Logical Information Modeling of Web-accessible Heterogeneous Digital Assets

Kshitij Shah and Amit Sheth
Large Scale Distributed Information Systems Lab
415 GSRC, University of Georgia, Athens GA 30602-7404 USA
<kjshah, amit>@cs.uga.edu, http://lsdis.cs.uga.edu

Abstract

This paper introduces the MREF framework for representing and correlating information at a higher semantic level than is possible with Web-based information systems today. The role that metadata plays in this framework is described, together with a metadata based infrastructure to support our media independent information correlation paradigm. To keep it consistent with evolving standards, broader acceptance and ease of implementation, MREF abstraction is structured on top of RDF and XML. Its central role in the context of the InfoQuilt system, for exploiting heterogeneous digital media using a federated and scalable architecture, is briefly described.

1. Introduction

Exploiting increasing amount of diverse Web-accessible data is recognized as an important problem and an interesting challenge. A federated approach being pursued in the InfoQuilt project involves distributed and independently managed (a) InfoQuilt servers (where information requests are processed) and (b) metadata repositories. Here we wish to share the important role metadata can play. Through the Web-based infrastructure, data (including organized information repositories of various media) and users can of course be locally or globally distributed. A key feature of InfoQuilt is to support media independent information requests for search and retrieval without restructuring, relocating, or reformatting federated data sources.

In this paper, we focus on two key objectives: *representation* of information and information requests, and semantic–level information *correlation* of heterogeneous digital assets. Information requests are queries over heterogeneous media assets represented at a higher level of abstractions, possibly in media-independent manner and supporting ontologies or domain-specific terminology. Key component of our approach involves representing and correlating information artifacts and information requests at a logical level, and the corresponding strategies for processing MREFs.

The current Web infrastructure allows for artifacts (URLs) to be linked via HREFs (Hyperlink References) which are physical (hard) relationships between Web artifacts. Metadata REFerence links (MREFs), on the other hand, allow logical relationships between Web artifacts. MREFs could be used anywhere that HREFs could be used. The MREF is a representation of an information request and would be processed when the

page that embeds the MREF is viewed. Simplistically, MREFs can be viewed as an extension of the view concept in relational databases, extended to the Web infrastructure for providing access to heterogeneous or multi-media repositories.

In the above context, data modeling and related knowledge processing issues of specific interest are:

- need for comprehensive representation and modeling of metadata,
- potential use of knowledge representation for correlating metadata (of heterogeneous media), and
- use of ontologies (where knowledge representation has already played important role) to deal with terminological differences between terms in the information requests and those in metadata and data

In this paper we will focus on the first two issues[1]. Our contribution in this paper is to lay the groundwork for a framework for correlating and representing information at a higher semantic level. We propose to build this on top of the emerging RDF [RDF] and XML [XML] standards. The MREF framework thus provides a higher level of abstraction over these evolving standards. In this paper we discuss the need for this abstraction layer and provide arguments for this higher level of information representation and correlation. We then provide a structure for doing this illustrated with some examples.

In Section 2 we will discuss the background issues needed for better logical descriptions of user searches. Here we will give a brief overview of metadata and introduce the MREF concept. A high level view of the InfoQuilt system will be sketched to put MREFs in perspective. In Section 3 we will provide a view of the abstraction layers in information management and we will introduce a framework for specifying these high level MREF expressions. Section 4 provides some examples of MREFs and, finally, related work is discussed in Section 5.

2. Background

In this section we will provide some background that will lay the foundation for the ideas discussed in the subsequent sections. In section 2.1 we will discuss a metadata classification and then discuss the basic concepts behind MREF construct in section 2.2. In section 2.3 we will give a very high level overview of the InfoQuilt architecture which will show how MREFs play a central role in this system.

2.1 What is metadata

Metadata represent information about the data in individual databases and data repositories. They may represent relationships between individual media objects. These metadata descriptions may be extracted using various mappings/extractors (e.g., see,

[1] Our treatment of the third issue appears in [MKIS96, MKSI96]. Related work can also be found in projects such as SIMS and InfoSleuth.

