Water Information and Data Subcommittee
Workgroup 4: Data Exchange Methodologies

Data Exchange Issues and Proposed Approaches

(Last updated 3/7/2013)

1.0 Introduction

The Water Data Exchange (WaDE) is a project initiated by the Western States Water Council (WSWC), in coordination with the Western Governors' Association (WGA), the U.S. Department of Energy National Labs, and the Western States Federal Agency Support Team (WestFAST). The purpose of the project is to better enable the exchange of water availability, water planning, and water use data between the states, with federal agencies, and with the public. The project will also include working with federal agencies to make federal data more available to state water planners, which would assist in the development of their water plans.

The Water Information and Data Subcommittee (Subcommittee) is a subcommittee formed under the direction of the Water Resources Committee of the WSWC. The Subcommittee has representatives from the WSWC states as well as ex-officio (non-voting) members from the federal agencies. At the October 2011 WSWC meeting in Idaho Falls, the Water Resources Committee agreed to a proposal by the Subcommittee to begin collaborating on the WaDE project. They also agreed to form four workgroups under the direction of the Subcommittee. These four workgroups would lay the groundwork for the development of the data exchange. Workgroup 4 (Data Exchange Methodologies Workgroup) has been tasked with evaluating the current technologies available for enabling web-services-based data exchanges. This workgroup will also make recommendations to the Subcommittee on a proposed approach, and work through the Subcommittee to help the states implement the exchange. As part of that task, the workgroup has compiled this document, which lists issues that will need to be addressed in order to fully implement the exchange. Issues described in this document are organized in the following categories:

1. System Design and Deployment: This category explores the issues surrounding system design and implementation (i.e., the software and database platform, security, potential design approaches, implementation, and long-term maintenance).
2. Governance: This category describes issues related to how state agencies and federal partners will work together to establish the WaDE project and its derivatives, how to implement changes as needed, and how to incorporate new partners.
3. Participation: This category will explore the issues surrounding state and federal involvement in the data exchange, as well as identifying incentives for participation. The degree to which data are made available to the public is also explored in this category.
4. Services: This exchange will rely upon a set of agreed-upon web services. How those services function will depend heavily on the system design and security constraints. The workgroup will discuss the core capability of the services, their expected performance and response time, and methods for managing the large volume of data coming from various sources.

5. Funding: The last category will focus on issues related to long-term funding for this effort. The initial design and development costs are being paid for, in part, from a Department of Energy grant through the WGA. Funding for system maintenance, adding new partners, and responding to new requirements will need to be discussed. The workgroup should also evaluate the potential cost savings for implementing such an exchange.

The purpose of this document is to identify the process by which the workgroup develops its recommendations for the Subcommittee. All issues identified in this document will have one or more proposed solutions. For each of those solutions, the pros and cons will also be identified. Finally, the workgroup will make a recommendation for addressing each issue.

2.0 Background

This section provides background information on the project to assist the workgroup in coming to a common understanding of the goals and core design of the data exchange.

2.1 Data Exchange Description

WaDE will provide a state-to-state, state-to-fed, and fed-to-state data exchange. The data exchanged will better support water planners in developing and implementing their water plans. The design of the system is such that data should flow seamlessly in a computer-to-computer fashion. Participants in the data exchange should not need to wait for a manual step in order to receive the data that is being requested. Data owners should also be able to control what data are shared. Participants benefit by being able to access data from other partners using this seamless approach. Participants also benefit by having their data published quickly and easily in a common format via web service and also using centralized data portal, thereby facilitating data access to the public and other interested parties.

Figure 1 is a diagram of the proposed system architecture. Participants in the data exchange would include state and federal partners. Each partner would agree to implement a core set of capabilities. Each partner would also have access to the data available from other data partners. Any resources (i.e. models, software code, etc.) developed under this project would also be available to any of the data partners free of charge. The model for the data exchange design will consist of a number of key components:

1. Each partner will run a data exchange 'Node.' A Node is a computer connected to the internet with the ability to respond to requests coming from outside the agency's firewall. All nodes should respond in the same manner, when asked the same question. This allows for one system to query all nodes without having to build special interfaces for each node. Nodes should also manage security and access to the data. Node software need not be identical as long as the responses are similar.

