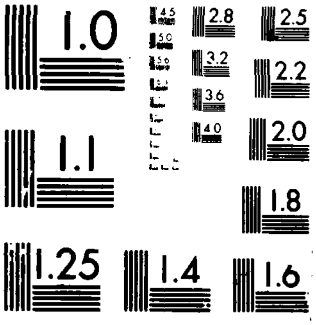
AD-A081 987   ILLINOIS UNIV AT URBANA-CHAMPAIGN COORDINATED SCIENCE LAB   F/G 9/2
BROWSING IN LARGE DATA BASES.(U)
FEB 79   D D DANKEL   N00014-75-C-0612
UNCLASSIFIED   WP-17   NL

1 OF 1
AD
A081987

END
DATE
FILMED
4-80
DTIC


1.0
4.5
5.0
5.6
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

(9) WORKING PAPER 17 (14) WP-17

# (6) Browsing in Large Data Bases.

(10) Douglas D. Dankel, II

Advanced Automation Group
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801

(11) Feb[redacted] [redacted]79 (12) 21

## Abstract

The formation on data bases containing information on mechanical systems for trouble shooting purposes has become increasingly popular and important. Examination of these data bases by humans can be very costly. A system called BROWSER is currently under development to heuristically search a data base containing information on Navy aircraft with little or no human intervention. BROWSER searches the data base guided by models and heuristics looking for interesting patterns or configurations. The user is then notified of the existence of these patterns.

Key Words and Phrases:

Heuristic Search, Pattern Identification, Data Modeling, Data Bases, Pattern Recognition, Alerters

(15) This work is supported by the Office of Naval Research under Contract N00014-75-C-0612 ncw

Working papers are informal papers intended for informal use.

BROWSER is an automated system for troubleshooting with data bases. It is based on Douglas Lenat's work on automated discovery in mathematics [LENA76, LENA77b] and is designed with a modified production system architecture [DAVI75, LENA77a]. BROWSER explores a data base using models describing the data and a large collection of data dependent and data indepentent heuristics.

The exploration of the data base is performed by defining subsets of data which are then examined using simple statistical techniques. BROWSER looks for regularities within a subset such as recurring sequences of data or unusually high occurrence rates of a particular datum. When comparing two subsets of data BROWSER looks for statistically different rates in the occurrence of a particular datum. These discoveries or facts about the data base which are found by BROWSER are reported to the user. The user could then be given the option of creating alerters [BUNE75, COHE77] to monitor new data as it enters the data base for the occurrence of these facts. This exploration of the data base is best described by the pattern recognition term of feature extraction [ANDR72].

BROWSER is designed to operate with little or no human intervention and is intended to be the prototype for a future portable data base system.

Why Data Bases?

Marvin Denicoff [DENI77], when discussing naval research programs in artificial intelligence, stated:

> The 1972 objective was to push strongly in the direction of activist systems which are capable of carrying on extensive English language dialogs, which have an automated capacity for comprehending relationships across data, for making inferences, for using that facility to take initiative, in real time, to alert managers to the impending occurrence of critical events. Knowledge systems used to anticipate and prevent crisis, not computers used to reconstruct the history of past disasters is the ultimate Large Data Base research objective.
>
> From the user perspective, the research includes ... the development of a natural language front end to enable direct contact with computers by high level managers -- this interaction ranging from "one liners" to sophisticated on-line browsing and

simulation; building a capacity for anticipating crisis situations through the setting and monitoring of automatic alarms or triggers.

The development of "activist systems" such as BROWSER is needed due to the time comsuming nature of browsing as a human task. Consider the tasks of trying to find similarities between the causes of the crashes of several aircraft of a common type, the causes or early warning signs for various diseases, the characteristics of tax returns which should be examined by an audit or advanced warnings of severe weather conditions. All of these tasks require a large amount of data and valuable time spent to sift through the data. A browsing computer system can also perform these tasks.

Only limited examination has been made of browsing systems. Douglas Lenat [LENA76, LENA77b] has developed AM, a system for automated discovery in mathematics. AM is a self-driven system which starts with a knowledge base of mathematical concepts. It applies a large collection of heuristics to this knowledge base to increase the knowledge by either filling in unknown information in the knowledge base or expanding it. This approach is similar to that used by a mathematician in mathematical research: using a large collection of heuristics to define and study new concepts.

