ChemReg: Using Case-Based Reasoning to Support Health and Safety Compliance in the Chemical Industry

Kirk D. Wilson

Air Products and Chemicals, Inc.
7201 Hamilton Blvd.
Allentown, PA 18195-1501
wilsonkd@apci.com

Abstract

ChemReg is a large knowledge-based system used by Air Products and Chemicals, Inc. to support compliance with regulatory requirements for communicating health and safety information in the shipping and handling of chemical products. This presentation concentrates on one of the knowledge bases in this system: the case-based reasoner. The case-based reasoner addresses the issue, How can proper communication of public health and safety information be insured while allowing for rapid and cost-effective product evaluation in the absence of actual hazard testing of the product? Estimates of hazard data from similar products are generated for new products using an existing relational database as a case library. Implementation of the cased-based reasoner in rules and objects using a commercial KBS shell is described. While some refinements remain, the performance of the case-based reasoner has met expectations.*

ChemReg is a knowledge-based system begun by Air Products and Chemicals, Inc. of Allentown, Pennsylvania in 1993 to support compliance with regulatory requirements for communicating health and safety information in the shipping and handling of chemical products. Today the ChemReg system consists of three knowledge bases, a large relational database of product information, and an on-line system that is used throughout the United States and Europe. The purpose of the system is to automatically generate safety information for transporting and handling chemicals.

Background: Purpose and Justification of ChemReg

The chemical industry is a highly regulated industry. Every hazardous chemical product must have a set of shipping descriptions that conform to large bodies of strict regulations issued by the Department of Transportation (DOT), the International Maritime Organization (IMO), and the International Air Transport Association (IATA) as well as a material safety data sheet (MSDS) that conforms to Occupational Safety and Health (OSHA) and ANSI standards and which provides a detailed description of the health hazards of the product along with proper handling and emergency procedures. Penalties ranging from delayed shipments to heavy fines or even incarceration of corporate officials can result from failing to comply with transport regulations. Moreover, inadequately prepared MSDSs can lead to substantial product liability suits against the company if the product is involved in an industrial accident.

The ultimate goal of these regulations is to insure proper communication of health and safety information for the protection and safety of the public. Air Products is committed to the initiative of the Chemical Manufacturers Association (CMA) known as Responsible Care.® This initiative focuses on the safe manufacturing, distribution, use, recycling, and disposal of chemicals. Proper communication through accurate shipping descriptions and full disclosure of hazard information in the MSDS plays a central role in fulfilling the obligations under Responsible Care.

Maintaining shipping descriptions and MSDSs constitutes a significant effort in the chemical industry. Most corporate systems are intensely manual. When shipping regulations change, which occurs periodically, a hectic period ensues as transportation groups scramble to reevaluate thousands of shipping descriptions. The principal problem with MSDSs is keeping their content up to date and consistent, since they are usually maintained as separate documents in word-processor files that have been written over a long period of time.

These problems were well known to Air Products. Air Products and Chemicals is a four billion dollar, international gas and chemicals company. About one-third of the revenues of the company are generated from the sale of chemicals. The product lines consist of a substantial number of multi-component, formulated

products, including polymer emulsions, catalysts, specialty performance additives, and amine products. In 1993 the chemical industry confronted a major change in DOT shipping regulations. For Air Products, manually reevaluating the affected products for new shipping descriptions was deemed to be impractical. At the time Air Products had already investigated automating the process of generating shipping descriptions. The goals of an automated system were:

- To significantly reduce the time for new products to go through the regulatory compliance process from the 10 to 30 days it was taking. This goal would have to be reached while retaining 100% compliance with the regulations.

- To provide a consistent level of service without increasing staff in spite of rapid new product introductions and acquisitions of new product lines. In addition, the system had to allow efficient response to changes in the regulations when significant numbers of products need to be reevaluated.

The decision was made to develop a knowledge-based system to fulfill these goals and to meet the new regulatory requirements. Programming on the transportation knowledge base, using AION,[1] began in early 1994, just nine months before the regulatory changes were to take effect.

