Expert Systems in Chemistry - Applications to Data Quality

Stephen R. Heller and Douglas W. Bigwood
USDA, ARS, MDCL
Bldg 007, Room 56
Beltsville, MD 20705-2350 USA

INTRODUCTION

Expert Systems (ES) is a field of Artificial Intelligence (AI) which have attracted considerable attention in the field of chemistry over the past two to three decades. The DENRDAL system for structure elucidation from mass spectral data (1) was the first ES developed in chemistry. Since then, there have been many other systems developed, but few, if any, are in regular operational use. The reason for this, in the opinion of the authors, is the lack of usefulness of the ES. For example, the DENDRAL program was able to solve problems that were too simple for use other than in the classroom.

Our efforts in using ES in chemistry have focused around the area of data quality in analytical chemistry. Providing consistent and objective evaluation of published scientific data is critical for planning future analytical studies and effective use of data. In this paper we will discuss three projects, the SELEX ES, a spectroscopy knowledge base for structure elucidation ES, and lastly a data property ES.

SELEX

Our first project undertaken in this area involved using a commercial expert system shell to create the SELEX system (1). The ES created was a computer system of approximately 200 rules to evaluate and quantitatively rate published data on selenium in foods. The evaluation scheme uses five general categories for its rule-making process: number of samples, analytical method, sample handling, sampling plan, and analytical quality control. For each selenium value to be evaluated, ratings are assigned in each category by the expert system based on input which is derived from the information reported in a given paper. A Quality Index (QI), which is derived from the ratings, is a measure of the reliability of a given selenium value over all categories for a given study. The concepts used in developing SELEX have the potential of establishing criteria for assisting journal editors and their reviewers in their evaluation of many manuscripts submitted for publication.

Increasing interest in the selenium intake of Americans due to the potential relationship of selenium to cancer prevention has generated a need for the compilation, evaluation, and improvement of data on selenium in foods. Reasons for undertaking this work include the concern with the uneven quality of the data and lack of support documentation. A set of criteria were developed to evaluate the quality of existing, peer-reviewed, published selenium data (2). A manual system for post publication

evaluation of selenium data (3) using these criteria proved successful in identifying foods for which the quality of data was poor or for which there were no acceptable data. However, this manual system was more tedious, more time consuming, and less consistent than desired. Consequently an expert system, SELEX, was developed to automate the evaluation process. Developed directly from the previously established criteria, this expert system provides users with several advantages over the manual system. These include speeding the evaluation process and production of more consistent numeric ratings. Development of the expert system also allows users who have less expertise than the domain experts to generate ratings.

For each food within a study, a rating is assigned in each of five different categories. These five categories are: number of samples, analytical method, sample handling, sampling plan, and analytical quality control. The ratings assigned by SELEX, the selenium mean, and ancillary information from the publication are written into a computer file which can be read by a SAS (Statistical Analysis System) program which determines the Quality Index (QI), selenium mean, and Confidence Code (CC) for each particular food. The QI is determined from the five ratings, and with a few exceptions, is equal to the simple mean of the five numbers. The ratings and QI range from 0 to 3. A QI of 1.0 or greater indicates that the selenium mean is considered acceptable. All acceptable means for a particular food are averaged to yield a grand selenium mean for that food. The CC

(A, B, or C), derived from the sum of the QI's, represents the confidence that can be attributed to the grand selenium mean.

Using the concepts and methods created for the development of the process of evaluating published selenium data, we have considered the broader implications of these methods. It is hoped that the concepts, principles, and rules developed for the selenium data evaluation system will be considered by journal editors and their reviewers for use in their pre-publication review process. At the least, this work indicates that better defined procedures are possible for analytical chemical data evaluation. By employing such techniques it is anticipated that a better dialog could be developed between the journal editors and authors.

