SCCWRP #0878

San Diego Integrated Regional Water Management Data Management System Basic Design Recommendations

31 July 2015

Table of Contents

All mention of proprietary systems are used as examples, no endorsement is implied.

Executive Summary

This project, funded by Proposition 84, is derived from the 2013 SD-IRWM's planned goal to develop a regional web-based Data Management System (DMS). The demand for such a system stems from the need for data to be consistent and sharable to meet data requirements and desires of various agencies and organizations in the region. A comprehensive DMS for the region can help improve understanding of the monitoring resource allocations throughout a watershed by providing a complete picture of where data are available throughout the region, for what parameters and at what frequency. This information can facilitate more effective use of existing data while simultaneously supporting better decision-making regarding where, when and what data are most relevant to filling information gaps.

One of the underlying factors driving the interest in a web-based DMS for the SD-IRWM is the lack of a single, consistent source for water monitoring data for the region. Before proposing the development of a new system, the strengths and weaknesses were considered for several existing systems used in the region. Features found desirable in some of these systems include: Consistency of data and quality control, system scalability, use of open source software tools, and a federated DMS architecture.

Two stakeholder workshops were convened to provide input to this recommendation report. The overarching priority for a regional DMS was identified as a system that serves to support watershed health and sustainability. Stakeholders identified three priority functions for the DMS: communication to a range of audiences; provide access to monitoring information; and streamlining the permitting processes. Three criteria identified for prioritizing these tasks were: The system should meet the needs of stakeholders and provide benefits shared by multiple stakeholders in the region; promote interoperability of systems; and build on innovative technology to optimize data gathering, analysis and sharing, and is user-friendly.

The Advisory Workgroup recommends the SD-IRWM develop the desired DMS as a federated data system, using an open source software platform. A federated data system will allow most participating data generators to continue using whatever DMS platform they already have in place and minimize alterations to current their existing systems and workflows while providing a

platform for sharing data with the regional DMS. Data incorporated into this system should follow open data standards evaluated by and overseen by a governing body. The initial focus should be on identifying a limited, common set of data relevant to assessment and support of watershed health and sustainability with emphasis on data collected and utilized by the majority of participating organizations. Where data are already available via an existing DMS, opportunities to harvest data from such systems should be the preferred option. Because historical data generally will require significant effort to prepare for inclusion in a DMS, we recommend against including historical data as an early priority. To build a strong architectural foundation, it is imperative that appropriate accommodations for well-developed, standards-based metadata be included in the DMS design and implementation. Additionally, the DMS should adopt open data standards to ensure interoperability among federated systems.

Development should be parsed into manageable subcomponents based on the time and cost associated with each aspect of the data management system. Hosting of the system should incorporate elastic cloud computing to allow the system to adapt to workload changes as needed. The web-based data system must have an intuitive and easy-to-use interface which make data discovery and download as easy as possible. Data outputs vary widely as their intended uses, some of the core system outputs to be considered include: simple aggregated datasets that consolidate data from multiple organizations into a consistent, comparable format for download or summarized data which is provided in calculated or visual formats representing the status, trends or other characteristics of the watershed. The web interface and data transfer services should be developed using available open-source software.

Finally, a governance structure should be established to define priorities and data requirements for data documentation, QA/QC availability, data sharing, including data privacy and sharing agreements. The governance group would also oversee development of system specifications for development, hosting and maintenance of the DMS. Because any DMS developed will require initial funding to establish and ongoing funding to maintain and support the system, it is essential to plan for these needs prior to initiation of system development. An investment of time and effort early in the planning process is essential to the long term success of the system.

Summary Recommendations:

Recommendations provided in Chapter 2: Tasks for a Regional Water Data Management System

1. Develop a clear understanding of existing data, data gaps and data required to promote watershed health and sustainability.
2. Focus initially on a limited, common set of data which are collected and utilized by the majority of participating organizations.
3. Harvest data from existing systems where available and when possible, use web-services to access data on-the-fly.
4. Do not include historical data which does not conform to system standards as an early priority. This data requires a significant effort to prepare for inclusion in a DMS.
5. Identify the data requirements of users at each stage of the data workflow and develop the DMS to provide, locate and access those data efficiently and effectively.

Recommendations provided in Chapter 3: Design and Structural Recommendations

6. Use a federated data system structure for the SD-IRWM data management system.
7. Subject historical data to the same metadata requirements as all other data in the system.
8. Adopt existing metadata standards to maximize compatibility with existing State and Federal data systems.
9. Adopt open data standards to ensure interoperability among federated systems.
10. Develop a governance structure to oversee data sharing concerns, including data privacy and quality control protocols.
11. Prioritize data types based on: (1) those that address watershed health and sustainability, (2) those that are readily available, and (3) those that are of high quality.

12. Develop simple, informative output tools that address planning and management priorities of the stakeholders. Complex output and analysis capabilities should only be developed after the core data is successfully registered to the global DMS catalog.
13. Develop a mechanism to host data for participants in the DMS who lack internal capacity to do so themselves.
14. Provide access tools for users to discover and retrieve data, and for administrative management of the DMS system using a web interface.
15. Provide an intuitive and easy-to-use interface based on user input and testing to ensure an effective user experience.
16. Use open source software tools and standards as a basis for the federated data structure.
17. Acknowledge, and plan for, integration with existing investments of partners' data systems. Existing investment and software products may already incorporate the required functionality to allow for integration with a federated data management system.
18. Use elastic cloud computing at the host organization to provide for adaptive infrastructure as the system and workload demands necessitate.
19. The DMS should provide end users and data system managers with data status information and provide managers the ability to indicated anticipate time until available.
20. Parse development of the DMS into manageable subcomponents based on the time and cost associated with each aspect of the data management system.

Recommendations provided in Chapter 4: Governance and Database Management Strategy

21. As a first step, develop a DMS governance structure to define priorities and requirements for data formats, QA/QC, documentation, data availability maintenance and funding.

22. Designate a staff position with specific responsibility to conduct system maintenance and updates to the global data catalog.
23. Develop written memorandums of understanding among participating organizations. These should address data and metadata standards and procedures should a member organization fall short of meeting expectations.
24. Plan regular system reviews every 3-5 years to evaluate the effectiveness and future priorities to guide maintenance and development of the DMS.

