GMP Data Warehouse – a Supporting Tool of Effectiveness Evaluation of the Stockholm Convention on Persistent Organic Pollutants

Jakub Gregor, Jana Borůvková, Richard Hůlek, Jiří Kalina, Kateřina Šebková, Jiří Jarkovský, Ladislav Dušek, Jana Klánová

HAL Id: hal-01852613
https://inria.hal.science/hal-01852613

Submitted on 2 Aug 2018

GMP Data Warehouse – a Supporting Tool of Effectiveness Evaluation of the Stockholm Convention on Persistent Organic Pollutants

Jakub Gregor[1,2,*], Jana Borůvková[1], Richard Hůlek[1,2], Jiří Kalina[1,2],
Kateřina Šebková[1], Jiří Jarkovský[2], Ladislav Dušek[2], Jana Klánová[1]

[1]Research Centre for Toxic Compounds in the Environment, Masaryk University,
Brno, Czech Republic
[2]Institute of Biostatistics and Analyses, Masaryk University,
Brno, Czech Republic
gregor@iba.muni.cz

Abstract. The Stockholm Convention on Persistent Organic Pollutants is multilateral environmental agreement focused on selected persistent organic pollutants (POPs) for which the contracting Parties must adopt measures to eliminate or reduce their production and use or minimise the unintentional releases. One of the tools for the effectiveness evaluation of the Stockholm Convention is Global Monitoring Plan for Persistent Organic Pollutants (GMP) – a project that aims to collect global data on POPs concentrations in selected environmental matrices. This paper introduces an information system GMP Data Warehouse, which was developed in order to provide user-friendly tools for the collection, storage, analyses and visualisation of data from international POPs monitoring activities.

Keywords: Stockholm Convention, Global Monitoring Plan, POPs, Database, Data collection, Visualisation, Analysis, Standardisation, Data Structure.

1   Introduction

The Stockholm Convention on Persistent Organic Pollutants was adopted on 22 May 2001 in Stockholm, Sweden, and entered into force on 17 May 2004. The convention is focused on selected persistent organic pollutants (POPs) – chemicals that represent a significant risk for the environment and living organisms, including the humans.

Although POPs form a heterogeneous group from the chemical point of view, their common characteristics include acute or chronic toxicity and high resistance to transformation processes, which makes them capable of long-range transport and accumulation in tissues of the living organisms.

The Stockholm Convention (SC) and its annexes currently (2016) contain 26 selected POPs (or their groups), for which the contracting Parties must adopt measures to eliminate or reduce their production and use or minimise the unintentional releases. In other words, the list contains both chemicals that were or have been intentionally

produced and used (e.g. DDT and other POP pesticides in agriculture, polychlorinated biphenyls in industry) and chemicals that are unintentionally formed and released during anthropogenic processes (e.g. production of dioxins during combustion processes).

Naturally, adoption and application of (legal/technical) measures for the reduction of environmental burden by POPs is not the only step that should be implemented; these measures should be also continuously evaluated in terms of their feasibility and effectiveness. Effectiveness evaluation of the Stockholm Convention is defined in its Article 16. Among others, this article requires establishment and operation of the Global Monitoring Plan for Persistent Organic Pollutants (GMP) – a tool for the collection of global data on POPs levels, assessment of their spatial and temporal trends and thus generating information on whether the environmental burden by POPs decrease and measures adopted by the Convention are effective in reality.

2   Stockholm Convention's Global Monitoring Plan for POPs

GMP implementation was officially endorsed in 2007 by the decision. SC-3/19, of the 3rd meeting of the Conference of the Parties to the Stockholm Convention. Main objective of the GMP is global collection of available data on POPs concentrations in selected matrices – ambient air, human milk, human blood, and water (for hydrophilic substances). Article 16 of the SC expects collection of data from existing environmental monitoring programmes however it allows for capacity building assistance and establishment of new monitoring programmes only in those regions with data gaps.

Request for harmonised and standardised data collection framework and their assessment is set up in a document entitled "Guidance on the Global Monitoring Plan for Persistent Organic Pollutants" [1]. The first draft document was approved simultaneously with the establishment of the GMP framework in 2007 and has undergone several updates since then, due to the addition of several chemicals to annexes of the Convention and need for detailed updated data handling and assessment guidelines.

