Data Management for Environmental Monitoring

Richard Hůlek¹, Miroslav Kubásek¹, Jakub Gregor¹

Abstract
A huge effort on data management must be made to discover long term trends, patterns and to enable performing advanced statistical analysis of collected data from environmental monitoring networks. Creating a robust ICT architecture, designing data structures for heterogeneous entities and processes, guarantee quality of data at all levels of data repository as well as implementation of all security needs is a challenging task. During development of GENASIS (Global ENvironmental ASsessment and Information System) these needs were identified, resolved and successfully implemented. This work briefly describes core concepts, current state and near future roadmap of GENASIS data management system, which is used – among others – as a base tool for National Centre for Persistent Organic Pollutants, Stockholm Convention Regional POPs Centre for Central and Eastern Europe, as well as for research and education activities of RECETOX (Research Centre for Toxic Compounds in the Environment).

1. Motivation
A huge effort on data management must be made to discover long term trends, patterns and to enable performing advanced statistical analysis of collected data from environmental monitoring networks. Creating a robust ICT architecture, designing data structures for heterogeneous entities and processes, guarantee quality of data at all levels of data repository as well as implementation of all security needs is a challenging task. During development of GENASIS² (Global ENvironmental ASsessment and Information System) these needs were identified, resolved and successfully implemented (Dušek et al., 2010), (Holoubek et al., 2011). This work briefly describes core concepts, current state and near future roadmap of GENASIS data management system, which is used – among others – as a base tool for National Centre for Persistent Organic Pollutants, Stockholm Convention Regional POPs Centre for Central and Eastern Europe, as well as for research and education activities of RECETOX (Research Centre for Toxic Compounds in the Environment) (Klánová et al., 2009), (Urbánek et al., 2010).

2. State of the Art
Current publicly available systems focused on environmental monitoring lacks of complexity. There are many institutions developing different systems from the fields of hydrology, air monitoring, soil contamination etc. But most of these systems are not capable of combining multiple types of monitored matrices or lack of visualisations and interpretation tools which allows users to explore collected information.

Main goal of GENASIS system is to fill this gap and provide robust platform for multi-matrix environmental monitoring together with well established data and quality management as well as publicly available analytical and interpretation software tools – all together as an ecosystem.

¹ Masaryk University Brno, Institute Biostatistics and Analyses, Kamenice 126/3, 62500 Brno, Czech Republic
² http://www.genasis.cz/
3. **Expand the Focus**

The initial version of GENASIS was launched in 2010 (Dušek et al., 2010). Primary focus of this environmental database was air monitoring data. Publicly accessible analytical and interpretation tools were developed together with the database, as well as content rich web portal with important information about POPs, national and international legislative concepts, case studies and scientific articles.

Needs to expand the focus to multiple matrices (such as soil, sediments or human tissues matrices) and enhance provided analytical tools have recently increased. Thanks to new contracts with data providers the amount of incoming data increased multiple times. During 2011 every piece of the GENASIS data management system was revised and in many cases redesigned and new innovative features were introduced (Holoubek et al., 2011), (Jarkovský et al., 2012).

4. **Approach**

4.1 **Robust architecture**

Conceptual scheme of system architecture from first version of GENASIS system was used and further enhanced. Core concept of a primary data repository, data warehouse and publicly available presentation layer proved its qualities as a robust architecture design.

Primary data repository is used to store data coming from particular data providers in form of individual records with additional description by set of standardised code lists.

From primary data repository are data simultaneously transformed through validation and analytical processes to a data warehouse layer where records are used in pre-processed form to support presentation and analytical outputs.

Presentation layer consist of set of analytical tools which enable users to discover valuable information in stored datasets.

4.2 **Used Technologies**

MS SQL Server is used as a core database, which serves both as a primary data repository and data warehouse. Pure entity relational modelling concepts are used with business intelligence layer to support analytical outcomes. Close integration with ESRI’s ArcGIS Server has been done to enable spatially oriented operations and to enhance spatial visualizations frequently used in the GENASIS case studies module\(^3\). Map layers are consumed from publicly available sources, own spatial services or generated directly from the data warehouse database in a real time. Several new technologies have been incorporated apart from currently used AJAX-based frontend technologies, such as Adobe Flex/Flash and ArcGIS API for Flex to support spatial visualizations.

4.3 **Database management**

Most data structures of GENAIS core database were redesigned and a new model was implemented to provide more robust and universal model to process a broad range of environmental monitoring activities over multiple matrices (such as air, water, soil, sediments or human tissues). Both database structure and code lists content are based on compound ontologies, common standards and best practices, which were thoroughly discussed with domain experts (Komprdová et al., 2011).

---

To be able to store such a heterogeneous data additional code lists had to be added to describe all aspects important to further analysis. Most important information is for example which method has been used to collect given sample (sampling method), which exact method has been used to measure concentration of given parameter (prepare method, analysis method) and so on.

GENASIS in its second version currently contains database tables to store information about:

- Institution – which does the monitoring (for example RECETOX);
• Project – within the monitoring activities are done (MONET CZ, MONET CEEC...);
• Site – where the monitoring is actually done with rich description of the site (latitude, longitude, altitude, landuse, slope...);
• Matrix – which matrix has been sampled (air, water, soil...);
• Sample – describes when and how (sampling method, prepare method, analysis method) was sample created;
• Value – contains exactly measured concentration, information about LOQ (level of quantification) and unit.

Every single database table mentioned in a very simplified form above is extended by se set of code lists which allows detail description of heterogeneous environmental data stored in database. There is nearly any attribute which allows entering free text (expect descriptions and notes). Attributes are derived from code lists content to be able to classify data and create clusters. Thanks to that database model grew up to about 60 code tables and about 20 data tables.