## Development of Browsing Systems

The development of data base browsing systems can be accomplished by breaking down the knowledge required to search a data base. This knowledge is easily broken into two parts: the knowledge which remains constant across all types of data bases and the knowledge which is dependent upon the particular data in the data base. Data independent knowledge includes how to define and create subsets of data, how to explore the subsets using statistical techniques and what types of information are needed about each subset. Types of data dependent knowledge include how the data is organized in the data files, which data fields are more important to examine, which data values are most interesting, and how the objects from which the data was gathered are composed.

In the development of BROWSER several design principles were formulated. These include:

1. Little if any natural language interaction is required. Most interactions with a user can be performed using menus or very simple question/answer formats. More complex user interactions can be developed depending upon the particular application.
2. The knowledge needed about each subset of data is clearly defined. This knowledge is represented as

"faceted concepts". Each subset of data is described by a concept consisting of a number of facets. Each facet holds characteristic knowledge about the subset.

3. The basic operation performed by BROWSER is defining new data subsets and then examining these subsets for regularities and differences. This operation can be broken into two basic types of actions: actions to fill in facets of a concept which are empty and actions to explore and examine the subsets of data associated with the concepts. This exploration and examination can suggest that new concepts and their associated data subsets be created.

4. The knowledge required by BROWSER can be broken into data dependent and data independent knowledge. The data dependent knowledge consists of sets of heuristics and models and is used by the system to decide on which data to perform actions. The data independent knowledge consists of a set of heuristics which perform the operations of the two basic actions described above. The data independent knowledge is designed to apply across all data bases. This allows a new browsing system to be developed for a different data base by changing only the data dependent

knowledge.

5. The heuristics are independent units of knowledge and, in general, do not strongly interact.

6. The set of heuristics need not grow as new subsets of data are defined. The growth of the system will occur in the definition of new subsets and the discovery of knowledge about these subsets.

7. The models provide knowledge on what data are contained in the data base and how the data are organized. They provide the intelligence of the browsing.

8. The models are used to show where data specific heuristics apply within the data, organizing this knowledge so it can be used efficiently. In this way, the models are used to store the heuristics.

## Human Troubleshooting

Troubleshooting is a term normally used in conjunction with mechanics and machinery. In that context troubleshooting is defined as the "... ability to locate a malfunction in a piece of equipment." [HIGH55] In order to troubleshoot a piece of malfunctioning equipment "... the individual must (1) have some knowledge of how the system functions normally, (2) obtain information about the current state of the system, (3) relate the information he gets to his conception of the normal system, his past experience with malfunctions of this or similar systems, and his theoretical knowledge of functional relationships embodied in the system, and (4) formulate and test hypotheses as to the probable source(s) of the malfunction." [GRIN53]

When troubleshooting with a data base the same techinique is used. Rather than attempting to locate a problem in a piece of equipment so the equipment can be repaired, in a data base one tries to locate a problem which has occurred and to forewarn of its future occurrence.

BROWSER is implemented with a data base containing information on maintenances which have been performed on navy aircraft [TENN78]. BROWSER attempts to find sequences of maintenance actions which preceded a particular maintenance action. Such an action might be the failure of a component in a system. If such a pattern can be found, it could be used to give warning of the possible failure of that component. When one of the preceding events occurs the part which is due to fail could be repaired or replaced eliminating possible adverse conditions which might later occur due to the failure of the component (e. g., the aircraft crashing). Additionally, BROWSER can compare aircraft with very good maintenance histories to aircraft with very poor maintenance histories. The intent of this comparison is to find differences which can lead to modifications on the aircraft with poor histories to improve their system performance.

For humans to perform these tasks is extremely time consuming due to the large amounts of data involved. The computer is able to perform the complex searching and examinations much quicker than a human, making it ideal for the task. But to perform these tasks it needs the complex knowledge used by the human when he/she searches.

## Important BROWSER Components

BROWSER has three significant components. The first is how known facts and knowledge are represented. This knowledge is contained in the data dependent and independent heuristics and in the models. Second is the controlled execution of tasks within the system. For this, BROWSER uses an agenda to store an ordered list of all tasks which are to be executed with the highest priority task always being executed first. Finally, new knowledge is represented and stored in concept trees. Each concept represents a subset of the data and contains several facets which contain information on the subset.