[SSK95, KSS95]) associated with the various types of digital data. In this paper, we consider the following types of metadata (see [KSS95] and [B98] for two classifications, [BKS98] for a review of research and standards on metadata of digital media):

- Content-independent metadata: This type of metadata is independent of the content of the artifact or document[2] it describes, e.g. location, date-of-creation etc.
- Content-dependent metadata: This type of metadata captures the information content of the document. We define three types of content-dependent metadata.
  - Content-dependent metadata: This type of metadata depends directly on the document content, e.g. keywords appearing in a document, colors appearing in an image document. One method of representing content-based metadata is using a collection of attribute-value pairs. A discussion of attribute-based access for textual data is discussed in [SKL95]. The attributes chosen may be media specific (e.g. color) or media independent (e.g. location, relief).
  - Content-descriptive metadata: This is a special case of Content-dependent metadata where the content of a document is described in a manner which may not be directly based on the contents of the document. Examples of content-descriptive metadata for images may be found in [OS95, KKH94] where textual annotations are associated with images and are used to correlate information across image and textual documents.
  - Domain-specific metadata: This is a special case of content-descriptive metadata typically represented in an attribute-based manner where the attributes used to characterize documents are domain-specific in nature, e.g. relief for the Geographical Information Systems domain. In some cases the metadata may be obtained from domain-specific ontologies and may be represented using various knowledge representation and information modeling alternatives.

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

2.2 MREF

MREFs are views defined using metadata of various types and of various media. As with HREFs, an end user may only see a link on a (possibly dynamically created) Web-page. However, MREFs can represent information requests or views involving keyword-based, attribute-based and content-based specifications involving various types of metadata.

They are treated as virtual objects in the InfoQuilt system. In relational databases a view is an abstract model that does not exist as a static object in the system. A SQL query is one way of representing a relational view. Other representations can be constructed for the same abstract view object. Usually a view itself is materialized when the query, or some other representation, is processed by the system. Alternatively, it is possible to have

[2] The terms information artifact and document are used interchangeably. For us they represent any Web-accessible objects.

materialized views that hold data prior to the submission of a query. As described in this paper, the MREF abstract metadata based view is analogous to a relational view. The MREF objects are treated as virtual objects that can be referenced from Web objects. As with traditional views, they can be materialized in the system at run time or can be precomputed. Furthermore, parts of MREFs can be precomputed and others materialized at run-time.

Two simple examples of the MREFs are given at [SK96]. A third, more complex example is used later in this paper.

2.3 InfoQuilt Architecture

The InfoQuilt system has its roots in the InfoHarness [SSK95] system, which was commercialized as the Adapt/X Harness system at Bellcore. The InfoHarness system provided the proof-of-concepts for the various building blocks that form the core of the InfoQuilt system. InfoHarness addressed the system level issues of metadata extraction and management. However, InfoHarness was not intrinsically distributed. InfoQuilt elevates the ideas and goals of the InfoHarness system to a higher level of abstraction. It focuses on the logical representation and correlation of encapsulated information artifacts and is fully distributed from the ground up. MREFs play a central role in the InfoQuilt system. A high level view of the InfoQuilt system (see Figure 1 for its architecture) is useful for sketching a complete picture of how MREFs (discussed in detail in the next section) provide the glue for metadata enabled information management and resource discovery. The functionality of the various subsystems is discussed next.

Encapsulator Agents: These mobile agents are responsible for determining the type of the underlying information artifacts to be encapsulated, and processing the artifacts themselves to extract content-dependent and content-independent metadata. This extracted metadata is modeled as a RDF object and is handed over to the metadata store (metabase).

Metabase: This is a persistent RDF object store. The metabase provides functionality to process keyword, attributed, and content-based queries. The metabase also provides support for structural modeling of the metadata repository to aid in user browsing, visualization, etc.

Metadata Directories: These sites manage information used by various components of InfoQuilt for dereferencing MREFs based on their metadata components and maps this information to specific metabases. The metabases themselves register the metadata that they serve with these directories. MREF representations are stored here. The user agents dereference MREFs embedded in Web objects to MREF representations stored in these directories. These are merely representations (as described in the next section); their instantiation is done at run time. Based on various temporal factors and the state of the underlying information artifacts themselves, the MREFs could have different

instantiations at different points in time. The InfoQuilt MREF namespace management is also done here. The broker agents use these directories as described next.