Cumulative number of records is reaching millions and has been growing continuously.

Database scheme implementation was enhanced and well balanced so that it is able to process amounts of millions of records in a real time during intensive calculations and aggregations.

Performance was – in particular cases – a problem of the previous version of the GENASIS system.

All database tables in GENASIS system implements versioning framework to keep track all modifications and deletes. Having good performance as a main goal versioning framework is optimized so that historical records are automatically moved to secondary database. This auto-purging mechanism helps to keep main database clean and fast while querying.

4.4 Data Management Administration Panel

We were also focused on the development of new tools (Administration Panel) to automate repeating and time consuming operations, mostly during the phase of data preparation before their import into the primary data repository. Administration Panel helps to keep data security and quality at a high level.

There are three main areas covered by available features of Administration Panel:
• Code lists – browse, search and export content of particular code lists.
• Data imports – Administration Panel helps its users to create import template for new data and import them into database.
• Data exports – batch exports are currently under development and will offer exporting big amounts of data in primary form from database to for example Excel sheets.

Administration Panel is password protected desktop application which is dedicated for internal GENASIS team members to simplify their work and reduce human made errors.

4.5 Workflow of data import

Adding new data (samples and measured values) into database or just updating them was always a problematic issue. All steps were done by database operator in cooperation with data manager. This process was very slow and prone to errors.

Important concept into whole data management workflow was introduced by import batches. Import batch is process during which are imported new data to the primary database. It is started by data manager through Administration Panel and new data are imported into database from filled import template files by software application automatically. Whole process has several steps so it is easy to control the current
state. Into import processes are involved guarantees whose confirm with their expertise and experience the correctness of all imported data. Once import batch is confirmed, all data are made visible and publicly available in analytical outcomes on GENASIS web portal.

Import workflow enables simultaneous work of data manager with data analysis team without any interruptions and without affecting content of the whole database with unfinished import batches which is most important.

States of import workflow:
- Importing of samples
- Samples imported, waiting for confirmation
- Samples confirmed by guarantee
- Importing of values
- Values imported, waiting for confirmation
- Values confirmed by guarantee
- Import batch confirmed – data are published and cannot be modified any more
- Import batch cancelled

Figure 2
GENASIS Administration Panel Application

4.6 Data Security Management

Adding more and more functions and cooperation with increasing number of subjects has introduced a new type of security problems (visibility, read/write permissions in projects/institutions). New permissions
of each individual system users were implemented through the Administrator Panel application as well. Data are protected against inappropriate usage by providing aggregated values for the public use only. Every record has its owner so appropriate security polices can be managed individually. These features are used in case of sensitive data.

4.7 Data Quality and Quality Assurance

There is a big need to ensure data quality so that GENASIS can serve its main purpose – as a decision making support tool.

Data quality is regularly checked and formal protocols of personal responsibility were implemented so for every record in GENASIS database is some person identifiable and responsible.

In any case of mistake original data can be easily tracked back thanks to version management framework implemented on the level of primary database.

Only users with assigned permissions can process given tasks and all activities are logged into log files automatically.

Essential rules and constrains are implemented and the lowest level of system – in database structure. GENASIS database heavily uses of combination tables which provide by its content a set of allowed combinations of other code list values. For example if there is a value measured on a sample sampled by passive air monitoring device it cannot be stored with µg/dm3 unit.

4.8 Document, Record & Content Management

Because all documents, metadata and other information can’t be stored in the database of information system itself, there is a good reason to incorporate some kind of content management system into workflow. Content management system enables its user to share files, keep all files versioned and protect them to be corrupted by simultaneous modifications by multiple editors. GENASIS system team members are using Alfresco⁴ content management system to manage meta data, external and internal code lists, conversion tables, protocols, project documentation and all other necessary documents in one place.

4.9 New tools and Case Studies

Publicly accessible analytical and interpretation tools⁵ were enhanced to profit from new data structures and to allow comparative analysis of multiple parameters or matrices which is quite unique

- Monitoring overview
- Sampling frequencies
- Reported values
- Time series
- Seasonality analysis
- Correlation analysis

GENASIS web portal (www.genasis.cz) was redesigned and its content was completely revised to provide up-to-date information. New interactive and visually rich case studies have been published.

---

⁵ http://www.genasis.cz/analyse_2.0/
5. Results

During the described second implementation phase of the GENASIS system its capabilities were expanded and proper data management workflow over this project was built.

In the near future and next development phase GENASIS will continue to receive new data from its data providers according to signed contracts and research activities. New case studies will be introduced and published on the GENASIS web portal. Development will continue to closer integration with spatial analysis software to support server-based spatial analysis available in a real time for web users. Other development branch will be connecting GENASIS data to other systems on a higher level than primary data. We will focus on spatial and time aggregations. GENASIS will be used as a validation scenario of the Ta-Too project as well.

6. Conclusions

GENASIS information and incorporated data management system is capable to resolve various needs of environmental monitoring, long term data collection and advanced analytical processing including time series and trend analysis to support decision making processes on a national and international level. Currently is content of GENASIS database mostly focused of POPs monitoring programmes but it can easily contain any concentration-like data.

Even though GENASIS as an information system is not introducing any great innovations it represents well established and robust platform for data management for environmental monitoring which is as a whole set quite unique solution in broader context.

Acknowledgments

Development of the GENASIS system is supported by the CETOCOEN project (CZ.1.05/2.1.00/01.0001) of the European Structural Funds, and the INCHEMBIOL project (MSM 0021622412) of the Ministry of Education of the Czech Republic.

Bibliography


Klánová, J. (et al.) (2009): Towards the Global Monitoring of POPs - Contribution of the MONET Networks. RECETOX, Masaryk University, Brno
