Jul 12th, 11:30 AM - 11:50 AM

Global Monitoring Plan Data Warehouse as a tool for worldwide assessment of persistent organic pollution

Jiří Kalina
Masaryk University, Research Centre for Toxic Compounds in the Environment, kalina@mail.muni.cz

Jiří Hřebíček
Masaryk University, Institute of Biostatistics and Analyses

Richard Hůlek
Masaryk University, Institute of Biostatistics and Analyses

Jana Borůvková
Masaryk University, Research Centre for Toxic Compounds in the Environment

Jana Klánová
Masaryk University, Research Centre for Toxic Compounds in the Environment

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

Global Monitoring Plan Data Warehouse as a tool for worldwide assessment of persistent organic pollution

Jiří Kalina[1], Jiří Hřebíček[2], Richard Hůlek[2], Jana Borůvková[1], Jana Klánová[1]

[1]Masaryk University, Research Centre for Toxic Compounds in the Environment,
Kamenice 735/5, 625 00 Brno, Czech Republic
[2]Masaryk University, Institute of Biostatistics and Analyses,
Kamenice 126/3, 625 00 Brno, Czech Republic
Corresponding author: kalina@mail.muni.cz

Abstract: In 2004 the Stockholm Convention came into force, which eliminates a manufacture and use of persistent organic pollutants. Global Monitoring Plan is an important executive component of the convention, providing framework for an evaluation of its effectiveness. It ensures collection of comparable monitoring data on POPs from around the world in order to identify changes in POPs concentrations over time and their eventual transport. There are several programs focused on a monitoring of persistent organic pollutants in diverse environmental matrices over the globe, each working with a different methodology of sampling and analyses. Results of these programs are reported in 6-years periods in form of 5 distinct regional monitoring reports prepared with a different design and content, which does not enable a comparison of the outcomes and analysis contained data. Taking into account a practice from the first reporting campaign (2003–2008) a joint team of Institute of Biostatistics and Analyses and Research Centre for Toxic Compounds in the Environment of Masaryk University prepared an efficient web-based system (data warehouse) for data collection, treatment, storage and presentation, which allows both to create unified outcomes for the regional monitoring reports and analyse data from all individual sources in one place. During its introductory use in 2014, it was filled by nearly 50,000 records from around the world and now serves as a unique publicly available source of global data on POPs.

Keywords: Stockholm Convention, Global Monitoring Plan, GMP, persistent organic pollutant, POP, data warehouse, scientific repository.

1 INTRODUCTION

Persistent organic pollutants (POPs) are chemical compounds, usually man-made, which are typical by their very long lifetimes in the environment. Environmental persistence is one of their properties, derived from a stable chemical structure, which was very often also the reason of their production and use in form of pesticides, insulators, coolants, dyers, lubricants, additives and in many other technical applications. Many of POPs are known by their harmful impacts on human health. Due to their persistence and lipofility, they tend to bioaccumulate within the food chains and express carcinogenity or endocrine disruption of different levels.

Stockholm Convention (SC) was signed by 128 parties in 2001 in order to protect the environment and human health from harmful effects of POPs. It is in force from 2004, turning an approach to the POPs all around the world by an elimination or significant restriction of their manufacturing and application. Most of production and use of listed compounds was stopped, but a huge amount of POPs, although still slowly decreasing in general, is still present in the environment. Thus a detailed and reliable analysis of the still developing situation is needed to the evaluation of SC effectiveness. The Conference of the Parties of SC established an arrangement of the Global Monitoring Plan (GMP) to

provide a comparable monitoring of the presence and levels of the listed POPs as well as their regional and global environmental transport (Klánová, 2014).

Ambient air, human blood, breast milk and ambient water were selected as core matrices for the assessment of the persistent organic pollution. There are several programs of sampling of these matrices around the world, whose results were reported by individual representatives of parties to the GMP. In total there are data from 9 programs of active and 5 programs of passive air sampling, 3 programs of fresh and 19 programs of marine water sampling and 2 programs of human matrices sampling. Such number of data providers raises a question about the comparability of reported data and the ability to pass individual pieces of information into a coherent picture of worldwide POPs pollution. This is not possible without appropriate tool for data collection, processing and final assessment.

