N87-29132

INTELLIGENT DATA MANAGEMENT

WILLIAM J. CAMPBELL
SPACE DATA AND COMPUTING DIVISION
DATA MANAGEMENT SYSTEM FACILITY
NASA/GSFC

ABSTRACT

INTELLIGENT DATA MANAGEMENT IS THE CONCEPT OF INTERFACING A USER TO A DATABASE MANAGEMENT SYSTEM WITH A VALUE ADDED SERVICE THAT WILL ALLOW A FULL RANGE OF DATA MANAGEMENT OPERATIONS AT A HIGH LEVEL OF ABSTRACTION USING HUMAN WRITTEN LANGUAGE. THE DEVELOPMENT OF SUCH A SYSTEM WILL BE BASED ON EXPERT SYSTEMS AND RELATED ARTIFICIAL INTELLIGENCE TECHNOLOGIES, AND WILL ALLOW THE CAPTURING OF PROCEDURAL AND RELATIONAL KNOWLEDGE ABOUT DATA MANAGEMENT OPERATIONS AND THE SUPPORT OF THE USER WITH SUCH KNOWLEDGE IN AN ON-LINE, INTERACTIVE MANNER. SUCH A SYSTEM WILL HAVE THE FOLLOWING CAPABILITIES:

- AN UNDERSTANDING BETWEEN SCIENCE APPLICATIONS AND STORED DATA
- THE ABILITY TO CONSTRUCT A MODEL OF THE USERS VIEW OF THE DATABASE, BASED ON THE QUERY SYNTAX
- THE ABILITY TO TRANSFORM ENGLISH QUERIES AND COMMANDS INTO DATABASE INSTRUCTIONS AND PROCESSES
- THE USE OF HEURISTIC KNOWLEDGE TO RAPIDLY PRUNE THE DATA SPACE IN SEARCH PROCESSES
- AN ON-LINE EXPLANATION SYSTEM THAT WILL ALLOW THE USER TO UNDERSTAND WHAT THE SYSTEM IS DOING AND WHY IT IS DOING IT

SUCH A SYSTEM WILL BE GOAL ORIENTED RATHER THAN PROCEDURE ORIENTED SUPPORTING VARYING LEVELS OF ABSTRACTION OF THE DATA OF INTEREST TO THE USER.

INTELLIGENT DATA MANAGEMENT

Presentation at the
Computer Science/Data Systems Technical Symposium
Leesburg, Virginia

April 16-18, 1985

William J. Campbell
Data Management Systems Facility
NASA/Goddard Space Flight Center

THE INFORMATION EXPLOSION PROBLEM/SOLUTION

PROBLEM

- A PRIMARY PROBLEM IN EARTH SCIENCE RESEARCH IS COPING WITH THE MASSIVE AMOUNTS OF REALTIME AND ANALYTICAL DATA BEING GENERATED
- THE AMOUNT OF AVAILABLE INFORMATION AND DATA ALREADY EXCEEDS THE SCIENTIST'S ABILITY TO MANAGE AND USE IT
- USING PRESENT TECHNOLOGIES ONLY A SMALL PERCENTAGE OF THE DATA WILL EVER BE UTILIZED

SOLUTION

- THE APPLICATION OF ARTIFICIAL INTELLIGENCE (AI) TO EARTH SCIENCE INFORMATION MANAGEMENT OFFERS THE ONLY REAL POSSIBILITY OF REVERSING THIS TREND

WHAT IS ARTIFICIAL INTELLIGENCE (AI)

FROM THE EARTH SCIENCE VIEW, WE DEFINE AI AS:

