Interoperating Geographic Information Systems M F Goodchild, M J Egenhofer, R Fegeas and C A Kottman (eds). Kluwer.

# 0 CHANGING FOCUS ON INTEROPERABILITY IN INFORMATION SYSTEMS: FROM SYSTEM, SYNTAX, STRUCTURE TO SEMANTICS

Amit P. Sheth

## 1. Introduction

Interoperability has been a basic requirement for the modern information systems environment for over two decades. How have key requirements for interoperability changed over that time? How can we understand the full scope of interoperability issues? What has shaped research on information system interoperability? What key progress has been made? This chapter provides some of the answers to these questions. In particular, it looks at different levels of information system interoperability, while reviewing the changing focus of interoperability research themes, past achievements and new challenges in the emerging global information infrastructure (GII). It divides the research into three generations, and discusses some of achievements of the past. Finally, as we move from managing data to information, and in future knowledge, the need for achieving semantic interoperability is discussed and key components of solutions are introduced.

Data and information interoperability has gained increasing attention for several reasons, including:

- excellent progress in interconnection afforded by the Internet, Web and distributed computing infrastructures, leading to easy access to a large number of independently created and managed information sources of broad variety;
- increasing specialization of work, but increasing need to reuse and analyze data, leading to creation of information and knowledge, and their subsequent reuse and sharing.

Any attempt to give a broad survey of information systems interoperability is difficult and complex for several reasons, including different levels of requirements, the variety of approaches, the number of technical areas involved, the large literature, and the large number of example systems. For summative and pedagogical reasons, we make some broad observations and avoid paying attention to many exceptions.

We propose to view information system evolution in the context of interoperability as occurring in three generations: Generation I which covers the period to roughly 1985, Generation II which covers the period of a decade through 1995, and Generation III which covers a period yet to be bounded since 1996. Our discussion primarily draws from database and information system research, while noting that relevant work can also be found in information retrieval, knowledge-based systems and AI, multimedia, and other disciplines. Also very little attention is given in this chapter to several application domains, for example, geographic information systems (Goodchild et al. 1997), in which interoperability has received significant attention.

One of the enduring approaches to studying the key interoperability issues in multidatabases and federated databases of the first generation has been to use the fundamental dimensions of distribution, heterogeneity, and autonomy (Sheth and Larson 1990). The same dimensions also provide a good starting point for studying interoperability issues in the subsequent generations. After a brief discussion of the distribution and autonomy dimensions, we will focus on the heterogeneity dimension as we study interoperability issues in the three generations.

### *1.1. Distribution*

The scope of interoperability during the first generation was primarily departmental and almost always within a company. Usually, the multidatabase systems involved just a few databases and computer nodes, either connected point-to-point or in a local area network. With the significant impact of the Internet and advent of the Web, the scope of interoperability during the second generation has been enterprise-wide as well as inter-enterprise. It was not unusual to find tens of computers and data repositories involved in a second-generation system. In the third generation, with significant improvements in communication technology, global information infrastructure, and distributed computing infrastructure, the dimension of distribution of data has achieved a very broad scope—from a single system to global. As the distributed nature of data and information is often hidden from the end users, the system developers face several new challenges. A few of the noteworthy challenges involve increasing use of large amounts of data and information sources—particularly involving visual data, use of a wide variety of communication modes with a variety of bandwidths, and a larger optimization space involving varying capabilities of the component systems. Compared to the first generation systems, the issue of optimization has received less attention in the second generation.

### *1.2. Autonomy*

The organizational entities that manage different information sources (including database systems, DBSs) are often autonomous. Those who control an information source are often willing to let others share the data only if they retain control. Thus it is important to understand the aspects of autonomy and how they can be addressed when a database system participates in a federation or shares its data with new users or applications.

Let us look at a classification of autonomy issues in the context of federated database systems (adapted from Sheth and Larson 1990), noting that these can be adapted to other architectures by considering the various types of information sources and information system components involved. A component participating in a federation may exhibit several types of autonomy, including design, communication, association and execution.

*Design* autonomy refers to the ability of a component to choose its own design with respect to any matter, including

- the data or information being managed (i.e., the Universe of Discourse or domain),
- the representation (data model, query language) and the naming of the data elements (or the ontology used),
- the conceptualization or semantic interpretation of the data (or the context),
- constraints used to manage the data,
- the functionality of the system,
- association and sharing with other systems (see association autonomy below), and
- the implementation (e.g., record and file structures, concurrency control algorithms).

*Communication autonomy* refers to the ability of a component to decide whether to communicate with other components. A component with communication autonomy is able to decide when and how it responds to a request from another component. *Execution autonomy* refers to the ability of a component to execute local operations without interference from an external entity and to decide the order in which to execute external operations. Thus, an external system cannot enforce an order of execution of the commands on a component with execution autonomy. Execution autonomy implies that a component can abort any operation that does not meet its local constraints and that its local operations are logically unaffected by its participation in a federation. Furthermore, the component does not need to inform an external or federated system of the order in which external operations are executed and the order of an external operation with respect to local operations. Operationally, a component exercises its execution autonomy by treating external operations in the same way as local operations.

*Association autonomy* implies that a component has the ability to decide whether and how much to share its functionality (i.e., the operations it supports) and resources (i.e., the data it manages) with others. This includes the ability to associate or disassociate itself from the federation and the ability of a component to participate in one or more federations. Several first-generation systems in the database area paid significant attention to the autonomy issue because they also attempted to support updates. In comparison, few second-generation systems have considered update issues. Although autonomy of components has been assumed in almost all systems, there has been virtually no attention paid to the challenges posed by the various aspects of the autonomy dimension.

### *1.3. Heterogeneity*

Many types of heterogeneity are due to technological differences; for example, differences in hardware, system software (e.g., operating system), and communication systems. Researchers and developers have been working on resolving such heterogeneity for many years. Figure 1 shows one perspective on heterogeneity. Focusing on the crucial dimension of heterogeneity and corresponding solutions leads us to discuss different levels of interoperability—*system*, *syntax*, *structure*, and *semantic*. In this classification, we consider differences in machine-readable aspects of data representation, also referred to as formatting, to be relevant to syntactic heterogeneity. We consider representational heterogeneity that involves data modeling constructs to be relevant to structural interoperability. Schematic heterogeneity that particularly appears in structured databases is also an aspect of structural heterogeneity.

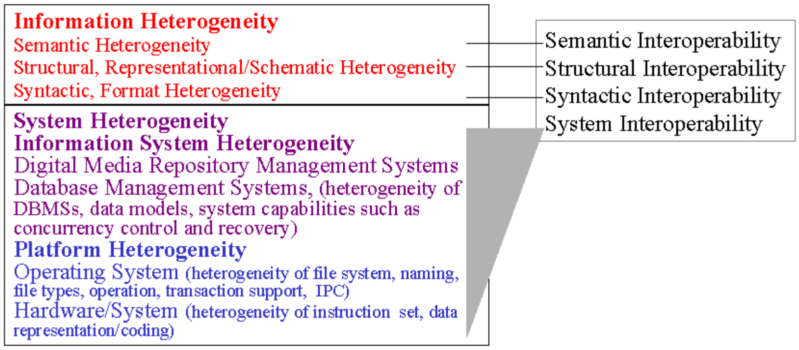

Information Heterogeneity
Semantic Heterogeneity
Structural, Representational/Schematic Heterogeneity
Syntactic, Format Heterogeneity
System Heterogeneity
Information System Heterogeneity
Digital Media Repository Management Systems
Database Management Systems, (heterogeneity of DBMSs, data models, system capabilities such as concurrency control and recovery)
Platform Heterogeneity
Operating System (heterogeneity of file system, naming, file types, operation, transaction support, IPC)
Hardware/System (heterogeneity of instruction set, data representation/coding)
Semantic Interoperability
Structural Interoperability
Syntactic Interoperability
System Interoperability

Figure 1. Heterogeneity in information systems

During the second half of the 1970s, we saw the ability to deal with hardware, operating systems, and communications heterogeneity; although with evolution in each of these, new issues have to be continuously addressed. During the 1980s, we saw significant progress in managing heterogeneity and support interoperability or integration in environments with structured databases and traditional database management systems (DBMSs). There is a large body of work during the first generation in dealing with heterogeneity associated with data models or schematic issues, DBMSs including query languages, concurrency control, commit and recovery, etc.

