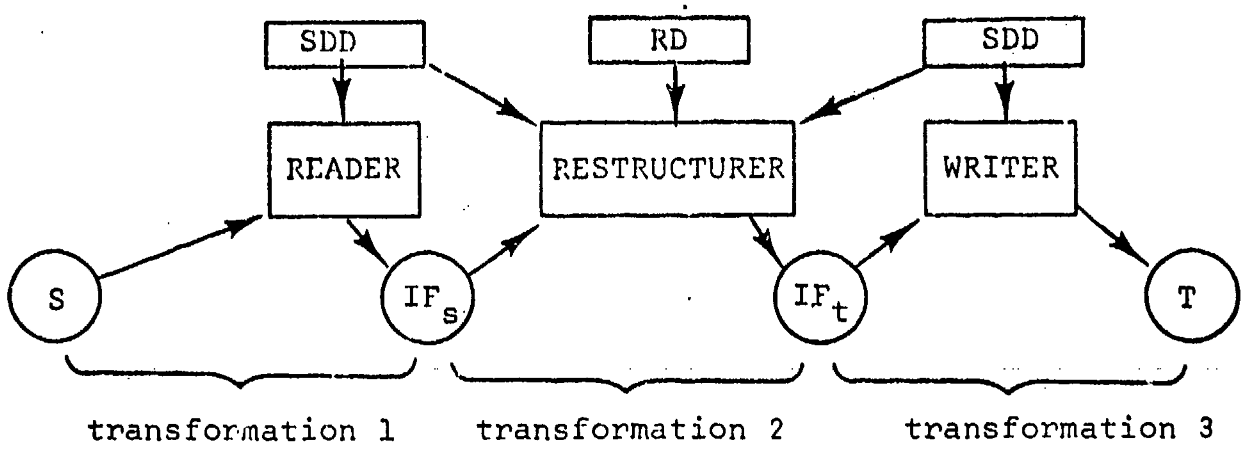
The Restructurer component is the most complex since it must perform sophisticated transformations in which large quantities of source data may be used to produce one instance of the target data. Both the Data Translation Project and SDDTTG have found it necessary to create an intermediate form for the source and target data but for different motivations. The SDDTTG calls this intermediate form a Translator Internal Form (TIF) and its objective is to be completely self-describing internal representation of source data. The internal form used by the Data Translation Project is termed a Restructurer Internal Form (RIF) and its objective is to facilitate the restructuring process.

Within the translation process (Figure 1.3), three distinct transformations on the data can be identified:

(i) Reading the source and producing the Internal Form of the source (IF_s)

(ii) Reading the IF_s and producing the Internal Form target (IF_t)

(iii) Reading the IF_t and producing the target



Data Transformations in Translation Process

Figure 1.3

2.0 SUMMARY OF ACCOMPLISHMENTS

Substantial progress has been attained in the development of sufficient descriptive mechanism and in the implementation of generalized data translators. Extending the basic declarative approach of data description languages, a Stored-Data Definition Language has been developed and successfully implemented to drive the translation process. Using similar technology, a Translation Definition Language has been specified to describe the logical transformation between data bases. Both of these results are basic research contributions.

With respect to the development of data translators, the initial development model proposed in 1972 proved sound and served as the basis for the implementation of a series of increasingly more general data translators. Developing the translator model immediately focused into three research directions:

1. A data accessing component
2. A data restructuring component
3. A data constructing component

Research on the data accessing component resulted in the development of a generalized access model which was driven by a high level device description. Investigation in data restructuring resulted in the formalization of data reorganization function which clearly delineated in the restructuring and reformatting capabilities. In order to develop a data restructurer, a model of data sufficiently general to handle hierarchical, network, and relational structures was developed. A set of restructuring operations based on this model was specified by three levels of abstraction: schema modification, instance operations, and value operations. The latter component identified a further research area--that of target file, evaluation and optimization.

Current research in the translation area is focusing on evaluation and selection of good structures for the target data base. Additional topics of research include interfacing a normal form of data to a source DBMS with the intention of decomposing the accessing problems into smaller subproblems that lend themselves to solution.

