6. Data Handling

6.1 Objectives and priorities

The results from the GMP will be used to determine trends from monitoring of POPs globally to support the effectiveness evaluation of the Stockholm Convention. A primary goal is therefore to obtain (comparable) data that are capable of revealing trends over time in emissions and/or exposure to contaminants of concern, in the various regions.

Effective sharing and delivery of necessary data and information by contracting parties is essential to achieving this objective. The data provided need to:

- be relevant, to the objectives of the evaluation of the Stockholm Convention;
- have sufficient quality and level of detail;
- be consistent and comparable over time;
- be transparent, and to the greatest possible degree public and unrestricted.

6.2 Data policy

6. 2.1 Terminology [need to be checked for consistency with other parts of the document]

To avoid confusion, it is important that some basic terms and concepts that are used in this document are defined so that they are understood to mean the same thing by all parties.

- ***Primary GMP data***: are the results of measurements made on samples collected under the auspices of the GMP, or other programmes that are compatible with the goals of the GMP. They include both measurements of POPs in specific samples, and measurements of other covariables relating to these samples (e.g. biological covariates), that are necessary to interpret the POPs data in a meaningful way, including the location and timing of sampling.

- ***GMP meta-data*****:** are any other data or information that describe the *primary GMP data* in some way. This can include information on the methodologies employed (e.g., for sampling and analysis) and the laboratories responsible for a particular set of analyses, or the design and implementation of programmes that contribute to the GMP, etc.

- ***Supplementary data:*** Are any other data or information that may be accepted for use in the Stockholm Convention evaluation process. This might include relevant information and/or data from published sources (e.g. the peer reviewed scientific literature, existing assessment, etc), results of modelling activities that may assist the data interpretation and evaluation, or results of research activities that may be relevant to interpreting the *primary GMP data* in a valid and meaningful way (e.g. process studies, food-web studies, etc.). Such data will comprise an important contribution to the Stockholm Convention evaluation process, especially in the initial period where the necessary data management infrastructure is still under development in some regions.

Primary GMP data (and *supplementary data* where these concern monitoring results from e.g. published sources) can be further sub-divided between:

- ***Un-aggregated data:*** individual sample measurement values (e.g. the concentration of CB153 in the liver tissue of a specific individual fish, sampled at location x at time y).

- ***Aggregated data:*** (statistically) summarised data, e.g. averaged values that summarise the measurements on a number of individual samples.

6. 2.2 Data policy

The GMP data handling activities should promote transparency of process, both with respect to the data themselves, and how they are treated and analysed. The GMP data policy should also have the goal of ensuring access (for the purposes of the Stockholm Convention evaluations) to the most relevant and up-to-date information available.

In considering potential public access to data, a distinction is usually made between *un-aggregated data*, *aggregated data,* and high level *meta-data*. Sensitivity with regard to making data publicly available generally decreases in the order *un-aggregated data* > *aggregated data* > high level *meta-data*; with high-level *meta-data* normally not subject to any restrictions.

Part of the data generated under the GMP will already be in the public domain, being made available for public access soon after their generation. Other data, however, may be restricted; for example, subject to a moratorium to allow scientists responsible for the data to publish their results before the data are made public.

Use of data for the purposes of the Stockholm Convention evaluations should not compromise the rights of the data owners. Data owners should therefore be fully informed of how their data will be used, and what parts of the data or results will be made public and when in order to ensure that they are in agreement. Furthermore, full and appropriate acknowledgement of data sources should be a key part of the data policy.

To facilitate the above, for all data delivered from the GMP:

- the data owners should be identified (note: this not always the same as the data provider);
- any conditions relating to restrictions to making the data publicly accessible should be properly described (by the data owners);
- the required citation/acknowledgement to the data should be provided (by the data owners).

6.3 Data to be reported

Minimum data reporting requirements are required to ensure consistency both within datasets over time and among the datasets between regions.

Ideally, *unaggregated data* (individual sample measurement values) should be reported. Where data are reported as statistically *aggregated data* (averages):

- the type of statistical average concerned (e.g. average, geometric mean, median) should be clearly indicated, and
- the data should also include an estimate of variability (standard deviation, standard error, confidence interval, etc.).