During the 1990s, the emergence of distributed computing, middleware technology, and standards has allowed us to increase focus on the heterogeneity that is intrinsic to data (or media). This has particularly supported syntactic and structural interoperability, and allowed us to address issues at the information level. As the future information system increasingly addresses the information and knowledge level issues, it will increasingly require semantic interoperability. Semantic interoperability requires that the information system understands the semantics of the user's information request and those of information sources, and uses mediation or information brokering to satisfy the information request as well as it can.

The remainder of this chapter provides an overview of the three generations of systems, with emphasis on the heterogeneity dimension of support for different levels of interoperability. Table 1 provides an overview of the three generations in terms of a variety of criteria.

| | **Generation I** | **Generation II** | **Generation III** |
|---|---|---|---|
| **Level of interoperation concern (new emphasis underlined)** | system, data | system, data, information | System, data, information, knowledge (incl. social), process |
| **Types of interoperability emphasized** | system (computer system and communication); limited aspects of syntax and structure (data model); transparency of location, distribution, replication, data models | syntax (data types and formats), structure (schematic, query languages and interfaces) | semantic (increasingly domain-specific) |
| **Dominant interoperability architecture** | multidatabases or federated databases | federated information systems, mediator | mediator, information brokering |
| **Scope of system interoperability** | handful of interconnected computers and databases | tens of systems on a LAN, databases and text repositories | enterprise-wide and global scope |
| **Software and information system architecture** | terminal access, point-to-point; also mainframes and minicomputers with remote access, client-server (two-tier); | client-server (three tier); | network, distributed, and mobile |
| **Communication infrastructure on which system interoperability solutions are built** | proprietary (IBM domination), TCP/IP | TCP/IP, http, CORBA | Internet/Web/Java, distributed object management, component, but increasingly higher-level such as multi-agent, mobile |
| **Types of data** | structured databases and files | structured databases, text repositories, semi-structured and structured and data in generic (e.g., SGML, | all forms of digital media with increasing support for visual/spatio-temporal/ scientific/engineering |

| | | | |
|---|---|---|---|
| | | HTML) and domain specific formats | data; |
| **Dominant information source/system model** | Relational and E-R | object-oriented | component-based; multi-modal |
| **Data/ information interoperability approaches** | structural and data model, data representation | understanding of a variety of metadata, comprehensive understanding of schematic heterogeneity | comprehensive use of metadata, increasing emphasis on semantics and ontology supported approaches |
| **Interoperability techniques (representative samples)** | data-level relationships, common/canonical data models, mappings, database exchanges, remote database interfaces, query transformations, schema translation, schema integration | schematic and metadata-level relationships, wrappers, extractors, single ontology, metadatabase, schematic heterogeneity, multidatabase consistency, mediators | multiple ontologies, information or semantic level relationships, context, media-independent information correlation, inter-ontological relationships, metadata consistency |
| **Key human roles in supporting interoperability** | data(base) administrators or experienced users, knowledgeable data structures and models, software developer written access programs | software developers to generate wrappers and mediators (with some toolkits) involving data level issues | domain experts for ontologies and for generating information correlations |
| **Access options** | database query language (SQL) for structured databases, keyword accesses for textual data/files | keyword-based attribute and (limited) content-based access, (limited) ontology-based access, | multimedia views; visual interfaces; information requests that are media-independent, multi-ontology based, context-sensitive and domain-specific |
| **A few representative applications** | integration of business databases or public databases | digital library, integrated access to heterogeneous data for a software team | digital earth, environmental phenomena, multi-step and multi-modal intelligence analysis |
| **One representative complex query** | Find a four star restaurant with less than $25 average cost that serves Mediterranean food in Richmond (a multidatabase query on distributed structured databases) | Find flowers suitable for winter gardens that look like *this* <image> with a soft smell (a keyword-, attribute-, and content-based query on text and image data repositories) | Find a block of land with urban land cover and moderate relief and population greater than 5000 and area greater than 1000 sq ft suitable for a strip mall (a query with terms whose meanings are understood by the system, and may involve multi-step processing against multi-modal data) |
| **Research prototypes** | ADDS, DDTS, Interbase, Mermaid, MRDSM, Multibase, | GARLIC, Harvest, HERMES, InfoHarness/Visual- | |

| | | | |
|---|---|---|---|
| | Omnibase, see Sheth and Larson (1990) for more examples | Harness, Information Manifold, InfoSleuth, RUFUS, SIMS, TSIMMIS, … | |
| **Products (Companies)** | UniSQL/M (UniSQL), Mermaid (Data Integration), DataJoiner (IBM), OmniConnect (Sybase) | AdaptX/Harness (Bellcore), (Junglee), TIE (Tesserae), (Excalibur) | |

Table 1: An overview of three generations of introperability R&D

## 2. First generation

By the 1980s, corporations had amassed large amounts of data in different departments to serve different applications, and on computers with different hardware and software (including DBMSs). The mantra of "data is a corporate resource" drove needs for exchanging and sharing the data between departments and within enterprises. Perhaps the most representative work in this generation occurred in the context of multidatabases (Litwin et al. 1982), or federated database systems (Heimbigner and McLeod 1985; Sheth and Larson 1990). We will focus on some of the key understandings gained during this period (Drew et al. 1993; Elmagarmid 1992; Elmagarmid and Pu 1990; Elmagarmid et al. 1998; Hsiao et al. 1993; Kambayashi et al. 1991; Kim 1995; Ram 1991; Schek et al. 1993; Sheth 1987).

Much of the emphasis during this generation was on achieving system interoperability, in particular by addressing the heterogeneity due to differences in DBMSs, with some work on syntactic heterogeneity and schematic heterogeneity as appropriate to the structure of databases. Correspondingly, the emphasis was on data management and data (as opposed to information or knowledge). Let us discuss these briefly, as good understanding on these issues has been achieved.

Each DBMS has an underlying data model used to define data structures and constraints. Both representation (syntactic as well as structural issues including constraints) and language aspects can lead to heterogeneity. Following are some of the issues that received significant attention from the database and information systems research community.

**Differences in structure**: Different data models provide different structural primitives (e.g., the information modeled using a relation or table in the relational model may be modeled as a record type in the network database model.). If the two representations have the same information content, it is easier to deal with the differences in structures. For example, address can be represented as an entity in one schema and as a composite attribute in other schema. If the information content is not the same, it may be very difficult to deal with the difference. As another example, some data models (notably semantic and object-oriented models) support generalization (and property inheritance) while others do not.

**Differences in constraints**: Two data models may support different constraints. For example, the set type in a schema based on a network database model may be partially modeled as a referential integrity constraint in a relational schema. However, a network database model supports insertion and retention constraints that are not captured by the referential integrity constraint alone. Triggers (or some other mechanism) must be used in relational systems to capture such semantics.

**Differences in query languages**: Different languages are used to manipulate data represented in different data models. Even when two DBMSs support the same data model, differences in their query languages (e.g., QUEL and SQL), or different versions of SQL supported by two relational DBMSs) could contribute to heterogeneity.

**Differences in the system aspects of the DBMSs**: Examples of system-level heterogeneity include differences in transaction management primitives and techniques (including concurrency control, commit protocols, and recovery), hardware and system software requirements, and communication capabilities.