It is well known that the quality of much of the scientific literature is often lower than desired. There is probably far more poor and irreproducible research being published than there should be. Lide, here at the Yokohama ICIK conference and elsewhere (5), rather bluntly points out that the "scientific literature contains vast amounts of data collected for a specific purpose and presented by authors to support their conclusions... Unfortunately, the quality of the data preserved in the literature leaves much to be desired. This becomes apparent when data on a much-studied subject are systematically retrieved... The measurements for (about 200 values of the thermal conductivity of copper as a function of temperature) were analyzed by the Center for Information and Numeric Data Analysis

and Synthesis at Purdue University. The scatter of these data illustrates the pitfalls of relying on a single value retrieved from the literature." Can the scientific community find a way to improve the peer review process? Based upon this system for published data on selenium in foods, it appears this is a goal that is achievable, at least in certain cases.

DATA QUALITY CRITERIA

For each of the five areas or categories used in the evaluation process (1), a detailed description of the criteria was prepared using knowledge of accepted analytical methodology, sample handling procedures, and quality control measures for selenium, as well as a knowledge of statistical methods, including statistically based sampling methods. As stated above, the ratings ranged from 3 (highest and most desirable) to 0 (lowest and unacceptable). For example, the evaluation criteria for the analytical method category are:

Rating 3 (Highest)
The official fluorometric method (reference provided) or other method was used and is documented by a complete write-up with validation studies for the foods analyzed. This includes use of an appropriate Standard Reference Material where available, 95-105% recoveries on a food similar to the samples analyzed which were reported in the same or another paper, and the selenium concentration above the quantitation limit of the method.

Rating 2

A modified fluorometric or other method was used and is partially documented, but validation studies for the foods analyzed are incomplete. There must be as least 90-110% recoveries on a food similar to the samples analyzed which were reported in the same or another paper, or good recoveries but no statistics are given in the paper, and/or the authors have used another method (official fluorometric, isotope dilution, or neutron activation analysis) on the same sample with good agreement (which is defined as within 10%).

Rating 1

A non-fluorometric method was used and is only partly described. Recoveries were either 80-90% or > 110% on a food similar to the samples analyzed, or even better recoveries were obtained or a comparison method was used on food samples with only a somewhat related nature to the sample in question.

Rating 0 (Lowest)

The method used for selenium analysis was not documented or referenced or the reference was inaccessible. No validation studies were performed or selenium levels found in the food sample by the test method compared poorly to those found by the comparison method (>10%).

With the above definitions it is expected that trained evaluators

will derive the same ratings when they examine published reports on selenium studies.

SELEX IMPLEMENTATION

The initial SELEX implementation was written in ART (the Automated Reasoning Tool) on a VAXStation II. The main inferencing mechanism was backward-chaining (deductive reasoning), although approximately 10% of the rules were forward-chaining (inductive reasoning). The system was driven backwards from the so-called "rating rules" which generated an integer rating from 0 to 3 for each of 5 major categories. The system was rewritten as completely forward-chaining due to the fact that the automatic goal generating mechanism of ART produced unacceptable slowness in response time to users. The forward-chaining ART version was then converted to CLIPS (the C Language Interfacable Production System) (3), a forward-chaining rule-based system which uses the Rete pattern-matching algorithm also used by ART and the computer language OPS5. An example of two rules from SELEX are shown in Figure 2, which gives both the computer code as well as the English translation.

CLIPS was written by NASA's Artificial Intelligence Section, Mission Planning and Analysis Division at the Johnson Space Flight Center (4). CLIPS provided three immediate benefits. First, the CLIPS syntax is based closely on ART syntax so that SELEX could be ported quickly. Second, because CLIPS was written

in standard C, it will run on any machine which has a suitable C compiler. This is particularly important in light of the fact that ART runs on a limited number of computers. Third, the source code was provided along with a built-in mechanism for adding functions so that extending and customizing CLIPS for SELEX was easily accomplished. For example, two extensions to CLIPS provide SELEX with the capabilities of verifying user input and keeping an audit trail file which contains the sequence of questions and the user's input for each session. The final system consists of approximately 200 rules and currently is implemented on VAX VMS and IBM PC MS-DOS machines, such as the IBM AT and Toshiba 3100/5100. In fact, we use the Toshiba portable computer to provide most of the demonstrations which we give of SELEX.

