Revised IMPACT Paper 5 to develop a Water Monitoring IT Vision (July 27, 2004)

Adapted from IMPACT Paper 5: Data Management (Draft) May 9, 2003
Authors: Ken Lanfear/USGS, Karen Klima/EPA, Ellen McCarron/FDEP, Julie Utter/FDEP, Robert King/EPA

A Vision for Enhancing Use of Information Technology in Water Quality Monitoring

I. Importance of Water Quality Data/Information Management

State, Interstate, Tribal and other organizations need to compile and manage water data and analytical reports so that the information is understandable and available to decision-makers, stakeholders, and public audiences. Assessments of watersheds typically require data from multiple political jurisdictions and sources that must be integrated. Data management has evolved significantly in the past decade due to the rise of the Internet, more emphasis on enterprise architecture, and world events that demonstrate the need for better security. This paper suggests and illustrates an approach for managing data on the environment that enables an organization to work with data partners to set priorities, address major water pollution issues, and report status and trends more effectively.

Management of water data is essential to a successful monitoring program. It must capture and preserve various types of data (chemical, physical, biological, fish tissue, toxicity, sediment chemistry, habitat, and land use) from various sources, for various water types (rivers streams, lakes, groundwater, estuaries, and oceans). It should allow streamlined data entry and retrieval, meet data standards, and include metadata while providing effective agency and stakeholder use and public access to the data.

Data management covers a variety of activities associated with collecting, developing, maintaining, and operating data systems to support a State's water quality management program. Data management is recognized by EPA as one of the ten basic elements of a State water monitoring program (Elements of a State Water Monitoring and Assessment Program, March 10, 2002). It is one of the 8 cogs in the water monitoring program wheel developed by the National Water Quality Monitoring Council.

Despite the growing importance of a water quality data management system, only 23 out of the 44 responding States reported adequate data management systems, when surveyed by the Association of State and Interstate Water Pollution Control Administrators (ASIWPCA) in 2002. The National Academy of Public Administration report *Understanding What States Need to Protect Water Quality (Keiner et al., 2002)*, found that data management is a state's second largest need, over 15 percent of a typical state's total resource needs. It recommended that "states should be prepared to make substantial investments in improving their environmental data systems and installing advanced information technologies and GIS that will enhance the collection and analysis of water

quality data so they can ultimately reduce the costs of managing their water programs by focusing their efforts on the most important water pollution problems."

The nature of data management is complex. Not only must data be gathered, but it must be managed, synthesized, interpreted and analyzed. The data flow spans related activities and other programs and agencies. Effective water data management must overcome technical, social, and organizational barriers and work with partners to integrate widely differing data formats, systems, data standards and metadata.

The cogs in the National Water Quality Monitoring Council's Monitoring Framework (see Figure 1) are inter-related, where it is not always clear where one cog ends and the next begins. Organizations need a clear and logically consistent method to help them manage data. This paper proposes a broad, comprehensive, approach to understanding the entire data flow within a water quality management agency/organization before taking action. By first studying the flow of data, a manager can then look for opportunities to adopt more cost effective and efficient information technology (IT) practices to reduce redundancies, eliminate duplication of effort, promote data sharing, and enhance collaboration and more informed decision making