## Knowledge Sources

Three sources of knowledge are used by BROWSER: knowledge of the data base, knowledge of the operation of BROWSER, and knowledge of the concepts. The most important knowledge source is knowledge of the data base. This knowledge is stored in the models and data specific heuristics. Five general types of models exist in BROWSER (see Figure 1). The first four model types provide detailed knowledge about the data in the data base while the fifth provides general knowledge on data structures useful for examining the data.

The lowest level models are the Data Base Models. These models provide information on the format of the data within the data files. A Data Base Model exists for each type of data file in the data base. The models contain information on the data fields which compose the data file. This information includes the field's retrieval name (used in the data base query language), an equivalent english name, the starting position of the field in a data record, the length of the field, and the type of value contained in the field (numeric or character). The Data Base Models are used mainly by the data base retrieval system for obtaining the data from the data base.

The Data Specific Models provide information on how the various data fields are related to each other. These models show the semantics of the data fields. Again, one Data Specific Model exists for each type of data file in the data base.

The Specific Models are of two varieties. The first shows how the data records, described semantically in the Data Specific Models, combine to form larger units known as conceptual units. In BROWSER an example conceptual unit is a performed maintenance. Each performed maintenance consists of several actions, such as the removal of an aircraft part, the repair of a part or the installation of a part. Each action is a data record in the data base which combines with other actions to form a single unit described by the conceptual unit of a maintenance action. The second type of Data Specific Models describes the objects from which the data was gathered. For BROWSER these models are of the F-4 and A-7 aircraft.

Typical Models are the most general data specific models. They describe generalizations of the Specific Models. Common knowledge which applies across Specific Models that can be generalized is stored in the Typical Models. BROWSER's Typical Models include a model of a typical aircraft and the model of an aircraft's history. The typical aircraft model represents common knowledge which exists about all aircraft.

The aircraft's history model shows how the conceptual units of performed maintenance combine to form the overall record of an aircraft.

The final model type is that of the General Models. They are used to provide information on the typical types of data structures which exist within the data. These models include a network, a tree, and a list.

Each of the model types are interconnected. The interconnections show how components of the different models relate to each other and provide means for the sharing of knowledge stored in the models. The shared knowledge is represented in the structure of the models and in heuristics which are attached to the models. The heuristic knowledge is attached to the places in the models where it is applicable. These heuristics provide information on what is important within the model; the models provide the structure of how the knowledge is interrelated.

The second source of knowledge in the system is general system knowledge. This knowledge deals with how the agenda of tasks which the system must execute is organized and how the system focuses its attention on particular tasks. Included here are all of the various system checks and actions which must be performed. This knowledge is represented in 10 to 20 general heuristics separately coded and numerous other heuristics embedded in the system code.

The final source of knowledge in BROWSER is knowledge of the concept trees. This knowledge is stored as heuristics concerned with creating new concepts to describe subsets of data, filling in the facets of concepts so facts discovered about the subsets are not lost, and exploring the data through the use of a statistical package. This heuristic knowledge is represented by a total of approximately 100 heuristics.

## The Agenda

The second significant feature of BROWSER is the use of an agenda to control the execution of tasks within the system. The basic operation performed by BROWSER is to define new subsets of data and explore them. At any one time BROWSER might have 100 or more subsets of data that it is considering to investigate, many additional subsets which are currently being examined, and some subsets which have already been fully examined. The system must have some means of keeping track of all of the tasks associated with the subsets which must be performed. The agenda serves this purpose.

The agenda is an ordered list of tasks which BROWSER is currently considering for execution. When examining a subset of data certain tasks must be performed before others. The agenda assures the proper ordering of these tasks. It allows the subsets of data to be explored in an orderly fashion. Those subsets which appear to be most rewarding are examined first. At any given time only a few tasks of the hundreds on the agenda will appear to be most worthwhile to examine. The agenda allows these tasks to be examined first.

The ordering of the tasks is determined by several factors. Certain tasks must be executed before others, for example a definition of the data contained in a subset must exist before that subset can be retrieved from the data base and the subset must be retrieved before it can be examined for regularities. Additionally, each task has a list of reasons why it should be executed, each reason has a value and the sum of all of the values, the task priority, gives the position of the task within the agenda.

The basic procedure of BROWSER is as follows. All tasks are sorted according to their priorities. The highest priority (most worthwhile) task is removed to be executed. The heuristics which are relevant to the execution of this task are gathered from all of the knowledge sources in the system: the models and their associated heuristics, the concept tree heuristics and the general system knowledge. These heuristics are then executed. When all of the relevant heuristics have been executed the cycle is repeated.

