International Environmental Modelling and Software Society (iEMSs)
8th International Congress on Environmental Modelling and Software
Toulouse, France, Sabine Sauvage, José-Miguel Sánchez-Pérez, Andrea Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2016-proceedings

An Information System for Groundwater Data and Modelling

Kamil Nešetřil[a] and Jan Šembera[a]
[a]*Technical University of Liberec, Czechia, http://msp.mti.tul.cz, kamil.nesetril@tul.cz*

Abstract: The subsurface is usually heterogeneous and characterized by scarce data – therefore it is a reasonable approach to try to fully exploit any available data and develop multiple but simple groundwater models. It is for this purpose we have developed the Hydrogeological Information System (HgIS). Its purpose is to load data from available data sources of any kind (semi-structured data), to visualize and analyze data (to support formulation of alternative conceptual models) and to implement simple procedural models. HgIS handles different kinds of data (observations, spatial interpretations, documents and other files) in different ways. HgIS is mostly built upon the Pentaho platform. It uses ETL tool Pentaho Data Integration to load data to the database. The database PostgreSQL uses our data model. Observations can be displayed in the online map application. To formulate conceptual models, data can be explored in 3D hydrogeological visualization software EnviroInsite. The approach is simplified due to the compatibility of our data model with that of EnviroInsite. We have designed multiple reports and developed multiple analyses and models (identifying redox processes, hydrochemical type of water, average hydraulic gradient and multicriterial analysis). The architecture of the system corresponds to a spatial business intelligence solution (GeoBI) – a combination of business intelligence and GIS. Business intelligence technologies and tools have not previously been applied to groundwater data.

***Keywords**:* environmental data management; business intelligence; Pentaho; ETL

1 INTRODUCTION

Hydrogeologists solving practical groundwater problems are facing significant uncertainty, which is given by the lack of relevant data and the ambiguity of their interpretation. Mathematical models are used to understand the ongoing processes and to support decision-making. Because of the uncertainty of the data and its interpretation it is appropriate to create multiple simple models – a "multiple model ensemble" (Uusitalo et al. 2015) – that can be later developed into more complex models. Even simple models can help to select a strategy of further exploration and collection of data and to support preliminary decision-making.

In this context simple models can be closed-form solutions and spatial analyses, which often make use of aggregated data and use simplified assumptions (e.g. geometry of the modelling domain is simplified to a single rectangle). Such models usually compute for example: balance (e.g. of water or chemicals), water flux (Darcy's law and continuity equation), flux of solutes or residence time. Slightly different are spatial computations performed mostly by geographical information systems (GIS) – e.g. the pumping rate based on the distance from the fringe of the contaminant plume. Available screening modelling tools for contaminant transport (e.g. length of steady-state contaminant plume) can be considered as simple ones too. To support the formulation of conceptual models it is necessary to visualize and analyze data and to perform data aggregation and common calculation and estimations (e.g. average hydraulic gradient, geochemical background or redox conditions from chemical composition).

The above issues mentioned require an information system that facilitates collection of data from various semi-structured sources, visualization and analysis of data in order to create alternative conceptual models and implement the corresponding simple procedural models. Interoperability with third-party modelling software is also a requirement.

2 ANALYSIS

The data needed to develop groundwater models are of different types. They are time-dependent and fully 3D. The source data exist in different formats as databases, data exchange formats (e.g. DBF, XML or flat files), archive data (e.g. MS Excel or MS Word), spatial data (e.g. ESRI SHP, KML or geodatabases). These data need to be retrieved into a single data structure to be used together. Highly structured data usually do not contain any interpretation or additional knowledge. Therefore it is necessary to adequately store and process all types of data. Some should be saved in a structured form so it can be further used (creating graphs, tables, maps, cross sections, etc.). Other data are used ad-hoc in the form it was obtained in so it is sufficient just to be stored and accessible – e.g. in the file system. The data and processes are depicted in table 1 that stands for the data flow diagram.

