Improving data management and dissemination in web based information systems by semantic enrichment of descriptive data aspects

Steffen Gebhardt\textsuperscript{a,}\textsuperscript{*}, Thilo Wehrmann\textsuperscript{b}, Verena Klinger\textsuperscript{b}, Ingo Schettler\textsuperscript{b}, Juliane Huth\textsuperscript{a}, Claudia Künzer\textsuperscript{b}, Stefan Dech\textsuperscript{b}

\textsuperscript{a}Department of Remote Sensing, Institute of Geography, Julius-Maximilians-University Würzburg, Am Hubland, 97074 Würzburg, Germany

\textsuperscript{b}German Aerospace Center, German Remote Sensing Data Center, Münchener Straße 20, 82234 Wessling, Germany

\textsuperscript{*}Corresponding author at: German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), Münchner Straße 20, 82234 Wessling, Germany. Tel.: +49 8153 28 1715; fax: +49 8153 28 1445. E-mail address: steffen.gebhardt@dlr.de (S. Gebhardt).

\textbf{A B S T R A C T}

The German–Vietnamese water-related information system for the Mekong Delta (WISDOM) project supports business processes in Integrated Water Resources Management in Vietnam. Multiple disciplines bring together earth and ground based observation themes, such as environmental monitoring, water management, demographics, economy, information technology, and infrastructural systems.

This paper introduces the components of the web-based WISDOM system including data, logic and presentation tier. It focuses on the data models upon which the database management system is built, including techniques for tagging or linking metadata with the stored information. The model also uses ordered groupings of spatial, thematic and temporal reference objects to semantically tag datasets to enable fast data retrieval, such as finding all data in a specific administrative unit belonging to a specific theme.

A spatial database extension is employed by the PostgreSQL database. This object-oriented database was chosen over a relational database to tag spatial objects to tabular data, improving the retrieval of census and observational data at regional, provincial, and local areas. While the spatial database hinders processing raster data, a "work-around" was built into WISDOM to permit efficient management of both raster and vector data. The data model also incorporates styling aspects of the spatial datasets through styled layer descriptions (SLD) and web mapping service (WMS) layer specifications, allowing retrieval of rendered maps. Metadata elements of the spatial data are based on the ISO19115 standard. XML structured information of the SLD and metadata are stored in an XML database.

The data models and the data management system are robust for managing the large quantity of spatial objects, sensor observations, census and document data. The operational WISDOM information system prototype contains modules for data management, automatic data integration, and web services for data retrieval, analysis, and distribution. The graphical user interfaces facilitate metadata cataloguing, data warehousing, web sensor data analysis and thematic mapping.

1. Introduction

An Integrated Water Resources Management (IWRM) Decision Support System (DSS) supports operational management and strategic policy-making and planning by monitoring and analyzing the existing situation, as well as the forecasting of future conditions and disasters (e.g. early warning system) (Plate, 2007; Gourbesville, 2008). Recent technologies in information system development enable creation of genuinely open architecture through standard components (Dunfey et al., 2006). These advances in open source technologies and applications in geoinformatics now permit development of spatial data infrastructures (SDI) and internet applications for comprehensive data sources and complex business processes.

Information systems for sustainable and adaptive Integrated Water Resources Management (IWRM) depends on the availability and accessibility of information for system analysis and modelling, social learning, and sustainable decision making (Flügel, 2007). Spatial data and related technologies such as geographical information systems (GIS), global positioning systems (GPS), or remote sensing provide information on abiotic, biotic and socioeconomic conditions, and are proven crucial for effective collaborative decision-making in the IWRM context (Kiehle, 2006; Flügel, 2007; Mansourian et al., 2008). Large amount of observational and modelling data with extensive geographical and temporal

---

0098-3004/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2010.03.010
coverage were collected in the past (McGuire et al., 2008; Schindler and Diepenbroek, 2008), and data processing requirements will continue to increase due to the advance of new sensors and automatic data acquisition tools (Butenuth et al., 2007). For IWRM implementation, sophisticated and scalable data assessment and information management systems are required (Flügel, 2007).

At the same time, data integration and rapid retrieval are hindered by multidisciplinary approaches with large data heterogeneities due to:

- Semantic heterogeneity due to different research disciplines, languages, terminologies, and perspectives (Lutz et al., 2009).
- Structural heterogeneity due to different data types (e.g. spatial data, text documents, graphs, and images).
- Syntactic heterogeneity among geographic datasets due to different native standards and formats for similar data, and usage of monolithic and proprietary systems for geospatial analysis (Bernard et al., 2003).
- Variation in levels of detail for data descriptions (metadata).

As a result, there is a diversity of data content, data models, data acquisitions schemes, as well as spatial data types (Butenuth et al., 2007) that creates a substantial work load of manually reformatting and semantically registering data before exploring, processing and visualising are possible (McGuire et al., 2008). Effective spatial data standards are necessary to establish such a system, and increase the ability to share spatial data and preserve its original meaning (Guptill, 1994). Syntactic and structural transformation approaches adequately handle these heterogeneities, however, they are not adequate for resolving semantic differences (Lin and Ludäscher, 2003). The uses of ontologies are considered a possible solution for semantic heterogeneity problems (Wache et al., 2001; Buccella et al., 2009). In general, data models are expanded to include semantic information either by entity relationships, object oriented links or the use of ontologies, and all of these share a requirement of methodologies or frameworks for semantic matching (Buccella et al., 2009). The fundamental task in data integration is the ability to recognize corresponding information in heterogeneous datasets and to describe the mapping between them, hence the two essential tasks are semantic enrichment and mapping discovery (Sotnykova et al., 2005).