Data collection for GMP purposes is performed in a six-year interval and the results including interpretations, conclusions, and recommendations are published in five regional reports (UN regions) and one global monitoring report. The reports are prepared by Regional Organisation Groups (ROGs) and a Global Coordination Group (GCG). These bodies consist of experts, representatives nominated by the Parties to the SC, with a significant knowledge in environmental monitoring and data management.

3   GMP Data Warehouse development and tools

It is noteworthy that the collection of data from different sources (i.e. monitoring programmes) is not a trivial issue. Each national or international programme has a specific purpose, different design and methods of sampling and chemical analyses, the data are collected and stored in different forms and structures. Hence, their global comparison is a challenge that emerged immediately after the first data collection

campaign in 2008. It is therefore necessary to set up a really robust and flexible methodology that allows merging such heterogeneous data together and performs their comparison, assessment and analysis. The published monitoring reports of 2008 showed that there is a strong need for a standardised electronic tool for both data collection and their presentation.

Establishment of a single data warehouse, in which all GMP data would be stored, is proposed in Chapter 6.5.2 of the GMP Guidance document. The GCG, ROGs, and Secretariat of the SC considered, supported and approved establishment of such at their meetings in 2011–2013. The concept and development of the electronic Global Monitoring Plan Data Warehouse (GMP DWH) was endorsed by the decision of the Conference of the Parties to the Stockholm Convention SC-6/23 at its 6th meeting in May 2013. Consequently, they authorised two research departments of the Masaryk University (Brno, Czech Republic) to design and develop a comprehensive information system for the collection, analysis and visualisation of GMP data in accordance with the following principles:

- fully electronic data processing
- standardised parametric data structure, standardised predefined code lists
- defined processes for the data input, validation, approval, and publication
- defined hierarchy of users and user rights with respect to access to the system and data processing
- connection of the database with visualisation and export tools
- online access to all tools and functions
- system of user support, help desk

Goal of the GMP DWH is to provide long-term reliable and cost-effective information and services to global community, support POPs monitoring activities and data management under the Stockholm Convention and offer tools for collection, storage, organisation, comparison, analysis, and evaluation of performance in relation to monitoring programmes on POPs. The objectives of the online GMP DWH are therefore twofold:

- provide user friendly tools for storage and analyses of data from international monitoring activities under the Global Monitoring Plan of the Stockholm Convention on Persistent Organic Pollutants and make POPs data visualisation available for regions and programmes that require support in data management, and
- contribute to the effectiveness evaluation of the Stockholm Convention by compiling and visualising results of global monitoring activities on POPs.

Development of the GMP DWH system was completed in 2014 and the system was handed over to the global community and main user (Secretariat of the Stockholm Convention). The GMP DWH structure has been designed to incorporate state of the art knowledge and expertise in building knowledge-based infrastructures. It encompasses data input, storage, processing (compiling and archiving) of both primary data as well as aggregated data, including supplementary data in cases where no primary data are made available. The system holds data on POPs in four core matrices: air,

human milk, human blood, and water. By respecting the requirements of uniform and harmonised presentation of data, all outputs of the GMP DWH are shown on the visualisation portal (http://visualization.pops-gmp.org/2014/).

The GMP DWH architecture consists of three layers guiding the data flow from the initial upload to the final publication. Each layer is connected with appropriate tools and processes of the data flow, user administration, security, user support etc. (Fig. 1) Data repository and data visualisation are the main parts of the system accessible by a wider group of users. No additional software is required to use the system, because it is implemented fully online and is accessible through standard internet browsers (Internet Explorer, Mozilla Firefox, Google Chrome). It is highly recommended to use their recent versions, which are freely available and ensure proper functioning, particularly for visualisations and graphic outputs.

