# Agent-Based Semantic Interoperability in InfoSleuth

Jerry Fowler†
Brad Perry†

Marian Nodine†
Bruce Bargmeyer‡

InfoSleuth Project†
MCC
Austin, Texas
{jfowler,nodine,bperry}@mcc.com

‡Office of Information Management
Environmental Protection Agency
Washington, D.C.
Bargmeyer.Bruce@epamail.epa.gov

## 1 Introduction

The InfoSleuth™ Project [1] at MCC has developed a distributed agent architecture that addresses the need for semantic interoperability among information sources and analytical tools within diverse application domains [4, 13]. InfoSleuth is being used as a significant component of the Environmental Data Exchange Network (EDEN)[2]. The current EDEN pilot demonstration enables integrated access via web browser to environmental information resources provided by offices of these agencies located in several states.

At the application level, InfoSleuth provides for semantic interchange among users by allowing an application developer to express the concepts and relationships of the application domain in high-level terms that are then translated into the low-level types of database schemas or semantic analyses of text and image resources. At the system level, InfoSleuth employs accepted standards where possible, to simplify data interchange and communication among processes.

To apply InfoSleuth in a specific application domain, it is necessary to identify the key elements of the business environment of the application, and create or discover an appropriate ontology for the domain, as well as identify the kinds of data that will be appropriate to the application. For the EDEN pilot demonstration, we are concentrating focus on sharing of information relating to remediation of hazardous waste contamination.

Several difficult problems are made apparent by an application such as EDEN:

- The different contexts in which users may examine data affect the way in which they wish to query the system and display the results.
- Ontologies used in semantic mapping must be adequately abstracted from physically available resources to ensure that new information sources can map to the same ontology. However, this abstraction ensures that exact mapping will be relatively rare.
- Many of the slots (attributes) contain values that are taken from one or more external ontologies. To cope with this satisfactorily requires traversing between and converting among multiple ontologies.
- Issues of uncertainty and imprecision of data are compounded by dirty data, aggregation and abstraction of data, dealing with multiple copies (some of which may be preferred over others), and mapping at both the schema and ontology levels.
- The semantics of the ontological concepts may be incomplete, uncertain, or evolving; thus it may be difficult to capture in an ontology.
- The embedded semantics of results generated based on solutions to any of these issues engenders a need for explaining to a user how a particular result came to be.

We do not address all of these issues in this paper.

[1] The InfoSleuth Project ended June 30, 1997, and is currently in phase two, called the InfoSleuthII Project. Some of the work described in this paper has come under the auspices of both projects. However, in the remainder of the paper we refer to both projects as simply "InfoSleuth". http://www.mcc.com/projects/infosleuth

[2] EDEN is a collaborative effort of three United States Government agencies, the Environmental Protection Agency (EPA), the Department of Defense (DOD), and the Department of Energy (DOE), with the European Environment Agency (EEA).

## 2 InfoSleuth Overview

InfoSleuth is an agent-based system designed to integrate heterogeneous, distributed information sources and tools via the use of common *ontologies*. In other words, a set, or community, of InfoSleuth agents collaborate at a semantic level to execute information gathering and analysis tasks, where the underlying information sources can be diverse both in their structure and content.

An InfoSleuth application is a collection of agents, coded in Java for portability and compatibility with popular Web browsers. The agents communicate via Knowledge Query Manipulation Language (KQML) [8], which implies communication at the semantic level over ontologies. The ontologies themselves are structured vocabularies representing the schematic metadata of a particular application domain. InfoSleuth agents employ the Open Knowledge Base Connectivity (OKBC) language standard [5] to communicate information about their ontologies and the constraints on the concepts in the ontologies.

Each agent in InfoSleuth provides a set of services that can be described as a set of tasks over the domain of InfoSleuth interaction.

