Target area based relational database system for managing earth observation information

T. Pyhälahti [a,*], S. Anttila [a], S. Korhonen [a], R. Teiniranta [a], M. Hynninen [a],
J. Pulliainen [b], J. Vepsäläinen [a], K. Kallio [a], S. Koponen [b]

[a] Finnish Environment Institute (SYKE), P.O. Box 140,00150 Helsinki, Finland – timo.pyhalahti@environment.fi
[b] Helsinki University of Technology, Laboratory of Space Technology, P. O. Box 3000, FIN-02015 HUT, Finland

Abstract – Monitoring of environment requires several sources of information to be combined: "General knowledge" of the monitored area, measurement information on it and naturally previously made assessments of the same area. To use earth observation in monitoring of for example lake water quality, traditional images have limited value: How could they be used for example when one wishes to draw a graph of time series from several lakes in order to compare them? Indicator values stored in a database would be required. To solve these data use problems in water quality monitoring, ULAPPA system was developed using target area approach. It starts from the fact that usually the areas to be monitored are clearly defined: Lakes, fields etc. have more or less well-defined locations and boundaries. They have distinctive a priori characteristics as individual entities or according to their type, in situ measurements are made on them, they are being observed by several types of satellites and all this data is combined using several different algorithms, possibly with several versions of algorithm parameters. Sometimes model assimilation requires access to uninterpreted, usable satellite data. Management of usable earth observation data is critical for optical data, as clouds often obscure the target area. All these needs can be integrated according to user needs using the target area approach.

Keywords: satellite remote sensing, earth observation data, information system, GIS, data archiving, data processing

1. INTRODUCTION

1.1 General monitoring requirements

Due to the elementary role of lakes, rivers and the Baltic Sea coast in the Finnish nature and their vulnerability of anthropogenic influences, there are extensive monitoring systems to support water management. Eutrophication is currently the main environmental problem. The European Union Water Framework Directive (WFD) (EU, 2000) imposes new challenges on the national monitoring systems. WFD requires the member states to take actions in order to reach ecological status "good" in for example coastal seas and lakes by 2015. This involves both characterization of water bodies in order to decide what "good ecological status" exactly means in each case and implementation of monitoring systems to verify the ecological status. In any case, there will be several factors to be taken into consideration when water quality classification is made.

Lakes cover relatively large proportion, about 10%, of the Finnish territory. They are mostly small and shallow, and the large ones are typically divided into sub-basins with peninsulas and islands. Thus, instead of large continuous areas of interest, there are lots of small scattered areas with individual properties. The complex Baltic Sea waterline with scattered archipelago and varying strong effect of the rivers results in similar individual areas in the coastal sea.

For WFD and other monitoring requirements, the water areas should be monitored individually, and the monitoring methods should be consistent over periods much longer than typically expected technical life of a single satellite. In addition to access to measurement results of different quality parameters one should be able to see trends in their time series. Finally, it is not just one type of parameters one is interested in, but one should be able to combine several different types of measurements and other sources of information. Naturally, these types of complex requirements are not typical for water quality monitoring only.

1.2 Target Area Approach

As described in the previous chapter, raster images from satellite data and thematic maps derived from them do not answer to all requirements of the monitoring systems. For creating the required information services, target area approach (Pyhälahti *et al*, 2002) was developed. The general idea is to gather satellite observations from pre-defined areas of interest, the *monitored entity*. In the current application they are lakes and parts of the Baltic Sea, in other applications of target area approach they could be fields, forest parcels etc stationary geographical objects.

Not all parts of the monitored entity may be monitorable with the available observing resolutions used instruments. It is not likely there will be methods capable of providing qualitative water information from mixed land/water pixels. These areas naturally depend on the resolution of the used instrument. Area deemed observable of a monitored entity with a certain resolution is called *target area*.

It is worth noting that the target area is often not likely to be homogeneous in its properties. Municipal, industrial or agricultural loads, depth and current structure or water inflow may cause significant differences in the ecologic and geophysical characteristics of different parts of the monitored entity. Due to problems in using mixed pixels in deriving useful information, different parameters – in case they are measured with different kinds of instruments – may be available for significantly different sub-areas of the whole

* Corresponding author.

** The work was co-financed by TEKES and the Academy of Finland

monitored entity. Thus it is necessary not only to derive monitoring information of the whole area, but also to have separate statistics for these special areas of interest. Thus, within target areas of the monitored entity, several potentially overlapping *monitoring areas* are defined. All monitoring results are to be connected with some monitoring area, and users will select the results they need from the pre-calculated results associated to a single monitoring area. In addition, monitoring areas are the key entities in integrating other than remote sensing data to the system. These include in situ sampling, weather-related data as well as a priori information on the monitored entity, such as expected normal variation of water quality parameters in a certain lake or lake type.