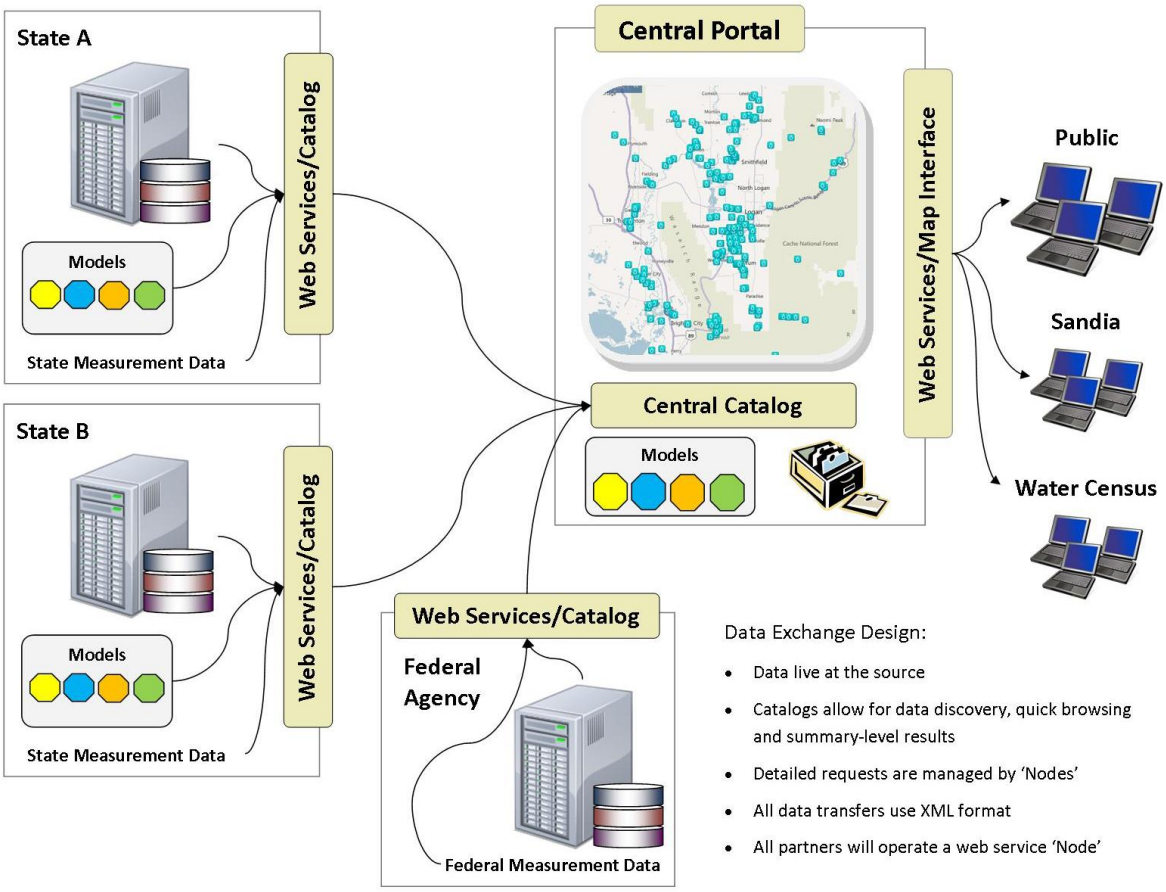

Figure 1. Conceptual System Design

2. Each partner will maintain a queryable 'Catalog' on top of their data. The Catalog provides an inventory of the data available in a partner's data system. Catalogs are meant to be summary level information and should be organized by an agreed upon method (preferably geographic). Catalogs allow other partners to be able to discover data without having to search through entire data sets. Once a user has narrowed in on the data that they want via the catalog, they can then use the catalog to retrieve it. Catalogs can also be a convenient way to indicate whether a particular data set is public or private, or whether it should be made available to other partners only.
3. Each node will support a core set of 'Services.' These Services will be defined by the workgroup. Examples of these services could include examples like 'Authenticate a User',

'Provide Synchronous Access to the Data', 'Provide Asynchronous Access to the Data', 'Receive Data From Another Source', 'Check the Status on a Request', or 'Provide Catalog Information.' The workgroup will need to define specific services that would fall under these categories, along with the input parameters and the returned format.

4. The system, where appropriate, will incorporate models to provide additional data products. Model incorporation into the system will be evaluated by the workgroup. Under an ideal scenario, models could be developed to directly incorporate data provided via this system. Model output would then be available to other partners.
5. A Central Portal with a corresponding Central Catalog will be developed that will have the ability to consume the summary information available in partner catalogs, display that information on a queryable map interface, compile a user's request for specific data, request the data from the respective nodes, and return a report or dataset to the user. The portal is a key piece in demonstrating the value of the data flow. The central catalog would allow the portal to perform quickly enough for ease-of-use by the client.