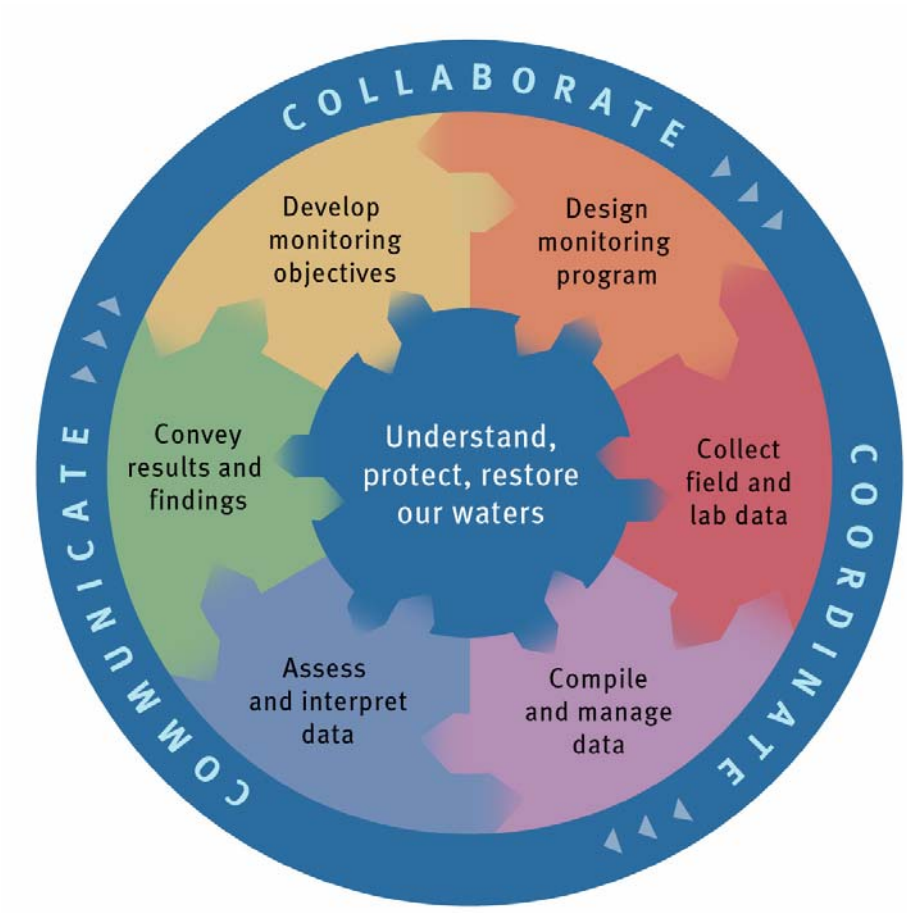

Figure 1. – NWQMC Framework for Water-Quality Monitoring Programs

II. The Basic Components of an IT Vision for Monitoring

The second part of this paper will describe how water quality monitoring can be viewed from an information technology (IT) perspective, following the flow of data and information through the cogs of the NWQMC Monitoring Framework.

A data flow diagram shows how data will flow from collection to decision making. It explains the role of data and its relationship to each component of the monitoring program, along with internal and external factors. It helps program managers make implicit understandings explicit, without the techno-babble that often intimates and confuses. Its basis is the logic model process that has been used for 20 years in program evaluation. ["Logic Models: A Tool for Telling Your Program's Performance Story", *Evaluation and Program Planning*, Volume 22, Number 1, February 1999, by John A. McLaughlin and Gretchen B. Jordan.]

Managers, partners and stakeholders are able to see exactly which data activities lead to what outputs critical to an effective monitoring program. This should be a team effort, reflecting the shared responsibility for data management. As the building process begins it will become evident that there are multiple realities or view of data management. Developing a shared vision of how data is supposed to flow will be a product of persistent discovery and negotiation between and among stakeholders.

Data flow is usually set forth as a flow chart, or network, that captures the logical flow of data, its interconnections or disconnects. It describes the linkages among program resources, activities, inputs, outputs, customers reached and outcomes. Boxes represent the significant processes or activities that are then explained with abbreviated text and linked with arrows to show sequence, inputs, outputs, and outcomes. One activity can lead to one or more different outputs. The processes can be described at many levels of detail.

Develop monitoring objectives.

Water quality data base management is driven by an organization's business, mission and monitoring objectives. The design of an information management system should heed an organization's rules, regulations, customers, and management practices. Customers, including researchers and those involved in regulating pollution, must have ready access to the data for analysis. Many different environmental programs or state agencies may use, and therefore, share, the costs of developing, maintaining and using water quality data. Upgrading data systems or maintaining a website may also be costs that are shared with others or conducted together for all of a state's environmental programs. Computer usage information also may need to be connected with cost accounting systems of the agency. Many government agencies, and all Federal agencies, now require formal capital asset planning for long-term investments in computing infrastructure. Typically, systems must be evaluated for such measures as return on investment and total cost of ownership. When these factors are considered, managers tend to favor enterprise architectures that offer common services and consolidated purchases.