This paper introduces the German–Vietnamese WISDOM project and information system. It summarises the WISDOM information system architecture which comprises the data model, data management, data integration, web services and applications and graphical user interface. We focus on the data model, including structuring and organising heterogeneous data within a common database management system, and how metadata are semantically linked to tabular data and sensor observations, fostering rapid search, and visualisation.

1.1. The WISDOM project

The German–Vietnamese WISDOM project is to design and implement an information system for the Mekong Delta, containing hydrology, sociology, information technology, and earth observation information. Research activities focus mainly on water resources, river systems and water related hazards through mapping products, knowledge acquisition and management of water resources and economic utilization and sustainability, as well as the information system development and data integration. The WISDOM system enables end users to perform analyses on very specific questions, such as regional planning activities, and facilitates integration of the dispersed datasets. As new data is acquired and existing data is updated, the stored data and new data increase in value by providing new information, as well as transforming the presentation of this information into a suitable manner. This integration and continuous access strongly encourages the various organisations to share their data.

1.2. Use cases of WISDOM information system

The overall goal of the WISDOM information system is to integrate its functions and data into existing workflows in the Vietnamese administrative units. The main idea was to establish a widely autonomous data management and query system following operational data flows to minimise end users’ IT skill level and user errors. Two user cases were identified: (i) user based data query (Fig. 1) and (ii) autonomous generation of domain specific information products, e.g. reports (Fig. 2).

Within the first case, users have two roles: both data producers and data users (Fig. 1). Both input their data and develop information products within their business processes, e.g. hydraulic modelling. The information system is a central entry point for data centric works to support users in managing, query and accessing data, and the system facilitates the exchange of all water data types with other partners. To do so data standardisation and uniform interfaces are of utmost importance. Besides

![Diagram for human interface access to data within WISDOM information system.](image-url)
simple data access, users are interested in custom data products including data union and intersection, map compilation and report generation in various formats. These custom product requirements add processing capability specification to the information system.

The second use case shown in Fig. 2 is the system component access to data resources. In creating autonomous products, system data selection logic can be implemented either on the process side or on the data access side. We chose the second approach because applied processes should implement no additional logic beyond data processing functionality, to maintain data reusability. For autonomous system processing, there are additional interfaces required to encapsulate direct data services of semantic enabled access to data resources (e.g., newest precipitation for a limited area, available land cover maps for a specific scale, data representation for a selected media type).

1.3. Data heterogeneities

With the heterogeneity of research themes and partners in the project, the variety of results and data, increase along with an increase in heterogeneity in data content, data types and models. Besides the generation of spatial data such as remote sensing products and model outputs, a wide range of types of documents, schemes, questionnaires and correspondence without a direct spatial basis are included. The data landscape within the project can be categorized as follows:

- spatial data in vector and raster data model;
- remote sensing data and derived products such as land cover classification, water turbidity, inundation mapping, or precipitation data;
- temporal data from sensors such as buoy and other sensors measuring, water levels, water flow, salinity and temperatures;
- various temporal point-based sample measurements of water quality indicators such as pesticide concentrations and endocrine disruptors in waterways;
- forecast model results such as water levels, water flow directions, water inundation areas, and pesticide distributions;
- statistical data within the administrative levels (e.g., national, provincial, or even household level);
- institutional maps by the various organizations working in the water sector.

1.4. System requirements

As an information systems for Integrated Water Resource Management, the system must be structured so that it can generate information from data provided by the various data producers (GWP and INBO, 2009). The platform should be capable of managing broad range of data types described above. The main components include:

- enterprise wide Geographic Information System (GIS) database: these are vital tools for managing and translating spatial data into maps, graphs, indicators and charts;
- on-line tools to manage a catalogue of data descriptions (metadata);
- resource manager decision-support and modelling tools;
- web portal for sharing and disseminating information.

Further requirements were derived for the information system design based on the described data landscape and user desired outputs, especially for data management.

1. Operational data flow: the system must function in an automated fashion capable of integrating and managing heterogeneous data with little or no human interaction.
2. Semantic enabled queries: the system must allow users to explore data using efficient and intelligent search options by theme, geographic and temporal search attributes.
3. End users’ business processes: the system should provide information products to support decision making processes at different administrative scales.
4. Direct access to data: data and information should be distributed to users using common web standards and protocols, such as WMS, WFS, FTP and HTTP.
5. Standard software packages: information products should be distributed in a common file formats, such as PDF, JPG and ASCII documents for loading to a common software.

2. WISDOM system architecture

The information system is designed using a classical, web-enabled multi-tier approach, such as that the data tier is separated from logic tier and presentation tier. Data tier and the middleware communicate using various web services thus enabling existing web resources within the WISDOM information...
system, as well as allowing access to other applications. Business logic is implemented in a Java web application, and thus supports data browsing and access. User can access the information system using standard web browsers.