As already stated, SELEX derives ratings for five major categories of evaluation: number of samples, analytical method, sample handling, sampling plan, and analytical quality control. Information is gathered by SELEX by a process of intelligent questioning of the user. The system was designed so that only pertinent questions are asked. The responses are provided in accordance with information derived from the publication containing the selenium value to be rated. Depending upon the responses, SELEX can produce a rating for each category from as few as 6 and as many as 65 answers. Approximately 90% of the questions require only a yes or no response with the remaining 10% requiring numeric input. A portion of a sample session with

SELEX is shown in Figure 2. As soon as SELEX has enough information to determine a rating for each of the five categories, the ratings are written to a file along with associated information such as a publication reference number and a description of the food. Periodically, this file is merged with a master file containing information from previously evaluated data. The master file is then analyzed with a SAS program which calculates a QI, a mean selenium value for each food, and a Confidence Code (CC) for that mean. The CC is derived from the QI's for all acceptable selenium values pertaining to a particular food.

SELEX VALIDATION

During development, SELEX was validated in two distinct ways. First, several of the 65 post-1960 selenium publications which reported original analytical selenium data for foods (from 33 different journals, reports, proceedings, and books) which have been manually evaluated by the domain experts were run through SELEX. In instances where there was a difference between the manual rating assignments and the computer expert system ratings, the differences were compared. When necessary, existing rules were clarified or changed. Also, if needed, additional rules were written to assure a correct evaluation. Second, hypothetical cases were run through the system to validate decision paths which were not encompassed by actual data from the publications. Ongoing validation will continue until the domain

experts are satisfied that SELEX performs at an acceptable level.

SELEX BENEFITS

There are several benefits over the original manual rating system. They are:

1. The manual system and the rules developed for SELEX incorporate knowledge from several domain experts who have complementary expertise. Therefore, the knowledge base is both broader and deeper than if only one expert had been used. With these rules incorporated in SELEX, publications can be rated by users who have less expertise than the domain experts.

2. During the process of formally defining the rating criteria as a rule set for SELEX, it was necessary to refine or restate some of the original criteria in more detail. Therefore, SELEX should produce more consistent results.

3. The formalization of the knowledge base facilitates its transfer to other users.

4. SELEX speeds the evaluation process and automatically maintains detailed records (audit trail) for each session.

5. SELEX reduces the "human error" factor by minimizing

transcription, data entry, and calculation errors. The determination of a rating for a category, e.g., analytical method, results from the synthesis of several pieces of information. SELEX minimizes the errors that may be caused by the omission of information.

6. Since new publications with selenium data are evaluated intermittently, SELEX eliminates the need for the users to continually refamiliarize themselves with the complex set of heuristics.

The overall benefit, of course, is that SELEX will improve the definition and evaluation of the quality of the information available to identify any selenium-cancer correlation, since the results will be more accurate using an automated (objective method) rather than a manual one.

SPECTROSCOPY KNOWLEDGE BASE

Considerable research has been undertaken in the area of expert systems in spectroscopy (6-8). The goal of these systems has been the elucidation of the structure of an unknown molecule from spectral data. The fact that these systems have not produced sufficient positive results to justify their everyday use is, in our opinion, due the enormous difficulty of the problem. Complete structure elucidation is an admirable goal, but owing to the current lack of sufficient knowledge for input and use by

such interpretation systems, we believe it is an unobtainable goal.

With the number of chemicals reported in the literature exceeding 8 million, and with only some 10,000 - 150,000 available spectral fingerprints, spectral library identification poses certain intrinsic challenges. Computer based structure elucidation methods have made impressive improvements in the last few years. However, the fact remains that without a major breakthrough, further enhancements are likely to be difficult. The potential for developing a knowledge base of spectral correlations to aid as a tool in furthering structure elucidation methods is clearly great. As Enke (9) has recently pointed out "one can expect that traditional structure elucidation tools (including human experts) will fail to extract all the valuable analytical information within a reasonable time interval". Such comments as these have led us to initiate a project which we call the ARS Spectroscopist, or ARS SPEC for short. The goal of this project is to develop a comprehensive knowledge base of spectral-structure correlation rules. We expect the knowledge base will cover all fields of spectroscopy. To start with we are using CNMR, MS, HNMR, and plan to use IR. The overall view of the ARS Spectroscopist is shown in Figure 3. ARS SPEC will accept spectral and other data and output a list of substructures which are likely to be present or absent. From this list one could then go on and use programs such as CONGEN (6), or the structure generation portion of CHEMICS (7) or CASE (8), in order to get a

possible complete structure.