3.1   Data repository

The data repository refers to a database in which all data are stored and administered. Data are collected and inserted in a predefined structure, which is partly identical for all monitored environmental matrices (ambient air, human milk, human blood, water); however, it also reflects certain specificities of the individual matrices and nature of the data (primary vs aggregated). The individual items of data structure are logically grouped into three levels (Fig. 2). The top level ("Site") defines and describes the place in which sampling was performed. The middle level ("Sampling attributes") contains information about the sampling period and employed methods. The third level ("Measurement") provides data about measured chemical substances, their concentrations, statistical variability etc. Such predefined structure ensures a full comparability of collected data and also prevents gaps in their completeness.
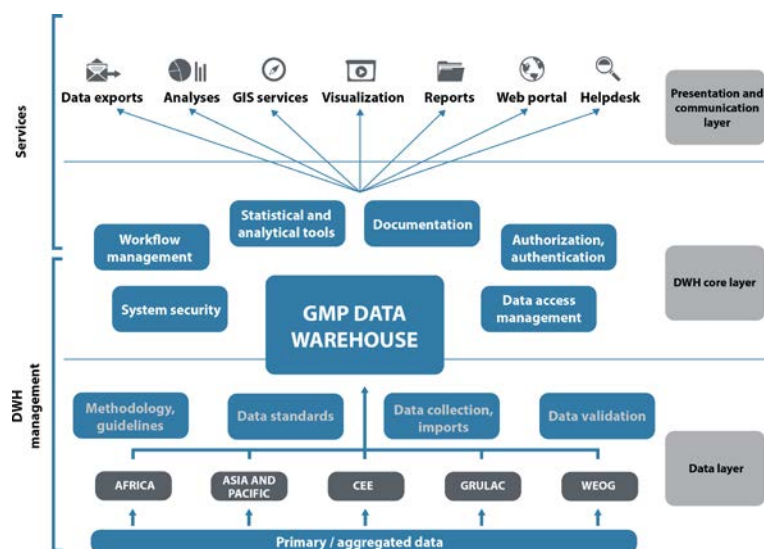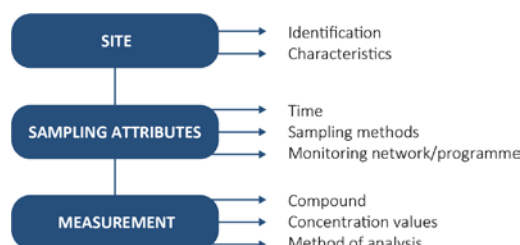


Fig. 1. GMP Data Warehouse architecture

Fig. 2. Individual items of the GMP DWH data structure are grouped into three layers

The GMP DWH is designed to store aggregated data with one-year granularity. This consensus works under the assumption that a large part of data from the monitoring programmes are not available in their primary format and that POPs concentration show, particularly in temperate regions, fluctuations due to seasonal variations in the course of the year. However, data providers are allowed to send/upload both primary or aggregated data using online forms or MS Excel sheets of a predefined structure. The data sets are subsequently processed, validated and aggregated (where applicable) by the GMP DWH administrators. The data insertion process is shown in Fig. 3.

Validated data are subject to approval process of the relevant geopolitical expert group ROGs; it is only relevant ROG members are allowed to approve particular data records for publication (in monitoring reports, online visualisation, data exports etc.) or reject any public use of these data in relation to the GMP.

The previous paragraphs imply that the system is accessed by various users and groups of users in terms of data management (data providers, administrators, ROG members) and geographical affiliation (individual UN Regional Groups). Management of users and user rights is therefore one of the crucial points in the system design. Access rights must be precisely defined for each individual user so that he/she could only view and process those data records that fall under his/her competence from both of the above-mentioned points of view. Data flow and the definition of user roles and rights are further complicated by the fact that the same user can be, for example, a representative of data-providing institution and a ROG member at the same time, or a European institution may provide data to ROG Africa etc. Development of the processes and rules for the management of users, their roles and rights (Fig. 4) was therefore another significant issue. This is closely linked to data security, which is guaranteed by a data security management certificate according to the ISO/IEC 27001:2014.