2 GLOBAL MONITORING PLAN DATA WAREHOUSE STRUCTURE AND FUNCTIONALITY

Five GMP regional organization groups (ROGs) are responsible for implementation of the global guidance document (UNEP, 2007) and the GMP within their regions (UN, 2014), taking into account regional realities. ROGs are operational units for data and information gathering, analysis, and preparing the regional outcomes. They collect data for six years long reporting campaigns (first period GMP I 2003–2008, second period GMP II 2009–2014) and publish it in a form of a regional monitoring report, following an agreed format (UNEP, 2007).

In order to achieve the GMP objectives – evaluation of POPs concentration levels and their trends – ROGs covers both existing international and regional programs and activities and separate studies/publications that may contain relevant information, and also should be open to public scrutiny to identify missing sources. Moreover, ROGs allow countries representatives to review the proposed data sources that may be related to their national situation. The regional reports are focused on collection and analysing samples of the four matrices, comprising wide spectrum of sampling methods, techniques and frequencies and demanding very cautious approach to their comparison and assessment.

However the form of regional reports does not allow any kind of direct comparison, standardization and calculation of statistics such as global levels of the pollution, concentration gradients, trends etc. In order to assess the data using any mathematical or statistical methods, it was necessary to gather all data into one database, enabling to treat it in a uniform manner.

Based on the tasks identified by the Global Coordination Group for GMP and the Secretariat of the Stockholm Convention, the Research Centre for Toxic Compounds in the Environment (RECETOX) and the Institute of Biostatistics and Analyses (IBA), Masaryk University, Czechia, performed a content analysis of the GMP I collection campaign data (up to 2008) in 2012. In a next step, an electronic tool for data storage and on-line visualization was prepared as a first part of a comprehensive IT solution for the second (2009–2014) and next collection campaigns (Gregor et al., 2013).

A complex data collection, storage and visualization web-accessible system was finished during the period 2012–2014, enabling to all ROGs and their authorized representatives to input data on POPs in both the primary and annually aggregated form, retreat and validate it and produce a comprehensive outline of the final regional report. Moreover, once the data are stored within the system, it is available for a variety of statistical analyses accessible via web browser to the general public. This system was of wide use during 2014 when the second collection campaign regional reports were prepared and published.

The system features three layers of data management – basement layer of data collection, directly linked with a treatment and storage layer. These layers form together the GMP data warehouse (DWH), completed by the assessment and visualization layer, making accessible all results of implemented statistical and analytical methods.

