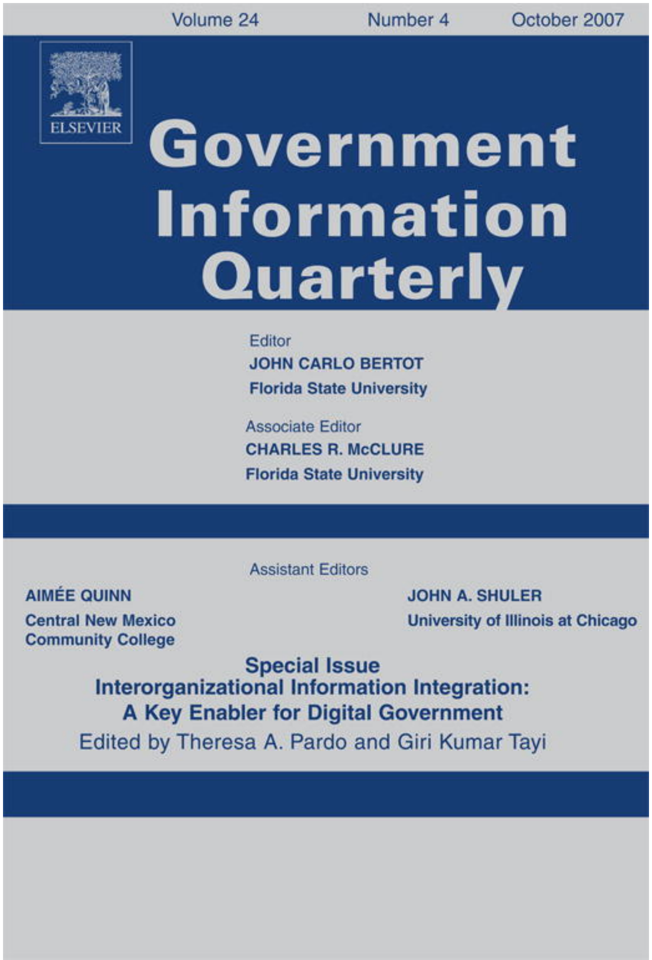


Volume 24 Number 4 October 2007

ELSEVIER

Government Information Quarterly

Editor
JOHN CARLO BERTOT
Florida State University

Associate Editor
CHARLES R. McCLURE
Florida State University

Assistant Editors
AIMÉE QUINN
Central New Mexico
Community College

JOHN A. SHULER
University of Illinois at Chicago

Special Issue
Interorganizational Information Integration:
A Key Enabler for Digital Government
Edited by Theresa A. Pardo and Giri Kumar Tayi

Available online at www.sciencedirect.com

ScienceDirect

Government Information Quarterly 24 (2007) 716–735

Government Information Quarterly

Semantic integration of government data for water quality management

Zhiyuan Chen [a], Arrya Gangopadhyay [a], Stephen H. Holden [b,*], George Karabatis [a], Michael P. McGuire [c]

[a] *ITE 431, Information Systems, University of Maryland, Baltimore County (UMBC), 1000 Hilltop Circle, Baltimore, MD 21250, USA*

[b] *SRA Touchstone Consulting Group, 1920 N Street, NW, Suite 600, Washington D.C., 20038, USA*

[c] *Center for Urban Environmental Research and Education, University of Maryland, Baltimore County (UMBC), 1000 Hilltop Circle, TRC 102, Baltimore, MD 21250, USA*

Available online 16 July 2007

Abstract

Normative models of e-government typically assert that horizontal (i.e., inter-agency) and vertical (i.e., inter-governmental) integration of data flows and business processes represent the most sophisticated form of e-government, delivering the greatest payoff for both governments and users. This paper concentrates on the integration of data supporting water quality management as an example of how such integration can enable higher levels of e-government. It describes a prototype system that allows users to integrate water monitoring data across many federal, state, and local government organizations and provides novel techniques for information discovery, thus improving information quality and availability for decision making. Specifically, this paper outlines techniques to integrate numerous water quality monitoring data sources, to resolve data disparities, and to retrieve data using semantic relationships among data sources taking advantage of customized user profiles. Preliminary user feedback indicates that these techniques enhance quantity and quality of information available for water quality management.

* Corresponding author. Fax: +1 202 338 6106.
E-mail addresses: steve_holden@sra.com (S.H. Holden), georgek@umbc.edu (G. Karabatis).

0740-624X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.giq.2007.04.004

1. Introduction

Normative models of e-government typically assert that horizontal (i.e., inter-agency) and vertical (i.e., inter-governmental) integration of data flows and business processes represent the most sophisticated form of e-government, delivering the greatest payoff for both governments and users. With this sophistication, though, comes great complexity of design and implementation that spans the domains of information systems, information policy and public administration. This paper concentrates on the integration of data supporting water quality management as an example of how it is possible to address this complexity in setting and implementing water quality policy.

Data integration problems arise from the implementation of the Federal Clean Water Act (CWA, under Sections 303(d) and 305(b)), which requires states, territories, and authorized tribes to report on the water quality status of jurisdictional waters every 2 years. These CWA mandates create unique data needs and problems, such as how to interpret information derived from multiple sources, of variable quality, using different formats, and collected according to different protocols and procedures. Existing tools for managing these data are not integrated nor do they provide any sort of data analysis capability to allow water resource managers to make informed decisions. The combination of both organizational and data complexity creates fundamental challenges to developing policies, based upon a robust information stream, which are responsive to a wide range of stakeholder interests. Such a problem setting represents the kind of challenge that sophisticated e-government systems are supposed to address.

We describe an interoperable system that builds upon the current digital government literature, allowing users to integrate water monitoring data across many organizations into a data warehouse for subsequent knowledge discovery. Our system makes the following contributions:

1.1. Strikes a balance between breadth and depth in data integration

Numerous organizations and individuals (e.g., volunteers) collect water quality data. Ideally we should integrate all possible data into a uniform format, but in practice this is difficult to accomplish because data sources do not agree on a universal format. We describe a hybrid approach that integrates the metadata (including information on how to access data sources, when and where the data are collected, which parameters are monitored, etc.) of all data sources but fully integrates data only from key sources, such as data from the Environmental Protection Agency (EPA) or U.S. Geological Survey (USGS). Thus, we enable water quality managers and policy makers to search the integrated metadata, to locate any source of interest, and to manually download data from that source; or access the fully integrated data as if the data were from a single source.

1.2. Enables policy makers to retrieve customized views of water quality data

Many aspects of water quality decision making (i.e., optimizing policy outcomes and minimizing public costs) depends on complete and consistent data. The decision making process, includes many diverse interests, including government agencies with different oversight and regulatory responsibilities. This leads to disparate, incomplete, and unintegrated data stores on water quality, which may result in sub-optimal decisions based on incomplete or poor quality data. This prototype recognizes the need for different stakeholder views for both data input and output in the decision making process. Establishing user profiles through the data retrieval interface allows different stakeholders and decision makers to customize data queries consistent with their interest or legal and policy responsibilities. Preliminary feedback from users indicates that such customized queries based on profiles and the other tools discussed below increases the breadth, depth, and quality of water quality data available for decision making.

1.3. Exploits semantic relationships

Identifying a water quality problem often requires analysis of data from multiple sources that are related semantically. For example, users looking for possible explanations for a recent change in fish population in a stream may need to examine all data sources semantically related to fish population, such as stream temperature, impervious surfaces, elevation, land cover, etc. It would be tedious and impractical for users to find all this information by themselves. We represent such relationships among data sources using semantic networks, which assist users in locating related sources.