We are proposing here a new strategy for chemical structure elucidation. For the purpose of this discussion structure elucidation problems can be divided into two categories, real problems and contrived problems. Real problems are those which are encountered everyday in analytical chemistry labs throughout the world. Contrived problems are those which are usually found in text books or are restricted to an arbitrary class of compounds (e.g., straight chain amines), the solution of which makes a good lecture, but is never used by a practicing chemist.

Usually when one goes to a spectroscopy expert for help the result is a collection of suggestions, ranging from comments on specific functional groups (or chemical substructures) being present or absent, to suggestions as to what additional data should be obtained which would be useful in solving the problem. Rarely does one get a quick and complete answer from the expert. Based on this situation, the strategy to create an expert system to do the same has developed. Thus, it is being proposed that the goal of this work is to provide the user with a list of suggestions, based on the existing knowledge base, as to what to do next. The goal of the system is then not to completely solve the problem, but rather to offer expert help and advice.

In Figure 3 it is seen that given a piece of spectral data will

be associated with a chemical substructure and vice versa. One will be able to go into the system using either the structure or data. As can be seen from Figure 4, the system is being designed to both predict structure features, as well as indicate what additional spectral data should be obtained to assure the accuracy of a prediction. In this we hope to be able to improve the likelihood that such a system will actually be accepted and used by the spectroscopy community as an aid in structure elucidation. With the scarcity and high cost of trained experts in the field of structure elucidation, the ARS SPEC has the potential of being a useful application of ES in chemistry.

ARS DATA EVALUATOR

Contamination of the groundwater in the USA is a serious concern and with the extensive use of pesticides and other chemical by the agriculture community, it is desirable to be able to predict the potential for chemical contamination. Developing models for these studies require the best possible data in order to assure the best predictions of groundwater contamination. The weak link in any modeling activity has generally been found to be the quality of the data used as input into the model. Thus, when our organization initiated a new model in this area it became necessary to provide a database of physical and chemical properties of pesticides used in the USA. An examination of the literature and discussions with modelers and pesticide chemists

quickly lead to the conclusion that neither a database, nor a organized collection of evaluated database of pesticide properties was available publicly. As part of our efforts in developing a pesticide property database (PPD), it was clear that there was a need for an objective evaluation system to examine the data found in the literature, as well as data from labs throughout the country which was unpublished. For example, we found that the solubility of a widely used herbicide, Alachlor, had a reported aqueous solubility value of 140 mg/kg at 23 degrees Celsius in one well known handbook from the United Kingdom and 242 mg/kg at 25 degrees Celsius in a second widely used handbook published in the USA. Thus, the ARS Data Evaluator concept was developed and work initiated to develop an ES for data property evaluations. Figure 5 shows the overall outline of the Data Evaluator. As can be seem from this figure, one problem is that many physical properties are estimated or calculated values, not experimental data. Thus the system had to be structured so that any value, whether experimental or theoretical, could be handled within the one system.

REFERENCES

1. D. Bigwood, S. R. Heller, W. R. Wolf, A. Schubert, and J. M. Holden, "SELEX: An Expert System for Evaluating Published Data on Selenium in Foods", Anal. Chim. Acta, 200, in press (1987).

2. J.M. Holden, A. Schubert, W.R. Wolf, and G.R. Beecher, "A System for Evaluating the Quality of Published Nutrient Data: Selenium, a Test Case", Food and Nutrition Bulletin, 9 (Suppl. - Food Composition Data: The User's Perspective), (1987).

3. A. Schubert, J. Holden, W. R. Wolf, J. Am. Diet. Assoc., "Selenium Content of a Core Group of Foods based on a Critical Evaluation of Published Analytical Data", 87 (1987) 285.

4. Gary Riley or Chris Culbert, NASA/Johnson Space Center, Mission Planning & Analysis Division, Artificial Intelligence Section - FM72, Houston, TX 77058. The software is available from COSMIC Software Catalog, 1987 Edition, page 270, # M87-11021, and costs $217.00 (including documentation). Address requests to NASA's Computer Software Management and Information Center, The University of Georgia, Computer Services Annex, Athens, GA 30602 USA (Telephone - 404-542-3265).