Chapter 1: Background

1.1: Purpose of the Integrated Regional Water Project

The Department of Water Resources, which governs the San Diego Integrated Regional Water Management (SD-IRWM) Implementation Grant Program, requires inclusion of a data management component for all IRWM plans. This project, funded by Proposition 84, is derived from the 2013 SD-IRWM's planned goal to develop a regional web-based Data Management System (DMS). The demand for such a system stems from the need for data to be consistent and sharable in order to meet data requirements and address the desires of various agencies and organizations in the region. As recommended by stakeholders and the Advisory Work Group, the overarching purpose for the DMS is to advance stakeholder recommendations.

The DMS should provide:

1. Simplified access to existing data sources;
2. Direct access to SD-IRWM-generated data;
3. User-defined interactive access to key data sets;
4. Efficient sharing of data resources; and
5. Effective integration and use of data resources.

1.2: Need for Coordinated Water Monitoring Information

Water data is utilized by multiple organizations and agencies for a variety of purposes. Environmental regulations such as the Clean Water Act require water quality monitoring for a variety of beneficial uses. Monitoring data is essential to determine when a water body is meeting regulatory objectives. If a waterbody does not comply with regulations it may be listed as "impaired," triggering a number of responses such as additional monitoring or remediation actions. Organizations may monitor for other reasons, such as non-regulatory research purposes (e.g. University research or special projects). Water quantity and flow are also commonly monitored as a means to understanding water supply.

The most common questions that arise from monitoring data typically relate to status and trend analyses. Specific questions may include: “What is the current quantity or condition of our water?” and, “Are our waters getting better or worse for a particular measured value?” Depending on the objectives, monitoring data may be gathered from direct measurements, such as quantities of specific chemicals in a water sample. Assessments may also be indirectly obtained using an index, such as biological indices computed from data about the community of organisms in a water body. Water monitoring data is useful in both status and trend analyses.

When data sets are stored and managed by multiple organizations, it becomes increasingly difficult to locate, compile, and standardize data in order to gain a comprehensive picture of a region’s water resources. Without appropriate communication and oversight, multiple organizations could be monitoring the same data parameters at the same locations, while other regions of the watershed are left unmonitored. The SD-IRWM region includes 11 watersheds in San Diego County, with portions of several watersheds extending beyond the county’s borders (Figure 1-1).

Figure 1-1: The San Diego IRWM region includes 11 watersheds which lay fully or partially in San Diego County.

A shared DMS for the SD-IRWM region would provide a single source for access to quality monitoring data. A comprehensive DMS for the region would improve understanding and most effectively use monitoring resource allocations throughout a watershed. Stakeholders involved in this project also suggested this regional effort will be a good model for other regions.

A common data management system will facilitate discoverable and accessible watershed data for the region. The regional DMS provides a shared platform for assessment and decision-making while reducing unnecessary duplication of data collection efforts.

This information can facilitate more effective use of existing data while simultaneously supporting better decision-making regarding where, when and what data are most relevant to filling information gaps. To develop a truly valuable data system, participation from a diverse set of participants, as represented by the SD-IRWM is essential. With widespread collaborative development and participation of a regional DMS, a truly beneficial and compelling resource can be developed and maintained to serve the needs of the region. A shared resource including a common catalog of regional data and easy-to-use interface for data discovery throughout the SD-IRWM, will provide a valuable and comprehensive source for the discovery and access to watershed data of known and documented quality for the SD-IRWM region.

1.3: Stakeholder Engagement

Stakeholders for this process were engaged at multiple levels. The planning team engaged directly with the IRWM's Regional Advisory Committee at several meetings throughout the process to update them on progress. An advisory work group (AWG) consisting of representatives from multiple agencies and organizations met monthly throughout the project to provide key insight and guidance as this recommendations report was developed. A broad call and direct requests for AWG members were conducted in late 2013, and several additional public agencies and organizations were invited to participate in the workshops. Two full-day stakeholder meetings (Figure 1-2) were held to gather input from the wider community, and finally, two

public outreach meetings provided opportunities for comments from the community. Please refer to Appendix A for membership lists of the planning team, advisory workgroup and stakeholder meeting participants.

Figure 1-2: Participants at the first stakeholder meeting provide input to the advisory workgroup and planning team.

1.3.1: Advisory Work Group

As outlined in the SD-IRWM Charter, the project's Advisory Workgroup members were selected based on demonstrated water policy experience and proven water data management expertise. Members were drawn from local public agencies, state government, and non-governmental organizations, university research institutes, and private consulting and engineering firms. The Advisory Workgroup was tasked with the following:

1. Identify major data management efforts, including duplicative efforts and information gaps, involving members' organizations, agencies, and interests.

2. Help connect the project to a broader community of water resource data managers and data users.
3. Recommend priority needs and design parameters for a regional, web-based water data management system.

The Advisory Workgroup provided strategic guidance with regard to policy and technical issues with input from the Stakeholder Group gathered through workshop participation. The Advisory Workgroup also provided recommendations on the final definition of watershed health and sustainability (see section 2.3) and development of the outline and content of this recommendation document.

1.3.2: Stakeholder Workshops

As described above for the Advisory Workgroup, participants in the stakeholder workshops were selected based on their expertise concerning a variety of topics. The role of participants in the two stakeholder workshops was of broader scope than the Advisory Workgroup with the objective of obtaining a wide range of ideas and input to inform the definitions and functional characteristics of an ideal DMS for the SD-IRWM. Because stakeholders expressed desires for a regional DMS were not constrained by realities of time, funding or technical complexity, their input was subsequently prioritized and structured to inform the Advisory Workgroup's discussions and final recommendations to this document.

1.4: Existing Databases and Unmet Needs

There are a substantial variety of local, state-wide and national water data systems in use by agencies and organizations throughout California. One of the underlying factors driving the interest in a web-based DMS for the SD-IRWM is the lack of a single, consistent source for water monitoring data within the region. Organizations use a variety of different, and typically incompatible, data systems or even paper-based systems, developed for their own specific purposes. The development of multiple systems with varied objectives results in a disjointed regional data landscape. This situation is far from unique to the San Diego region. Often data systems are highly varied among the various public, private and non-governmental organizations within a given region.

1.4.1: Identification of Existing Systems

Before proposing the development of a new system, the strengths and weaknesses were considered for several existing data management systems used in the region. These systems were reviewed for two purposes: First, to determine if any of the existing platforms might serve as a basis for the envisioned DMS; and second, to describe some of the features and lessons learned from existing systems that could potentially inform the functional specifications of a DMS specifically designed to meet the needs of the SD-IRWM. While the following list is far from comprehensive, it includes several systems already in use in California that are familiar to many of the stakeholders.

