A Multi-layered System Architecture for Environmental Monitoring Data Management – Taiwan's Experience

Yu-Chi Chu[1]

Abstract

This paper presents a comprehensive system architecture which adopts a multi-layer approach and is based on the web-based platform for environmental monitoring data management, ranging from data collection to applications. This architecture is designed in a modular and flexible fashion. As an example of practicality, the second phase Taiwan Air Quality Monitoring Network (TAQMN-2) has been implemented using the proposed system architecture to justify its feasibility. The overall structure of TAQMN-2 is described, and two major subsystems, data collection manager (DCM) and data access manager (DAM), are illustrated in detail. Taiwan EPA benefited greatly from the results of TAQMN-2, not only in terms of streamlining the air quality data management but also in upgrading work performance on improving the air quality in Taiwan. The usage of proposed architecture is not restricted to air quality data management; it is also suitable for other domains of environmental science with limited tuning, adjustments, or augmentation.

1. Introduction

Environmental monitoring data is the foundation for understanding and managing our living planet. It also plays an important role in supporting environmental decision-making. The tasks of collecting, managing, and accessing a large volume of diverse monitoring datasets has become ever more challenging due to the expansion of various newly available instruments, measurements, as well as diverse and complicated data storage systems. Over the past few years, a number of studies and efforts have been made on the issue of data management for environmental monitoring; what seems to be lacking, however, is a comprehensive and integrated system architecture that can streamline and consolidate the data management processes for environmental monitoring. The processes should include the data collection, data computation, data storage, as well as data dissemination and applications.

This paper presents a comprehensive system architecture, which adopts multi-layer approach and is based on the web-based platform for monitoring data management, ranging from data collection to applications. Our purpose is to consider the streamlining of data management processes for environmental quality monitoring. As an example of practicality, the second phase Taiwan Air Quality Monitoring Network (TAQMN-2) has been implemented using the proposed system architecture to justify its feasibility.

The following section describes the overall characteristic and features of a multi-layer system architecture mainly for environmental monitoring data management. Based on the proposed architecture, section 3 will illustrate the details of the TAQMN-2 project, including the overall structures and two major components, data collection manager and data access manager. A number of concrete results produced by TAQMN-2 are given in section 4. We will conclude with discussion on the future work in section 5.

[1] Department of Environmental Monitoring and Information Management,
Environmental Protection Administration, Taiwan, R.O.C.

2. System Architecture

As shown in Figure 2.1, the system architecture adopts a multi-layer approach to facilitate seamless interaction from the data collection to data applications of environmental monitoring. Each layer is comprised of several components which interact within the layer and communicate with other layers. We explain the roles and functionalities of each layer as follows:

- *Collection layer*: comprises a number of apparatuses such as sensors and instruments responsible for measuring environmental quality. These apparatuses are connected to a programmable logic controller (PLC) and form the front line for monitoring data collection and gathering. The PLC uses a standard interface to adapt digital/analogy signals coming from the apparatuses and convert the signals to numerical data. There is an on-site database dedicated to storing these data, converting them into XML format files, and then transmitting them to the transmission layer. Unlike traditional data acquisition systems (DAS) which usually use proprietary formats and protocols, our system adopts a series of open standards such as HTTP and TCP/IP for connecting monitoring apparatuses and IT devices, making the collection layer an Internet-based application.
- *Transmission layer:* (i) apply virtual private network (VPN) technology which integrates with the comprehensive firewall and router features. It supports connection for a secure data transmission over the Internet. (ii) use Microsoft Message Queuing (MSMQ) technology as a data transmission platform that bridges the collection layer and manipulation layer. The MSMQ provides guaranteed messaging, priority-based messaging and enables applications running at different times to communicate across networks that may be temporarily offline.
- *Manipulation layer:* performs several procedures for data pre-processing. When data is gathered from the monitoring sites, it moves through the staging server along with an intermediary store. The staging server is the place where all collected monitoring data is put together and prepared for loading into the database. It is like an assembly plant or a construction area. In this area, we examine collected data, perform the various functions for data quality assurance and resolve inconsistencies. Once the data is finally prepared, it will temporarily reside in the intermediary data stores waiting to be loaded into the monitoring database.
- *Storage layer:* mainly consists of a commercial database management system (DBMS) along with a set of toolkits that may construct and maintain the metadata for underlying monitoring data. As the monitoring data increases with different formats both in syntax and semantics, the metadata acts like a nerve centre in the storage layer. Furthermore, in order to provide data services more efficiently and widely, this layer may also aggregate data and functionalities from other sources such as the meteorological offices which may provide data on environmental resources in particular weather forecast information.
- *Application layer:* provides a number of applications and services for data retrieval, analysis, as well as data disseminations. Since the system uses a web-based as a standard user interface, a variety of advanced applications such as data mining and data visualization can also be attached as part of the systems. This makes our system more flexible and scalable than client-server architecture. The user interface for most applications uses Web standards: XHTML and CSS are used for presentation; JavaScript and ASP.NET are used to handle user interaction. In addition, Web Services related standards and technologies such as SOAP, WSDL, and AJAX are adopted to deal with the communication and integration issues with other environmental data resources.

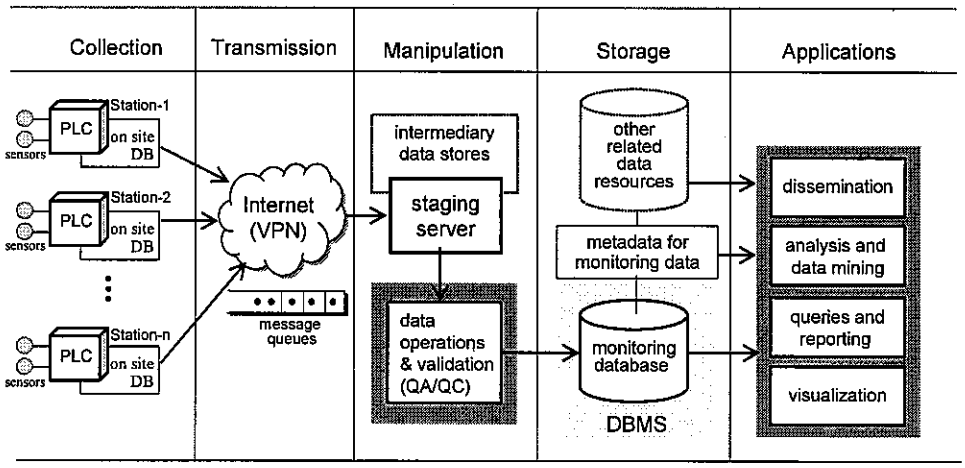


Figure 2.1 System architecture for environmental monitoring data management – ranging from data collection to data applications.

Our approach can be easily applied to different subjects of environmental monitoring such as water quality monitoring and hazardous waste monitoring. We also believe that our solutions are applicable to other domains of environmental science. For example, the data management problems for ecology and natural resource observation may also be applied to the proposed system architecture with limited tuning, adjustments, or augmentation.

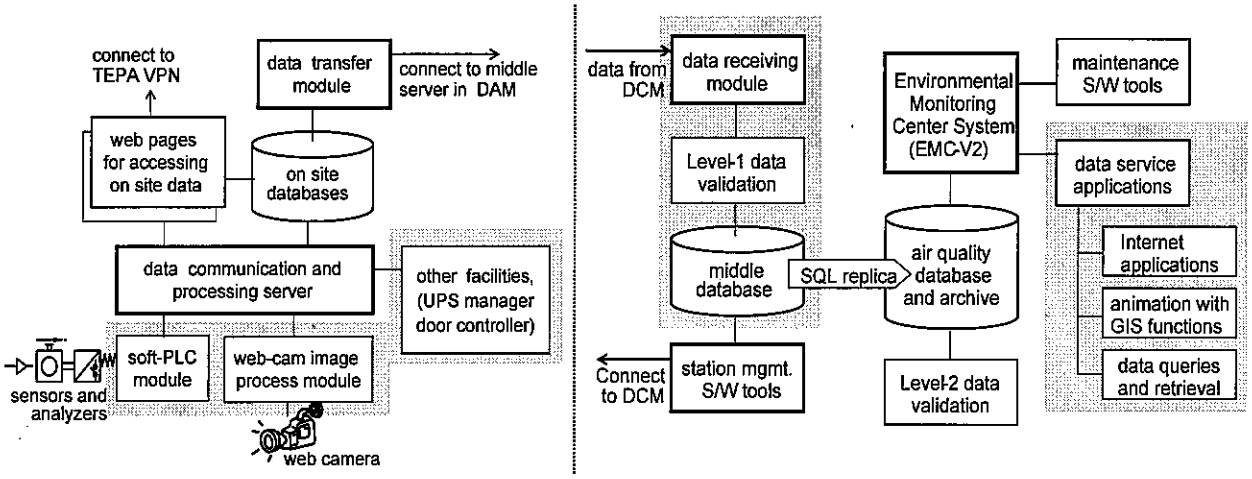
3. TAQMN-2 Project

Air quality and water quality are the primary targets of Taiwan's environmental monitoring efforts. Over the last ten years, Taiwan has developed diverse, advanced monitoring systems. These systems not only provide residents with an abundant source of information on environmental quality, but also provide opportunities for international technology exchange and environmental cooperation.