A number of prototypes, including some of the well-known systems such as Multibase, Mermaid, DDTS, ADDS, or MRDSM addressed a variety of technical issues such those listed above. However there has been very little commercialization resulting from these efforts, and the success of commercial systems has been limited. Reasons are both technical and business (see Sheth 1995 for a discussion). Elmagarmid et al. (1998) provide a more recent review of technical work during this generation.

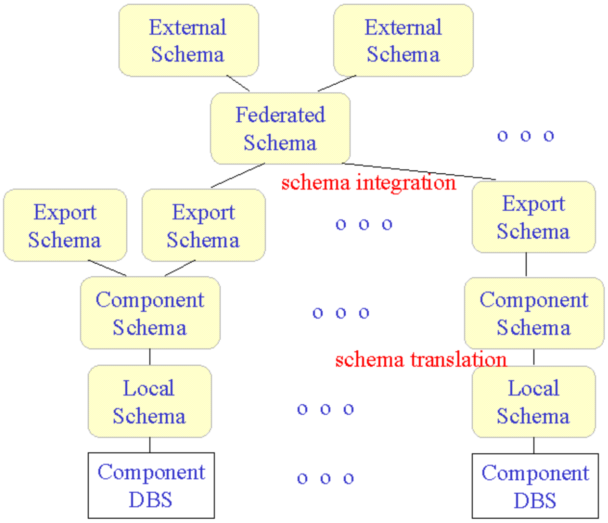


Figure 2. Schema architecture for a federated database system

Figure 2 shows how some of the distribution, autonomy, and heterogeneity issues for integrating component databases can be handled (see Sheth and Larson 1990 for more details). Briefly, the local schemas can represent data in the data model of respective DBMSs. To be able to compare the data objects modeled in different data models, one has either to perform direct and pairwise comparison—something like comparing apples and oranges—or to convert the schemas to a common or canonical model, preferably with an expressive power exceeding that of models for component databases, and then compare objects. Defining export schemas allows handling of one aspect of autonomy. Integration of export schemas into federated schemas allows for integrated or uniform access to objects managed by multiple component databases. Defining external schemas allows for handling additional types of heterogeneity.

DBMS-level heterogeneity covers only a small set of heterogeneity related to structure databases. Figure 3 shows one classification of a variety of conflicts related to achieving interoperability among or integration of multiple databases managed by traditional DBMSs (Kim et al. 1993; Sheth and Kashyap 1993).

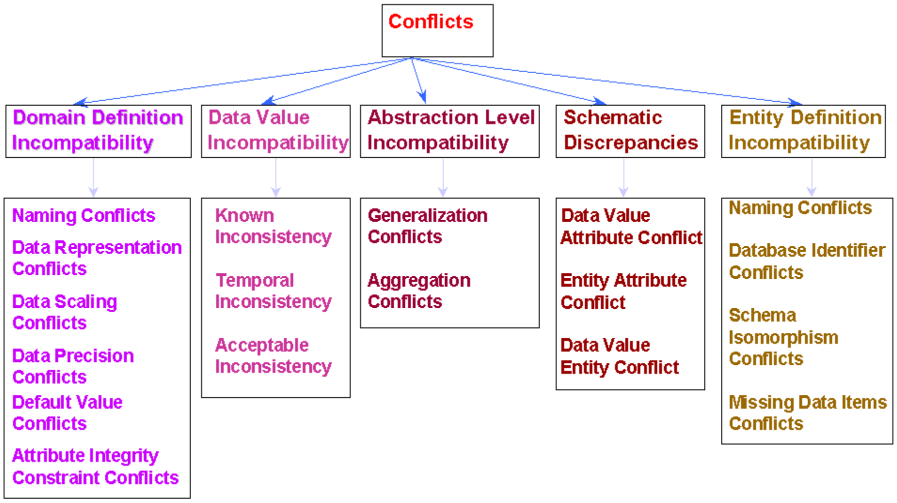

Conflicts
Domain Definition Incompatibility
Data Value Incompatibility
Abstraction Level Incompatibility
Schematic Discrepancies
Entity Definition Incompatibility
Naming Conflicts
Data Representation Conflicts
Data Scaling Conflicts
Data Precision Conflicts
Default Value Conflicts
Attribute Integrity Constraint Conflicts
Known Inconsistency
Temporal Inconsistency
Acceptable Inconsistency
Generalization Conflicts
Aggregation Conflicts
Data Value Attribute Conflict
Entity Attribute Conflict
Data Value Entity Conflict
Naming Conflicts
Database Identifier Conflicts
Schema Isomorphism Conflicts
Missing Data Items Conflicts

Figure 3. A classification of conflicts in structured databases

We believe that some of the lasting contributions of research in this generation are:

- understanding of the data models for structured databases, and the techniques to deal with data modeled with different models (so-called schema translation issue);
- techniques for schema integration (Batini et al. 1986) and understanding the complexity (and at times the futility) of performing integration involving real-world schemas (Sheth 1995);
- understanding of and finding ways to deal with schematic or representational heterogeneity, and the early realization of the distinction between schematic/representational issues and semantic issues in dealing with relationships and possible integration of data managed by different sources (or

DBMSs; Drew et al. 1993; Meersman and Mark 1997; Sheth 1991; Sheth and Kashyap 1993);

- issues of managing consistency among data managed by different databases, and corresponding research in advanced transaction models (Elmagarmid 1992) and multidatabase transactions execution (Georgakopolous and Rusinkiewicz 1994)

The predominant architectural framework during the first generation systems was that of the federated databases (Heimbigner and McLeod 1985; Sheth and Larson 1990). A federated database system (FDBS) architecture consists of a schema component, such as the one shown in Figure 2, interspersed with processors (such as transforming processors to help with heterogeneity, filtering processors to help with autonomy, and constructing processors to help with integration of distributed data). Two broad subcategories of FDBS architectures emerged:

- a loosely-coupled architecture that provided for a more dynamic or flexible federation (where it is easier for a component to join or leave), usually with no support for updates, but with reliance on more end user involvement and less transparency of resources; and
- a tightly-coupled architecture that provided for more stable federation, with ability to support updates and distributed transactions as well as better integration of data managed by components, but requiring more systems administrator involvement for the tasks such as schema integration

One of the major issues that this generation addressed but that subsequent generations have yet to consider is that of updates, and correspondingly data consistency and use of transaction mechanisms.

## 3. Second generation

During the second generation two very important trends brought extraordinary opportunities for interoperability and exploitation of data: (a) proliferation of a variety of data—from structured database, and semi-structured data, to digital media, including visual media (Gupta and Jain 1997), and (b) spread of the Internet and emergence of the Web. Applications such as digital libraries (Paepcke et al. 1998) and electronic commerce provided the context of interoperability.

Some of the key trends and achievements of this generation are (Papazoglou and Schlageter 1998; Sheth and Klas 1998):

- technology for dealing with heterogeneity of systems, data, and representational levels;
- support for a broader variety of data—not just structured databases, but also text, semi-structured, and unstructured (including image and video) data;
- use of a broad variety of metadata to support interoperability and integration; and
- use of knowledge representation and reasoning, especially for handling terminological differences.

### *3.1. Metadata*

Metadata are usually defined as data about data. Often they is more than that, involving information about data as they is stored or managed, and revealing partial semantics such as intended use (i.e., application) of data. This information can be of broad variety, meeting if not surpassing the variety in the data themselves. Metadata can be regarded as an extension (albeit a significant one) of the concept of the schema in structured databases. They may describe, or be a summary of the information content of the individual databases in an intentional manner. They typically represent constraints between the individual media objects that are implicit and not necessarily represented in the databases themselves. Some metadata may also capture content-independent information like location and time of creation. Examples of what we consider media types are structured data (data in relational or object-oriented databases), textual data (of different formats, such as Word files, source code, etc.), images (of possibly different modalities such as X-Ray, MRI scan), audio (of possibly different modalities such as monaural, stereophonic), and video. Sheth and Klas (1998) give an extensive discussion on types of metadata and their applications in managing and exploiting various digital media.