2.2 Types of Services

In deciding how the services will be developed, the workgroup should consider two types of services that are currently in use: 1) SOAP Services, and 2) RESTful Services. For the sake of automated transfer of data in as seamless an approach as possible, the workgroup should avoid such data transfer technologies as: FTP, Delimited Flat File, Web Scraping, Database Backups, or other proprietary formats. If RESTful services are used, then the generated URL should be discoverable by the user. This allows for the user to retrieve the same data again by saving the URL. An example of this can be found at http://www.waterqualitydata.us/.

SOAP Services: Simple Object Access Protocol (SOAP) is a simple XML-based protocol to let applications exchange information over HTTP. Responses are contained in an envelope that describes the response as a SOAP response. SOAP services are usually accompanied by a Web Service Definition Language (WSDL) description of the service which allows other applications to bind to the service and acquire data.

RESTful Services: Representational State Transfer Services (REST) also use HTTP to exchange information, but rather than encoding the response and the request via a WSDL, the requests are encoded using a URL that has been defined by the service. Any application that can pass the URL to the web server will receive an XML or even a JavaScript Open Notation (JSON) response. It is important that the REST URLs are visible to the requesting client. Access to the URL allows the user to retrieve the same or similar sets of data that were previously retrieved on a repeat basis. Optimally, users can be given a URL generator webpage where desired parameters are entered and the necessary URL is returned. The USGS has a good example of this functionality on their website, found at http://waterservices.usgs.gov/rest/DV-Test-Tool.html.

2.3 Categories of Data

There are two categories of data that will be evaluated under this project. Each of these categories may require slightly different solutions for the implementation of a data exchange. The first category of data will be termed 'derived' or 'value added' data. These data can be described as data that are not direct measurements, but rather reflect interpretations, evaluations, or decisions based on other datasets. Examples of this type of data include water appropriations, consumptive use estimates, or water planning data. Each of these data types is derived by evaluating other datasets to arrive at a conclusion. An important piece of information that must be captured for this data is the methodology used by the agency to arrive at their conclusion so that others can understand the context of that data.

The second category of data will be termed 'measured' data, and for the purpose of this project, these data will most often be 'time-series' in nature (i.e. a set of measurements taken over a period of time). Although some of these measurements may still be derived from a number of measured values, they can still be quantified as a specific measurement at a specific time that can be evaluated along with other measurements over that same time period. Examples of data that fit within this category include stream gaging data, snow depth data, reservoir height data, precipitation data, etc.

Although there are some similarities between these two categories of data, there are enough differences that different data exchange approaches should be considered. For example, there may be specific security concerns around the derived data that wouldn't apply for the measured data. Measured data may also be significantly more voluminous and not have the same metadata requirements as the derived data. Considering the different aspects of these two categories of data, the workgroup should evaluate which approach works best for each, and not necessarily be constrained with coming up with one solution that meets the needs of both types of data.

3.0 Issues

To assist in the process of making a recommendation to the Subcommittee, the following issues have been identified and evaluated. Through discussion within the workgroup, possible solutions for each of these issues have also been identified, along with the pros and cons for each solution. Final recommendations have also been made. As new events arise, the Subcommittee should be able to make adjustments to these recommendations as necessary. These recommendations constitute the best professional judgment of the members of the workgroup, and should be taken as such.

3.1 System Design and Development Issues

1. **Platform/Software**: A key design decision will be what development software the various components will be written in. Two key factors should be considered in this decision: 1) What development software can the states support, and 2) If the goal of this project is to turn

over these components to other federal agencies, what development software can they support. Some preliminary discussion within the workgroup leaned toward making sure that whatever software is used should be platform independent. This requirement would lean to a Java development environment as opposed to a .NET environment. However, research would still need to be done on whether all the states can support a Java environment. There is a possibility that the services deployed at the states will need to be developed in both Java and .Net. This decision needs to be made for the following components: 1) Node, 2) Services, and 3) Central Catalog. However, since the states will not be running the Central Catalog, the decision on its software platform need not be tied to what the states can support.

Possible Solutions:

1a. Develop using a Java Environment

Pros: Cross-platform, might be more in line with some states and USGS's computing environment.

