GENASIS System Architecture

Richard Hůlek, Jiří Jarkovský, Miroslav Kubásek, Jakub Gregor, Jiří Hřebíček, Ladislav Dušek, Jana Klánová, Kateřina Šebková, Jana Borůvková, Ivan Holoubek

GENASIS System Architecture

On the way from environmental data repository towards a research infrastructure

Richard Hůlek[1,2], Jiří Jarkovský[1], Miroslav Kubásek[1], Jakub Gregor[1,2], Jiří Hřebíček[1], Ladislav Dušek[1],
Jana Klánová[2], Kateřina Šebková[2], Jana Borůvková[2], and Ivan Holoubek[2]

[1]Institute of Biostatistics and Analyses, Masaryk University,
Kamenice 126/3, 625 00 Brno, Czech Republic
{hulek, jarkovsky, dusek, kubasek, gregor, hrebicek}@iba.muni.cz
[2] Research Centre for Toxic Compounds in the Environment, Masaryk University,
Kamenice 753/5 (A29), 625 00 Brno, Czech Republic
{klanova, sebkova, boruvkova, holoubek}@recetox.muni.cz

Abstract. GENASIS (Global ENvironmental ASsessment and Information System) is the web environmental information system which is an environmental data repository that provides comprehensive information on chemical contamination of the environment by Persistent Organic Pollutants (POPs). GENASIS combines data from long-term environmental monitoring programmes operated by RECETOX with validated data from partner institutions, and provides data archives, data management services and analytical processing of data. In the past few years the GENASIS system has undergone rapid development and grew up from the data repository into a scientific data infrastructure.

Keywords: GENASIS, information system, web portal, system architecture, environmental monitoring, POPs, Stockholm Convention

1 Introduction

The environment is significantly impacted by human activities. For example the occurrence of many dangerous diseases such as respiration problems or various types of cancer is directly affected by environmental conditions, particularly by artificial contamination of the environment by toxic chemical compounds. Tracking current state and evaluating long-term trends of chemical contamination is subject of interest of many monitoring programs over the globe [1-2].

Information and communication technology (ICT) tools for collection and storage of environmental data, ensuring their quality and security as well as tools suitable for statistical processing and evaluating of data are needed, preferably within an environmental information system. Such an information system (IS) plays an important part in decision making processes, effectiveness evaluation, research activities as well as in education.

adfa, p. 1, 2011.

GENASIS (Global ENvironmetal ASessment and Evaluation System) provides data repository capacities optimized for monitoring of important elements of environment (ambient air, surface water, soil, sediments or even human tissues), statistical processing tools directly connected to data repository, reporting tools and an information and publication platform in form of web portal. All its ICT tools, features and services are wisely integrated together to create its own "ecosystem" [1-2].

The GENASIS system and its software components are subject of this paper which describes the overall system architecture, individual component's design and features and the way they are interconnected and used.

The GENASIS system is developed by two cooperating institutions: The Research Centre for Toxic Compounds in the Environment (RECETOX) and the Institute of Biostatistics and Analyses (IBA), Masaryk University, Czech Republic.

Since the first time the idea of the GENASIS system occurred, it has been intended to serve not only for internal purposes of RECETOX but also for a broader group of users from the Stockholm Convention parties [3] and a wider range of activities. This vision impacted all decisions about the system architecture, its components and the way the system is being developed.

2 GENASIS System Architecture

The GENASIS system architecture is inspired by several design patterns and technology standards: modular design, serviced oriented architecture (SOA) [4] and a data warehouse principle [5].

Modules of GENASIS are designed to gather similar sort of system functionality into one system component with a defined interface, so that the inner implementation would be hidden inside of the module and could be updated or changed by different module implementation which supports the same required interface.

Services in terms of SOA are independent system components which provide some kind of functions with additional value (services) to other components. Services must have well defined interfaces and usually communicate through standardized communication protocols and data formats.

The data warehouse is a specific form of relational database which is optimized for complex queries upon large amounts of data integrated together from various sources.