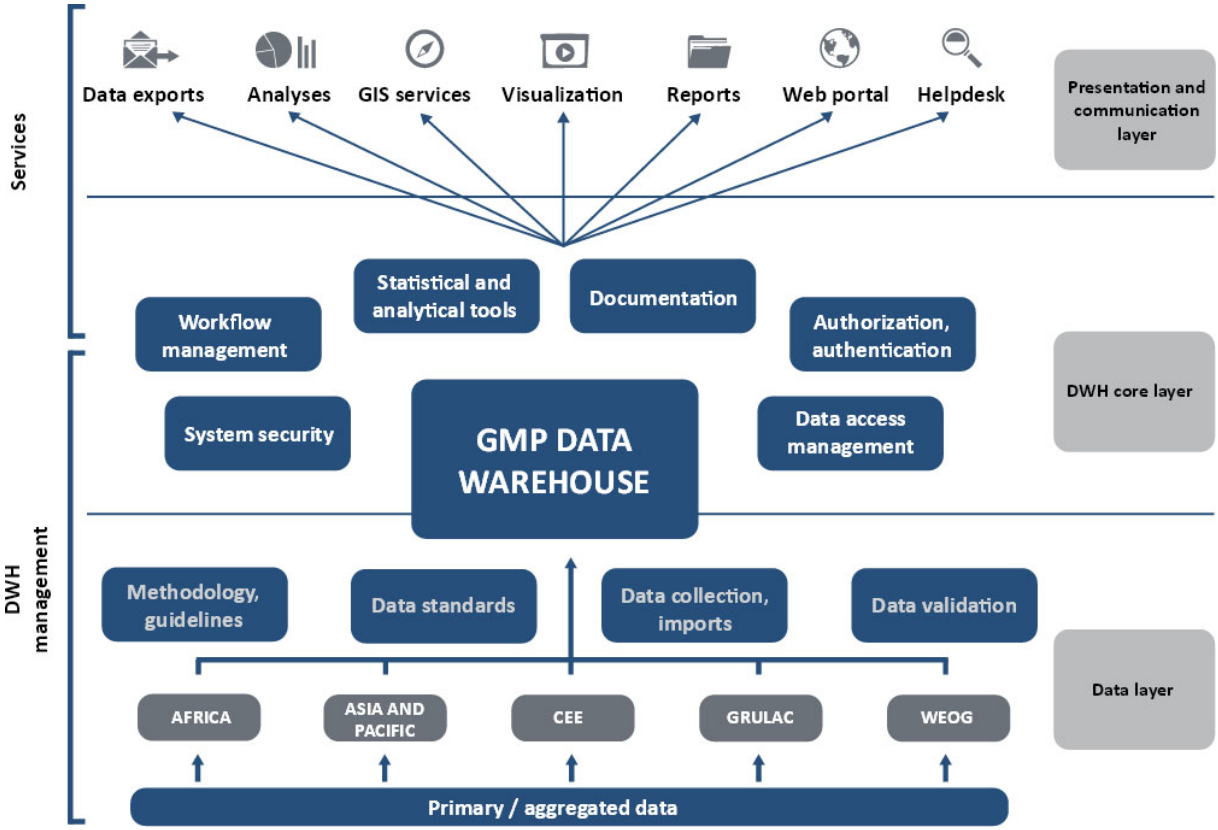

Figure 1. Scheme of GMP DWH and visualization system.

2.1 Data input points

There are three options of data input implemented in the data layer. Direct input of annually aggregated values from POPs monitoring using the web-interface available for authorized representatives of the ROGs, filling a MS Excel template for each of the four matrices by primary or aggregated data or data transfer from publicly available data source upon request of the ROG (both aggregated and primary data again).

The system encompasses also several data treatment modules, such as identification of outliers (with a final manual check), computation of sum parameters (sum of 6 indicator PCBs, sum of DDT congeners and derivatives), weighted sums of toxic equivalents (TEQ of polychlorinated dibenzodioxines and dibenzofurans) and sums of different fractions of selected matrices (gas and particle fractions of air and water sampling). Results of passive sampling are recalculated using the Global Atmospheric Passive Sapling template, developed continuously by Environment Canada (Harner, 2014).

Specific assumptions are required for the annual aggregation step. A result of the aggregation is computed as an arithmetic mean of a batch of primary values within one calendar year, although several other statistics, such as minimal and maximal value within the year, median, geometric mean, standard deviation, frequency of sampling, number of samples etc. are computed and stored within the DWH core layer. A minimal number of primary values to aggregate is chosen with respect to site and compound specific characteristics, such as matrix, known or expected fluctuations over the year, possible outlaying values etc. In case, that the aggregation step is realized within the GMP DWH, an infrastructure of Genasis data repository with all available features is deployed and directly linked to the GMP DWH. The Genasis repository with its modules for data aggregation, recalculations and other data management features was well described in (Jarkovský et al., 2013), (Hůlek et al., 2013), (Kalina et al., 2015).

Table 1. Monitoring programs that provided data to GMPII reporting campaign (2009–2014).

