Tools for Collection, Analysis and Visualization of Data from the Stockholm Convention Global Monitoring Plan on Persistent Organic Pollutants

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

Tools for collection, analysis and visualization of data from the Stockholm Convention Global Monitoring Plan on Persistent Organic Pollutants

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

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

Abstract. The Global Monitoring Plan for persistent organic pollutants is an important component of the effectiveness evaluation of the Stockholm Convention and its main objective is assessment of long-term changes in POPs concentrations in core matrices – ambient air and human tissues (milk, blood). This paper summarizes results of activities of the Research Centre for Toxic Compounds in the Environment and the Institute of Biostatistics and Analyses, Masaryk University, Czech Republic, which have been performed on the basis of the mandate given by Global Coordination Group for GMP and Secretariat of the Stockholm Convention: content analysis of the GMP monitoring reports published in 2009, on-line visualization tool for browsing and analyzing collected data from the monitoring reports, and proposal of a design of future data collection campaigns.

Keywords: Stockholm Convention, Global Monitoring Plan, POPs, database, data collection, visualization.

1 Background

Multilateral environmental agreements on chemicals management have brought requests for developing instruments that are able to collect or evaluate data and establish and/or optimize risk management related to increasing levels of chemicals in the environment. Such instruments can be employed on both national and international level.

Development and adjustment of systems for the collection, analysis, and visualization of environmental data must cope with relatively high heterogeneity of collected data, e.g. in terms of data sources (institutions, projects, purposes), different matrices (ambient air, water, soil, sediments, human tissues), or chemical parameters (isomers, degradation products). It is therefore important to strictly define data structure and

code lists to ensure reliability of all collected and analysed data and provide detailed guidance and support to all users participating in the data collection process. Furthermore, sufficiently complex conceptual models, advancing development of formal ontologies for environmental and epidemiological data acquisition systems are needed [1,2].

The Global Monitoring Plan (GMP) for persistent organic pollutants (POPs) is an important component of the effectiveness evaluation of the Stockholm Convention (SC) and its main objective is the assessment of long-term changes in POPs concentrations in core matrices – ambient air and human tissues (milk, blood). A methodological guidance document specifying the methods to be used and compounds to be monitored was published in 2007[3]. The first global data collection occurred in 2008 and the results were published in five regional monitoring reports in 2009.

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 MU), Masaryk University, Czech Republic, performed a content analysis of the GMP reports[4], prepared electronic tools for storage and on-line visualization of data published therein[5], and proposed a comprehensive IT solution for future data collection campaigns.

2 Content analysis of GMP monitoring reports

The first step in the analysis of the regional GMP reports was to determine the relation of all reported parameters to the Stockholm Convention. The classification strictly followed scope of the Stockholm Convention and the GMP Guidance document and their amendments in time. Considerations were also given to the reported data content and relevance of existing records for further data collection (Fig. 1).

1. 12 initial POPs included in the Stockholm Convention in 2001, their congeners, isomers and degradation products specified in the GMP Guidance document (2007) – 58 parameters;
2. Additional 10 POPs listed in the Stockholm Convention in 2009 and 2011 and specified in the updated GMP Guidance document (version 2009) – 7 parameters;
3. All other compounds, their sums and toxic equivalents related to the Stockholm Convention but not specified in any of the GMP Guidance documents – 84 parameters;
4. Compounds found in the GMP reports but not related to the Stockholm Convention (i.e. PAHs) – 22 parameters.

The analysis revealed several serious challenges related to data standardization in the GMP report. The reports suffered from the lack of standardized taxonomy for POPs, their isomers, transformation products and summations. Heterogeneity of the data was further enhanced by reporting various toxic equivalents (TEQ) (based on WHO TEF values from various years) rather than concentrations of the individual polychlorinated dibenzo-p-dioxins (PCDDs), Polychlorinated dibenzofurans (PCDFs) and poly-

chlorinated biphenyls (PCBs). Unclear identification of units, time and spatial scales of the reported concentrations as well as insufficient specification of aggregated data belonged to other frequently identified drawbacks. The review, however, resulted in conclusion that the available GMP data can be used for baseline statistical processing. Data can be analysed as annually aggregated time series or at least as relevant point estimates, prepared for comparison with the next data collection campaign.