2.1. **WISDOM Geodata exchange format (WGEF)**

The WGEF standard for exchanging spatial data defines data formats, coordinate reference system, spatial metadata, map styling description, as well as the container format. The specified file formats are ESRI Shapefile for vector and GeoTIFF for raster data. The description of geographic metadata follows the ISO19115 standard. The standard coordinate reference system for project areas is defined as Universal Transverse Mercator (UTM) Zone 48 North with World Geodetic System 1984 (WGS84) ellipsoid and datum. The dataset map styling information is described using the Styled Layer Description (SLD) standard. The WGEF format requires a zip file container of spatial data, metadata, map SLD, and optionally, a preview image of the data.

2.2. **Metadata generator**

The metadata generator is a standalone graphical user interface that uses templates to load data providers’ (project partners) input to valid ISO19115 XML files. This manually input data includes author contact information, and data producers and dataset information such as title, content, abstract, lineage, temporal range, category and keywords. The Extended Markup Language (XML) files are stored in the WGEF zip container.

2.3. **Data Entry Portal (DEP)**

The DEP is a Java application interface for automatic incorporation of spatial data into the management system. This application is triggered by the submission of a new WGEF zip file. In summary the DEP performs the following steps:

- **Analysis**: evaluates whether the WGEF file is complete and valid according to the standard definitions.
- **Cleaning**: automatically generates metadata. Depending on the spatial data model and file format, the appropriate vector or raster information is extracted, including file size, geographic extent and coordinate system. This application merges automatically and manually generated metadata using XSLT transformations. This step also renders thumbnail images using either the submitted SLD information or user generated preview images.
- **Dissemination**: the new dataset is integrated to the data management system for query and access. This step includes registering and storing to the spatial database, file system, and XML database.

The present DEP is a platform independent Java application in command line executable form. The DEP will soon be available as web service within a Web Processing Service (WPS).

2.4. **Data management**

The data management layer comprises the object-relational database management system PostgreSQL, along with its spatial extension PostGIS. This layer realizes all project data, both spatial and non-spatial, in the developed data models for storing and management. XML structured data, e.g. metadata and SLD information, are stored within the open-source XML database eXist. Other file types, including raster data, documents, and images, are distributed through an Apache Webserver.

So far in this paper we have focused on an introduction to data modelling and management components of the WISDOM information system. Detailed descriptions on all the above components are given in the subsequent sections.

2.5. **Services and protocols**

Web services are used for communication between middleware and data tier to add scalability and extensibility. Open Geospatial Consortium (OGC) compliant services are used for spatial data access, especially Web Mapping Service (WMS). Data processing requires direct access to raw data through Web Coverage Service (WCS) for raster data and Web Feature Service (WFS) for vector data. All three services are implemented using Mapserver software by University of Minnesota. Necessary processes, e.g. report generation and data intersection, are encapsulated by Web Processing Service (WPS) as a control interface to link operationally processes and data. Within WISDOM information system all resources are dynamically linked to ensure up to date data access and data distribution over the web. For this reason additional services were implemented based on OGC services to follow representational state transfer (REST) style architecture (Fielding, 2000). These interfaces embed additional logic to directly access certain datasets in a specific format prescribed by user profile. The following resources are currently guided by these logic routines within the WISDOM prototype:

- Search and access by water related organisations.
- Query legal and administrative documents specific to the organisation.
- Access time-series data from measurement stations within the administrated region.
- Generate thematic maps based on statistical and economical data for administrative units.
- Query metadata specific to the organisation or user, e.g. geographical and technical metadata, quick look images, and resource location.

2.6. **Web application**

The WISDOM prototype web application was developed as Struts 2 application on an Apache Tomcat Webserver to communicate with both the database and the data services. Action Servlets are used to control requests and responses between the client (user and web browser) and server. Current WISDOM prototype actions control communication with the database to:

- manage user authorization (login, logout) and overall session management;
- retrieve spatial and thematic reference object hierarchies;
- identify datasets based on administrative region, theme and time parameters;
- retrieve sensor descriptions and positions based on measurement types and time period specifications for thematic mapping.

In addition, servlets are communicating with the REST services to:

- retrieve WMS images for dataset- and thematic mapping;

---

2.7. Graphical user interface

The graphical user interface constitutes of an interactive internet browser application. It is based on the ExtJS\(^4\) JavaScript library. The OpenLayers\(^3\) JavaScript framework has been adopted as map client. The application combines the components of (1) data exploration and access, (2) metadata catalogue, (3) web mapping, (4) sensor measurement exploration, (5) thematic mapping, (6) geo-statistical analysis and (7) yellow pages. Figs. 3 and 4.

3. Modelling different aspects of data

Now, we focus on a detailed description of the developed and implemented data models through an indepth explanation of the various aspects of data modelling and the implementation of the data management system using object-oriented spatial databases and XML databases.

Each dataset is modelled as an entity sharing common descriptive attributes (date of creation, data provider, etc.) and specific, technical data aspect attributes (compare Fig. 5). The data aspect provides more detail, serving as a metadata model. Data aspects used here include (i) geometric aspects of whether the data are raster data or vector data, (ii) data transfer aspects which describe parameters necessary to establish the transfer of datasets using common standards like web mapping services (WMS), (iii) data styling aspects defining the graphical representation of datasets, (iv) tagging and registration of datasets by using geographical and site-specific metadata and (v) data attributes.

