WHITE PAPER

Considerations for Regional Data Collection, Sharing and Exchange

Bruce Schmidt, StreamNet Program Manager
and
The StreamNet Steering Committee

June 1, 2009

BACKGROUND and PURPOSE

The need to share environmental data has grown significantly due to multi-agency programs like ESA recovery and shared management responsibilities. Agencies and projects collect data for their specific needs, but wider scale programs often require shared data from multiple sources. Data should be maintained and accessible for long term use and not lost when a project ends or staff changes. Achieving these ends will require action at various levels from the field to policy.

Environmental data are time consuming and expensive to collect, and should be utilized to the greatest advantage in managing and enhancing resources. To accomplish that goal, data need to be available for wider use beyond their initial local purpose. Public funding of sampling further emphasizes the need to make the resulting data available to others and the public. This guide outlines basic actions needed by various entities from data collection in the field to agency programs, funding programs, and policy levels to facilitate wide scale sharing and use of data. These recommendations are also summarized as checklists in Appendix C.

This is a general guide, independent of the purpose or use of the data, intended as a "nuts and bolts" description of the steps needed to establish a comprehensive approach to data sharing. The focus is more on the container than the contents. It is intended to provide a checklist of all aspects of data creation and use, even though many agencies and projects may already be adept at various aspects of it. It can inform development of new data management approaches and systems, or allow comparison of existing systems to these recommended components. The guide does not prescribe specific actions but attempts to list the issues and discuss the various paths available for addressing them. It relates to data sharing approaches as they currently are. Ideally, in the future data sharing will become a routine part of wide-scale, multi-agency monitoring programs rather than the current more *ad hoc* sampling.

ROLES AND RESPONSIBILITIES

Various entities have roles and responsibilities in effective data sharing. Executives at the regional policy level need to make basic decisions about priorities (which data should be shared) and provide specific policy guidance. Funding entities, including regional and federal agencies, can negotiate the specifics of data creation, management and sharing and enforce them in contracts. Agencies that conduct data creation in the field are responsible for meeting their statutory mandates and providing guidance and resources to their field staff for applying agency and regional policy and funding entity guidance. Individual field samplers are responsible for implementing that guidance as data are created. The individual sampler and data creating agency roles are closely aligned. And, regional scale database management projects can provide technical services and perform many required data sharing functions.

- Agencies and field samplers.

Many agencies and programs collect environmental data in support of their missions and mandates, including state, tribal and federal fish and wildlife agencies and programs, state and federal environmental quality agencies, state and federal land management agencies, etc. The sampling is done in the field by various agency staff, project staff or consultants. Many of the sampling and data management recommendations discussed here are influenced by agency policies, support capabilities and internal requirements. Agency policy should also provide guidance to their respective sampling projects in order to implement these guidelines.

Due to different purposes, different environments, and historic data, it will not always be possible to standardize sampling and data management among agencies, even though that would simplify data consolidation and sharing. There may be several ways for agencies to implement these recommendations. Therefore these recommendations are intended to urge samplers and agencies toward maximizing standardization to the degree practicable, but to managing the data to facilitate consolidation and sharing when standardization is not feasible or possible. Where preexisting requirements (agency, funder or legal) are in effect, they should take precedence over this general guidance. This guide is intended to provide recommendations to fill the gap where no specific requirements are currently in place or being followed. It may also be used by organizations that currently have data management systems and guidance in place as a means to compare and evaluate existing practices, and to potentially supplement or streamline processes.

- Funding entities

Various agencies and entities fund field sampling to create environmental data. For this guide, funder recommendations relate to entities that provide contract funds to others to do work, such as the Northwest Power and Conservation Council's (NPCC) Fish and Wildlife Program funded by the Bonneville Power Administration (BPA), state programs (Oregon Watershed Enhancement Board (OWEB), Washington Salmon Recovery Fund Board (SRFB), etc.), federal programs (Pacific Coast Salmon Recovery Fund (PCSRF)) and individual federal agencies that fund work outside their agency (e.g., Environmental Protection Agency (EPA), U.S. Forest Service (FS), Bureau of Land Management (BLM), etc.) Work done within these agencies by agency staff would be considered under the Agencies and Field Samplers sections.

All entities that fund environmental sampling have the ability to negotiate or establish specific requirements. These may relate to agency mandates, policies, legislation, and technical considerations. In some cases it may be appropriate to negotiate sampling methodology to meet each party's data needs. Funding entities may wish to specify data management requirements in contract language to assure that data are maintained and shared appropriately and not lost at the end of a project. Such language could apply to all projects to assure compliance with national programs (such as feeding water quality data to a national database), or be project specific. Language could be original, or could reference one or more published documents. Contracts could include relevant recommendations from this guidance document, with recognition that there may be multiple means to accomplishing these objectives, and different procedures may be appropriate for different kinds of environmental data or agencies.

- Policy level

Policy level guidance relates to decisions made at the executive level by the heads of involved state, tribal, federal and regional entities. Since a goal is to establish regional consensus on monitoring programs and data sharing, a collaborative approach to establishing formal data management and sharing guidance is important. Any policy level collaborative group should include the agencies and organizations that create environmental data, use data collected by others, and fund monitoring and data management activities. Policy level issues may include setting priorities for which kinds of data to be shared and addressing other policy level questions, including those posed later in this guide.

- Database management projects

A number of regional scale database management projects are available to provide advice and data management services (see Table 1 for a partial list). These usually specialize in specific kinds of data or meeting specific program needs. In some cases, these projects can perform data management and sharing tasks for other projects and agencies, and can be consulted to take advantage of their technical expertise. Incremental costs for these services may often be lower than developing similar expertise or capability in house.

RECOMMENDED ACTIONS

The following actions represent a series of recommendations or steps to consider as part of a comprehensive approach to data management and data sharing. Many actions can have several suitable approaches or options. In some cases, one approach may be identified as best or ideal, but final decisions may depend on the specific needs or capabilities of a given agency or project. The various entities often have different rolls within each recommended action.

1. Standardize sampling to the degree possible

Many different agencies and projects collect similar kinds of data, but often with different objectives, approaches or methods. This reflects the longstanding nature of many monitoring programs, individual agency mandates, different purposes for sampling (addressing different

questions), and the need to function effectively in local conditions. At the same time, broad scale issues like ESA recovery, subbasin planning and multi-jurisdictional management are best served when relevant data from all sources can be combined and analyzed seamlessly.