Each heuristic is defined with a left and right hand side. The left hand side represents conditions which must be satisfied and the right hand side contains actions to be performed. If the conditions of the left hand side are satisfied the actions of the right hand side of the heuristic are executed. This execution can cause new tasks to be suggested, values to be given to facets of a concept, suggestions that new concepts be created or suggestions of new reasons for existing tasks.

Some additional comments should be made concerning the heuristics. Because a large number of heuristics exist in BROWSER and only a small portion of them will be relevant at any time, the heuristics are stored in such a way that only those heuristics which apply to the current task will be gathered and executed. This storage and retrieval of relevant heuristics allows all of the relevant heuristics to be executed because little time is wasted on pursuing worthless heuristics. The heuristics associated with the data models are the most relevant and give important information to the rest of BROWSER. The general system heuristics and some heuristics associated with the facets tend to be of a bookkeeping nature not intended to discover new information but rather to provide necessary actions within BROWSER.

## Representation of New Knowledge

The final significant feature of BROWSER is how new knowledge is represented and stored. This new knowledge consists of facts gathered by BROWSER from the data base. This knowledge is stored as faceted concepts on concept trees. Initially the data base is described by a small set of concepts: ALL-PLANES (the entire data base), CRASH (data on aircraft which have crashed), NON-CRASH (data on aircraft which have not crashed), HI-MAIN (data on aircraft requiring excessive maintenance), and LO-MAIN (data on aircraft with good maintenance histories). BROWSER creates more subsets and their associated concepts by examining existing subsets of data and offering ideas on how the existing subsets might be restricted to form smaller subsets or, possibly, generalized to form larger subsets. Each new subset is described by a concept in a concept tree with its facets holding the important and relevant information desired about the subset. Figure 2 show the initial concept trees with an additional concept added. Note that the subsets of data described by the concepts can intersect in their coverage of the data base, but that new subsets are always created from existing subsets as shown by concept X in Figure 2.

The concepts hold the relevant information needed about the data subsets. They represent the subsets and show their relationship to the other subsets. Each concept consists of a number of facets which describe it. Typical facets include: NAME - the name used to refer to the subset; DEFN - the definition of the data contained within the subset; INTEREST - a numeric measure of how interesting the subset of data is; GENERALIZATIONS - pointers to more general (larger, containing) subsets; SPECIALIZATIONS - pointers to more restricted (smaller, contained) subsets.

## Performance

Two browsing systems were developed - a dumb system which contains a bare minimum of knowledge of the data base and a smart system which contains enough knowledge to intelligently browse through the data. The dumb system contains all of the general knowledge of browsing, knowledge of the concept trees and the Data Base Models for the data files contained in the data base. This system either selects data ranges or particular data which prove to be statistically significant or randomly selects data when none prove to be significant. The smart system contains all of the knowledge in the dumb system plus the Data Specific Models for the particular data files in the data base. The higher level models have not been implemented as yet, due in part to the limited amount of aircraft data contained in this initial implementation of BROWSER.

While the difference in knowledge appears to be only slight, the smart system is able to focus its attention on important data and relationships much more quickly. Theoretically the dumb system is capable of discovering everything which the smart system finds. The discovery, however, could take a significant amount of time due to the randomness of its operation.

## Conclusion

The current implementation of BROWSER is designed to test the feasibility of browsing systems. While it has appeared to be successful in its initial testing, several needed changes have become apparent, both for the current system and for similar systems which might be developed.

1. Only a limited amount of data dependent knowledge currently exists in BROWSER. The limitation is due to the small data base that BROWSER currently examines. More of the available data on the aircraft should be added to the data base with the knowledge needed to describe it. Higher level models can then

be implemented, tested, and studied for their usefulness.

2. A more extensive front-end could be provided allowing user specification of new heuristics and/or models, specific tasks to be performed or restrictions on the type of discovered knowledge which should be reported to a given user. The front-end could also be extended to provide more knowledge on why it performs specific actions and how it actually performs them.

3. The system could monitor its own heuristics and their performance. This self-monitoring might offer suggestions leading to improvements in the performance of the overall system. Using this technique the dumb system might be capable of discovering data dependent knowledge contained in the smart system or the smart system might discover new unknown data dependent knowledge.