The goals of ChemReg dovetailed with another long standing goal of Air Products: to have a central database to store product information. Thus, the basic architecture of ChemReg was born (Figure 1): an on-line system, written in PowerBuilder,[2] was distributed to business managers and laboratory technologists in order to add technical data to a central database, which the knowledge base reads in order to generate shipping descriptions. The knowledge base writes shipping descriptions back to this database where they are available to the order-entry system, which produces bills-of-lading, and to the container labeling system. The transportation knowledge base was completed in 1994 in sufficient time to meet the new DOT requirements. In addition, the knowledge base generated shipping descriptions under IMO (sea) and IATA (air) regulations.

The transportation knowledge base has accomplished its operational goals: (1) it has reduced the time required to process regulatory information for new products to two days (the minimal time to accommodate plant labeling operations) and (2) it has allowed the transportation professionals to meet business demands for rapidly introducing new products and to respond quickly to regulatory changes without increasing staff.

---

[1] AION® is a registered trademark of PLATINUM *technology, inc.*.
[2] PowerBuilder® is a registered trademark of Sybase, Inc..

A further knowledge base was added to the system in 1995 for generating (*authoring*) the contents of the MSDS. Unlike shipping descriptions, whose content is a concise identification of the hazards that a chemical presents during transportation, MSDSs are lengthy documents covering a myriad of topics related to health hazards that a chemical could present in the workplace and contain detailed guidance for handling the chemical. The ChemReg MSDS knowledge base represents the content of the MSDS as a semantic network of the various topics covered in the MSDS. Rules for each node of the network generate what is said about a particular topic for a product.

Both knowledge bases share a common design philosophy of mapping classificatory rules directly into knowledge-base code. These rules, which determine the hazard classification of the product, are typically stated with respect to results of tests performed on *products*. For example,

- A product may be classified on the basis of a physical property such as its flash point, e.g., that it is a Flammable material.

- Regulatory rules are also based on the product's harmful effects on living organisms. Such properties include oral, dermal, and inhalation acute toxicity as well as corrosivity and irritation effects.

However, it is time-consuming and expensive to perform a complete set of tests on every product. It is estimated that to conduct tests on all its products would cost Air Products $1MM to $4MM (million). But not testing for all hazards could allow possibly hazardous classifications to be missed, thereby contravening the goal of 100% compliance. Thus, both knowledge bases incur the same issue: *how can proper communication of public health and safety information be insured while allowing for rapid and cost-effective product evaluation?*

All chemical companies face this issue. The issue was initially addressed in ChemReg by requiring estimates to be entered when actual test data was unavailable. However, this approach proved to be unsatisfactory. The challenge became one of finding a means to alleviate the manual entering of estimates and yet of insuring that there would be reliable and auditable data available to the transportation and MSDS knowledge bases to apply the classificatory rules in full compliance with the regulations.

The Solution: Automated Case-Based Reasoning

If products are allowed to assume the role of cases, this challenge becomes a classic case-based reasoning problem. Case-based reasoning (CBR) is a problem solving strategy based on finding a known, similar case on which to base the solution of a new problem. Figure 2 presents the flow of case-based reasoning tasks used in ChemReg. The following tasks are executed:
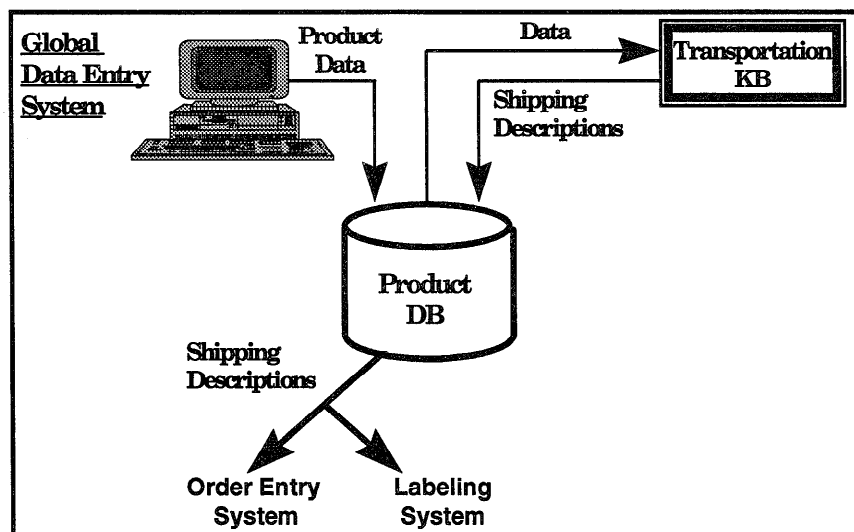
Figure 1. *Initial Configuration of the ChemReg System. The knowledge base runs twice daily as a "batch" program.*

1. An **Analyze Case** task determines whether data for a hazard is relevant for classifying the current product. Rules determining relevancy identify the indexes for retrieving products from a relational database. Indexes consist of the type of actual data to be sought and, if appropriate for the hazard, the principally hazardous component in the current product that causes the hazard to be relevant.