There is growing regional interest in employing common sampling methods among agencies to facilitate comparability and sharing of like kinds of data, but adopting field methods that adhere to regionally recommended protocols may require altering existing, sometimes longstanding sampling approaches. Agencies need to decide whether to ask their field staff to adopt regionally recommended sampling methods or to maintain existing practices.

Complete standardization is difficult to achieve due to variability in the purposes for sampling and the environments being sampled. Also, absolute adherence to standards can stifle innovation or improvement of methods. However, actions to limit the number of acceptable sampling protocols, both within and between agencies, and fully describing the sampling protocols used would significantly ease compilation of data sets from multiple sources and enhance data compatibility for broader scale use. The recommended approach is to participate in appropriate wide scale collaborative efforts to establish agreements on a limited number of sampling methodologies. Alternatively, field sampling could be consolidated into regionally agreed upon coordinated monitoring programs, also developed through a collaborative process. Collaborative efforts will require participation by all interested parties, including the agencies that conduct field sampling and the entities that utilize data from multiple sources,

- Agency actions:
    - To maximize data comparability, sampling agencies should utilize consistent sampling methodology to the greatest degree practicable. Ideally, methods should at least be standardized within each agency. The goal should be to provide the most consistent and useful information at an agency and a regional scale.
    - If agencies can not or choose not to adopt regionally recommended standard sampling protocols, they should make that decision known so that regional emphasis can shift to focus on means to consolidate the data produced by different methodologies.
    - Provide agency perspective and expertise by participating in collaborative regional efforts to recommend standard sampling protocols or create coordinated monitoring programs. Collaborative efforts should serve to select a limited set of recommended appropriate methodologies.

- Field sampler actions:
    - Follow agency guidance and adhere to established sampling protocols and methods as much as possible. Avoid developing new sampling approaches independently. If this is unavoidable, then modified or newly created protocols should be described and provided to regional collaborative bodies for review and evaluation.
    - Describe and document the specific sampling protocols or method manuals you followed in all publications and data descriptions. Prototype tools are being developed at the PNW regional scale that should simplify this task (e.g., PNAMP Protocol Manager). If sampling is done consistently, then describing the method is a one-time effort.

- Record any adjustments to or deviations from established sampling protocols. Many things can affect actual sampling, such as weather, equipment malfunction, flow, changes, etc., and any resultant changes to standard approaches must be recorded so that subsequent users of the data can understand the context.

- Funder actions:
  - If project data are to be shared, funders should negotiate with project sponsors to ensure sampling methodology meets funder and sponsor needs and is appropriate for the sampling environment. Contract language can be used to assure agreed methods are used.

2. Follow existing data management guidance documents

Data management standards relate to how data are defined, coded, error-checked, documented, recorded, published and shared. Consistent use of established standards simplifies and improves the ability to combine and share data. Currently available guidance includes "Best Practices" documents for reporting location and time information (http://www.nwcouncil.org/ned/time.pdf), for creating a data dictionary (http://www.nwcouncil.org/ned/DataDictionary.pdf), and for developing a data management plan (http://www.nwcouncil.org/ned/Checklist.pdf). Participation in collaborative groups to create additional guidelines and standards is encouraged. These standards relate to common types of information that describe or qualify the sampling effort. They do not dictate the specific environmental metrics to be measured.

- Agency actions
  - Adopt specific Best Practices recommendations as standard procedure for agency staff.

- Funder actions
  - If adherence to specific Best Practices is important to the project, contract language can be used to specify required practices.

- Field sampler actions
  - Follow the Best Practices for managing data as specified by agency and funder.

3. Automate data capture and management, to the degree possible

Computerized data capture and management is becoming cheaper and more effective, reliable, and efficient. Ideally, data should be entered into electronic format in the field or immediately afterward, and then flow into an agency-wide data system. Such systems provide multiple benefits at all levels: immediate and accurate data entry; data validation on entry; automatic generation of metadata; local control over data management and updates; canned analyses or standard outputs to analysis programs, canned reports at the field and agency levels; automatic data consolidation agency-wide; support for comprehensive analysis at the agency level; and automatic translation and output into regional data sharing formats.

Costs for developing systematic approaches to data management are decreasing, and often the largest challenge isn't expense but expanding the data management focus to an agency-wide perspective. Concepts (and sometimes computer code) can be obtained from agencies already using the technology and adapt it for use. Assistance from regional database projects is often available to organizations planning and developing data management systems.

- Agency actions
  - Work toward developing comprehensive data management systems for the high priority types of data for the agency. These can include field data entry devices, data validation routines, agency wide databases, etc. An iterative, modular approach by data type would be least expensive and is recommended.
  - Adopt a partnership approach between biological staff and IT specialists to design and construct agency wide data systems and other tools.

- Field sampler actions
  - Field test data input devices at the field level, as they become available. Participate in system development as opportunities arise. Field level input is critical to ultimate system success. Provide feedback early in system development and testing.
  - Adopt use of tools to input data in the field or immediately after collection.

- Funder actions
  - Support agency efforts to deploy field data collection tools and develop consolidated data systems, since these will be critically important and are prerequisites to feeding data into any regional scale data sharing approach.

- Database project actions
  - Develop field data capture applications on request of agencies.
  - Assist agencies with design and programming of agency database systems and tools.

4. Use common coding and formatting, and describe in a data dictionary

Many agencies have established code lists for common sampling elements (for example, species) that may be mandatory. There are few code lists adopted among agencies, however. In the absence of agreed-upon region-wide coding systems, some database projects have developed systems to support combining and storing data from multiple sources for dissemination. Examples include StreamNet's (www.streamnet.org) "Data Exchange Format" and the Pacific Northwest Water Quality Data Exchange's (http://www.ecy.wa.gov/pnwdx/pnwdx_main.htm) "Data Exchange Template," and there are specific standards for managing coded wire tag and PIT tag data established by the Regional Mark Processing Center (http://www.rmpc.org/) and the PIT Tag Information System (http://www.ptagis.org/), respectively.