1.4. Resolves disparities in resolution of spatial, temporal, and format

Water monitoring data are collected in a variety of formats, units (metric or SI), and spatial and temporal granularities. For example, one data set may measure stream flow in cubic feet per second while another data set uses cubic meters per second. Furthermore, land cover data are typically captured at a 30-meter resolution, but stream chemical and biological data are collected at a specific point along a stream segment. This leads to disparities in terms of spatial granularities. We resolve format disparities through the development of conversion tools (which convert one format into another) and we resolve the spatial and temporal disparities using spatial–temporal join/aggregation operations (which aggregate lower level location areas into higher level ones).

We describe the conceptual model of a data warehouse to store the fully integrated water quality data for advanced decision support to assist water quality managers.

The remainder of the paper is organized as follows. In Section 2, we provide information related to the government mandate for water quality management, and in Section 3, we discuss related work in e-government and data integration. Section 4 presents our methodology and technical approach to using data integration to support policy makers. Section 5 describes the usability study findings of users who experimented with the prototype system for water quality management. The final section, Section 6, summarizes our findings from this research and identifies future work.

2. Background

The Federal Clean Water Act (CWA, under Sections 303(d) and 305(b)) requires states, territories, and authorized tribes to report on the water quality status of jurisdictional waters every 2 years. In order to do this, states are required to utilize all readily available data and information, including chemical, physical, and biological data. Any waters determined to be impaired must be listed on an impaired waters list (the 303(d) List), including the cause of impairment and any known sources. The 303(d) List is then submitted to the U.S. Environmental Protection Agency (EPA) biennially for review and approval.

These CWA mandates create unique data needs and problems, such as how to interpret information derived from multiple sources, of variable quality, using different formats, and collected according to different protocols and procedures. Moreover, the increasing use of biological community data (i.e., fish and aquatic insects) by states has been useful for determining water body health but has not been as successful in identifying pollution causes and sources (see EPA guidance for developing Stressor ID tools, http://www.epa.gov/ost/biocriteria/stressors/stressorid.pdf). Government agencies and water monitoring councils have designed several tools to address these data and information needs of the water resource managers, including (1) the EPA's STORET database (http://www.epa.gov/storet), (2) the USGS National Hydrography Dataset (http://nhd.usgs.gov/), and (3) the Maryland Water Monitoring Council's (MWMC) Clickable Map (http://cuereims.umbc.edu/website/mwmc/). These tools are not integrated nor do they provide any sort of data analysis capability to allow water resource managers to make better informed decisions. In our work we expand on these currently available digital government tools as well as give government officials and the public access to powerful data mining techniques generated through user interaction and feedback.

The MWMC is an organization created in 1995 to foster cooperation among groups involved in all types of water monitoring activities in Maryland. The Council serves as a statewide collaborative body to help achieve effective collection, interpretation, and dissemination of environmental data related to issues, policies, and resource management involving water monitoring. The MWMC has representatives from State and local governments, colleges and universities, and private groups. An attempt has been made by the MWMC to integrate monitoring site metadata via the MWMC Clickable Map. This is a Web-based geographic information system (GIS) utilizing a metadata form that allows users to submit their metadata, view monitoring site locations spatially, and query a limited set of metadata. The main shortcoming of the current system is that integration is largely left up to the individual user. To gain access to multiple sources of monitoring data for a certain area, users must query the Clickable Map to first identify monitoring sites within their area of interest. Then they need to gather the contact information for each site from the metadata record and contact the person responsible for each monitoring site. For areas that may include sites from many organizations, this would be a daunting task. Furthermore, once the user is able to acquire the data, there are numerous integration problems.

3. Literature review

E-government literature typically relies on both organizational and systems integration as the vehicle for delivering many of the expected benefits of automating manual and paper process in the public sector. Traditional stove-pipe information and service delivery typically put government organizational structure before user needs, such that users had to know that the Department of State issued passports or that water quality data might be found in organizations as diverse as the EPA, USGS and the MWMC. Instead, e-government is supposed to insulate users from such government complexity through a user-friendly Web interface.

Stage or maturity models of e-government illustrate that, over time, providing users with seamless information and service delivery involves a greater degree of complexity across several dimensions of e-government. These models (Baum & Maio, 2000; Hiller & Bélanger, 2001; Layne & Lee, 2001) suggest that e-government capabilities begin modestly and initially provide static, one-way information but grow more sophisticated and add interactive and transactional capabilities. The models predict an ultimate evolution of e-government that includes horizontal and vertical integration and the development of true portals and seamlessness around the concept of "life events".

Three models of e-government maturity point this out, but in somewhat different ways. The Gartner model (2000) displays, in some detail, the policy, technology, data, and organizational issues that must be resolved for organizations to progress to higher levels of e-government maturity with an attendant increase in benefits for both government organizations and end-users. Layne and Lee (2001) highlighted the fact that achieving more mature levels of e-government requires higher levels of both technology and organizational complexity. See Fig. 1 for Layne and Lee's four-stage model of e-government. Lee, Tan, and Trimi (2005) have adapted this model to further stress the importance of seamless data integration to assess the practices of leading e-government countries. Hiller and Bélanger (2001), in particular, stress increasing levels of data integration required for true transformational e-government but warn that such data integration raises significant privacy issues when the data involve personally identifiable information. What these models imply, but only sometimes make explicit, is that the complexity of these various forms of integration have likely resulted in many organizations reaching the highest level of e-government maturity (Moon, 2002; Norris & Moon, 2005; West, 2004).

Related research has identified the challenges of data integration for e-government and related decision making but has provided little in the way of specific solutions to the problem. Much like with water quality, responsibility for policy making and regulatory enforcement of air quality standards is shared among federal agencies and across levels of government. The technical challenges in data integration to support policy making are profound as noted by Pantel, Philpot, and Hovy (2005) when they state, "[T]he resulting massive data heterogeneity makes it impossible to effectively locate, share, or compare data across sources…" (p. 43). A recent workshop examining eco-informatics and policy identified five specific information technology challenges facing environmental decision makers. Three of the challenges, data presentation problems, geographic data gaps, and tool problems (relating to the lack of data standards tools for meta data) relate directly to this paper and remain largely unaddressed in the literature (Cushing & Wilson, 2005).