5. D. R. Lide, Jr., "Critical Data for Critical Needs", Science, 212 (1981) 1343.

6. R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, and J. Lederberg, Applications of Artificial Intelligence for Organic Chemistry - The DENDRAL Project; McGraw-Hill, New York (1980).

7. S. Sasaki and H. Abe, in Computer Applications in Chemistry, S. R. Heller and R. Potenzone (eds.), pp. 185-206, Elsevier, New York (1983).

8. M. E. Munk, C. A. Shelley, H. B. Woodruff, and M. O. Trulson, Z. Anal. Chem., 313, 473(1982).

9. C. G. Enke, A. P. Wade, P. T. Palmer, and K. J. Hart, "Solving the MS/MS Puzzle: Strategies for Automated Structure Elucidation", Anal. Chem., 59, 1263A (1987).

Figure 1 - Sample SELEX Rules

Two rules are used to determine a rating for sample handling. The first rule asserts a rating from information that has been obtained from the user. The second rule is an example of a rule which queries the user for information. Each rule is followed by an English translation.

```
(defrule Rating-sample-handling-10
         (declare (salience 100))
         (seeking-rating sample-handling)
         (homogenization-validation-data optimal)
         (moisture-level-documented false)
         =>
         (assert (rating sample-handling 2)))
```

Translation of rule Rating-sample-handling-10:

If you are seeking a rating for sample handling and the homogenization validation data is optimal and the moisture level was not documented, then the rating for sample handling is 2. NOTE: This rule has a declared salience of 100. The system will "fire" this rule ahead of rules with lower salience. In this case we want rating rules to fire ahead of information gathering rules such as the one below (rules with no declared salience are assigned a default salience of 0) because once SELEX can determine a rating, no further information is needed. This exemplifies one key element of expert systems - intelligent questioning.

```
(defrule Food-preparation-documented
         (seeking-rating sample-handling)
         (or (perishable-food false)
             (shipping-and-storage-appropriate true)
             (shipping-and-storage-documented false))
         (not (food-preparation-documented ?))
         =>
         (if (y-or-n-p 3060 0 "Was the food preparation
                                        documented")
             then (assert (food-preparation-documented true))
             else (assert (food-preparation-documented false))
                  (assert (food-preparation-appropriate true))))
```

English translation for rule Food-preparation-documented:

If you are seeking a rating for sample handling and either the food is not perishable or the shipping and storage procedures were appropriate or the shipping and storage procedures were not documented and it is not known whether or not the food preparation was documented, then ask the yes-or-no question "Was the food preparation documented?". If the answer is yes then assert that the food preparation was documented or else assert that the food preparation was not documented and assume that the

food preparation was appropriate.

Figure 2. Part of a typical session with SELEX. This portion represents the rating process for sample handling for a hypothetical example. (The answers the user provides are underlined.)

```
============================================================
Now seeking a rating for sample-handling for selenium.
============================================================

Was the sample handling procedure documented?
Response (Y or N): Y
Was the sample food perishable?
Response (Y or N): Y
Were the shipping and storage procedures documented?
Response (Y or N): N
Was the food preparation documented?
Response (Y or N): Y
Was the method of food preparation appropriate?
Response (Y or N): Y
Was only the edible portion of the food analyzed?
Response (Y or N): Y
Was homogenization of the sample required?
Response (Y or N): N
Was the sample moisture level documented?
Response (Y or N): Y
Was the moisture level of the sample appropriate?
Response (Y or N): Y

The rating for sample-handling is 2.
```