Cons: Many states may not have the in-house experience to work with Java. There are not as many scientific libraries as there are with other platforms (i.e. Python).

1b. Develop using a .Net Environment

Pros: Broader adoption within the state IT communities.

Cons: Not cross-platform.

1c. Develop using both

Pros: We could get broader adoption by the states.

Cons: It would be very difficult to maintain two sets of code.

1d. Develop services using Python or PHP

Pros: Cross platform, simpler scripting languages

Cons: Not as widely used,

Recommendation: After some discussion, it was decided that the programming language should be open-source, and that the option of developing in multiple platforms would not be sustainable. The specific language that the WSWC will use will depend upon various capabilities of each, with the goal of maintaining the code as open source. The WSWC is currently evaluating Hypertext Pre-Processor (PHP) scripts that sit on top of development databases as the driver of RESTful web services.

2. **Platform/Database**: A database will need to be deployed alongside the node and services at each of the states. There is no one consistent database platform (i.e. Oracle, SQL Server) that is used across all of the western states. The enterprise level database platforms may also be cost prohibitive for the WSWC to support on their servers. A multi-database platform

approach may be necessary. As such, the database platform specific capabilities should only be used in limited cases. This will allow the basic components to be transferable from one database platform to another. Alternatively, the group could recommend identifying a core group of states that are consistent in their database platforms and use them to pilot the system, allowing for further expansion once the initial proof-of-concept has been completed.

Possible Solutions:

2a. Develop using Oracle

Pros: Very robust. Standard database for several states.

Cons: Expensive for the WSWC to maintain. Not all states have Oracle databases.

2b. Develop using SQL Server

Pros: Very robust. Standard database for several states.

Cons: Expensive for the WSWC to maintain. Not all states have Oracle databases.

2c. Develop Open Source

Pros: Lower cost, and in line with goal of keeping product as an open-source product.

Cons: No states have currently implemented open-source databases.

2d. Take a multi-platform approach

Pros: Meet the needs of many states.

Cons: May be difficult to maintain several different versions of the database.

2e. Target like-platform states to include in 'proof-of-concept' project

Pros: Allows the WSWC to achieve success with minimal up-front investment.

Cons: May build to capability that is not repeatable in all states.

Recommendation: After discussion, the workgroup recommends that initial development be in open-source. This allows the WSWC to have a very robust database for minimal cost during the development of the database and database models. For deployment at the states, databases are similar enough that additional platforms could be developed with minimal effort from WSWC's open-source version. However, the original database setup should be done with this in mind (saved scripts, functions, etc.). Platform-specific database capabilities should not be used. The WSWC has chosen PostgreSQL as the development database. The WSWC has the capability to develop the database in either Oracle or SQL Server if necessary to meet specific state needs, however, the original design and development will be done in PostgreSQL.

3. **Security**: The data exchanged via this project may have varying levels of security. Some of the data will require some level of accrediting in order to access the data. Other data sets may be fully open to the public. Other security concerns include being resistant to denial-of-service attacks, maintaining the integrity of the data, properly documenting the source of the data, and ensuring that the data are accessible to the right people. If a security model is to be implemented that requires users to provide credentials, then some central system would need to be developed to manage those users. Within the EPA's Exchange Network, this existing capability already exists in the form of the Network Authentication and Authorization Services (NAAS). This central service maintains the list of allowed users on the network along with information on which services that they can access. If the group decides that a similar security model would be needed for WaDE, then they should consider making use of the NAAS to implement a security model.

   Possible Solutions:

   3a. All services are open

   Pros: No security model is needed.

   Cons: Can't restrict access to specific data other than not making it available at the state-node level.

   3b. Some services are open, others require credentials. The services that require credentials would need to be identified. There is also the possibility that the same services could be implemented in both ways, but for the credentialed client a fuller data set would be returned.

   Pros: Greater security flexibility if needed.

   Cons: More expensive to implement and maintain. The WSWC would have to maintain a list of user credentials.

   3c. All services require credentials

   Pros: Allows for only specific users to access the data.

   Cons: Doesn't allow for public access. This would not meet the requirements of the project.

   Recommendation: After discussion, the workgroup decided to go with option 3a: All services are open. The workgroup felt that none of the data that would be shared under this project are sensitive. Data that are sensitive (i.e. drinking water intake locations, and in some states industrial water supply use) would not be shared.