- The user agent maintains a user's state, and provides the system interface that enables a user to communicate with the system independently of location.
- Broker agents match requests for services or information with agents that can provide them. Similar capabilities are described in [15].
- The ontology agent serves the set of ontologies supported by the InfoSleuth application and provides details of the ontology upon demand.
- Resource agents translates queries and data stored in some external data repository between their local forms and their InfoSleuth forms. The mapping done by a resource agent is similar to the mapping that is done traditionally between an internal schema and a conceptual schema in a multidatabase.
- Value mapping agents help convert queries and results between common acceptable forms and the canonical form defined in the EDEN ontology. The mappings done by a value mapper typically are either useful in multiple domains or are too complex or sophisticated to be addressed using traditional mechanisms.
- The multi-resource query agent handles the decomposition and distribution of sub-queries to various resource agents and then recomposes the results.

Numerous other agents perform special functions including specialized data aggregation and event detection.

Agents communicate and reason about each other's capabilities in terms of a shared ontological model of information management to resolve user requests. Requests are posed in terms of an ontology, called the "domain ontology of the application," that provides a semantic framework for information activities in the domain of the user's interest. Dynamic growth of agent communities is supported by means of semantic brokering, which allows agents to identify potential collaborators based on their advertised capabilities. The distribution of the agent community places low demands on the computation and storage power of a user's local machine, and means that access to resources that have registered with the broker is independent of the user's location. In addition, the user needs to know nothing about the physical location or structural characteristics of any resource (although it is within the system's power to report this information).

## 3 Real–World Concerns

The government participants in the EDEN project find the acquisition, use, and dissemination of environmental information to be of increasing strategic importance. Furthermore, congressional mandates have required increased interagency cooperation in sharing data regarding environmental remediation efforts. In these circumstances, where numerous legacy databases exist, each with differing schema and often with different database management software, InfoSleuth provides a natural way of integrating data from the various sources by means of a common ontology.

InfoSleuth provides adaptability that may help EDEN participants to address new congressional mandates or citizen information requests. Changing business requirements dictate that the domain ontology of an application will change, and that data sources may come, go, and evolve. These data sources were not necessarily designed to fit together, as in a distributed database. Neither are they explicitly integrated together using, say, schema integration techniques. This means that sources with dissimilar schemas and lexicons must be conceptu-

| | | | | |
|---|---|---|---|---|
| CERCLIS | Oracle | EPA | Crystal City, VA | Superfund site profiles |
| Hazdat | Sybase | EPA/CDC | Atlanta, GA | Toxicology information |
| ITT | MS-Access | EPA | MCC, Austin, TX | Remediation technology |
| EDR | Oracle | EPA | MCC, Austin, TX | Environmental Data Registry |
| ERPIMS | Oracle | DOD | Brooks AFB, TX | Air Force site profiles |
| IRDMIS | Oracle | DOD | Aberdeen, MD | Army site profiles |
| ERIP | Oracle | DOE | Idaho Falls, ID | DOE site profiles |
| OREIS | Oracle | DOE | Oak Ridge, TN | DOE site profiles |
| Basel | MS-Access | EEA | MCC, Austin, TX | Basel Convention transport data |

Table 1: EDEN Data Sources

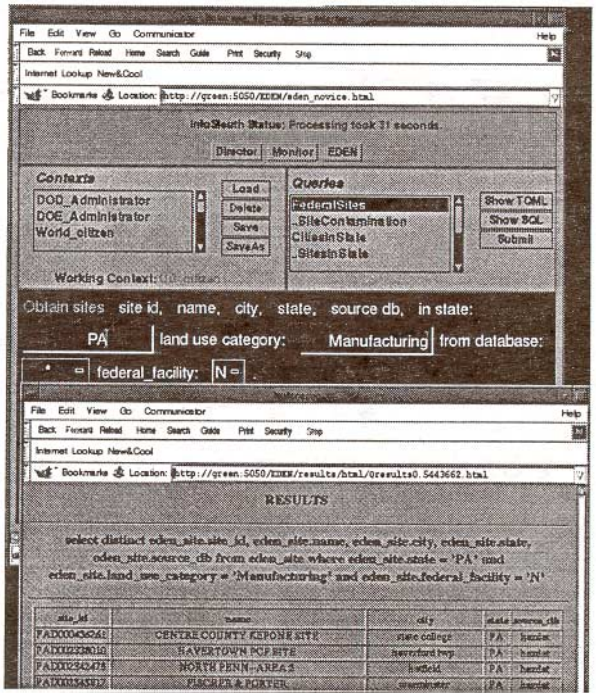


Figure 1: The TQML Browser displaying a query. Its results are displayed in a second browser window when received. Query terms marked with asterisks are removed from the generated SQL.