The Taiwan EPA began monitoring air quality in 1980. At the beginning there were nineteen stations situated in Taiwan's major cities. In 1993, the government developed the Taiwan Air Quality Monitoring Network, abbreviated TAQMN. The scope was to install up to 66 air quality monitoring stations, two mobile vans equipped with monitoring facilities, one quality assurance laboratory, and a data management centre. This network has been operating since September 1993. In recent years, aside from replenishing the network with more sophisticated equipment and expansions we have also been dedicated to converting the former standalone systems into a more flexible and configurable environmental information system. In our vision, the new age of environmental monitoring systems need not only provide qualitative and instantaneous records, they need to be useful analytically and able to forecast more accurately. Many air quality control measures have to be based on the outcomes of monitoring processes.

TAQMN-2 was a four-year project from 2002 to 2006 to enhance the capability of air quality monitoring in Taiwan, adopting the proposed system architecture for the whole spectrum of data management. TAQMN-2 successfully demonstrates the applicability of our approach and verifies the expected benefits by describing a compelling case in the domain of environment data management.

TAQMN-2 has led the establishment of 76 monitoring stations in different regions of the country. The pollutants monitored in TAQMN-2 stations include of PM_{10}, carbon monoxide, sulfur dioxide, nitrogen dioxide, and ozone. The meteorological instruments produce parameters, such as wind direction, wind speed, temperature, dew point and precipitation, which can make the air quality forecasting more accurate. Some sophisticated equipment has been added to the list in order to measure acid rain, hydrocarbons, $PM_{2.5}$, ultraviolet-type B, etc.



(a) data collection manager (b) data access manager

Figure 3.1 TAQMN-2 overall structures

3.1. Overall structures

Figure 3.1 depicts the overall structures of TAQMN-2. First, we implement a subsystem called data collection manager (DCM) to collect the data from each station. The stations feed the measurements into the DCM, which also collects data from other monitoring facilities such as the web-cam images (for security consideration) and the power system monitoring data. After the DCM receives hourly measurements, it deposits them into a collective on-site database and uses MSMQ technology to transmit the data. Secondly, in the manipulation layer, there are software modules to validate measurements, tools to check and do data maintenance. Thirdly, by using the data access manager (DAM) in the storage layer, different applications can access the data accordingly. In the application layer, there are web applications to provide the on line statistics and provide information for air quality forecast. For data dissemination, TAQMN-2 can transfer the statistics and forecast data via web pages, toll free phones, TV, radios, and newspapers.

To keep TAQMN-2 as an open and flexible architecture which allows different types of hardware and software work together seamlessly, we employ the most common standards and the most widely used products. Table 3.1 summarizes the implementation environment of TAQMN-2, ranging from hardware platform to software development tools. Although most system software such as operating systems and DBMS are commercial products, it is easy to migrate to a free software environment like Linux and MySQL if the scale of monitoring systems is of suitable size.

Table 3.1 TAQMN-2 implementation environment

	Data collection manger (DCM)		Data access manager (DAM)	
	gathering and collecting	communication and processing	middle server	database server
H/W	Soft-PLC	Xeon (DP) 2.4GHZ	Xeon MP 1.5GHZ	HP RX-5670 HP MSA-1000
OS	Build-in Kernel	Win2000 Server	Win2000 Server	Win2000 Server
Data storage	Text Files	Oracle 9i DMBS	Oracle 9i DMBS	Oracle 9i DMBS
Development tools	Ladder, Java and C language	MS .Net and Java language	MS .Net, Java and SQL	MS .Net, Java and SQL

3.2. Data collection manager (DCM)

As shown in Figure 3.1(a), the DCM is comprised with a number of mentoring facilities, equipments and software modules. For collecting monitoring data, we connect most of the air quality sensors and analyzers to a SoftPLC. The SoftPLC has three modules, digital input (DI), analog input (AI) and high level language (HLL), that adopt different methods to collect data and convert the data into a text file, then store the file in the built-in memory of SoftPLC ready to be accessed by the data communication and process server. Some specific devices such as web-cams, UPS, and door controllers are connected to data communication and process servers directly since their interface are relatively simple and ready to connect to the Ethernet.

Data communication and processing are the main services in the DCM. The DCM gathers all monitoring data, computes hourly average values for each measurement, and uses SQL procedures to store the data into an on-site database. For ensuring the correctness of each analyzer, the DCM periodically issues a series of collaboration commands to trigger the SoftPLC and corresponding monitoring analyzers to perform collaboration procedures. In general, the collaborated information is sent and the results of collaborations are collected in XML format. Therefore, they can easily be integrated into the on-site database and be presented on web pages.

The DCM also generates web pages allowing users to access the row data, the status of equipment, as well as collaboration records and the web-cam images stored in the on-site database by a common web browser. However, this access can be done only via the VPN within TEPA for security considerations.

On the top right of Figure 3.1(a), the data transfer module is in charge of sending the measurement data to the DAM for further processing. The data transfer module adopts Microsoft Message Queuing (MSMQ) technology as the mechanism of data transmission. This mechanism can deal with the re-transmission process appropriately when the system confronts a network disconnect problem. Therefore, we won't lose any data from the DCM.

3.3. Data access manager (DAM)

The DAM can be divided into two subsystems, middle server and database server, based on the functionalities of services. The design concept of the middle server is to overcome the difficulties when writing mass data (like time series data) into a database, since it might slow down the performance of whole systems. The middle server plays the role of staging servers in the manipulation layer described in Section 2. The database server is dedicated to storing detailed data and is re-

sponsible to the application developments. Thus, it can be viewed as an reflection of the storage layer and application layer described in Section 2.

As shown on the left of Figure 3.1(b), the data receiving module accepts data from the DCM, and triggers the Level-1 data validation procedures, which include data effectiveness checking and thresholds checking. All the validation rules are stored in a predefined profile. The inspected data are then stored in the middle database. The middle server is also responsible for the configuration management of each monitoring station. A set of software tools in the middle server are dedicated to performing management functions remotely. For example, users with certain privileges can issue commands to reset the parameters stored in an on-site database.

Once the data passes Level-1 validation and are stored in middle database, we adopt a "read-only materialized view replication" process to maintain data consistency between the middle server and database server. Then, we perform Level-2 validation in the database server to ensure other data quality factors such as data completeness, consistency and credibility.

The Environmental Monitoring Centre System (EMC-V2), as shown on the top middle of Figure 3.1(b), operates at the heart of the TAQMN-2. EMC-V2 supplies the user with a robust interface for performing a number of specialized functions. The principles of EMC-V2 are to provide a generic mechanism for accessing air quality data and serving as a focal point for data application developments. A set of software tools have been implemented and operated successfully for data maintenance, data services and application developments. We will demonstrate some of the concrete results in Section 4.

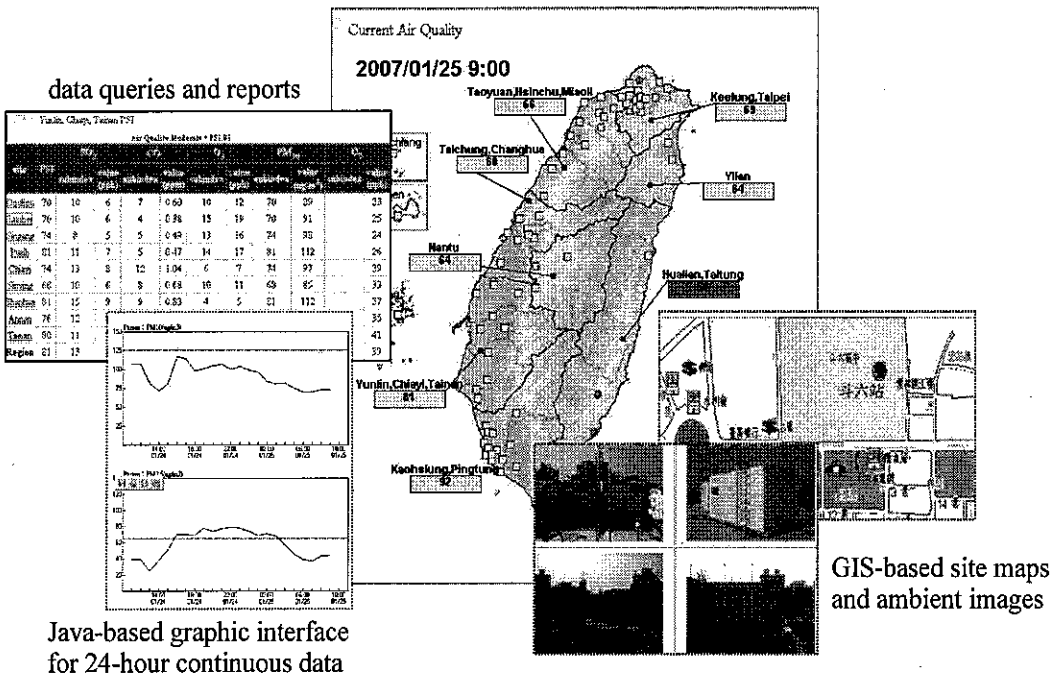
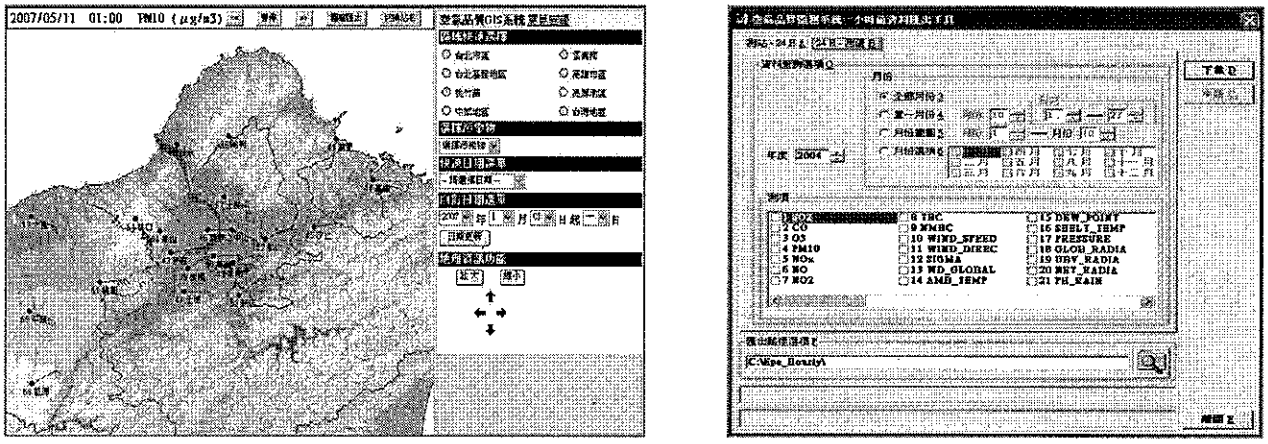