Furthermore, the dataset can also be semantically enriched by semantically registering the dataset to (i) spatial reference objects like the geographic area in which the observations or social data are collected, i.e. within a political jurisdiction, making the information more useful to decision makers, (ii) thematic references, linked to simple ontologies realized as hierarchical groupings, which increase the speed of retrieval and accessibility of hydrologic, environmental, or social data, (iii) temporal references which link the datasets to time periods, such as “rainy season” or “day-” or “night-time”. Semantic registration and tagging is necessary for a faster and more “intelligent” access of the data, for both user defined and automatic data queries. In the WISDOM information system, semantic enrichment and mapping datasets to semantic information is so far realised by simple dataset entity relations. Structured hierarchies representing spatial, thematic and temporal references have been created. The mapping of dataset entities to these is done automatically during data ingestion by spatial and time intersection functions and group–theme relations.

3.1. Spatial reference model

The WISDOM system must be capable to store all spatial and non-spatial datasets from the different sources of information such as hydrologic modelling and monitoring, earth observation or census survey. User specifications require the datasets are spatially searchable and retrievable in a quick simple manner. Correspondingly, data should be tagged not only with plain coordinates (centre point or extent), but with even more “meaningful” unit names for end users, such as administrative unit names. The use of administrative names, in turn requires that each spatial dataset is semantically enriched by machine-readable metadata that can delineate the spatial region bounding or intersecting field observations (e.g. administrative areas). Furthermore, the WISDOM information system must accommodate processing point observations, such as hydrologic time series from sensor measurements, and integrate multi-temporal statistical data, as in the case of census data. The system should also be flexible in allowing changes in the territorial boundaries of administrative areas over time, i.e. changes in the reference geometries over time. In short, the system must:

- accommodate unique “spatial reference objects” of either point, line, or polygon type,
- assure non-redundant management of administrative “spatial reference objects” having capabilities of “is-part-of/consist-of” properties, representing provincial, municipal, and local levels for different time periods according to different municipal reforms,
- semantic tagging of datasets for cross referencing to the “spatial reference objects”, e.g. administrative units,
allow tagging of non-spatial data to point and polygon reference objects, such as measurements to sensor point locations and census data to administrative boundary polygons.

The concept of “spatial reference objects” is realized within a spatial reference table. This table stores the actual geometry of the reference objects, either as point, line or polygon type, as also names, hierarchical level and unique identifiers for each of the objects. Currently, this table manages the administrative units for each level and the sensor observation web locations within the Mekong Delta.

The implemented administrative scheme is structured according to administrative levels, building a hierarchical classification, indicating that one or more objects are related or within another object. Starting, for example, at the national level is subdivided into regional objects, which consist of provinces in the next lower level. The provinces are further subdivided into districts which consist of communal areas. Having established a simple hierarchy reflecting the hierarchical regional structure, we can infer the “is-part-of/consist-of ” relationship between the different levels.
Each regional reference entity is described by unique identifiers, actual geometry, codes and names. The hierarchical structure is defined by assigning a parent to each entity. The parent is the geometric object at the next higher level, which contains the referring child. The concept of spatial reference objects ensures that unique geometric entities are stored separately. For applications such as interactive thematic mapping, these objects are joined with their corresponding attributes within database queries to arrange interrelated entities (e.g. population in districts for 2005). Geometric objects (administrative units) might change over time, and these changes will impact the spatial analyses, such as area normalised population density. In our approach, dated entities are retained in the spatial reference table to preserve the temporal relationship between geometric entity and actual attributes at the time of entry. That means the database stores the administrative regions before and after multiple municipal reforms in Vietnam that occurred over past decades. This temporal data also ensures consistent management of geometric objects with respect to the variability of related attributes over time.

As an example application of this hierarchical structure and the temporal database, we performed a query of specific water sensors within a specific time period to derive the total suspended sediments (Table 2). The sensor network web in the Mekong Delta is equipped with over 100 water quality sampling stations and hydrological sensors which monitor water levels, water discharge, suspended sediment, temperature and electric conductivity throughout the flood season. Similar to administrative reference objects, individual sensors are stored by coordinate location, name and codes in the spatial reference table (see Table 1). Physical measurements are stored as attributes of the sensor, and related through the sensor point objects id. Every integrated measurement belongs to the sensor and attendant module (i.e. description of measurement type). For example, sensor T7 at object id 146 collects high temporal resolution data for total suspended sediments (TSS), electric conductivity and temperature.