- California Environmental Data Exchange Network (CEDEN)
- Surface Water Ambient Monitoring Program (SWAMP)
- California Integrated Water Quality System (CIWQS)
- Geotracker
- Beachwatch
- California Geoportal
- West Coast Governors Alliance, Ocean Data Portal (WCGA ODP)
- California Data Exchange Center (CDEC)
- Integrated Water Resources Information System (IWRIS)

1.4.2: Summary of existing systems

(Appendix B provides a more detailed summary for each of these systems)

While these systems provide examples of some useful best practices and approaches, no single existing system appears to be an appropriate repository for the data needs of the SD-IRWM in their current or anticipated states. In most cases, these systems, and others developed by individual stakeholder organizations, were designed with little or no consideration for sharing of data between systems. Individually, each of these integrates data from multiple stakeholders, but was developed with a specific program or application in mind. Therefore each lacks the necessary foundation of high quality data infrastructure appropriate to serve as the basis for design of the interactive system envisioned by the SD-IRWM stakeholders. In many cases, these

repositories were focused on accumulating data in a particular format (a valuable feature), but this was done with little consideration of to how others would be able to use/view the data. Nonetheless, these useful examples of a number of the expressed functionalities and additional, valuable features that should be considered when developing final design specifications for a regional DMS.

Examples of desirable features found in these systems include:

- Consistency of data and quality control,
- System scalability,
- Use of open source, and
- Federated DMS architecture.

Another characteristic to note is the administrative or governing structures utilized. While this is not necessarily built in to the technical architecture of these systems, the organizational entity behind a DMS system has pros and cons as well. Several of these systems are governmental. Therefore they have the presumably long-term support that comes with a government agency, as well as the associated relative certainty for financial sustainability. By contrast, systems developed by non-governmental organizations or private entities have an assumed higher likelihood of being "unplugged" during funding gaps and/or changes in staffing, and potentially lack with the needed expertise to keep systems functioning. Ultimately, when reviewing these and other existing systems in light of the design features derived through the stakeholder process, there is not a single existing system that precisely meets all the needs of the SD-IRWM water data users. However, several of these systems provide valuable insight and examples of components that can be included in the proposed federated system. In particular, the WCGA ODP comes close. As an open source system, it provides a platform that can be duplicated and modified to achieve most of the most desired design features without requiring development of a new system from the ground up. A summary of desired capabilities by system is provided in Table 1-1.

Table 1-1: Capability summary of desirable features in databases reviewed.

Database	Consistent data and Quality Control	Scalable System	Opensource Platform	Federated Architecture
CEDEN	X	X		
SWAMP	X	X		
CIWQS	X	X		
Geotracker	X	X		
Geoportal	Metadata only	X	X	X
WCGA ODP	Metadata only	X	X	X
CDEC	X	X		
IWRIS		X		X

Several national level efforts provide potential examples for development of water data system design, including the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) Hydrologic Information System (HIS), USGS Water Data for the Nation, National Water Information System (NWIS) and the Environmental Protection Agency's Water Quality Exchange Network (WQX). While a number of their design features, and system protocols could provide valuable examples for a regional DMS, given their national focus, size and complexity these systems were not reviewed in detail here. Links to these data management systems and standards are provided below for reference.

Web links: CUAHSI homepage: www.cuahsi.org

CUAHSI Hydrologic Information System (HIS): his.cuahsi.org

USGS Water Data for the Nation (NWIS): waterdata.usgs.gov/nwis

EPA WQX: www.epa.gov/storet/wqx

Chapter 2: Tasks for a Regional Water Data Management System

Based on the goals identified in chapter one, two stakeholder workshops were convened to provide input to the AWG in identifying and prioritizing tasks a regional DMS should be designed to accommodate. Additionally, the stakeholders provided input to the AWG regarding a comprehensive definition of watershed health and sustainability for which a regional DMS could help to provide data necessary to achieving that goal.

2.1: Priority Tasks and Criteria for Selection

Project stakeholders identified the tasks for the DMS to serve at the first stakeholder workshop. The criteria for prioritization of those tasks were based on the following four criteria:

1. Meets needs/provides benefits that are shared by multiple stakeholders in the region;
2. Promotes interoperability of systems;
3. Builds on innovative technology to optimize gathering, analysis and sharing ; and
4. User-friendly

2.2: Watershed Health, Sustainability and Priority Tasks

Watershed health and sustainability was agreed upon by Advisory Workgroup members as the overarching priority for the DMS to support. In addition to watershed health and sustainability, the DMS will also work to support these other priority tasks, identified by the stakeholders:

- Monitoring information: Data should be accessible in a manner that allows for multiple means of discovery including question-driven, location-based and thematic queries. Inventorying existing efforts will provide clarity regarding data that exist and where data gaps are present. This information will facilitate more effective and efficient water quality planning, include periodic synthesis of monitoring efforts, analysis for regulatory compliance and elimination of redundancy to better allocate limited resources; and
- Streamlining permitting processes: Permitting is a data-driven process which often reviews monitoring data or requires additional monitoring as a condition of a permit.

A regional DMS would streamline the process by providing agency staff and permittees relevant information about existing data and monitoring programs relating to the location and parameters of interest.

- Communication to a range of audiences: Currently data are not accessible in a manner that is readily available and understandable to a wide range of stakeholders. Many systems are designed for resource managers with domain knowledge as opposed to the general public or elected officials. A well-designed system would effectively communicate to all of these constituencies.

2.2.1: Additional Tasks

The AWG recognizes a number of additional topics for which a regional DMS may provide value. These tasks follow from the above priorities but were not prioritized. To ensure a successful implementation, we recommend limiting the initial DMS design to a limited set of data and objectives and considering additional topics in the future. Additional topics suggested for the DMS included:

- Drought planning and management;
- Climate change planning and management; and
- Flood management.

Although some of the data types necessary to explore these additional topics may be included in the initial development, planning for additional resources to address these additional tasks specifically should be planned for and identified when they are undertaken by the SD-IRWM. In particular the capture of the required spatial and temporal dynamics of the above tasks can be expected to require somewhat more complex DMS design considerations. Data for these topics is important to long term planning and management of water resources in light of extreme weather and climate change. When considering these topics in context of a regional DMS, emphasis should be placed upon the uses of data for planning and management. In the case of flood management, systems are already in place. These current systems function effectively to meet the needs of agency staff who require flood information. Therefore, it will be important to avoid duplication of these existing systems, and instead focus efforts on mechanisms to make

the underlying data contained within them more readily available to the water management community.