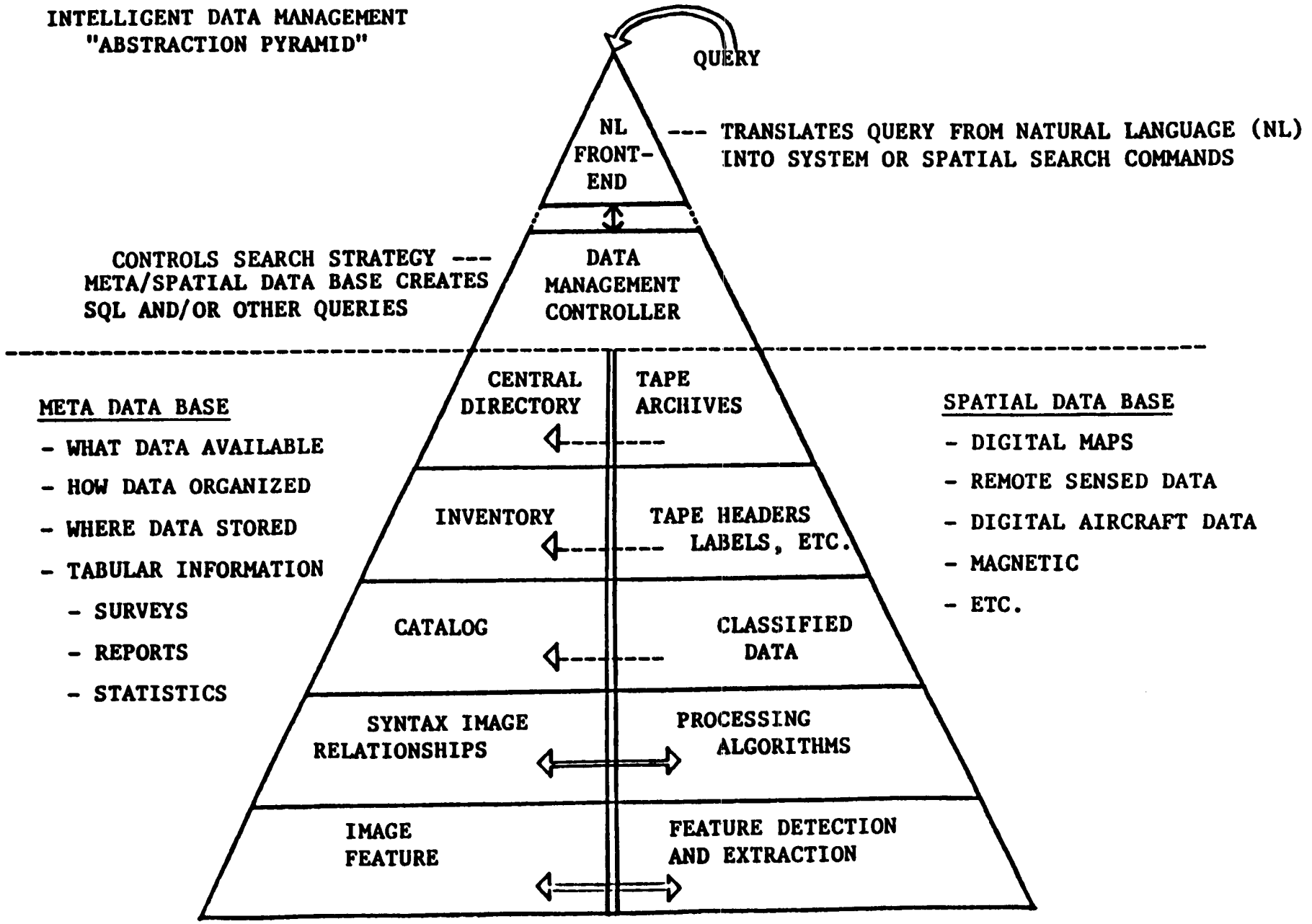
- A VALUE ADDED SERVICE THAT EXTENDS THE OPERATIONAL CAPABILITIES OF A USER BY PROVIDING A SYSTEM THAT CAPTURES HUMAN KNOWLEDGE FOR SUPPORTING PROCEDURAL, OPERATIONAL AND RESEARCH TASKS IN THE AREA OF DATA MANAGEMENT FOR EARTH SCIENCE RESEARCH

HOW INTELLIGENT DATA MANAGEMENT CAN SUPPORT EARTH SCIENCE

- PROVIDE VALUE-ADDED SERVICES TO SYSTEMS THAT SUPPORT EARTH SCIENCE RESEARCH
- PROVIDE THE ABILITY TO MINIMIZE THE INFORMATION PROCESSING AND MANAGEMENT TIME, WHILE MAXIMIZING THE RESEARCH AND APPLICATION TIME
- THE MAJOR CONTRIBUTIONS ARE EXPECTED TO BE:
  - THE UNBURDENING OF THE SCIENTIST FROM THE MORE MUNDANE TASKS OF DATA IDENTIFICATION, SELECTION AND MODIFICATION
  - THE EXTENSION OF HUMAN COGNITIVE PROCESSES TO OPERATIONS THAT SUPPORT THE MANAGEMENT, ANALYSIS, UNDERSTANDING AND APPLICATION OF EARTH RELATED DATA
  - THE CAPTURING OF HUMAN KNOWLEDGE INTO SYSTEMS THAT CAN BE USED BY NON-EXPERTS