The criterion we use to classify metadata (Kashyap et al. 1995) is the extent to which they are successful in capturing the (data and information) content of the information asset (also called artifact or document in different contexts) represented in various media types. The level of abstraction at which the content of the assets is captured is very important. We believe that to capture the semantic content (i.e., at a level of abstraction closer to that of humans), it is important for the metadata to model application domain-specific information. Metadata descriptions present two advantages:

1. They enable the abstraction of representational details such as the format and organization of data, and capture the information content of the underlying data independent of representational details. This represents the first step in reduction of information overload as intentional metadata descriptions are in general an order of magnitude smaller than the underlying data.
2. They enable representation of domain knowledge describing the information domain to which the underlying data belong. This knowledge may then be used to make inferences about the underlying data. This helps in reducing information overload as the inferences may be used to determine the relevance of the underlying data without accessing the data.

One of several classifications of metadata (Kashyap et al. 1995) is as follows (also see Boll et al. 1998; Lagoze et al. 1996):

**Content-independent metadata**: This type of metadata captures information that does not depend on the content of the asset with which it is associated. Examples of this type of metadata are location, modification date of a document and type of sensor used to record a photographic image. There is no information content captured by these metadata but these might still be useful for retrieval of assets from their actual physical locations and for checking whether the information is current or not.

**Content-dependent metadata**: This type of metadata depends on the content of the asset it is associated with. Examples of content-dependent metadata are size of a document, maximum number of colors, number of rows, number of columns of an image. Content-dependent metadata can be further subdivided as follows:

**Direct content-based metadata**: This type of metadata is based directly on the contents of an asset. A popular example of this is full-text indices based on the text of the documents. Inverted tree and document vectors are examples of this type of metadata.

**Content-descriptive metadata**: This type of metadata describes the contents of an asset without their direct utilization. It often involves use of knowledge or human perception or cognition. An example is denoting the fragrance of an image containing a flower. Another example of this type of metadata is textual annotations describing the contents of an image.

**Domain-independent metadata**: This type of metadata captures information present in the document independent of the application or subject domain of the information. Examples of these are the C/C++ parse trees and HTML/SGML document type definitions.

**Domain-specific metadata**: Metadata of this type are described in a manner specific to the application or subject domain of information. Issues of vocabulary become very important in this case as the terms have to be chosen in a domain-specific manner. Examples of such metadata are relief, land cover from the GIS domain, and area and population from the Census domain. In the case of structured data, the database schema is an example of such metadata. Another interesting example is domain-specific ontologies, terms from which may be used as vocabulary to construct metadata specific to that domain.

Metadata may be precomputed (and possibly stored in a database) or they may be computed when needed (at query-processing time), in which case they may be represented by a computation, procedure, or method (e.g., an image processing routine giving values for land-cover metadata of a satellite image, executed when needed; Kashyap 1997).

Use of different types of metadata leads to different query or information access options. For example, three forms of information access, as supported in the VisualHarness system (Mudumbai 1997; Shah et al. 1997) are keyword-based access, attribute-based access, and content-based access. Table 2 shows relationships between various aspects of metadata for image data as exploited in the VisualHarness system. Furthermore, information access may involve iterative and customized (through use of different relative weights) use of one or more of these, as shown in Figure 5. Other types of information access options are possible, including standard, mediated, and immersive browsing (Grosky et al. 1998), semantic associative search (Kiyoki et al. 1998), and mixed media (Chen et al. 1998). Understanding the role of metadata and its use is likely to be one of the most enduring legacies of this generation of systems.

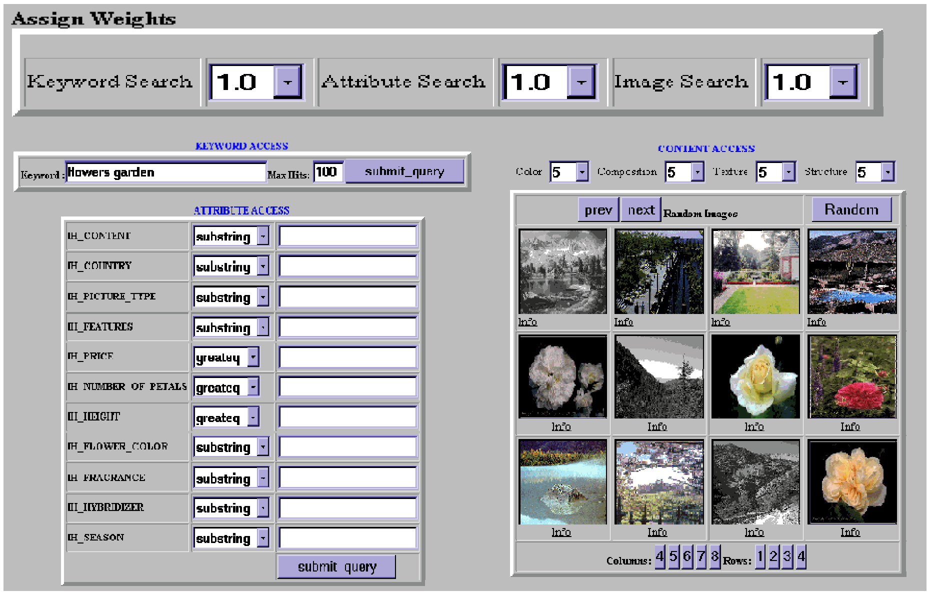
Some of the systems in this generation adapted the FDBS architecture to federated information systems architecture, where the DBMSs as managers of component information sources (databases) were replaced by a broader variety of information systems, including simple access protocols to access data, a broad variety of DBMSs (from network DBMSs to relational DBMSs to object-relational and object-oriented DBMSs), DBMSs specialized to manage specific digital media types (predominantly images), the Web itself as a manager of semi-structured data, and even expert systems. However, the mediator architectures (Wiederhold 1992) were clearly the dominant ones, involving wrappers for encapsulating heterogeneous information sources to provide more uniform interface to the rest of the world, and mediators to provide a broad variety of value-added services (Wiederhold 1997).

| **Metadata** | **Content proc. used** | **Semantic (domain-specific) info. used** | **Typical format of metadata** | **Primary access option** | **Example for the metadata of this type from image data** |
|---|---|---|---|---|---|
| **Content-indep.** | No | No | Attribute values | Attribute-based | Image height, width, size, date of creation, etc. |
| **Content-dep.** | Yes | No | Feature values | Content-based | Color, composition, texture, structure etc. (domain-independent) |
| **Content-based** | No | Yes | Attribute values | Attribute-based | Color (based on human interpretation), height (of the real-world object depicted by the image), fragrance, hybridizer of a flower; model, category etc. of an aircraft (domain-specific) |
| **Content-descr.** | May be | Either | Text and indices on text | Keyword-based | Descriptions of a flower, an x-ray, an aircraft (domain-specific); general descriptions of the image itself (domain-independent) |

Table 2: Image metadata classification and uses

During this generation, we saw increasing standardization or adoption of ad hoc standards, resulting in significant progress towards achieving system, syntactic, and structural interoperability. Acceptance of the Internet as a standard for interconnections between the systems, and evolution of infrastructures and middleware that support distributed computing (RMI, CORBA, and DCOM), and database connectivity (e.g., ODBC, JDBC etc. for relational databases), have had significant positive impact on achieving system interoperability. Syntactic interoperability includes the ability to deal with formatting and data exchange as supported by standards such as HTML for most current web-accessible documents, Z39.50 for bibliographic data, and MPEG-1 for pixel-level representation of image data. At the structural level, standardization for data modeling such as ANSI SQL, and object modeling standards and methodologies such as ODMG and UML have helped. Structural interoperability is further promoted by the use of IDL for distributed objects. Structural and a limited form of semantic interoperability are achieved by adoption of general purpose metadata standards, such as Dublin Core (OCLC 1997), as well as metadata standards in various domains such as bibliography (e.g., LCC and DDC; Beard and Smith 1998), space and astronomy, and geographical and environmental (e.g., FGDC and UDK; Günther and Voisard 1998).