As it comes to remote sensing observations, it does not really matter how many overlapping monitoring areas are defined to cover any point. For example, the same set of satellite observations useful for deriving average concentration of chlorophyll in a certain bay is going to be a subset of the set of satellite observations for calculating the same average for the whole lake. For satellite observations, the idea is to store only the useful satellite observations, and to store them as 'packages' according to their location. This is accomplished by dividing the target areas to parts according the desired monitoring area borders, creating a set of *resolution areas*. The borders of different resolution areas will thus follow the borders of target areas for different resolution instruments as well as other necessary borders (ecologically selected, administratively feasible etc.) The monitoring areas will then be constructed simply as unions of these resolution areas.

1.3 ULAPPA

Target area approach is the term for the described methodological approach in using remote sensing in practical monitoring systems. The general technology developed for practical implementation of it is called SADB (Satellite Archive Data Base). The implementation of SADB to the specific problem of monitoring Finnish lakes and coastal sea is called ULAPPA. *Ulappa* is a Finnish word for large open water area – the type of area from where daily medium resolution satellite observations are available.

In addition to administrative boundaries, resolution is the main parameter in defining monitoring and resolution areas in ULAPPA system. Resolution areas with higher resolution form buffer zones around coarser resolution areas, which are only observable with higher resolution level instruments. In ULAPPA these levels are currently 250-300 m (MODIS channels 1,2), 300-500 m (MERIS; MODIS 1,2), 500-1000 m (MERIS, MODIS 1-7) and 1000 m (MERIS, MODIS, SeaWiFS, AVHRR etc.)

2. METHODOLOGY

The SADB (Satellite Archive Data Base) methodology is based on relational database and XML technology. It is of generic nature, thus the structure and mechanisms of the system do not change if the observed areas are water, fields or forests. The system is currently implemented on a Windows SQL 2000 server, and it is fairly complex, consisting of ca. 200 tables. The queries available for the database users are typically handled by the data access software layer, thus masking the complexity of the database.

However, the real complexities of storing detailed information on for example a priori information on observed target areas are to be stored within XML data structures within the relational database text fields. XML (Extended Markup Language) is a markup language for documents containing structured information. Structured information contains both content (words or in this case typically parameter values) and indication of what role that content plays. XML was created to use structured documents over the net. Basically, XML documents are text data containing markup strings and the actual data content. This approach has the benefit of both not having to try to generate even more complex relational database structure and not to limit the options available for storing the information.

2.1 Geographical definition of monitoring areas

A resolution area is typically a polygon in a vector GIS database and monitoring area is a union of these non-overlapping resolution areas. Normally a single resolution area belongs to several monitoring areas. Typically one monitoring area is defined to consist solely of this single resolution area. To avoid having to constantly rely on potentially heavy GIS searches for location, a geographical envelope is stored to database for both resolution and monitoring areas including the maximum and minimum coordinate values on both directions.

2.2 Observational data from direct remote sensing of the target area

The primary purpose of defining resolution areas is to store information of the usable satellite measurements from the specified area. There are basically two methods of doing this: One way is to select the centre points of the unique measurements generated by the satellite instrument which fall into these areas, check them for cloud contamination etc and store the point values for further processing. The observations falling on resolution areas that demand higher resolution than the current instrument uses are always neglected. The other way is to rectify the satellite observations into a georeferenced image, to mask out cloudy and otherwise unusable areas and finally to check if there are useful measurements within defined resolution areas. Instead of actual satellite measurement, this method stores a link to the usable data in a rectified source data image.

For further processing of the earth observation data over the resolution areas it is crucial to be able to know what the stored data actually is. SADB links together information on individual instrument, the platform on which it operates, characteristics of the data it produces and naturally configuration of the wavelengths, bandwidths, resolutions, calibration coefficients etc. used in the measurement. For synthetic aperture radar (SAR) instruments changes of operating modes affecting resolution, measured polarization and other parameters are done routinely from one acquired image to the next one. Some optical instruments, specially some airborne imaging spectrometers, which here can be used in the same way as satellite instruments, can be reconfigured too to measure different wavelengths. Calibrations of all instruments, even if planned stationary, typically change during instrument lifetime. Even if the instrument is a member of a series of instruments, such as AVHRR or MODIS, and it is nominally supposed to have stable characteristics from one instrument to another, it will require individual descriptions for accurate error characterization, since there will be unique variations in the instrument itself.