Table 1. Data flows – data sorted from structured to unstructured

Data source →	Storage →	Usage	Content		
Structured and semi-structured data – obser-vations (databases, files)	Data warehouse	Reporting, visualization incl. geological profiles and cross sections, export	Data		
Spatial interpretation of data, other geodata	Standard-based storage	Maps, GIS		Information	
Documents	Stored with metadata	Ad hoc			Know-ledge
Other files	Storage, accessibility				

2.1 Existing EDMS

The above-mentioned requirements are met to a certain degree by “Environmental Data Management Systems” (EDMS). Some of these are EQuIS (earthsoft.com), SiteFX (earthfx.com), GW-Base (ribeka.com), WISKI (kisters.net), EnviroData (geotech.com), Oasis-montaj (geosoft.com), HydroManager (waterloohydrogeologic.com) or ESdat (esdat.net). Those tools usually have an excellent graphical user interface and are able to import dozens of data exchange formats. But they are not flexible enough to create new data imports because they do not contain an easy to use highly adaptable ETL (extract, transform and load) module. Some existing EDMS contain some reporting engine (e.g. SSRS, Telerik or Crystal Reports) to create high-quality reports. Without an ETL module they cannot efficiently combine data operations (e.g. aggregation), analyses and simple visualization (reporting). Those shortcomings are overcome by the Hydrogeological Information System (HgIS) presented here.

2.2 Business Intelligence

Groundwater information management can be described as loading of both archive and actual data (that are not modified anymore) from diverse (structured and semi-structured) sources; visualization of data in tables and graphs (downloadable in common formats as MS Word and MS Excel), data analysis, and model development. The same description corresponds to a completely different discipline – to Business Intelligence (BI), where BI uses data about a company to support its manager’s decision-making.
Therefore HgIS utilizes Pentaho – the BI platform. It is a Java based product of Pentaho Corporation with an open-source version (community.pentaho.com). The Pentaho platform contains ETL – Pentaho Data Integration PDI aka. Kettle. Reports designed by Pentaho Report Designer (PRD) can be run on a local computer or on the BI application server Pentaho Business Analytics (PBA). PBA facilitates users to design dashboards, analyze OLAP cubes etc. The Pentaho platform can be easily integrated or embedded to other applications.

3 HYDROGEOLOGICAL INFORMATION SYSTEM HGIS

HgIS is an information system developed at the Technical University of Liberec in the Czech Republic. Its purpose is to load data from the available data sources of any kind, to visualize and analyze data

(to support formulation of alternative conceptual models) and to implement simple models based on the data. Table 2 shows how specific kinds of data are managed in HgIS. Although it is focused on groundwater, it is also being used for broader range of environmental data.

Table 2. HgIS architecture

Data source	➡	Storage	➡	Usage
Observations (XML, MS Excel, flat files, SQL databases)	ETL (Pentaho Data Integration)	Data warehouse (PostgreSQL)	BI Platform (Pentaho)	Reporting, visualization, procedural models
			ETL (Pentaho Data Integration)	Complex visualization (EnviroInsite)
Spatial data (ESRI SHP, KML, raster images etc.)	GIS (QGIS), ETL (Geo-Kettle) etc.	Spatial database (PostGIS) + georefe-renced images	Map server (GeoServer)	Online map application, desktop GIS (e.g. QGIS)
Documents		Reference management software (Zotero)		Ad hoc
Other files		File system		

The management of unstructured data (documents and other files) can be performed using existing tools. Software development was focused on structured data – as is described in the rest of this paper. An earlier version of HgIS was presented in the paper (Nešetřil and Šembera 2014), that focuses on data management and contains more details on the data model and on the online map application. It also contains an illustrative schema and screenshots of HgIS.

3.1 Database – data warehouse

It is reasonable to use an existing data model for the newly developed information system. We have reviewed available data exchange standards and data models as Ground Water Markup Language (GWML), other application schemas of Geography Markup Language (GML), INSPIRE, Hg2O, Arc Hydro Groundwater, Data Model of National Groundwater Information System, H+ and some others. None of those data models were used. Besides some other issues, some were not suitable for needs of groundwater practitioner; some were too concise or not sufficiently documented. All data models, data exchange formats and data models of EDMS were reviewed and used as an inspiration for the developed data model.

Visualization of the hydrogeological data on a desktop computer can be easily performed with EnviroInsite from EI LLC (enviroinsite.com) – low priced software in .NET. It can be used to display maps (including localized tables and graphs), technical documentation of boreholes, geological cross-sections, 3D geological models and interpolation in 2D and 3D. The data model of HgIS is based on an existing data model of EnviroInsite. Therefore the database and the visualization software have consistent data structures that reduce the need for non-unique data transformation, and so it does not confuse users.