The GENASIS system architecture (Fig. 1) consists of:

- **Data repository** is the core of the whole system and is composed of a database of primary data and a data warehouse.
- **Controllers and engines** are server side software components which ensure data quality and security and other features. A *data warehouse engine* transforms for example all updated data when any changes are performed on primary data. Design of internal controllers and engines was implemented to ensure strict separation of concerns.
- **Services** allow controlled and secured access for particular users or other system components to stored data (and data services); spatially oriented services are used for spatial analysis, computations and for map publishing (spatial services); ser-

vices focused on data management provide features for importing data and to setup security constrains (data management services); features of advanced statistical computations are provided by computational services.

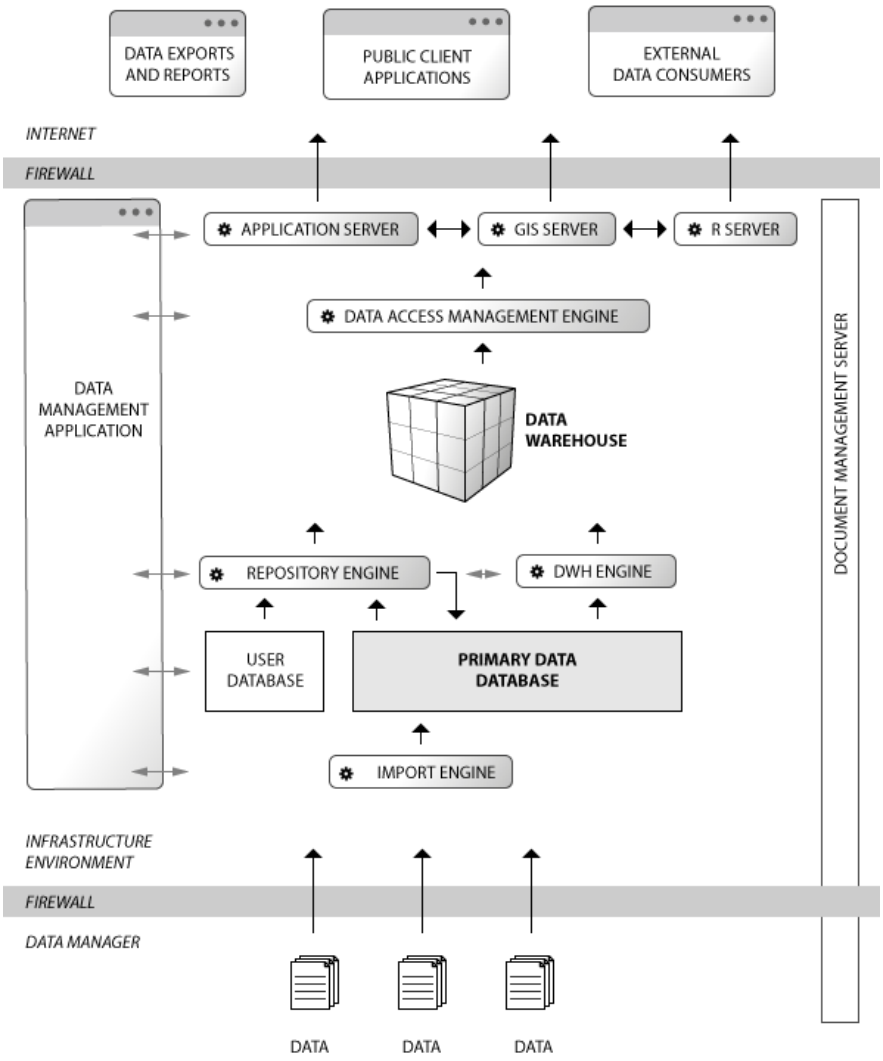

Fig. 1. GENASIS system architecture (Source: Authors)

- **Client applications** offer visualization of data stored in the data repository and enable users exploring the content of the database, perform basic evaluation analysis and export their outputs for further use.

- **Admin Panel** is used by data managers for data import, configuration of security constrains and export of raw - data into data sheets.
- **Supporting tools** such as document management system are used to maintain project documentation, workflows and to archive data packages provided directly by individual data providers.

Individual components of the GENASIS infrastructure are described in detail in the following subchapters.

2.1 Data Providers and Data validation

Data validation is an important process in gathering data from various environmental monitoring networks from different data providers. Before importing to the database the data must be validated, its nomenclature and code lists harmonized as well as transformed into a data format which is supported by the import engine [2], [6], [8].