4. **Implementation:** The envisioned design (see section 2.1 above) would have some of the components deployed at the states, and other components deployed at WSWC. All components would be developed and tested in the WSWC environment. Once testing is complete, the Node, database, and service layer would be deployed at the states. The WSWC would assist the states in implementing the components. For the Node, the group should

consider leveraging existing capabilities that exist at many of the state environmental quality agencies. All of the states have developed Nodes under EPA's Exchange Network Program. Although these nodes are intended for state-to-EPA transactions, they have been proven to work for other implementations, including state-to-state and even USGS-to-state interactions.

Possible Solutions:

4a. Use Exchange Network Nodes for services interactions

Pros: Can use existing infrastructure and capabilities.

Cons: Not all of the capability is needed (see discussion on security). The Exchange Network is based upon secure transactions. Since the proposed services would all be open, without authentication, the Exchange Network approach wouldn't work. The agencies that have the Nodes are not the same as the agencies with the data that we're interested in for this project.

4b. Use Exchange Network Node as a starting point, but develop our own capability

Pros: Make the best use of work that's already been done.

Cons: The concept of a 'Node' as defined by the Exchange Network may not be necessary. What really needs to be developed is an endpoint that responds to service requests. The group has decided that the service requests should all be handled as RESTful Services. The Exchange Network is built upon SOAP.

5c. Develop own node technology

Pros: Allows for the most flexibility to meet our needs.

Cons: It may take more time to develop our own capability. Developed services may not be compatible with other data sharing efforts.

Recommendation: The Exchange Network is in the process of developing a REST specification. The WaDE project should build capability that matches this specification. This allows WaDE to meet the system requirements, while at the same time being compatible with other data sharing efforts.

5. **Maintenance:** Since the components are distributed across the states, we'll need to develop a plan for maintaining the components that takes this distributed nature into account. This project is planned as a pilot demonstration project with the intent that it will feed into follow-on projects being led by the USGS. Over the long-term, it is not clear if the WSWC will continue to be the developer and maintainer for this application. One of the core purposes of this project is to develop standards by which data can be exchanged. All components should

be developed in such a way so that they can be maintained by the states themselves if necessary, so long as they can continue to provide information and services that meet the standards developed under this project.

Products can be licensed as either Open Source GPL (General Public License) which requires that the code stay open-source, including derivatives of the code, or they can be licensed as Open Source BSD (Berkeley Software Distribution) which allows for more commercial use of the code.

Recommendation: Products developed under this project will be open source, and can be incorporated or changed by any state partner so long as the changes are shared back with the open-source community. The group recommends that the WSWC use a GPL license for the WaDE project. The group also recommends that the WSWC use an open source code repository for the code developed under WaDE. One possible repository would be to use GitHUB (https://github.com/), which is free for open source projects.

6. **Governance:** Governance will determine how decisions are made regarding the system. A formal process should be established that clearly defines how the various entities participate in the process, who makes the final decisions, how to deal with conflicts, and how to engage new partners. At the early pilot stages of this project, governance is less critical since our main goal is to demonstrate the capability. As more partners come on board, however, governance becomes increasingly important. Approaches will need to be established for change management. For example, any changes to the schema or services would need to be vetted by the entire community. New versions should consider the effect on existing versions. Governance allows for the benefit of changes to be balanced against the cost of implementing those changes across the entire community. If the group decides to follow the Exchange Network approach, then there is a pre-existing governance structure that could be embraced. EPA, in coordination with the Environmental Council of States (ECOS), has established a governance structure that allows for equal participation between the states and the feds on governance bodies that evaluate technology, overall direction, and communication. Members of this community are made up of representatives from state environmental agencies or IT agencies, but the potential exists to open up the representation beyond those groups.

Recommendation: During the pilot phase of this project, the main governance body is the Water Information and Data Subcommittee within the WSWC. WSWC staff report out to the Subcommittee on regular intervals. Issues that require governance decisions will be raised with the Subcommittee. At the conclusion of the pilot, and as more partners enter into data sharing agreements, governance can be expanded beyond the Subcommittee. Products such as the XML Schema and Flow Configuration Document should get Exchange Network approval. This will allow for a broader participation over time, and also allow for the project to get valuable expert opinion on the products being developed. The Exchange Network is developing a REST specification, and the project should follow that specification if it determined by the Subcommittee that it will meet the needs of the project. Federal partners should also be considered for participation in governance once the pilot is concluded. USGS and the Sandia

National Lab should be key players in the governance both throughout the pilot and at the conclusion of the pilot.