2. **Retrieve Cases** creates the appropriate SQL statements for retrieving candidate similar products and their components. Indexes identified in step 1 are translated into SQL WHERE clauses. Retrieved cases are parsed into appropriate knowledge-base classes.

3. **Determine Match** applies similarity metrics based upon the way Air Products safety experts would choose the most closely matching similar product. ChemReg must measure *structural* and *performance* similarity, rather than matching on specific product properties. Products are primarily defined by their structure, i.e., by their components. While the compositional level of a particular component may be considered a specific feature, each component may or may not be matched in another product. Other metrics measure performance similarity by assessing the aggregate effect and interaction of the components in a mixture.

4. **Adapt Case** adapts the hazard data found on the selected similar product to the current product when there still exists a significant difference between the current product and the selected similar product.

In ChemReg, these steps must be performed *without human intervention*.

The basic process behind case-based reasoning was not new to the product safety experts on the project. Indeed, when MSDSs were manually written, a "similar" product's MSDS was often selected as the model on which to base a new MSDS. However, the manual selection of a similar MSDS could not be exercised rigorously in a reasonably acceptable amount of time, and it became clear very early during testing that the case-based reasoner was able, in a significantly reduced time, to do a more thorough and accurate job of selecting similar products than could the experts themselves.
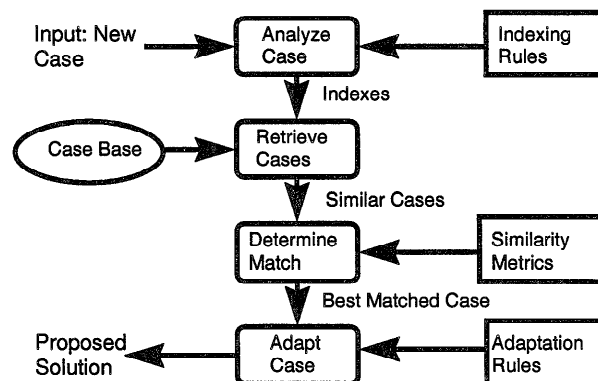

Figure 2. *Problem solving strategy of case-based reasoning in ChemReg. Adapted and abridged from Riesbeck and Shank, p. 32.*

The case-based reasoner provides several operational benefits: it makes estimates consistently and it tracks what product and reasoning were used as the basis of the position for addressing classificatory issues in the absence of actual test data. Indeed, the Environmental Protection Agency (EPA) endorses the use of similar products in this
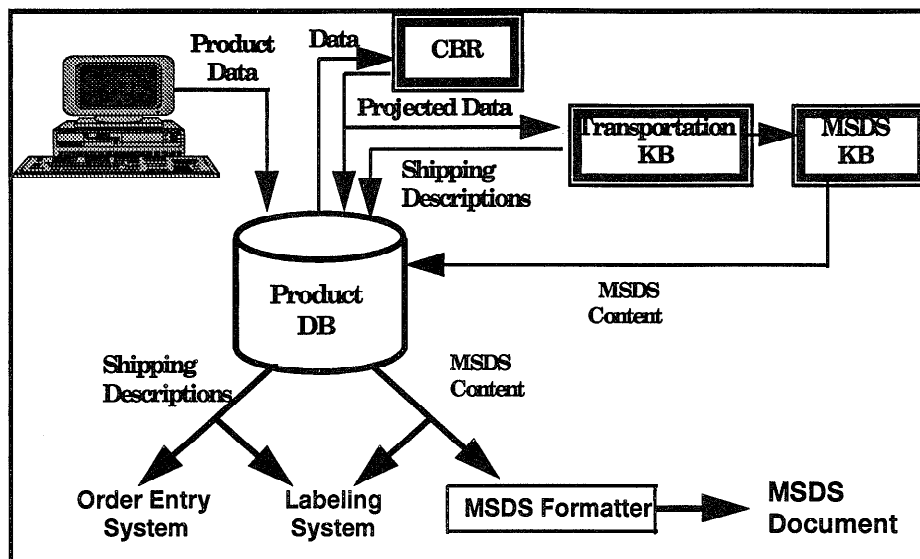
Figure 3. *Configuration of ChemReg with the MSDS knowledge base and the case-based reasoner (CBR).*