Figure 4.1 A variety of views and displays of TAQMN-2

4. Results and lesson learned

Figure 4.1 shows some results produced by the TAQMN-2.[2] The country map shown in Figure 2 displays the current PSI value of each region. The data is updated hourly. The user can select a region to view detailed data. For example, when the region of Yunlin-Chiayi-Tainan is selected for viewing, all detailed measurements of the stations in the region are displayed as shown on the left side of Figure 2. This resembles the results of a data query or a report requirement. When user selects a specific monitoring station, the past 24 hours of data will be displayed using a Java-based graphic interface. In addition, the location of each station and the images for the surrounding status can also be displayed using a GIS-based toolkit.



(a) GIS with wind speed and wind direction (b) an interface for data access

Figure 4.2 Examples of the data services and applications

A JAVA-based GIS application has been implemented to support the decision-making of air quality forecasting. As shown in Figure 4.2(a), we integrate the GIS maps with measured meteorological data such as wind direction and wind speed to visualize the weather conditions in different areas. This is very helpful for the daily work of air quality forecasting.

Since all time series data are stored in the database server and managed by the EMC-V2, by using the interface shown in Figure 4.2(b), data can be aggregated over time, either by specifying a defined period or by using time related grouping functions such as monthly, weekly and so on. Moreover, data can be selected or filtered according to a specified condition (measured value, time period, as well as certain areas and districts). The results for a specified requirement can be shown on screen, printed out as a report or downloaded by users in EXCEL file or text file format depending on user requirements. These capacities make TAQMN-2 a very flexible and robust system, which can satisfy the requirements of a variety of users, ranging from professional to the general public.

5. Conclusions

This paper presented TAQMN-2, a web-based air quality monitoring system designed and implemented nationwide in Taiwan, with functions ranging from data collection to data retrieval.

[2] For details, please visit http://www.epa.gov.tw/taqmn-2/ or http://210.69.101.141/emce/

TAQMN-2 has been implemented based on the proposed multi-layer architecture for environmental monitoring data management.

The main advantages of our approach are that it:

- replaces conventional DAS with an open standards web-based server attached with a PLC system making monitoring work more flexible
- provides a platform for automatically tracking data lineage while monitoring workflow execution
- supports tasks to systematically collect diverse monitoring data
- requires minimal modifications to migrate existing monitoring systems
- improves the scalability of environmental monitoring projects, in terms of both ease of adding and modifying components in each layer

Environmental data management, analysis, communication and evaluation are essential components of environmental characterization and decision making. IT and related technologies such as DBMS, the Internet, and associated Web technologies have become an integrating force for these components. In future developments, we will adopt XML and related technologies such as Web Services SOAP, and UDDI, based on the concept of multi-layer architecture, to construct a loosely couple platform for integrating heterogeneous environmental monitoring data. In addition, some of the data mining methods are considered to be associated with our platform for shaping environmental data to be more useful to support the environmental decision-making.

References

Chu, Y. C., et al, (2002), Integrating Environmental Information through Data Warehousing with Data Quality Assurance, *Proceedings of EnviroInfo'2002,* Vienna, pp. 368-375.

Manel, P., et al, (2004), Designing and building real environmental decision support systems, *Environmental Modeling & Software*, 19 (2004), pp. 857-873.

Pokorny, J., (2006) Database architectures: Current trends and their relationships to environmental data management, *Environmental Modeling & Software*, 21 (2006), pp. 1579-1586.

Schimak, G., (2003) Environmental data managemen and monotoring system UWEDAT, *Environmental Modeling & Software*, 18 (2003), pp. 573-580.

Taiwan EPA, (2004) *Air Quality Monitoring Data Center Implementation Project Report*, EPA-92-L105-02-234. (in Chinese)

Taiwan EPA, (2006) *Enviromental Policy Monthly*, May, 2006

Triantafyllow, A.G. *et al.* (2005) Design of a web-based information system for ambient environmental data, *Journal of Environment Management*, (2005), pp. 1-7.

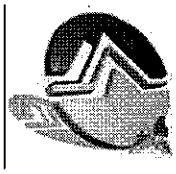
行政院環境保護署
Environmental Protection Administration
Executive Yuan, R.O.C. (Taiwan)

A Multi-layered System Architecture for Environmental Monitoring Data Management Taiwan' Experience

Yu-Chi Chu, Ph.D
Dept. of Environmental Monitoring and Information Management
EPA, TAIWAN

EnviroInfo 2007 Conference
Warsaw, Poland
September 12-14, 2007

Introduction (1/2)

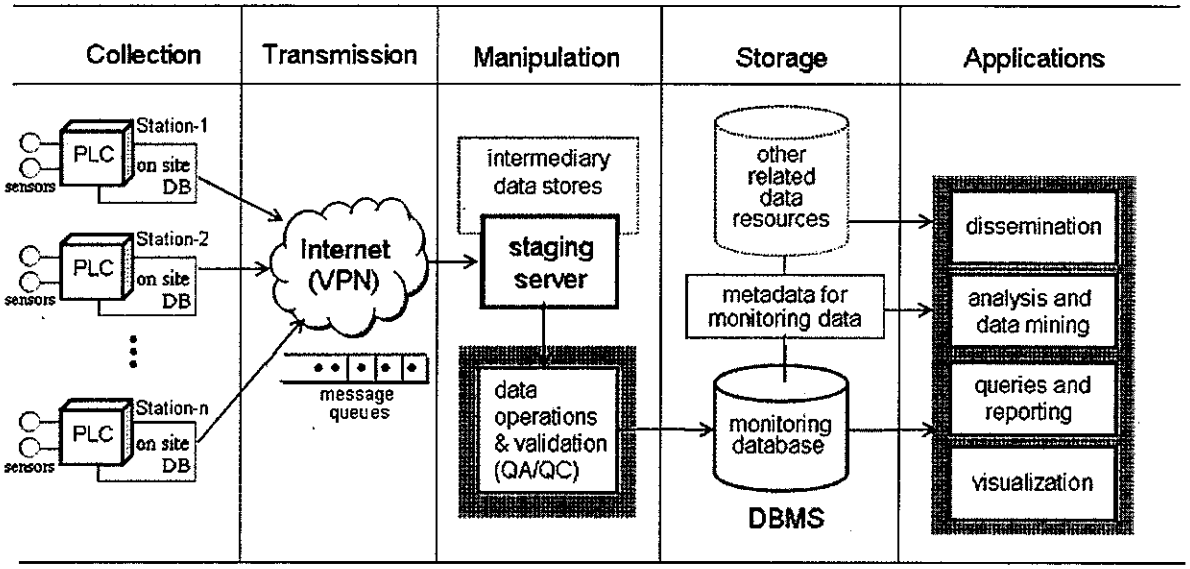


- Environmental monitoring data plays an important role in supporting environmental decision-making
- The task of environmental data management becomes ever more challenge
- We need a comprehensive and integrated system architecture that can **streamline** and consolidate the data management processes for environmental monitoring

Introduction (2/2)

- We present a system architecture
  - using multi-layer approach
  - based on the web-based platform
  - ranging from data collection to applications
  - purpose: **streamlining of data management processes for environmental monitoring**
- Taiwan Air Quality Monitoring Network (TAQMN-2) has been implemented using the proposed system architecture

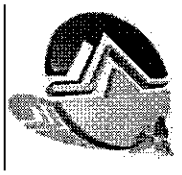
System Architecture (1/6)



System Architecture (2/6)

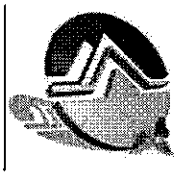
- **Collection layer**
  - comprises a number of apparatuses connecting to a programmable logic controller (PLC), forming the front line for data collection and gathering
  - an on-site database dedicated to storing these data, converting them into XML format
  - adopts a series of **open standards** such as HTTP and TCP/IP for connecting monitoring apparatuses and IT devices