KNOWLEDGE-BASED EXPERT SYSTEMS

- WOULD SUPPORT THE USER WITH EXPERT KNOWLEDGE IN DATA SEARCH AND FEATURE IDENTIFICATION
- WOULD PROVIDE PROCEDURAL KNOWLEDGE TO SUPPORT DATA MANAGEMENT SYSTEM OPERATION AND EXPLANATIONS
- WOULD PROVIDE SELF CHECKING AND CORRECTION PROCESSES
- SUPPORT THE COLLECTION OF HEURISTIC KNOWLEDGE ON A REALTIME ON-GOING BASIS
- WOULD SUPPORT APPROXIMATE REASONING TO INFER CONCLUSIONS THAT ARE NOT EXPLICITLY STATED BY THE USER (I.E. FUZZY QUERIES)
- WOULD USE HEURISTIC SEARCH PROCESSES THAT LEAD TO SHORTCUTS BY EARLY PRUNING OF LARGE PORTIONS OF UNWANTED DATA SPACE


INTELLIGENT DATA MANAGEMENT
"ABSTRACTION PYRAMID"
QUERY
NL FRONT-END
--- TRANSLATES QUERY FROM NATURAL LANGUAGE (NL) INTO SYSTEM OR SPATIAL SEARCH COMMANDS
CONTROLS SEARCH STRATEGY --- META/SPATIAL DATA BASE CREATES SQL AND/OR OTHER QUERIES
DATA MANAGEMENT CONTROLLER
META DATA BASE
- WHAT DATA AVAILABLE
- HOW DATA ORGANIZED
- WHERE DATA STORED
- TABULAR INFORMATION
- SURVEYS
- REPORTS
- STATISTICS
CENTRAL DIRECTORY
TAPE ARCHIVES
INVENTORY
TAPE HEADERS LABELS, ETC.
CATALOG
CLASSIFIED DATA
SYNTAX IMAGE RELATIONSHIPS
PROCESSING ALGORITHMS
IMAGE FEATURE
FEATURE DETECTION AND EXTRACTION
SPATIAL DATA BASE
- DIGITAL MAPS
- REMOTE SENSED DATA
- DIGITAL AIRCRAFT DATA
- MAGNETIC
- ETC.

NATURAL LANGUAGE

- NATURAL LANGUAGE MEANS AUTOMATED HUMAN LANGUAGE UNDERSTANDING
- NATURAL LANGUAGE PROCESSING FOR SUPPORTING EARTH SCIENCE INVOLVES THE DEVELOPMENT OF COMPUTER PROGRAMS WHICH CAN ANALYZE HUMAN LANGUAGE AND PERFORM THE APPROPRIATE ACTION ON THE INFORMATION CONTAINED IN THE TEXT OR UTTERANCE
- PRESENT NATURAL LANGUAGE APPLICATIONS ARE LIMITED TO AREAS THAT REQUIRE A LIMITED DICTIONARY (E.G. DATABASE QUERIES)

THE RELATIONAL APPROACH

DATABASES

TAPEINFO				
NOTAPE	TAPETYPE	TITLE1	TITLE2	TITLE3
1003	SDT	NIMBUS6...	BY...	DATA...
1004	SDT	NIMBUS6...	BY...	DATA...