2.3: Watershed Health and Sustainability Defined

Through extensive discussion with stakeholder workshop participants, the following statements encompass the shared, and purposely comprehensive vision of a definition for watershed health and sustainability in the San Diego IRWM region. A comprehensive definition of watershed health and sustainability was developed at the request of the stakeholders' because they felt a definition of these concept had not been adequately developed for the San Diego Region.

1. Physical, biological and chemical aspects of health are maintained, and support the full suite of beneficial uses, including wildlife and recreation.
2. Ecological, social, and economic systems are resilient.
3. Water resources, quality, and supplies are managed effectively and efficiently.
4. Groundwater resources, including recharge areas and basins, are protected and managed effectively and efficiently.
5. Water quality requirements are realistic, appropriate, and support the full suite of beneficial uses.
6. The quality of water that is upstream of and drains into reservoirs and other water bodies designated for drinking water supply is protected.
7. Reservoirs that are integrated with municipal water treatment facilities are distinguished from those that release water downstream.
8. Upstream development, land use and zoning, and urban development in the watershed are compatible with natural watershed functions.
9. Stormwater collection systems manage volumes of water delivered to ecosystems during rain events to optimize ecosystem health and comply with permits.
10. Beaches, oceans, riparian areas, and rivers are swimmable and fishable.
11. Native plant and animal species are protected, while non-natives are reduced.

12. Open spaces and parks are preserved and managed for multiple uses.
13. River and reservoir systems, wetlands, coastal areas, and estuaries contribute to civic pride, recreation, food security, and economic activity.
14. Public health and safety are secure.
15. Threats to watershed health are known and addressed.
16. The public is well educated about its role in maintaining healthy watersheds, has appropriate access to data, believes management efforts are effective, and actively participate in stewardship activities.
17. Water resource managers and the public think and act readily with upstream-downstream connections in mind.
18. Water resource managers, regulators, and agencies consistently coordinate their efforts at the watershed and regional scales, and partner where possible.
19. Monitoring is question-driven, baseline conditions and performance metrics exist, and quantitative trigger points/thresholds clearly define when impairment requires action.
20. Best practices that contribute to watershed health and sustainability are reasonable and regularly updated.
21. Financial resources are sufficient to continue activities that promote health and sustainability.

2.4: Priorities for a Regional Data Management System

> ***Recommendation 1:***
>
> *Develop a clear understanding of existing data, data gaps and data required to promote watershed health and sustainability.*

The overarching purpose of a DMS is to advance watershed health and sustainability, as defined above. Therefore, each of the components of the watershed health and sustainability definition outlined above should be analyzed for a clear understanding of existing and needed data to inform the development of the DMS. It is not the purpose of this technical recommendation to assess currently available data or data generators. Rather, the recommendation is to highlight the importance of this first step to ensure that all agreed-to data structures, including associated metadata, are selected based on how each will address components of the watershed health and sustainability definition. Data relevant to the SD-IRWM may fall into one of three categories:

1. Data which one or more organizations collect and which are well-known by others. Such data may already be available through an organizational website or personal contact with the data generator.
2. Data which one or more organizations collect but which are NOT well-known by others. Such data are often generated as part of collections efforts falling outside regulatory or permit processes, and therefore, may not prepared and/or submitted in readily discoverable and shareable formats.
3. Data which are relevant and valuable for watershed assessment or analysis but which are NOT collected by any organization. In these cases, data may not be collected due to a lack of resources or due to a lack of knowledge about the existence of a data gap.

While data falling into the first category above may be readily integrated into a shared DMS, the existence of data in this category does not necessarily make it the highest priority for inclusion. It is not the purpose of this technical recommendation to assess and prioritize currently available data or data generators. Rather, we raise this point is to highlight the importance of

determining the needed data as a first step with an eye to how each will serve to address one or more of the 21 points of the definition of watershed health and sustainability. Additionally, data should be assessed to ensure that all agreed-to data structures (including associated metadata) are met.

> ***Recommendation 2:***
>
> *Focus initially on a limited, common set of data which are collected and utilized by the majority of participating organizations.*

2.4.1: Functional steps prior to developing a regional DMS

Before construction or assimilation of any data system for use by the SD-IRWM, some functional tasks that have been identified and are recommended to be addressed in support of this effort include:

1. Review parallel local, state and federal efforts for overlapping monitoring requirements.
2. Review existing data systems to identify opportunities to leverage data using web services in a federated model rather than duplicating systems or programs which already, effectively capture these data.
3. Review any idle or defunct data systems from which valuable data or lessons learned can be gleaned.

Several stakeholders raised questions regarding the scope of the DMS and which data sets will be included in the system. While it is not the purpose of this report to specify the particular data parameters and analytes to be captured and cataloged, it will be important to assess the value of each data type considered for inclusion. The initial focus should consist of identifying a limited, common set of data which address the objective of watershed health and sustainability and are collected and utilized by the majority of the

> ***Recommendation 3:***
>
> *Harvest data from existing systems where available and when possible, use web-services to access data on-the-fly.*

participating organizations, and particularly those which are not readily available through an existing, online DMS.

When some data is already available, opportunities to harvest data from existing systems via web-services as part of a federated data architecture should be the preferred option. Additional considerations raised by stakeholders relating to scope were: Should the envisioned DMS include agricultural waiver data, and, should the envisioned DMS include land use data (e.g. GIS data)? While there is value in all these data, expanding the scope of the SD-IRWM DMS to directly include these data types is not recommended. Instead, supporting data should be obtained through existing systems or web-services and the SD-IRWM should focus its efforts on the water related data of highest priority to addressing issues of Watershed Health and Sustainability. Every effort should be made to leverage existing systems where feasible, rather than duplicating them, given resources to build and maintain a DMS are finite.