The original data model of EnviroInsite (9 tables) was extended to 36 tables because the EnviroInsite data model contains the data relevant for visualization only. The original tables were extended by additional fields and the model was further normalized. It contains data on: observation objects*, characterization of geological layers, technical construction of wells, definition of observed quantities*, action levels, definition of vertical intervals*, measurements tied to vertical intervals (e.g. chemical analyses or head measurements)*, measurements tied to specific depth (e.g. geophysical logging), sampling conditions, conversion of units (e.g. mg to g) and quantities (e.g. nitrate to nitrogen), anti-aliasing, time intervals, metadata, lookup tables etc. Tables containing data noted with asterisk (*) are organized to the snowflake schema.

We are using the PostgreSQL (postgresql.org) database management system. Interpretations and non-point data (arcs, polygons etc.) are stored in the PostgreSQL due to the spatial extension PostGIS (postgis.net). That spatial data and georeferenced images are served via GeoServer (geoserver.org) as the mapping service (e.g. WMS or WMTS), that can be loaded as a basemap to our web application or to any GIS.

3.2 ETL

The data are loaded to the database (data warehouse) by the ETL tool PDI. Data transformations in PDI can be implemented without coding through an intuitive graphical user interface and run also in command-line interface or on the ETL server. We implemented the loading of following data:

- analyses from laboratory information management system “Labsystém” (xBase files),
- geologic description and water quality from Czech Geological Survey (MS Access and XML files),
- geology of boreholes (MS Word documents created by a Geobanka software),
- flat files with precipitation and temperature from a watershed authority (text files via FTP server),
- database format of EnviroInsite (MS Access, MS Excel),
- general cross-table (MS Excel) and
- formats from some other data vendors (groundwater pumping, river discharges etc.).

Subsequent transformations provide data cleaning, anti-aliasing, validation and loading to the database. Coordinate conversion and loading of data to PostGIS is performed by GeoKettle – a spatially enabled fork of PDI. Ad-hoc loading is performed by common GIS software (e.g. QGIS – qgis.org). PDI is also used to export data to third-party simulation tools.

3.3 Visualization

Due to the compatibility of our data model with that of EnviroInsite, data can be easily exported to MS Access or MS Excel file and can be visualized in EnviroInsite. This is suitable for professional hydrogeologists to develop conceptual groundwater models. Stakeholders and other nonspecialists can view data in a web application we have developed. The application combines tables, a graph and a map on a single screen. Different map layers are provided as mapping services.
We have developed the following reports:

- The graph and the table of time development of an arbitrary quantity in arbitrary observation points and basic descriptive statistics.
- Profile of geologically documented borehole.

3.4 Analyses and models

We have developed the following analyses and the models that are reusable because of their general purpose and connection to the database.
Some analyses are utilizing PDI. Results are stored directly in the database as separate quantities:

- Data aggregation (e.g. total annual precipitation computed from daily precipitation, minimal monthly discharge in a year). Aggregations can be computed easily using PDI step “Group By”.
- Computation of the hydrochemical type of water (based on major cations and anions) – e.g. Ca-Mg-HCO3.