There are also other very important parts of the system, such as the statistical and analytical tools for data validation, correct aggregation, and last but not least, the assessment of time trends in POPs levels. For this purpose, a special R package [2] was developed in cooperation with the Environment Canada, which is able to perform all necessary steps. The software is described in details on a special website (http://www.genasis.cz/time-series/), which also provides a practical demonstration of the package [3]. See Chapter 3.3 for details.
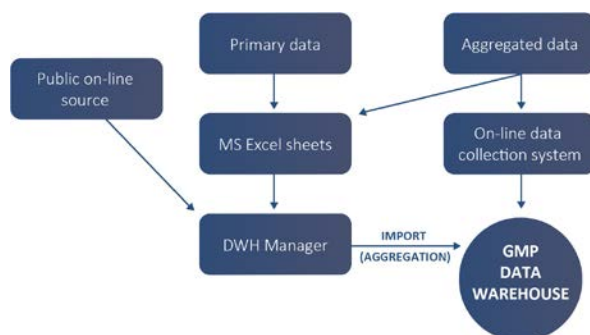
Fig. 3. The process of data insertion into the GMP DWH
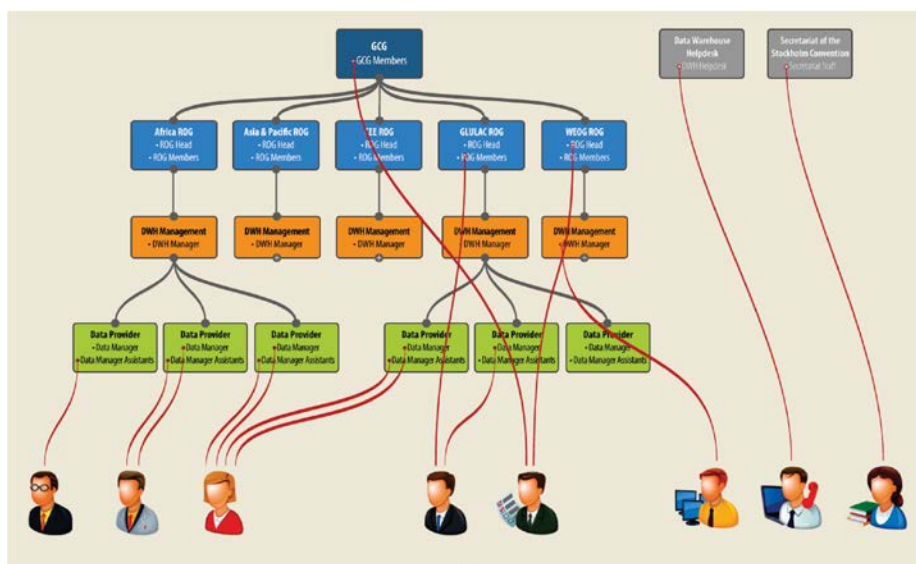
Fig. 4. Assignment of user roles and user rights in the GMP DWH

3.2   Data visualisation

Work with large data sets often requires some type of visualisation in order to understand their nature and distribution, to compare different data sets, or to identify outliers. The website http://visualization.pops-gmp.org/2014/ [4] was designed to provide such basic overview of data in the GMP DWH; moreover, it also offers advanced analytical and mapping tools that enable the user to browse and to analyse data from different points of view:

- "Spatial distribution" – a map overview of sites for which data are available. The sites can be stratified according to a set of criteria.
- "Data availability" – a plot describing data availability in time (years in which sampling ran) and for individual POPs (compounds that were analysed).

- "Summary statistics" – assessment of measured concentrations of a particular compound at individual sampling sites and years by means of descriptive statistics. The user can switch between various types of central value (mean, median) and variability (percentiles, min–max, standard deviation).
- "Time series" – three different views on time trends of POPs levels globally. Two of them are maps that display statistical assessment of long-term trends, i.e. whether the trend is increasing, decreasing, or statistically insignificant. The third tool is a plot that shows annual values of POPs concentrations in time, trend regression and outputs of statistical tests.
- "Data exports" – this provides an overall summary of selected dataset i.e. contributing monitoring programmes, sites, and analytical methods included in the current data set, and also tools for export of the dataset to files that may be processed in MS Excel.