2.4.2: Historical Data

Another theme brought forward by the stakeholders was the value of including historical data. In this context, historical data refers to any data collected before data standards for the DMS have been established. However, some more recent data (perhaps within the last decade or so) may comply with the same, or very similar, standards as those adopted for the DMS, making them easier to integrate. In general, including historical data as an early priority is not recommended unless there is a critical need for those data to answer core questions (e.g. establishing baseline conditions). Historical data often require significant effort to prepare for inclusion in a DMS. The addition of historical data may require key-punching to convert paper documents into a structured data format of the database. Additional effort may be required to identify and document quality analysis/quality control (QA/QC) and collection or laboratory methods establishing the required metadata format for the DMS. Even if historical data are already contained in a digital database, interpreting and

> ***Recommendation 4:***
>
> *Do not include historical data which does not conform to system standards as an early priority. This data requires a significant effort to prepare for inclusion in a DMS.*

translating such data into the required formats may not be worth the effort required. Although the preparation and incorporation of historical data can provide great value when assessing historical or baseline conditions, this should not be as high of a priority as current and future data streams, particularly in the near-term. Inclusion of any historical data set should be carefully considered based on whether the particular historical data set is of sufficient quality to be comparable and analytically valuable in contemporary data analysis.

> ***Recommendation 5:***
>
> *Identify the data requirements of users at each stage of the data workflow and develop the DMS to provide, locate and access those data efficiently and effectively.*

2.4.3: Participants in a Regional Data Management System

In assessing the purpose of data to be captured, significant consideration was given to assessing the needs of the key participants of the DMS. Four types of participants involved at various stages of the data workflow were identified. These DMS participants were broadly categorized by their role in the functionality of the system into three categories:

1. Data Generators;
2. Data Users;
3. Data Consumers.

One individual or organization may serve in any one or several of these three roles simultaneously as a multi-function stakeholder in the process. Detailed descriptions of these roles are provided on the following page. A core emphasis identified by stakeholders is the need for the DMS design specification to be directly tied to requirements identified by data generators, data users, and public audiences while recognizing that each audience does not have the same needs and responsibilities. A well-designed system must provide mechanisms to effectively support the necessary and varied functionality required by each of these sub-groups of end users. If stakeholders in the DMS do not see a direct benefit to their participation with the system including its ability to effectively serve their specific needs, they will not use it or support the effort to develop and maintain the system.

Data Generators: These individuals or organizations are responsible for the collection of data in the field or laboratory. Their role includes inputting data in the proper format (often within their role as the sampling agency or organization). They also prepare data for the DMS including documentation of the associated metadata (QA/QC, methods, etc.). Stakeholder participants made a point of the need to keep the database easy for data to be submitted and to remain connected with their respective data sets. An additional, special case of generator is the ***Data Intermediary.*** These individuals or organizations may include those who maintain data registries, data brokering services, data indexing services, value-added data processors and others who provide digital data in relevant formats.

Data Users: These individuals or organizations are interested in the application of the data stored/retrieved from the DMS. The stakeholder groups also made note that ease of extraction and specificity in a data request be specially noted in these design recommendations. These participants require data that is reliable and well-documented, including information to allow for integration of data from multiple data generators. Data should be in a consistent and comparable format and easily ingested into the desired analysis tools such as statistical software, environmental models, or spatial analysis in GIS.

Data Consumers: These individuals make up the broadest group of participants. Data consumers represent interests ranging from the general public, resource managers or elected officials. This audience typically requires data in processed/interpreted formats communicated for a specific purpose, such as the implementation of data usage for program designs, regulatory compliance or decision-making. While this groups needs should be considered in developing a DMS, they should not be a primary emphasis. Data applications are specific to each consumer's needs and are best developed as external applications that draw data provided by the DMS as opposed to being an inherent part of the design. Data requirements for such applications will benefit from the adoption of open data standards that support developing custom applications in the future.

Multi-function stakeholders: In many cases, the same individual or organization will fall into multiple categories. For example, an agency may be a data generator for a portion of a watershed while also being a user of data contributed by other entities operating in the same watershed. They may also wish to communicate summary data or visualizations to their management or the public.

Chapter 2: Recommendations

1. Develop a clear understanding of existing data, data gaps and data required to promote watershed health and sustainability.
2. Focus initially on a limited, common set of data which are collected and utilized by the majority of participating organizations.
3. Harvest data from existing systems where available and when possible, use web-services to access data on-the-fly.
4. Do not include historical data which does not conform to system standards as an early priority. This data requires a significant effort to prepare for inclusion in a DMS.
5. Identify the data requirements of users at each stage of the data workflow and develop the DMS to provide, locate and access those data efficiently and effectively.

Chapter 3: Design and Structural Recommendations

This chapter provides specific technical recommendations for consideration in developing a regional DMS for the SD-IRWM, as described in the following sections. Stakeholders developed and prioritized a list of suggested design features (Appendix D). In the following design recommendations we incorporate the highest rated of these priorities to the degree feasible given the current state of technology and what we believe is achievable within a realistic timeframe and at reasonable cost.

3.1: General Design Principles

The Advisory Workgroup recommends the SD-IRWM develop the desired DMS as a federated data system, using an open source software platform. Data incorporated into this system should follow open data standards evaluated by and overseen by a governing body. Design specifications and implementation of the DMS should be phased over a period of time (see section 3.6) according to the availability of resources. Since technology changes relatively rapidly, no specific software or hardware platform, though is recommend at this time.

> ***Recommendation 6:***
>
> *Use a federated data system structure for the SD-IRWM data management system.*

For many years, the technological standard in the data management world was to create a single, all-encompassing data repository. These centralized data systems would store all relevant data on a powerful server with substantial storage space. A repository, (also called a data warehouse) typically requires a trained system administrator to maintain the server's function and efficiency by installing updates to security software and operating system, monitoring system diagnostics and load balancing (tasks which have no relationship to the domain specific data and data structures). Centralized architectures often over-allocate time, money, and resources during the design phase in anticipation of eventual needs or require significant influxes of resources when the hardware reaches its limitations. Today, high-speed computer networks, inexpensive computer hardware and data standards that provide for interoperability of data and systems, have resulted in better options.

The recommended architecture is the federated data system. A federated system is best described as a system of systems and does not actually store the data in one location. Instead, it accesses data from participating data providers in response to parameters of a data request made by a data user. The request process can happen in a variety of ways, though most commonly this is through a browser-based web interface or phone/tablet application (app).

Rather than retrieving and storing all available data from all data generators in advance, a federated system performs a focused search using the parameters of the data request to the global data catalog as specified by the client (Figure 3-1). Only matching data from each data provider participating in the federated system is retrieved and delivered to the client. Data is pulled from each agency's source database (stored in their local format), and processed through a series of operations on the federated database server to display (or provide for download) a single data set in a consistent format. These compiled data may be delivered directly to the client, or to a client application via a web service for processing into a report or visualization.