3.0 SPECIFIC ACCOMPLISHMENTS

3.1 Stored-Data Definition Language Research

The goal of a Stored-Data Definition Language (SDDL) is to describe the logical and physical characteristics of stored-data in a complete precise manner. Research in the synthesis and development of a Stored-Data Definition Language occurred primarily in 1971. The research on the synthesis of the language is documented in Taylor [1971] and Sibley and Taylor [1973].

To fully use the language as a tool, the langauge descriptions must be analyzed to ensure that they conform to the langauge specifications. Since the language is voluminous and inherently complicated, several research problems were identified. These are discussed in Metrick [1976].

3.2 Research Model for Translation

The development model for a data translator is presented in Fry et al [1972a]. This model is further enhanced in Fry et al [1972b], Sibley and Merten [1972], Merten and Fry [1974], and Fry [1974]. The basic model identifies three major processes: accessing the source data, reorganizing the source data into the target data, and constructing the target data. These processes correspond respectively to the Reader, Restructurer, and Writer components of the translation model. Each of these components have identified major research topics and are further described.

3.2.1 Reader

The Reader accepts the source file described in the SDDL, accesses the physical structure, and constructs a normal form. The initial efforts of the Reader research adopted a suggestion from Taylor [1971] where a string or pattern match was made with the input string. This, however, proved to be insufficient, and a generalized access model was developed (Yamaguchi [1975] and Frank and Yamaguchi [1974]). This model was driven by a high level device description. The language developed specifies access paths and addressing mechanisms for secondary storage.

3.2.2 Restructurer

The Restructurer is driven by the Translation Definition Language and performs logical translation of the data. The research of restructuring included the formulation and formalization of reorganization (Fry and Jeris [1974]). The formulation identified the two ends of the reorganization spectrum, reformatting and restructuring.

In the area of data base restructuring, results have been obtained in the specification of data models for restructuring, the formulation of restructuring operations, and the development of semantics for restructuring functions. Fundamental to the restructuring of data bases is a model of data which is rich enough in semantics to specify unambiguous restructuring transformations, but practical enough to perform the transformations efficiently. Navathe and Merten [1975] analyzed the Relational Model and discovered that the problems of mapping the source data to the normalized representation of the Relational Model outweigh the model's facility to use powerful manipulation languages.

Navathe and Fry [1976] and Navathe [1976] used a simplified version of the CODASYL data model to base their formulation of restructuring operations for hierarchical data models. They defined the restructuring process by three levels of abstraction: schema modification, instance operations, and item operations. At the schema level, three basic restructuring types were identified--Naming, Combining, and Relating. These types were defined by eight restructuring operations which serve to form the primitives for a Restructuring Language. The eight operations were further defined at the next level by eighteen data instance operations for the specification of restructuring algorithms. Finally, seventeen low level item operations were defined to manipulate the data base.

A further contribution has been the development of a data model for restructuring data bases [Deppe 1975]. This model not only handles the more complex network structure relationships, but in addition, allows the expression of how the various data constructs in the model are implemented. This result facilitates the specification of unambiguous structures so that meaningful transformations can be made in the generalized restructuring environment. The restructuring model of data uses a two level modeling process with a mapping between the levels. The first level, the Information Model specifies the relationships and information concepts of the real world by describing Entities and binary relationships among the Entities. The next level, the Data Model, defines the implementation of the Information Model structures by defining how the various structures are realized in systems. This two level approach provides sufficient information for the restructuring algorithm to make intelligent decisions about the restructuring operations specified by the user and the implied transformation on the data.

3.3.3 Writer

The Writer, the conceptual inverse of the Reader, constructs the target data base. Less research has been accomplished on the writing process itself, instead the research has focused on the optimization and choice of structure for the target which is further described in Section 3.3.

3.3 Optimization of Data Bases

Research on the optimization of the target structure began with Severence [1972] and Severence and Merten [1972] who described a simulation model capable of choosing an initial storage structure based on the criteria cost storage, retrieval speed, and data item usage. Later work (Yao [1974], Yao and Merten [1975]) developed an analytic model which selects an optimal file organization. The model uses usage parameters, environmental constraints, and a set of cost equations to achieve an optimal solution. Other research has focused on analysis and synthesis of file designs (Das and Teorey [1976], Yao et al [1976], Teorey and Das [1976]).