After satellite data stream is received in the receiving station, several steps of processing have already been conducted before the user of the data receives it and adds it to the SADB sys-tem. If the original data is not uncalibrated instrument digital readings, but for example below-the-atmosphere reflectances or even interpretation results to some geophysical parameters, it is likely the processing methods are developed and thus changed over time. The SADB system allows several levels of processing of the original data to co-exist in the database as alternative sources of interpretation.

It is not advisable to store data that is already flagged useless in the original data processing, completely contaminated with cloud cover etc. There are often several plausible ways of pre-processing the data for extracting the usable data. Sometimes an iterative approach may be-come handy: An automated filtering of completely useless data is followed by careful semiautomatic extraction of the remaining, less obviously flawed data. The relational database system allows both of these to co-exist – additionally other types of filtering based on for example proximity to target area borders or limits of measurement angles could be applied and the results stored as alter-native datasets for further processing. Reason for having these multiple datasets could be increased sensibility of some algorithms to for example atmospheric errors or measurement circumstances as compared to more robust methods.

2.3 Other types of observational data

In addition to direct satellite measurements of the monitoring areas, there are other measurements or measurement/modeling combinations, which provide information concerning these areas, even if they are not specific to them. For example meteorological atmospheric models can provide gridded estimates on solar radiation at a certain time. The resolution of this data is much coarser than the dimensions of the typical monitored areas. However, it is possible to find suitable values from it using location information and data format specific methods for searching the stored data. Certain other types of data, such as ultraviolet radiation data are available as time series for certain coordinate points, as the coarse resolution satellite measurements over time in the neighborhood of these points are binned together. The logical step is then to assign each monitoring area to the closest coordinate point in order to get the best representative UV measurements.

Last but not least, there are the in situ ground measurements. There is an immense amount of different types of them; they could even include on-ground photographs or lists of plant and animal species observed in the area at certain time. As said, variable contents can be stored in XML or binary format to the SADB database. Some of these observations can be used directly in automated satellite observation interpretation, some are merely intended for end user personal interpretation. Typically these observations can be considered as point-wise, such as water samples, but for example flow-through measurement devices produce point measurement tracks. These measurements are assigned to all monitoring areas they are considered to represent.

2.4 A priori data and information

The monitored entity – lake, sea area, agricultural field, forest parcel… - is typically quite well known even before a single remote sensing observation has been made. In addition to it's geographical borders it typically has well defined ownership and administrative definitions, perhaps names and/or codes in several languages and coding systems and possibly additional geographical or geophysical information. Even if this information may be useless in processing remote sensing observations, the user of the data may find it extremely useful for combining the results in order to fulfill the actual purpose of monitoring. Second type of information is related to the monitored phenomenon, such as anticipated range of variation in optical properties in the observed water body or anticipated spectra of bare soil of the area in different situations. This data can be utilized directly in interpreting remote sensing observations into geophysical values of interest, for example in lake-wise optimized detection algorithms. Third type of monitoring area specific a priori information is related to modeling of the actual observed phenomenon or assimilation of the related remote sensing observations to these models. These model parameters are often difficult to measure, so are likely to be merely results of mathematical model calibrations.

In addition to this kind of monitoring area specific information there is data that is common to and independent of all monitoring areas of the same type. General model parameters and spectral optical property data such as absorption spectra of certain species of phytoplankton are good examples of these. Even the structure of a model to be used in each single monitoring area can be described in this kind of information. SADB allows even linkage between these general parameters or models and the monitoring-area-specific data.

There is typically a lot of information that should be available for the end user of the data. However, the information may reside in several different information systems, whose databases are maintained by several different organizations. It would not be feasible to try to copy and to maintain a synchronous copy of all these data in a single database system. Instead, a method of storing links and methods of retrieving data concerning single monitoring areas is implemented in SADB methodology.

2.5 Grouping and connecting monitoring areas

The Finnish lakes are often comprised of several interconnected basins, thus separate monitoring of the single basins as well as generation of monitoring indicators for the whole water body is required. Different lakes can be classified according to ecological, bio-optical, geographical etc properties. It is necessary to associate these lakes and basins in the database in order to attach information on these associations. In practice the relevant monitoring areas are grouped together separately for each purpose, and the properties intended for describing them are assigned to these groups.