There is also a rich body of existing work on information integration problems beyond e-government (Levy, Rajaraman, & Ordille, 1996; Miller et al., 2001; Papakonstantinou,

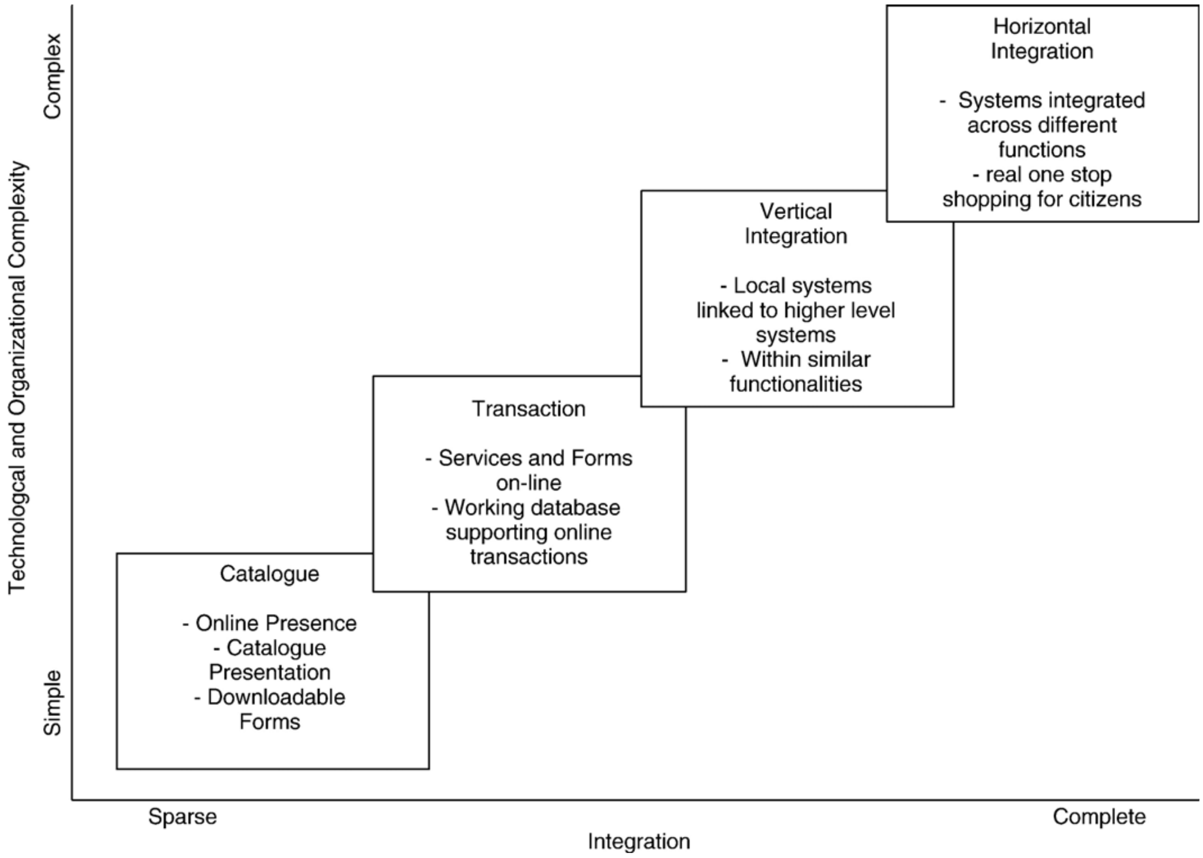

Horizontal Integration
- Systems integrated across different functions
- real one stop shopping for citizens
Vertical Integration
- Local systems linked to higher level systems
- Within similar functionalities
Transaction
- Services and Forms on-line
- Working database supporting online transactions
Catalogue
- Online Presence
- Catalogue Presentation
- Downloadable Forms
Technologcal and Organizational Complexity
Simple
Complex
Sparse
Complete
Integration

Fig. 1. Four-stage model of e-government, by Layne and Lee (2001).

Garcia-Molina, & Ullman, 1996). All such work takes a *deep integration* approach that assumes there is a global schema covering all data sources (called *mediated schema*). The mediated schema hides all schema discrepancies and provides an overall umbrella covering all data sources. The major problem of having a global mediated schema is when there are many sources (as in the case of water monitoring), it is extremely difficult to let them all agree on the mediated schema. More recently, there exists work on decentralized data sharing (Berberidis, Angelis, & Vahavas, 2004; Halevy, Ives, Suciu, & Tatarinov, 2003; Hyperion, n.d.; Tatarinov & Halevy, 2004). However, the main problem of such work is that the user often has to go through several sources to reach the desired data source.

There has been much effort by the ecology research community to integrate their data (KNB, n.d.; SEEK, n.d.; Stiglitz, Orszag, & Orszag, 2001). These systems take a *shallow integration* approach where only metadata is integrated, not original data from the data sources. These systems allow users to store metadata of data sets in a centralized database and to select data sets using a metadata keyword search or SQL-based search. Such systems avoid the problem of defining a global-mediated data schema and allow researchers to share data in an ad hoc way. A standard called Ecological Metadata Language (EML, n.d.) (knb.ecoinformatics.org/software/eml) has been developed by the ecology discipline and has been used in these projects. EML is implemented as a series of XML document types that can be used in a modular and extensible manner to document ecological data.

There are three major limitations of existing research on data integration. First, there is little empirical work that demonstrates how it is possible to use data integration to provide highly integrated e-government information products and services. Second, identifying a water quality problem often requires analysis of data from multiple sources with semantic relationships between them but there is limited support assisting users find data sources semantically related to their interest. Third, there is no system that supports both shallow and deep integration; thus, users either have to assume all data sources conform to the same mediated schema (when using deep integration) or have to integrate data by themselves (when using shallow integration).

4. Research design and methodology

In this section, we describe our design and methodology for semantic data integration of government repositories containing information on water quality. First we present our system architecture. We then identify our techniques for semantic data integration (both deep and shallow), the use of semantic networks, and our current prototype system. Then, we describe our approach with the data warehouse and we describe its multidimensional schema for water quality data analysis.

The main difference of our approach compared to existing work is that we support both deep and shallow integration approaches. Deep integration provides a universal view of data and allows users to utilize our analysis tools. This is appropriate for many users such as the public and government agencies interested in water quality. However, applying deep integration approach to all data sources is very hard to achieve in practice because of the problems we mentioned that are associated with a required mediated schema. To address these problems we support shallow integration, which is a much lighter weight approach that converts metadata to a universal format and allows users to query the metadata and locate data sources of their interest, which they can further download manually.

We designed and implemented a prototype system to address the above problems. To validate the correctness of our methodology, we gathered preliminary user feedback from environmental researchers, policy makers, and students. We asked them to use our system, and answer a set of data queries, while at the same time evaluate usability aspects of the system. More specifics of this study are described in Section 5.

4.1. Overview of the architecture

An examination of Fig. 2 provides a view of the prototype system built for and examined through this research, but it also depicts the level of disaggregation of the data prior to the prototype. In effect, the water monitoring community displayed in the upper left hand corner of the figure represents the “before” picture. Prior to the development of the prototype, the water monitoring community relied on individual data repositories maintained by volunteers, various levels of government and even members of the public. What follows is an overview of the tools previously available to the water monitoring community and how this research has built on that

