MODELLING, MANAGEMENT AND DISTRIBUTION OF HETEROGENEOUS DATA
FOR A WEB BASED INFORMATION SYSTEM

Steffen GEBHARDT\textsuperscript{a}, Thilo WEHRMANN\textsuperscript{b}, Verena KLINGER\textsuperscript{b}, Juliane HUTH\textsuperscript{a} and Claudia KÜNZER\textsuperscript{b} \\
\textsuperscript{a}Department of Remote Sensing, University of Würzburg \textsuperscript{c/o \textsuperscript{b}German Aerospace Center, German Remote Sensing Data Center} \textsuperscript{82234 Wessling, Germany} \textsuperscript{Tel.: +49 (0) 8153 28-1715 Fax: +49 (0) 8153 28-1445} \textsuperscript{E-mail: steffen.gebhardt@dlr.de}

KEY WORDS: Decision Support System, Spatial data modelling and -management, Data heterogeneities, Data standardization, Data distribution, Web-based Geographic Information Systems

ABSTRACT: Efficient and sustainable Integrated Water Resources Management (IWRM) requires computer based Decision Support Systems (DSS) supporting operational management and strategic policy-making and planning. Spatial data and related technologies provide information on abiotic, biotic and socioeconomic conditions which are crucial for effective collaborative decision-making in the IWRM context. The more high quality data are available the more a DSS can support decision-making processes by customized, question- and problem driven information products. The German-Vietnamese Water-related Information System for the Mekong Delta (WISDOM) 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. With that, large amounts of heterogeneous data from the above fields including spatial data, statistics or ground-based observations have to be managed by the system. At the same time these data shall be identified by meaningful user driven requests, like datasets belonging to a specific theme, for a specific administrative region, and within a defined period in time. 

This paper introduces the data models and –management components of the web-based WISDOM system. The presented data models cover all relevant aspects on both, the semantic- and actual dataset level. They enable efficient management of spatial, statistical or observational data. Our approach of semantic tagging of datasets allows for the efficient and simple data identification and access. Our models are driven by existing international standards for metadata description (ISO19115), Map Styling Information (OGC Styled Layer Description) or dataset distribution (e.g. Web Map Service and Web Coverage Service). The data management utilizes an object-relational database management system also supporting management and processing of spatial data, a XML database for managing and querying XML-structured information as also Web-based Distributed Authoring and Versioning (WebDAV) for managing files collaboratively on web servers.

1. INTRODUCTION

Information systems for sustainable and adaptive Integrated Water Resources Management (IWRM) depend 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, Valadan Zoje et al. 2008).

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. It enables end-user to perform analyses on very specific questions, such as in regional planning activities, and facilitates integration of the dispersed datasets. 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.

This multidisciplinary approach hinders smooth data integration and rapid retrieval due to large data heterogeneities which are (i) semantic heterogeneity due to different research disciplines, languages, terminologies, and perspectives (Lutz, Sprado et al. 2009); (ii) structural heterogeneity due to different data types (e.g. spatial data, text documents, graphs, and images); (iii) 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, Einspanier et al. 2003); and (iv) variation in levels of detail for data descriptions (metadata).

Those are yielding in a diversity of data content, data models, and data types (Butenuth, Gosseln et al. 2007) creating huge work load of manually reformattting and semantically registering data before exploring, processing and visualizing (McGuire, Gangopadhyay et al. 2008). Effective spatial data standards are necessary to establish such a system (Guptill 1994). Syntactic and structural transformation approaches handle these heterogeneities, however, they are not adequate for resolving semantic differences (Berg, Stengel et al. 2007). The use of ontologies is considered a possible solution for semantic heterogeneity problems. 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, Cechich et al. 2009).

This paper introduces the data standards, models and –management components of the web-based WISDOM system with focus on the spatial data. The presented data models cover all relevant aspects on both, the semantic- and actual dataset level.

2. THE WISDOM INFORMATION SYSTEM ARCHITECTURE

The WISDOM architecture follows a web-enabled multi-tier approach. Data tier and 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 supports data browsing and access. User can access the information system using standard web browsers. The system components in general comprise:

- **Data standards**: The WISDOM Geodata Exchange Format (WGEF) defines data formats (ESRI Shapefile, GeoTIFF), coordinate reference system (EPSG), spatial metadata (ISO19115), map styling description (SLD), as well as the container format for individual spatial raster or vector datasets (ZIP).
- **Data Entry Portal (DEP)**: interface for automatic incorporation of spatial data into the management system. This application comprises analysing data for standard compliance, automatic metadata generation and disseminating the dataset to the data management system.
- **Data Management System**: 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.
• **Services and Protocols:** Web services are used for communication between middleware and data tier to add scalability and extensibility. Implemented data access services are in compliance to OGC are WMS, WCS and WFS. Data processing tools are encapsulated by Web Processing Service (WPS). Additional services were implemented based on Representational State Transfer (REST) style architecture.

• **Graphical User Interface:** The graphical user interface constitutes an interactive internet browser application. 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.