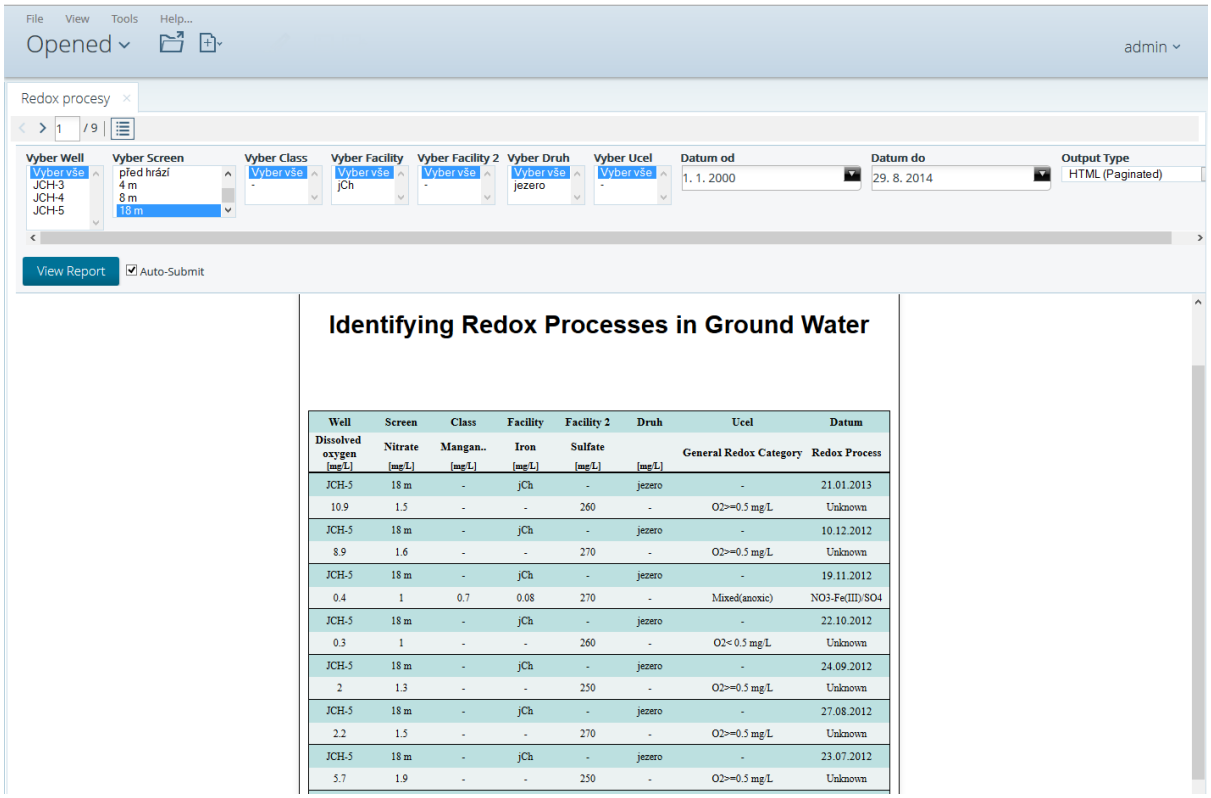
Some analyses and a model are utilizing PDI and formulas in Pentaho Reporting (OpenFormula), results are depicted in reports:

- Identifying redox processes in ground water from chemical composition (dissolved O_2, NO_3, Mn^{2+}, Fe^{2+}, SO_4^{2-} and sulfides) without measured Eh and pH (Chapelle et al. 2009) – Figure 1.
- Multicriterial analysis assessing water quality trends in correspondence to eutrophication. Aggregated values of quantities (nitrogen/phosphorus ratio, saturation of oxygen, pH etc.) were compared to estimated limits. The trend of the sum of the weighted logical values (overall score) indicates the trend in water quality.

Some analyses were performed in external tools (data were exported with PDI):

- Average hydraulic gradient was calculated from the hydraulic heads of selected boreholes. Consecutively seepage velocity and retention time were computed. This calculation is performed in MS Excel spreadsheet (Devlin 2003) by matrix formulas.
- To support geological interpretation of thousands of exploratory boreholes from former mining area we tested an automated classification of the detailed text characterizations of strata. We used RapidMiner software (rapidminer.com).

The above-mentioned analyses support conceptual model developments. The same techniques can be used to implement simple procedural site-specific groundwater models.



Well	Screen	Class	Facility	Facility 2	Druh	Ucel	Datum
Dissolved oxygen [mg/L]	Nitrate [mg/L]	Mangan.. [mg/L]	Iron [mg/L]	Sulfate [mg/L]	[mg/L]	General Redox Category	Redox Process
JCH-5	18 m	-	jCh	-	jezero	-	21.01.2013
10.9	1.5	-	-	260	-	O2>=0.5 mg/L	Unknown
JCH-5	18 m	-	jCh	-	jezero	-	10.12.2012
8.9	1.6	-	-	270	-	O2>=0.5 mg/L	Unknown
JCH-5	18 m	-	jCh	-	jezero	-	19.11.2012
0.4	1	0.7	0.08	270	-	Mixed(anoxic)	NO3-Fe(III)/SO4
JCH-5	18 m	-	jCh	-	jezero	-	22.10.2012
0.3	1	-	-	260	-	O2< 0.5 mg/L	Unknown
JCH-5	18 m	-	jCh	-	jezero	-	24.09.2012
2	1.3	-	-	250	-	O2>=0.5 mg/L	Unknown
JCH-5	18 m	-	jCh	-	jezero	-	27.08.2012
2.2	1.5	-	-	270	-	O2>=0.5 mg/L	Unknown
JCH-5	18 m	-	jCh	-	jezero	-	23.07.2012
5.7	1.9	-	-	250	-	O2>=0.5 mg/L	Unknown