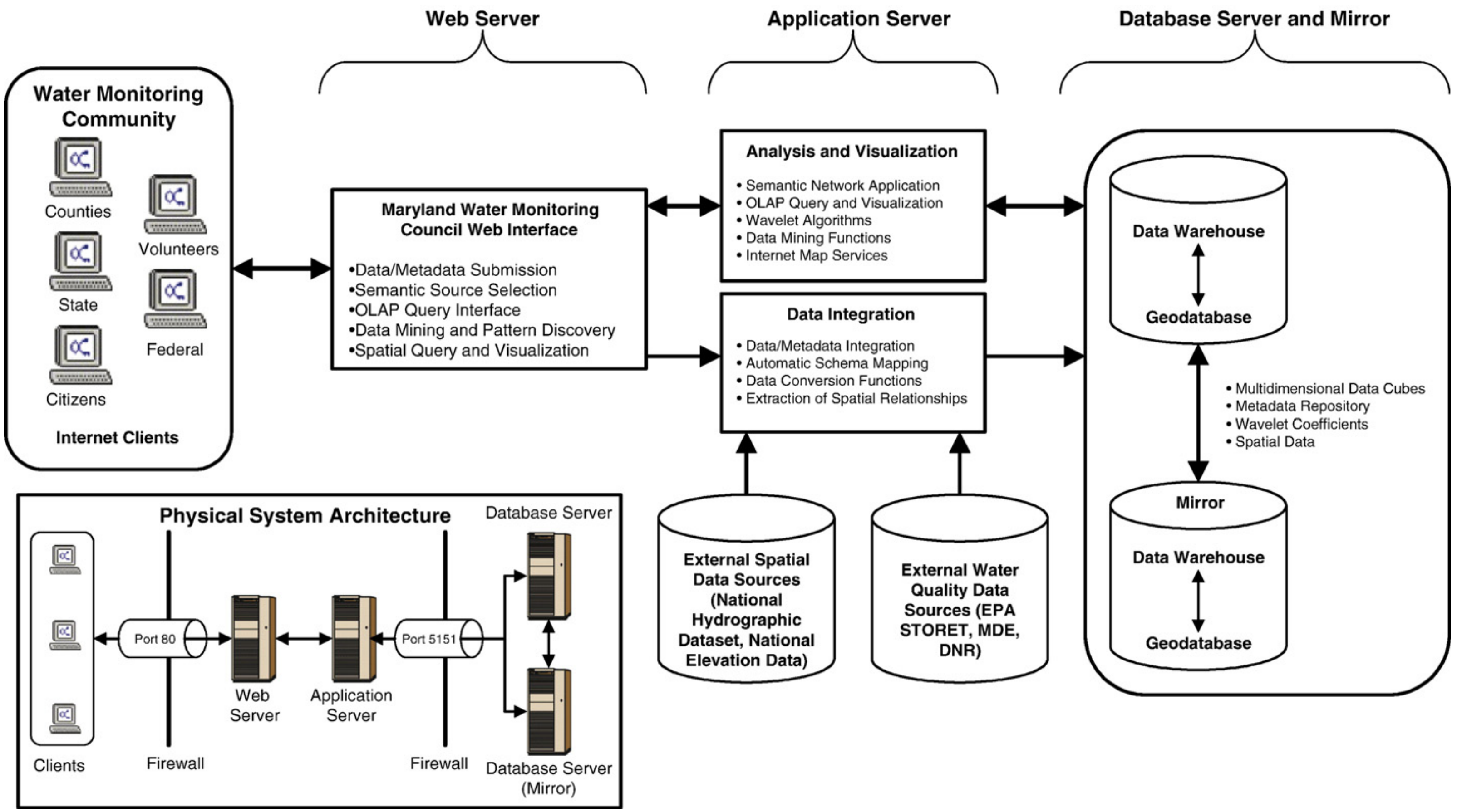

Web Server
Application Server
Database Server and Mirror
Water Monitoring Community
Counties
Volunteers
State
Federal
Citizens
Internet Clients
Maryland Water Monitoring Council Web Interface
•Data/Metadata Submission
•Semantic Source Selection
•OLAP Query Interface
•Data Mining and Pattern Discovery
•Spatial Query and Visualization
Analysis and Visualization
• Semantic Network Application
• OLAP Query and Visualization
• Wavelet Algorithms
• Data Mining Functions
• Internet Map Services
Data Integration
• Data/Metadata Integration
• Automatic Schema Mapping
• Data Conversion Functions
• Extraction of Spatial Relationships
Data Warehouse
Geodatabase
• Multidimensional Data Cubes
• Metadata Repository
• Wavelet Coefficients
• Spatial Data
Mirror
Data Warehouse
Geodatabase
External Spatial Data Sources (National Hydrographic Dataset, National Elevation Data)
External Water Quality Data Sources (EPA STORET, MDE, DNR)
Physical System Architecture
Database Server
Clients
Port 80
Firewall
Web Server
Application Server
Port 5151
Firewall
Database Server (Mirror)

Fig. 2. Conceptual system architecture.

foundation to provide an architecture for data integration that supports water quality management through e-government.

The Maryland Water Monitoring Council (MWMC) Clickable Map is currently being used as a tool to integrate water quality monitoring site metadata. It is a collaborative system that facilitates information sharing and monitoring program coordination among federal, state and local agencies, research, private, and volunteer monitoring programs. The Clickable Map is accessed through a Web-based geographic information systems (GIS) interface running on a single Web server using ESRI ArcIMS® software. Users can select a monitoring station via the on-line map and instantly receive metadata associated with that station. The map also includes a Web form (at http://cuereims.umbc.edu/mwmc_new/trunk/www/index.asp) that standardizes monitoring program metadata submitted by users.

Fig. 2 shows the conceptual and physical architecture of our system. The system consists of a Web interface component, a data integration component, a data warehouse/geodatabase component, and an analysis and visualization component. The Web interface allows users to submit data and metadata, to select data sources via a semantic network, to perform analyses on data stored in multidimensional structures, such as Online Analytical Processing (OLAP)[1] data cubes, to mine data and discover patterns, and to use a GIS interface to visualize and query spatial data. It is expected that this component will be mainly used by decision and policy makers. The data integration component consists of applications to integrate data and metadata, provides automatic schema mapping, converts data to a canonical format, and extracts spatial relationships. Once the data are integrated, it is stored in the data warehouse. The data warehouse component consists of multidimensional data cubes, a metadata repository, and additional mining components.

4.2. Data integration

We integrate all data sources using the shallow integration approach and allow experts to search for data sources of their own interest. We apply deep integration approach to a subset of widely used data sources (e.g., the EPA STORET database (http://www.epa.gov/storet), the USGS National Hydrography Dataset (http://nhd.usgs.gov/), and Maryland state agency data. These data sources will be further stored in a data warehouse for query and analysis purposes.

4.2.1. Shallow integration

We address two main issues for shallow integration: (1) integrating metadata from different sources into a universal format and (2) providing a semantic-based source selection interface for users to find data sources of interest.

[1] On-line Analytical Processing (OLAP) provides technical approaches and tools to quickly answer queries that operate on multidimensional data (data cubes). For example, date, region, and product are three dimensions forming a data cube of sales data.

4.2.1.1. Integration of metadata. The current Clickable Map Web interface allows data monitoring stations to input metadata, including the location of monitoring stations, parameters being monitored, units, monitoring types (e.g., chemical, physical, biological, etc.), the duration of collection, etc. However, the metadata format of the Clickable Map is different from the metadata format of other data sources such as the NHD and STORET.

Our approach is to convert the metadata from all these sources to the EML format. EML is a well-known metadata standard developed by and for the ecology community. We chose EML for two reasons. First, EML has been used in several well-known projects (KNB, n.d.; SEEK, n.d.; Stiglitz et al., 2001) for metadata integration. Second, EML is implemented as a series of XML document types that can be used in a modular and extensible manner to document ecological data. XML has become the de facto standard of exchanging information over the Internet. Currently the conversion of metadata to EML is manual, although we are addressing this problem and plan to make the conversion (semi)automatic.

4.2.1.2. Selection of semantics-based sources. The main problem of Clickable Map and other existing systems (KNB, n.d.; SEEK, n.d.; Stiglitz et al., 2001) for integrating environmental data is the limited support in assisting users to select data sources that are semantically related to their interests, as we already mentioned.