### 3. DATA MODEL

A spatial dataset is modelled as an entity sharing (i) common descriptive attributes (date of creation, data provider, etc.) and (ii) specific, technical data aspect attributes as also (ii) semantic dataset aspect parameters. Technical data aspects 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 by Web Mapping Services (WMS), (iii) data styling aspects defining the map representation of datasets, (iv) metadata, and (v) data attributes. Semantic attributes register the dataset to (i) spatial reference objects like the geographic area in which the dataset is situated, i.e. within a political jurisdiction; (ii) thematic references, linked to simple ontologies realized as hierarchical groupings representing the thematic dataset domain (e.g. hydrologic, environmental, or social data). Semantic registration and tagging enables faster and more “intelligent” access of the data. 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 and thematic 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 objects

A major requirement of the WISDOM system is the dataset identification and access by regional search parameters, i.e. the administrative unit of interest. The system therefore must: (i) accommodate unique “spatial reference objects” of either point, line, or polygon-type; (ii) assure non-redundant management of administrative “spatial reference objects” at different levels for different time periods according to different municipal reforms; (iii) semantically enrich datasets for cross-referencing to the “spatial reference objects”, e.g. administrative units; and (iv) 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 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. 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). Every dataset, either spatial or non-spatial is registered to the respective reference object using cross table relations.
3.2. Thematic reference objects

Similar to the Spatial Model hierarchy, 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 main topics such as Hydrology, Landcover and Landuse, Infrastructure, or Population and Economy. In keeping with the properties of ontology, each of these topics is divided into finer subdivisions, following a class-like structure. For example, the topic “Hydrology” contains the child “Water quality monitoring” which is composed of finer categories, of which one grandchild is “Physical”. 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).

3.3. Dataset entity model

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

3.3.1. Spatial reference aspect

Every dataset makes reference to at least one object in the spatial reference table. Referencing of a spatial dataset to the respective “spatial reference objects” is done automatically during the data integration process of the Data Entry Portal (DEP). Thereby, 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. The spatial reference for non-spatial datasets is defined manually in compliance to the developed exchange standards.

3.3.2. Thematic reference aspect

Each individual spatial dataset belongs to a specified group (e.g. watermask, water turbidity, and landcover), 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. The cross reference registers each dataset to its respective themes at different levels, for example a “watermask” belongs to “Hydrology” at the highest level, “Inundation Monitoring” in the next level and, finally to “Water Distribution” at the lowest level. The benefits of this modelling approach are (i) adding thematic contextual information to spatial and non-spatial datasets using hierarchies and (ii) enabling dataset queries through meaningful thematic groups, and (iii) maximizing the users' ability to browse the data catalogue using thematic groups.

3.3.4. Geometric aspect

Our system utilizes the PostgreSQL database with the PostGIS spatial extension. 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. Since raster data are 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. The geographic extent of both raster and vector datasets is stored as simple polygon geometry in the “footprint” table (coverage aspect). 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) fast query and dataset identification using the same spatial query parameters.

3.3.5. WMS layer and map styling aspect

WISDOM models WMS services as single entities with their attributes describing those above plus the metadata elements service web URL, layer limits, supported formats and units, extent, scale, contact information, and WMS layer elements, according to the WMS specifications. In the data model, a dataset entity referencing a “WMS layer” entity in a 1:1 relation includes the referenced dataset entity id to the metadata elements, and the addition of styling parameters. Each spatial dataset in the WISDOM information system, either raster or vector, is related to a SLD specification. A layer requested as WMS by the Graphical User Interface is represented as map layer using the defined SLD symbology, and with its specific metadata and query attributes. The actual SLD files are stored to the xml database.

3.3.6. Geographical metadata aspect

Metadata describing spatial data allow users to identify data sets not only by bibliographic information, such as authors, title, date, publisher, but also by spatial or temporal coverage, parameters used, and data quality. The ISO19115 standard provides xml structured metadata for each spatial dataset in the WISDOM system. The actual xml files are stored to the xml database.

4. DATA ACCESS WORKFLOW

Using the developed graphical user interface datasets can be immediately queried by defining regional, thematic and temporal request parameters. Submitting the query calls the respective REST service and forwards the request as translated SQL statement to the database. The response is received by the GUI tabular element which represents all identified dataset entities (Figure 1). At the same time the user selected a table row, another action executes a REST service retrieving the ISO19115 metadata. The response provided filtered information on title, abstract, generation, credits, contact, download and preview image. Spatial data can be selected for interactive webmapping. By doing so a REST service is called which WMS service definition on the fly and delivers the rendered dataset map layer as WMS GetMap request to the map client (Figure 1).

5. CONCLUSIONS

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 and as many user requests via the web client. Many spatial datasets, in both vector and raster format were integrated in the database. Implementing the management of different data aspects advanced data query and data distribution algorithms. 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 regions administrative name, theme description, and time ranges.
Figure 1: Components of web GUI of the WISDOM Information System that enables data query and exploration.

6. REFERENCES