Among a large number of systems representing this generation, three classes of systems stand out: systems focusing on information integration or uniform access to heterogeneous repositories, systems providing more dynamic architecture or query processing mechanisms to process a user service or information request on demand, and systems that address domain-specific or semantic-level issues. A brief review of these three classes of systems follows.



Assign Weights

Keyword Search 1.0 Attribute Search 1.0 Image Search 1.0

KEYWORD ACCESS

Keyword : flowers garden Max Hits: 100

ATTRIBUTE ACCESS

| | | |
|---|---|---|
| IH_CONTENT | substring | |
| IH_COUNTRY | substring | |
| IH_PICTURE_TYPE | substring | |
| IH_FEATURES | substring | |
| IH_PRICE | greateq | |
| IH NUMBER OF PETALS | greateq | |
| IH_HEIGHT | greateq | |
| IH_FLOWER_COLOR | substring | |
| IH FRAGRANCE | substring | |
| IH_HYBRIDIZER | substring | |
| IH_SEASON | substring | |

CONTENT ACCESS

Color 5 Composition 5 Texture 5 Structure 5

Figure 4. Keyword-, attribute-, and content-based access in VisualHarness

A small representative set of systems that support access to heterogeneous and distributed information sources includes TSIMMIS (*http://www-db.stanford.edu/tsimmis*), Information Manifold (*http://www.research.att.com/~levy /imhome.html*), GARLIC (*http://www.almaden.ibm.com/cs/showtell/garlic*), InfoHarness/VisualHarness (*http://lsdis.cs.uga.edu/infoharness*), SIMS (*http:// www.isi.edu/sims*), HERMES (*http://www.cs.umd.edu/projects/hermes*), and InfoSleuth (*http://www.mcc.com:80/projects/infosleuth*). SIMS, HERMES, and InfoSleuth also share capabilities with the third class as they support domain modeling or ontology, coupled with knowledge-based reasoning. WebSQL is an example of the second class of systems. There are also a number of systems that focus on specific media types or application domains, for example, QBISM (Arya et al. 1994) and CoBase (*http://burton.cs.ucla.edu/*) for image data management, and THETIS (Houstis et al. 1997) and Drew and Ying (1998) for environmental data and geographical information systems, etc. For brevity, we only discuss some of these systems.

TSIMMIS (Garcia-Molina et al. 1995) uses a mediator approach to combine information from several sources containing textual and semi-structured data. Data sources are encapsulated using wrappers or translators that logically convert the data to a common information model by translating information requests and results to this common model. The mediator layer above the wrappers is responsible for routing queries to appropriate sources and for post-processing the results. An important focus of the system is to generate wrappers and mediators automatically for a set of specified rules. Thus, TSIMMIS provides a framework for users to specify information integration, which may be done manually or in a semi-automated manner.

The Information Manifold (Levy et al. 1995) is a system for retrieval and organization of information from disparate (structured and unstructured) information sources. The architecture of Information Manifold is based on a knowledge base containing a rich domain model that enables describing the properties of the information sources. The user can interact with the system by browsing the information space (which includes both the knowledge base and the information sources). The presence of descriptions of the information sources also enables the user to pose high-level queries based on the content of the information sources. One area of focus in the project has been to optimize the execution of a user query expressed in a high-level language, which might potentially require access to and combination of content from several information sources.

The InfoHarness system (Shklar et al. 1995) and its commercial counterpart AdaptX/Harness (Sheth 1996) provide browsing as well as keyword- and attribute-based access (involving typed attributes and logical operators) to a broad variety of Web-accessible heterogeneous documents (e.g., unstructured text, semistructured text such as AP news and emails, word processor and source code files, data accessible through well-defined interfaces such as NNTP news group server), and relational databases. Some of the key features are extensibility of metadata and corresponding dynamically-created query interface with typed attributes, logical remodeling of information space (such as browsing or searching source code files by function signatures), and support for multiple, third-party indexing strategies. The VisualHarness system (Mudumbai 1997; Shah and Sheth 1998) further extended it to support images using third-party visual image retrieval engines, and customizable access involving keyword-, attribute-, and content-based access.

WebSQL (Mendelzon et al. 1997a) is a high-level declarative query language for extracting information from the Web. WebSQL takes advantage of multiple index servers without requiring users to know about them, and integrates full text with topology-based queries. This enables definition of the content of domain-specific text indexes. WebSQL is used to define logical views on the unstructured global repository of Web-accessible documents. A level above Web SQL is the Web Semantics Query Language (WSQL) (Mendelzon et al. 1997b). This system has different layers of abstraction and provides mechanisms for describing the data that are available, for discovering the existence of data relevant to a problem, and for accessing discovered relevant data. WSQL has constructs for source discovery via controlled Web navigation, source registration in domain-specific catalogs, associative selection of sources from existing catalogs, and uniform access to data stored in heterogeneous sources.

There are several systems that employ the metadata-based semantic view of the world and employ an ontological layer above it. In the SIMS project (Arens et al. 1996) a model of the application domain is created using a knowledge representation system to establish a fixed vocabulary describing objects in the domain, their attributes, and relationships among them. For each information source a model is constructed that indicates the data model used, query language, network location, size estimates, etc., and describes the contents of its fields in relation to the domain model. Queries to SIMS are written in the high-level uniform language of the domain model. SIMS determines the relevant information sources by using the knowledge encoded in the domain model and the models of the information sources. These information sources are determined at run time based on their availability at that time.

The HERMES (Adali and Subrahmanian 1994) system follows the mediator architecture for semantically integrating different and possibly heterogeneous information sources (including those containing visual data) and reasoning systems. This integration is done using mediators that are very similar to the ones in the TSIMMIS system described above. Mediators, in HERMES, are logical guidelines of how information from different sources will be combined and integrated. In this framework, external information sources are abstracted as domains which execute certain functions with pre-specified input and output types. These domains are accessed in mediators using a logic-based declarative language. The system also provides a uniform environment for adding new external sources to existing mediators.

The InfoSleuth system (Bayardo et al. 1997) views an information source at the level of its relevant semantic concepts, thus preserving the autonomy of its data. Information requests to InfoSleuth are specified generically, independent of the structure, location, or even existence of the requested information. InfoSleuth filters these requests, specified at the semantic level, flexibly matching them to the information resources that are relevant at the time the request is processed. The InfoSleuth approach is to specify a common ontology for a domain, and local mappings from individual database schemas to the common ontology. These mappings can be thought of as views of the data that simplify query specification for selecting information. Given an appropriate set of mappings for a particular knowledge discovery task, the InfoSleuth system provides query support for selecting relevant information. It also pre-processes and transforms the underlying database data into records whose attributes consist of concepts from the ontology. Early emphasis in the InfoSleuth system was on harnessing structured databases and support for text data is also reported.

## 4. Third generation

One thing that has not changed for the third generation is that we are once again faced with more distribution, more autonomy, and more heterogeneity among the accessible information, information sources, and users. With the progress in global interconnectivity, we now need to deal with more heterogeneous information consisting not only of a broader variety of digital data, but also operations and computations (such as simulations) that can create new data and information. The scale of the problem has changed from a few databases to millions of information resources, and the new resources are added independently to the accessible set of resources, as other resources change rapidly or disappear. Currently favorite strategies that depend on keyword-based access or involve only representational or structural components of data are usually found to provide a poor quality of result, and their lack of precision leads to increasing information overload. We fully expect increasing standardization and interoperability at system, syntactic, and structural levels to address many issues—for example, see Paepcke et al. (1998) for relevant work in the domain of digital libraries. However, the key challenges to be faced are at the semantic level, where people would increasingly expect the information systems to help them not at the data level, but at the information, and increasingly knowledge levels.