FILEINFO		
PB	FILE	NOTAPE
17400.0	2	1003
17401.0	3	1003
17457.0	2	1004
17459.0	3	1004

RECINFO											
DATE	TIME	LON	LAT	ALT	ZEN	PB	QUALITY	ELECTR	ILLUMIN	CALIB	SCAN
790103	124549	158.57	-1.00	1112.50	153.63	17400.0	0	ON	NIGHT	NO	OFF
790105	145629	25.90	-0.96	1112.30	153.97	17459.0	0	ON	NIGHT	NO	OFF
790117	110204	2.75	10.43	1105.00	34.11	17457.0	0	ON	NIGHT	NO	OFF
790107	105349	4.82	10.84	1103.00	35.59	17401.0	8	OFF	DAY	YES	ON

QUERIES ON DATABASES

```
select (t.tapetype, t.title1, t.title2, t.title3) as result
from tapeinfo t,
     fileinfo f,
     recinfo r
where t.notape = f.notape
  and f.pb = r.pb
  and r.zen = 153.63
  and r.calib = 'NO'
```

NATURAL LANGUAGE INTERACTION

USER	SYSTEM
1. WHERE IS THE HIGHEST SOIL EROSION POTENTIAL IN THE CHESAPEAKE BAY WATERSHED?	1. WHAT DOES EROSION MEAN?
2. EROSION MEANS	2. WHAT TEMPORAL RANGE DO YOU WISH?
3. PAST SIX YEARS	3. DO YOU WISH ME TO CONSIDER ALL RELEVANT DATA TYPES?
4. YES, BUT PRIORITIZE WITH THE MOST CURRENT	4. SYSTEM NOW TRANSLATES AND PASSES QUERY TO DATA MANAGEMENT CONTROLLER

DATA MANAGEMENT CONTROLLER

1. REVIEW QUERY AND IDENTIFIES WHAT INFORMATION IS NEEDED, IF IT IS AVAILABLE AND WHERE IT IS LOCATED

2. INTERACTS WITH APPROPRIATE DATA BASE AND SEARCH STRATEGY TO DETERMINE IF QUERY IS SATISFIED

3. IF ANSWER TO QUERY IS NOT RESIDENT AT TOP OF "ABSTRACTION PYRAMID" CONTROLLER THEN COMPARES DATA SELECTED AGAINST DATA MISSING OR INSUFFICIENT

4. CONTROLLER THEN POSTULATES ALTERNATIVES AND SUGGESTS THESE TO USER - IF CONCURRENCE, SYSTEM CONTINUES TO SATISFY QUERY. (THIS IS A REVIEW OF CENTRAL DIRECTORY, CATALOG AND INVENTORY)

5. CONTROLLER REVIEWS EXISTING DATA AND SETS UP OPERATIONAL PROCEDURES FOR DATA THAT NEEDS PROCESSING FROM "RAW" FORM TO "CLASSIFIED" FORM

   THIS INCLUDES:

   -- REGISTRATION
   -- CORRELATION (GEOLOGY WITH HYDROLOGY)
   -- PLOTTING VARIANCE AMONG DATA SETS
   -- SPATIAL MODELING
   -- IMAGE GENERATION

   RESULTS WILL BE RETURNED AND DISPLAYED AT THE USER'S WORKSTATION FOR CONCURRENCE OR FURTHER ITERATION

6. THE ABOVE OPERATIONS WILL BE A CONTINUOUS SERIAL DIALOG BETWEEN USER, "CONTROLLER" AND THE RULES CONTAINED IN EACH DATA BASE AND LEVELS OF "ABSTRACTION." CERTAIN PROCESSES MAY BE PERFORMED ON SPECIALIZED MACHINES (AP, SYMBOLIC, MPP, ETC.) AS APPROPRIATE. IT WILL ALSO BE THE JOB OF THE CONTROLLER USING "PROCEDURAL KNOWLEDGE" TO DETERMINE EACH PROCESS