With this design, measurements by sensor are simply filtered with specific time and module parameters such as the following SQL statement for the results shown in Table 2:

```sql
SELECT DISTINCT m.date, m.sensor, m.value, mod.name FROM vector.spatial_reference AS v, core.sensor_measurement AS m, core.measurement_module AS mod WHERE m.sensor = s.id AND m.module = mod.id AND mod.name = "tss" AND s.id = 146 AND m.date > '2008-06-09 16:00:00.0' AND m.date < '2008-06-09 19:00:00.0' GROUP BY m.sensor, m.value, m.date, mod.name;
```

This approach allows intersecting spatial reference objects of multiple categories. One can think of integrating hydrological vector geometries which represent the upper Mekong River versus the Lower Mekong catchment areas. Every spatial dataset is registered to multiple objects of the various categories (e.g., administrative unit, catchment area). Cross referencing can be implemented using further cross tables offering many-to-many (n:m) relationships. The proposed referencing approach speeds all spatial queries because the time consuming spatial functions of intersects and contains relationships are called only once during dataset registration. Even complex geometries (e.g. water masks with ten of thousands small, distributed polygons) are efficiently accessed.

### 3.2. Thematic reference model

Ontology-enabled semantic data integration enables easier user-access to the desired information and speeds data processing. Similar to the Spatial Model hierarchy (above), thematic groupings in ontology-like order within the WISDOM information system operate within three levels. Each entity is defined by a unique identifier, code and name. The highest level comprises five main topics: (i) environment, (ii) infrastructure, (iii) demographics (i.e., population and economy), (iv) knowledge management, and (v) water management. In keeping with the properties of ontology, each of these topics was divided into finer subdivisions, following a class-like structure. For example, the topic “Environment” contains the child “Land cover and land use” which is composed of finer categories, of which one grandchild is “Agriculture”. Simple parent–child relationships are used to define the hierarchy of the entities including all characteristics of a hierarchy like inclusiveness, inheritance and transivity (Kwasnik, 1999). Table 3 exemplarily demonstrates the structure of the applied thematic model.

### 3.3. Temporal reference model

Temporal attributes consist of (i) static windows, representing annual patterns and planning cycles, e.g. the seasons and crop
3.4. Dataset entity modelling

A dataset is an entity having common descriptive attributes, which is further described by the specific data aspect attributes. The common attributes are the least subset of core attributes described above, while the data aspect attributes are divided into: (i) temporal, spatial, (ii) thematic, (iii) geometric, (iv) Web Map Service and layer, (v) graphical and data style (vi) graphical metadata, and (vii) geographical metadata aspects.

In our implementation, a Universally Unique Identifier (UUID) and unique ID assure distinct access to each dataset. The reference to the data provider points to a table of organisations of data providers that provides name, address and the contact information.

3.4.1. Temporal reference aspect

A dataset is defined by the date created and the time range over which it is valid. As noted above, the time range can be static or dynamic. For example the value for the population density for the year 2004 is considered valid during the whole year, while water masks and rainfall estimates are only valid for the day of acquisition, respective of the observation for that specific moment only. A soil map may not necessarily have a set date for an updated issuance of the map, in which case the ending year is infinite. We assume that new observations will update the old. Furthermore, a dataset entity is registered against the temporal scheme by cross referencing. A dataset can hence be related to e.g. “autumn”, “flood season” and “field data collection 2008” at the same time.

3.4.2. Spatial reference aspect

Every dataset makes reference to at least one object in the spatial reference table through their respective ID. Table 4 provides an illustration of registering a dataset within the administrative regions in which it belongs, in this case the dataset with ID 403 is referenced at district, provincial and regional administrative levels and the ID allows this dataset to be queried through the regional “label” search parameter.

Referencing of a dataset to the respective “spatial reference objects” is done automatically during the data integration process of the Data Entry Portal (DEP). Therefore, the spatial extent of the dataset is spatially intersected with all administrative spatial reference objects. Datasets are identified efficiently on the fly and do not require time intensive intersection at runtime when users define web request. The coordinate reference system (CRS) for the dataset is defined by the European Petroleum Survey Group (EPSG) code.

<table>
<thead>
<tr>
<th>Dataset id</th>
<th>Dataset uid</th>
<th>Reference entity id</th>
<th>Reference name</th>
</tr>
</thead>
<tbody>
<tr>
<td>403</td>
<td>b33c503f-4038-44b7-9232-1c99065041de</td>
<td>843</td>
<td>H. Châu Phú</td>
</tr>
<tr>
<td>403</td>
<td>b33c503f-4038-44b7-9232-1c99065041de</td>
<td>333</td>
<td>Mekong Delta</td>
</tr>
<tr>
<td>403</td>
<td>b33c503f-4038-44b7-9232-1c99065041de</td>
<td>632</td>
<td>Bac Lieu</td>
</tr>
<tr>
<td>403</td>
<td>b33c503f-4038-44b7-9232-1c99065041de</td>
<td>631</td>
<td>An Giang</td>
</tr>
<tr>
<td>404</td>
<td>4ff0e48b-164a-4696-af16-21d8f082e79</td>
<td>333</td>
<td>Mekong Delta</td>
</tr>
<tr>
<td>404</td>
<td>4ff0e48b-164a-4696-af16-21d8f082e79</td>
<td>631</td>
<td>An Giang</td>
</tr>
</tbody>
</table>

3.4.3. Thematic reference aspect

Each individual spatial dataset belongs to a specified group (e.g. watermask, water turbidity, and land cover), and these are registered within the thematic scheme. These dataset groups are managed through a single table, using an n:m cross reference between groups and thematic scheme, as shown in Table 5. The cross reference registers each dataset to its respective themes at different levels, for example a “watermask” belongs to “Environment” at the highest level, “Hydrology” in the next level and, finally to “Water level” at the lowest level. Also, spatial datasets are related to multiple themes within the same thematic level, e.g., the “River network” which belongs to “Environment”, and “Infrastructure” at the highest level.