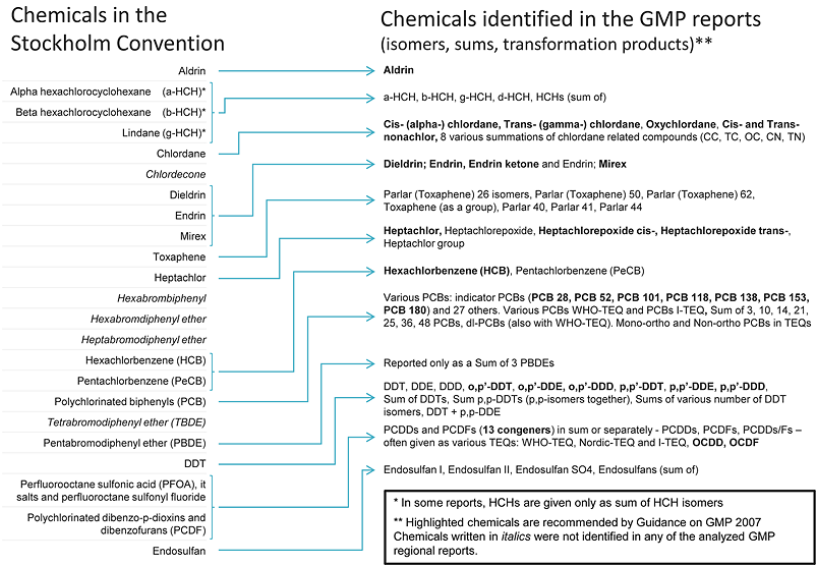

Fig.1. Overview of chemicals identified in GMP reports

3 GMP database

Data from the GMP regional reports were transformed into a form that was suitable for both database development and analytical processing. Data had to be extracted from textual form, tables, citations and charts into a common standardized data structure with shared property lists (matrices, chemicals, units…) in several consecutive steps:

1. **Identification of collected data.** First of all, an overall inspection of all regional GMP reports was performed. The main goal was to identify reported matrices, measured chemicals (compounds), used values and aggregation characteristics as well as sampling frequency and time ranges.
2. **Datasets volume identification.** Determination of volume (range and size) of datasets was equally important, because an appropriate storage technology had to be chosen. A relational database engine was used as the main storage solution in order to enable fast sorting, aggregation and selection.

3. **Design of a common data structure.** For the purposes of the data extraction from particular regional reports, a common data structure was used allowing for maintenances of all collected values – both directly measured and aggregated.
4. **Extraction of data from reports into a common data structure.** During the process of data extraction from regional reports, a number of problems had to be solved. All missing information had to be looked up in the supporting text of the reports or was completed based on the context knowledge. Most of the problems were caused by the non-existent standardized form for data collection.
5. **Data validation.** The database entries were double-checked after the digitalization. Manual validation was applied particularly to problems with text labels and variables, which could not be automatically solved by analytic approaches. Second step of validation included analytical tools to identify missing values, extreme values, and duplicities.
6. **Data pre-processing.** Some records in GMP reports contained one value for multiple years, or countries (sub-regions). Such values were individually assigned to all covered countries and/or years.
7. **Data aggregation.** For the purposes of analytical outcomes from regional reports, annual aggregation of values was used. Aggregations were calculated and prepared using specialized statistic software (SPSS). Import of re-calculated data into the database was done by automated procedures to ensure data consistency and integrity. Different aggregation approaches were used for:

   - Primary and aggregated data obtained from the GMP reports
   - Different matrices (air and human tissues)

8. **Data import.** MS Excel sheets were used as data transfer medium and imported into MSSQL Server Database by means of standard embedded functions and features. Microsoft tools and data files were used in all steps to ensure maximal compatibility during the whole process of data transfer and processing.

Aggregated values originating in GMP reports are stored in a relational database consisting of 10 basic entities: UN Region; Country; Site; Matrix; Compound; Parameter; Record; Aggregation; Percentile; Unit.