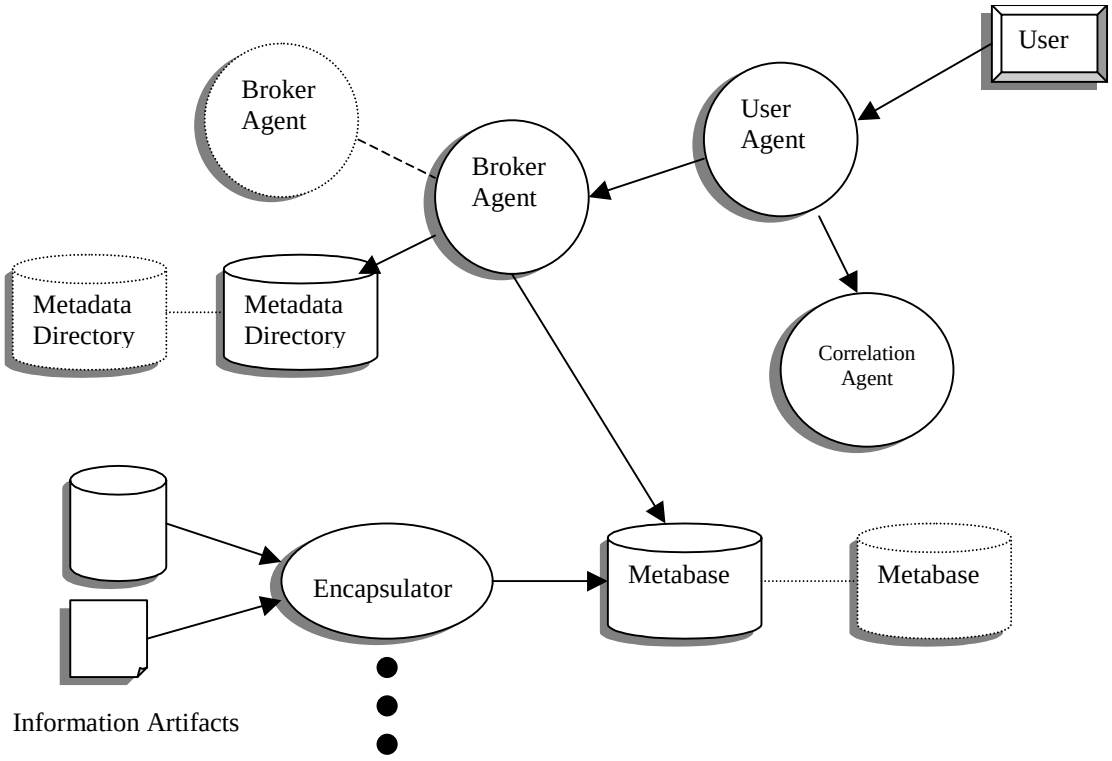

Figure 1: High Level InfoQuilt Architecture

Broker Agents: These agents are responsible for decomposing the MREFs into partial queries. The broker agents interact with the metabases for this purpose. The broker agents also decide which metabase(s) to contact for a given MREF component based on the user profiles, traffic, etc. These brokers are also responsible for merging the results that come back from the various metabases. Further correlation can be done by interacting with the correlation agents.

Correlation Agents: These agents are ontology managers and correlate MREFs based on the respective ontologies.

User Agents: MREF construction, interpretation, and translation are done here. As described above, MREFs can be embedded in Web objects (e.g., HTML files) or can exist as standalone artifacts themselves. The user agents are responsible for constructing user queries and for initial MREF processing.

3. Logical Description and Expression of Web Accessible Heterogeneous Media Information

This section discusses the various abstraction layers for information artifacts. The use of metadata for logical correlation is discussed in the MREF layer.

3.1 Abstraction layers in information management

Figure 2 shows four different layers of representation. The bottom layer is the data layer. This data could be of:

- Heterogeneous type, e.g., text, image, audio etc.
- Heterogeneous representation, e.g., GIF, JPEG, PNG etc.
- Heterogeneous encoding, e.g., compressed, uuencoded etc.

Moreover the same data could have multiple instances, e.g. mirrors, multiple representations and multiple encodings. The data can be local or widely distributed, as is the case with most web repositories today. The diverse environments that end-users or information consumers utilize to view or access the data introduce one more level of abstraction.

Metadata pertaining to the underlying data can be utilized to manage the different abstraction layers. This corresponds to the second layer in the Figure 2. Metadata can be viewed as an insulator in this case, insulating the heterogeneity of the underlying data format and distribution from the information consumer.

Figure 2 also distinguishes between the various components in the metadata layer. Indices on the data are one form of metadata. Indices usually contain content-specific metadata. The primary aim of these indices is to provide an efficient search capability over the underlying data. Full text (keyword) indices are used to search large text based repositories. Many Web search sites maintain such large keyword indices for the HTML or text component of the Web. Indices can also be built on other data types, e.g., image and video, although, technologies for creating and managing these are not as advanced as those for keyword based approaches. Creation of data indices is usually performed as a pre-processing step.

The second component of the metadata layer is the attribute metadata store. This metadata is either content-specific or content-descriptive. Information about the data artifact's location, type, representation, encoding, etc., can be viewed as attribute metadata. In its simplest form these could be tag-value pairs, whereas, more complex attribute metadata, e.g., nested or hierarchical attributes, could be applicable in some cases. This metadata can be used to provide structured searches over the underlying data artifacts. This metadata, like indices, is also pre-computed and stored in the metadata repository.