The benefits of this modelling approach are (i) adding thematic contextual information to spatial datasets using hierarchies, (ii) enabling dataset queries through meaningful thematic groups, and (iii) maximizing the users’ ability to browse the data catalogue using thematic groups. The thematic hierarchy was developed in conjunction with all project partners for their data in respect of resulting structures.

<table>
<thead>
<tr>
<th>Product Id</th>
<th>Product name</th>
<th>Theme Id</th>
<th>Theme name</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>River network</td>
<td>1</td>
<td>Environment</td>
</tr>
<tr>
<td>16</td>
<td>River network</td>
<td>2</td>
<td>Infrastructure</td>
</tr>
<tr>
<td>16</td>
<td>River network</td>
<td>8</td>
<td>Hydrology</td>
</tr>
<tr>
<td>16</td>
<td>River network</td>
<td>10</td>
<td>Transportation</td>
</tr>
<tr>
<td>16</td>
<td>River network</td>
<td>14</td>
<td>Irrigation network</td>
</tr>
<tr>
<td>16</td>
<td>River network</td>
<td>29</td>
<td>Water flow</td>
</tr>
<tr>
<td>17</td>
<td>Watermask</td>
<td>1</td>
<td>Environment</td>
</tr>
<tr>
<td>17</td>
<td>Watermask</td>
<td>8</td>
<td>Hydrology</td>
</tr>
<tr>
<td>17</td>
<td>Watermask</td>
<td>28</td>
<td>Water level</td>
</tr>
</tbody>
</table>

3.4.4. Geometric aspect

Most of the data integrated into the system are spatial originating from field surveys, digitized maps, earth observation, sensor measures, statistical yearbooks, and from pre-processed and analysis of existing data.

The primary advantage of spatial databases lies in the ability to combine GIS and relational database management system (RDBMS) data. However, the trade-offs are in the limitations imposed on the data model, i.e., within a geospatial (GIS) database, the difficulty of integrating models is greater, such as integrating Land Surface Models (for estimating evapotranspiration), some meteorological formats, and algebraic and topological operations to generate two dimensional vectors and three dimensional surfaces. There are several open sources and proprietary database management systems that are extended with spatial management modules, e.g.
PostgreSQL\(^5\) and PostGIS\(^6\), MySQL\(^7\), Oracle Spatial Extension,\(^8\) or IBM DB2 Spatial Extender.\(^9\) In spatially enabled database systems, geometry is handled as a special data type. The OGC defines standard GIS object types as “Simple Features Specification for SQL”, required manipulation functions, and two metadata tables “spatial_ref_syst” and “geometry_columns” (Ryden, 2005). As noted above, the disadvantage of extended RDRM systems is the lack of more comprehensive data models, such as models supporting topology and raster data management. A few proprietary solutions already partially support these features through ArcSDE\(^{10}\), Oracle Spatial and Rasdaman.\(^{11}\)

Our system utilizes the PostgreSQL database with the PostGIS spatial extension. Every spatial vector dataset is registered in the “geometry_columns” table by name (UUID), geometry column, spatial dimensions, vector feature type, and coordinate reference system (CRS). The actual vector datasets are stored in separate tables consisting of geometric coordinates, which form the geometry of elementary features, and columns for the dataset objects attributes. Specific meta-information for describing a vector dataset (e.g. number of objects and attributes, data quality indices, and data scale range) is stored in a “vector_data” table (feature aspect).

Since raster data is currently not supported by the PostGIS extension, information relevant for managing raster datasets (raster aspect) utilizes “raster_data” tables which include URL to raster file storage, raster dimension, number of bands, and raster format. Each entity in this table inherits the dataset by an established foreign key constraint to the “dataset” table. Raster data can originate from remote sensing data products and (GIS) geo-statistical analysis (e.g. interpolations). Those raster data from analyzing remote sensing data, e.g. land cover classification and water turbidity, are further described by source image data, such as satellite sensor and scene acquisition date.

The geographic extent of both raster and vector datasets are stored as simple polygon geometry in the “footprint” table (coverage aspect). Each footprint entity is related to the respective dataset entity using the UUID. The benefits of such a footprint registration are (i) spatial registration of raster data through the same management information as vector data, (ii) uniform registration of dataset geographic locations with polygons, and (iii) through this organisation fast query and dataset identification using the same spatial query parameters.

3.4.5. WMS and layer aspect

The OGC Web Map Service Interface Standard (WMS) provides a simple HTTP interface for requesting geo-registered map images from one or more distributed geospatial databases. The response to the request displays one or more geo-registered images in a client mapping application (de la Beaujardiere, 2006). The OGC WMS specification supports very basic styling options by advertising a preset collection of visual portrayals for each available data set (Lalonde, 2002; de la Beaujardiere, 2006). The OGC Styled Layer Descriptor (SLD) standard is used to portray outputs from Web Map Servers, Web Feature Servers and Web Coverage Servers (Lalonde, 2002). SLD is an XML schema for describing the appearance of map layers and capable of describing the rendering of vector and raster data. With SLD visual parameters (Bertin, 1984) for each map layer component, either vector objects or raster data, are specified.