4. Multiple users could be allowed. Each user could specify his/her specific interests and desires. The system would then perform its normal operations and would report specific information only to the users which specifically requested it.

5. In its exploration, BROWSER might compute some values from the data which itfinds to be extremely

valuable or find some data fields which appear highly interrelated. These discoveries could lead to a restructuring of the data base. Heuristics could be added to look for such occurrences.

6. BROWSER could act as a data validator in addition to its other operations. While exploring the data subsets it could look for possible errors in the data, reporting these errors to a user for correction.

The development of BROWSER and similar systems can provide important ideas for the development of more complex systems. Valuable information about the objects which the data describes could be provided by these systems with a significant savings in costs for the users.

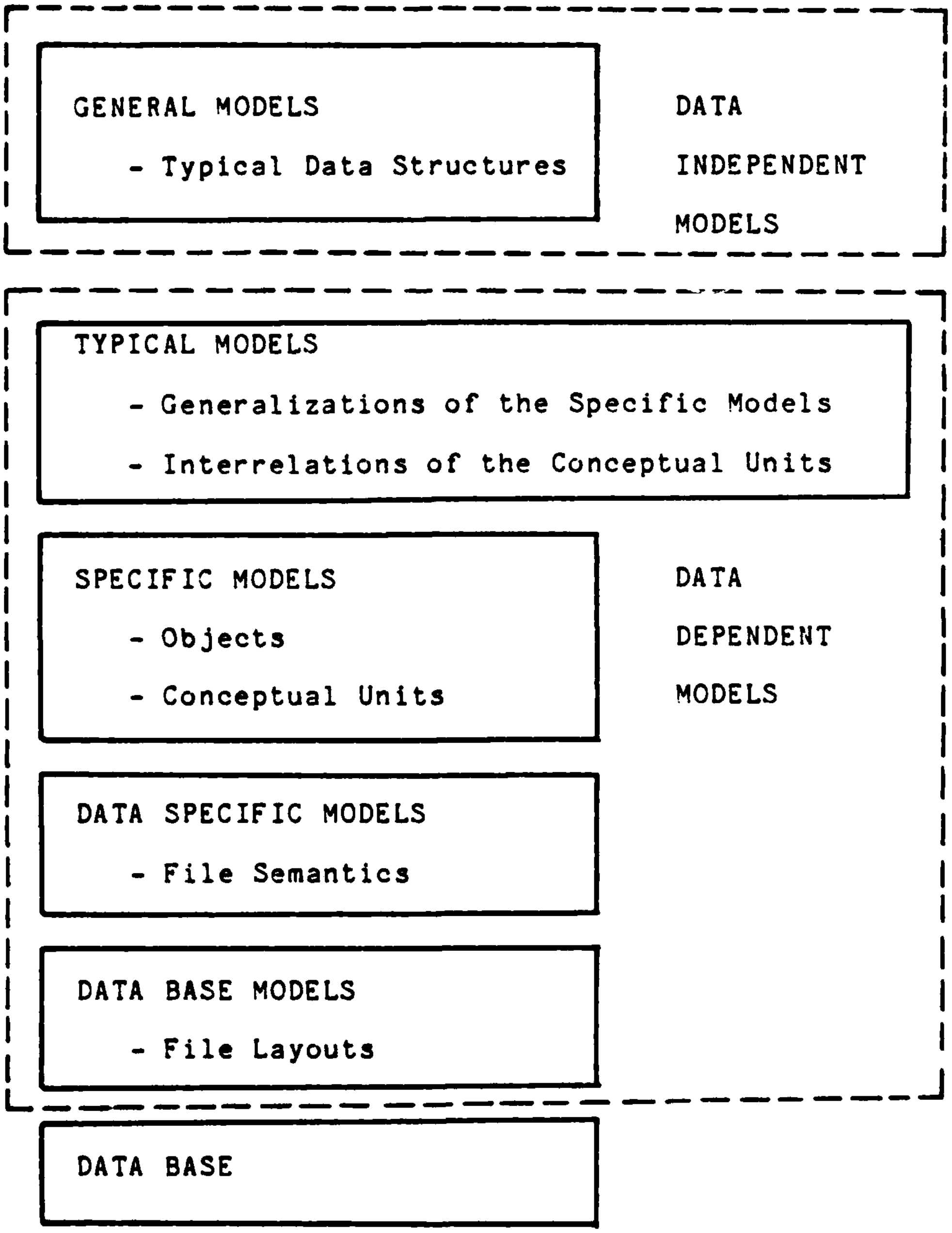


Figure 1. The Model Types

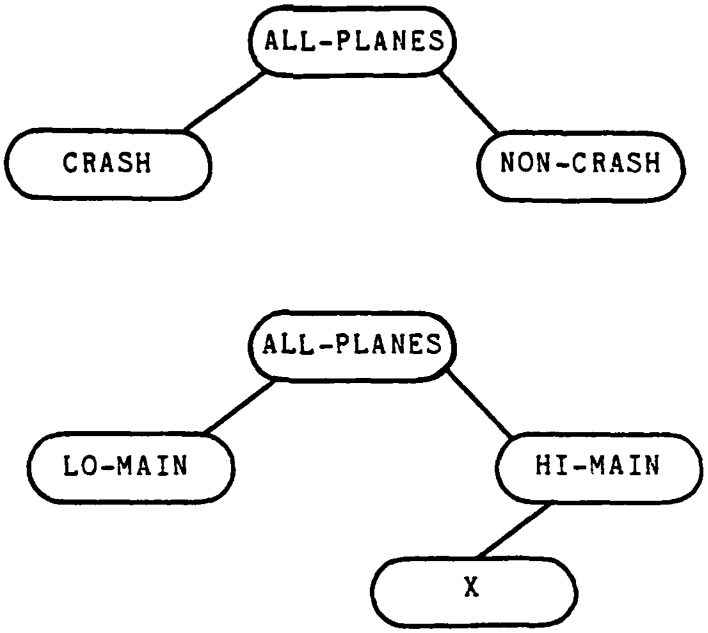


Figure 2. Initial Concept Trees with an added Concept X

## References

[ANDR72] Andrews, H. C., INTRODUCTION TO MATHEMATICAL TECHNIQUES IN PATTERN RECOGNITION, Wiley-Inter Science, N. Y., 1972

[BUNE75] Buneman, O. P. and H. L. Morgan, "Alerting in Database Systems: Concepts and Techniques", The Wharton School, University of Pennsylvania, Dec. 1975

[CONE77] Cohen, S. F., "Alerters in Network Databases", The Wharton School, University of Pennsylvania, Feb. 1977

[DAVI75] Davis, R. and J. King, "An Overview of Production Systems", STAN-CS-75-524, Computer Science Department, Stanford University, Oct. 1975