The above-mentioned modules (Fig. 5) are further enriched by other functions to facilitate the work with the system. Graphical outputs can be directly adjusted according to user needs (ordering values, type of values) and they can also be exported into a PNG format in order to be used in further work. Each combination of filters and each view on the selected data set are defined by a unique code which is included in the URL of a particular view. This significantly enhances linking and sharing of information through various internet communication channels.
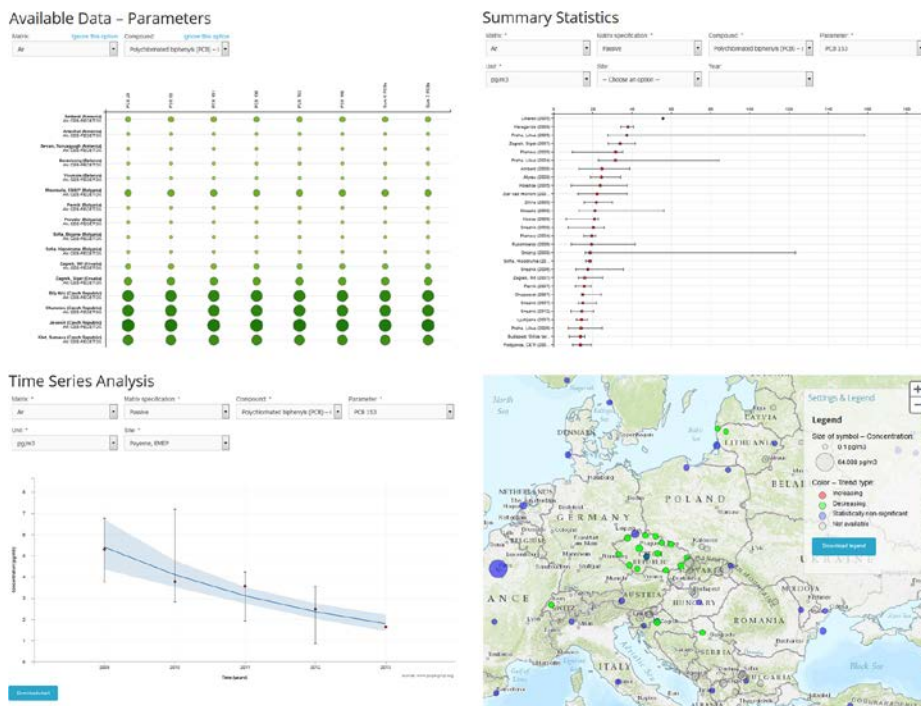

**Fig. 5.** Example of graphical outputs in the GMP DWH visualisation tool

3.3 Employed information technologies and approaches

The GENASIS system [5] was used as a platform for handling primary data and their transformation into the form and structure required by the GMP Guidance [1] and GMP DWH standards (see Chapter 3.1). An OpenCPU technology was used for integration of the R software environment into the GENASIS infrastructure [6]. The R package developed in cooperation between the Masaryk University and Environment Canada [2] ensured harmonisation of data on the level of data standards and metadata. In particular, this tool was used for the recalculation of air-passive sampling data to volume-based values, which ensured their comparability with data from the air-active monitoring programmes. Further steps involved the calculation of derived parameters required by the GMP Guidance (sums of related compounds, toxicity equivalents – TEQs) and the summation of individual sampled fractions (for air data). Another (unpublished) R package was used for the annual aggregation of primary data.

The data visualisation tool employs its own database layer, which is updated in 24-hour intervals on the basis of the central data warehouse content. The data and data views are partially pre-prepared and pre-calculated to enable a faster and more fluent operation, calculations and loading. Almost all charts displayed within the data visualisation are custom-made and have been implemented using the D3.js library, since common available visualisation libraries do not contain required types of charts.