In information retrieval studies, query expansion techniques have been used to address the problem of finding relevant articles (Jing & Croft, 1994; Mitra, Singhal, & Buckley, 1998; Qiu & Frei, 1993; Xu & Croft, 1996). Query expansion adds to the query terms statistically co-occurring with the search term. For example, the term “lung excision” and “smoking” will be added to a search term “lung cancer” because in many medical articles these terms co-occur. The SEEK project (Bowers, Lin, & Ludascher, 2004; Bowers & Ludascher, 2004) also uses a predetermined ontology for ecology concepts to expand query terms.

4.2.2. Prototype system

We have developed a prototype system (Chen, Gangopadhyay, Karabatis, McGuire, & Welty, 2007) consisting of software modules that belong to the analysis and visualization and integration components of Fig. 2. This prototype system differs from existing data integration systems on two important aspects. First, we capture the relationships on the data source level of granularity rather than on the term level because spatial and temporal information is available at that granularity. For example, although the terms “stream temperature” and “fish population” are relevant, the temperature of a stream in Maryland is not relevant to the fish population of a stream in California. In our prototype system, we used semantic networks to capture the semantic relationships between data sources and help users select additional data sources. This work expands on the specification of relationships among database objects stored in heterogeneous database systems (Frazier & Sheth, 1985; Georgakopoulos, Karabatis, & Gantimahapatruni, 1997; Karabatis, Rusinkiewicz, & Sheth, 1999; Rusinkiewicz, Sheth, & Karabatis, 1991).

These semantic relationships form a semantic network of related information, which assists users to discover additional information, relevant to their search but possibly unknown to them. We identify this property of the system as extremely valuable to government and decision makers. We realize that some dependencies may not be captured initially in the metadata repository, especially when semantic incompatibilities prevent direct identification of data (such as problems related to synonyms, homonyms, etc.). Nevertheless, missing dependencies are captured and added to the metadata repository by observing usage patterns by users who interact with the semantic network, as we will describe later. The notion of dependencies (relationships) between concepts is also related to the topic maps (TopicMap, n.d.) or concept maps and semantic Web (W3C, n.d.) for XML and Web documents containing metadata about concepts.

A semantic network consists of nodes corresponding to data sources and edges corresponding to relationships between sources. Each edge is also associated with a relevance score identifying the degree of relevance between the two connected nodes. Fig. 3 shows a semantic network for the above fish example. The semantic network can be used to recommend additional data sources of possible interest to the user. For example, if a user is interested in the fish population (fish counts) data source, the semantic network in Fig. 3 would recommend fish counts, vegetation, stream chemistry, and stream temperature data sources. Internally, the semantic network utilizes conditional probabilities to identify the relevance score between any two nodes in the network (Chen et al., 2007). The values of these conditional probabilities identify nodes that are related with one another (either directly or indirectly) through all possible paths in the semantic network. These nodes identify data sources that are relevant with the initial query and thus, recommended by our system.

Fig. 4 shows the front–end of the query system incorporating the semantic network capability. It illustrates an example where a user entered “fish” and the semantic network responded with three additional and relevant data sources that contain information the user might be interested in. Note that several nodes in the network were not selected by our system (such as elevation, stream flow, impervious surfaces, and meteorology) because these nodes were considered not to be quite as relevant to the initial user query as defined by the conditional probabilities connecting the various nodes starting with the data source node containing the user supplied keywords. When a user clicks on the “View Network” button, our

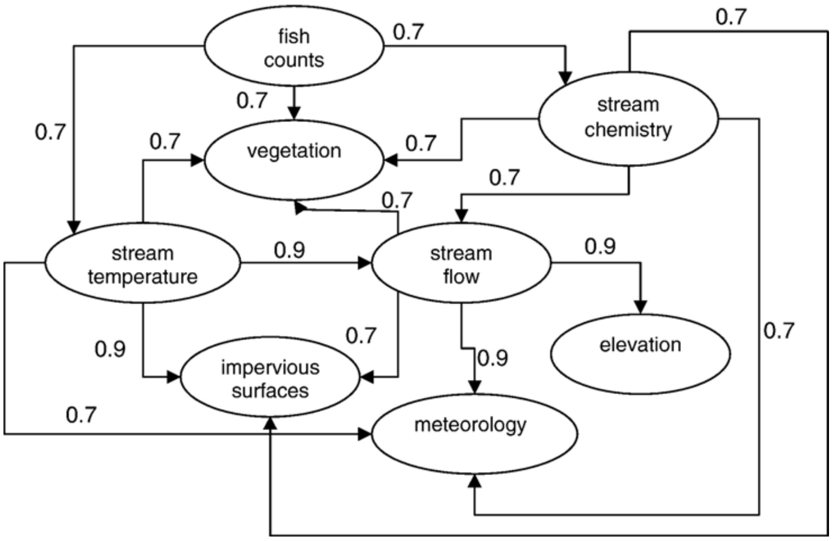
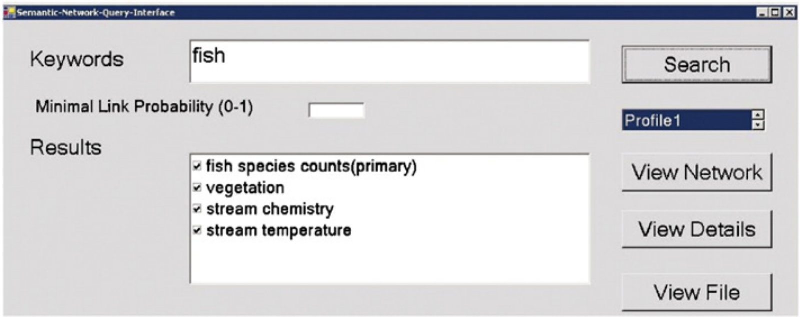


Fig. 3. An example semantic network.


Semantic-Network-Query-Interface

Keywords: fish

Minimal Link Probability (0-1)

Profile1

Results

- ✓ fish species counts(primary)
- ✓ vegetation
- ✓ stream chemistry
- ✓ stream temperature

Fig. 4. Semantic network recommendations for vegetation input.

system displays the semantic subnetwork containing just the selected nodes. The user can also define the lowest degree of relevance between nodes to be possibly considered for semantic selection; this is the "minimal link probability" (MLP) in the user interface. The system preset value is 0.5, but a user may increase it or decrease it depending on the number of relevant data sources that he or she is looking for, fine-tuning the number of recommended data sources. A higher MLP value indicates that a user prefers highly relevant sources to be retrieved, thus the system (in general) will return fewer but highly related recommendations. On the contrary, a lower MLP indicates that the user is interested in relevant data sources but not very strongly related; thus, the system will most likely recommend more data sources.

The second important aspect is that we utilize methods to constantly refine the semantic network based on user access patterns of the sources. For example, suppose many users select the fish population source and also look at the vegetation data source. Based on such usage, we refine the semantic network and add an edge between these two data sources such that the vegetation data source will be recommended to subsequent users who access fish population.

We also consider the issue that different users of our system may not have the same interests or regulatory and oversight responsibilities. Instead they may need to utilize different semantic networks (if available) pertaining to their own special interests in the data. In addition, users from different government agencies would prefer to view data from a perspective related to their organization or agency's goals and responsibilities. Within the same governmental agency there are even different types of users with different job duties and they would also prefer to have access to the data according to their own views. For example, a stream chemist may not be interested in the Elevation data source, but an urban developer would certainly focus on it. We address this problem by creating different user profiles, each corresponding to a separate semantic network, reflecting the data needs and interests of a particular stakeholder. Initially, representatives from various stakeholder entities define a set of profiles. Before using our system, users from specific governmental entities will select a profile most appropriate to the particularities of their job. They have the ability to change this selection at any time. For each profile, we also track the usage patterns by users and collect information that is used to dynamically refine and augment the semantic network based on these patterns. Therefore, although an initial profile may not completely satisfy every user, it will adapt to user preferences over time.