Air (monitored at sites unaffected by local contamination) and human tissues (breast milk or blood) have been identified as the priority monitoring matrices under the GMP. However, the data handling routines should also accommodate results from monitoring of other types of environmental sample identified under the GMP (bivalves, tissues and organs of other biota, etc.). Where data on priority or secondary GMP matrices are not available, some flexibility will be retained to allow use of other relevant data, for example POPs levels in food, etc.

6.3.1 Contaminants data

Contaminants of concern are those that are identified under the Stockholm Convention GMP (see Chapter 2). To the greatest extent possible, data should be reported for individual compounds or congeners or isomers.

Data on contaminant concentrations should be reported together with a clear indication of both the units and the basis of determination (wet weight, lipid weight, etc.). Recommended units and basis of determination for GMP priority matrices are as follows:

	Air	**Human milk and blood**	**Tissues and organs of other biota**
All POPs except PCDD/PCDF	pg/m^3	ng/g lipid	ng/g lipid
PCDD/PCDF	fg/m^3	pg/g lipid	pg/g lipid

pg/g = pico-grams per gram = 10^{-12} = nano-g/kg

fg/g = femta-grams per gram = 10^{-15} = pico-g/kg

6.3.2 Co-factors and methodological information

In addition to reporting of data on contaminant concentrations in the various media, the goals of the GMP require that sufficient supplementary data and information are also reported to allow valid interpretation of, for example, time-series datasets. This includes, for any individual dataset, reporting:

- the sampling location(s) concerned (including site description)
- the time of sampling (or the time period represented by the dataset)
- data on other factors that may be relevant to interpretation of temporal trends (for example, age/size of animals sampled, volumes of air sampled, information on smoking or dietary habits of the sampled populations, methods employed, etc.)
- data on parameters to allow conversion between reporting basis (e.g. %lipid and methods used for lipid determination)
- information on methodologies employed for sampling and analysis, QA/QC routines
- information on results of laboratory performance in (international) intercalibration exercises and laboratory performance testing schemes

Further details of the reporting requirements will need to be determined when the monitoring programme has been specified in greater detail.

6.3.3 Limit of detection, limit of quantification

Definitions of the limit of detection (LOD) and the limit of quantitation (LOQ) are defined in Chapter 5 of this document.

Non-detects should normally be reported as 'less than the LOD', the value of which has to be reported; i.e. if the limit of detection is 0.5 ng/g lipid, a non-detect should be reported as <0.5 ng/g lipid. [If another method is used it has to be clearly specified, see Section 5.2.2].

6.3.4 Derived parameters

Derived quantities, such as normalized or adjusted values or parameters such as TEQs or sums of congeners should normally be produced by those responsible for evaluating the data, on the basis of the reported data for individual congeners, etc.

If it is agreed that derived values should/may be reported, then a detailed definition of the methodology to be applied should be provided, including description of how to incorporate values below the detection limit, TEF to be applied, etc.

For TEQ calculation in the case of PCDD/PCDF analysis, it is strongly advised that upper bound and lower bound values be reported in keeping with the recommendations by JECFA (Joint FAO/WHO Expert Committee on Food Additives).

6.4 Data quality

Prior to being accepted for use in the Stockholm Convention process, it is recommended that data should be accepted, through an independent evaluation, as having 'appropriate quality'.

Data quality requirements shall be the same for all regions; where necessary the objective will be to build capacity not to reduce requirements to the lowest common denominator.

Data quality evaluation involves several components at different stages:

- Data should be evaluated at source as being of appropriate quality before they are reported. This includes application of appropriate methodologies and QA/QC routines during sampling and within the laboratory. Data should be scrutinized by the laboratory generating them and thereafter by a national coordinator, who among other things should check that the data have been correctly transcribed and compiled and are complete with respect to the reporting requirements
- Upon reporting, where the possibility exists, data should be subject to data quality checking at, for example, data centres – where routines should be available for checking completeness of data submissions and may be available for conducting basic checks including inter-component comparisons (e.g. relative concentrations of different parameters/congeners) and cross-comparisons of data from different sources. Data centres should provide data quality feedback to data sources.
- Finally, the data, confidence intervals and all supporting information on QA, sampling and analytical methods, etc. should be evaluated by a regional quality review panel responsible for accepting the data for use in the Stockholm Convention Evaluations.
- A system may need to be developed for flagging data that, e.g., lack appropriate QA/QC information, do not fulfil all quality criteria, or are between the LOD and the LOQ, but which may still be acceptable for some purposes in the Stockholm Convention evaluation process.