Matrix	Specification	N. of POPs	N. of Sites	First Year	Last Year	Monitoring Program
Air	Active	16	5	1992	2010	AMAP
Air	Active	14	1	2006	2012	Chemicals in Environment, Japan
Air	Active	21	37	1998	2013	China National POPs Monitoring
Air	Active	12	18	1991	2011	EMEP
Air	Active	13	5	1990	2013	IADN
Air	Active	21	1	1996	2013	Kosetice
Air	Active	16	1	1992	2012	Northern Contaminants Program
Air	Active	12	8	2009	2013	POPs Monitoring in East Asian Countries
Air	Active	6	8	2004	2012	Toxic Organic Micro-Pollutants Program
Air	Passive	12	66	2004	2010	GAPS
Air	Passive	3	11	2011	2012	GAPS-GRULAC
Air	Passive	23	41	2010	2013	GMP UNEP
Air	Passive	15	13	2010	2012	LAPAN
Air	Passive	17	122	2003	2014	MONET
Milk	Pooled	13	27	2007	2011	China National POPs Monitoring
Milk	Pooled	23	79	1987	2014	GMP UNEP
Water	Fresh water	1	6	2013	2014	MONET
Water	Fresh water	1	2	2013	2013	China National POPs Monitoring
Water	Fresh water	1	5	2014	2014	GMP UNEP
Water	Marine water	1	12	2011	2013	China National POPs Monitoring
Water	Marine water	1	1	2014	2014	GMP UNEP
Water	Marine water	1	25	2008	2008	Alcor
Water	Marine water	1	4	2008	2008	Amundsen
Water	Marine water	1	19	2007	2007	ANTXXV
Water	Marine water	1	34	2010	2010	ANTXXVII/1
Water	Marine water	1	17	2010	2011	ANTXXVII/2
Water	Marine water	1	25	2009	2009	ARK-XXIV/3
Water	Marine water	1	35	2009	2009	Endeavor
Water	Marine water	1	12	2005	2005	Ga 442
Water	Marine water	1	39	2005	2005	Ga 446
Water	Marine water	1	92	2010	2011	Malaspina
Water	Marine water	1	18	2007	2007	Maria S. Merian 2007
Water	Marine water	1	49	2008	2008	Maria S. Merian 2008
Water	Marine water	1	23	2005	2005	Oden 2005
Water	Marine water	1	26	2007	2007	Oden 2007
Water	Marine water	1	42	2007	2007	Polarstern 2007
Water	Marine water	1	39	2008	2008	Polarstern 2008
Water	Marine water	1	25	2010	2010	Snow Dragon

For the second reporting campaign (GMPII 2009–2014), the DWH was widely used to collect all relevant data from ROGs and produce regional monitoring reports. Klánová et al. (2013) shows, that there is about 10 active and 7 passive air-monitoring programs measuring POPs over longer time, whose results are included into the GMP. Human matrices are sampled by UNEP in collaboration with

World Health Organisation (WHO) and water samples originate from individual research programs realized between 2005 and 2014. Table 1 summarizes all POPs monitoring programs, that provided data to the GMPII.

Between January 2014 and August 2014, the data from all 5 ROGs were collected in the primary or aggregated form, processed and stored in GMP DWH. In total 47,895 annually aggregated values were collected, most of them (74.3%) for ambient air sampling. Since the DWH input interface allows to import any older data, including those from GMP I, the oldest measurements are from 1987, up to 2014. From 26 November 2014 Stockholm Convention encompasses 23 substances or their groups, all these compounds were considered as being of interest. This gives 24 groups of compounds and 109 individual parameters defined by CAS number. More detailed description of collected data is a part of the Content analysis section.

2.2 Data processing and presentation

The uppermost part of the GMP DWH system is the presentation layer consisting of analytical and visualisation tools, useful both for the creation of regional monitoring reports of ROGs and for general presentation of the GMP results to scientific public. There are five branches of description analysis, ranging from GIS visualizations of sampling sites and regions, metadata analysis of data availability in time and different compounds up to summary statistics (central tendencies, spread), time series analysis (direction and significance of exponential trends) and various forms of data exports. Most of the analytical tools are equal to the Genasis repository outcomes described in previous works (Jarkovský et al., 2013), (Hůlek et al., 2013), (Kalina et al., 2015).

3 GLOBAL MONITORING PLAN II CONTENT ANALYSIS

The total number of annually aggregated values within the second reporting campaign was 47,895, which is more than thrice the number of records of the first campaign (2003–2008), containing 14,769 records. No data for human blood were reported in 2014, breast milk was sampled by UNEP in collaboration with World Health Organisation (WHO). Water samples were collected from 17 individual research cruises and 3 fresh water monitoring programs. Most of data within the DWH are on air sampled by both the active and the passive ambient air samplers. Detailed distribution of samples is recorded in table 2.

Table 2. Descriptive statistics of GMP II data.

Matrix	Sub-matrix	N. of POPs	N. of countries	N. of Sites	N. of records	First Year	Last Year
Air	active	104	21	83	16,397	1990	2013
	passive	106	82	246	19,186	2003	2014
Water	fresh	4	10	13	31	2013	2014
	marine	6	-	524	1,122	2005	2011
Human milk (pooled)		98	79	106	11,159	2003	2014

Total number of countries reached 112, which is most of the 179 countries which have ratified SC before 2014. This allows first time to assess the POPs pollution on a really global scale, comprising robust data from all regions worldwide. Compared to GMP I reporting campaign, the length of time series measured on stable sites reached values (about 10 years or more), which are sufficient to identify temporal trends (decrease or increase) with high degree of significance. Such robust base of monitoring data allows to make reliable conclusions about the pollution and achieve the GMP objectives. In the next section, the main results of GMP DWH data assessment are presented.