ally integrated in a dynamic manner. In these circumstances, value mapping and semantic translation, reasonably well understood in the context of schema integration, become dynamic problems that must be addressed in a flexible way. An agent-based system allows new functionality to be incorporated in an existing design, as well as allowing new or modified designs to be developed within the architecture of a functioning system.

A factor with a significant effect on the usability of a distributed information system such as EDEN is database size. While this may be better described as an engineering issue than a semantic one, it has a profound effect on the semantics of queries that can be considered acceptable to the system. Without addressing this issue in some way, no large-scale system can be viable. We are developing declarative definitions of semantic constraints on classes in the ontology such that a user agent can discourage or forbid a user from posing a query on a particular class without specifying an adequate constraint in the `where` clause.

# 4 Viewing Contexts

The EDEN pilot demonstration attempts to address these needs to query over heterogeneous resources by creating a resource agent for each of the many resources that maps to the common ontology, we provide access to all of these resources using a single ontological query framework. Although InfoSleuth supports the retrieval and extraction of concepts from text resources, none has currently been identified within the EDEN project. The heterogeneity of the pilot project at present is characterized in Table 1.

By creating multiple viewing contexts over portions of the ontology we allow different users with different needs to access different kinds of data in different ways. To address the need for for flexible query interfaces that allow the declarative construction of useful parameterized queries over the EDEN ontology, we have developed a query interface that manipulates a Template-based Query Markup Language (TQML) for specifying a mapping between natural language query fragments and SQL over an ontology, and representing the parameters through entry fields or domain-valued menus and list boxes. These query specifications are delivered to the browser by the User Agent; the user interface then populates choice lists from the user's locally materialized view of the ontology and uses the currently selected values to build the correct SQL query. This query is then passed to the InfoSleuth agent system to retrieve the appropriate in-

formation from EDEN resources. A sample TQML query specification and the appearance of the user interface it creates are shown in Figure 1.

We anticipate that sophisticated interaction with the user can improve the ability to deliver semantic content in the face of uncertainty with respect to results that may differ in reliability or granularity. We are addressing this by means of result annotations using Extended Markup Language (XML). These annotations will be linked to type-specific display objects that allow the user to drill through a result by querying the individual agent responsible for a particular component of the result. In this way, the user could learn, for example, on what raw data an individual transformation were based, or what sources contributed to a specific component of a result; armed with this information, the user could then modify the parameters of the semantic transformations and re-process the query.

# 5 Domain Ontologies

EDEN uses its domain ontology to support the ability of a user to communicate with other users and with data resources in the user's own terms. Whereas the use of a federated database model requires that the conceptual schema must be updated if new resources are added or old resources deleted, an InfoSleuth ontology is built independently of the form and availability of the actual data. This frees the user from the need to understand details of database schemas or to learn parts of the ontology irrelevant to the user's current needs.

The ontology used in the EDEN pilot project focuses principally on the relationships between contaminated sites, the wastes that cause the contamination, and technologies used to remediate specific kinds of contamination in specific media at each site. To derive a set of lexical terms for the ontology itself, we are in the process of incorporating terms from the EEA's General European Multilingual Environmental Thesaurus [1] (GEMET) into our domain ontology. GEMET provides a foundation for standardized vocabulary in EDEN and forms the basis for translation of queries and results between roughly a dozen languages from three continents.