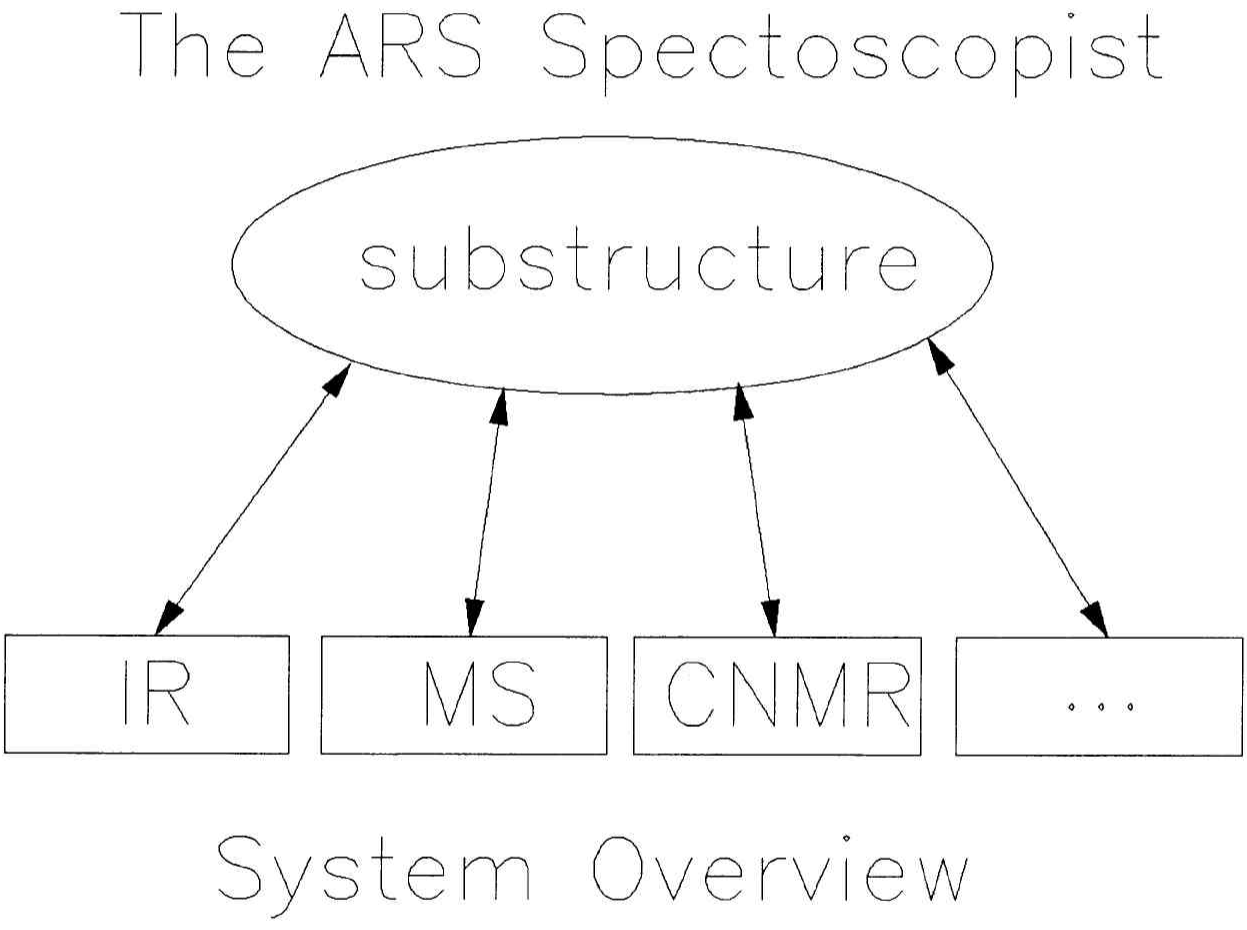
The ARS Spectoscopist
substructure
IR
MS
CNMR
...
System Overview

Figure 3

Figure 4. Sample session for the ARS Spectroscopist

Please enter the data you have as it is requested.

CNMR - Enter a chemical shift and multiplicity
(or None if not available)
User Response: 55,2 S

MS - Please enter peaks and intensities as pairs separated
by commas (or None if not available)
User Response: 31,10 45,5

IR - Please enter absorption range (cm-1) and intensity,
separated by a comma (or None if not available)
User Response: 1300,1000,40 2850,2800,10

From the data provided it is suggested that your sample contains:
A methoxy group

The Probability of this is: 85%

It would be helpful if you could obtain a HNMR spectrum of this sample to see if there is a peak in the spectrum which corresponds to that of the hydrogen atoms of the methoxy group.