situation in preference to the rival method of projecting the toxicity of a mixture from the toxicity of that mixture's components. The EPA has stated,

> The basic assumption in the recommended approach is that risk assessments on chemical mixtures are best conducted using toxicologic data on the mixture of concern *or a reasonably similar mixture*.[3]

However, the EPA provides no guidelines for determining reasonable similarity, preferring to leave that issue to "case by case" judgment (p. 34023).

Prior knowledge-based system approaches to authoring MSDSs have employed the technique of projecting hazard classifications from components.[4] Chemreg *automates* the similar product approach for the first time in the chemical industry.

The current configuration of the ChemReg system is shown in Figure 3. The on-line system is used by hundreds of Air Products employees throughout the United States and in several facilities in Europe. The knowledge-based system, consisting of the three knowledge bases, runs twice daily as a Windows application and processes those products on which maintenance has been performed since the last cycle. In particular, the case-based reasoner generates the data, called "projected data", that is used by the transportation and MSDS knowledge bases to make classifications.

If no similar product can be found for a product, the product is "kicked out" to the product safety experts with a recommendation that testing be performed on it. It is

also possible for the experts to override the results of the case-based reasoner by specifying an estimate for a test value on the product as input to the case-based reasoning process. However, the need to do this counts as a failure for the case-based reasoner.

In the following section, we describe the architecture for implementing the case-based reasoner in AION using traditional KBS technology (rules and objects).

Case-Based Reasoning Architecture

Implementation of the ChemReg case-based reasoner involves three layers of inferencing: (1) the inference engine itself which is provided by the development software (AION) and which provides backward and pattern-matching forward chaining, (2) a task-control layer representing the problem-solving strategy of CBR (Figure 2) and which sequences the execution of tasks, and (3) the task-specific rules. Task-specific rules determine the relevancy of hazard data for classifying a product (relevancy rules), measure the similarity of products (similarity metrics), and adapt the selected similar-product data to the current product (adaptation rules).

The following points summarize the physical design of the knowledge base:

- The task control layer consists of a functional decomposition of the steps necessary to complete processing. Execution is controlled via procedural code, which invokes control rules that send messages to knowledge base objects to initiate the problem-solving tasks.

- Hazards are represented as classes in a hierarchy (Figure 4).

---

[3] Environmental Protection Agency, p. 34020 (emphasis added).
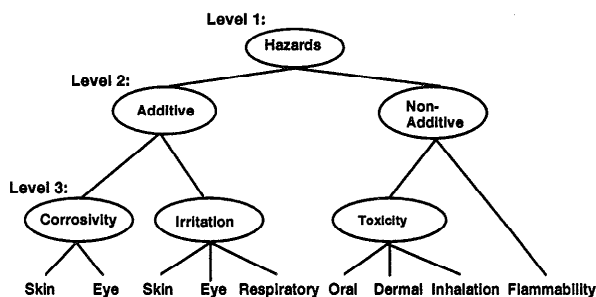[4] Most notably, The Lubrizol Corporation, see Sorani and Lauer.

Figure 4. *Class hierarchy of hazards. Leaf nodes are classes that contain slots for holding test results from a similar product.*

- Backward chaining rules identify the indexes for retrieving products for consideration by the case based reasoner.
- Cases, which consist of products and their components, are represented as instances of classes. Candidate similar products have a slot for accumulating their similarity scores.
- All similarity metric rules are expressed by pattern-matching, forward chaining rules (Figure 5). These rules join the knowledge-base classes constituting a case. Thus, a single rule can process a particular metric involving all candidate similar products.

```
IFMATCH Current_Comp, Similar_Comp,
        Similar_Products
 with Current_Comp.comp_id = Similar_Comp.comp_id
 and Similar_Comp.prodcode =
                       Similar_Products.prodcode
THEN
    send (CalcCompDifference to Similar_Products
        with Current_Comp.comp_percentage,
          Similar_Comp.comp_percentage)
END
```