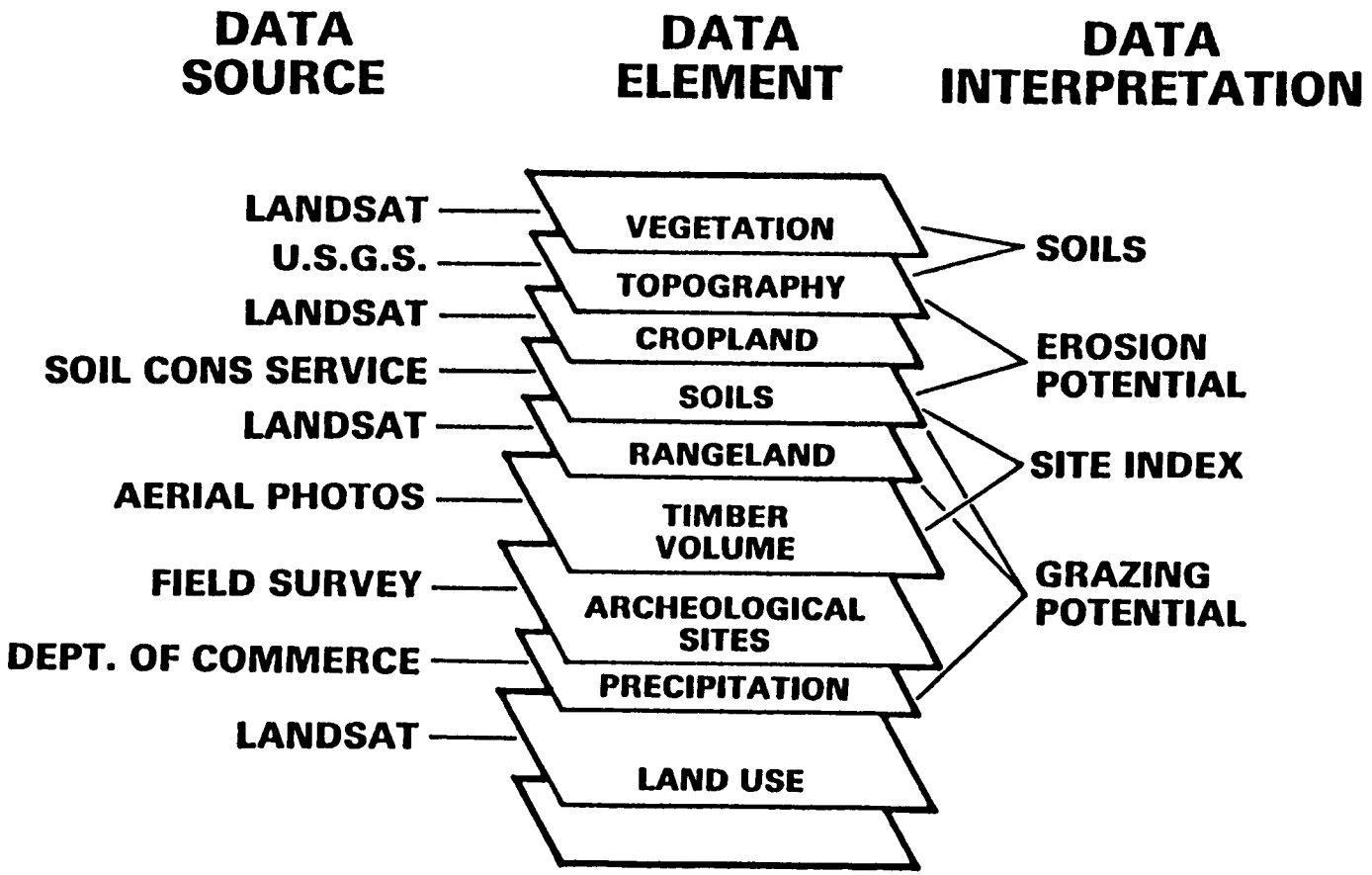

DATA SOURCE
DATA ELEMENT
DATA INTERPRETATION
LANDSAT
U.S.G.S.
LANDSAT
SOIL CONS SERVICE
LANDSAT
AERIAL PHOTOS
FIELD SURVEY
DEPT. OF COMMERCE
LANDSAT
VEGETATION
TOPOGRAPHY
CROPLAND
SOILS
RANGELAND
TIMBER VOLUME
ARCHEOLOGICAL SITES
PRECIPITATION
LAND USE
SOILS
EROSION POTENTIAL
SITE INDEX
GRAZING POTENTIAL

FIGURE 4. QUADTREE STRUCTURE (FROM HUNTER AND STEIGLITZ, 1979)