In addition to QA/QC considerations relating to the accuracy of the results themselves, QA/QC routines need to be implemented to ensure that quality is maintained during the data exchange process. Data compilation and data reporting include a number of steps where (considerable) potential exists for introducing errors: data punching, application of algorithms used in data conversion of transformation, data communication, etc. This is especially so when data are transferred beyond the 'horizon' of those who are most familiar with the data and therefore best placed to spot apparent discrepancies, i.e. those responsible for collecting/generating the data. It is therefore recommended that:

- an appropriate chain of custody be established from the data originator to the data assessment group. This chain should be as short as possible.
- at each point of transfer in the chain, those responsible for delivering and receiving the data should sign-off to confirm that the data have been correctly and accurately transferred. In practise, this involves (a) data recipients confirming that data delivered to them meet the necessary requirements and specifications for delivery, (b) data recipients preparing summary data products (maps, summary statistics, etc.) that will allow data errors or discrepancies introduced during the transfer to be detected, which are returned to the data deliverer (c) the data deliverer examining these products and confirming that the data appear to be correctly

transferred. Ultimately, any GMP data evaluations/products should be returned to the data sources for their comment/confirmation.

6.5 Data flow and 'storage facilities'

6.5.1 Scope

The main goal of the GMP data strategy is to compile *un-aggregated - primary GMP data.* Un-aggregated data permit data to be treated in a transparent and consistent manner according to agreed assessment methodologies. If these methodologies are modified or further developed at some point in the future, the availability of *un-aggregated - primary GMP data* provides the best possibilities for re-calculation or for repeating previous data treatment. *Aggregated data* provide much more limited potential for re-analysis or for combining data from different sources. Most data derived from *supplementary information* will be aggregated (unless they are otherwise accessible as *un-aggregated data* from data centres/archives).

That part of the *GMP meta- data* that detail methodologies employed in the collection and generation of the *primary GMP data,* as well as laboratory intercalibration/testing scheme results should follow the *primary GMP data* and also be reported to *data centres***,** as well as being made available in an appropriate form to data assessment groups. Since intercalibration/ performance testing results available from the organizers of these exercises are often referred to an (undisclosed) laboratory code system, these results will need to be reported by the laboratories themselves, along with the measurement data.

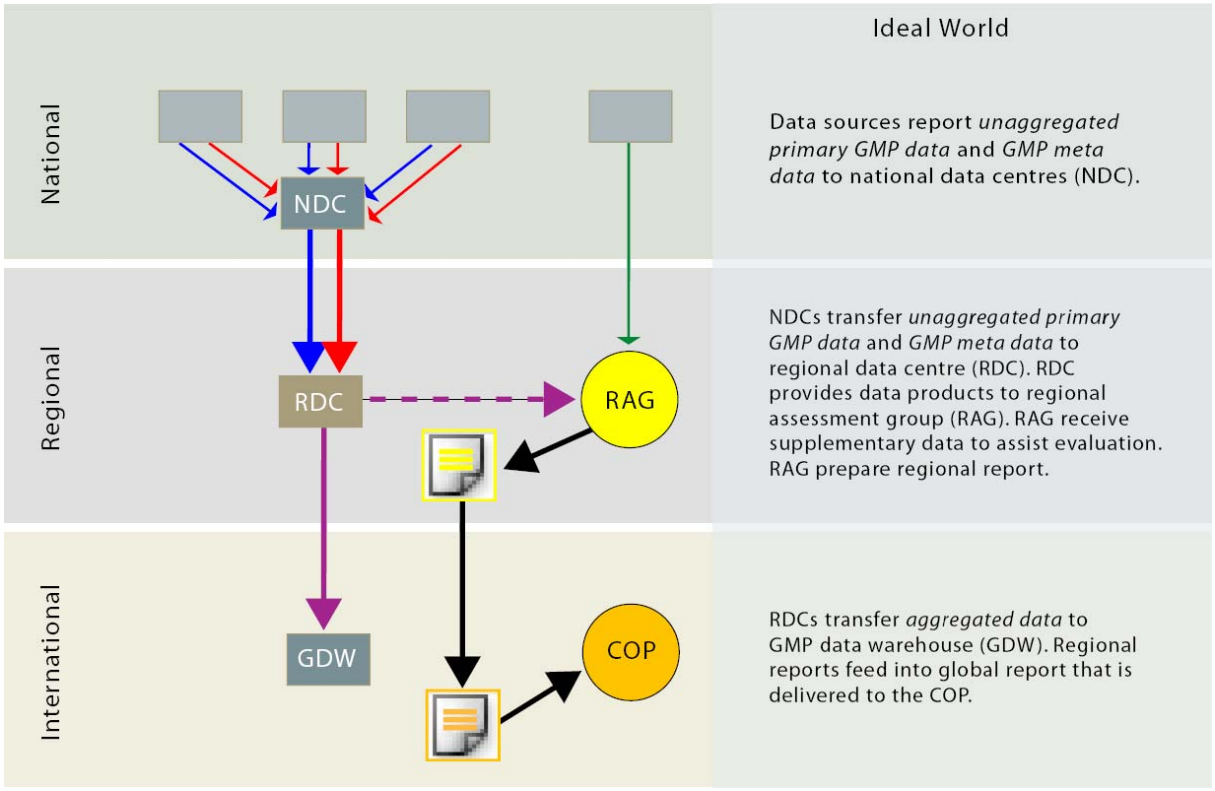
The data flow for the GMP outlined here focuses on reporting and compilation of data at the international level. Organization of data compilation and reporting at the national level is assumed to be the responsibility of contracting parties. However, contracting parties requiring assistance to build capacity in this respect may look to the GMP for such assistance, including exchange of experience between parties and countries.

