Implications of the data management roadmap *Collaborative Information Management to Support Ongoing Assessments of VSP, Hatchery, and Tributary Habitat Effectiveness for Columbia River Basin Anadromous Salmon*

DRAFT 2010/03/31

The roadmap document outlines the immediate and longer term approaches needed to consolidate data to support assessments under the FCRPS BiOp. The actions described in the document will help to guide data collectors (biologists) and data managers in establishing procedures and infrastructure to accomplish the goals. The recommended approach has a number of implications for how participating agencies and entities will need to adjust their operations, particularly in regard to changes in the way agencies collect, manage, describe and share data, and specific actions that will be asked of them.

Many of the actions described in the roadmap document were also presented in more specific detail in a white paper entitled Considerations for Regional Data Collection, Sharing and Exchange by the StreamNet project (Schmidt, et. al., 2009). Based on both documents, the following actions are likely to be required in accomplishing the assessments under the BiOp. These should be understood up front to avoid unpleasant surprises when the actual work is initiated.

Data Collecting Agencies

Long Term

The roadmap document indicates that the **long term** goal is for agencies (primarily state and tribal) that collect the data that support the assessments to provide the data in a standardized format through an Exchange Network approach, as developed by the US EPA. In order to implement such a long term approach, the following actions and capabilities will be required:

1. Data collection methods will have to be standardized, to the degree possible, across agency lines. This means that agency biologists will need to reach agreement across agency lines on the best overall approach for conducting each priority sampling type, and then modify sampling accordingly. This will likely result in changes to field procedures for many or most agencies going forward. It will eventually lead to the need to translate legacy data so that it can be analyzed over the longer timeframe along with the new data. Note that there will be flexibility depending on specifically what data is required. For example, if the primary data need is a derived metric, there may be different field methods suitable for acquiring the raw data to calculate the required metric.
2. Agencies will need to decide on the scope under which they will provide data. The large management agencies that have responsibilities in more than one Columbia Sub-region must decide whether they wish to provide data on a consistent, agency-wide basis in all sub-regions, or allow separate approaches and database systems to develop in each sub-region. An agency-wide approach would remove the likelihood of divergent data definitions and coding systems among sub-regions, and simplify the overall management of agency generated data. This is both a long term and immediate need.

3. Data should be managed in comprehensive databases for each major type of sampling, either on an agency-wide basis or a sub-region basis (as discussed in #2). This will require development of database management systems and changed procedures at the field level to load data to the agency databases. This will be required because the data will have to be served to the Exchange Network over the Internet in XML format, as defined in an agreed upon Data Exchange Format. While several agencies are already working toward this goal, none have fully completed the task for all of the types of sampling that will be required.
4. Development of the standardized formats for exchanging the priority data will require time from both biologists and data specialists. Agencies will need to commit biologist time to ensure that the standardized formats adequately represent the data that will be used in the assessments. This is also an immediate need.
5. Managing data in comprehensive agency databases by major sampling type will require that all data adhere to agency defined data definitions and codes. Agencies that do not already have such code systems will need to develop them. If such code systems do not already exist, it would be good to adopt already existing regional or national coding systems rather than developing another new system. Or, simply adopt the Data Exchange Format as the basis for the agency coding approach, with the understanding that the DEF standard will likely include only the 'common denominator' elements, and the agencies will wish to include additional information in their agency systems.
6. All data that will be shared through the Exchange Network must be accompanied by descriptive metadata. This will represent new work for field biologists to describe the details of how they sampled and what their data sets contain. As data are consolidated into the agency databases, metadata will need to be prepared for the larger consolidated data sets. And, as data are analyzed and derived data are produced, these will also need to be described with metadata describing the analytical methods used.
7. Each agency will need to decide what specific data it will exchange. A key decision is whether to provide raw data along with derived estimates, or only the derived data. This should be worked out at the Sub-regional Workshops.
8. Post and maintain the data and metadata, in approved Data Exchange Format, as XML on the Internet. There are various ways to accomplish this. An agency-wide database system should be capable of doing this. Or, agencies can post data by using an umbrella organization (e.g., CRITFC) or a regional database project (e.g., StreamNet).

Immediate / short term

Since none of the agencies have completed agency wide database systems for all of the required types of sampling data, there will be a need for intermediate approaches to support the initial assessments, and it will be critical for the initial reporting timeframe to be clearly stated. These approaches may vary by agency, but the following components may be helpful:

1. Since current data are based on existing sampling methods that are not standardized across agencies, biologists and data managers should work together to determine what data can be appropriately shared and combined. It may be necessary to limit the data exchanged to derived

estimates, which can be based on different sampling methods. Analysts may need to evaluate each approach, however, to determine whether the statistical accuracy is adequate for the task.