In some scenarios it may not be possible to pre-compute all the possible metadata that would be used by the information consumer as part of the information request. Images, to take a simple example, are feature rich. When dealing with geographic images, for example, we could extract a wide range of features (metadata) ranging from domain-

independent features like color, texture, structure, etc., to domain-specific features like land cover, elevation, etc. Domain specific features would also be dependent on the context of the information request, e.g., if the information request were related to a geographic survey then the features of interest would be vegetation, population, etc. All the possible features could not be pre-determined in such a case, as the same image would be used in a variety of domains across a wide range of information requests. Parameterized routines would be used in such cases to compute the features at runtime. These can be viewed as dynamic metadata. Various optimization/efficiency issues need to be addressed in this case.

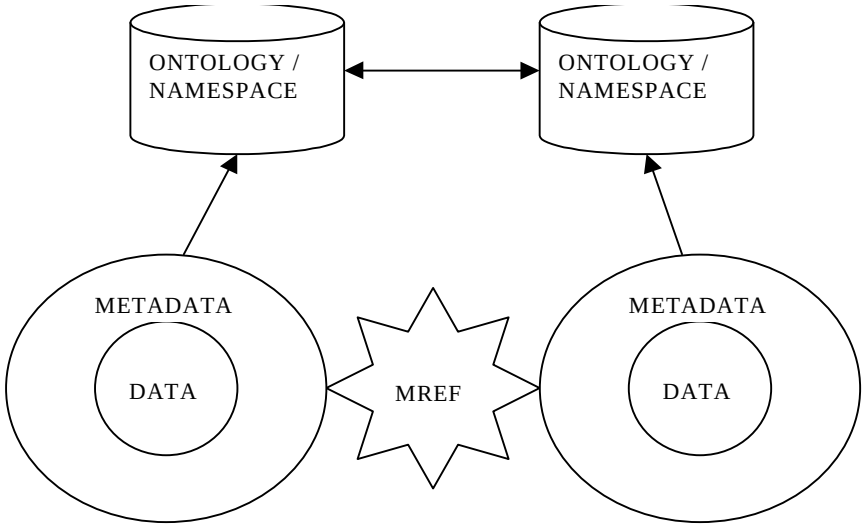

Figure 2: Abstraction Layers

In Figure 2 the layer above the metadata layer is the MREF (Metadata Reference) layer. The current Web infrastructure allows for artifacts (URLs) to be linked via HREFs (Hypertext References) which are physical (hard) relationships between Web artifacts. MREFs, on the other hand, allow logical relationships between Web artifacts. MREFs could be used anywhere that HREFs could be used. The MREF is a representation of an information request and would be processed when the page that embeds the MREF is viewed.

The highest layer in this structure is the ontology. The information request could contain terms, which have different terminological contexts in different domains, e.g., “cricket” could convey very different meanings to a zoologist and a sports enthusiast. The applications would also need to use multiple ontologies and maintain mappings between terms/concepts across these ontologies. This issue is not discussed in this paper further.

3.2 Building logical, semantic webs

The Web as it exists today is a graph of information artifacts and resources. The graph nodes are represented by embedded HREFs. These enable the implicit linking of related (or unrelated) web artifacts. This web is very suitable for browsing but provides little or

no direct help for searching. Web crawlers and search engines try to impose some sort of an order by building indices on top of the web artifacts, which are primarily textual. These efforts face an ever-increasing problem of scalability resulting in lower precision and incomplete coverage. However, in this scenario we can trivially say that a keyword query imposes a correlation (logical relationship) at a very basic (limited) level between the artifacts that make up the result set for that query.

Metadata is the key to this correlation. For a keyword query we can conceptually view the keyword index as content-dependent metadata and the keywords in the query as specific resource descriptors for the index, the evaluation of which would result in a set of correlated resources. To be more general, we need a framework for expressing *metadata based, media independent correlation* across federated digital media.

How much of the correlation is done automatically by the query processing system? The level of automation usually depends (inversely) on the information content captured in the metadata. How meaningful is the correlation? This, on the other hand, depends (directly) on the information content captured. For query processing systems to adequately address these design considerations, it is desirable to move towards location-independent, media-independent, and content-dependent methods of correlation specific to the domain of information [SK96].

One approach to represent MREFs is to use the Resource Description Framework (RDF) [RDF] as the underlying framework. MREFs are not bound to RDF in the sense that other frameworks can be used with the processing required for that framework dependent on the application. In this paper we describe MREF processing in the InfoQuilt [KSS95] system.

RDF emphasizes facilities to enable automated processing of Web resources. RDF enables serializable representations of a directed graph, which is basically the 3-tuple data model for representing named properties and their values. These properties serve both to represent attributes of resources (and in this sense correspond to usual attribute-value-pairs) and to represent relationships between resources. One way of expressing RDF statements is using the eXtensible Markup Language (XML) [XML]. Just as HTML is the standard for Web publishing, XML is poised to become the standard for Web-based applications.