C-2

USER	INTELLIGENT INTERFACE	DBMS
1. SELECT SPATIAL DATA FOR THE CBWS FOR 1983, 84, 85	1A. NATURAL LANGUAGE PROCESSOR (NLP) TRANSLATES THE QUERY INTO SQL AND LAUNCHES QUERY TO TEST FOR VALID SYNTAX ONLY	2B. DATA MANAGEMENT SYSTEM ACCEPTS QUERY AND FAILS TO FIND CHESAPEAKE BAY LOCATION
	. DATA MANAGEMENT CONTROLLER LOOKS FOR RESPONSE AND DETERMINES THAT QUERY FAILED. . NLP RESPONDS THAT IT DOES NOT UNDERSTAND THE LOCATION (CBWS). . DATA MANAGEMENT CONTROLLER AMENDS RESPONSE TO INCLUDE APPLICATION OF QUERY AND MINIMAL DATA NECESSARY	
2. USER PROVIDES LAT/LONG INDICATES IT IS FOR SOIL EROSION AND REQUIRE 10 IMAGE MIN.	2A. DATA MANAGEMENT CONTROLLER REVIEWS RESPONSE AND BASED ON KNOWLEDGE BASE (KB) ASKS USER MINIMUM TIME BETWEEN IMAGES	
3. USER RESPONDS 3 MONTHS	3A. DATA MANAGEMENT CONTROLLER CREATES AN ENGLISH QUERY AS FOLLOWS SELECT DATA SETS FROM LANDSAT CREATED FROM THE TM SENSOR, IN BAND 2, 3, 4 FOR QUARTERS 1, 2, 3, 4 FOR YEARS 1983, 84, 85 FOR LAND AREA LAT/LONG QUERY TRANSLATED BY NLP	3B. DATA MANAGEMENT SYSTEM SEARCHES AND FINDS DATA FOR USER

RULES FOR 3A

1. IF QUERY SUPPORTS SOIL EROSION ANALYSIS THEN THE TYPE OF SENSOR THAT SHOULD BE USED IS INFRARED IMAGING

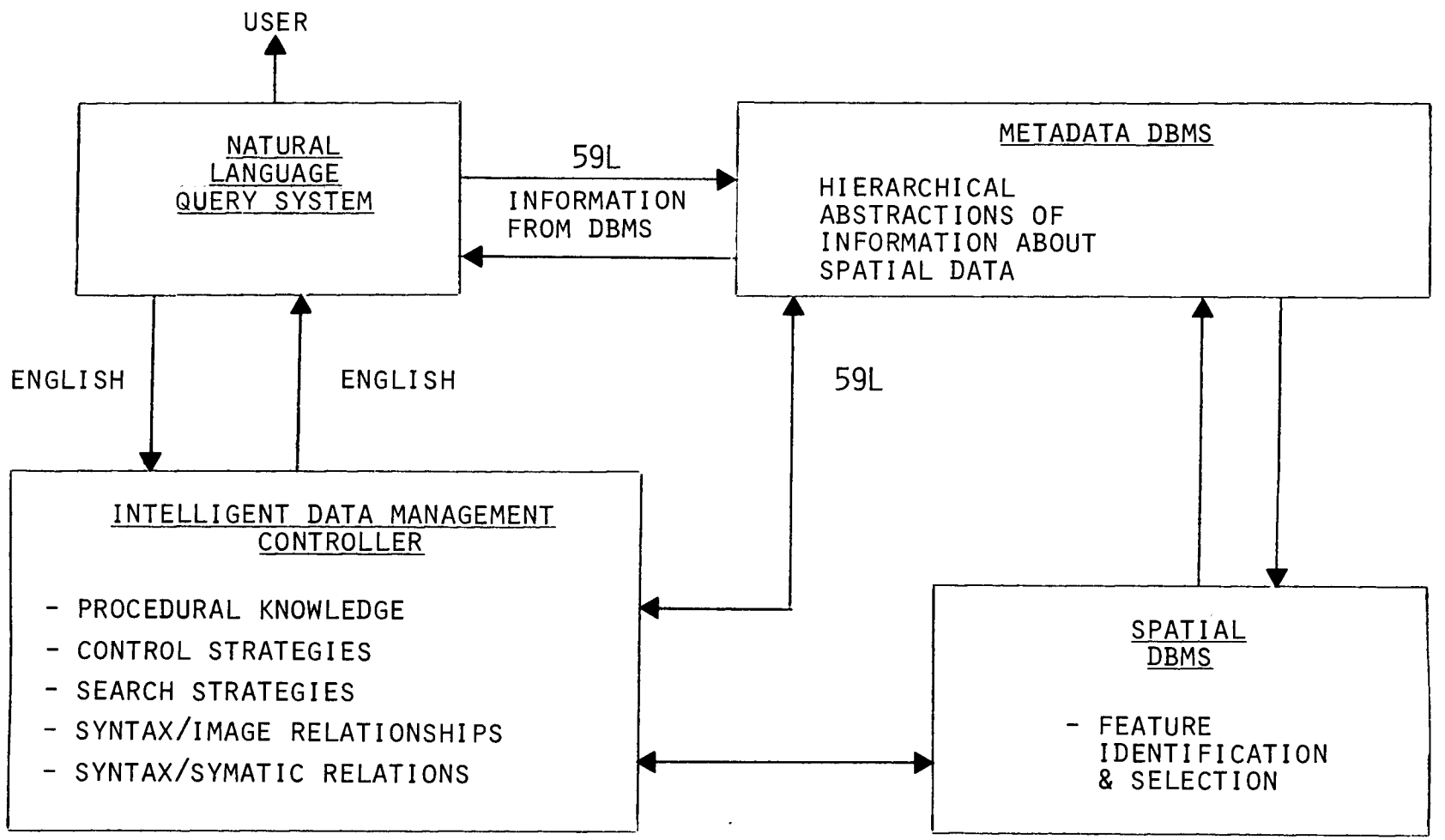
2. IF YEARS FOR QUERY IS AFTER 1983 THEN THE IMAGE DATA SHOULD BE THE THEMATIC MAPPER OF LANDSAT