6.5.2 GMP data storage (compilation and archiving)

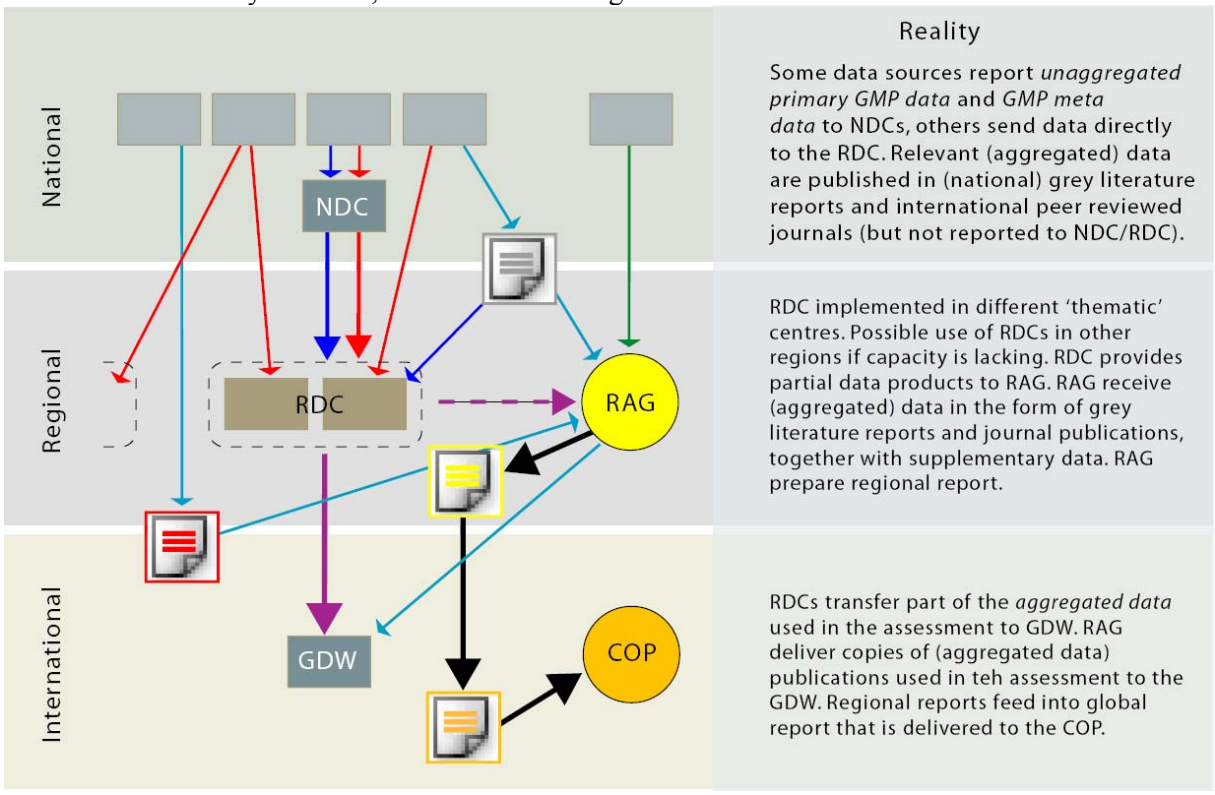
The data reporting model that is being suggested involves compiling and archiving *primary GMP data* within a 'regional data repository' in each of the 6 geographic regions.