MREFs are specified as RDF statements. Figure 3 gives an idea of how XML, RDF, and MREFs are related. RDF is mainly aimed at describing and exchanging meta-information about a resource (artifact) or a set or resources and their relationships. We consider MREFs to be resources that logically correlate other resources based on their metadata. Specifically, MREFs are RDF assertions with the MREF constraints modeled as RDF properties. The intrinsic flexibility of RDF enables the constraints to be embedded within the MREF or to be external (e.g., constructed by a user interface based on an MREF template). Higher level mappings can also be done by associating the MREF resource with a schema URI using the XML namespace mechanism. This mapping associates the

MREF properties (constraints) to MREF namespaces. These schemas can then be mapped to ontologies that are known to InfoQuilt.

Let us consider a simple example here and defer the more extensive examples to a later section. Let us consider a query that has a list of free form keywords, "winter rose" and a list of attribute value pairs, "color –> red", "fragrance –> slight".

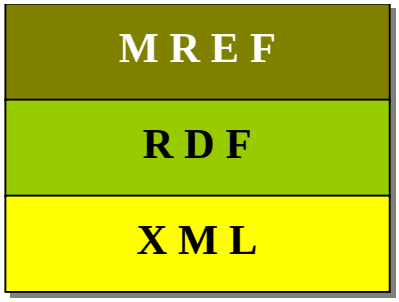

Figure 3: XML, RDF, and MREF

```
<?namespace href="http://www.foo.com/IQ" as="IQ"?>
<?namespace href="http://www.w3.org/schemas/rdf-schema" as="RDF"?>
<RDF:serialization>
      <RDF:bag id="MREF:12345>
            <IQ:keyword>
                  <RDF:resource id="constraint_001">
                     <IQ:threshold>0.5</IQ:threshold>
                     <RDF:PropValue>winter rose</RDF:PropValue>
                  </RDF:resource>
            </IQ:keyword>
            <IQ:attribute>
                  <RDF:resource id="constraint_002">
                        <IQ:name>color</IQ:color>
                        <IQ:type>string</IQ:type>
                        <RDF:PropValue>red</RDF:PropValue>
                  </RDF:resource>
            </IQ:attribute>
            <IQ:attribute>
                  <RDF:resource id="constraint_003">
                        <IQ:name>fragrance</IQ:color>
                        <IQ:type>string</IQ:type>
                        <RDF:PropValue>slight</RDF:PropValue>
                  </RDF:resource>
            </IQ:attribute>
      </RDF:bag>
</RDF:serialization>
```

This query can be represented in the RDF model as shown in Figure 4.

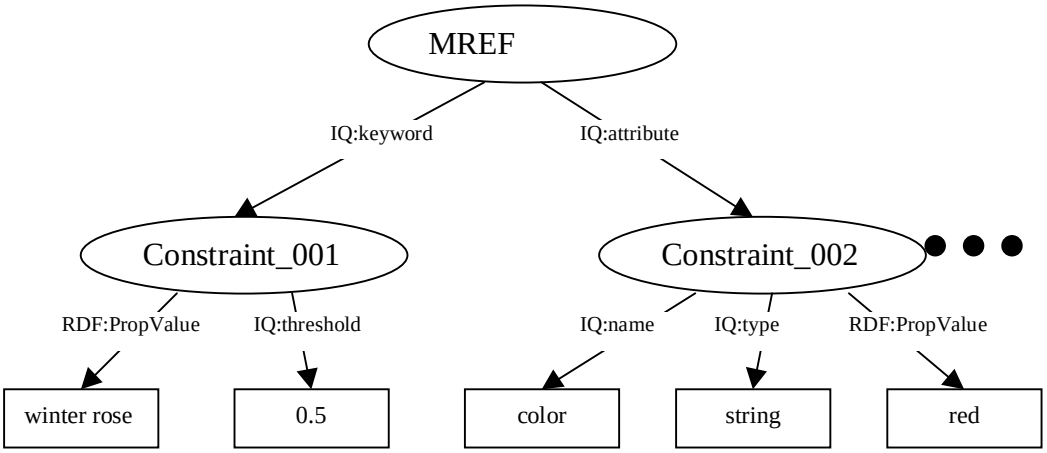

Figure 4: RDF Model for MREF

In this model MREFs are treated as bags of constraints. We treat the MREF itself as a virtual resource. Thus, MREFs can be embedded in Web pages or can exist as separate entities. The processing and instantiation of MREFs is described in detail in the following sections.