[DENI77] Denicoff, M., "Office of Naval Research Programs in Artificial Intelligence", 5th International Joint Conference on Artificial Intelligence, M. I. T., Cambridge, Mass., Aug. 1977, pp.947-948

[GRIN53] Grings, W. W., J. W. Rigney, N. A. Bond, and S. A. Summers, "A Methodological Study of Electronics

Trouble Shooting Skill: I. Rationale for and Description of the Multiple-Alternative Symbolic Trouble Shooting Test.", Technical Report No 9, Project Designation NR153-093, Contract NONR-228(02), University of Southern California, Aug. 1953 summarized in Standlee, L. S. and W. J. Popham, N. A. Fattu, "A Review of Trouble Shooting Research, Research Report No 3", Institute of Educational Research, Indiana University, Dec. 1956

[HIGH55] Highland, R. W., "A Guide for Use in Performance Testing in Air Force Technical Schools", Technical Memorandum, ASPRL-TM-55-1, Lowry Air Force Base, Colorado, Jan. 1955 summarized in Standlee, L. S. and W. J. Popham, N. A. Fattu, "A Review of Trouble Shooting Research, Research Report No 3", Institute of Educational Research, Indiana University, Dec. 1956

[LENA76] Lenat, D. B., "AM: An Artificial Intelligence Approach to Discovery in Mathematics as Heuristic Search", SAIL-AIM-286, Stanford University, July 1976

[LENA77a] Lenat, D. B. and J. McDermott, "Less than General Production System Architectures", 5th International Joint Conference on Artificial Intelligence, M. I. T., Cambridge, Mass., Aug. 1977, pp. 928-932

[LENA77b] Lenat, D. B., "Automated Theory Formation in Mathematics", 5th International Joint Conference on Artificial Intelligence, M. I. T., Cambridge, Mass., Aug. 1977, pp. 833-842

[TENN78] Tennant, H., "The PLANES Database", Working Paper 14, Advanced Automation Group, Coordinated Science Laboratory, University of Illinois, June 1978

IED

80