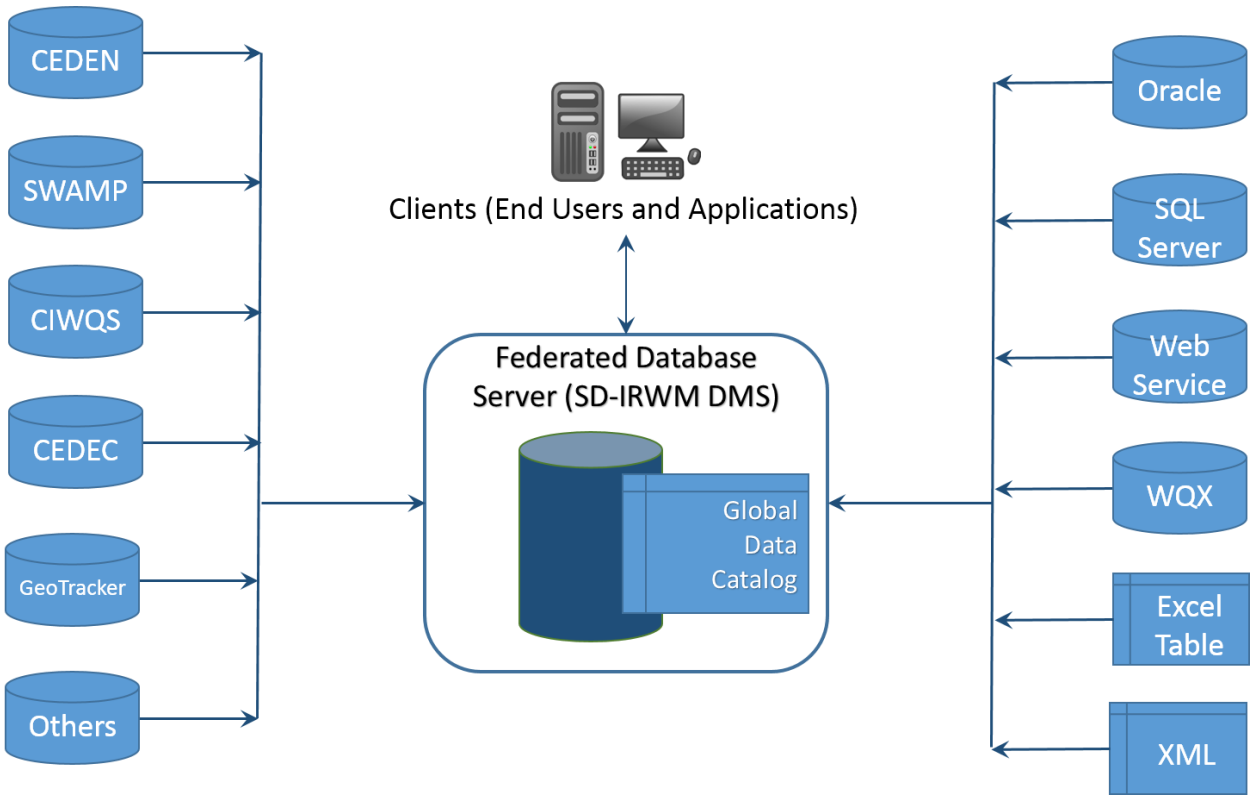

Figure 3-1: A federated system design provides for integration of multiple database systems and platforms into a virtual database through a client interface appearing as a single database to the end user.

3.1.1: Metadata

Metadata can broadly be described as data about data. It serves as the documentation for each contributed data set or data service and captures valuable information to track information including when data was collected, by whom, with what methods and where and how it is stored and accessed. Metadata provides a basis for addressing two essential data characteristics for the client:

1. The relevance of available data in meeting the needs of the end user (e.g. a data type that addresses the question or topic of interest as described by the definition of watershed health and sustainability); and
2. The spatial, temporal and quality of the data, (e.g. the date, time, location, QA/QC, generating organization, contact information, etc.).

Metadata is essential to properly locate and retrieve relevant data. Many databases fail because metadata is not prioritized during the planning stages. Given the definition of watershed health and sustainability is such a widely encompassing concept, metadata will play an important role in ensuring that data is properly categorized and cataloged so it can be retrieved and assessed correctly with respect to each component of the definition.

The choice to capture historical data can prove challenging. Often, the ways and means by which data was collected and recorded in years past was not as detailed as would be today (for example, the achievable detection limits may not be sufficient to meet modern reporting requirements, or data may have missing or insufficient QA/QC). The most notable challenge in considering when, or even if, to include historical data would be reconciling stored calculated values. The values for particular projects, or the statistical methods used, may change over time, let alone throughout decades of past data from multiple agencies and organizations. Therefore, comparing a value calculated by contemporary methods may not have analogous meaning to one calculated many years ago.

As important as metadata is for current data sets, it is equally as important for historic data sets. Unfortunately, the availability of metadata for historic data sets is often missing or incomplete. Many older data sets, particularly those collected without the intention of sharing

the data with other organizations, simply did not record metadata in the first place. While sometimes important information may still be available by interviewing long-time staff members who may have personal knowledge of the data, staff turnover and the passing of time lead to questionable recall of the methods and conditions under which sampling, analysis, and reporting took place. When available, however, metadata equally functions to avoid data misuse, misrepresentation, and/or mis-categorization.

3.1.2: Historical Data

When the topic of metadata was addressed during stakeholder workshops, many stakeholders brought up concerns for historic data and posed questions as to whether it would be included in the DMS along with its metadata. As mentioned in section 2.4.2, historical data can present several challenges, including high cost and effort to integrate into a shared DMS. We recommend inclusion of historical data be given low priority during the initial development of the DMS. If particular historical data is prioritized for inclusion in the DMS it should be subject to the same metadata requirements as any other data in the system. When possible we recommend historical data be held for inclusion during later stages of the DMS development. Selection of high-value historical data should be prioritized over data that may be of limited value, either due to a lack of known and documented quality and/or comparability with contemporary data.