4. Example

In this section we illustrate the MREF concept with an example of media-independent domain-specific correlation. We use HTML 4.0 [HTML4] for these examples. Let us first consider a scenario of a site location and planning application supported by a Geographic Information System [SK96] and a correlation query illustrated in the following example:

```
<HEAD>
<OBJECT declare
        id="mall-loc"
        type="application/x-mref"
        data="
<?namespace href="http://www.foo.com/IQ" as="IQ"?>
<?namespace href="http://www.w3.org/schemas/rdf-schema" as="RDF"?>
<RDF:serialization>
      <RDF:bag id="MREF:mall-loc”>
            <IQ:attribute>
                  <RDF:resource id="constraint_001">
                        <IQ:name>population</IQ:color>
                        <IQ:type>number</IQ:type>
                        <IQ:operator>greater</IQ:operator>
                        <RDF:PropValue>500</RDF:PropValue>
                  </RDF:resource>
            </IQ:attribute>
            <IQ:attribute>
                  <RDF:resource id="constraint_002">
```

```
                        <IQ:name>area</IQ:color>
                        <IQ:type>number</IQ:type>
                        <IQ:operator>greater</IQ:operator>
                        <IQ:units>acres</IQ:units>
                        <RDF:PropValue>50</RDF:PropValue>
                  </RDF:resource>
            </IQ:attribute>
            <IQ:attribute>
                  <RDF:resource id="constraint_003">
                        <IQ:name>region-type<pe/IQ:color>
                        <IQ:type>string</IQ:type>
                        <RDF:PropValue>block</RDF:PropValue>
                  </RDF:resource>
            </IQ:attribute>
            <IQ:attribute>
                  <RDF:resource id="constraint_004">
                        <IQ:name>land-cover<pe/IQ:color>
                        <IQ:type>string</IQ:type>
                        <RDF:PropValue>moderate</RDF:PropValue>
                  </RDF:resource>
            </IQ:attribute>
            <IQ:attribute>
                  <RDF:resource id="constraint_005">
                        <IQ:name>relief<pe/IQ:color>
                        <IQ:type>string</IQ:type>
                        <RDF:PropValue>moderate</RDF:PropValue>
                  </RDF:resource>
            </IQ:attribute>
      </RDF:bag>
</RDF:serialization>
">
</OBJECT>
</HEAD>

<BODY>
To identify potential locations for a future shopping mall, all regions
having a population greater than 500 and area greater than 50 acres
having an urban land cover and moderate relief
<OBJECT classid=”http://foo.bar.com/iq.mref”
          standby=”Loading MREF...”
            data=”#mall-loc”>
can be viewed here.
</OBJECT>

</BODY>
```

In the above example the MREF is located within the same document as the reference. An alternative would be to have the MREF as a separate object and referenced from within this document as shown below:

```
<BODY>
To identify potential locations for a future shopping mall, all regions
having a population greater than 500 and area greater than 50 sq feet
having an urban land cover and moderate relief
<OBJECT classid=”http://foo.bar.com/iq.mref”
          standby=”Loading MREF...”
```

```
            data="./mrefs/mall-loc.mrf">
can be viewed here.
</OBJECT>

</BODY>
```

The MREF object could be on any arbitrary server (the *data* element could take any URI as a value). The MREF can also be supplied by a metadata directory or constructed dynamically by a user query interface. In the information brokering paradigm, any information supplier or information broker can specify or manage a MREF object or publish a Web-page with embedded MREFs [S97].

The processing of the above MREF results in the structured information (area, population) and the map of the regions satisfying the above constraints being included in the HTML document. The level of indirection that the MREF provides hides the heterogeneity of the media in which the underlying data might be stored. Data satisfying the specified constraints can be found, for example, in independent repositories containing census data and TIGER/Line database as shown in Figure 5. Mapping the domain specific attributes– relief, land-cover, area and population– to the underlying data requires access to processing over additional image and structured databases.