Figure 5. *General similarity metric. CalcCompDifference calculates a similarity score based upon the compositional level of a particular component in the current product vs. its level in a similar product.*

- Similarity scores are calculated by a simplified form of nearest neighbor weighting:

$$score = weight \text{ X } sim(f_\text{I}, f_\text{R})$$

where *weight* is the importance factor of feature f for the hazard under consideration and *sim* is the objective measurement of the difference between that feature in the input case and the retrieved case respectively. Scores for individual features are summed.[5]

- The metric for measuring the similarity of products when considering some types of hazards differs from the metric for measuring similarity with respect to other hazards. Therefore, the ChemReg case-based reasoner needs to accommodate *context sensitive matching*.[6] Context sensitive matching employs metric rules that reference Level 2 of the hazard class hierarchy (Figure 4). Dynamically created knowledge-base objects track similarity scores of the same product within different contexts.
- Further selection heuristics are applied to refine the selection of the highest scoring similar products. Selection heuristics are constructed on the same pattern as similarity metric rules (Figure 5), but (1) may require matching on levels in the hazard class hierarchy below Level 2 (see Figure 4) and (2) may reject a candidate similar product out-rightly.
- Adaptation rules are implemented in functions that apply the relevant data from the selected similar product to the current product. The CBR technique used is *parameter adjustment*.[7]

Figure 6 shows an example of the reasoning that is followed by the case-based reasoner when considering a case of oral toxicity.

Development, Performance, and Maintenance

Four types of task-specific rules had to be elicited from the experts:

Relevancy rules. These rules identify when a hazard is relevant for classifying a product according to the regulations. Each type of hazard has five to ten relevancy rules.

Similarity metric rules. Similarity metrics measure the difference between structural and performance features of the new product and a similar product. These rules, which calculate a numeric score for the similarity between structural and performance features of two products, are either *general* and apply the same metric to all cases (Figure 5) or *specific*, in which case they apply different metrics depending upon the hazard under consideration (context sensitive matching). ChemReg currently measures similarity on four axes of compositional and performance similarity. In general, metrics for additive hazards must make an assessment of a product's performance characteristics, while metrics for non-additive hazards must be more sensitive to compositional structure.

Selection Heuristics. These rules insure that the highest scoring similar product meets conditions of acceptability. They express the experts' judgment for selecting similar products in anomalous situations. The need for this layer

---

[5] Compare Kolodner, 355.

[6] Kolodner, pp. 328-329.
[7] Kolodner, pp. 404-407.

The Knowledge Base executed the following activities when processing Product NEWPROD Air Products' New Product.

Oral Tox	Oral toxicity data is relevant for product NEWPROD because there is a component that is classified as an oral toxin. The most severely hazardous component for oral toxicity is Pentamethyldiethylenetriamine at 95.00%.
Oral Tox	For oral toxicity the following products were evaluated: S_PROD01, S_PROD02, S_PROD_03, S_PROD04.
Gen Comp	Product S_PROD01 is being evaluated for oral toxicity from a base score of 216.12.
Gen Comp	Product S_PROD02 is being evaluated for oral toxicity from a base score of 195.00.
Gen Comp	Product S_PROD03 is being evaluated for oral toxicity from a base score of 163.50.
Gen Comp	Product S_PROD04 is being evaluated for oral toxicity from a base score of 120.80.
Max Comp	Product S_PROD02 was incremented to 585.00 for oral toxicity because it contains Pentamethyldiethylenetriamine at 100.00% vs current prod concentration of 95.00.
Max Comp	Product S_PROD03 was incremented to 490.50 for oral toxicity because it contains Pentamethyldiethylenetriamine at 58.50% vs current prod concentration of 95.00.
Rej NonAdd	Product S_PROD03 was rejected for oral toxicity because it contains N,N-DIMETHYLCYCLOHEXYLAMINE at 41.50%, which is a more severely hazardous component.
Rej NonAdd	Product S_PROD04 was rejected for oral toxicity because it contains N,N-DIMETHYLCYCLOHEXYLAMINE at 84.20%, which is a more severely hazardous component.
Comp Rej	Product S_PROD01 was rejected because it contains Pentamethyldiethylenetriamine at 11.25% which is below the minimum concentration for this component of 48.33%.
Oral Tox	Product S_PROD02 was selected as the similar product for oral toxicity for product NEWPROD.