3.4 Operational Aspects of the Data Translator

The Data Translation process operates on two data bases (the source and target). Since the translation process may require a large amount of time, research efforts were directed to restart and recovery (Sayani [1972]) and microprogramming translation operations.

3.4.1 Microprogramming and its Relevance

With a view towards making data translation and restructuring more efficient, a research effort was initiated in the microprogramming area. Investigations were directed toward enhancement of microcoding and translation functions which could benefit from microcoding. DeWitt [1976] achieved some results in determining when two or more microoperations could be executed concurrently, thereby achieving further efficiencies at the microcode level. His approach utilized machine independent Control Word Model to define the semantics for the control words in microprogrammable computers. DeWitt [1975] discusses the applicability of microprogramming to the translation of data and the conversion of data base management systems. Many areas of applicability were found in which efficiencies could be realized through the increase in the level of control that microprogramming affords. The specification of a high level microprogramming language could alleviate some of the developmental problems and also take advantage of the concurrency available in most microprogrammed machines.

3.5 Extensions to the Translator Model

The translation of data is only one aspect of the conversion problem. Another conversion problem is involved in the translation of data base procedures. Research in this area has resulted in some initial formulations of the problem [Kintzer (1975)].

3.6 Related Research-Data Base Management Systems

The Data Translator's process may change as DBMSs evolve. Research in future directions of DBMSs are discussed in Fry [1975] and Fry [1973].

REFERENCES

TAYLOR 1971

TAYLOR, R. "A Data Definition and Mapping Language." Ph.D. Dissertation, University of Michigan, Ann Arbor, 1971.

FRY ET AL 1972a

FRY, J. P., ET AL. "A Development Model for Data Translation." Proc. 1972 ACM SIGFIDET Workshop on Data Description and Access.

FRY ET AL 1972b

FRY, J. P.; SMITH, D.; and TAYLOR, R. "An Approach to Stored Data Definition and Translation." Proc. 1972 ACM SIGFIDET Workshop on Data Description and Access.

SAYANI 1972

SAYANI, H. H. "A Decision Model for Restart and Recovery from Errors in Information Processing Systems." Ph.D. Dissertation, University of Michigan, Ann Arbor, 1972.

SEVERENCE AND MERTEN 1972

SEVERENCE, D. G. and MERTEN, A. G. "Performance and Evaluation of File Organization through Modelling." Proc. 1972 ACM National Conference, 1972.

SEVERENCE 1972

SEVERENCE, D. G. "Some Generalized Modelling Structures for Use in Design of File Organizations." Ph.D. Dissertation, University of Michigan, 1972.

SIBLEY AND MERTEN 1972

SIBLEY, E. H. and MERTEN, A. G. "Transferability and Translation of Programs and Data." Proc. COINS 1972 Symposium, Dec. 1972.

SIBLEY AND TAYLOR 1973

SIBLEY, E. H. and TAYLOR, R. "Data Definition and Mapping Language." Comm. ACM (Dec. 1973).

FRY 1973

FRY, J. P. "Towards the Specification of Requirements for and Utilization of Data Base Management Systems." Presented at IRIA International Seminar on Data Base Management Systems. Rocquencourt, France, 1973.

MERTEN AND SIBLEY 1973

MERTEN, A. G. and SIBLEY, E. H. "Transferability and Translation of Data." SIGPLAN Notices, Sept. 1973.

FRANK AND YAMAGUCHI 1974

FRANK, R. F. and YAMAGUCHI, K. "A Model for a Generalized Data Access Method." Proc. 1974 National Computer Conference, May 1974, pp. 45-52.

FRY 1974

FRY, J. P. "On the Implementation of a Physical Data Model for Data Translation." Presented at the International Workshop on Data Structure Models in Information Systems, forthcoming.

FRY AND JERIS 1974
FRY, J. P. and JERIS, D. W. "Towards a Formulation and Definition of Data Reorganization." Proc. ACM SIGFIDET Workshop on Data Description, Access and Control, May, 1974.