The properties attached to individual monitoring areas should be preferred over the same type of properties assigned to groups, as these monitoring area values may contain some information, which is actually adjusted to the real properties of the described area.

In addition to having several basins in one water surface level, the Finnish lakes are typically connected to each other with rivers or narrow straits. As the water surfaces are on different levels, there is a flow of matter typically always in the same direction. This kind of connections is formed pair wise in between two monitoring areas with clear statement of direction. A single monitoring area can contain several types of connections including connections in return direction. Individual properties, such as model parameter values, can be attached to all of these. An interesting possibility is to connect monitoring areas of lake or coastal water to monitoring areas of fields or forests for modeling different kinds of water transport for hydrological purposes.

2.6 Interpretation results

The interpretation results may be single values, data in XML or binary format or references to result images. For monitoring areas, there are two main modes of detection: 'Area value' treats the monitoring area as homogeneous, thus describing things as mean value of parameter within area, standard deviation of concentration, percentage of area covered by certain type of vegetation etc. 'Event' type of results are more point-wise phenomena, which in principle do not concern the whole monitored area. Such events detected from satellite images as oil spills, ships, icebergs or even large trees in open terrain have a certain location within the monitoring area, and naturally none or multiple of such events can be detected. The interpretation results of events may be connected to each other physically (ship – oil spill), logically (detected oil spill – modeled drift trajectory of the same spill) or temporally (same oil spill in consecutive images).

As there will be a large variety of different types of results, SADB enables storage of documentation on the method as well as grouping the methods according to user requirements. SADB enables a single interpretation result type to be a member of one or more classes, and even the classes may have a hierarchical system. As validation results, research and algorithm development are likely to modify the original algorithm, versions of similar result types should be grouped under the same type class. The result type classes may be ranked according to preference for use for higher hierarchical classes. Suitability classes could be defined to for example different types of lakes, in which the accepted interpretation methods and their rankings could be substantially different.

For normal user – who typically only requires some classification or approximate value - the estimates can be characterized in different ways, such as:

1. Ranking the different estimation methods according to their applicability for the individual monitoring area or group of them, or according to applicability for a certain task. Discontinuities in data quality are to be expected as the best available instrument and algorithm change.
2. Comments on estimated accuracies can be attached to individual measurements, algorithms or algorithms on specific monitoring areas or groups of them. Naturally, accurate estimation of algorithm performance can be difficult.
3. Syntheses can be derived from multiple types of information. One must be prepared for a possible re-calculation of this true value on the basis of a new set of information – and to keep the old value still available for comparison.

3. CONCLUSIONS

As the structure of the SADB relational database becomes quite complex, the users of the system do not need to be exposed to it directly. Effective data access layer software should be common to all users and applications accessing or storing information to it. SADB itself contains method for multilingual documentation of the stored data, methods etc. but it is up to the SADB database managers to make sure information is available and correct. Efficient SADB management tools are still to be developed.

As it is possible to group monitoring areas, interpretation result types etc in very many ways, careful consideration is required in order to provide the user logically consistent and user-need-driven associations. It is an important task for database management to make sure all the groupings actually really contain the data they are supposed to.

Even if target area approach results in complex systems like SADB, there are several benefits to it. As ULAPPA is currently transferred from a demonstration phase to operational phase, these benefits will be more clearly visible. Similar systems would benefit the access to drainage basin property information, which is seen as very important input for water quality modeling in Pyhälahti *et al* (2005). In the long run it will be more feasible to have one SADB methodology to implement target area approach to water monitoring in ULAPPA and potentially terrestrial monitoring in some other system than to re-invent the wheel each time.

4. REFERENCES

4.1 References from Journals

Pyhälahti, T., Anttila, S., Korhonen, S., Teiniranta R., Hynninen, M., Pulliainen, J., Vepsäläinen, J., Kallio, K., Koponen S. User requirements for water quality remote sensing products and means of meeting them, 31st International Symposium on Remote Sensing of Environment: Global Monitoring for Sustainability and Security, June 20-24, 2005 Saint Petersburg, Russia

Pyhälahti, T., Teiniranta, R., Kallio, K., Vepsäläinen, J., Koponen, S., Pulliainen, J. Target Area Approach In Remote Sensing Of Lakes And Coastal Zone. Ocean Optics XVI, Nov 18-22, 2002, Santa Fe, USA

4.2 References from Websites

EU, 2000. Directive 2000/60/EC of the European Parliament and the Council establishing a framework for the community action in the field of water policy (EU Water Framework directive (WFD)) http://europa.eu.int/comm./environment /water/water-framework/index_en.html