Even a casual user of the Web is aware of the rapid increase in the amount and diversity of information available online. However, what is creating an even bigger challenge is the increased expectations of the user in terms of understanding of the context of the user's information need, increasing availability of semantically rich visual and new media, and a corresponding need to support semantic-level interoperability. The problem of information overload has turned the challenge of "So far (schematically) yet so near (semantically)" (Sheth and Kashyap 1993) faced by the previous generations into "So near (syntactically and structurally) yet so far (semantically)".

Although there are several uses and interpretations of *semantics* in information systems, our view is that future information systems will need to support a more general notion that involves relating the content and representation of information resources to entities and concepts in the real world (Beech 1997; Meersman 1997; Sheth 1997). That is, the limited forms of operational and axiomatic semantics of a particular representational or language framework are not sufficient (see Paepcke et al. 1998 for a relevant discussion on syntax and some types of semantics). Semantic interoperability will then support high-level (hence easier to use), context-sensitive information requests over heterogeneous information resources, hiding system, syntax, and structural heterogeneity. In essence, we need an approach that reduces the problem of knowing the contents and structure of many information resources to the problem of knowing the contents of easily-understood, domain-specific ontologies, which a user familiar with the domain is likely to know or understand easily.

Foundational research leading to building the third generation of information systems has been carried out in several umbrella projects and initiatives, including Knowledge Sharing Effort (*http://www-ksl.stanford.edu/knowledge-sharing*), Intelligent Integration of Information (*http://mole.dc.isx.com/I3*), and the Digital Library Initiative (*http://www.cise.nsf.gov/iis/dli_home.html*). Systems belonging to the third class of the second generation have also made contributions that the third generation systems can build on. Increasing standardization at different levels of information systems architecture for corresponding type of interoperability also plays an important role. Some of the examples are as follows.

- *System*: IIOP for interactions between distributed objects and components, KQML for interaction between agents;
- *Syntactic*: XML for all forms of Web-accessible data;
- *Structural*: RDF for general purpose description of information sources, various object models for web-based information exchange (Manola 1998), MPEG-4 for structural or object-level description video, MHEF-5 for multimedia and hypermedia, KIF for knowledge representation, OKBC for distributed knowledge bases;
- *Semantic*: MPEG-7 (still in progress) with likely support for limited forms of semantics with identification of context, objectives requirements, and applications.

We now focus our attention on a discussion of possible enablers of semantic interoperability. In particular, we identify three enablers and capabilities:

**Terminology (and language) transparency**: This will allow a user to choose an ontology of his or her choice (e.g., one based on LCC for querying bibliographic data or FGDC for geospatial data), while allowing the information source to subscribe to a related but different ontology (e.g., an ontology based on DDC or UDK, repsectively. The latter recognizes some overlap between geospatial data sets and environmental data sets, and their respective modeling).

**Context-sensitive information processing**: The information system will recognize or understand the context of an information need and use it to limit information overload, both by formulating more precise queries used for searching information sources and by filtering and transforming the information before presenting it to the user.

**Semantic correlation**: This will allow the representation of semantically-related information regardless of distribution and heterogeneity (including various forms of media) by the user or the third party, and their use for obtaining all forms of relevant information anywhere.

Three key components of a possible solution are metadata (especially domain-specific and content-based metadata), contexts, and ontologies (Kashyap and Sheth 1998). We briefly discuss their role in developing semantic interoperability solutions. One key aspect of the third generation (operation or process interoperation) will not be discussed for brevity.

### *4.1. Ontologies and terminology transparency*

An ontology can be defined as a specific vocabulary and relationships used to describe certain aspects of reality, and a set of explicit assumptions regarding the intended meaning of the vocabulary of words (Gruber 1991; Guarino 1998). Among various other classification schemes and structures, including keywords, thesauri, and taxonomies, ontologies are often viewed as allowing more complete and precise domain models (Huhns and Singh 1997). Support and use of multiple, independently-developed ontologies is important for developing scalable information systems with multiple information producers and consumers (e.g., Arens et al. 1996; Dao and Perry 1996; Genesereth and King 1995; Kashyap and Sheth 1998; Khang and McLeod 1998 for need and use of multiple ontologies). One challenging issue in supporting semantic interoperability is how to allow both users and providers to subscribe to existing ontologies of their choice or create a new one (Kashyap and Sheth 1998). Processing an information request represented in terms of one ontology in an environment with information resources that subscribe to different (but related and relevant) ontologies may involve using inter-ontological relationships, such as synonym, hypernym, homonym, and other possibly domain-specific relationships. This work also requires understanding of and containing loss of information in multi-ontology query processing (Mena et al. 1998). One early example of research along these lines is the OBSERVER (sub)system (*http://siul02.si.ehu.es/~jirgbdat /OBSERVER*), which is a component of the InfoQuilt system (*http://lsdis.cs.uga.edu/infoquilt*).

### 4.2. Context

In characterizing the similarity between objects based on the semantics associated with them we have to consider the real-world semantics (RWS) of an object. It is not possible to completely define what an object denotes or means in the model world. We propose the *context* of an object as the primary vehicle to capture the RWS of the object. Understanding of the context of the information request can help the system to distinguish between whether the term *cricket* refers to an insect or a sports game.

Adapting from research in AI and Knowledge-Based systems (e.g., Shoham 1991), linguistics and other fields, modeling and representing context can lead to several benefits in dealing with information overload in a global information infrastructure (GII; see Kashyap and Sheth 1998 for more details):

- *Economy of representation*: In a manner akin to database views, contexts can act as a focusing mechanism when accessing the component databases or information sources on the GII.

- *Economy of reasoning*: Instead of reasoning with the information present in the database as a whole, reasoning can be performed with the context associated with an information source.

- *Managing inconsistent information*: In the GII, where information sources are designed and developed independently, it is not uncommon to have information in one source be inconsistent with information in another. As long as information is consistent within the context of the query of the user, inconsistency in information from different databases may be allowed.

- *Flexible semantics*: An important consequence of associating abstractions or mappings with context is that the same two objects can be related to each other differently in two different contexts. Two objects might be semantically closer to each other in one context as compared to the other.

There are several proposals for representing context. We believe that an effective approach needs to bring together metadata, user profiles, information modeling abstractions, and ontologies, as well as to allow their dynamic construction to model application domain and user needs. Besides their modeling and representation, a key challenge includes the ability to reason about or compare contexts (e.g., Kashyap and

Sheth 1996; Lee et al. 1996; Ouksel and Naiman 1994). While there are many representations and associated reasoning techniques, practical application of context in GII is expected to be a key research challenge for achieving semantic interoperability in information systems.

### 4.3. Information correlations

One of the key applications of semantics in GII is to represent or specify information requests and semantic level information correlations regardless of the media (and other heterogeneity) and locations of information sources. These can involve queries over heterogeneous media assets represented at a higher level of abstraction in media-independent manner, using metadata and ontologies.

Two approaches to *representing* information correlations between independently-managed networked resources are Metadata Reference Links (MREFs; Shah and Sheth 1998; Sheth and Kashyap 1996) and Distributed Active Relationships (DARs; Daniel et al. 1998). They provide an initial step in specifying information correlation between heterogeneous digital media. Specifically, MREFs allow subscription to one or more ontologies in their specification, and the meta-information used in specifying an MREF is mapped to views involving keyword-based, attribute-based, and content-based specifications involving various types of metadata of heterogeneous digital media. Specification and processing based on information correlations can be easily integrated with the Web technology. For example, MREF could be used anywhere a hypermedia link (HREF) is used, and its specification and processing can be supported using an RDF and XML-based infrastructure. However, many challenges remain in extending the current proposals to include non-standard resources such as datasets and procedures, integrating information correlation representation and processing with context and context mediation, and processing them efficiently in a very large information space.