3.1 Visualization layer

Since it is not possible to present here all the results of analyses implemented in the upper layer of GMP DWH, only the main outcomes are listed. From the total number of 106 compounds, only a

polychlorinated biphenyl 2,4,4'-trichlorobiphenyl (PCB 28) was chosen being one of the most often measured compounds from that treated by SC.

An extensive data filter (data selection module of the web data browser) is implemented as an initial point of any analysis. A definition of working set of the data is appropriate, despite it is possible to work with the full set of data contained within DWH. The outcomes of the visualization layer are arranged into 5 groups. First two groups characterize a data availability according space (GIS interactive module with circle marks for individual sampling sites/countries), time and groups of compounds. These analyses do not work with the concentration values, but only the metadata. There are bubblecharts and Gantt charts, showing the availability of data in selected categories and time periods.

Third and fourth groups of outcomes are focused on the concentration of POPs. Summary statistics allows to display different kinds of aggregated values for each combination of environmental matrix, year, site and compound in a form of boxplot. This enables to compare different locations and time periods and estimate levels of pollution in different regions. Figure 2 depicts such a comparison of pollution levels between the five UNEP regions.

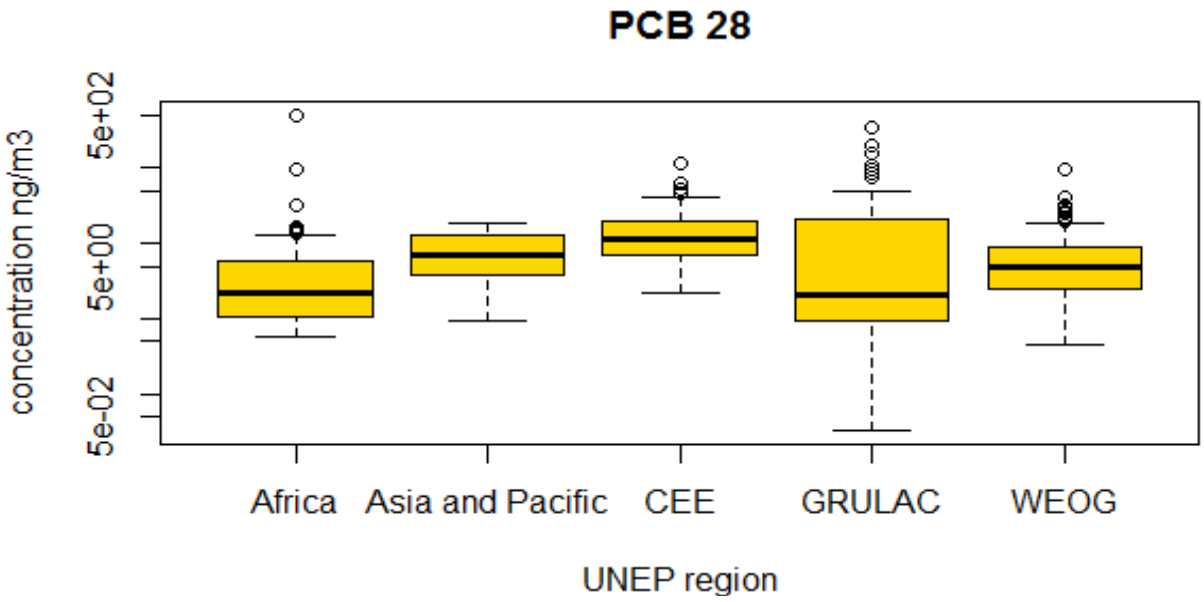

Figure 2. Levels of PCB 28 ambient air concentration measured by passive sampler between 2009 and 2014 according to UNEP regions (CEE = Central and Eastern Europe, GRULAC = Group of Latin America and Caribbean Countries, WEOG = Western Europe and Others Group). The thick black line denotes median, the yellow box denotes 25th to 75th percentile range, the whiskers denote minima and maxima except of outlying values denoted as black rings.