System Architecture (3/6)



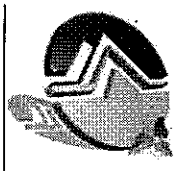
- **Transmission layer**
  - apply virtual private network (VPN) technology which integrates with the comprehensive firewall and router features
    It supports connection for a **secure data transmission** over the Internet
  - use Microsoft Message Queuing (MSMQ) technology as a data transmission platform that bridges the collection layer and manipulation layer

System Architecture (4/6)



- **Manipulation layer**
  - When data is gathered from the monitoring sites, it moves through the **staging server** along with an intermediary store
  - In this area, we examine collected data, perform the various functions for **data quality assurance** and resolve inconsistencies
  - Once the data is finally prepared, it will temporarily reside in the intermediary data stores waiting to be loaded into the monitoring database

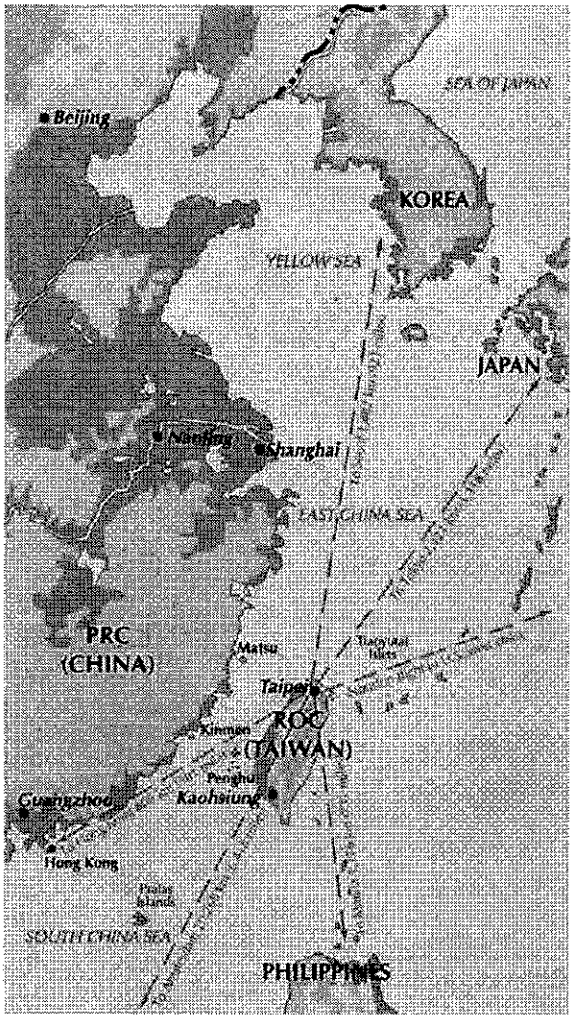
System Architecture (5/6)



- **Storage layer**
  - consists of a commercial database management system (DBMS) along with a set of toolkits to maintain the metadata for monitoring data
  - in order to provide data services more efficiently and widely, this layer may also aggregate data and functionalities from other sources such as the meteorological offices which may provide data on environmental resources in particular weather forecast information

System Architecture (6/6)

- **Application layer**
  - provides a number of applications and services for data retrieval, analysis, as well as data dissemination
  - a variety of advanced applications such as data mining and data visualization can also be attached as part of the systems
  - The user interface for most applications uses Web standards: XHTML and CSS are used for presentation; JavaScript and ASP.NET are used to handle user interaction

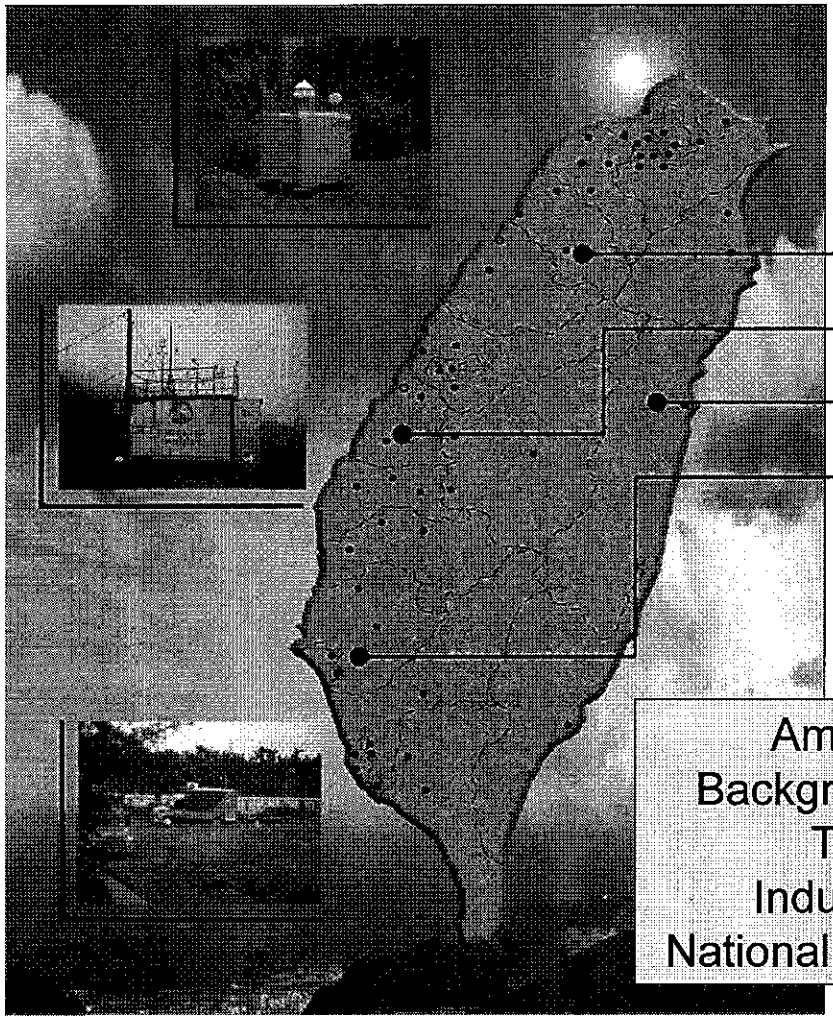


TAIWAN

- Located in the Western Pacific about 160 km off China's southeast coast, midway between Japan and the Philippines
- An area of approximately 36,000 sq. km (about 394 km long and 144 km wide)
- Total population around 23 million people (average density 621 person/km²)

TAQMN-2 Project (1/5)

- **T**aiwan **A**ir **Q**uality **M**onitoring **N**etwork
- an automated and computer aided air quality monitoring system
- Established in 1993 with 66 stations, 2 mobile vans equipped with monitoring facilities, 1 QA/QC Lab
- Expanded to 76 stations in 2006



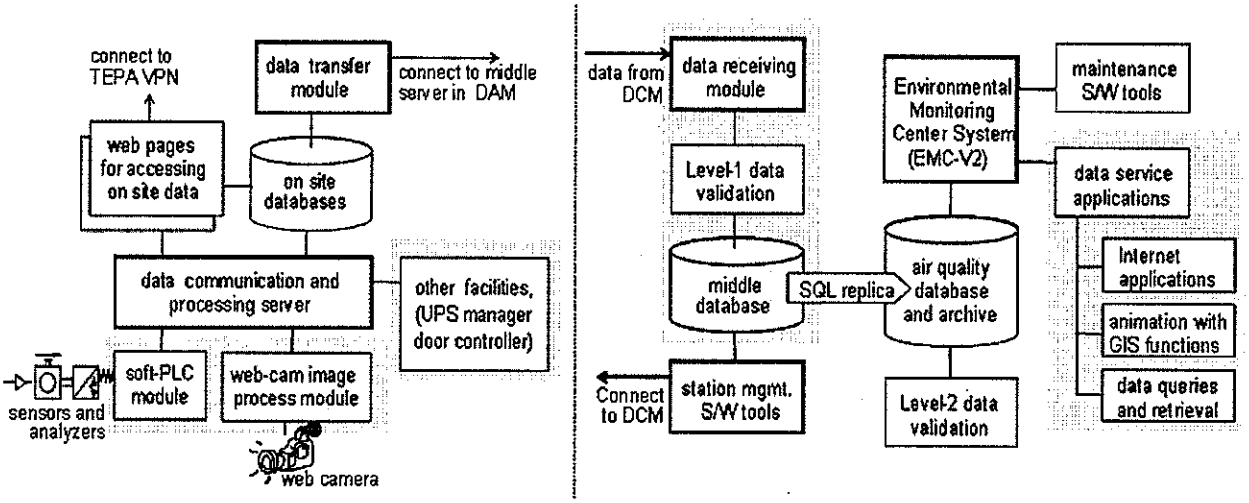
monitoring stations in various locations

Northern	25
Central	18
Eastern	4
Southern	26
	73
Offshore Islands	3

Ambient -	58
Background -	3
Traffic -	7
Industrial -	4
National Park -	4

TAQMN-2 Project (2/5)