One concept that illustrates several of the problems faced in EDEN is the measurement of levels of a contaminant at a site. Our resources include, at one extreme, those that have taken great care to provide detailed information not only on concentration, but also times and methods of both sampling and analysis, precise location and sample depth, and much other data necessary for a scientist to evaluate the progress of a cleanup effort. At the other extreme, there may be a single measurement for an entire site, or merely the indication that the chemical is present and being addressed in a remediation effort. The simple transformations on these slots are those relating to unit conversions and translation of geographic coordinates. Others relate to the meaning of an average concentration value, the relative accuracy of measurements made using different analytical techniques, and the meaning of values at or near the level of detection of a particular analysis. One interesting transformation involves the comparison of values over time when differing detection levels are the by-product of improved technology. Another involves comparison of qualitative and quantitative results. For example, where only the presence of a chemical is indicated, it might be inferred that the chemical has been detected at levels requiring federally mandated action. This might be factored into the transformation.

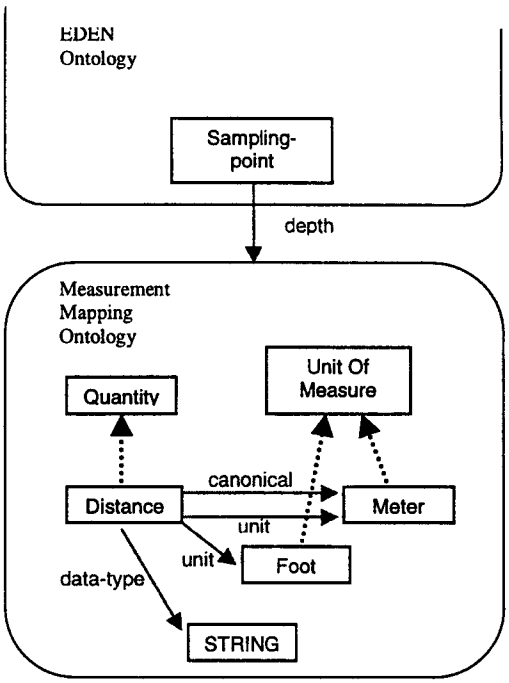


Figure 2: Composition of the EDEN and measurement mapping ontologies

An enhancement to InfoSleuth under current development will support construction of complex ontologies from smaller component ontologies. This reduces redundancy of expression and allows tools tailored for one component ontology to be used in many application domains. A salient example of smaller ontologies that might be incorporated into

the larger EDEN ontology is the ontological fragment relating to value mapping. Other natural candidates are units of measure, chemistry knowledge, and geographic metadata. An example relating to units of measure is shown in Figure 2.

# 6 Uncertainty

Data heterogeneity has many attributes. Differing data type, quantity, granularity, and quality each pose challenges. Extracting concepts from multimedia data to return them in the same result with data queried from structured relational databases implies an inherent difference in certainty about the accuracy of the results. Aggregating or summarizing large amounts of data can become critical not just to semantic matching, but also to efficient performance of the system. Statistical methods appropriate to these goals introduce uncertainty into the semantic equation. Aggregation and summarization can address the issue of differing granularity; to accomplish this it may be necessary to construct a semantic lattice of ontological terms and attempt to reason over a set of least common ontological terms. Again, the need to deal with uncertainty appears.

A "traditional" view of *uncertainty in heterogenous data* assumes that a probability, or membership function, can be attached to data items; then statistical methods can be used to aggregate and propagate uncertainty as information is combined from multiple sources. Unfortunately, this traditionalist view does not cover the range of uncertainty issues one encounters when piecing together information from actual information sources. Within our various applications of InfoSleuth for heterogeneous data gathering, we have found the issue of uncertainty and imprecision to manifest itself at various levels, only one of which can be addressed by traditional probability tactics. The following list describes the range of uncertainty issues that must be addressed in actual heterogeneous data gathering applications:

- *Varying levels of information aggregation*: The most common type of uncertainty we have encountered across heterogeneous data sources is that of data existing at different, though related, levels of granularity and aggregation. A simple example from the EDEN domain is the case where one data source has information about `individual chemicals` and another about `hazardous waste groups`. In this case, the ontology needs to contain consistent information hierarchies that allow for data sources to advertise information at the appropriate level of granularity.