Temporal trends are presented in two forms. In case of an overview, a scalable map is shown, with circles denoting all sites of the measurement. Size of the circle corresponds with median concentration at the site, its colour indicates a statistical significance and direction of the trend. There is also a form of the map with barplots instead of circles, which shows directly the concentrations on each site. The circles/barplots are clickable, linked to another time series analysis, consisting of a time plot for individual matrix, site and compound. In this plot all annual values are depicted as points, equipped by confidence intervals and interlaced by exponential curve, suggesting the long-term development of pollution at the site. One exemplar of the trend map (PCB 28) is shown in figure 3.

Taking into account, that all the reported data are publicly available (Monitoring reports, 2014), the last analytical tool was constructed to serve as a designer of data exports. This allows to choose a desired structure of an output table which encompasses all data included in the data selection. The table is displayed directly on the website and also it is available to download in a form of a Microsoft Excel .xls file.

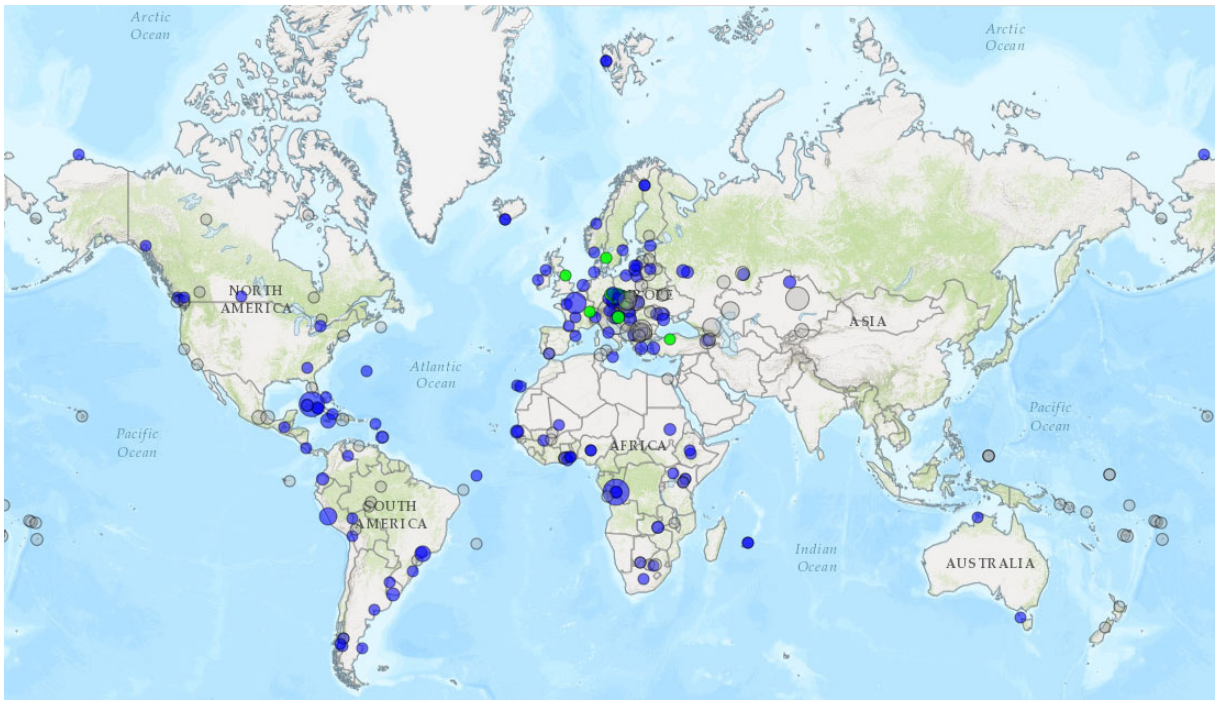

Figure 3. Concentration levels and trends of PCB 28 in ambient air (passive sampling). The size of the circle depends on a level of pollution, colour of the circle represents trend: green denote statistically significantly decrease, red circles increase, blue circles insignificant trend and grey circles are individual measurements (trend is not available).

4 CONCLUSION

The development of the publicly-available and unified system for the collection, treatment, storage and publication of data on persistent organic pollution was one of the biggest tasks and urgent needs to enhance visibility of the GMP and to facilitate improved interpretation, spatial visualization and modelling of available monitoring data, as identified in (Klánová et al., 2013). During the second half of 2013 the system was developed and immediately after its finishing, it was 8 months in 2014 in charge, collecting data from ROGs in frame of second reporting campaign (GMP II 2009–2014). It was used for data treatment and aggregation likewise for generation of outcomes composing main parts of regional monitoring reports (Monitoring reports, 2014). Furthermore the system is prepared to serve as a collection and assessment tool also for the upcoming third reporting campaign (GMP III 2015–2020) with its expected progress in four next years.