This approach to measurement will enable the program manager and stakeholders to assess how well the program is working to achieve its short term, intermediate, and long term aims and to assess those features of the program and external factors that may be influencing program success.

Design monitoring program.

The design of the monitoring program determines the type of data collected, its scale, and its level of precision and confidence. It determines how data will be generated, how it is to be stored and retrieved, how it is to be analyzed and interpreted, and how the resulting information will be conveyed to the appropriate decision-maker in a timely manner and in a format that relates to the decision at hand. For example, a State monitoring program will most likely integrate several monitoring designs in a tiered approach to address management decisions at multiple scales (e.g., fixed station, intensive and screening-level monitoring, rotating basin, targeted, and probability design). An integrated design produces multiple types of data (e.g., chemical, biological, physical, sediment, groundwater, and stream flow) at various geographic scales and incorporates an array of tools (e.g., water quality and landscape modeling and indicators). In these ways, the monitoring design sets forth fundamental requirements for the data flow and supporting data management system(s).

Collect field and lab data.

Data entry procedures should be tailored to the collection process. Field crews need to enter relatively small amounts of data about stations visited on a trip; they may do this with laptops, personal data assistants (PDA's), cell phones, or they may transcribe written field notes. Samples sent to a laboratory for analysis, or recording media recovered from field sites, must eventually be matched to collection information. Event timing, buffering capability, and redundancy of the entire chain of transmission become critical issues in designing real-time systems, because real-time data must be captured when it arrives or it may be lost forever. Finally, accepting data from partners may involve batch processing to accept large amounts of data.

Data input necessarily involves a limited number of trusted, or at least identified, users whose identity must be validated through security procedures. The data management system also must provide record locking so that, for example, two people can't attempt to modify the same record simultaneously.

Blunders are incorrect data that result from such events as human error, equipment breakdown, or environmental conditions beyond the range of sensors. The object of quality assurance is to reduce the incidence of blunders in the database while allowing for genuine environmental variability and uncertainty.

The difference between variability and blunders is not always obvious. 238 °C is an impossible temperature for ambient water, but what about 32.8 °C? Both could result from mis-transcribing 23.8 °C. The former is easily caught by software that “knows” 238

°C exceeds the boiling point of water. Catching the latter may require more sophisticated tests, such as comparisons with recently recorded values or values at nearby stations. Quality assurance must be careful not to confuse a legitimate outlier with a blunder.

Some blunders can, and should, be caught and corrected upon data entry. All entered data should at least be checked for reasonableness. Seeing obvious, glaring blunders can cause users to distrust all data. Blunders also can confound display and analysis routines: try to plot a series with that 238 °C temperature, and all you'll likely get is a horizontal line with a spike. One of the most effective quality-assurance methods is simply to look at the data in a list or graph. As Yogi Berra said, "You can see a lot just by looking." Unfortunately, the looking often comes long after the data have been in the system. Therefore, the system should, provide some means indicating or removing questionable data and recording this fact.

Data may be assigned a provisional status until quality control checks are completed. This is important for real-time data, which usually comes straight from a field sensor with only minimal checking. Periodic calibration of sensors in the laboratory or field may necessitate later adjustment of data after they have been reported.

For Federal Government systems, the Office of Management and Budget has issued rules for quality assurance. Basically, the rules require a quality assurance process to be in place with procedures for investigating and resolving specific problems identified by users.

Compile and manage data

Water data must be stored so that it can be readily retrieved for analysis, interpretation, and public access. The ASIWPCA survey of State Water Monitoring Programs examined the predominant methods for storing data for types of data collected. These methods included electronic databases, spreadsheets, floppy disks or CDs, and paper files. The survey results showed that while States and Interstate agencies are increasingly storing data in electronic databases, a small number of agencies still use paper files as their predominant means of storing data. Additionally, historic data originally contained in paper files may not be converted to electronic files or databases due to resource limitations.

Database options include building your own or using national databases (such as EPA's STORET and the U.S. Geological Survey's (USGS's) NWIS). EPA's new STORET (STOrage and RETrieval) system provides an accessible, nationwide central repository of water information of known quality. In the future, EPA will require that all States use STORET either directly or indirectly (e.g., via the Central Data Exchange (CDX)). See www.epa.gov/storet for further information on STORET, including system updates for users and instructions on how to download data via the Web.