In addition to the regional data centres, a single GMP 'data warehouse' will be established to compile and archive aggregated data, data products and results, including *supplementary data* that are used in the Stockholm Convention evaluations. A primary purpose of the GMP data warehouse will to provide transparency to the process, facilitating access to the data and results that are the basis for any conclusions of the (sufficiency and effectiveness of the) evaluations. The GMP 'data warehouse' could also function as the data centre for maintaining the database of *meta-data*, including meta-data on GMP implementation in the various regions, and information and documentation that may be required by assessment groups concerned with, for example, data quality evaluations, such as information on laboratory performance.

This 'ideal' solution for a particular region is shown in the following flow-chart (1).



In reality, however, this ideal solution is unlikely to be achieved. The following flow-chart (2) below shows the more likely situation, at least for some regions for some time to come.



Due to the desirability of ensuring that data are handled in centres with appropriate expertise to 'understand' the data concerned, it may well be appropriate, also in regions with well-developed existing data centres, that rather than a single physical location, the regional 'data repository' is implemented through a limited number of specialist thematic centres; as few as necessary to cover the

type of data involved, with preference being given to centres that are capable of serving as regional centres for multi-disciplinary datasets (blood/milk, etc.).

If appropriate 'data centres' cannot be identified in one or more regions, a temporary solution should be identified to facilitate data handling while the necessary capability is being established within the region; one possible option being to use facilities that may exist in neighbouring regions.

Capacity building for GMP data management activities will be essential in several regions. One way to efficiently implement this would be to establish 'model' solutions in some regions and then consider possibilities for 'technology transfer' (e.g. these model centres make their existing database developments available to other centres – under some suitable licensing agreements to avoid infringing intellectual property rights) and staff training to implement data centres in other regions. Effort will also need to be expended to support data management capability at the data sources, both to educate data sources in the needs and requirements of the GMP and to 'realise' the data delivery; this also is not just a problem for developing areas but also a major obstacle to data flow in areas with existing programmes and data flow. It is critical that data reporting is an integral part of GMP (monitoring) implementation at every level – from simple pilot projects to national activities in the most advanced countries – data management should not be an 'add-on' exercise. It should be recognized that data-management may consume up to 5-10% of a monitoring programme finances; however, without this investment the other 90% of the expenditure is largely wasted.

6. 5.3 Selection of GMP data 'centres'

Selection of GMP data 'centres' should take account of the following:

- Data should be compiled in centres that are founded on a basis that will secures their continuing existence and stability over a long-period of time (decades at least); centres lacking a secure long-term funding perspective should be discounted.
- Data should be compiled at centres where the in-house staff possesses the appropriate expertise, both in terms of data management and understanding of the types of data being handled.
- Data should be compiled at centres possessing the necessary technical resources and equipment for the required data handling, including communications and transfer of data, secure data storage (including on-site and off-site back-up), preparation of data products, etc.
-

The GMP is envisaged as a long-term activity. In some cases several years of data will be required before reliable interpretation of trends can be achieved. Disruption to the data management process through frequent changes in the (location of or operations at) data storage facilities should be avoided.

A number of 'data centres' or 'programmes' exist today that could be considered either as candidate GMP 'data storage facilities' within a region, or as centres that could partner or facilitate capacity-building of 'storage facilities' in other regions. Some of these are presented in the table in Annex 1.

6.5.4 Standardized data exchange and reporting systems

Reporting of data in a manner that is technically feasible and reasonably convenient for all parties concerned, minimizes potential for errors and ensures that all reporting requirement are met is a major challenge.

GMP data exchange will probably involve use of a wide variety of formats. Data reporting systems should therefore aim to be as flexible as possible, while at the same time trying to promote the maximum possible degree of standardization. Some constraints will need to be imposed to ensure that data reported meet the minimum requirement with regard to content and level of detail.

Compilation of data according to agreed standards is also important if they are to be used in connection with modelling activities, for example for the understanding of environmental transports within and between regions. If properly implemented, the GMP 'data warehouse' will constitutes a potential source of data that can be used for model validation, etc. However, this subject is not addressed further in this guidance document.