Each spatial dataset in the WISDOM information system, either raster or vector, is defined by a SLD specification. The information on WMS services and layers stored in the current WISDOM information system dynamically supports different WMS services. Different interfaces were developed with the Mapserver software to transform the service, layer and dataset information from the database to valid WMS service configurations which are then accessible via web or specific WMS clients. A requested layer is represented using the defined SLD symbology and with its specific metadata and query attributes.

3.4.6. Graphical and data style aspect

The OGC Web Map Service (WMS) specification supports very basic styling options by advertising a preset collection of visual portrayals for each available data set (Lalonde, 2002; de la Beaujardiere, 2006). The OGC Styled Layer Descriptor (SLD) standard is used to portray outputs from Web Map Servers, Web Feature Servers and Web Coverage Servers (Lalonde, 2002). SLD is an XML schema for describing the appearance of map layers and capable of describing the rendering of vector and raster data. With SLD visual parameters (Bertin, 1984) for each map layer component, either vector objects or raster data, are specified.

Each spatial dataset in the WISDOM information system, either raster or vector, is defined by a SLD specification. The information on WMS services and layers stored in the current WISDOM information system dynamically supports different WMS services. Different interfaces were developed with the Mapserver software to transform the service, layer and dataset information from the database to valid WMS service configurations which are then accessible via web or specific WMS clients. A requested layer is represented using the defined SLD symbology and with its specific metadata and query attributes.
an easy to use graphical XML editor called “WISDOM Metadata Generator” that quickly develops valid ISO19115 metadata.

Remaining metadata not assigned manually are subsequently extracted by the Data Entry Portal. Depending on the spatial data model and file format respective vector or raster information are extracted regarding file size, geographic extent and coordinate system, as mentioned above. The manual and automated metadata elements are merged using XSLT transformations to complete metadata XML files which are then submitted to the DBMS. To date two approaches are incorporated for managing metadata in the WISDOM model. First is the adoption of the Geonetwork\textsuperscript{12} database model to the WISDOM dataset model. The second approach manages all the metadata XML in the native XML database eXist. The reason for both are: (a) efficient query and manipulate of XML files using XQueries in eXist database, and (b) reduced conflicts and contingences by the simple adoption a complete OGC CSW compliant metadata catalogue application.

4. Results and examples

At the date of paper submission, the system was implemented as a prototype with as the system architecture and all components described in the introduction section. The system is available at the WISDOM project website.\textsuperscript{13}

4.1. Automatic integration of spatial data

Through the Data Entry Portal approximately 500 spatial datasets of vector and raster data have been integrated into the database with the automated assist from 20 different providers covering 23 different dataset groups.

4.2. Data identification

Using the developed graphical user interface datasets can then be immediately queried as presented in Fig. 4. On the left hand side the user defines his regional, thematic and temporal request parameters, where the tree representation of the regions and themes are dynamically created out of the spatial and thematic reference objects in the database. Submitting the query calls the respective Struts 2 servlet action and forwards the request as translated SQL statement to the database. The following example shows the generated SQL for a user request for datasets on hydrology for the “Can Tho” region which are available in 2008. The response of the servlet action is received by the central GUI tabular element which immediately queried as presented in Fig. 4. On the left hand side the user interactively defined the theme of the map, and subsequently generates a dynamic WMS the same as for the dataset mapping. Again the response is the GetMap request results directly send to the map client of the web GUI.

4.3. Metadata extraction

At the same time the user selected a table row, another action executed a REST service based on the selected dataset UUID retrieved from the ISO19115 metadata. The response provided filtered information on title, abstract, generation, credits, contact, download and preview image drawn to right hand side information window in the GUI (Fig. 6).

4.4. Dynamic WMS generation

After selecting the dataset, the spatial data was added to the map client of the web GUI. By doing so a REST service is called which based on the respective UUID of the dataset generates a WMS service definition on the fly and delivers the rendered map layer as WMS GetMap request to the map client (see Fig. 4). Therefore the REST service queries the database for the required WMS layer information for the respective dataset and also the WMS service definitions, WMS metadata and the SLD style information from the eXist database. This service can also be called from other WMS client software, hence the user is not necessarily restricted to the web GUI for data visualisation.

4.5. Sensor measurement visualisation

Currently, 83 sensor locations are managed as spatial reference objects. About 500,000 sensor measurements from the 1980s until today are related to those. A REST service receives parameters of sensor id, thematic measurement, timeframe, and output type specifications as URL (e.g. http://mekongvm2:8181/hydrostation/1136/51/01-01-85/01-01-95?type=li ne&output=image&width=1200&height=900). The result is an image with the sensor measurement representation as image. The definition of request parameters is done within the “Sensor Toolbox” of the web GUI, thereby the actual REST service call is executed by clicking on the sensor position in the map. The resulting image is visualised as popup for the respective sensor position (see Fig. 4).