- *Credibility and pedigree of information*: Since end-users are inevitably aware of the distributed nature of the information, they must be provided with enough ancillary information to establish credibility, or trust, in the information products. In other words, a heterogeneous data product cannot be represented as simply a standard "database result," it must be delivered with additional metadata describing the pedigree of the information itself and providing reference hooks to permit the user interface to query the responding agents about the nature of their results.

- *Aggregation accountability*: Closely related to information pedigree is the issue of accountability, or traceability, of information as it is aggregated and combined in an information network such as InfoSleuth.

- *Comparison operators and value domains*: Vertically aggregating, or joining, information from multiple sources requires the information network to establish a comparable value domain between the sources. In EDEN, we often find "many to many" mappings among values. An example is chemical names, codes, and groupings.

- *Information summaries*: A related issue to information granularities found in heterogeneous sources is that of information summaries desired by end-users. It is often the case that an end-user only wants an abstract summary of the information space with appropriate pointers to the detailed information should it be deemed necessary.

- *Probabilities and membership functions*: The final type of uncertainty we have encountered in our applications is the traditional view itself; that is, data from heterogeneous sources only approximately represents an ontological concept. There may be a probability or membership function that describes the degree of this representation. The information network must be prepared to combine and propagate probability measures as information aggregation is performed.

In InfoSleuth we have experimented with solutions to each of the above "uncertainty issues" in various application domains. All are applicable to

the EDEN project. Four techniques have arisen within InfoSleuth that help us harness these issues. The fourth, value mapping, is treated in the next section.

- *Hierarchical ontologies*: The InfoSleuth ontology model facilitates modeling an information domain with super/subclass and part-of relationships. Given the chemical versus chemical waste group problem cited above, we model this in the EDEN ontology with the following artifacts:

```
class Substance
class Waste (subclass-of Substance)
class Chemical (subclass-of Substance,
                part-of Waste)
```

  This allows us to query over `Substance` when the level of granularity in the data sources does not matter; yet we can always query at the appropriate level of granularity when necessary. Domain ontologies that contain a rich aggregation hierarchy allow resources to advertise the correct level of detail so that applications can query at the correct level of aggregation.

- *Information tagging*: One of the most important techniques incorporated into the InfoSleuth system is that of *information tagging*. As outlined above, a consistent and expressive "reporting structure" is paramount for any heterogeneous data gathering application. Within EDEN, every information product is tagged with the originating source (or sources) to which the product can be attributed.

- *Fuzzy functions*: A few of our prototype applications have required an approach where certain slots in the ontology are "marked" as "uncertainty measures" and the agents perform fuzzy joins whenever comparing or aggregating values in these slots.

## 6.1 Schema and Value Mapping

Ontologies specify a canonical representations both of the concepts in the application domain and of value-domains for the actual domain elements. Data represented in other value-domains can be mapped into the ontology's canonical value-domain by both resource agents and user agents so that they may relate values expressed in the conceptual domains in the ontology to data as stored in real world databases and as perceived by users.

To make this discussion concrete, conceptual domains represent types of values in specific contexts; for instance, chemical compounds. Each conceptual domain may have a set of value domains, one of which is canonical; for instance, chemical name ("Mercury"). Alternative value domains may include Chemical Abstracts Service (CAS) registry number (7439-97-6), "raw" CAS number (dashes removed—7439976), and common name ("quicksilver"). Within an ontology, each class has one or more slots, where each slot has a conceptual domain name with a canonical value domain over which all agents communicate when referring to that slot.

This mapping problem has several manifestations, which we relate roughly in the order they have been addressed in the literature (insofar as they are addressed at all):

- *Traditional:* Mapping between schemas can take place by imposing a view on that information and/or defining functions to translate the data from one value domain to another, as is currently done in relational databases, e.g., [9]. This type of mapping provides little support for semantics, but rather relies on the structure of the data.

- *Ontology-based:* A common ontology is defined with well-specified semantics for the concepts it describes. Mapping between a schema and an ontology is done on a semantic level [12]. Since the relationship between data items in a schema and semantic concepts in an ontology is often sloppy, the study of how to do this also addresses issues such as hyponyms and hypernyms, and the uncertainty that they introduce.