3.2 Participation

Recommendation: Participation in the data exchange would be open to any state that wants to deploy the services. States would only share data that would be publicly available without any restrictions. Services will be open to the public without any authentication requirements. All data within the Portal will be labeled as being owned by the organization that provided it.

3.3 Services – PRELIMINARY

The group recommends the following set of services as the 'core' set of services that each state would run. The below service descriptions are preliminary, and are likely to change. These services will be further defined and finalized in the Flow Configuration Document (FCD). The FCD should be considered the final source for the service definitions.

Service Name: GetReportUnits
Service Type: Catalog
Brief Description: Returns the reporting units for a given state or lat/long box. This service would allow a user to map the reporting units, and have an understanding of what data are available in each reporting unit.
Input Parameters:

Parameter Name	Description	Type	Required?
State	Unique Identifier for the state (i.e. UT)	String	No
Lat/long box	Min/max lat/long to describe an area of interest separated by commas.	String	No

Restrictions: A user would need to provide either a state or a lat/long box.
Return: The service would return data in the WaDE Catalog Schema format.

Service Name: GetAllocationCatalog
Service Type: Catalog
Brief Description: Return summary information of site specific data available. Information would be grouped by 8-digit HUC, County, or Reporting Unit. Information could be returned for the entire state, or for individual reporting units.
Input Parameters:

Parameter Name	Description	Type	Required?
State	Unique Identifier for the state (i.e. UT)	String	No
County	Unique Identifier for the county (state FIPS code)	String	No
HUC	Unique Identifier for the 8-digit HUC	String	No

Restrictions: A user would need to provide either a state, county, or HUC

Return: The service would return data in the WaDE Catalog Schema format.

Service Name: GetAllocationDetails
Service Type: Synchronous
Brief Description: Returns the detail information for a given site, selection of sites, or sites within a given geographic area
Input Parameters:

Parameter Name	Description	Type	Required?
County	Unique Identifier for the county (state FIPS code)	String	No
HUC	Unique Identifier for the 8-digit HUC	String	No
ReportingUnit	Unique Identifier for a state Reporting Unit	String	No
AllocationID	Unique Identifier for an allocation. Multiple allocations can be provided by separating with a comma.	String	No
BeneficialUse	Text describing the beneficial use that the user is interested in.	String	No

Restrictions: A user must provide at least one input parameter. Beneficial Use can be used to further restrict the request, but can't be provided as the only input parameter.
Return: The service would return data in the WaDE Schema.

Service Name: GetSummaryDetails
Service Type: Synchronous
Brief Description: Returns the summary data for a given reporting unit or series of reporting units.
Input Parameters:

Parameter Name	Description	Type	Required?
Lat/long box	Min/max lat/long to describe an area of interest separated by commas.	String	No
ReportingUnit	Unique Identifier for a state Reporting Unit	String	No
DataType	String specifying the type of data for which the summary will be made for (i.e. allocation, availability, or consumptive use).	String	No

Restrictions: A user must provide at least one input parameter.
Return: The service would return data in the WaDE Schema.

Versioning: As the data schema is anticipated to change over time, each major release will also contain a version number as part of the base URL.

3.4 Funding

Efforts under this project are currently being funded by a Department of Energy grant to the Western Governors' Association. Although these resources will allow for the development of the pilot project and to demonstrate the capabilities, it will likely not be enough to maintain the

system over the long term. This needs to be considered as the system is designed and requirements are finalized. Some follow-on funding options are available, but the group will need to discuss which of these options to pursue, if any.

1 – USGS Water Census: Under the original plan for the Water Census, the USGS had envisioned a grant program for the states to help them build out their data management capabilities to better support the Water Census. However, Congress has not yet funded the Water Census to a level to support this grant program.

2 – EPA Exchange Network Grants: EPA provides approximately $10 million every year to the states and tribes to support the development of the Exchange Network. These grants are principally used to help the states implement regulatory data exchanges (improving the data flow for those programs where EPA requires the states to provide information to EPA as part of their delegated program). However, they have expressed interest in the past in expanding the network into new and innovative areas. This project may fit that case. Also, there is significant linkage between water quantity and water quality. Improving the sharing of the water quantity data would be a significant benefit to the water quality community.

Recommendation: Both of these options are viable funding opportunities. The Exchange Network funding has more immediate potential since that activity is funded, whereas the USGS grants have not yet been funded. This project appears to be in line with EPA's plans for Phase II of the Exchange Network. Applications are due November 9th, 2012 and the recipients are anticipated to be announced in April of 2013.