We also developed a semiautomatic method to construct a specific semantic network for water monitoring domain. This method uses a set of manually generated rules to guide semantic network construction. For example, one rule could be "add an edge with 0.9 relevance score between two sources that overlap spatially and contain the terms "fish" and "stream temperature." We developed a source selection interface in the Clickable Map using the constructed semantic network to record the usage patterns of users and fine-tune the semantic network. So starting from a small network, which was manually created, we record user queries and user patterns to expand the network. Additionally, when there is no activity for a specific edge in the network for long periods of time, the edge "ages" and its relevance score decreases. Therefore, we emphasize active parts of the network, while inactive ones may become passive. However, users can submit new edges between data sources to the system that will adopt them after an expert environmental scientist concurs. The aging algorithm is described in detail by Chen et al. (2007).

4.2.3. Deep integration

We address three main issues for deep integration: (1) mapping the schema (metadata) of individual sources to the global mediated schema, (2) converting data at individual sources to a canonical form, and (3) extracting spatial relationships from data.

4.2.3.1. Automation of schema mappings. This is a very important component of our system since it automatically maps schemas of individual sources to the mediated schema; therefore, users and policy makers see a consistent view of the sources. Our approach is similar to the ones by Doan, Domingos, and Halevy (2003) and Subrahmanian and Garrett (2002) and more recently by Madhavan, Bernstein, Doan, and Halevy (2005). In essence, we store existing mappings from one source to another as templates, with each template specifying the type of parameter being monitored.

4.2.3.2. Conversion of data to a canonical form. Water monitoring data are collected in a variety of units (metric or SI). Moreover, various data sources may have different spatial and temporal granularities, and the data may be collected at different time intervals as we already mentioned. Programs to convert different units into the same unit are essential in such architectures.

4.2.3.3. Extraction of spatial relationships. We also extract two types of spatial relationships between data items and store them in the data warehouse: (1) the relationships between a water monitoring site and a stream reach and (2) the relationship between stream reaches. This is quite useful since each monitoring site in a stream reach can provide a link to other nearby monitoring sites and also a link to other monitoring sites in different stream reaches through the stored relationship between stream reaches. In effect, such spatial relationships allow users and policy makers to identify the existence of similar events in different related geographical areas.

4.3. Data warehouse for water quality management

We have designed a data warehouse for the purposes of decision support for water quality government managers. Data in repositories which have been selected for deep integration also reside in our data warehouse. The warehouse facilitates (1) analysis of water quality measurements at multiple hydrographic spatial hierarchies; (2) aggregation of data at multiple spatial hierarchies; and (3) enhancement of the data with spatial relationships and analysis of OLAP queries.

Water quality management often requires the analysis of spatial relationships between the observed phenomena and their placement within the landscape. This requires the incorporation of an extensive spatial component to the data warehouse that facilitates the execution of data mining queries at multiple spatial and temporal resolutions. Fig. 5 shows a data warehouse model (star schema) that we have implemented for stream water monitoring data.

This model consists of various measurements of water chemistry, stream characteristics, indices of biological integrity, and fish species counts (population). Also, there are a number of spatial and temporal dimensions (hierarchies) included in the model, the most interesting of which are the site and the hydrography dimensional hierarchies. Using the site dimension, it is possible to connect monitoring point locations along the stream network using the stream reach dimension (NHD Reach Dimension) and calculate the approximate distance between them. The hydrography dimension consists of drainage areas such as the sub-watershed (NHD Sub-watershed Dimension), which represents the land area that drains to a particular point along a stream segment. Since this is the lowest level spatial unit in the watershed dimensional hierarchy, other spatial distributions such as land cover, which have a major impact on water quality impairment (Sponseller, Benfield, & Valett, 2001), are summarized at this level. The remainder of the hierarchy consists of progressively larger aggregations of sub-watersheds to watersheds to sub-basins and to basins (left side of Fig. 5). The hydrography spatial hierarchy allows water quality managers to roll up to summary results in progressively larger catchment areas and drill down to the details of each individual monitoring site. The ability to traverse the spatial hierarchy dimensions allows for the resolution of multiple spatial incompatibilities, by aggregating data to the least common level of hierarchy and performing a join at that level. Similarly for the time dimension hierarchy, we represent data in the aggregated level of hierarchy. Thus, while performing spatiotemporal joins/aggregations, at the same time we eliminate the spatiotemporal disparities. This model can be easily extended to include spatial relationships between watersheds and to incorporate other types of water quality monitoring data such as water chemistry, contaminants, and sediments.

4.3.1. Analysis of data in the data warehouse

Our system allows many different types of monitoring data to be analyzed and provides further integration of monitoring data across space and time. Interested users in general and government employees in particular may dynamically access data at multiple spatial hierarchical levels. However, this approach creates an aggregation problem: the detail of each monitoring point is lost as aggregation units become larger. For example, if there are two sites in the same sub-watershed and one site has a count of 15 for a certain species of fish and the other site has a count of 0 for the same species, when these two sites are aggregated to the sub-watershed level using a summation function, the total count is 15 but it is not clear how many fish were present at each site. A methodology that utilizes multidimensional wavelet transforms can be used in such cases, which allows for aggregation of data at multiple