- *Reasoning:* Sometimes the conversion of values requires sophisticated reasoning or computation. This might occur, for instance, if data is measured over intervals, but the duration and boundaries of the intervals differ between information sources. The reasoning involved may be captured independently of the ontology. This is desirable especially when inferencing rules or computations are changing as new semantic knowledge is discovered.

- *Multi-ontology based:* In this world, the concepts in the common domain take their values from some external ontology. This occurs, as described previously, with chemical names. Here, mapping is explicitly specified among the external ontologies as relationships between a term in one ontology and the related terms in

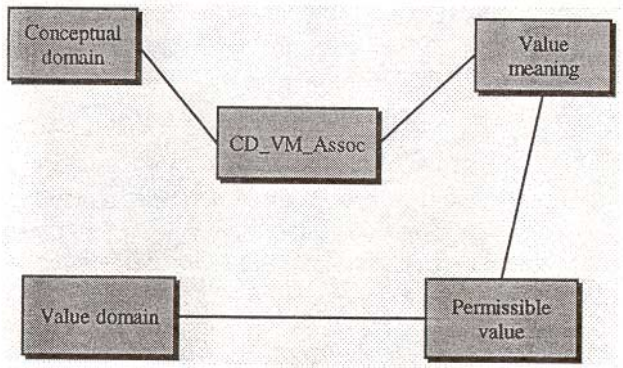


Figure 3: Value mapping schema from the EDR.

each other ontology. This process may be done independently of the shared ontology that describes the information being integrated.

- *Changing equivalences:* The values of specific attributes may take one form from some (changing) equivalence class. This case differs from the previous one in that membership is not fixed (i.e., one value per value domain), but may be very flexible. This type of mapping often occurs with hand-entered data, where people may use different abbreviations or misspellings for the same item (e.g., "sulfuric acid", "sulph. acid", "sufluric acid"). Again, many of these equivalences may be derivable from multiple information sources and populated independently of any specific shared ontology.

InfoSleuth addresses these problems at two levels, which are roughly divided between the resource agents and the value mapping agents. Resource agents implement tràditional and ontology-based mapping strategies for their own data. For some common types of mapping in (and outside of) EDEN, we have defined *mapping ontologies* that capture transformations among other ontologies. An example of this is the distance mapping ontology shown earlier in Figure 2.

### Value Mapping Agents

Infosleuth takes the approach of encapsulating common or sophisticated value mapping services into separate *value mapping agents*. These services include mappings that involve reasoning, the use of changing equivalence classes, and mappings that involve multiple ontologies. These agents map query terms and data to and from canonical value domains. Users query and view data in whichever value domain they prefer, and their user agents perform the value mapping necessary to communicate with other agents in the canonical value domain. Furthermore, resource agents advertise the canonical value domain even if they internally use a different value domain. This naturally implies that a resource agent's first act may be to request the address of a value mapping agent to help produce an advertisement that uses the canonical value domain.

There are numerous value domains in EDEN that warrant value mapping, including environmental media (e.g. soil, groundwater), land use categories, element of site characterization, chemical identifiers (name, CAS code) and state/province identifiers. The value mapping agent used in EDEN takes advantage of an important EPA tool with which to address data heterogeneity in the environmental domain, the Environmental Data Registry (EDR). The EDR is a reference implementation of the ISO/IEC1117 meta-data registry standard. It is a structured set of data types and related value sets that can be used both as a standard for data representation among cooperating agencies (development of the EDR is being related to the DOD's Defense Data Dictionary System in order to achieve a synergistic benefit of the two agencies' expertise), and as a resource for value mapping. A view of the EDR is shown in Figure 3.

## 7 Related Work

Work in federated and multi- database systems has a long history [9, 16]. This initial work was limited in that incorporating new information sources was difficult. Progress in this area was made with the introduction of mediation to facilitate integration, e.g. in [17]. Recent researchers have begun addressing the application of agent technology to the problem of heterogeneous data access [7, 13], which further facilitates the integration of data sources.