All data from data providers are saved in a protected archive before any modifications are made.

Individual steps of data preparation are performed manually by specialized staff.

2.2 Import Engine

The import engine is the system component which ensures proper import of data to the primary database. It is controlled via an admin panel application through which data sheets in a standardized format are passed in. Imported sheets are parsed to elementary records and further processed. The main responsibilities of the import engine are:

- Data transfer from data manager's PC to the main server infrastructure;
- Format validation of data from the input;
- Error detection and proper notification handling;
- Storing data into the database.

The import engine is capable of handling tens of thousands records in a row; thus, the data managers can import data very efficiently. Data are imported into the database in batches so that they could be processed as groups. The overall process of the data import is divided into several stages:

- Import of samples;
- Samples imported, waiting for confirmation;
- Samples confirmed by guarantee;
- Import of values;
- Values imported, waiting for confirmation;
- Values confirmed by guarantee;
- Import batch cancelled;
- Import batch confirmed – data are published and cannot be modified any more.

Only data from confirmed import batches are used in the primary database and other components of the GENASIS system.

Import engine is implemented in a way that all operations with data are transactional –it means that a principle "all or nothing" is applied and any single error leads to cancellation of the whole dataset import. This guarantees integrity of all imported data.

2.3 Primary Data Database (Repository)

The primary database – called repository as well – is the core of the system. It is designed to work as a storage capacity of heterogeneous data from environmental monitoring networks. Data structures allow storing data from ambient air, soil, water, plants, sediments and other types of environmental matrices from both long-term and short-term monitoring programs of the Stockholm convention [3].

Data are organized into a logical hierarchy: Institution – Project (Monitoring network) – Site – Matrix – Sample – Parameter – Value – Unit. Beside this core structure there are data structures to provide more information about classification, categorization and detailed description of data. Samples are described according to the matrix, sampling method, analysis method and other factors. Sampling sites are located by spatial coordinates, described in text by the expert who performed measurements and classified by a set of parameters such as pollution source descriptors, geography, pedology or hydrology categorization. Measured values are stored with proper information on limit of quantification (LOQ) and the way it was determined.

Uniform storage data structures and standardized code lists ensures standardization of data formats and their content so that comparability of individual records would be ensured.

2.4 User Database

The user database serves as a helper database of the system. It contains information about user accounts and access rights. While nearly all operations in the system are registered, thanks to the user database it is possible to track authors and editors of records and their changes. The user database is important in connection with data access management module where it provides a flexible access control list (see detailed description in chapter 2.8).

2.5 Repository Engine

The repository engine expands functionality of the primary database. It is controlled by an admin panel application and integrates procedures over the primary database such as quality assurance heuristics and transformation procedures.

Quality assurance heuristics are single purpose server side computer programs which attempt to identify potential quality leaks which cannot be discovered during the phase of data import (such as duplicate records, missing records in sequences etc.).

Transformation procedures convert values stored in the database into transformed parameters. A good example of such transformation is converting concentration of a particular dioxin compound to its toxic equivalency (TEQ) or calculating sum of indicator PCBs (polychlorinated biphenyls). Transformations are performed directly upon the database so that the data managers would not be required to calculate additional parameters manually and import them separately.

Repository engine scripts are designed to run in batches and their execution is scheduled by a central update time table or is controlled by the admin panel application.

2.6 Data Warehouse

The data warehouse database contains transformed and pre-processed data prepared to be delivered to the end users of the system by any of the top level system applications or services (chapters 2.9, 2.10, 2.11). It integrates data from various monitoring programs into one common data structure optimized for reporting and data analysis and it represents a platform for decision-support data.

Analytical codes are added to elementary records to enable a broader range of selections and filters. Complex calculations and aggregations are calculated in advance in order to speed up results delivery.

The concept of the data warehouse enables integration of external data sources regardless its physical data structure.

Thanks to separation of the data warehouse database from the rest of the repository, it is easy to perform changes in the primary database without affecting the data warehouse database.

2.7 Data Warehouse Engine

The data warehouse engine connects the primary data database and the data warehouse. It is responsible for its creation and performs updates of the data warehouse based on changes in the primary database. Update procedures are scheduled by a central scheduling mechanism or are executed by the admin panel application.