- Overall structures



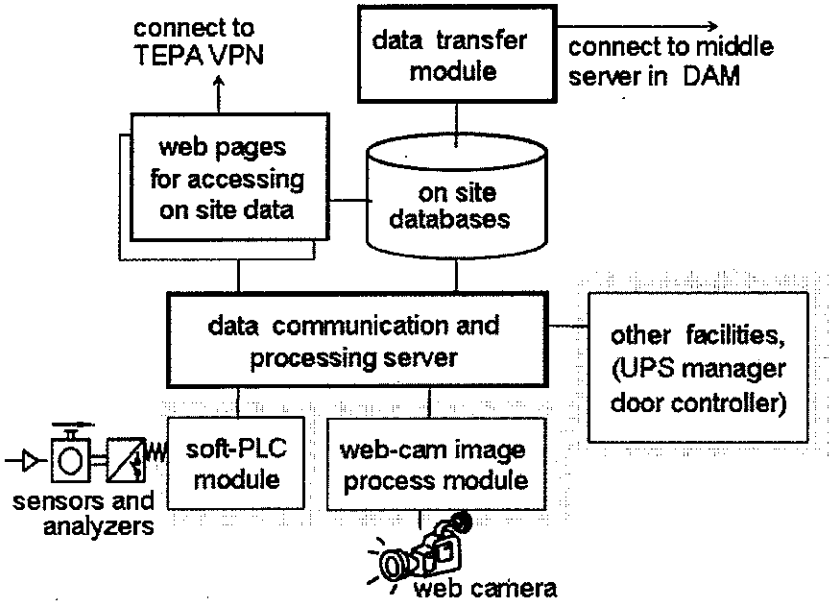
TAQMN-2 Project (3/5)

- Implementation environment

	Data collection manger (DCM)		Data access manager (DAM)	
	gathering and collecting	communication and processing	middle server	database server
H/W	Soft-PLC	Xeon (DP) 2.4GHZ	Xeon MP 1.5GHZ	HP RX-5670 HP MSA-1000
OS	Build-in Kernel	Win2000 Server	Win2000 Server	Win2000 Server
Data storage	Text Files	Oracle 9i DMBS	Oracle 9i DMBS	Oracle 9i DMBS
Development tools	Ladder, Java and C language	MS .Net and Java language	MS .Net, Java and SQL	MS .Net, Java and SQL

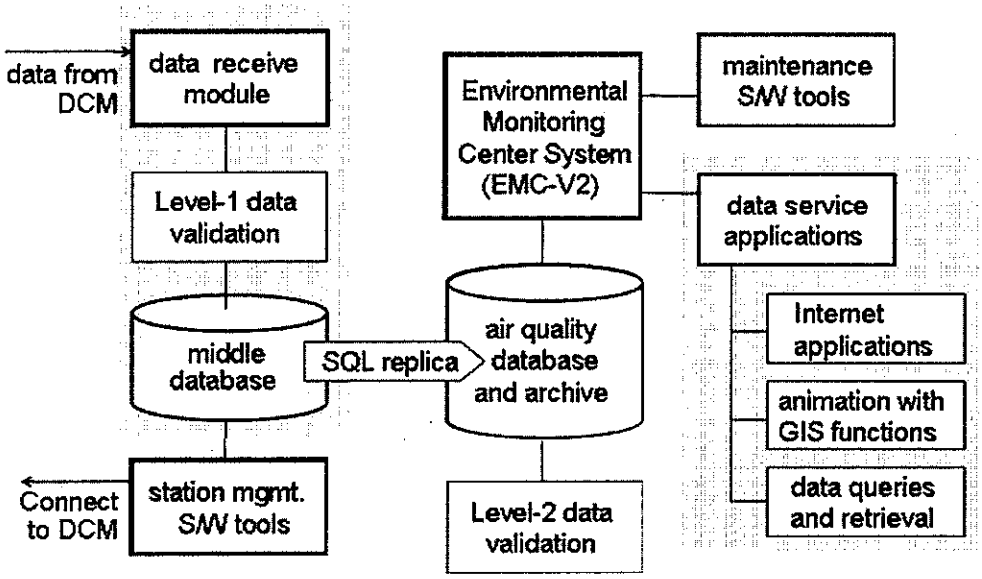
TAQMN-2 Project (4/5)

- data collection manager (DCM)



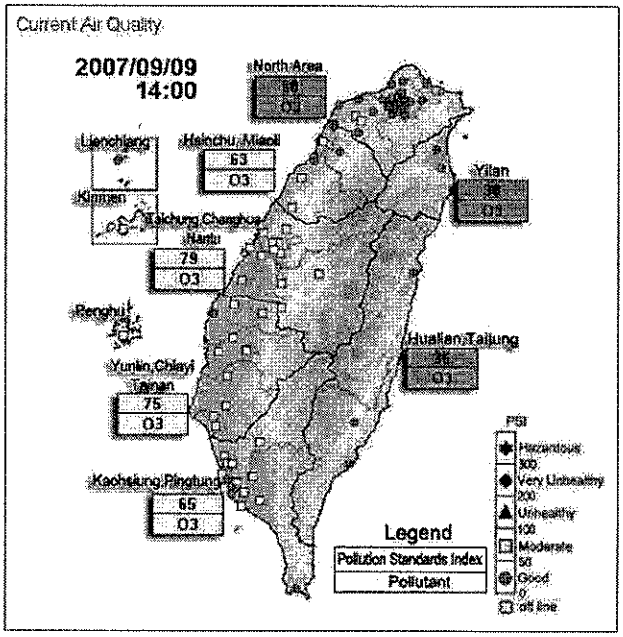
TAQMN-2 Project (5/5)

- data access manager



Results and lesson learned

- The country map displays the current PSI value of each region
- The data is updated hourly
- Users can select a region to view detailed data



http://210.69.101.141/emce

- When the region of Yunlin-Chiayi-Tainan is selected for viewing, all detailed measurements of the stations in the region are displayed

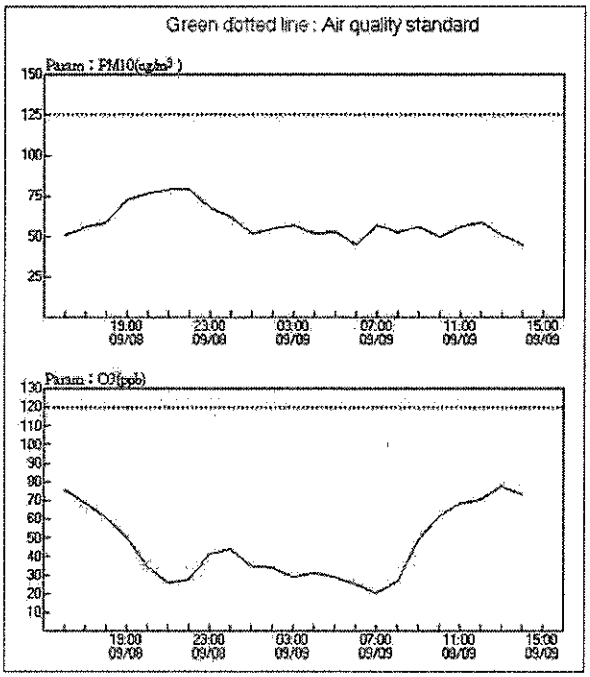
> Home > Air Quality Monitoring Network

Yunlin, Chiayi, Tainan Area PSI

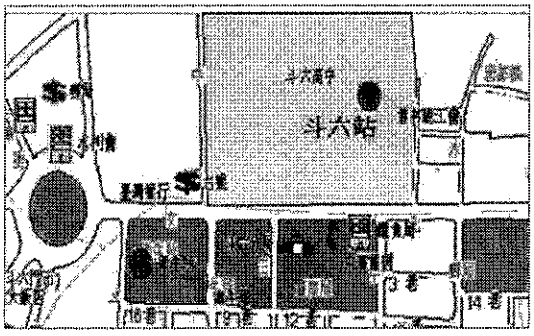
Air Quality:Moderate, PSI:75

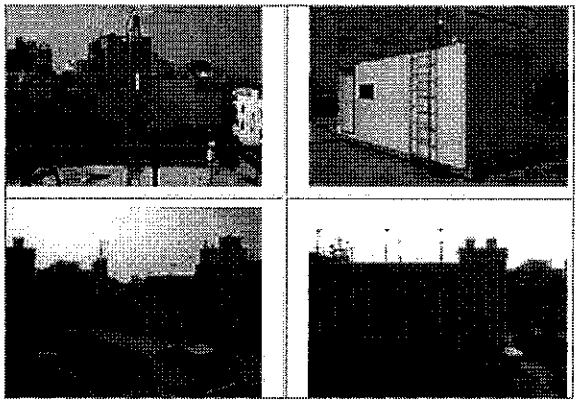
SITE	PSI	SO_2		CO		O_3		PM_{10}		NO_2	
		sub-index	value (ppb)	sub-index	value (ppm)	sub-index	value (ppb)	sub-index	value (ug/m^3)	sub-index	value (ppb)
Douliou	65	6	4	5	0.46	65	78	54	58		10
Lunbei	74	7	4	5	0.42	74	89	62	75		11
Singang	56	5	3	4	0.40	52	63	56	62		10
Puzih	58	7	4	5	0.41	56	67	58	65		12
Chiayi	51	6	3	5	0.45	51	61	51	52		11
Sinying	64	6	4	5	0.45	64	77	58	67		10
Shanhua	69	6	4	4	0.32	69	83	52	55		9
Annan	54	6	3	5	0.48	44	53	54	58		20
Tainan	81	8	5	6	0.53	81	97	51	52		26
Region	75	7	4	5	0.49	75	90	59	69		19