Figure 1. Pentaho Business Analytics with the analytic report of water quality

Components of HgIS are being used in a state enterprise (that carries out recultivations) for data management and for the water balance model of a lake. HgIS is deployed and used within research projects at the Technical University of Liberec.
Aforementioned principles are being used also for simulation of nanomaterial transport in the subsurface. For the prediction of nanomaterial (especially zero-valent iron nanoparticles) behavior in the subsurface, it can be useful to use merged parameters that are describing macroscopic behavior (not specific physical and chemical processes). Besides common groundwater parameters (hydraulic conductivity, hydraulic gradient etc.) two more parameters are necessary: retardation (scaling the velocity of migration) and the first order reaction rate (scaling the mass decrease). Such parameters can be determined from column experiments and can be directly used in common solute transport codes. This strategy requires collection of data for specific soil types.

4 CONCLUSIONS

Heterogeneous subsurface environments are usually characterized by scarce data – therefore a common approach is to fully exploit available data and develop multiple but simple models. For this reason we have developed Hydrogeological Information System HgIS. Its purpose is to load data from any available data sources (also semi-structured data), to visualize and analyze data (to support formulation of alternative conceptual models) and to implement simple models based on the data. We used the tools (Pentaho platform) that enable “power users” to customize and extend the system. The architecture of the system corresponds to a spatial business intelligence solution (GeoBI) – a combination of business intelligence (BI) and GIS. Therefore it can be used also for geographical analyses and management of big data sets. BI technologies and tools have not been applied for groundwater data before. Groundwater practitioners have worked with GIS software for decades but not with BI tools. Our effort is to introduce BI to the groundwater community. HgIS is available commercially, upon request (contact the corresponding author).

Currently we are working to:

- Integrate all components (Pentaho platform, online map application and unstructured data) to a single graphical user interface.
- Automate processes and set event-based reporting with “solutions” (.xaction files).
- Design interactive dashboards.
- Simplify design of reports and dashboards for business users by creating an abstract business layer (Pentaho Metadata) including localization.

ACKNOWLEDGMENTS

The contribution was prepared with the financial support of FP7 project GUIDEnano "Assessment and mitigation of NM-enabled product risks on human and environmental health: Development of new strategies and creation of a web-based guidance tool for nanotech industries", no. 604387 and with the support of the Technology Agency of the Czech Republic via the project no. TA04020207. Work of programmers D. Krejbich, T. Jodas, D. Kendik, P. Štírek, J. Hadač and contractors is greatly acknowledged.

REFERENCES

Chapelle, F.H., Bradley, P.M, Thomas, M.A., McMahon, P.B, 2009. Distinguishing iron-reducing from sulfate-reducing conditions. *Ground Water*. 47 (2), 300–305, doi:10.1111/j.1745-6584.2008.00536.x

Devlin, J.F., 2003. A spreadsheet method of estimating best-fit hydraulic gradients using head data from multiple wells. *Ground Water*. 41 (3), 316–320, doi:10.1111/j.1745-6584.2003.tb02600.x

Nešetřil, K., Šembera, J., 2014. Groundwater data management system. In: Gómez, J.M, Sonnenschein, M., Vogel, U., Winter, A., Rapp, B., Giesen, N. (Eds.), EnviroInfo 2014 – ICT for Energy Efficiency : Proceedings of the 28th International conference on informatics for environmental protection. Oldenburg, Germany, pp. 301–306, http://www.iai.kit.edu/ictensure/site?mod=litdb&subject=art&pid=X13287035&action=detail

Uusitalo, L., Lehikoinen, A., Helle, I., Myrberg, K., 2015. An overview of methods to evaluate uncertainty of deterministic models in decision support. *Environmental Modelling & Software*. 63, 24–31, doi: 10.1016/j.envsoft.2014.09.017