4 On-line data visualization tool

In addition to the content analysis of GMP monitoring reports and transferring the published data into the database, an on-line data visualization tool was developed to allow easy searching, browsing, and analysing the GMP data in their annually aggregated form (www.pops-gmp.org). The browser provides easily accessible information on the performance of monitoring programmes in individual countries and/or regions sorted by matrices, time, and compounds. The database underlying the visualization tool was supplemented with additional data from other public sources, in particular from large environmental monitoring programmes, as some conclusions of the re-

gional monitoring reports refer to them (EMEP[1], AMAP[2], IADN[3], MONET[4]). The visualization comprises of three descriptive and three analytical tools:

1. **Monitoring overview.** An interactive map displaying monitoring activities in the individual years.
2. **Available data – parameters.** Regional availability of monitoring data on compounds of interest in key matrices.
3. **Available data – years.** Regional availability of monitoring data in selected time periods.
4. **Reported values.** Reported concentrations in the key matrices. For air, only the concentration values reported from the background (urban, suburban, rural, remote, mountain and polar) sites were included.
5. **Regional backgrounds – data validation.** A 6-step validation procedure of data from ambient air monitoring.
6. **Regional backgrounds – inter-regional variability.** Statistical evaluation of reported background atmospheric concentrations.

Fig.2. On-line data visualization of GMP data – tool 4: reported values. Available at www.pops-gmp.org.

The visualization tools were developed using Adobe Flash and ArcGIS technologies. The user-defined outputs are available for download as a graphic file (PNG format); the underlying data from tools 4–6 are also available in text format which is easy to import to MS Excel.

The website has been designed particularly for purposes of the Stockholm Convention effectiveness evaluation by UNEP, as well as by environmental professionals and lay public.

---

1 European Monitoring and Evaluation Programme
2 Arctic Monitoring and Assessment Programme
3 Integrated Atmospheric Deposition Network
4 Passive Ambient Air Monitoring Network

5 Design of upcoming data collection campaigns

Considering the drawbacks identified in the 2008 GMP reports, an updated design of data collection and processing for future data collection campaigns was proposed by RECETOX and IBA MU and approved by experts from the GMP Global Coordination Group and Regional Organization Groups. The proposed design includes, among others, an electronic data capture system, a defined parametric data structure for both primary and aggregated data, and predefined code lists for most of the items to ensure maximum standardization and uniformity of reported data to allow reliable analysis and reporting.

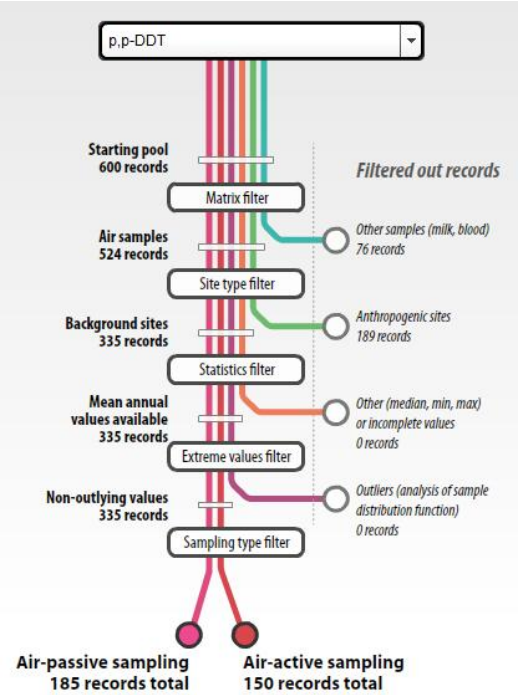

Fig.3. On-line data visualization of GMP data – tool 5: a 6-step validation procedure of data from ambient air monitoring. Available at www.pops-gmp.org.

Structure of the data capture system and the code lists are adjusted to data from monitored environmental matrices (ambient air, human tissues) and to both primary and aggregated values. Options to check and correct the GMP1 data, should need be, are also included.

There are two principal ways for data recording into the database, each with specific data fields:

1. Data sets of primary POPs concentration data in a given site and at given time points; these records can be also uploaded automatically through standardized import MS Excel sheets.

2. Data sets of aggregated data on mean (median) annual POPs concentrations with proper variability measures (minimum, maximum, percentiles, standard deviation).

The GMP data warehouse has been built on an on-line web-based information system with a central data repository to collect relevant data (TrialDB system). TrialDB system was developed in cooperation with the Yale University [6-8].