- When user selects a specific monitoring station, the past 24 hours of data will be displayed using a Java-based graphic interface, measurements of the stations in the region are displayed

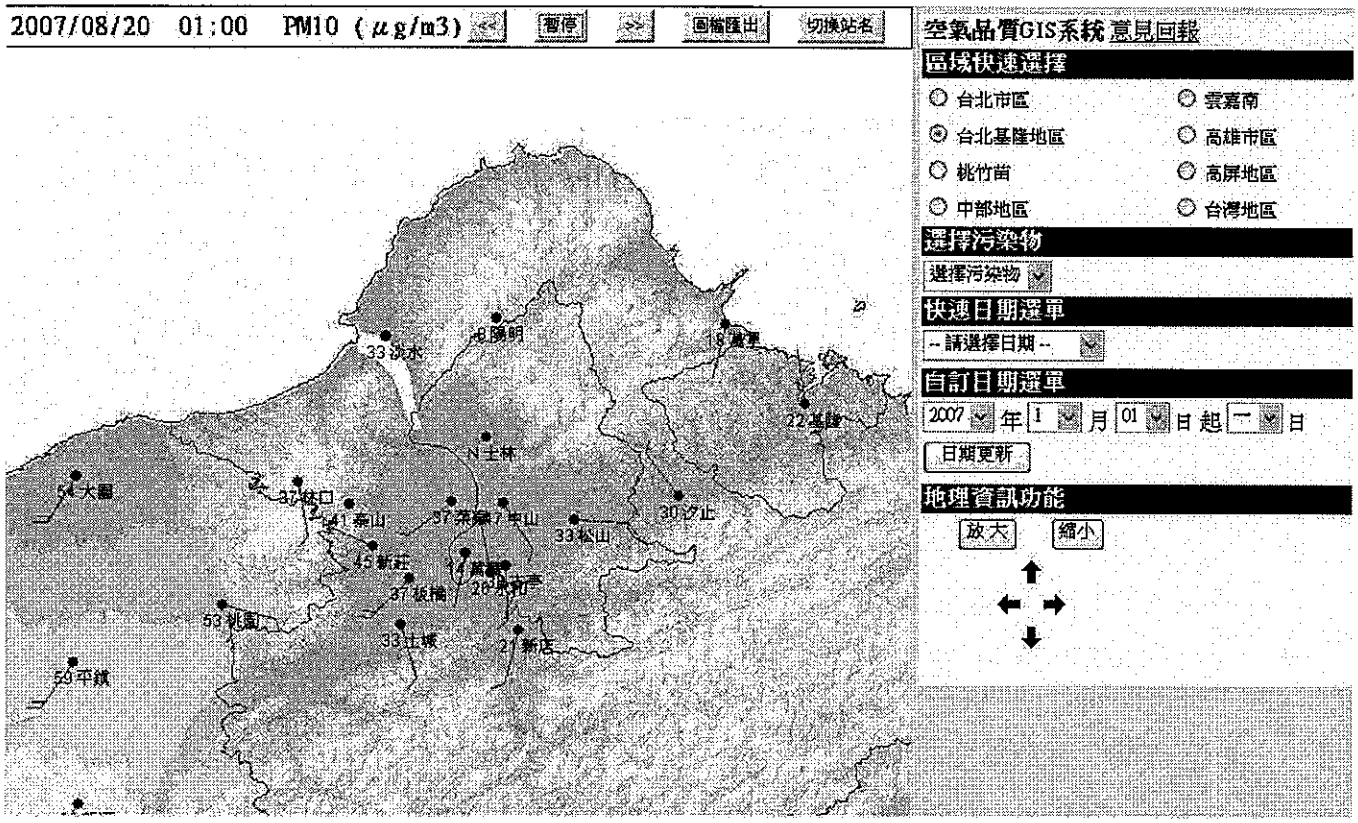


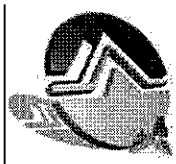
- the location of each station and the images for the surrounding status can also be displayed using a GIS-based toolkit



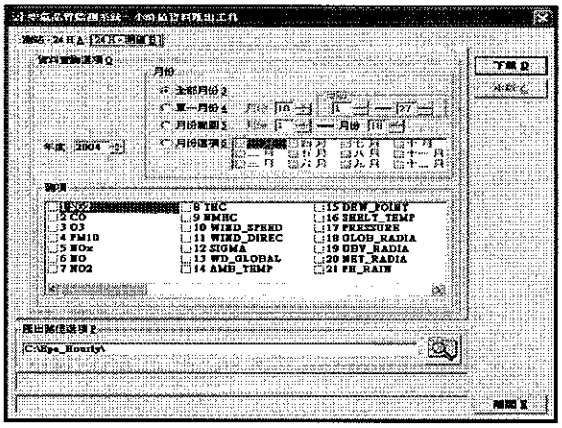


- **GIS with wind speed and wind direction**

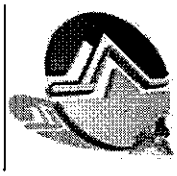




- **An interface for data access**
  - data can be aggregated over time, either by specifying a defined period or by using time related grouping functions such as monthly, weekly and so on
  - data can be selected or filtered according to a specified condition (measured value, time period, as well as certain areas and districts)

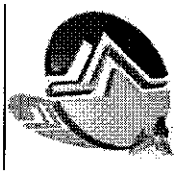


Conclusions



- The main advantages:
  - replaces conventional DAS with an open standards web-based server attached with a PLC system making monitoring work more flexible
  - provides a platform for automatically tracking data lineage while monitoring workflow execution
  - supports tasks to systematically collect diverse monitoring data
  - requires minimal modifications to migrate existing monitoring systems
  - improves the scalability of environmental monitoring projects, in terms of both ease of adding and modifying components in each layer

Future work



- adopt XML and related technologies such as Web Services, SOAP, and UDDI, based on the concept of multi-layer architecture, to construct a **loosely couple platform** for integrating heterogeneous environmental monitoring data
- some of the data mining methods are considered to be associated with our platform for **shaping environmental data** to be more useful to support the environmental decision-making

Thank you!

21st International Conference on Informatics for Environmental Protection
September 12-14, 2007, Warsaw Poland

An Ontology-Driven Approach for Harmonizing and Integrating Environmental Information

Yu-Chi Chu,[1] Su-Mei Huang,[2] Chin-Cheng Lien,[2] Chen-Chau Yang[3]

Abstract

The paper describes an ontology-driven methodology to be used for the integration of heterogeneous environmental information that may be distributed in government agencies. The methodology adopts a series of systematic processes to construct a localized domain ontology which may be used as the driven force or guidelines for environmental information integration. As an example application, the Taiwan EPA's Environmental Data Repository Project (EDR) is used to demonstrate the feasibility and applicability of the proposed methodology.

1. Introduction

For making sensible, justifiable, and legally correct decisions to protect our earth, both government agencies and private sectors need detailed information regarding the current state of the environment and ongoing developments. Currently it is very difficult to share environmental data since the information typically resides on geographically disparate and heterogeneous databases (systems). These systems often do not facilitate access by secondary users and frustrate attempts to draw data together to form a more comprehensive understanding of environmental conditions and actions. Therefore, there is a major demand for appropriate systems and adequate tools to provide integrated information for managing the issues of environmental protection.

The term ontology was originally used in philosophy to describe a theory of "being and existence." In the field of artificial intelligence it was adopted to describe knowledge models that provide definitions of vocabulary used to describe a certain domain. Hence, the use of ontologies for the explication of implicit and hidden knowledge is a possible approach to overcome the problem of semantic heterogeneity. During the past years, ontologies have gained importance in many areas of applications, like interoperability, design of intelligent systems, and lately for knowledge management and electronic commerce. In addition, ontologies are being used for a more precise specification of the semantic information content of the underlying data as well as of the user's information needs. Some domain specific ontologies are also being viewed as vehicles of capturing semantic information content independent of the underlying syntactic and structural representation of the data.

The following section describes the proposed methodology with an example application. The systematic processes for ontology extraction, alignment, as well as integration and enrichment will be illustrated in detail. Section 3 offers some concluding remarks.

[1] Department of Environmental Monitoring and Information Management, Environmental Protection Administration, Taiwan, R.O.C.
[2] Department of Computer Science, SooChow University, Taiwan, R.O.C.
[3] College of Electrical and Computer Engineering, St. John's University, Taiwan, R.O.C.

2. Methodology

For the purpose of integrating environmental information, we want ontologies to be practical and useful artefacts. This means that the effort required to construct new ontologies must be minimized and overall effort required to construct an ontology must be amortized over multiple uses and users. We take several steps towards attaining that goal. First, we provide an environment that assists users to build prototype ontologies by extracting knowledge from the existing information sources such as database schemas or DTD for XML documents. Secondly, the terms of concepts in the built prototype ontologies will align with existing common ontologies and some of the domain specific thesauri such as GEMET. The alignment may provide a semantic consistency among different ontologies. For example, the names of concepts facility-names, plant-names, site-names, and contaminant-sites will be aligned to a term in GEMET called "*site*." After this process, the ontologies will be called aligned ontologies. Thirdly, the aligned ontologies will be merged and combined to form an integrated ontology.