Data management includes following appropriate metadata and State/Federal geo-locational standards. Metadata – data about data – is important for finding data and

determining its suitability for use. In a water-quality database, the distinction between data and metadata is not very sharp. A dissolved oxygen value of 7.0 mg/l, for example, clearly is data, but station name, location, date, sampler, etc. – actually, the bulk of the record – could all be considered metadata.

The Advisory Committee on Water Information (ACWI), through the National Water Quality Monitoring Council, has published *Data Elements for Reporting Water Quality Results of Chemical and Microbiological Analytes*. (http://wi.water.usgs.gov/pmethods/elements/elements.html. The recommended data elements are grouped into 7 major topics:

1. Contact
2. Results
3. Reason for sampling
4. Date/time
5. Location
6. Sample Collection
7. Sample Analysis

One problem is that, for historical data, some important data elements simply were never recorded. Debates remain about whether some elements should be required or optional. The ACWI data elements are not organized into a formal schema of a data set, but USGS and EPA are discussing the need for a standard interchange schema.

Deciding how much metadata to include involves a trade-off between data entry and data retrieval. Data providers often object to excessive metadata requirements that slow the data collection and entry process. Users, on the other hand, rarely complain about having too much metadata. Unfortunately, the need for metadata may not become apparent until long after the data are collected. For example, recording the detection value for an analytical procedure may seem unimportant at the time to someone doing compliance monitoring; it could later become a critical to a researcher doing trend analysis.

All data systems require some minimal data display and transmission capabilities. Regardless of its display capabilities, a water database must be able to send data in a variety of standard formats. Protocols for exchanging water-quality data in XML are yet to be established. For the time being, delivery in a tab-delimited format (which is easily imported into spreadsheets) may be sufficient for many applications.

To use water data, users must first find it. Larger databases, such as STORET and NWIS, are easy to find because major search engines index USGS and EPA websites and many popular web pages link to them. This may not be true for the databases of smaller agencies. It's important that at least the "home page" of a database be designed to interact with search engines. Sites such as "Search Engine Watch"[1] (http://searchenginewatch.com/) offer help in this regard.

[1] Mention of commercial names is for identification only and does not constitute endorsement.

Search engine “spiders” – programs that read and index web pages – usually can not get through database interfaces. When faced with constructs such as, "Type a station number or name here," they have no clue what to do. Thus, while a search for "water quality data" on most search engines would quickly lead to the home pages of STORET and NWIS, a search for "Potomac River water quality" is more problematic. Google, for example, finds real-time data for stations on the Potomac River in Maryland only because the URL, http://waterdata.usgs.gov/md/nwis/current/?type=quality, is a link on other pages that are indexed by Google, not because Google checked the whole NWIS database.

Data users typically are looking for station data, a set of observations collected at a particular station or group of stations. Metadata may be useful as a screening tool; a user, for example, may want to see only data collected by a certain analytical procedure. User behavior for data retrieval rarely mirrors that of data collection. Thus, the output interface to a database may need to be very different than that used for input. Operations to capture data act in the dimension of time, usually the present. Data retrieval, on the other hand, tends to focus on the spatial dimension, the station. These characteristics can create some difficult demands on systems designed to serve both purposes. The National Water Information System (NWIS) of USGS, for example, is split into a collection side, called NWIS, and a distribution side, called NWISWeb. While this creates two copies of the data, which must be carefully synchronized, it vastly simplifies the distribution effort.

Assess and interpret data.

The return on investment of a data management system is not fully realized until the data are analyzed, interpreted, and the resulting information used to make informed decisions. An abundance of data might be available, but it is not useful unless it is evaluated to determine what story is being told. The ability of an organization to assemble, analyze, assess and interpret data for decision making is affected by how the data is compiled, documented, quality-assured, and combined with data from secondary sources, collected for a variety of purposes under a variety of quality control practices.