3. IF THE APPLICATION IS FOR SOIL EROSION AND THE DATA DESIRED IS IN THE INFRARED THEN THE INFORMATION DESIRED IS BANDS 2, 3, 4

4. IF THE INFORMATION DESIRED IS BY YEAR THEN THE NUMBER OF IMAGES MUST BE SELECTED BASED ON MINIMUM DURATION BETWEEN IMAGES

5. IF THE APPLICATION IS FOR SOIL EROSION THEN CLOUD COVER MUST BE LESS THAN 30%

INTELLIGENT DATA MANAGEMENT CONCEPT



Software Management Environment

Frank McGarry - Goddard Space Flight Center

There exists an overall RTOP effort, supported by Code R, which is attempting to define, assess and integrate software measures and tools into an environment that will aid the management process for software development. During this briefing, examples of three specific areas of work are discussed. These areas are:

1. Research into the development of Software Design measures
2. Research into the development of Software Specification measures
3. Attempts at integrating identified measures and models into a 'Dynamic Management Information Tool'(DYNAMITE)

A. Design Measures

By closely studying and monitoring numerous software development projects at Goddard, several promising approaches to assessing design characteristics have been developed. One is a mechanism by which the classical measures of 'strength' and 'coupling' are extracted at design time and are used to predict reliability and overall quality of the software product. Early indications imply a high correlation with overall reliability and productivity of the end product.

A second design measure being investigated is one that tracks the evolution of the general architecture of a software system during the design process. Early results imply that projects with apparent disparity between the evolution of 'data structure' and 'control structure' will be more likely to have reliability problems later.

B. Specifications Measures

Attempts were carried out at developing measures by which software specifications could be used to predict complexity, reliability and other characteristics of the software product. One approach was utilized where objective measures (or counts) of such items as number of external file requirements, size of specifications, number of processes, etc. resulted in a dead end; no reliable measures were identified.

A second approach, called the Composite Specification Model (CSM), was applied to an ongoing project at Goddard and the early indicators show promise for this approach as a means for truly analyzing software complexity during the requirements phase.

C. Dynamic Management Information Tool (DYNAMITE)

A software tool has been designed which attempts to support the user in assessing software quality as well as predicting future events such as

schedules and cost. The tool has at its disposal a set of historical data on completed projects, a set of measures and models developed by this research effort, as well as an 'expert system' which contains a set of software development 'rules', also developed by this and other research projects.

This tool exists in the prototype stage and currently contains about 100 rules. It utilizes the KMS inference engine and currently has the capability of performing some very basic predictions and assessments of active projects where some very basic development information is made known to it.