For new sampling programs collecting data elements that are already included in a regional exchange format or regional database system, we recommend use of that coding and format system. For sampling efforts already using other formats, project sponsors should work with the appropriate database project (Table 1) to ensure that the data can be output in a common

exchange format for data sharing. For data types without adopted regional scale formats, participation in collaborative efforts to develop common coding and formats is encouraged. It is important that data are defined consistently within and among agencies, and this is simplified by utilization of agency-wide data systems. Some apparently similar data types can be incompatible if defined differently, for example, using different length and diameter definitions of what constitutes "Large Woody Debris." To maintain efficiency, it would be helpful to prioritize which attributes are most important to standardize for sharing on a wide-scale perspective. If at all possible, new coding systems and data definitions should not be created, but should be adopted from existing efforts.

Most critically, all data definitions and codes should be described in a data dictionary for each data set. This is simplified if each type of data is standardized within the agency, requiring only one dictionary for each type of sampling. A data dictionary is a critical component for describing a data set and making it understandable to others. The dictionary needs to include definitions of all data elements plus information on units of measure, format, field sizes, acceptable values; data coding and lookup tables; and information about the table structure and relationships if in a relational database. Additional information about developing data dictionaries can be obtained from *Best Practices for Data Dictionary Definitions and Usage* (http://www.nwcouncil.org/ned/DataDictionary.pdf) or a regional database project.

- Agency actions
  - Utilize standard code lists for common data elements within the agency.
  - Develop agency wide code lists and data dictionaries, by type of sampling.
  - When adopting new code lists, work with other agencies to adopt a common set of codes that are consistent among agencies. Try not to create any new, individual data coding systems.

- Field sampler actions
  - Adhere to standard code lists as established by your agency.
  - When developing new code lists, work with others to adopt a regionally consistent set of codes. Try not to create any new individual data coding systems.
  - Follow the appropriate agency data dictionaries, or if there aren't appropriate ones for your type of sampling, develop a data dictionary for each data set.
  - Provide a copy of (or link to) the data dictionary in the metadata (Recommended Action 5).

- Database project actions
  - Provide agencies and projects with existing data definition and code lists, as requested.
  - Assist agencies and projects with development of data dictionaries, as requested and within the scope of the data types addressed by the database project.

- Funder actions
  - Address any need for specific data coding and a data dictionary through negotiation with the project sponsor. Specific needs can be included in contract language related to metadata (Recommended Action 5).

5. Describe your data so that others can understand and use them

For every data set there should be a set of descriptive information that allows others to fully understand the data and how to use them. Such descriptive information, or data about the data, is referred to as "metadata." This is a technical requirement of all approaches to automate data transfer. Metadata includes information about who collected the data, what data were collected, how the data were collected, how the data elements are defined and coded, what purpose they serve, where and when they were collected, and where the data reside and can be accessed.

For geographic data for use in a GIS, the data should adhere to the minimum metadata standards as prescribed by the Federal Geographic Data Committee (FGDC, www.fgdc.gov/metadata), which should be familiar to all GIS professionals. Tabular biological data should be described following the FGDC Biological Data Profile. Descriptions of various metadata creation tools are available through National Biological Information Infrastructure (NBII) of USGS at http://metadata.nbii.gov/portal/community/Communities/Toolkit/Metadata/FGDC_Metadata/Tools/

These standards indicate that only a small portion of potential metadata is absolutely required, but minimal required data often don't provide sufficient information. We recommend a somewhat larger set of minimal metadata (Appendix A). Full metadata would be even more useful and would minimize subsequent requests for additional information about a data set, but full metadata is not required by FGDC. The amount of metadata required could be scaled based on regional priorities for sharing specific kinds of data. Assistance with developing metadata is available from NBII at http://www.nbii.gov/portal/community/Communities/Toolkit/Metadata/ and from regional database projects for specific types of data (Table 1). Agency-wide data systems would be useful in automating creation of metadata.

- Agency actions
    - Adopt as agency standard practice that all data sets should be accompanied by descriptive information (metadata).
    - Phase in the requirement for metadata. The task of creating metadata only appears daunting at first. Once a few data sets have been described it becomes much simpler to create metadata for additional data sets because the majority of descriptive fields can simply be copied over from existing metadata.
    - Metadata for existing or historic datasets can be developed sequentially over time. Note that much of the metadata can be cut and pasted from previous sets of metadata, making the job easier over time, as only the differences need to be newly described.

- Field sampler actions
    - Include a set of descriptive information with all data sets. Note that after one data set has been described, it is often possible to simply copy the descriptions for additional data sets, with only a few basic pieces of information like location or species changing. Updating descriptive information in subsequent years is often simply a matter of adjusting dates and describing any unusual events for the subsequent year.

- Funder actions
    - Contract language can be used to require that metadata be prepared and supplied with data sets created and supplied under the contract.

- Database project actions
    - Share expertise with projects and agencies regarding metadata creation, as requested.

6. Publish the metadata

Not only should metadata be included with every dataset, metadata for every data set that will be shared should also be publicly available so that the metadata and data can be found by online searches. Posting the metadata on the Internet (ideally as Extensible Markup Language – XML) is a prerequisite for being able to locate the data through online clearinghouses or portals. There should be a long term commitment to keeping the metadata updated along with the data over time unless it is a completely static data set. Several approaches to publishing metadata are available.

a. Publish the metadata as a web service. This is the preferred approach for ongoing projects, since it results in only a single copy of the metadata being made available. The web service is registered with the desired clearinghouses (e.g., NBII) and/or portals (e.g., the NED Portal at http://gis.bpa.gov/Portal/, Geospatial One Stop at http://gos2.geodata.gov/wps/portal/gos, etc.) which then simply point to the original metadata, resulting in only one copy of the metadata to keep current. Assistance with establishing a web service can be obtained from agency GIS programs, agency IT departments, regional database projects, etc.

b. Use an intermediary project to host metadata as a web service. Regional database projects (Table 1) that work with the kinds of data being developed by your project (such as StreamNet for fish abundance data, PNW Water Quality Data Exchange for water quality data, etc.) can often host metadata as a web service for clients. This can be useful if the project does not have the necessary technology or staff, or is not a long term project. Regional database projects may be able to provide long term maintenance for the metadata, and serve as the single place to contact to update the metadata.

c. Publish the metadata on a clearinghouse or portal. This approach can be expeditious if a project does not have the necessary technology available to host its own web service. For example, the NED Portal has an online tool to create and post metadata. Be aware that other portals often harvest and republish metadata, resulting in duplicate copies, creating a burden to locate and update all duplicates as changes are made to the data set and metadata.