Since the primary database and the data warehouse have different data structures, update procedures of the data warehouse engine implement all necessary transformations, calculations and aggregations.

Procedures implemented in the data warehouse engine are designed to be executed in batch mode because these tasks usually take a longer time to be executed.

2.8 Data Access Management Engine

Since the GENASIS system and its infrastructure has been used for various purposes and it contains data from more than 20 large monitoring programs as well as internal research experiments to date, there are various requirements to control access to individual record sets. Not all data are suitable for public use.

A custom data access management was implemented in the GENASIS system to meet specific needs of data protection. It is controlled by the admin panel application and it creates an access control list (ACL) for each user based on rules which arise from logical data structure, access scopes (private data / protected data / public data), data ownership and combinations of these factors.

The access control lists are partially pre-processed in advance (when some changes in repository occur) to speed up results delivery when large amounts of data are queried.

2.9 Application Server and Data Services

The application server provides an environment for various applications which are built on top of the GENASIS infrastructure. Applications typically access the data warehouse database or rarely primary database and perform further data transformations and calculations and pass results to the presentation tier of particular application's user interface. The application server provides pre-configured connectivity to other components of the system (such as GIS server or R server) as well as the appropriate computation resources.

2.10 GIS Server and Spatial Services

A GIS server extends functions of the GENASIS system by adding support for spatial data visualizations, map distribution and performing spatial analysis in real-time. The GIS server also allows linking additional information to stored data based on spatial references.

2.11 R Server and Computation Services

The R (R Core Team) server platform [6] was integrated into the GENASIS infrastructure to make use of its computations capabilities. Algorithms already implemented in free software programming language R are used instead of custom-created implementations of non-trivial statistical tests and models.

The integration of the R server improves the level of delivered services and speeds up the development of new analytical tools of the GENASIS system.

The R server consumes data from the data warehouse as well as from the primary data database and delivers results to various components of the GENASIS system. Analytical applications make use of the R server computation services for example to calculate time series statistical tests; the repository engine uses the R server component for parameter transformations.

2.12 Publicly Available Applications

Once various data from different sources are validated and imported into GENASIS repositories it is desired to analyse them in terms of existing patterns and trends. An

important additional value of the GENASIS system consists in analytical and interpretation tools which are directly connected to the data repository and perform statistical computations in real time.

The GENASIS on-line data browser [7] offers visualization of data from both national and international POPs monitoring projects of the Stockholm convention. Among the most important projects are Košetice EMEP station, MONET network and others. At present (June 2013), the database contains POPs data from 20 projects, 4 continents, 58 countries, 295 compounds and over 606,000 records collected during more than 20 years (first records date 1988) of continuous POPs monitoring. Monitored compounds cover most of these listed under the Stockholm Convention on Persistent Organic Pollutants [3].

In contrast to real-time general purpose web based statistical and interpretation tools, there are case studies [8] which deal with a specific topic or a selected part of the data repository. Similarly to the data browser they allow for online analysis of selected data, which are processed and presented in a structure and level adjusted to the particular purpose and study and are supplemented with further information on a given topic, methodology, and interpretation of obtained results.

2.13 External Data Consumers – Interoperability

Thanks to service-oriented and modular architecture the GENASIS system is capable of sharing data with external data consumers at various levels. The system can provide direct access to primary or aggregated data from the whole repository or parts of it. Besides data alone, results of analyses such as trend tests or maps of spatial distribution of contamination by particular compounds can be also provided.

2.14 Data Exports

Data exports in standardized format are an important function of the GENASIS system. Data exports are used in research, scientific publications and for both national and international reporting activities of Stockholm Convention CEE Regional POPs Centre [9].

2.15 Admin Panel

The admin panel is a standalone application connected to the GENASIS server infrastructure and is used by data managers to manage data in the GENASIS database and surrounding system components.

Through the admin panel data managers are able to:

- Import data into the database;
- Execute update procedures in the data warehouse engine;
- Execute transformations and quality assurance heuristics in the repository engine;
- Manage access to data and maintain user accounts;
- View and export data.