The problem of mapping between representations was irrefutably identified with the development of the ANSI/SPARC three-schema architecture [2]. This framework posed the goal of composing information from heterogeneous sources using a conceptual schema, formed by integrating the schemas of the component databases [3]. Schema integration techniques were used to develop the conceptual schema, and data was translated to and from the conceptual schema. Early approaches to schema and data transformation include views [6] that do structural mapping, and functional mappings on

the data [9]. These approaches are limited by the awkwardness of maintaining the conceptual schema.

Later systems such as Carnot use a shared ontology as a common basis for querying and sharing information [10, 11]. Typically, the integration of information sources begins with the definition of shared vocabulary defining the *semantic* concepts (ontology). Individual information systems then map their information onto this ontology. Naturally, this process is an uncertain one, as the structure and implied semantics of the information in the local source may not necessarily match the semantics of the ontology [12]. Unfortunately, using ontologies does not completely solve all issues of merging information, specifically falling short in places where semantic information is incomplete, uncertain, or changing [14]. Wrestling with these issues in the EDEN project motivated us to develop agents designed specifically for value mapping.

# 8 Conclusion

Our development work does not yet address many of the issues raised in this paper. Although there is much still to do before an EDEN system based on InfoSleuth can be deployed for the use of government personnel or citizens, the pilot project shows good promise as an agent-based system to address some of the concerns of semantic interoperability raised by the participants in the EDEN project.

# References


[1] European Environmental Agency. General european multilingual environmental thesaurus. *http://www.eea.dk/Locate/GEMET/default.htm*, 1997.

[2] The ANSI/X3/SPARC dbms framework: Report of the study group on data base management systems. *Information Systems*, 3, 1978.

[3] C. Batini, C. Lenzerini, and S. Navathe. A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys*, 18(4), 1986.

[4] R. Bayardo et al. Infosleuth: Semantic integration of information in open and dynamic environments. In *1997 ACM SIGMOD*, pages 195–206, Tucson, AZ, May 1997.

[5] V.K. Chaudhri et al. Open knowledge base connectivity 2.0. Technical Report KSL-98-06, Stanford University, 1998.

[6] U. Dayal and H.-Y. Hwang. View Definition and Generalization for Database Integration in a Multidatabase System. *IEEE Transactions on Software Engineering*, SE-10(6):628–645, Nov 1984.

[7] K. Decker and K.P. Sycara. Intelligent adaptive information agents. *Journal of Intelligent Information Systems*, 9(3):239–260, 1997.

[8] T. Finin, R. Fritzson, D. McKay, and R. McEntire. KQML as an agent communication language. In *Third Intl Conference on Information and Knowledge Management*, Nov 1994.

[9] Dennis Heimbigner and Dennis McLeod. A federated architecture for information management. *ACM Transactions on Office Automation Systems*, 3(3):253–278, Jul 1985.

[10] M. Huhns et al. Enterprise information modeling and model integration in Carnot. In *Enterprise Integration Modeling: Proceedings of the First Intl Conference*. MIT Press, 1992.

[11] J. Kahng and D. McLeod. Dynamic classificational ontologies: Mediators for sharing in a cooperative federated database. In *Proceedings of the first IFCIS Intl Conference on Cooperative Information Systems*, 1996.

[12] Vipul Kashyap and Amit Sheth. Semantic and schematic similarities between database objects: a context-based approach. *VLDB Journal*, 5, 1996.

[13] Marian Nodine, Brad Perry, and Amy Unruh. Experience with the InfoSleuth agent architecture. In *Proc. AAAI-98 Workshop on Software Tools for Developing Agents*, 1998.

[14] Aris Ouksel. Ontologies are not the panacea in data integration. *Distributed and Parallel Databases*, 7, 1999.

[15] Katia Sycara, Matthias Klusch, Seth Widoff, and Jianguo Lu. Dynamic service matchmaking among agents in open information environments. *SIGMOD Record*, 1999.

[16] G. Thomas et al. Heterogeneous distributed database systems for production use. *ACM Computing Surveys*, 22(3), Sep 1990.

[17] Gio Wiederhold. Mediators in the Architecture of Future Information Systems. *IEEE Computer*, 25(3):38–49, Mar 1992.