- Agency actions
    - Decide which data sets will be made available for sharing.
    - Establish an agency approach to publishing metadata, including decisions on how the metadata will be published (by the agency or through an intermediary database project) and promote the plan within the agency.

    - Provide metadata creation and validation tools, preferably a single agency-wide application. There are several of these on the Internet that are available for free download and use (see Recommended Action 5).

- Field sampler actions
    - Follow agency policy regarding publication of metadata.
    - Update the online metadata as data sets are updated.

- Funder actions
    - Contract language can be used to specify how metadata are to be made available under the contract after negotiation with the sponsor.

- Database project actions
    - If within project capabilities and area of project responsibility, host metadata and post as a web service on request of partner agencies. Capability and data type focus varies by project.

7. Assure control over data quality

Specific attention needs to be paid to quality control to assure that data are accurate and appropriate for their intended use. A variety of specific actions are needed at every step of the data cycle, from initial collection through ultimate use of the data. Quality Control procedures should be incorporated in the data collection process from the very beginning. In general the following suggested QC procedures and steps should be adopted to the degree possible:

- The sampling design should be reviewed by a statistician to insure that representative measurements can be made with appropriate accuracy and precision to minimize error within a desired level of confidence.
- Employ sampling methodology suitable for the intended purpose and for the environment being sampled.
- Follow a prescribed sampling protocol, and record any specific differences or deviations.
- Statistical techniques should be employed early during data collection to monitor the performance of the methods to successfully address issues of variation and repeatability and enhance the probabilities of obtaining accurate and precise measurements.
- Plan sampling and data coding to minimize the opportunity for data translation errors.
- Competent, well trained personnel should be used for sampling. Provide them with training and sampling manuals.
- Enter data into electronic format as quickly as possible (see Recommended Action 3). Potential actions could include:
    - Use handheld or other data entry tools in the field for original data recording, if possible.
    - Use redundant data capture, such as voice recordings along with direct electronic entry, to provide a back up when entering data directly in the field.
    - Use double entry to validate accuracy when entering data from forms.
    - Automate data entry to the degree possible (pull-down lists, pre-populated fields where possible, required formats, etc.)

    - Automate data validation to the degree possible, with built in range checks, required formats, review of summary statistics, etc.
- Back up data immediately; archive a copy in a safe, preferably different location.
- Review the data at the end of each sampling session for obvious errors.
- Errors discovered in the field or at any later review should be shared back to the data originators for correction in all versions of the data.
- Maintain close control of versioning of the data set. Document any changes made to base records
- Keep the data flow pathway as short as possible from collection to storage and ultimate use. For example, have a single official data set and send people to it rather than passing data sets from person to person.
- Limit the number of data processing steps to only once for each stage of treatment.
- Check for all data entry and other errors before reports are generated or the data are placed in permanent storage.
- Record all QA/QC steps and procedures used and include that information in the reports and metadata associated with the data. Include the Quality Assurance Project Plan if one was required by a funding entity.

Overall, data quality control can best be provided by those people most familiar with the data.

- Agency actions
    - Develop an agency wide Quality Assurance program and require compliance.

- Field sampler actions
    - Follow the principles and procedures in the above recommended QA steps and as contained in your agency's QC process.

- Funder actions
    - Contract language can be used to require that QA steps be clearly articulated and described in the data management plan (Recommended Action 8).

8. Prepare a data management plan

All projects that collect data should prepare a data management plan prior to data collection. Such a plan could be made a requirement for funding in contract language. Such a plan does not need to be lengthy, but it should clearly describe how data are going to be collected, stored, managed, quality assured and shared. Issues of sensitive data, timeliness of delivery, ultimate disposition of data, etc. should be detailed in the plan. Developing a plan assures that all steps in creating, managing and sharing the data are considered. One suggested approach to developing a data management plan is outlined in Appendix B. The *Checklist for Organizing Field Data Collection and Management of Data* (http://www.nwcouncil.org/ned/Checklist.pdf) may also be useful in developing a data management plan. The data management plan, or a link to it, should be included in the metadata describing the data set(s).

- Agency actions
    - Require that sampling projects develop a data management plan before initiating data sampling in the field. For ongoing monitoring that does not yet have a plan, request development of a plan prior to the next round of sampling. The plan should cover specific kinds of sampling, and apply to all such field efforts across the agency.

- Field sampler actions
    - Develop a data management plan for sampling activities, or use an agency plan.

    - Follow the steps outlined in the data management plan when collecting and managing data.

- Funder actions
    - Contract language could require that a data management plan be submitted to the funder prior to initiation of sampling. Any specific needed approaches to data sharing and management should be negotiated with the sponsor and included in the plan.

9. Prepare a data analysis plan

If the project will create summarized or analyzed data, the analysis approach used should be described in a separate Data Analysis Plan (for detailed analysis) or described specifically in the Data Management Plan (for data summarization or simple analysis). The data analysis plan, or a link to it, should be included in the metadata describing the data set(s). This plan may also be useful in the Methods section of any reports or publications that result.

- Agency actions
    - Assure that data summarization or analysis steps are described in a data analysis plan or in a section of the data management plan.

- Field sampler actions
    - Develop a data analysis plan or include information on the analysis procedure used in a section of the data management plan.
    - Follow the steps outlined in the data analysis plan when analyzing or summarizing data.

- Funder actions
    - Contract language could be used to require a data analysis plan or to include analysis procedures in the data management plan, depending on the needs of the funder or through negotiations with the project sponsor.

10. Plan to share data

There are several options for making data available so that others can obtain and use them for wider scale analysis. Which approach is chosen can be influenced by agency policies, available agency or project IT capacity, funder requirements, available assistance and support, and/or commitment to long term maintenance. At a minimum, the contents, location, availability, and

methods used to collect and analyze the data should be described in the metadata, which should be made publicly available (Recommended Action 5).