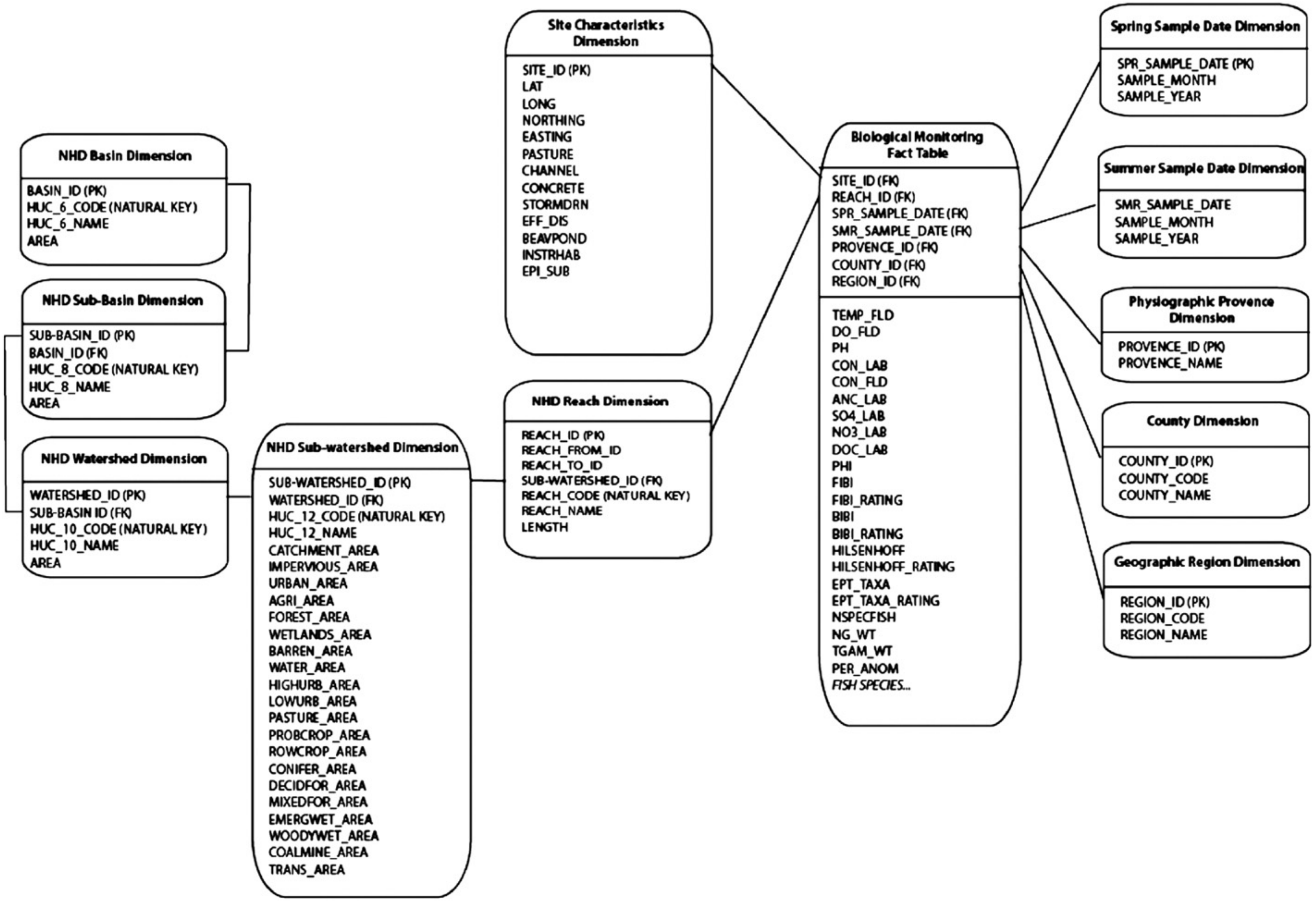


Fig. 5. Data warehouse for Maryland DNR biological stream survey data.

spatial hierarchies while preserving details at the lowest level; an interested reader can find detailed information by Daubechies (1992) and Goswami and Chan (1999); however, this discussion is outside the scope of this paper.

5. Experimental results

We conducted a preliminary usability study to validate our approach of using semantic networks to help environmental researchers find related data sources. The hypothesis of our study was that environmental researchers using our system would find answers to their research problems quicker, easier, and with more confidence as compared to not using our system. We also asked them to rate their satisfaction and overall experience using the semantic network system.

To set up our experiment, we asked an environmental researcher to act as an expert and define an initial semantic network containing information on pre-existing data sources on fish population, vegetation, stream chemistry, stream flow, temperature, elevation, meteorology, impervious surfaces, and stream flow (see Fig. 3 presented earlier). Once the initial network was created, we developed the relevant profiles (on fish species, stream network flow, and stream chemistry) and performed the following experiment. We asked 15 subjects, including environmental scientists, policy makers, and students, to answer the three tasks research questions listed below using our system:

- Question 1: Which data sets have an effect on stream flow?
- Question 2: Which data sets have an effect on fish species counts?
- Question 3: Which data sets have an effect on stream chemistry?

Using our system the subjects chose the appropriate profile and starting with a keyword search they explored the related sources in our system. We compared their answers with the correct answers provided by the expert environmental researcher and discovered that the subjects identified the correct answers for each question with 96.4% accuracy. We also asked them to explicitly rate our system on five categories. For each category, they were asked to compare it against the baseline of answering the same questions without our system (no semantic network techniques). The five rated categories were:

- Timeliness: How quickly they accomplished the requested task using our system.
- Easiness: How easy it was to answer the research questions.
- Satisfiability: How satisfied they were to complete the tasks.
- Confidence: The degree of confidence they had on their answers.
- Overall experience: How they rated the overall interaction with the system.

A 7-point Likert scale was used for the questionnaire. The scoring scale ranges from 0=*lowest score* to 7=*highest score*. Higher scores indicate better results. The result of the usability study confirmed our hypothesis and showed high degree of overall approval of the system. Fig. 6 illustrates the average score by the users for each category tested. They all ranked

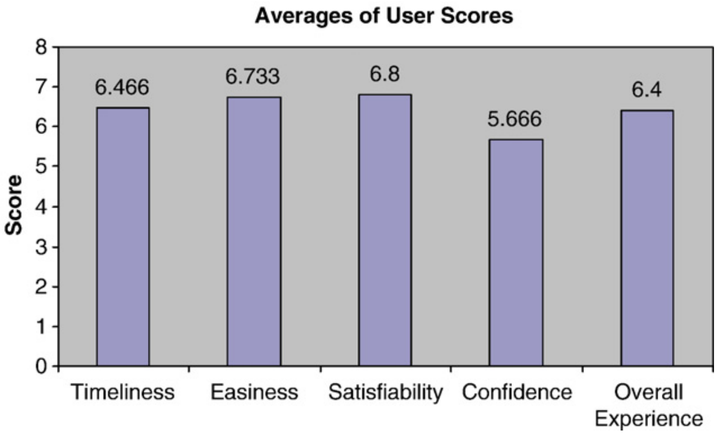


Fig. 6. Averages of user scores for each category in the usability study.

quite high ranging from 5.6=*confidence* to 6.8=*satisfiability*. The subjects scored their overall experience with the system at 6.4. The lowest score was recorded on confidence. When we asked the subjects about this rating they mentioned that there might be other data sources relevant to the ones being asked, which they might not know about, and are not included in our system yet.

The users who experimented with our system concluded that utilizing the semantic network approach surpassed the traditional exact query systems. In almost all cases we tested, the users were able to find the correct data sources to answer the research questions being asked. They also gave high scores on timeliness, ease of use, user satisfaction, confidence, and overall experience with the system. Therefore, our hypothesis was proven to be true.

6. Conclusions and further work

Data integration and systems interoperation are challenging but necessary tasks for government agencies and water monitoring councils that need to identify pollution sources and give the public access to such information. In this paper, we discussed how a prototype system allowed users to integrate water monitoring data across many government organizations and use the subsequent data integration and an enhanced interface for data retrieval to provide better information for water quality monitoring and related decision making.

The prototype used a dual approach to data integration, including both a shallow and deep level. This approach automatically identifies semantically relevant yet unknown information and supports the creation of user profiles for possible expansion of the high level queries with the ability to refine the profiles over time. Preliminary feedback from a usability study indicated that the prototype delivered improved data query capabilities and the users rated five measured categories quite high. Improvements included a wider array of data sets, customized views with the ability of the user to select or ignore related sets of data depending on their task or stakeholder position and a richer, data warehouse that brought together previously disparate, and unintegrated data sources.

As is the case with most prototypes there is more work to be done. Some of this work responds to the emerging and evolving needs of users as they gain familiarity with the capabilities of the system. As part of our continuing work, and based on the feedback of users, we will improve our system to provide an enhanced view of information through visualization and analysis tools. Such additions expand the information that can be presented to decision makers, provide them with additional knowledge, and enable better water management policy. We also plan to evolve our system beyond the prototype status, construct it more robustly, and make it available to the public.

There is also obviously a need to expand user testing and build on the existing profiles and related semantic networks. Also we plan to extend the use of the system beyond just the state of Maryland and incorporate metadata for additional states. This will not be difficult to do regarding concepts, but it will take time to collect the metadata and enable shallow and deep integration.