2. The scope of data exchange should be approached as for the long term. If DEFs or other standards are developed on a sub-regional basis, they will diverge, causing long term problems for agencies that operate in multiple sub-regions. An agency-wide approach would provide for internal consistency in agency data.
3. Since there initially may be few comprehensive databases, it may be necessary in the short term to accumulate the data directly from individual field offices. This will require efforts both by the individual biologists and data technicians/stewards working with them. The data technicians or stewards will have to develop databases to accumulate the data. Existing database projects should be able to handle this effort with small additions of staffing to handle the additional data types, and already have database expertise and infrastructure that can be applied to the task.
4. Development of the standardized formats for exchanging the priority data will require time from both biologists and data specialists. Agencies will need to commit biologist time to ensure that the standardized formats adequately represent the data that will be used in the assessments. This task is the same for both the immediate and long term.
5. The data technicians / stewards will need to work with the biologists to develop crosswalks between existing data definitions and coding over to the required Data Exchange Formats.
6. Metadata will be required in the short term, but since little metadata currently exists, each agency will need to work with regional interests to determine the minimum amount of descriptive information needed for the initial assessments. Initial efforts at metadata creation should probably begin with the derived data, and over time be developed for the raw data as well. In the future, the metadata from the raw data would be used to create the metadata for the derived data.
7. Each agency will have to decide what specific data it will exchange. A key decision is whether to provide raw data along with derived estimates, or only the derived data. This is the same for both immediate and long term.
8. Data and metadata will need to be conveyed to the analysts doing the assessments. Some of this may take the form of pilot exchange network approaches, where sufficient infrastructure and capability is in place. Or, existing regional database projects can serve the data through existing query systems. In other cases in may be necessary for the data technicians/stewards to load the data directly into spreadsheets appropriate to each analysis. In that case, the assembled data should also be archived so that it remains available into the future. The StreamNet Data Store can accomplish this task. The data should also be loaded into agency-wide database systems as they are developed.

Funding entities

Since this assessment effort represents new data work, there will be both immediate and long term needs for additional funding above normal agency operating budgets.

<u>Long term</u>

In order for an Exchange Network approach to function, the agencies will need to develop internal comprehensive agency-wide database systems (by type of sampling) to manage all of the agency's data and provide the web services to serve the data to the network. While the agencies may wish to develop such systems for internal purposes, the fact that these systems are required for regional scale data sharing suggests that some additional funding support may be appropriate. Funding needs will include infrastructure (hardware and software), staff time for system development, and may also require a small amount of ongoing staff time to manage the data and systems. Funders and agencies will need to discuss how to approach this funding need.

Immediate / short term

Since only a few agency database systems may be available for the initial assessments, there will need to be an alternative approach for accumulating and managing the data for the initial assessments. This will require several kinds of data management expertise to acquire and manage the data, development of some minimal database systems to support the assessments, and staff time to work on developing the Data Exchange Standards. These needs and strategies for meeting them should be a primary discussion at the Sub-regional Workshops. Much of this expertise already exists in regional database projects, with small amounts of additional data technician / steward time needed for the expanded number of data types included.

Policy Guidance

There are several kinds of policy guidance that will be needed.

An immediate need is for priority guidance on the specific data types needed for the assessments. This guidance will need to come from senior program managers and scientific analysts and will need to be a primary outcome of the Sub-regional workshops. The priorities will need to be very specific, down to individual data components so that the data specialists can focus their work and not be distracted by all of the other potential data that may be available.

Key decisions will be needed on how to protect the data originators in terms of right to first use of their data and to avoid improper use of the data. These, and other topics, will be incorporated into a data sharing agreement required for use of the Exchange Network. The specific components in the data sharing agreement will need to be approved at the executive level to establish the policies.

There will need to be policy level commitments from the participating agencies and entities to assure that the effort will proceed as planned and meet its objectives.

Summary

These implications are not major roadblocks, but need to be understood up front so that the partners are able to be fully ready to proceed when the effort begins. These implications need to be considered in order to estimate the amount of effort that will be required and to estimate what additional support may be needed. The above are general ideas intended to stimulate discussion and consideration of how each participant wishes to address them.