The amount and character of collected POPs data is unique in its broadness, encompassing measurements from 1987 up to 2014 from several matrices and 112 countries over the globe. Database of such complexity allows to compute very robust long-term trends, regional pollution levels, prepare comparisons and models of POPs transport, as well as other analyses in future. Although five groups of statistical tools were implemented directly on the website, there is a big potential of the dataset to serve as a base of forthcoming research works.

All the reported values in annually aggregated form, analyses of temporal trends, pollution levels, charts and maps are publicly available on a website of GMP DWH project (GMP, 2014), which allows to download both the raw data in the form of .xls files or the analytical outputs as bitmap pictures in a .png format.

REFERENCES

GMP, (2014). Global Monitoring Plan on Persistent Organic Pollutants. http://pops-gmp.org/ (last accessed 14.04.16.).

Gregor, J., Hůlek, R., Jarkovský, J., Borůvková, J., Šebková, K., Schwarz, D., Klánová, J., Dušek, L., 2013. Tools for collection, analysis and visualization of data from the Stockholm Convention Global Monitoring Plan on Persistent Organic Pollutants. In: Hřebíček, J., Schimak, G., Kubásek, M., Rizzoli AE. (Eds.), Environmental Software Systems. Fostering Information Sharing, Heidelberg, vol. 1, pp 222–229.

Harner, T., 2014. 2014_GAPS Template for calculating PUF and SIP disk sample air volumes_November 6 2014. https://www.researchgate.net/publication/267867691_2014_GAPS_Template_for_calculating_PUF_and_SIP_disk_sample_air_volumes_November_6_2014 (last accessed 14.04.16.).

Hůlek, R., Kalina, J., Dušek L., Jarkovský, J., 2013. Integration of R Statistical Environment into ICT Infrastructure of GMP and GENASIS. In: Hřebíček, J., Schimak, G., Kubásek, M., Rizzoli AE. (Eds.), Environmental Software Systems. Fostering Information Sharing, Heidelberg, vol. 1, pp 240–252.

Jarkovský, J., Dušek, L., Kubásek, M., Hůlek, R., Gregor, J., Hřebíček, J., Klánová, J., Šebková, K., Borůvková, J., Holoubek, I. GENASIS: On-line data browser for environmental monitoring and associated information systems, 2013. In: Hřebíček, J., Schimak, G., Kubásek, M., Rizzoli AE. (Eds.), Environmental Software Systems. Fostering Information Sharing, Heidelberg, vol. 1, pp 253–261.

Kalina, J., Hůlek, R., Borůvková, J., Jarkovský, J., Klánová, J., Dušek, L., 2015. Three levels of R language involvement in global monitoring plan warehouse architecture. In Hřebíček J.,Denzer R.,Schimak G.,Argent R.M. IFIP Advances in Information and Communication Technology. New York, vol. 1, pp 426–433.

Klánová, J., Harner, T., 2013. The challenge of producing reliable results under highly variable conditions and the role of passive air samplers in the Global Monitoring Plan. Trends in Analytical Chemistry, 46, 139–149.

Monitoring reports, 2014. http://chm.pops.int/Implementation/GlobalMonitoringPlan/MonitoringReports/tabid/525/Default.aspx (last accessed 15.04.16.).

UN, 2014. United Nations Regional Groups of Member States. http://www.un.org/depts/DGACM/RegionalGroups.shtml (last accessed 14.04.16.)

UNEP, 2007. Guidance on the Global Monitoring Plan for Persistent Organic Pollutants. http://www.pops.int/documents/meetings/cop_3/meetingdocs/inf14/GMP%20Guidance%20CD/Guidance.pdf (last accessed 14.04.16.)

UNEP, 2013. Results of the global survey on concentrations in human milk of persistent organic pollutants by the United Nations Environment Program and the World Health Organization. http://chm.pops.int/Portals/0/download.aspx?d=UNEP-POPS-COP.6-INF-33.English.pdf (last accessed 14.04.16.)