The preferred means of sharing data is via the Internet. At a minimum, the data should be posted in a machine readable format to allow subsequent data use, such as in a relational database or spreadsheet application, GIS files, or if as text, in a delimited file format (ASCII). Data provided in .pdf format or summarized in project reports are not sufficient for sharing data. Ideally, the data, or a link to request the data, would be available at all times.

Data files may be made available in various ways, including File Transfer Protocol (.ftp), links on a web page, an online database and data query system, an Internet Map System, a Distributed Data Base Management System (DDBMS) or some combination of all of the above. In all cases, the existence of the data and the means to acquire the data should be included in the metadata, in project reports, and, ideally, also described on a web page. For very large data sets too large for direct download, or subject to specific requirements for use, the means to obtain the data, along with any limitations, should be publicized on the project website and in the metadata.

Project sponsors have essentially two options for managing and posting data for sharing: 1) posting and maintaining the data directly themselves, or 2) posting the data through an intermediary. The preferred approach will depend on a number of factors including the needs of the project, the type of data being collected, the longevity of the project, available IT infrastructure, and the sponsor's desire to maintain the data and update them as necessary.

a. Posting data directly on the Internet

   A minimum data management infrastructure sufficient to support a project website and a commitment to maintaining project databases are required before a project or agency can post and maintain its data directly via the Internet. Data can be posted in database or spreadsheet format for direct download from a project website, through file transfer protocol (.ftp)or as XML. For larger or more complex data sets, the data owner may need to provide additional tools to query the database so that users can locate specific data within the overall database. The data owner also must be prepared to continually maintain and update the data and metadata as necessary. This would be simplified by utilizing an agency wide data system (see recommendation 3).

   Large data sets that require more extensive database management systems and more complex approaches to serving data, such as on-line data query tools and/or Internet Map Services require more specialized expertise and capabilities. These resources may be beyond the purpose and available level of support for some projects. Some projects may be short term or not sufficiently staffed to manage databases and data distribution functions into the future. In such cases, it could be more efficient and effective to utilize an intermediary to post the data.

b. Posting data through an intermediary

   Where field samplers or agencies do not have the time, resources or interest in maintaining data on the Internet over an indefinite time period, a number of options are available. Some

commercial sites, e.g., Google, will host simple spreadsheets or database files for low or no cost, and the URL for the data set can be publicized on a project website.

Another, more focused approach is to work through regional scale database projects to have the data posted and maintained on the Internet. A number of database projects consolidate, standardize and disseminate specific subsets of environmental data in the Pacific Northwest (a partial list is contained in Table 1). Some, like the Pacific Northwest Water Quality Data Exchange (PNWQDX) and StreamNet provide data hosting services for the kinds of data they specialize in, either directly or through their partner agencies. StreamNet also has an archive program (the Data Store, http://www.streamnet.org/online-data/datastore.html) that can accept and post any kind of data, and the StreamNet Library (http://www.fishlib.org/) will archive any natural resource related documents. A data hosting service makes the data available over the Internet and also publishes the metadata to make the data findable through portals.

In cases where there is a regional database project that specializes in the type of data being collected, the data, or a URL to the data, should be provided to the database project, even if the archiving function isn't needed. This assures that all data of the particular type are consolidated across agencies for maximal value to the broader region.

Individual database projects have different procedures for handling data, so project sponsors should contact the appropriate database project(s) early in their planning to discuss requirements, procedures and data formats (Table 1). For example, fish related data in the StreamNet database are usually managed by StreamNet project staff in the partner fish and wildlife agencies or are sent directly to the regional database at PSMFC if they are data that do not conform to the StreamNet data exchange format. Water quality data in PNWQDX are submitted through and maintained within databases in the state environmental quality agencies, and data can be housed in a host database for partners unable to serve data on the Internet.

- Agency actions
    - Decide on an agency approach to sharing data sets. Actions could include hosting data sets on the web directly, using an intermediary database project to host data, or some combination.
    - Contact the database project appropriate to the kind of data to determine the details of how to submit data.
    - Update data sets as appropriate for the type of data, either internally or at the database project.
    - Even if a regional database project is not used to host data, the database projects appropriate for the type of data should be notified of available data and updates so that they can point their users to the data.

- Field sampler actions
    - Follow agency directives regarding sharing data on the Internet. Complete QA procedures and post your data set on the Internet, transmit data to your agency for posting, or provide it to the relevant database project for them to post.

- If not covered under an agency policy, field sampling projects should contact the database project appropriate to the type of data to explore data hosting and to provide data and updates.

- Funder actions
  - Contract language should specify how data are to be shared, based on negotiations with the project sponsor.
  - The approach for long term management and update of the data should be negotiated with the project sponsor.
  - Negotiate with project sponsors to assure that the details of data sharing are included in the data management plan.

- Database management project actions
  - For database projects with these capabilities and within the scope of data types they address, host agency data sets and make them available on the Web, as requested. Incorporate data in project data systems, if appropriate, or host as independent data.
  - Negotiate with agencies, projects or funders for large efforts that might require additional resources.

11. Establish data sharing priorities and policies

A number of policy issues related to data collection, management and sharing need to be agreed upon collaboratively at a regional scale. Data collected or developed with public funds should be considered public data and should be made readily available to others. Within that premise, specific policy guidance is needed to address issues that cross agency authorities. Consensus on policy needs to be reached among all involved entities, including field agencies that collect data, agencies that use data from multiple sources, and funding entities. A collaborative approach is recommended. While there may be numerous policy issues in need of resolution, the following topics are of immediate importance.

a. Data sharing priorities

Agencies and projects collect data of many types for many purposes, but not all data sets are needed for regional scale sharing. Regional consensus on which specific types of data are highest priority for wide sharing would allow focusing efforts initially on those data sets of greatest wide-scale utility.

b. Timeliness of sharing data

Regional scale entities often need data quickly, while the data originators are busy and may need time to consolidate and quality check data or to fully analyze their data and complete manuscripts and management recommendations, leading to concern over early release of the data. A regional policy is needed to promote rapid sharing of data but protect the interests of the data originator. Timeliness standards will need to be flexible depending on various circumstances, such as whether the data are from annual monitoring or are part of a multi-year sampling design. Policy could indicate, as a general rule, that data from annual

monitoring should be made available prior to the subsequent round of sampling. But, there may be reasonable concerns over premature release of partial data from multi-year sampling designs, and a release schedule for such data may need to be negotiated. Absent regional policy, negotiations between project sponsors and funding entities will be needed as part of a process to develop required schedules for data availability. At a minimum, data release schedules should be addressed as part of the data management plan and in the metadata. Regional policy should indicate general timelines for sharing various kinds of data.