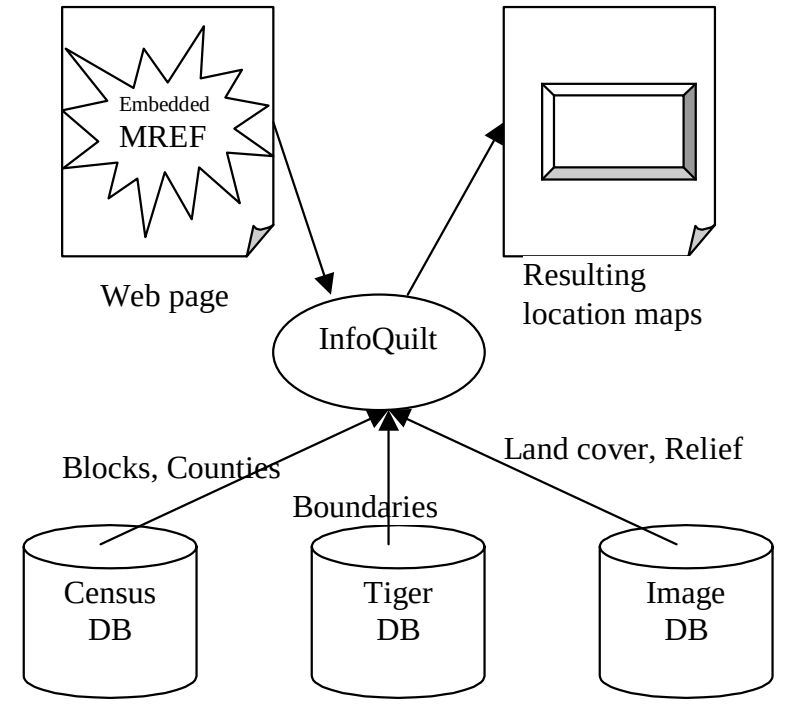

Figure 5: MREF correlation over heterogeneous sources

One of the key things achieved because of media-independent correlation is that the end-user (or the decision-maker exploiting information) is completely insulated from the

underlying heterogeneity. Information about the attributes population and area are obtained from a structured database storing Census Data, whereas land-cover and relief information is obtained from the image database storing USGS maps. Boundary information, which is necessary for correlation, is obtained from the underlying TIGER/Line Database. All these resources are encapsulated as RDF objects in some metabase. The broker agent maps the constraint attributes to the appropriate metabase. These InfoQuilt RDF objects might have the necessary metadata to satisfy the constraints or might have associated dynamic methods (also referred to as parameterized routines) to compute the necessary metadata. The parameters themselves may be supplied within the MREF or might be supplied as runtime parameters.

5. Related Work

WebSQL [MMM97] 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) [MMRT97]. Compared to our MREF-based framework, this system has different layers of abstraction and provides mechanisms for describing the data that is 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. The goals of this project are very similar to ours but we treat the semantic views as resources themselves.

There are several systems that employ the metadata based semantic view of the world and employ an ontological layer above it. The OBSERVER [MKSI96, MKIS96] system uses domain specific ontologies to specify ontological commitments or agreements between users and information providers. This is achieved by mapping real world concepts to data structures in the underlying repositories and providing translations of information requests across ontologies.

Several systems that support access of heterogeneous and distributed information sources include SIMS, InfoHarness, TSIMMIS, Information Manifold, HERMES and InfoSleuth. A subset of these which also support logical level of information modeling are briefly described next, with respect to their support for logical description of data and information correlation if any.

In the SIMS project [AKS96], 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.

TSIMMIS [CGH+94, GHI+95] uses a mediator approach to combine information from several sources containing textual and semi-structure 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 are responsible for routing queries to appropriate sources and for post-processing the results. An important focus of the system is to automatically generate wrappers and mediators 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.

Another system that is based on the mediator approach is the HERMES [AS94] system 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 type. 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 Information Manifold [LSK95] 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. The focus in the Information Manifold project however is 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 InfoSleuth [BBB+97] system 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.

There are also a number of systems that focus on specific media types, e.g., QBISM and CoBase, just to name a couple for image data management. We do not discuss these for brevity.

Key distinctions the InfoQuilt system intends to offer include a combination of the following:

- integral support for not only structured and semi-structured (including Web pages) data, but also visual media,
- support for very broad range of metadata and distributed management,
- logical correlation (the focus of this paper) of heterogeneous digital media information with associated mapping to Web-relevant evolving standards, with support for their management in an information brokering architecture, and
- support for media-independent information requests and their distributed information processing.

Conclusions

Web-related technologies have provided an infrastructure to share a broad variety of artifacts and information resources. These include structured data (as in Web-accessible databases), textual and semi-structured data, and increasingly visual data. A federated approach to exploiting these requires a powerful logical or semantic level mechanism for representing and correlating relevant information regardless of their media, powerful ways to express media-independent information requests (including keyword-based, attribute-based and content-based), and the corresponding ways to process them. This paper discussed MREFs, the central concept in our InfoQuilt system that aims to support the above-described federated approach. In particular, we focused on its RDF and XML based representation that can support ease in their communication/sharing and processing.

References

[AS94] S. Adali, V.S. Subrahmanian, Amalgamating Knowledge Bases, III: Distributed Mediators, International Journal of Intelligent Cooperative Information Systems, 1994.

[AKS96] Y. Arens, C.A. Knoblok, and W. Shen. Query Reformulation for Dynamic Information Integration, Journal of Intelligent Information Systems, 6 (2-3): 99-130, 1996.