4.6. Thematic mapping

Around 250 statistical datasets comprising around 9000 measurements from statistical yearbooks were integrated and related to spatial reference objects at country, regional, provincial and district level. With the “Thematic Mapping Toolbox” of the web GUI the user interactively defined the theme of the map, and these parameters defined the type and date of thematic data (e.g. Population Density, 2005) from the available data. When submitting a request, a REST service was called to join the thematic data to their respective spatial objects on database level and subsequently generates a dynamic WMS the same as for the dataset mapping. Again the response is the GetMap request results directly send to the map client of the web GUI.

5. Discussion

All project data, output data and graphics described were managed using the WISDOM system. The presented data models and the database implementation are robust for managing various spatial datasets and their distinct aspects as demonstrated through numerous tests on data integration (approximately 500 spatial datasets) and as many user requests via the web client. Many spatial datasets, in both vector and raster format as well as derived from statistical yearbooks and sensor and GPS measurements were integrated in the database.

Implementing the management of different data aspects advanced data query and data distribution algorithms. For example spatial datasets were searched by ISO19115 metadata using REST services and Geonetwork catalogue system. Further-

\textsuperscript{12} http://geonetwork-opensource.org/, last verified March 15, 2010
\textsuperscript{13} http://www.wisdom.caf.dlr.de
more, WISDOM system developed REST services allow for retrieval of these data aspects by mapping database attributes to URL based web resources, and serve all necessary information via common simple interfaces. Through geographical metadata and other data aspects, each dataset was retrieved as a WMS layer using common web or desktop clients such as OpenLayers, Gaia or ESRI ArcMap. The use of thematic, temporal and spatial grouped hierarchies enables a contextual description of datasets and allows for semantic enabled data query. The user retrieved datasets by defining meaningful search parameters such as the regions administrative name, theme description, and time ranges.

However, the prototype implementation appears insufficient in many ways. The relation of datasets to different hierarchic schemes apparently causes some redundancies. For example, a dataset which belongs to the group “water turbidity” is cross related to “physical constituents” at the highest level in the thematic scheme, but also related by its parent “Water quality” and the grandparent “Environment”. Efficiently, there should only one registered child that inherits the properties of his parents. Furthermore, future changes within the developed hierarchical structures based on extended demands will induce administration, maintenance, consistency and efficiency problems. It is difficult to integrate different sources of information using different ontologies and domains. Ontology merging, mapping and integration with our approach are not possible. Future developments will therefore strongly focus on the evaluation and adoption of different approaches of ontologies (i.e., single, multiple, or hybrid) and software based reasoning techniques. Existing solutions, e.g. as described by Buccella et al. (2009), will therefore be evaluated.

For present, datasets are ingested directly by the Data Entry Portal without checking for consistency, redundancy, and replacement. The staging procedures in future revision should incorporate assessing data quality and consistency based on common data warehouse strategies.

Another constraint is storing raster data outside the database. In our approach while registered spatially, temporally and thematically and managed equally, the actual raster data are stored as separate files. Today’s open source database systems do not allow for an efficient management of raster data. Hopefully, future developments in database structures will support raster data management to meet existing highly optimised, concurrent access from powerful query languages, and controlled by a gradient of user privileges.

Another key aspect is the presence of major redundancies in the database resulting from metadata management. While it has its advantages, adoption of the Geonetwork software and database model for full OGC CSW compliance, more efficient management of XML structured information is offered through xml databases. Thus, we stored ISO19115 metadata through the eXist XML database. So far, the goal of the WISDOM information system has not been to develop its own CSW client, but rather focus on implementing existing systems. Future evaluations will determine whether a full CSW metadata catalogue system should be an integrative part of the WISDOM information system.

The present data model does not cover all aspects of the data to the required degree of detail. For example, the definitions of meaningful information on raster datasets (attributes, descriptions, statistics, and accuracy) are still missing. The definitions and modelling of satellite sensors are also insufficient. Both of these aspects should be extended in accordance with OGC application scheme for Earth Observation products (Gasperi, 2007) and the OGC Sensor Web Enablement (SWE) initiative (Botts et al., 2006).

An additional constraint within the current data management model is that a sophisticated user authorization management subsystem does not exist. Further developments should focus on modelling user’s access to data through user group definitions. These authorization gradients can be implemented in a semantic way, as well, using ontologies.

The WISDOM data model and implementation appear to be transferable to similar applications, e.g. integrated water resources management (IWRM) and other projects concerning spatial data management with similar data characteristics, user, and system requirements. The application can be installed on any other computer or server which connects to the other available IWRM systems via Internet. The present limitations aside and with slight modification, the system was installed for several training activities within various organisations in Vietnam. Currently, these organisations are using the WISDOM prototype for managing their data in the described method. The prototype...
was successfully adopted to support a similar research project on water management in Central Asia.14

In the future, the focus should also be in the implementation of the distributed data management system at all organizational levels, populating with various datasets and thus, enabling data sharing. This would allow different organisations to handle their own data in order to obtain data properly and responsibly, in addition, the data providers will manage the data which will be available.

Acknowledgements

This research was undertaken in the framework of the German–Vietnamese research project “WISDOM”, funded by the German Ministry of Education and Research BMBF (FKZ: 0330777) and the Vietnamese Ministry of Science and Technology (MOST). The authors gratefully thank Christian Strobl and Michael L. Whiting for their valuable input to the manuscript.

References