### 4.4. Information-brokering architecture

It is hard to predict the architecture of the third generation of information systems. One proposed architecture, information brokering (Kashyap 1997; Kashyap and Sheth 1994), adapts and extends the concepts of (1) federated environments (Heimbigner and McLeod 1985; Sheth and Larson 1990) in which resources, metadata, and ontologies are created, administered, and enhanced independently; and (2) mediator architectures (Wiederhold 1992) which involve decoupling information creators and providers from information users and better semantic-level services and interoperability. However, we believe that the key to this generation of systems is their support for semantic interoperability, through exploitation of various forms of metadata, multiple ontologies, and contexts. Furthermore, we believe that before very general architectures that can support various domains can be developed, support for semantic interoperability demands that we focus on a specific domain first, such as GIS (Goodchild et al. 1997), and then extend what we learn to general-purpose and multi-domain environments. Figure 6 shows a schematic of a system for supporting semantic interoperability as described above in a geographical domain. It is too early to give representative examples of the third generation, but a few early efforts are described by Wiederhold (1996), and Papozoglou and Schlageter (1998); and see InfoQuilt (*http://lsdis.cs.uga.edu/infoquilt*).

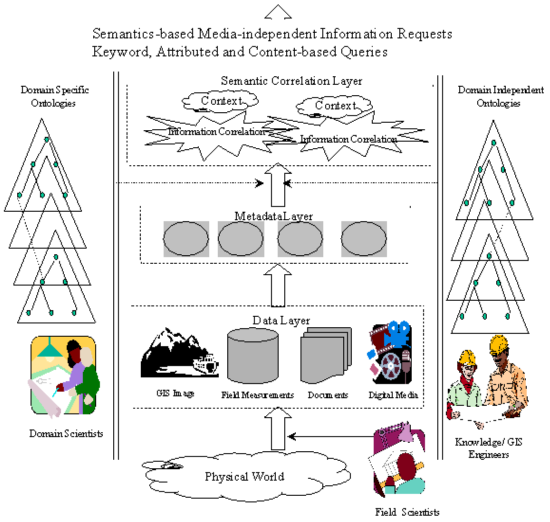


Figure 5. An architecture to support geographic information brokering

In closing, we believe that semantic interoperability is the key to progress towards a vision of Infocosm, a society whose members will have information anywhere, any time, and in many forms, for knowledge creation and use, effective decision-making, better learning, and more fun.

## Acknowledgements

Members of the InfoQuilt project, especially Vipul Kashyap, Kshitij Shah, Clemens Bertram, and Krishnan Parsuraman, have contributed to some of the ideas presented in this paper.

## References

Adali S, Subrahmanian V S 1994 Amalgamating knowledge bases, II: Distributed mediators. *International Journal of Intelligent and Cooperative Information Systems* 3(4): 349–383

Arens Y, Knoblock C A, Shen W 1996 Query reformulation for dynamic information integration. In Wiederhold G (ed) *Intelligent Integration of Information*. Kluwer Academic Publishers: 11-42

Arya M, Cody W, Faloutsos C, Richardson J, Toga A 1994 QBISM: extending a DBMS to support 3D medical images. *Proceedings of the 10th International IEEE Conference on Data Engineering, Houston, Texas*: 314-325

Batini C, Lenzerini M, Navathe S 1986 A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys* 18(4): 323–64

Bayardo R, Bohrer W, Brice R, Cichocki A, Fowler G, Helal A, Kashyap V, Ksiezyk T, Martin G, Nodine M, Rashid M, Rusinkiewicz M, Shea R, Unnikrishnan C, Unruh A, Woelk D 1997 Semantic integration of information in open and dynamic environments. *Proceedings of the 1997 ACM International Conference on the Management of Data (SIGMOD), Tucson*: 195-206.

Beard K, Smith T 1998 A framework for meta-information in digital libraries. In Sheth A, Klas W (eds) *Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media*. McGraw Hill: 341-365

Beech D 1997 Data semantics on the information superhighway. In Meersman R, Mark L (eds) *Database Application Semantics*. Chapman and Hall

Boll S, Klas W, Sheth A 1998 Overview on using metadata to manage multimedia data. In Sheth A, Klas W (eds) *Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media*. McGraw-Hill Publishers: 1–23

Chen F, Hearst M, Kimber D, Kupiec J, Pedersen J, Wilcox L 1998 Metadata for mixed-media access. In Sheth A, Klas W (eds) *Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media*. McGraw Hill: 319-340

Daniel R, Lagoze C, Payette S 1998 A metadata architecture for digital libraries. Proceedings of the Forum on Research and Technology Advances in Digital Libraries (ADL'98), Santa Barbara: 276-288

Dao S, Perry B 1996 Information mediation in cyberspace: scalable methods for declarative information networks. In Wiederhold G (ed) *Intelligent Integration of Information*. Kluwer Academic Publishers: 43-62

Drew P, Ying J 1998 Metadata management for geographic information discovery and exchange. In Sheth A, Klas W (eds) 1998 *Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media*. McGraw Hill: 89-121

Drew P, King R, McLeod D, Rusinkiewicz M, Silberschatz A 1993 Report of the workshop on semantic heterogeneity and interoperation in multidatabase systems. *SIGMOD Record* 22(3): 47–56

Elmagarmid A (ed) 1992 *Database Transaction Models for Advanced Applications*. Morgan-Kaufmann

Elmagarmid A, Pu C (eds) 1990 Heterogeneous databases. Special Issue. *ACM Computing Surveys* 22 (3)

Elmagarmid A, Sheth A, Rusinkiewicz M (eds) 1998 *Heterogeneous Distributed Databases*. Morgan Kaufmann Publishers

Garcia-Molina H, Hammer J, Ireland K, Papakonstantinou Y, Ullman J, Widom J 1995 Integrating and accessing heterogeneous information sources in TSIMMIS. *Proceedings of the AAAI Symposium on Information Gathering, Stanford, California:* 61-64

Genesereth M, King R (eds) 1995 *Reference Architecture, Intelligent Integration of Information.* Stanford University and University of Colorado. *http://logic.stanford.edu /achitecture/reference.html*

Georgakopoulos D, Rusinkiewicz M 1994 Using tickets to enforce the serializability of multidatabase transactions. *IEEE Transactions On Knowledge and Data Engineering* 6(1): 166–180

Goodchild M, Egenhofer M, Fegeas R 1997 *Interoperating GISs: Report of a Specialist Meeting Held under the Auspices of the Varenius Project, Panel on Computational Implementations of Geographic Concepts*. National Center for Geographic Information and Analysis, Santa Barbara

Grosky W, Fotouhi F, Jiang Z 1998 Using metadata for the intelligent browsing of structured media objects. In Sheth A, Klas W (eds) *Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media*. McGraw Hill: 123-148

Gruber T 1991 The role of a common ontology in achieving sharable, reusable knowledge bases. *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning, Cambridge:* 601-602

Guarino N 1998 Formal ontology and information systems. *Proceedings of the 1st International Conference on Formal Ontology in Information Systems [FOIS'98], Torino*: 3-15

Günther O, Voisard A 1998 Metadata in geographic and environmental data management. In Sheth A, Klas W (eds) *Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media*. McGraw Hill: 57-87

Gupta A, Jain R 1997 Visual information retrieval. *Communications of the ACM* 40(5): 70–79

Heimbigner D, McLeod D 1985 A federated architecture for information management. *ACM Transactions on Office Information Systems* 3(3): 253–278

Houstis C, Nikolaou C, Marazakis M, Patrikalakis N, Sairamesh J, Thomasic A 1997 THETIS: Design of a data repositories collection and data visualization system for coastal zone management of the Mediterranean Sea. *D-Lib Magazine, The Magazine for Digital Library Research,* November. *http://www.dlib.org/dlib/november97/thetis/11thetis.html*