> ***Recommendation 7:***
>
> *Subject historical data to the same metadata requirements as all other data in the system.*

3.1.3: Adopt Existing Metadata Standards

> ***Recommendation 8:***
>
> *Adopt existing metadata standards to maximize compatibility with existing State and Federal data systems.*

In order to build a strong architectural foundation, it is imperative (and thus recommended) that appropriate accommodations for metadata be made in the new DMS. Well-developed metadata standards and terminologies available such as those provided by the Open Geospatial Consortium's WaterML 2.0, provide an ideal standard information model to represent water observations data. Following existing OGC standards makes the regional DMS readily interoperable while addressing exchange requirements for the regional DMS and with other, external systems. Therefore, we recommend adoption of an existing standard such as WaterML 2.0 over the development of non-compliant, local standards. The Open Geospatial Consortium (OGC) has recently approved the Sensor Observation Service 2.0 Hydrology Profile Best Practice document which, among other components, defines the protocol by which WaterML 2.0 content is exchanged.

Web links: www.opengeospatial.org/standards/waterml
www.opengeospatial.org/pressroom/pressreleases/2119

3.2: Database Input

3.2.1: Data Collection and Sharing

In a data warehouse structure, data must be submitted to the central database prior to it being available to end users. Transfer of data from generators into the warehouse requires additional processing steps, and interaction with a system administrator. The difficulty with this approach is that it requires data generators to alter their own internal processes to match the formats required, which may be viewed as barriers to participation requiring extra, recurring cost and effort. Additionally, data delivery to the central system is not immediately accessible. As a result, data accessed by end users may be out-of-date, incomplete or erroneous compared to the version held internally by the data generator.

A federated system structure largely eliminates the above issues. While the initial development costs for a federated system are likely similar to a data warehouse, the recurring costs are significantly reduced given that individual organizations are in direct control of their own data updates. Data submission is an automated process, requiring initial set-up and minimal maintenance beyond what are internal data management processes already in place within the organization. When a data request is made by a client system, the collection of the desired data happens dynamically from the perspective of both the data generator and the data user with all data compilation handled through an automated machine-to-machine (M2M) interaction. The federated server reaches out to all affiliated systems via web-services, and using pre-defined translation tables developed using open data standards (i.e. HTML, XML, etc.) gathers and compiles the relevant data at that moment and delivers it to the data user. Since there is no permanent and centralized data storage server, the client always receives the most up-to-date data as provided by each contributing system. If a data generator adds or updates the data in their own system, it is immediately available to clients of the federated system. Additionally, because a federated system draws data from the original source, only one, authoritative copy of the data is maintained, thus eliminating the possibility of multiple, and potentially inconsistent, versions.

> ***Recommendation 9:***
>
> *Adopt open data standards to ensure interoperability among federated systems.*

Given the overall performance of a federated system is defined by the slowest node, consideration should be given to caching or mirroring of data from any individual data provider with an unacceptably slow or unreliable connection. Although this may lead to a slight possibility of not receiving the most current data, caching data on a frequent basis and particularly when changes are known to have been made, can minimize the likelihood of this occurring. Additionally, the system should provide an alert when data is obtained from a cached copy rather than the originating source.

The federated data system architecture will allow most participating data generators to continue using whatever DMS platform they already have in place, provided they are capable of

exposing data through a web-service. This approach will minimize alterations to current DMS and workflows while providing for sharing of specified subsets of data with the regional DMS (i.e. specified subsets of the data may be shared while other data in a participating system can remain private, thereby allowing each individual data generator to retain ownership and control over their data). Ensuring source data remains under control of the data generator eliminates the risk of multiple versions and/or incorrect or outdated copies of data to be available to end users of the system.

3.2.2: Data Sharing Challenges

> ***Recommendation 10:***
>
> *Develop a governance structure to oversee data sharing concerns, including data privacy and quality control protocols.*

Stakeholders voiced a desire to include mechanisms to control data sharing, including data privacy and quality control protocols in the system design. While some data sharing protocols may be addressed through technical design, many of these concerns are best addressed through the development of Memorandums of Understanding (MOU) or data sharing agreements in addition to data and metadata standards for participating organizations. The data sharing MOU should also address technical issues including data availability, expectations regarding the system speed and uptime for each node (or caching protocols for those which cannot achieve them) and for coordinated version control and updates to data. From a technical perspective, data sharing can be controlled by limiting data access via account access control (requiring users to establish an account on the system which provides a specific level of access to data). Additionally, because the global data catalog of a federated system allows queries of metadata, even in cases where a given dataset is not provided directly via the system, a search can return information about the restricted data and provide information on how it may be obtained.

While concerns regarding data sharing may be warranted, over time, as the value of a shared DMS builds and trust in data obtained from other participating organizations increases, concerns over data sharing will typically decrease. As trust among participating organizations builds there is potential to reduce redundancy in data collection (for example where two organizations

previously collected similar data in space and time to “checkup” on one another). This may allow for the reallocation of monitoring data collection and management to provide improved spatial and temporal coverage for the IRWM region. While it is recognized that some particular data types may be sensitive and therefore inappropriate for full and open access; our general recommendation is that data be as open and transparent as feasible. Therefore, a hybrid data access structure by which most data is freely available and some may be restricted is the preferred option. Developing a governance structure to examine these questions and policies in advance will provide the best mechanism to address such concerns.

3.2.3: Updating Data

While the initial development of a DMS should be expected to be significant regardless of the architecture, the ongoing maintenance for a Federated system should be significantly less. For example, occasional updates or corrections to data are inevitable. Changes to data sets occur for multiple reasons, including sample misidentification, contamination, and human error. During the planning stages of a DMS, mechanisms to update and track changes to the data are essential. Data stored using a centralized model requires changes are managed by both the data generator and by the administrator of the central system. By contrast, updates to data in a federated system are managed exclusively by the data generator and become instantly available to all end-users of the system reducing the need for extensive staffing and upkeep costs to maintain the central data catalog, discovery and retrieval system. There are staffing requirements to maintain the central registry and a data catalog to ensure they remain in sync with the data. The system should also provide a management dashboard for staff to monitor all services in the system, and alert them of any failures. With a well-maintained central registry, there is minimal risk of confusion caused by the existence of multiple, and potentially differing, versions of the data. All interested users of the data have access to the exact same data at any given time.

3.3: Data Processing/Transmission

3.3.1: Retrieval and Distribution

A key benefit of a shared DMS is to centralize the discovery and retrieval of relevant data. Currently, not all stakeholders make data available through a web-accessible format. As a result, users must locate data not only by searching multiple organizations' websites, but also through personal phone or email contacts. This process leads to a high likelihood that significant data may be excluded from an analysis, simply because the user does not know where to find it or if the data even exists. Even if one is successful in identifying, locating and obtaining the available data, there is no guarantee it will be usable due to differences in data structure, collection methods, QA/QC and comparability.