c. Right to first use of data

Related to timely release of data, the originators of data may be reluctant to release data before they have had the opportunity to publish results based on the data. A regional data use policy could allow for conditional release of data with a provision that limits subsequent publication or sharing of the data set until a specified date or after publication by the originator. Such a limitation could be enforced by requiring a signed data sharing agreement prior to data release, and the FGDC Biological Data Profile contains a field to specifically record such a requirement in the metadata.

d. Release of sensitive information

Handling sensitive data is another important consideration that would benefit from a regional policy. Any proposed policy should recognize legal and agency constraints, but should facilitate sharing of data to responsible parties. Policy decisions should include defining what constitutes "sensitive" information, and allowing restricting release of sensitive information only to agencies or entities with recognized need, or specifying that information may be released with sensitive information generalized, such as to protect individual site locations. The policy should also require that any restrictions on use of sensitive data be specified in the data plan and in associated metadata.

e. Regional approach to building a data sharing system.

There has been interest expressed in developing a regional (Columbia Basin to Pacific Northwest scale) data delivery system as a means of making environmental monitoring data widely available. While the technical aspects of how to create such a system are technical IT, not policy level, questions, there is a need to establish policy on what such a system should be tasked to do if undertaken, what data types should be included, comparability of data from different sources (data standardization or interoperability), and agency capabilities and needed support. Such policy discussions would ultimately need to deal with all of the steps outlined in this data sharing guide, as well as needed features and cost. It will be essential to include all entities involved from field data collection to regional scale in such a collaborative effort.

- Executive actions at a policy level
    - Utilize a collaborative process that includes data creating agencies, wider scale data users, regulatory agencies, and funding entities to develop and collectively establish policy related to the above issues and any additional data sharing issues as necessary.

- Funder actions
    - Appropriate policy actions as described in policy documents should be supported through contract language, as needed.

- Agency actions
    - Participate in collaborative efforts to establish policies.
    - Endorse regional policy documents, and adopt within agency
- Field sampler actions
    - Follow policy, as appropriate. Include appropriate policy in data management plans and describe in metadata, as needed.

- Database project actions
    - Participate in regional collaborative processes to provide data management expertise and IT recommendations.

CONCLUSION

This guide outlines various approaches that agencies and sampling projects can take to preserve their data and make them available for use by others. Adherence to the data management considerations outlined here would significantly improve the quality and availability of data for use beyond the data originators, and some recommendations would facilitate data use within originating projects or agencies. The extent to which these practices will be used depends on directions issued by project funders and voluntary adoption by agencies. They would be strengthened by development of a full suite of best practices, including example data sharing agreements and formal policy statements as outlined above.

To make a significant difference, the principles and recommendations from this general guidance document should be incorporated into agency and regional policies as appropriate, and become part of formal business practices. It would be beneficial to the region as a whole, data creating agencies and to any organization that uses data from outside its own collection programs to participate in collaborative efforts to refine and implement these recommendations.

Table 1. Partial list of database / data warehouse projects in the Pacific Northwest.

Name	Website	Data Types
StreamNet	www.streamnet.org	Fish abundance (redd counts, dam counts, hatchery returns, etc.), fish distribution, 100K hydrography, fish related facilities (hatcheries, dams, barriers, passage, screens, etc.), hatchery releases, age, Protected Areas, etc. Also will store and disseminate any other data.
Pacific Northwest Water Quality Data Exchange	http://deq12.deq.state.or.us/pnwwqx/	Water quality, soil and sediment quality, tissue analyses, and population data
Fish Passage Center	www.fpc.org/	Smolt migration (mainstem), upstream fish passage counts, real-time hatchery releases, hydropower releases, etc.
Pacific Fisheries Information Network	http://www.psmfc.org/pacfin/	Commercial fish harvest data
Recreational Fisheries Information Network	http://www.recfin.org/	Marine recreational fisheries data
Regional Mark Processing Center	http://www.rmpc.org/	Coded-wire tag marking and recovery data, marked fish releases, etc.
PIT Tag Information System	http://www.psmfc.org/content/view/47/186/	PIT tag release and recovery data.
Integrated Status and Effectiveness Monitoring Program	http://www.nwfsc.noaa.gov/research/divisions/cbd/mathbio/isemp/index.cfm	Pilot project to assemble fish and habitat data in the Wenatchee, WA, and John Day, OR, subbasins
Interactive Biodiversity Information System	habitat@nwhi.org	Wildlife life history information, terrestrial habitat information.

Appendix A

Suggested Minimum Contents for Metadata for Tabular Data

The following metadata elements represent information that is essential for understanding data and using them appropriately. This is adapted from the ODFW Data Clearinghouse (https://nrimp.dfw.state.or.us/DataClearinghouse/default.aspx?p=1) instructions for submitting data sets (http://nrimp.dfw.state.or.us/DataClearinghouse/DataTemplates/DC_RecordCreation_20June08.doc). These recommended elements are a subset of FGDC Biological Data Profile metadata but exceed the minimum FGDC requirement. They also include several items not specifically included in FGDC, as noted below. While the full suite of FGDC metadata provides the most utility, the basic information is covered here.

It may be appropriate to scale the amount of metadata to the degree of summarization included in any data set. For example, an agency-wide data set built from many data sets obtained from local offices would likely describe the origin of the data in general terms, while each of the original local data sets should have origin and methods explained in specific detail.

Further refinement of minimum metadata needs should be considered as part of establishing a regional level data sharing policy.

Citation Information

Title: = "Name of the dataset."
Originator/owner: = "The name of the organization or individual that developed or owns the dataset."
Pub. Date: = "The date when the data set is published or otherwise made available for release."
URL link: = “the URL link to access the data, or the URL to the project if the data are not available on line”

Contact Information

Contact Person: = "The person responsible for providing access to the data."
Submitting Agency: = “The name of agency responsible for the data.”
Contact Job Position: = "The job position of the person responsible for providing access to the data."
Contact Phone: = "The telephone number by which individuals can speak to the organization or individual."
Contact E-Mail: = "The email address by which individuals can speak to the organization or individual" (This element was not specifically included in the FGDC BDP)

Description

Abstract: = "A brief narrative summary of the dataset."
Purpose: = "A summary of the intentions with which the dataset was developed."