Complex generic environment for on-line data management is recommended for the development of on-line databases with multiple touch-points collecting primary data. Multi-tier architecture (client - application server - database server) is the most commonly used approach. A common web browser (i.e. Internet Explorer) is employed as a client in that architecture. Internet connection is then necessary for all participating users. Web browser is available as default equipment in all personal computers; thus there is no need to install any specialized software on the users' workstations. Users can access all functions of the system (data manipulation, entry, editing, viewing, reporting, analyses, etc.) through their web browser.

Communication between the client and the server is always realized via secured (encrypted) https protocol (128-bit encryption is used), and data security is guaranteed in a standardized manner: system administrator assigns access login and password to individual users. Access level can be specified separately for each account. The legal safeguarding is defined in contracts between the users and the provider.

The main advantages of the centralized on-line solution are:

- Regular monitoring by trained staff;
- Security and availability for sharing among clients (when permitted);
- No need of upgrades, because changes in the web application are made centrally;
- Monitoring the progress of the project and addressing potential problems as soon as possible.

The proposed system is equipped with a number of useful functionalities for data reporting, handling and evaluation. In addition, a track changes tool displays recent changes in a form or group of forms, a data validation tool provides a list of incomplete records and missing items and some other useful tools are available. A standardized workflow for data reporting from involved monitoring programmes and subsequent approval by Regional Organization Groups has been set up to allow effective completion of regional monitoring reports in late 2014.

6 Conclusions

The GMP data available in monitoring reports from 2008 can be used for baseline statistical processing. It is, however, necessary to design a standardized and parametric data collection system to allow reliable assessment of time trends in concentrations of POPs listed in the Stockholm Convention. As recommended by the Global Coordination Group and updated Guidance on GMP [9], the data should be obtained mainly from large environmental monitoring programmes and surveys (EMEP, AMAP, GAPS, IADN, MONET, WHO human milk surveys).

Acknowledgements. This work has been supported by the project TB010MZP058 "Development of the system for spatial evaluation of the environmental contamination".

References

1. Dušek, L., Hřebíček, J., Kubásek, M., Jarkovský, J., Kalina, J., Baroš, R., Bednářová, Z., Kánová, J., Holoubek, I.: Conceptual Model Enhancing Accessibility of Data from Cancer–Related Environmental Risk Assessment Studies. In: Hřebíček, J., Schimak, G., Denzer, R. (eds.) Environmental Software Systems. Frameworks of eEnvironment. IFIP Advances in Information and Communication Technology, vol. 359, pp. 461-479. Springer, Heidelberg (2011)
2. Jarkovský, J., Dušek, L., Janoušová, E.: Is On-Line Data Analysis Safety? Pitfalls Steaming from Automated Processing of Heterogeneous Environmental Data and Possible Solutions. In: Hřebíček, J., Schimak, G., Denzer, R. (eds.) Environmental Software Systems. Frameworks of eEnvironment. IFIP Advances in Information and Communication Technology, vol. 359, pp. 486-490. Springer, Heidelberg (2011)
3. 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)
4. Klánová J., Dušek L., Borůvková J., Hůlek R., Šebková K., Gregor J., Jarkovský J., Kohút L., Hřebíček J., Holoubek I.: The initial analysis of the Global Monitoring Plan (GMP) reports and a detailed proposal to develop an interactive on-line data storage, handling, and presentation module for the GMP in the framework of the GENASIS database and risk assessment tool. Masaryk University, Brno, Czech Republic (2012)
5. 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 [online]. Masaryk University,http://www.pops-gmp.org/visualization (2013)
6. Nadkarni, P.M., Brandt, C., Frawley, S., Sayward, F.G., Einbinder, R., Zelterman, D., Schacter, L., Miller,P.L.: Managing attribute--value clinical trials data using the ACT/DB client-server database system. J. Am. Med. Inform. Assoc. 5, 139-151 (1998)
7. Nadkarni, P.M., Brandt, C.M., Marenco, L.:WebEAV: automatic metadata-driven generation of web interfaces to entity-attribute-value databases. J. Am. Med. Inform. Assoc. 7, 343-356 (2000)
8. Nadkarni, P.M., Marenco, L.: Easing the transition between attribute-value databases and conventional databases for scientific data. Proc. AMIA Symp., pp. 483-487 (2001)
9. United Nations Environment Programme: Guidance on the Global Monitoring Plan for Persistent Organic Pollutants. Secretariat of the Stockholm Convention on Persistent Organic Pollutants, Geneva (2013)