At the outset of the paper, a statement of the challenges of making informed decisions on water quality policy underscored the need to rationalize the various sources, sets, and levels of data that are available to policy makers in this domain. The review of the literature of e-government and data integration established a linkage between advanced levels of e-government and underlying data integration for users external to government, stakeholders, and policy makers. The prototype system described in this paper documents a technical approach to data integration to support both monitoring and related policy making for water quality in Maryland. Preliminary user testing has demonstrated some of the potential benefits from more advanced levels of e-government, both vertical and horizontal integration, identified in Layne and Lee's (2001) four-state model of e-government. While this is clearly a first step in a much longer journey, the paper documents that is possible to achieve data integration as a means to delivery improved e-government information products and services.

Acknowledgments

This material is based upon work partly supported by the National Science Foundation under Grant No. BES-0414206 to C. Welty and M. Mcguire and by the United States Environmental Protection Agency under grants R-82818201-0 to A.J. Miller and CR83105801 to C. Welty. Although the research described in this article has been funded wholly or in part by the United States Environmental Protection Agency, it has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endoresement should be inffered.

References

Baum, C., & Maio, A. D. (2000). Gartner's four phases of e-government model.

Berberidis, C., Angelis, L., & Vahavas, I. (2004). *Inter-transaction association rules mining for rare events prediction*. Paper presented at the 3rd Hellenic Conference on Artificial Intellligence (SETN'04).

Bowers, S., Lin, K., & Ludascher, B. (2004). *On integrating scientific resources through semantic registration*. Paper presented at the Scientific and Statistical Database Management.

Bowers, S., & Ludascher, B. (2004). *An ontology-driven framework for data transformation in scientific workflows*. Paper presented at the International Workshop on Data Integration in the Life Sciences.

Chen, Z., Gangopadhyay, A., Karabatis, G., McGuire, M., & Welty, C. (2007). Semantic integration and knowledge discovery for environmental research. *Journal of Database Management*, *18*(1), 43−67.

Cushing, J. B., & Wilson, T. (2005, May 15–18). *Eco-informatics and natural resource management*. Paper presented at the 2005 National Conference on Digital Government Research Atlanta, GA.

Daubechies, I. (1992). *Ten lectures on wavelets*: Capital City Press.

Doan, A., Domingos, P., & Halevy, A. Y. (2003). Learning to match the schemas of data sources: A multistrategy approach. *Machine Learning*, *50*(3), 279−301.

EML. (n.d.). Ecological Metadata Language. Retrieved from http://knb.ecoinformatics.org/software/eml/

Frazier, G. L., & Sheth, J. N. (1985). An attitude-behavior framework for distribution channel management. *Journal of Marketing*, *49*(3).

Georgakopoulos, D., Karabatis, G., & Gantimahapatruni, S. (1997). Specification and management of interdependent data in operational systems and data warehouses. *Distributed and Parallel Databases, An International Journal*, *5*(2), 121−166.

Goswami, J. C., & Chan, A. K. (1999). *Fundamentals of wavelets: Theory, algorithms and applications*: John Wiley.

Halevy, A. Y., Ives, Z. G., Suciu, D., & Tatarinov, I. (2003). *Schema mediation in peer data management systems*. Paper presented at the ICDE.

Hiller, J. S., & Bélanger, F. (2001). Privacy strategies for electronic government. In M.A. Abramson & G. E. Means (Eds.), *E-Government 2001* (pp. 162−198). Lanham, MD: Rowman & Littlefield.

Hyperion. (n.d.). The Hyperion Project. Retrieved from http://www.cs.toronto.edu/db/hyperion/

Jing, Y., & Croft, W. B. (1994). *An association thesaurus for information retrieval*. Paper presented at the RIAO'94.

Karabatis, G., Rusinkiewicz, M., & Sheth, A. (1999). Interdependent database systems. *Management of heterogeneous and autonomous database systems.* San Francisco, CA: Morgan-Kaufmann.

KNB. (n.d.). Knowledge Network for Biocomplexity Project. Retrieved from http://knb.ecoinformatics.org/index.jsp

Layne, K., & Lee, J. (2001). Developing fully functional e-government: A four stage model. *Government Information Quarterly*, *18*(2), 122−136.

Lee, S. M., Tan, X., & Trimi, S. (2005). Current practices of leading e-government countries. *Communications of the ACM*, *48*(10), 99−104.

Levy, A. Y., Rajaraman, A., & Ordille, J. J. (1996). *Querying heterogeneous information sources using source descriptions*. Paper presented at the VLDB.

Madhavan, J., Bernstein, P. A., Doan, A., & Halevy, A. Y. (2005). *Corpus-based schema matching*. Paper presented at the ICDE.

Miller, R. J., Hernandez, M. A., Haas, L. M., Yan, L., Ho, C. T. H., Fagine, R., et al. (2001). The Clio project: Managing heterogeneity. *SIGMOD Record*, *30*(1).

Mitra, M., Singhal, A., & Buckley, C. (1998). *Improving automatic query expansion*. Paper presented at the ACM SIGIR.

Moon, M. J. (2002). The evolution of e-government among municipalities: Rhetoric or reality? *Public Administration Review*, *62*(4), 424−433.

Norris, D. F., & Moon, M. J. (2005). Advancing e-government at the grassroots: Tortoise or hare? *Public Administration Review*, *65*(1), 64−75.

Pantel, P., Philpot, A., & Hovy, E. (2005). Data alignment and integration. *Computer*, *38*(12), 43−50.

Papakonstantinou, Y., Garcia-Molina, H., & Ullman, J. (1996). *Medmaker: A mediation system based on declarative specifications*. Paper presented at the ICDE.

Qiu, Y., & Frei, H. P. (1993). *Concept-based query expansion*. Paper presented at the ACM SIGIR'93.

Rusinkiewicz, M., Sheth, A., & Karabatis, G. (1991). Specifying interdatabase dependencies in a multidatabase environment. *IEEE Computer*, *24*(12), 46−53.

SEEK. (n.d.). The Science Environment for Ecological Knowledge. Retrieved from http://seek.ecoinformatics.org

Sponseller, R. A., Benfield, E. F., & Valett, H. M. (2001). Relationships between land use, spatial scale and stream macroinvertebrate communities. *Freshwater Biology*, *46*(10), 1409−1424.

Stiglitz, J. E., Orszag, P. R., & Orszag, J. M. (2001). The role of government in a digital age. Retrieved May 26, 2002, from http://www.ccianet.org/digital_age/report.pdf

Subrahmanian, E., & Garrett, J. H., Jr. Two Experiences in Digital Government. Retrieved May 28, 2002, from http://216.239.51.100/search?q=cache:qWRKeboCPXsC:www.ctg.albany.edu/research/workshop/21-subrahmanian.pdf+Two+Experiences+in+Digital+Government&hl=en&ie=UTF8

Tatarinov, I., & Halevy A. Y. (2004). *Efficient query reformulation in peer-data management systems*. Paper presented at the SIGMOD.

TopicMap. (n.d.). XML Topic Maps (XTM) 1.0. Retrieved from http://www.topicmaps.org/xtm/

West, D. M. (2004). E-government and the transformation of service delivery and citizen attitudes. *Public Administration Review*, *64*(1), 15−27.

W3C. (n.d.). Semantic Web. Retrieved from http://www.w3.org/2001/sw/

Xu, J., & Croft, W. B. (1996). *Query expansion using local and global document analysis*. Paper presented at the SIGIR.