General Information

Project Name: = “The name of the project as used by the funding agency” (This element is not specifically included in the FGDC BDP. It would fit in “Supplemental Information”.)

Funding Entity/Program: = “The entity and program providing funds to collect or create the dataset.” (This element is not specifically included in the FGDC BDP. It would fit in “Supplemental Information.)

Project Number: = “The number assigned to this project by the funding entity.” (This element is not specifically included in the FGDC BDP. It would fit in “Supplemental Information”.)

Time Period: = "The year (and optionally month, or month and day) for which the data is applicable." (This should be broken into Start and End dates to fit FGDC fields)

Geo. Extent: = "General text description of the geographic location covered by the dataset."

Status: “Complete,” “In progress as of this date” or “Planned”

Keywords: = "Generalized keywords to aid in searching for this document."

Intended Usage: = “A description of the intended ultimate use of the data (e.g. management decision, technical publication, peer reviewed journal, etc.)” (This element was not specifically included in the FGDC BDP. This might be considered the same as the “Purpose” field)

Usage Caveats: = "Restrictions and legal prerequisites for using the dataset after access is granted."

Format: = “The native dataset format.”

Data Quality Information

Lineage-Source: = “A general description of the dataset source(s) and processing steps in its development.”

Appendix B

Outline for a Data Management Plan

A data management plan can be very helpful in assuring that all people involved in creating or using a data set understand how the data will be managed. This will avoid misunderstanding, especially among participants in a data collection program or between a funding entity and a project sponsor. It also assures that all steps in the process of collecting, storing, analyzing, using and sharing the data are thought through before field work begins, and assures that critical steps are not overlooked, especially the final disposition of the data so that data are not lost over time. Emphasis should be on the specific actions planned for handling the data once they are created. Details of the project, its purpose, sampling protocols used, etc. should be included in the metadata, and need only be referenced briefly in the data plan.

A data management plan should address the following items.

I. Project Description
   a. Title
   b. General description[1]

II. Contacts
   a. Project Leader
   b. Person responsible for collecting the data in the field
   c. Person responsible for entering the data
   d. Person responsible for managing (maintaining, changing, updating, correcting) the data after collection and entry

III. Data
   a. General Description
   b. Collection methods. Identify the manuals, standards or protocols being followed for data collection. If no formal protocol is followed, provide general description of method.[1]
   c. Data capture. Provide copy of field forms, or describe electronic tools. Are all needed data elements included in the forms?
   d. What standards are being followed for data management (standard coding schemes, formats, etc.)?
   e. Data dictionary (include data definitions, coding, units, and whether optional or required)
   f. QA process / procedures to be employed
   g. Data storage process and format (including data backup procedures)
   h. Where data will be stored (locally, and other databases) and versioned
   i. Data “ownership” or control (describe)
   j. Data analysis (how the data will be summarized or analyzed. If detailed analysis will be performed, write a separate data analysis plan)
   k. Access to data (who, how, describe any restrictions or limitations)
   l. Sensitive data (how this will be handled)
   m. Long term data storage and dissemination

IV. Schedules
   a. Description of data pathway and operations
   b. Schedule for each node in the data flow (a flow diagram may be helpful)
   c. Methods for tracking data status
   d. How and when data will be made available to others (schedule, rights of use, etc.)

V. Metadata
   a. Provide metadata or link to it, if available at project initiation, or
   b. Describe who will develop metadata, and when and where it will be available

[1] Details should be included in the metadata. Only a general description is needed here.

Appendix C

Summary Checklists

A. Recommendations for **field agencies**

1. *Standardize sampling to the degree possible*
   Increases consistency of data, eases compiling data from different offices into an agency or wider database. Especially important within an agency. Standardize within each specific kind of sampling to the degree possible
2. *Follow existing data management guidelines/standards*
   Use existing 'best practices' documents to guide how data are recorded to maximize consistency in recording and managing data
3. *Automate data capture and management, to the degree possible*
   Electronic tools can greatly speed data entry, improve accuracy, improve data flow and utility in the agency, and enable data sharing. These can include data capture devices, automated data validation, agency wide data systems, canned reports, etc.
4. *Use common coding and formatting and describe in a data dictionary.*
   Ideally, all offices and projects in an agency should use the same code lists and formats. Work toward adopting (or crosswalk to) regional or national code systems. Describe data element definitions, coding, units of measure, formats, and acceptable data ranges in a data dictionary.
5. *Describe all data sets with metadata*
   All data sets should include explanatory information so they can be understood and used appropriately by others. Metadata should meet or exceed minimum national standards. Initial metadata development can take time, but subsequent sets of metadata are easier and can mostly be created by cut and paste.
6. *Publish metadata*
   For data sets that will be shared, publish metadata as a web service in XML, allowing the metadata to be located through online clearinghouses and portals.
7. *Implement quality controls*
   A quality control process should be implemented and followed at all steps in data creation, management and use. System automation can simplify QA/QC.
8. *Develop a data management plan*
   Preparing a data plan prior to initiating sampling will ensure that data are entered, stored and used appropriately. This will avoid oversights and lost data.
9. *Develop a data analysis plan*
   If data will be analyzed, a description of the analysis approach should be written. For simple analyses or summarization, this can be included in the data management plan.
10. *Select an approach for sharing data*
    For data sets that will be shared outside the agency, determine the approach, either posting directly to the Internet, or posting through an intermediary database project.
11. *Establish data sharing and use policies*
    Agency policy makers should address issues such as priorities of data to share, subsequent use of agency data, timeliness of data release, treatment of sensitive data, and other policy level issues. Participation in collaborative efforts to develop regional scale data policies is recommended.