Hsiao D, Neuhold E, Sacks-Davis R (eds) 1993 *Interoperable Database Systems (DS-5*., North-Holland

Huhns M, Singh M 1997 Ontologies for agents. *Internet Computing* 1(6): 81-83

Kambayashi Y, Rusinkiewicz M, Sheth A (eds) 1991 *Proceedings of the RIDE-IMS'91: International Workshop on Interoperability in Multidatabase Systems.* IEEE Computer Society

Kashyap V 1997 *Information Brokering over Heterogeneous Digital Data: A Metadata Based Approach*. PhD Thesis, Rutgers University

Kashyap V, Sheth A 1994 Semantics based information brokering. Proceedings of the 3rd International Conference on Information and Knowledge Systems: 363-370

Kashyap V, Sheth A 1996 Schematic and semantic similarities between database objects: a context-based approach. *The Very Large Databases Journal* 5(4): 276-304

Kashyap V, Sheth A 1998 Semantic heterogeneity in global information systems: the role of metadata, context and ontologies. In Papazoglou M, Schlageter G (eds) *Cooperative Information Systems: Current Trends and Directions*. Academic Press: 139-178

Kashyap V, Shah K, Sheth A 1995 Metadata for building the MultiMedia Patch Quilt. In Jajodia S, Subrahmanian V S (eds) *Multimedia Database Systems: Issues and Research Directions.* Springer-Verlag: 297-319

Khang J, McLeod D 1998 Dynamic classificational ontologies: mediation of information sharing in cooperative federated database systems. In Papazoglou M, Schlageter G (eds) *Cooperative Information Systems: Current Trends and Directions*. Academic Press: 179-203

Kim W (ed) 1995 *Modern Database Systems: The Object Model, Interoperability and Beyond*. Addison Wesley

Kim W, Choi I, Gala S, Scheevel M 1993 On resolving schematic heterogeneity in multidatabase systems*. Distributed Parallel Databases International Journal* 1: 251-279

Kiyoki Y, Kitagawa T, Hayama T 1998 A metadatabase system for semantic image search by a mathematical model of meaning. In Sheth A, Klas W (eds) *Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media*. McGraw Hill: 191-222

Lagoze C, Lynch C, Daniel R 1996 The Warwick Framework: A container architecture for aggregating sets of metadata. Technical Report TR96-1593. Cornell University, Department of Computer Science. *http://cs-tr.cs.cornell.edu:80/Dienst/UI/2.0/ Describe/ncstrl.cornell/ TR96-1593*

Lee J, Madnick S, Siegel M 1996 Conceptualizing semantic interoperability: a perspective from the knowledge level. *International Journal of Cooperative Information Systems* 5(4): 367–393

Levy A Y, Srivastava D, Kirk T 1995 Data model and query evaluation in global information systems. *Intelligent Information Systems* 5(2): 121-143

Litwin W, Boudenant J, Esculier C, Ferrier A, Glorieux A, La Chimia, J, Kabbaj K, Moulinoux C, Rolin P, Stangret C 1982 SIRIUS: systems for distributed data management. In Schneider H-J (ed) *Distributed Data Bases*. North-Holland, Netherlands: 311–66

Manola F 1998 *Towards a Web Object Model*. Object Services and Consulting, Inc. *http://www.objs.com/OSA/wom.htm*

Meersman R 1997 An essay on the role and evolution of data(base) semantics. In Meersman R, Mark L (eds) *Database Application Semantics*. Chapman and Hall

Meersman R, Mark L (eds) 1997 *Database Application Semantics*. Chapman and Hall

Mena E, Kashyap V, Illarramendi A, Sheth A 1998 Domain specific ontologies for semantic information brokering on the global information infrastructure. Proceedings, International Conference on Formal Ontology in Information Systems (FOIS'98), Torino: 269-283

Mendelzon A, Mihaila G, Milo T 1997a Querying the World Wide Web. *Journal of Digital Libraries* 1(1): 68–88

Mendelzon A, Mihaila G, Raschid L, Tomasic A 1997b Locating and accessing heterogeneous data sources. *Proceedings of CASCON'97*

Mudumbai S 1997 *ZEBRA: Customizable, Extensible Metadata-based Access to Federated Image Repositories*. M.S. Thesis, Department of Computer Science, University of Georgia

Online Computer Library Center, Inc 1997 *Dublin Core Metadata Element Set: Reference Description*. Office of Research and Special Projects, Dublin, Ohio. *http://www.oclc.org:5046/research/dublin_core/*

Ouksel A, Naiman C 1994 Coordinating context building in heterogeneous information systems. *Journal of Intelligent Information Systems* 3 (2): 151-183

Paepcke A, Chang C, Garcia-Molina H, Winograd T 1998 Interoperability for digital libraries worldwide. *Communications of the ACM* 41(4): 33-43

Papazoglou M, Schlageter G (eds) 1998 *Cooperative Information Systems: Current Trends and Directions*. Academic Press

Ram S (ed) 1991 Heterogeneous distributed database systems. Special Issue. *IEEE Computer* 24(12)

Schek H–J, Sheth A, Czjedo B (eds) 1993 *Proceedings of the RIDE-IMS'93: International Workshop on Interoperability in Multidatabase Systems*. IEEE Computer Society

Shah K, Sheth A 1998 Logical information modeling of Web-accessible heterogeneous digital assets. Proceedings of the Forum on Research and Technology Advances in Digital Libraries (ADL'98), Santa Barbara: 266-275

Shah K, Sheth A, Mudumbai S 1997 Black box approach to image feature manipulation used by visual information retrieval engines. *Second IEEE Metadata Conference*

Sheth A 1987 *Heterogeneous Distributed Databases: Issues in Integration*. Tutorial Notes, 3rd International Conference on Data Engineering

Sheth A (ed) 1991 Semantic issues in multidatabase systems. Special Issue. *SIGMOD Record*

Sheth A 1995 *Multidatabase Interoperation: Perspective of Researchers and Practitioners*. Tutorial Notes, 11th International Conference on Data Engineering, Taiwan

Sheth A 1996 Bellcore's ADAPT/XHarness system for managing information on Internet and intranets. *Proceedings of the 22nd International Conference on Very Large Data Bases, Bombay, India*: 585

Sheth A 1997 Data semantics: What, where, and how? In Meersman R, Mark L (eds) *Database Application Semantics*. Chapman and Hall: 601-610

Sheth A, Kashyap V 1993 So far (schematically) yet so near (semantically). In Hsiao D, Neuhold E, Sacks-Davis R (eds) 1993 *Interoperable Database Systems (IFIP Transaction A-25, Proceedings of DS-5)* North-Holland: 283-312

Sheth A, Kashyap V 1996 Media-independent correlation of information: what? how? Proceedings of First IEEE Metadata Conference

Sheth A, Klas W (eds) 1998 *Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media*. McGraw Hill

Sheth A, Larson 1990 Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys* 22(3): 183–236

Shklar L, Sheth A, Kashyap V, Shah K 1995 InfoHarness: Use of automatically generated metadata for search and retrieval of heterogeneous information. Proceedings of CAiSE-95: 217-230

Shoham Y 1991 Varieties of context. In Lifschitz V (ed) *Artificial Intelligence and Mathematical Theory of Computation: Papers in Honor of John McCarthy*. Academic Press, Boston: 393–408

Wiederhold G 1992 Mediators in the architecture of future information systems. *IEEE Computer* 25(3): 38–49

Wiederhold G (ed) 1996 *Intelligent Integration of Information*. Kluwer Academic Publishers

Wiederhold G 1997 Value-added mediation in large-scale information systems. In Meersman R, Mark L (eds) *Database Application Semantics*. Chapman and Hall