Figure 6. *Cased-Based Reasoning Report.* 'Gen Comp' *messages result from the general similarity metric* (Figure 5); 'Max Comp' *messages represent the application of the specific similarity metric rule for Toxicity, and* 'Rej NonAdd' *and* 'Comp Rej' *messages result from selection heuristics. (Adaptation was not necessary in this example.)*

arose during system testing when the experts would reject the results of the objective metric calculations based on their interpretation of or knowledge about the data. The experts could articulate these heuristics only when confronted with an anomalous situation. At present, seven selection heuristics have been identified.

Adaptation rules. Adaptation rules define a *reasonable* way of estimating scores based on similar product data when there is still a significant difference between the current product and the selected similar product. Adaptation rules, therefore, have to (1) recognize when a "significant" gap occurs in the measurements of a structural or performance feature, and (2) define a procedure by which the severity of the hazard data from the selected similar product is to be increased or diminished. Both these points are matters of professional

judgment; thus the adaptation rules in ChemReg are heuristic. Adaptation rules proved to be the most difficult for the experts to articulate. Only one adaptation rule for toxicity mapped to previous manual procedures for handling regulatory compliance for new products. Each hazard type amenable to adaptation (Corrosivity, Irritation, and Toxicity) has an average of four (two pairs) of adaptation rules.

Excluding development of the relevancy rules, which were implemented independently of the similar-product logic in an initial version of the knowledge base, eliciting the similarity metrics, selection heuristics, and adaptation rules took approximately 40 hours of sessions with the experts over a period of five months. This period covered a substantial period of reviewing tests and refining the selection heuristics and adaptation rules.

A technical challenge arose with the use of SQL to retrieve cases (products) from a relational database due to the normalized structure of the database. Normalization leads to a complex case structure.[8] ChemReg does not require similarity-based retrieval mechanisms (e.g., fuzzy matching or nearest neighbor retrieval), since all similarity measurement takes place within the case-based reasoner itself. However, in this situation there is a performance tradeoff that must be resolved through the design of the indexes in order to keep the number of case records coming into the program across the network to a minimum while not excluding potentially relevant cases. In ChemReg, this tradeoff is resolved by retrieving only those products with actual data *within the same chemical family* as the current product: a specific chemical family is used for those product lines in which Air Products has a significant number of products; a general family is used to expand the search in more specialized product lines.

Nevertheless, ChemReg requires three passes over the database (1) to set-up (in the database) the list of products to be retrieved based upon applying the relevant indexes (a step involving a three table join), (2) to retrieve products and their components (a step involving a six table join), and (3) to retrieve pure products of the principally hazardous component regardless of chemical. It takes about 15 seconds to process a product through the case-based reasoner. On average, eight similar products are retrieved per product; however, this may involve 60 or more database records being retrieved depending upon the number of components comprising those products.

The case-based reasoner was fully implemented in June 1996. Performance has been tracked for acute toxicity (dermal, oral, or inhalation toxicity), skin corrosivity , and skin and eye irritation. Figure 7 summarizes the

performance criteria and actual results obtained during testing and the first four months of production. Air Products also tracked times when the case-based reasoner proposed estimates that, upon review, were preferred to the experts' own original estimates (column 4). The chart shows that in cases involving toxicity the case-based reasoner has exceeded expectations for its performance. Further refinement needs to be done in the area of corrosivity and eye irritation.

Type of Estimates	Performance Goal	Actual results	ChemReg judged better
Estimates for acute toxicity	90%	96%	62%
Estimates for skin corrosivity	95%	93%	43%
Estimates for skin irritation	95%	97%	38%
Estimates for eye irritation	95%	82%	36%

Figure 7. *Performance Results of the Case-Based Reasoner.*

Maintenance requirements differ for the three knowledge bases. Maintenance on the transportation knowledge base must be performed periodically as the transportation regulatory agencies promulgate revised regulations, which occurs about every two years. The MSDS knowledge base, on the other hand, was designed as a formal system for authoring the MSDS. Consequently, the product safety experts can play a more direct role in controlling the output of the knowledge base to meet their particular needs for changing the contents of an MSDS. Currently, there is a need to monitor the performance of the case-based reasoner and consider refinements to its reasoning, particularly its selection heuristics and adaptation rules.