Definition of a standardized format for use in data exchange between the regional 'storage facilities' and the GMP 'data warehouse' will probably be necessary in order that the data warehouse can serve its intended purpose.

The problems and costs involved in developing new data exchange systems, and reporting formats databases, and in adapting databases to accommodate new systems should not be underestimated. Maintaining existing databases is, in itself, a costly matter that may well require additional resources if centres are requested to handle larger volumes of data. All efforts should therefore be made to make the best possible use of existing developments/centres, and to avoid 're-creating the wheel'. Collaborating in data handling efforts with established programmes and 'buying' data handling services from existing operations will likely be more cost effective than setting up new systems from scratch in many regions, and avoid duplication, and the possible negative consequences for all parties associated with this. At the same time, the diversity in regional capabilities in this connection needs to be recognized. In some regions, new data handling capability may need to be developed. Here again, cooperation (e.g. partnerships) with existing well-functioning systems in other regions may well have advantages, both financial and in terms of time required to implement capacity.

6.5.5 Some complicating factors

There are a number of issues that need to be addressed, both in relation to data management and in a wider context within the GMP. Not the least of these is 'language'. It may or may not be practical to insist on use of a common language (e.g. English, or the most widely used language within a region). However, at a certain point in the path from data source to data warehouse, language barriers will need to be bridged. Data reporting is not a one-way process. Those responsible for compiling and archiving data, or for evaluating and assessing data will want to address questions back to data sources , requests for missing components, requests for clarification, etc. This also applies to technical aspects of data,for example PCB to one person may mean polychlorinated biphenyl and to another pentachlorobenzene,agreement on and adoption of standardised coding for use in data reporting should be a matter of priority.

Relevant data are potentially available from many sources, both 'official' (governmental) and other (e.g. universities, peer reviewed literature). The Stockholm Convention evaluations will presumably need to make use of data from several sources, not all of which will be available in the form of data files. The GMP data warehouse at least will need to be able to accommodate 'data' in several formats, including 'documentation' in electronic or hard-copy formats.

In addition to restrictions on data that may be imposed by the 'data owners' for proprietary reasons, some types of information are sensitive and subject to national legislation concerning data confidentiality. Data on humans is a case in point. Data restrictions will typically apply that prevent any data being identified with a particular individual – and therefore data that are made available for international exchange tend to have a high level of 'aggregation', which can conflict with the desire for detailed information. Conversely, some countries have legislation that requires that data are made public. Both of these situations need to be taken into account in developing the GMP data strategy.

6.6 Data analysis

To promote comparability among the regions, harmonized assessment tools (such as statistical methods for temporal trend evaluations) and products should be agreed. This again will need to be determined in association with the further elaboration of the monitoring programme and the associated

assessment methodology. Some international programmes (e.g., OSPAR, AMAP, EMEP) are already employing standardized methods that could be considered for adoption by the GMP.

The reliable identification of trends will require that statistical evaluation be carried out on the design of each national trend monitoring programme contributing to the GMP, to ensure that it is powerful enough to detect trends of interest. This will involve establishing the target accuracy of the analysis.

It should be kept in mind that the statistical power is likely to be reduced when data from several laboratories are combined. Given the expected variability, based on results of inter-laboratory studies, it is recommended to record site-specific trends in POPs concentrations based on results of single laboratories.

6.7 Cost and financial implications

The costs of establishing the necessary systems within individual countries to allow them to collect and report data to GMP regional data centres are almost impossible to estimate. They will depend on both the volumes of data involved and the existing capacity within the country concerned. The governmental structures and way in which relevant institutions are organized and funded are additional factors. These will vary widely from country to country. It is proposed that the arrangements within a country to deliver the data to the GMP are a matter for the countries concerned, and fall under their Stockholm Convention commitments. Where capacity is lacking, capacity building mechanisms should be applied to institute the required infrastructures.

With regard to operation of GMP regional data centres, this will similarly differ from region to region depending on the existing situation, and in particular the availability of existing data centres that could serve as the regional centre (or a 'thematic' component within a regional centre network). However, at this level the costs of operating the regional data centre(s) should be possible to estimate based on similar activities within other programmes. Costs essentially comprise two components:

- **Establishment costs**: the initial investments necessary to equip a data centre with the necessary technology, and to implement (develop or adapt) databases and data handling routines so that they meet the requirements of the GMP.