2.1. Systematic processes

The core component of our approach is the integrated ontology, which is dedicated to the domain of environmental protection. Figure 2.1 depicts the process of the proposed method. We explain the main steps of the method we propose for building an integrated ontology from existing knowledge sources such as the environmental databases owned by local governments and EPA at the central government level.

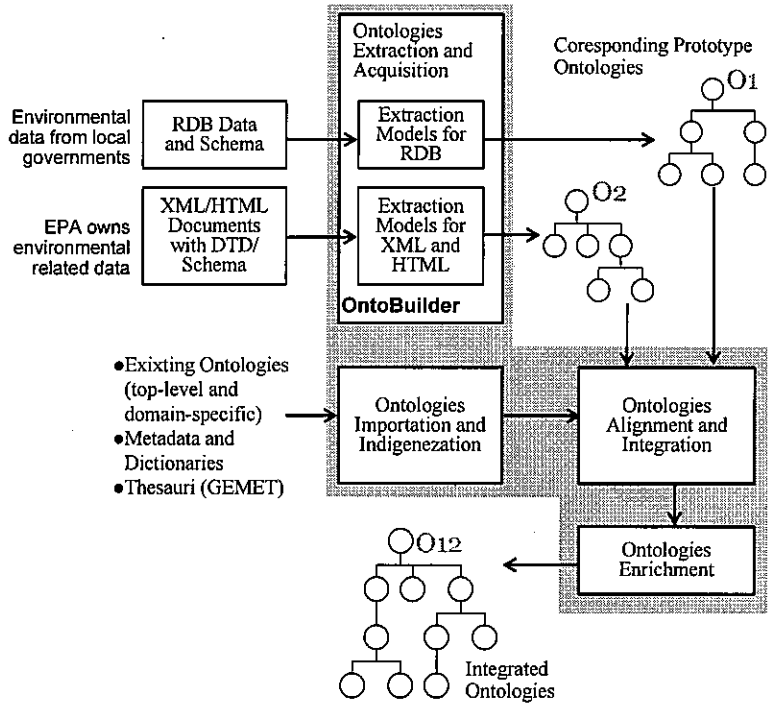


Figure 2.1 The process of the integration of environmental domain ontologies

1. *Selection of underlying information sources, standards, laws and regulations, classifications, etc.* In this step, we select the sources of information that we consider relevant to the

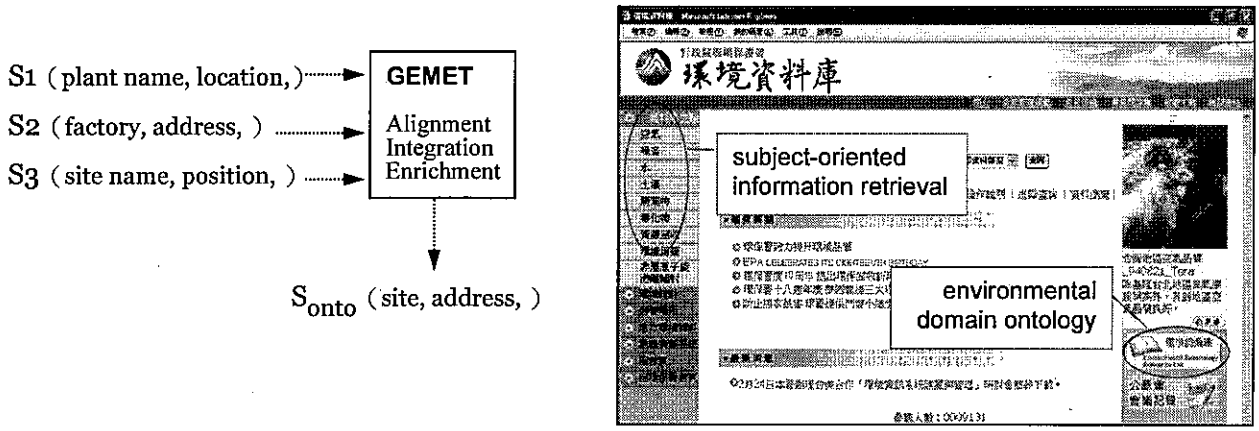
target domain, for example, environmental protection, or e-Business for a certain industry. They usually provide taxonomy of concepts and terminologies used in the domain.

2. *Ontology extraction and acquisition.* This step performs the process of knowledge acquisition from the sources of information previously selected and adapts them to form a prototype ontology for each knowledge source. This activity can be performed using tools such as Ontolingula and WebODE. We have been implementing a simple tool set for this activity called OntoBuilder, which focuses on the relational database schema and XML documents.
   In addition, we also employ the procedure of "protocol analysis" with domain experts. This procedure consists of asking users to describe various types of domain applications, the data used in such applications, and the terms used in their field. Several such sessions of protocol analysis may result in a standard set of terms and inter-relationships among these terms.
3. *Ontology importation and indigenization.* In this step, we import ontologies that are existing ontologies in target domains but might usually be used in other countries or areas. For example, in the domain of environmental protection, the OECD[4] countries commonly adopt a classification, which can be seen as an ontology for identification of environmental industry. Most countries in European area use GEMET, which is a thesaurus for unifying the terminology in environmental use. Those ontologies are helpful to construct new ontologies. However, when importing those ontologies, we need to tailor them to fit the feature in the area or country where ontologies will be applied. For example, if we want to adopt the OECD classification for use in Taiwan, we have to delete some of the items in the classification such as forest management, reforestation because they are not under the jurisdiction of the Environmental Protection Administration in Taiwan.
4. *Ontology alignment and integration.* This activity consists of two processes. First, align the prototype ontologies with imported ontologies or upper level ontologies. This process may include adjusting the name or terminology in the prototype ontologies and making them consistent with each other. Secondly, we combine and merge the prototype ontologies to form an integrated ontology.
5. *Ontology enrichment.* Traditionally, most of the ontologies merely represent taxonomy of concepts, where others may just include some attributes for them. In this activity, we will enrich the integrated ontologies with extra information where available.

2.2. Example application

We use the Environmental Data Repository Project (EDR)[5] of the Taiwan Environmental Protection Administration (TEPA) as a pragmatic example to illustrate the process of implementation. EDR is an integrated data warehouse system that provides a single point of access to data extracted from several major TEPA databases, including the Air Pollution Control System, the Water Permit Database, the Hazardous Waste Control System, and the Toxic Release Database. We construct the integrated ontology by extracting the domain knowledge from some of the information sources and aligning the concepts in the ontology with the laws and regulations of Taiwan EPA. The ontology becomes a major component to drive the integration process of the EDR project. Based on the implementation of a prototype system of EDR project, we may justify the applicability of our approach.

[4] Organization for Economic Co-operation and Development
[5] http://edb.epa.gov.tw



(a) An example of term integration (b) EDR homepage

Figure 2.2 Example application

As shown in Figure 2.2(a), each information source might maintain the data regarding the potential pollution sites. However, they do not use the same database schema and cause a conflict between the terms that is used to identify the sites. Ontology can assist us in overcoming this conflict, and develop a consistent view through which information can be integrated. Figure 2.2(b) shows the EDR homepage, which we have implemented based on the proposed methodology.

3. Conclusions

This paper presents a methodology based on ontology-driven related technologies to integrate environmental information. It is shown, to some degree, that ontology can provide assistance in solving the heterogeneous problems among diverse information sources. Our approach may serve as an infrastructure component for integrating environmental data with known, but differing, collections of data. As for future work, recent advancements including Web services, Ajax and knowledge management might be integrated with the proposed approach to design and implement a more sophisticated and practical system.

References

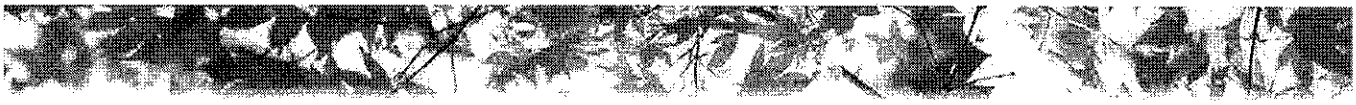
Chou, K. W., (2007) An ontology-based knowledge management system for flow and water quality modeling, *Advances in Engineering Software*, 38 (2007), pp. 172-181.

Chu, Y. C., (2001) *Integrating Heterogeneous Information Sources through Ontology-Driven Model and Data Quality Analysis*, Ph.D. Dissertation, National Taiwan University of Science and Technology.

Chu, Y. C., S. M. Huang, C. C. Yang, (2005) Ontology-based government information integration, *Proceedings of National Computer Symposium* (NCS 2005), Tainan, Taiwan (in Chinese)

Europen Environment Agency, GEMET web pages, http://www.eionet.europa.eu/gemet

Purvis, M. et al., A multi-agent system for the integration of distributed environmental information, *Environmental Modeling & Software*, 18 (2003), pp. 565-572.