[BBB+97] R. Bayardo, W. Bohrer, R. Brice, A. Cichocki, G. Fowler, A. Helal, V. Kashyap, T. Ksiezyk, G. Martin, M. Nodine, M. Rashid, M. Rusinkiewicz, R. Shea, C. Unnikrishnan, A. Unruh, D. Woelk, "Semantic Integration of Information in Open and Dynamic Environments". Proceedings of the 1997 ACM International Conference on the Management of Data (SIGMOD), Tucson, Arizona, May 1997.

[B98] K. Boehm, Metadata Handling in HyperStorM, Chapter 2 in [SK98].

[BKS98] S. Boll, W. Klas, and A. Sheth, Using Metadata to Manage Multimedia Data, Chapter 1 in [SK98].

[BS95] R. Brachman and J. Schmolze, "An Overview of the KL-ONE Knowledge Representation System," Cognitive Science, 9: 171-216, 1985.

[CGH+94] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. "The TSIMMIS Project: Integration of Heterogeneous Information Sources". In Proceedings of IPSJ Conference, pp. 7-18, Tokyo, Japan, October 1994.

[GHI+95] H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. "Integrating and Accessing Heterogeneous Information Sources in TSIMMIS". In Proceedings of the AAAI Symposium on Information Gathering, pp. 61-64, Stanford, California, March 1995.

[HTML4] HTML 4.0 Specification, http://www.w3.org/TR/REC-html40/.

[KSS95] V. Kashyap, K. Shah, and A. Sheth, Metadata for building the MultiMedia Patch Quilt, in S. Jajodia and V. Subrahmanian, eds, Multimedia Database Systems: Issues and Research Directions, Springer Verlag, 1995.

[KKH98] Y. Kiyoki, T. Kitagawa, and T. Hayama. "A Metadatabase System for Semantic Image Search by a Mathematical Model of Meaning." Chapter 6 in Managing Multimedia Data: Using Metadata to Integrate and Apply Digital Media, A. Sheth and W. Klas, eds. McGraw-Hill, 1998.

[LSK95] A.Y. Levy, D. Srivastava, and T. Kirk. Data Model and Query Evaluation in Global Information Systems. Intelligent Information Systems, 5 (2), September 1995.

[MKIS96] E. Mena, V. Kashyap, A. Illarramendi and A. Sheth, "Managing Multiple Information Sources through Ontologies: Relationship between Vocabulary Heterogeneity and Loss of Information," Proceedings of the workshop on Knowledge

Representation meets Databases (in conjunction with European Conference on Artificial Intelligence), Budapest, Hungary, August 1996.

[MKSI96] E. Mena, V. Kashyap, A. Sheth and A. Illarramendi, "OBSERVER: An approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies," Proceedings of the 1st IFCIS International Conference on Cooperative Information Systems (CoopIS '96), Brussels, Belgium, June 1996.

[MMM97] A. Mendelzon, G. Mihaila, T. Milo. Querying the World Wide Web, Journal of Digital Libraries 1 (1): 68-88, 1997.

[MMRT97] Alberto Mendelzon, George Mihaila, Louiqa Raschid, Anthony Tomasic, "Locating and Accessing Heterogeneous Data Sources," In Proceedings of CASCON'97, November 1997.

[OS95] V. Ogle and M. Stonebraker. "Chabot: Retrieval from a Relational Database of Images". IEEE Computer, special issue on Content-Based Image Retrieval Systems, V. Gudivada and V. Raghavan. eds, 28 (9), 1995.

[RDF] http://www.w3.org/Metadata/RDF/

[SKL95] A. Sheth, V. Kashyap and W. LeBlanc. "Attribute-based Access of Heterogeneous Digital Data," Proceedings of the Workshop on Web Access to Legacy Data, 4th International WWW Conference, Boston MA, December 1995

[SK96] A. Sheth and V. Kashyap, "Media-independent Correlation of Information: What? How?" Proceedings of First IEEE Metadata Conference, April 1996.

[SK98] A. Sheth and W. Klas, Eds. Managing Multimedia Data: Using Metadata to Integrate and Apply Digital Media, McGraw-Hill, 1998.

[S97] A. Sheth, "Semantic Interoperability in Infocosm: Moving Beyond Infrastructural and Data Interoperability in Federated Information Systems", Interop'97, Santa Barbara, December 3-5, 1997. [Also accessible at http://lsdis.cs.uga.edu/publications/]

[SSKS95] L. Shklar, A. Sheth, V. Kashyap, and K. Shah. Infoharness: Use of Automatically Generated Metadata for Search and Retrieval of Heterogeneous Information Proceedings of CAiSE '95, June 1995.

[XML] http://www.w3c.org/XML/