MERTEN AND FRY 1974
MERTEN, A. G. and FRY, J. P. "A Data Description Language Approach to File Translation." Proc. ACM SIGFIDET Workshop on Data Description, Access and Control, May, 1974.

YAO 1974
YAO, S. B. "Evaluation and Optimization of File Organizations through Analytic Modeling." Ph.D. Dissertation, University of Michigan, 1974.

DEWITTT 1975
DEWITT, D. "A Control Work Model for Detecting Conflicts Between Micro-operations." Microprogramming Workshop. Chicago, Ill., 1975.

NAVATHE AND MERTEN 1975
NAVATHE, S. B. and MERTEN, A. G. "Investigations into the Application of the Relational Model to Data Translation." Proc. 1975 ACM/SIGMOD Conference on the Managment of Data, May, 1975.

YAO AND MERTEN 1975
YAO, S. B. and MERTEN, A. G. "Selection of File Organization Using an Analytical Model." Proc. International Symposium on Very Large Data Bases. Framingham, Mass., Sept., 1975.

FRY 1975
FRY, J. P. "Significant Developments in Data Base Management Systems." Proc. Workshop on Data Bases for Interactive Design. Waterloo, Canada, Sept., 1975.

YAMAGUCHI 1975
YAMAGUCHI, K. "An Approach to Data Compatibility: A Generalized Data Access Method." Ph.D. Dissertation, University of Michigan, 1975.

KINTZER 1975
KINTZER, E. "Translating Data Base Procedures." Proc. 1975 ACM Annual Conference. Minneapolis, Minn., 1975, pp. 359-360.

DEPPE 1975
DEPPE, M. E. "A Restructuring Model to Specify the Semantics of Data Base Structures." Data Translation Project, Technical Report DT 75.2, Oct., 1975.

YAO ET AL 1976
YAO, S. B.; DAS, K. S.; and TEOREY, T. J. "A Dynamic Data Base Reorganization Algorithm." Submitted to ACM Transactions on Database Systems (TODS).

DEWITT 1976

DEWITT, D. "A Machine Independent Approach to the Optimization of Horizontal Microcode." Ph.D. Dissertation, University of Michigan, 1976.

NAVATHE AND FRY 1976

NAVATHE, S. B. and FRY, J. P. "Restructuring for Large Databases: Three Levels of Abstraction." Forthcoming in March issue of TODS.

METRICK 1976

METRICK, L. "Time Space Tradeoffs in the Automatic Programming of Finite State Machines." Ph.D. Dissertation near completion, University of Michigan, 1976.

TEOREY AND DAS 1976

TEOREY, T. J. and DAS, K. S. "Application of an Analytical Model to Evaluate Storage Structures." Submitted to 1976 ACM/SIGMOD Conference on Management of Data.

NAVATHE 1976

NAVATHE, S. B. "Towards a Methodology for Restructuring Databases." Ph.D. Dissertation near completion, University of Michigan, 1976.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFOSR - TR - 76 - Ø556

2. GOVT ACCESSION NO.:

3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): A THEORETICAL ANALYSIS ON DATA DEFINITION AND TRANSLATION,

5. TYPE OF REPORT & PERIOD COVERED: Final rept.,

6. PERFORMING ORG. REPORT NUMBER:

7. AUTHOR(s): A. G. Merten

8. CONTRACT OR GRANT NUMBER(s): AF-AFOSR [redacted] -2219-72

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
University of Michigan
Industrial Engineering Department
Ann Arbor, Michigan 48104

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:
9769-02
61102F

11. CONTROLLING OFFICE NAME AND ADDRESS:
Air Force Office of Scientific Research/NM
Bolling AFB, Washington, DC 20332

12. REPORT DATE: 1976

13. NUMBER OF PAGES: 14 (15p.)

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): AF-9769 9769Ø2

15. SECURITY CLASS. (of this report): UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Over the past four years of research for AFOSR, considerable progress has been made toward development of a data translation methodology. A model for implementing data translators has been formulated and verified through a series of increasingly more general data translators. Mechanisms for prescribing stored-data transformations and descriptions, a Stored-Data Definition Lanaguage, and Translation Definition Language to direct the data translator have developed.