- **Operating costs**: the costs to handle the GMP data on a routine basis, to receive data, apply QA/QC procedures, archive data in databanks, and produce required data products (in support of assessment activities). These are recurring costs, and primarily concern staff employment to handle the GMP datasets. These costs are partly a function of the volume (and complexity) of data involved.

Use of existing data centres can significantly reduce (or entirely eliminate) the need for *establishment costs*. *Operating costs* can also be substantially reduced by utilising data centres that are also used by other (international) programmes, thus avoiding the need to duplicate reporting of data that may serve several purposes/programmes; this also reduces the burden on the countries involved. Similarly, harmonization in data management procedures, data analyses and data products can all lead to cost-effective data handling solutions.

In some regions it may be possible to implement operation of regional data centres on the basis of cost sharing agreement between the countries in the region; in other cases, and also probably for the GMP 'data warehouse', this may need to be identified as a core activity requiring some central funding.

Several international programmes (AMAP, OSPAR, etc.) and their respective data centres (see Annex 1) should be able to furnish relevant information on financing of data activities that can be used as a basis for estimating costs of establishing and operating data (regional) centres.

Not included in the above, are the additional costs of 'data assessment' activities; for example convening expert groups to conduct evaluation and assessment of GMP data.

6.8 References

USDA Pesticide Data Program http://www.ams.usda.gov/science/pdp/Qc10.pdf
JECFA recommendations http://www.inchem.org/documents/jecfa/jecmono/v48je20.htm#3.2.3
ICES Environment data centre http://www.ices.dk/env/index.htm
ICES Reporting format http://www.ices.dk/env/repfor/index.htm
AMAP data collection http://www.amap.no/
UNEP GEMS/Water http://www.cciw.ca/gems/gems.html
Canada NPRI http://www.ec.gc.ca/pdb/npri/npri_home_e.cfm

Annex 1: Examples of existing data 'storage facilities' (suggest expanding table with more from other regions should be part of TWG activity)

Institute	Area of Expertise	Plus	Minus
Air data			
Norwegian Institute for Air Research (NILU)	Air monitoring data	Operating and developing monitoring databases for more than 3 decades; compile data from ca. 40 countries (Europe and Russia); data centre serves several other international programmes (AMAP, EMEP, OSPAR, HELCOM). Collaboration with data initiatives in Asia (EANET, Korea)	
Cooperative Program for Monitoring and Evaluation of Long-Range Transmission of Air Pollutants in Europe under Convention on Long-Range Transboundary Air Pollution (EMEP) (see NILU)	Synthesis of (regional) POPs data	Eurasia focus; all European countries plus Russia. Hemispheric transport and modelling activities	
Others			
Human milk/blood data			
AMAP human health group / Institut National de Santé Publique du Québec	Human tissue monitoring (blood and breast milk)	AMAP Human Health sub-programme data (Arctic focus); CHUQ coordinates QA/QC inter-comparison programme for laboratories involved in human blood monitoring (ca. 20 countries, Arctic, Europe, North and South America)	Data management activities targeted only to AMAP assessment needs at present
GEMS/Food	Human tissue monitoring (breast milk)		Data management activities in support of WHO breast milk surveys
Others			

Institute	Area of Expertise	Plus	Minus
Other GMP media – marine (biota, sediments)			
International Council for the Exploration of the Sea (ICES)	Marine monitoring data (abiotic/biotic)	Operating and developing monitoring databases for more than 3 decades; compile data from ca. 20 countries (focus on NE Atlantic region); data centre serves several other international programmes (AMAP, OSPAR, HELCOM). Reporting systems include internationally adopted coding systems and reporting of methodological and QA/QC information.	Reporting formats are detailed. Complexity of reporting formats has deterred reporting from some countries and potential data sources.
Others			
Other GMP media – freshwater, foodstuffs			
National Water Research Institute, Burlington, Canada	Freshwaters	Data centre for the UNEP GEMS/Water (Global Environmental Monitoring System/ Freshwater Quality Programme; global (ca. 70 countries)	Freshwater media are not GMP priority; mainly physical/water quality parameters for major rivers
GEMS/Food			
University of Alaska-Fairbanks (SYNCON)	Data management	AMAP Terrestrial/ Freshwater data centre (Arctic focus); Flexible data reporting systems; online database	Current status of operations?
Others			