The data flow affects an organization’s ability to assess and interpret the data, and the organization’s assessment methodology can influence and determine the data flow. These are the typical steps for developing an assessment methodology (Source: www.epa.gov/owow/monitoring/calm.html):

- Identify the required or likely sources of existing and available data and information and procedures for collecting or assembling it;
- Describe or reference requirements relating to data quality and representativeness, such as analytical precision, temporal and geographical representation, and metadata documentation needs;
- Include or reference procedures for evaluating the quality of datasets; and
- Explain data reduction procedures (e.g., statistical analyses) appropriate for comparing data to applicable water quality standards.

Organizations need to store analytical reports so that the information is understandable and accessible. EPA strongly recommends that all States use either the Assessment Database (ADB) or an equivalent relational database for storing water quality standards attainment status for each assessment unit. (See Appendix B of the 2002 *Integrated Water Quality Monitoring and Assessment Report Guidance*.)

Convey results and findings

As the data and resulting information reach closer to the final information user, the information can be placed in a format and IT portal relevant and timely to the decision-maker's needs. While all decisions are not standard (i.e. require careful interpretation relative to an uncertain set of new conditions), many of today's water quality management information needs lend themselves to development of standard reporting formats. For example, fish consumption and swimming beach advisories driven by human health standards and preparation of 303d lists, using water quality standards enforced at the time of the list's creation.

It can be envisioned that each decision-maker within an agency, from triennial standards review, through standards compliance and permit writing, to planning, financial aid, and 305b reporting, can have an IT mechanism, constructed on the foundation of an agency's data base, to greatly facilitate access to the information fundamental to the decision at hand. By using IT such information can be accessed 'just in time' enhancing the timeliness and efficiency of an agency's staff.

Collaboration, Coordination, and Communication

In the article *Managing Troubled Data,* Stephen S. Hale, et al. [Environmental Monitoring and Assessment, Kluwer Academic Publishers, Netherlands, 2003.] explains how partnerships that agree on the flow of data can overcome many technical barriers.

A variety of secondary providers may access the data and offer it to other users in a value-added form. It may be mutually convenient to service these providers with a batch download, rather than their retrieving through the standard interface. Providing data upon electronic request in a standard format is equivalent to the "just in time" delivery systems of many industries. In doing so, in a reliable manner, an agency can discourage users from downloading data and saving it offsite. Besides possibly being cheaper and more convenient for users, "just in time" delivery ensures users work with the most current and correct data.

Conclusions

While the above 'vision' of employing information technology in all phases of the NWQMC's monitoring framework suggests a potential for improving both the efficiency and effectiveness of a water quality management program, there are many developments that need to take place within each of the framework's cogs. During a special session at the Fourth National Monitoring Conference in Chattanooga, Tennessee, in May 2004,

discussing IT developments in water quality monitoring, it was noted that there are a relatively large number of efforts underway at present. Agencies and software developers are making progress in all phases of monitoring, often, however, in uncoordinated ways. For a complete listing of the efforts reported to the NWQMC, see the Council's webpage: http://water.usgs.gov/wicp/acwi/monitoring/.

IT, as is often seen in business, comes with a need to be logical and transparent in the processes and methods used to obtain information in support of water quality management decision making. Furthermore, to justify the expense of developing a 'supply chain software' approach to managing water quality data and information, with a high level of sophistication, within all phases of a management agency, there may be a need for agencies, as it has been for companies, to employ more common processes and methods in order to gain cost effectiveness from acquisition of monitoring IT software. This fact, in many ways, calls for water quality management agencies to work together in defining 'standard' data and information processes that can then be 'programmed' into a sophisticated and advanced data/information system.

As business of the global economy advances while employing the latest developments in IT, it is not hard to envision how water quality management must also make major strides in advancing its efficiency and effectiveness by employing the latest developments in IT. The recent round of budget cuts being imposed on many state water quality management agencies offers an opportunity to rethink data and information acquisition process and methods and foster a collective evaluation of the advantages of greater use of IT.

Hopefully, this paper presents a vision of IT in water quality monitoring that prompts continuing dialogue and technological development. The NWQMC is developing a communication mechanism to foster such a dialogue and, with support of key water quality monitoring organizations and agencies, hopefully a detailed and agreed to vision and agenda for improving IT in monitoring will emerge.

References

(to be completed)