2.16 Document Management Server

The document management server works beside the GENASIS system itself but it still can be considered as a part of the whole infrastructure. The document management system is not (yet) connected with any other GENASIS system component programmatically but plays an important role in the overall process.

In the document management server there is stored the whole project documentation, content of code lists, all metadata and description of used standards as well as configuration properties needed for proper system set up. There is also an data archive containing the data as provided by the data providers. The document management server has the advantage against regular file system folders of flexible access control management and version tracking of all documents.

2.17 Firewalls and security components

The GENASIS system infrastructure is protected across several levels. GENASIS infrastructure is operated in Masaryk University Brno, Czech Republic [9]. The university's subnet is monitored and there are two levels of firewalls in front of each server. All activities and services are provided in accordance with ISO certifications EN ISO 9001:2009, ISO/IEC 20000-1:2006, and ISO/IEC 27001:2006.

3 Conclusion

The GENASIS system is used not only for internal purposes of RECETOX activities; it represents a research infrastructure available for use by other institutions, hosting data from different monitoring programs and providing its features and services.

Core services of the GENASIS infrastructure were used during development of tools for collection, analysis and visualization of data from the Stockholm Convention Global Monitoring Plan on Persistent Organic Pollutants [10]. Infrastructure of the GENASIS system is also used for the purpose of national inventories on persistent organic pollutants in the Czech Republic to meet requirements given by the Stockholm Convention on Persistent Organic Pollutants [3].

Acknowledgements. This work was supported by the TACR project no. TB010MZP058 “Development of the system for spatial evaluation of the environmental contamination”.

References

1. Holoubek, I., Dušek, L., Klánová, J., Kubásek, M., Jarkovský, J., Baroš, R., Komprdová, K., Bednářová, Z., Hůlek, R., Hřebíček, J.: GENASIS Information System: A Global Environmental Assessment of Persistent Organic Pollutants. In: Hřebíček, J., Schimak, G., Denzer, R. (eds.) 9th IFIP WG 5.11 International Symposium on Environmental Software

Systems: Frameworks of eEnvironment, ISESS 2011, pp. 480-485. Springer, Heidelberg (2011)
2. Hůlek, R., Kubásek, M., Gregor, J. Data Management for Environmental Monitoring. In: 26th International Conference on Informatics for Environmental Protection, EnviroInfo 2012, pp. 631-637. Dessau, Germany (2012)
3. Stockholm Convention on Persistent Organic Pollutants (POPs): Interim Secretariat for the Stockholm Convention, United Nations Environmental Programme (UNEP) Chemicals: Geneva,Switzerland, http://chm.pops.int
4. Arsanjani, A., Zhang, L.J., Ellis, M., Allam, A., Channabasavaiah, K.: A service-oriented reference architecture. IT Professional, 9(3), 10-17 (2007)
5. Chaudhuri, S., Dayal, U.: An overview of data warehousing and OLAP technology. SIGMOD Record (ACM Special Interest Group on Management of Data), 26(1), 65-74 (1997)
6. R Core Team: R: A language and environment for statistical computing, http://www.R-project.org/
7. Jarkovský, J., Dušek, L., Kubásek, M., Kohút, L., Klánová, J., Hůlek, R., Gregor, J., Šebková, K., Borůvková, J., Hřebíček, J., Holoubek, I.: On-line data browser for environmental monitoring and associated information systems Masaryk University, http://www.genasis.cz/, (2011)
8. Komprdová, K., Komprda, J., Sáňka, M., Hájek, O., Hůlek, R., Jarkovský, J. Spatial modeling of pesticides concentrations and pools, Masaryk University http://www.genasis.cz/case-studies/pops-spatial-modeling/ (2011)
9. Stockholm Convention CEE Regional POPs Centre, http://recetox.muni.cz/index-en.php?pg=regional-pops-center
10. Masaryk University Brno, http://www.muni.cz
11. Hůlek, R., Jarkovský, J., Borůvková, J., Kalina, J., Gregor, J., Šebková, K., Schwarz, D., Klánová, J., Dušek, L.: Global Monitoring Plan of the Stockholm Convention on Persistent Organic Pollutants: visualization and on-line analysis of data from the monitoring reports, Masaryk University, http://www.pops-gmp.org/visualization/ (2013)