行政院環境保護署
Environmental Protection Administration
Executive Yuan, R.O.C. (Taiwan)

An Ontology-Driven Approach for Harmonizing and Integrating Environmental Information

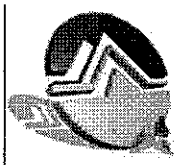
Yu-Chi Chu Dept. of Environmental Monitoring and Information Management, **EPA, TAIWAN**

Su-Mei Huang
Chin-Cheng Lien Dept. of Computer Science, SooChow University, **TAIWAN**

Chen-Chau Yang College of Electrical and Computer Engineering, St. John University, **TAIWAN**

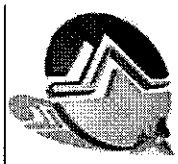
EnviroInfo 2007 Conference
Warsaw, Poland
September 12-14, 2007

Introduction



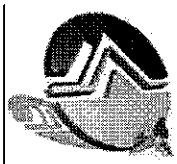
- Government agencies and private sectors need detailed information regarding the current state of the environment
- It is very difficult to share environmental data
  - resides on disparate databases
  - heterogeneity (**syntax** and semantics)
- There is a demand for appropriate systems to provide integrated environmental information

Heterogeneity issues



- **Structural heterogeneity**
  means that different information systems store their data in different structures.
- **Semantic heterogeneity**
  considers the **content of** an information item and its intended meaning.

Example -1



S_1
- Factory(facId, name, address, ...)
- Permit(facID, permitNo, description, ...)

S_2
- Plant(serialNo, plant-name, Plant-location, ...)

S_3

```
<Site>
    <Name> ... </Name>
    <Location>
        <latitude> ... </latitude>
        <longitude> ... </longitude>
    <Location>
    <WasteItem> ... </WasteItem>
        .
        .
</Site>
```

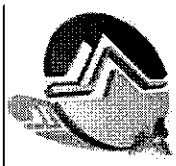
- the offices in EPA do not all use the same database schema
- with the growing amount of data on the Internet, we are facing with data that is not well designed but with little structure such as HTML pages and XML documents

Example -2

Ontology

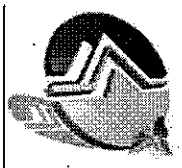
- Ontologies have gained importance in many areas of applications
  - interoperability, knowledge management
  - design of intelligent systems
- Used for a more precise specification of the semantic information content of the underlying data
- Domain specific ontologies are also being viewed as vehicles of capturing semantic information content

Methodology (1/2)

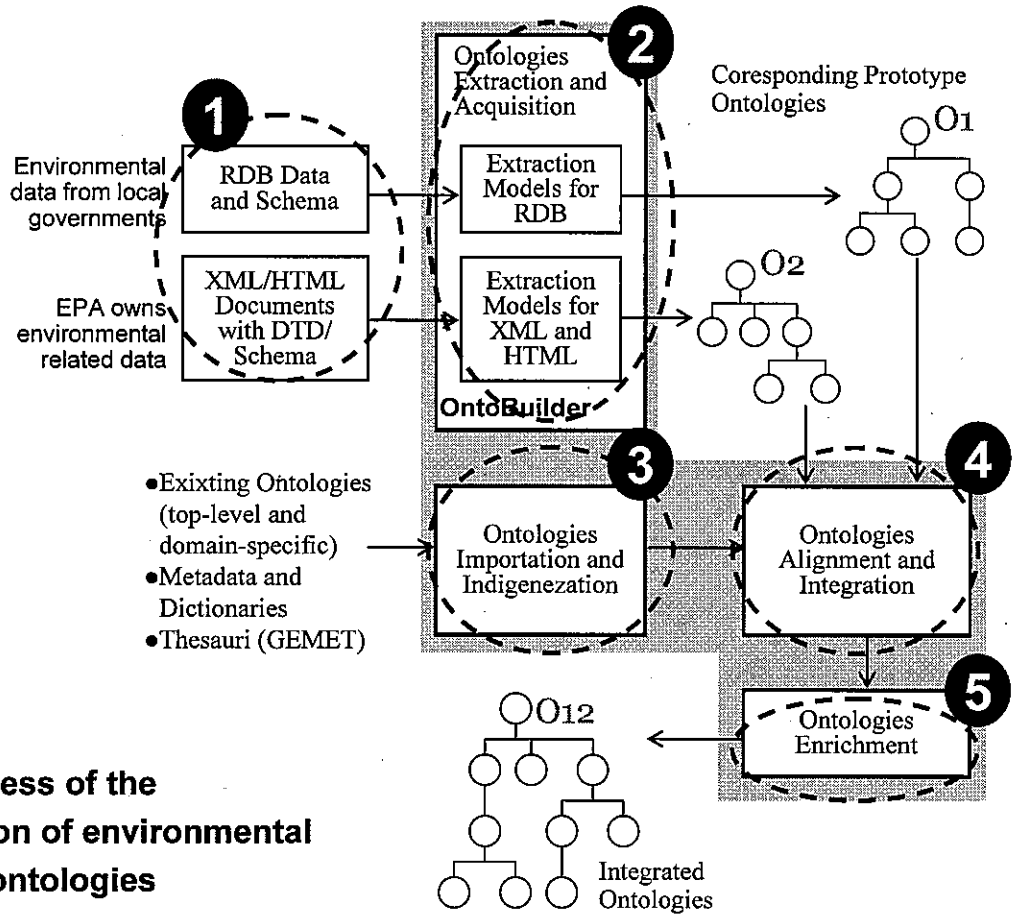


- Purpose and principles
  - ontologies to be practical and useful artefacts
  - the effort required to construct new ontologies must be minimized
  - overall effort required to construct an ontology must be amortized over multiple uses and users

Methodology (2/2)

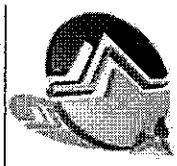


- First, we provide an environment
  - assists users to build **prototype ontologies**
  - extracts knowledge from the existing information sources such as DB schemas or XML documents
- Secondly, for the built prototype ontologies
  - align with existing common ontologies and some of the domain specific thesauri such as GEMET
  - provide a semantic consistency among different ontologies
- Thirdly, the aligned ontologies will be merged and combined to form an **integrated ontology**



The process of the integration of environmental domain ontologies

Systematic processes (1/5)



- Selection of underlying information sources, standards, laws and regulations, classifications, etc.
  - select the sources of information relevant to the target domain
  - usually provide **taxonomy of concepts** and terminologies used in the domain

Systematic processes (2/5)

- Ontology extraction and acquisition
  - knowledge acquisition from the sources of information
  - form a prototype ontology for each knowledge source
  - employ the procedure of "**protocol analysis**" with domain experts
    - asking users to describe various types of domain applications, the data used in such applications, and the terms used in their field

Systematic processes (3/5)

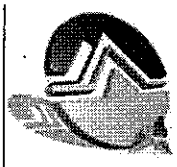
- Ontology importation and indigenization
  - import ontologies that are existing ontologies in target domains but might usually be used in other countries or areas
  - tailor them to fit the feature in the area or country where ontologies will be applied
  - For example:
    - OECD classification
    - GEMET thesaurus

    → localized to fit in other countries or regions

Systematic processes (4/5)

- Ontology alignment and integration
  - align the prototype ontologies with imported ontologies or **upper level ontologies**
    - adjusting the name or terminology in the prototype ontologies
    - making them consistent with each other
  - combine and merge the prototype ontologies to form an integrated ontology

Systematic processes (5/5)

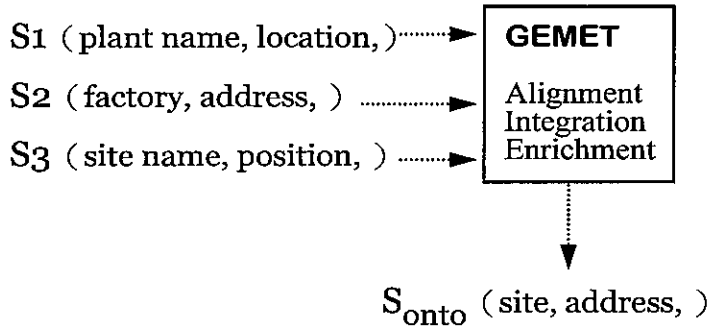


- Ontology enrichment
  - most of the ontologies merely represent taxonomy of concepts, where others may just include some attributes for them.
  - In this activity, we will enrich the integrated ontologies with extra information where available (attributes, features, …)

Example application

- Environmental Data Repository (EDR)
  - an integrated data warehouse system that provides a single point of access to data extracted from several major Taiwan EPA databases
  - construct the integrated ontology by extracting the domain knowledge from some of the information sources and aligning the concepts in the ontology with the laws and regulations of Taiwan EPA

an example of term integration



- each information source might maintain the data regarding the potential pollution sites
- conflict between the terms that is used to identify the sites
- ontology can assist us in overcoming this conflict, and develop a **consistent view** through which information can be integrated

EDR Homepage (http://edb.epa.gov.tw)

Conclusions

- a methodology based on ontology-driven related technologies to integrate environmental information
- It is shown, to some degree, that ontology can provide assistance in solving the heterogeneous problems among diverse information sources
- The approach may serve as an infrastructure component for integrating environmental data with known, but differing, collections of data

Future work

- recent advancements including:
  - Web services,
  - Ajax and Web 2.0 related technologies
  - Knowledge management

  might be integrated with the proposed approach
- design and implement a more sophisticated and practical system

Thank you!