Benefits and Future Enhancements

Regulatory compliance is a public and legal obligation; it is a cost of doing business. The business strategy must be to find the most cost-effective means for insuring full regulatory compliance and communication of the hazards of the product. ChemReg *guarantees* that shipping descriptions and MSDSs will comply with corporate policy and interpretation of the regulations. Moreover, periodic regulatory changes can be quickly and *consistently* implemented across the entire product line.

---

[8] Products and their components (the primary elements of a case) exist in a *m:n* relationship; compare Kitano and Shimazu, "…cases have to be represented as a flat record of *n*-ary relations" (p. 254).

The result is an assurance of superior hazard communication to the public and a reduction in the potential for significant loss due to non-compliance for the company.

Significant costs and effort would have been incurred under the manual system to bring shipping descriptions and MSDSs into compliance with the current regulatory environment. Developing the transportation knowledge base avoided an estimated $250,000 expenditure involving a substantial effort on the part of the product business managers to review all their products (a highly time-consuming activity to impose on product management) in order to comply with the original DOT regulatory revisions. Developing the MSDS knowledge base and case-based reasoner avoided a 1.5 to two person-year effort to implement the ANSI standard format for all Air Products' MSDSs and to insure consistency throughout all documents. In addition, the case-based reasoner has relieved an estimated $1MM one-time cost of selectively testing all products in order to develop complete health and safety information on each product in the whole product line.

While these initial savings have been nearly offset by development expenditures, on-going benefits for ChemReg may be estimated in terms of costs that would be incurred under the manual system to achieve a comparable level of quality and consistency. In other words, what would the cost be to Air Products to return to the manual system and expect the same level of performance? Air Products estimates that it would take at least one to two person-days of additional work on each new product to achieve the same quality of results. In addition, there is an expected savings of $50K to $70K in annual expenditures for testing new products. It is estimated that ChemReg will accomplish $1MM worth of otherwise additional work and of avoided expenditures over five years.

The long term project plan calls for ChemReg to become a global hazard communication system. In 1996 European road transport regulations (ADR) were added to the transportation knowledge base, and during this year the MSDS knowledge base is being extended to classify products according the European Economic Union (EU) classificatory scheme for generating the Risk and Safety phrases required on a European MSDS. Expansion to Latin America and the Pacific rim is anticipated.

Summary

ChemReg is a large knowledge-based system used by Air Products and Chemicals, Inc. to support compliance with regulatory requirements for communicating health and safety information in shipping and handling of chemical products. This presentation has concentrated on one of the knowledge bases in this system: the case-based reasoner. Previously written knowledge bases for generating shipping descriptions and material safety data sheets mapped classificatory rules directly into knowledge base code. To support regulatory compliance, however, this strategy required estimates to be entered for all hazard-related properties on every product. To alleviate this burden an automated case-based reasoner was written that generates estimates of hazard data from similar products using an existing product database as its case library. While some refinements remain, the performance of the case-based reasoner has met expectations.

Acknowledgments

References

Environmental Protection Agency: Guidelines for the Health Risk Assessment of Chemical Mixtures. 1986. Federal Register 51(185):34014-34025.

Kitano, Hiroaki and Shimazu, Hideo. 1996. "The Experience-Sharing Architecture: A Case Study in Corporate-Wide Case-Based Software Quality Control". In Leake, David B. ed.. 1996. *Cased-Based Reasoning: Experience, Lessons, & Future Directions*, 235-268. Menlo Park, California: AAAI Press.

Kolodner, Janet. 1993. *Case-Based Reasoning*. San Mateo, California: Mogan Kaufmann Publishers, Inc..

Riesbeck, Christopher K. and Shank, Roger C.. 1989. *Inside Case-based Reasoning*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Sorani, Giorgio and Lauer, Robert A.. 1991. "The Application of Artificial Intelligence in the Field of Chemical Regulation". In Smith, Reid and Scott, Carlisle, eds. *Innovative Applications of Artificial Intelligence 3*, 55-64. Menlo Park, California: AAAI Press.