B. Recommendations for **field samplers**

1. *Standardize sampling to the degree possible*
   Follow agency direction for sampling. Work toward adopting standardized field sampling methods to the greatest degree possible. Document methods used and record any adjustments or deviations from standard protocols.
2. *Follow existing data management guidelines/standards*
   Use existing 'best practices' documents to guide how data are recorded to maximize consistency in recording and managing data
3. *Automate data capture and management, to the degree possible*
   Enter field data into electronic format during sampling or immediately after. Archive a copy in a safe place. Use electronic data capture devices in the field, as available.
4. *Use common coding and formatting and describe in a data dictionary.*
   Follow agency direction. Use standardized code lists and formats to the degree possible. Do not create new coding systems. Adhere to agency data dictionary, or develop a dictionary to describe data definitions, codes, units of measure, formats.
5. *Describe all data sets with metadata*
   Include a set of metadata with each data set to describe the data. Use agency adopted format, or at least the items listed in this guide. Cut and paste from other metadata as much as possible to decrease workload, with only the differences newly described.
6. *Publish metadata*
   Follow agency process for publishing metadata.
7. *Implement quality controls*
   Follow established quality control procedures when entering and correcting data. Use automated data systems with built in quality checks (e.g., range checks, required formats, drop down lists, etc.) if possible. System automation can simplify QA/QC.
8. *Develop a data management plan*
   Preparing a data plan prior to initiating sampling will ensure that data are entered, stored and used appropriately. This will avoid oversights and lost data.
9. *Develop a data analysis plan*
   If data will be analyzed, a description of the analysis approach should be written. For simple analyses or summarization, this can be included in the data management plan.
10. *Select an approach for sharing data*
    Follow agency procedures for sharing data. Submit data sets to appropriate systems or entities.

C. Recommendations for **funding entities**

1. *Standardize sampling to the degree possible*
   Funder should recognize agency/sponsor selection of sampling methodology or negotiate methodology unless there is specific need for a particular protocol. Specific needs can be negotiated and stipulated in contract language.
2. *Follow existing data management guidelines/standards*
   During contract negotiations encourage project sponsor to follow existing data guidelines.
3. *Automate data capture and management, to the degree possible*
   Support project sponsors to able use of appropriate data capture and management tools, including field data capture tools and development of consolidated database systems. Specify if data are to be entered into a specific database.
4. *Use common coding and formatting and describe in a data dictionary.*
   During contract negotiations encourage project sponsors to use standardized data definitions and coding, if such exist for the data being collected. Require a data dictionary that provides data definitions, codes, units of measure, formats, etc. with the data.
5. *Describe all data sets with metadata*
   Request or require that metadata be provided with any data sets produced by a project.
6. *Publish metadata*
   Negotiate with the project sponsor to specify how metadata is to be disseminated.
7. *Implement quality controls*
   Funder should discuss quality control process with sponsor, and specify any required steps in the contract.
8. *Develop a data management plan*
   Funder should require preparation of a data management plan as part of project proposal.
9. *Develop a data analysis plan*
   Funder should require preparation of a data analysis plan as part of project proposal if sufficient detail will not be part of the data management plan.
10. *Select an approach for sharing data*
    Funder and sponsor should agree on how data will be shared and specify in data plan. Require that data be provided in a machine readable format, not in .pdf or summary reports.
11. *Establish data sharing and use policies*
    Funding entities should participate in collaborative efforts to establish regional scale data policies and priorities.

D. Recommendations for **executive policy makers**

1. *Standardize sampling to the degree possible*
   Within agency, establish policy regarding which standardized sampling protocols will be required, if any. Within region, work collaboratively toward regional scale policy recommendations regarding desired sampling protocols.
2. *Follow existing data management guidelines/standards*
   Within agency, establish policy on which guidelines are required.
3. *Automate data capture and management, to the degree possible*
   Within agency, determine approach to building data systems and deploying field data entry tools. Within region, collaboratively determine if support is needed to develop regional scale data dissemination capabilities in agencies.
4. *Use common coding and formatting and describe in a data dictionary.*
   Within agency, establish policy on which coding systems and data dictionaries are to be used. Within region, collaboratively support broad consolidation of data dictionaries.
5. *Describe all data sets with metadata*
   Within agency, develop policy on use of metadata.
6. *Publish metadata*
   Within agency, establish policy on publication of metadata.
7. *Implement quality controls*
   Within agency, establish data quality control policies.
8. *Develop a data management plan*
   Within agencies and regionally support need for data management plans.
9. *Develop a data analysis plan*
   Within agencies and regionally support need for data analysis plans.
10. *Select an approach for sharing data*
    Within agency, determine best policy approach toward sharing data, including which data should be shared and how. Within region, collaboratively determine policies related to desired approaches for sharing data.
11. *Establish data sharing and use policies*
    Within agency, determine policy on issues such as any need for data sharing agreements, timeliness of data release, sensitive data, etc. Within region, use collaborative approach to develop regional scale data sharing policies and programs. Explore regional scale data sharing strategies and goals (functions, not technical IT approaches) and needed functions for regional scale dissemination of data.

E. Recommendations for **database management projects**

1. *Standardize sampling to the degree possible*
   Provide assistance and advice if field sampling is relevant to the specific data program (e.g., PIT or CWT tagging). Otherwise, database projects have no role in field sampling procedures.
2. *Follow existing data management guidelines/standards*
   Provide advice or assistance, if requested.
3. *Automate data capture and management, to the degree possible*
   Provide technical assistance to agencies with developing internal data management systems and templates for field data entry, as requested and as possible within project scope and capabilities. Actions may include providing technical advice or actual development of data entry templates, local or agency database systems, data translation tools, and/or data output interfaces.
4. *Use common coding and formatting and describe in a data dictionary.*
   Provide technical assistance with data dictionary development, as requested.
5. *Describe all data sets with metadata*
   Provide advice and technical assistance with development of metadata, as requested and within scope and capability of database project.
6. *Publish metadata*
   Provide advice and technical assistance, as requested. Some database projects may be able to publish metadata for partner agencies.
7. *Implement quality controls*
   Provide advice and technical assistance, as requested.
8. *Develop a data management plan*
   Provide technical advice, as requested.
9. *Develop a data analysis plan*
   Provide technical advice as requested and if relevant to the database project.
10. *Select an approach for sharing data*
    Provide technical advice, discuss options, as requested. Include data in project database if of relevant type. Assist partner agencies by posting data if requested and within capabilities.
11. *Establish data sharing and use policies*
    Participate in collaborative data policy efforts to discuss technical options, provide expertise and make technical recommendations.