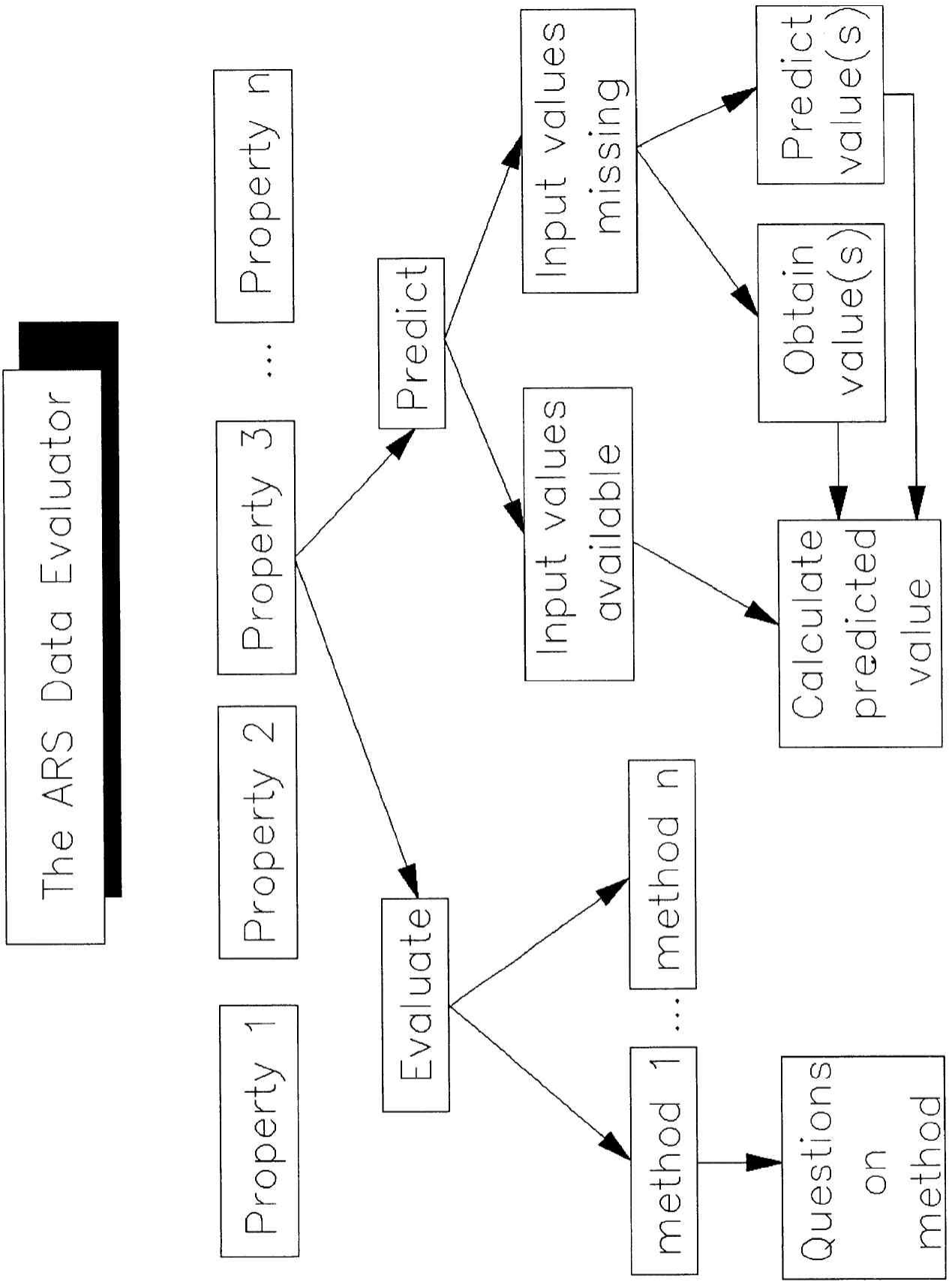
The ARS Data Evaluator
Property 1
Property 2
Property 3
... Property n
Evaluate
Predict
method 1
... method n
Questions on method
Input values available
Input values missing
Obtain value(s)
Predict value(s)
Calculate predicted value

Figure 5