Three different servers technologies are involved in the data visualisation: application/web server (Apache), ArcGIS Server and OpenCPU R-server. The application server provides web services and ensures the system security. The map server employs the ArcGIS for Server and ArcGIS for Javascript technologies to create map compositions that obtain information on sites from the database and display sites that correspond to the user's selection in map windows of the web application. The R-server performs real-time calculations of time trends for selected sites, compounds and period [7] (Fig. 6).
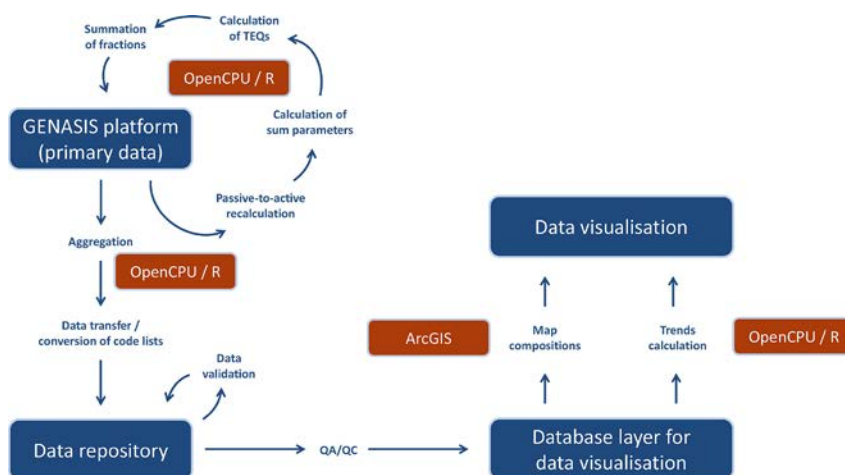
Fig. 6. Procedures and software tools employed in processing of the POPs data on their way from the primary database (GENASIS platform) to the GMP data visualisation

4 GMP DWH operation and user support

The final version of the GMP DWH was provided to the main user (Secretariat of the Stockholm Convention) in 2014, in which the second GMP data collection took place. In contrast to the first data collection period, these data were collected digitally in a parametric and structured form, which enabled both standardisation and comparability of outputs that were presented in regional monitoring reports. Monitoring reports were adopted and data visualisation was made publicly available at the 7th meeting of the Conference of the Parties to the Stockholm Convention in May 2015.

All tools and parts of the GMP DWH are accessible through the web portal www.pops-gmp.org/dwh. The website contains all important information about the system, links to individual tools, background information, user guides, overview of data structure and reporting spreadsheets (Fig. 7).

Fig. 7. Web portal of the GMP DWH (www.pops-gmp.org/dwh)

4.1 Providing user support during implementation of the GMP data tool

Help desk was introduced as soon as the first version of the GMP DWH emerged in 2013. This service supported all GMP DWH users and other relevant persons interested in the project. The service is available on weekdays from 8am to 4pm CET via e-mail and phone. Help desk operators either immediately provide requested information or assistance, or in case of a more complicated request, allocate the task to another expert/data manager. The help desk performance is in agreement with requirements of the international standard ISO/IEC 20000-1:2012 and guarantees that response will be provided within 24 hours.

The main tasks of the service include:

- technical assistance to all GMP Data Warehouse (GMP DWH) users
- transfer of requests to appropriate DWH managers
- administration of user accounts and associated rights
- contact point for data files from the providers, their formal validation and transfer to DWH managers
- contact point for error reports and suggestions for further software and services development

The list below shows the range and types of requested hotline services and support in 2014 (a period in which the data collection, validation, and approval were carried out and the visualisation tool was not yet publicly available):

- Reception and validation of data files from the Environment Canada, communication with persons responsible for data transfer (3 persons). Most issues discussed were related to correct data format and terminology (~ 10 requests)
- Additional correction in data already provided and entered in the database (~ 10 requests)
- Providing ROG members and their consultants with access to the database and visualisation, preparation of instructions and user guides (43 user accounts)
- Support to users experiencing technical problems with the visualisation tool. These technical issues were caused by older versions of web browsers used by users. (~ 5 requests)
- Identification of bugs in the visualisation tool in cooperation with the Secretariat members and other users (~ 15 requests)

4.2 Attendance of the GMP data visualisation website

Google Analytics were employed for the assessment of attendance of the GMP visualisation website. Table 1 summarises the main characteristics of users' behaviour when browsing the website during the first year from publication (from 1 May 2015 to 1 May 2016). The data indicate, among others, a relatively long duration of an average session and a high number of pages viewed. The highest attendance peaks were recorded on 11 June and 28 July 2016, when the GMP DWH was presented at the BRS Conventions' Twitter account and newsletter, respectively.

Table 1. Users' behaviour on the GMP data visualisation website in the first year of public operation (assessed by Google Analytics)

Parameter	Counts
Sessions	2,025
Users	818
Page views	30,012
Pages per session	14.82
Average session duration	11 min 8 sec

4.3 Current use of the GMP DWH

Functional and content updates of the system were temporarily discontinued on the date of its official publication and of publication of regional monitoring reports in May 2015. In the period preceding the final publication, the data visualisation module was mainly used by ROG members and consultants for data approval and compilation of the monitoring reports. It now serves as a supplementary and supporting tool to the monitoring reports that enhances browsing, viewing and analysing the global POPs data. It is also presented as a reference tool in the communication with new data providers and partners in order to identify new potential data sources worldwide.

5 Conclusion

A multimodular online GMP DWH was developed for data collection, processing and reporting for current and future GMP data collection rounds. The system is based on fully parametric data sheets to improve the quality of collected global data sets on POPs concentrations, to determine their fate in the environment over time and to strengthen the responsibility and visibility of data providers. The developed online tool for the Global Monitoring Plan is capable of visualising information available globally or from regional and country perspective. The GMP DWH is accessible for anyone with internet connection on a dedicated portal www.pops-gmp.org/dwh. The system is customised to function in all standard internet browsers (Internet Explorer, Mozilla Firefox, Google Chrome).

6 References

1. United Nations Environment Programme: Guidance on the Global Monitoring Plan for Persistent Organic Pollutants. Secretariat of the Stockholm Convention on Persistent Organic Pollutants, Geneva (2007)
2. Kalina, J., Klánová, J., Dušek, L., Harner, T., Borůvková, J., Jarkovský J.: genasis: Global ENvironmental ASsessment Information System (GENASIS) computational tools. Masaryk University, Brno (2014). https://cran.r-project.org/web/packages/genasis/index.html
3. Kalina, J., Jarkovský, J., Dušek, L., Klánová, J., Borůvková, J., Šnábl, I., Šmíd, R.: Time series assessment in the Era of Stockholm Convention & GMP. Masaryk University, Brno (2014). http://www.genasis.cz/time-series/

4. Hůlek, R., Borůvková, J., Gregor, J., Kalina, J., Bednářová, Z., Šebková, K., Melkes, O., Šalko, M., Novák, R., Jarkovský, J., Dušek, L., Klánová, J. Global Monitoring Plan of the Stockholm Convention on Persistent Organic Pollutants: visualisation and on-line analysis of global levels of chemicals in air, water, breast milk and blood. Masaryk University, Brno (2014). http://visualization.pops-gmp.org/2014/.

5. Hůlek, R., Jarkovský, J., Kubásek, M., Gregor, J., Hřebíček, J., Dušek, L., Klánová, J., Šebková, K., Borůvková, J., Holoubek, I. GENASIS System Architecture: on the Way from Environmental Data Repository towards a Research Infrastructure. In: Hřebíček, J., Schimak, G., Kubásek, M., Rizzoli, A.E. (eds.) Environmental Software Systems: Fostering Information Sharing. IFIP AICT, vol. 413, pp. 230-239. Springer, Heidelberg (2013).

6. Hůlek, R., Kalina, J., Dušek, L., Jarkovský, J.: Integration of R Statistical Environment into ICT Infrastructure of GMP and GENASIS. In: Hřebíček, J., Schimak, G., Kubásek, M., Rizzoli, A.E. (eds.) Environmental Software Systems: Fostering Information Sharing. IFIP AICT, vol. 413, pp. 240-252. Springer, Heidelberg (2013).

7. Kalina, J., Hůlek, R., Borůvková, J., Jarkovský J., Klánová, J., Dušek, L.: Three Levels of R Language Involvement in Global Monitoring Plan Warehouse Architecture. In: Hřebíček, J., Denzer, R.., Schimak, G., Argentm R.M. (eds.) IFIP AICT, vol. 448, pp. 426-433. Springer, Heidelberg (2015).