A shared DMS will help to alleviate many of these issues by providing the required infrastructure to manage, catalog and make available comparable data for the region, and also to serve as a single point for discovery and distribution. The need for data generators to submit their data to multiple regional, State and/or Federal data repositories as currently required by a variety of permit and regulatory requirements (e.g. WQX, CIWQS, CEDEN, SWAMP, etc.) may be reduced or eliminated. A federated DMS can deliver data to any of these systems eliminating the need for data providers to reformat their data to satisfy the submission requirements of multiple data systems. Stakeholders expressed interest in a DMS architecture able to automatically deliver their formatted data to meet multiple submission requirements. A well-designed region-wide DMS could provide a one-stop-shop to effectively and efficiently meet these needs and minimize duplicative efforts.

3.3.2: Efficient, Interoperable Data

Traditionally, using data from multiple organizations has been a challenge to end users. While different agencies may collect data for the same purpose, variations in the data formats and structures can result in data being recorded in completely different ways. Stakeholders highlighted a need for standardization of data ontologies, units of measurement and standard/documented methods as components of metadata included in the DMS. Standardization can provide for efficient data comparability and interoperability between systems, agencies and organizations. While standardization can be accomplished using either a

centralized or federated structure, however, the primary benefit of a federated structure is that interoperability is built into the system at the front end though M2M communication using open data standards. Centralized systems require data generators to compile this information at the level of each data submission or revision, which is much more time and labor intensive.

A regional DMS would provide a documented and consistent data delivery platform that could deliver the data to these other systems automatically. Achievement of automated data delivery would provide a DMS platform that is much more efficient system than any system currently in existence and would be considered a tremendous success by stakeholders in the region. As federated models have been gaining in popularity in recent years, there is a slow, but steady move away from a static centralized repository approach.

3.3.3: Types of Data Stored

A data type is a group of data records that follow a unified format. Recommended data types for the DMS to store, as identified by thematic workgroups during the stakeholder meetings, are listed below:

- **Benthic**: data pertaining to species identification, abundance, and habitat assessment based on ecological diversity
- **Chemistry**: data for samples analyzed in a lab setting
- **Continuous**: data collected on a frequent time scale (e.g. flow data collected every 15 minutes from a probe)
- **Field/Habitat**: field observational data; can be subjective
- **Location**: capture of geographic information such as latitude/longitude or site polygons or channel segment centerlines
- **Microbiology**: data associated with bacteria testing
- **Meteorological**: data such as storm patterns and dew point
- **Riparian Habitat Assessment**: data collected and used for riparian health and monitoring, including channel characteristics, riparian vegetation, canopy cover, etc.)
- **Toxicity**: data associated with toxicity testing.

The prioritization of which data types should be addressed should be based on several factors: First, which data are needed to effectively address the definition of watershed health and sustainability? Second, an assessment of which data are already available via web-services or can easily be made available to a federated architecture from an existing web-based DMS should be conducted. Finally, the overall readiness of the data should be assessed based on the associated metadata, data QA/QC to ensure effort is focused on initially incorporating high quality data and time and effort needed to prepare for inclusion. A review of the complexity of these data types allows us to assess the relative cost and complexity of developing and integrating particular data types as represented below (Table 3-1). Priority was assessed based on projections of data usage. By this logic, we assume that data types with higher demand warrant nearer term prioritization. These priorities may require adjustment based on the highest priorities defined at the time development of an initial system is undertaken.

Recommendation 11:

Prioritize data types based on: (1) those that address watershed health and sustainability, (2) those that are readily available, and (3) those that are of high quality.

Table 3-1: Relative effort (cost and complexity) of integrating various data types into a regional DMS.

	Near-Term Priority	Mid-Term Priority	Long-Term Priority
Low Cost	Field; Microbiology		
Medium Cost	Chemistry	Benthic	Stream Assessment
High Cost	Toxicity	Sensor Data (Continuous data stream)	Meteorological

3.3.4: Consolidated Searches

Stakeholders felt strongly that the DMS provide a mechanism to perform a single search to find all the data needed to answer questions about watershed health and sustainability. To accomplish this, the DMS would need to connect all data generators identified as holders of appropriate data including consideration for connections to other relevant DMS. A clear

advantage of the federated model is the global data catalog coupled with web-services to share data between and among systems. The need to individually visit and search multiple systems to obtain and then compile the required data will likely decrease in the coming years.

3.4: Data Outputs

Data outputs vary just as widely as their intended uses. The strength of a federated system structure is that it provides a platform upon which a wide variety of data output and visualization tools may be built. Making data discoverable and accessible using web-services provides the necessary infrastructure for anyone to build customized tools meeting their analytical requirements, independent of the core system. Several examples of output tools are provided, but specific recommendations are not given as the value of such tools is dependent upon specific user requirements which are not a focus of this stakeholder process. With applicability to the components defined for watershed health and sustainability, some of the core system outputs to be considered include:

> ***Recommendation 12:***
>
> *Develop simple, informative output tools that address planning and management priorities of the stakeholders. Complex output and analysis capabilities should only be developed after the core data is successfully registered to the global DMS catalog.*

1. **Simple Aggregated:** (e.g. tables) These are simple summary outputs of common, useful data. Table 3-1 shows an example of a simple, aggregated table listing all stations sampled for a particular analyte of interest. Tables representing lists of desired information may be used by an analyst to develop their own assessment or interpretation once queried from the data system.

Table 3-2: Example of a simple aggregated data output. This table lists all the stations sampled for Demeton, Total.

Demeton, Total	
Site Code	Site Name
12-351	IRV007 Receiving Water
406ItalianGardens	Italian Gardens
406NALF_SCI_REC1	San Clemente Island Receiving
406NALF_SCI_REF1	San Clemente Island Reference
409DC	Deer Creek _ ASBS
901EM	El Moro Canyon
901SO	San Onofre_ASBS

2. **Calculated/Visualized**: In some cases, the end user does not desire the raw data, but rather data that has been meaningfully summarized. This may provide the easiest means for a decision-maker or the public to digest data and understand the purpose for which it was collected. One such example of a well-developed data visualization interface is *Bay Delta Live* which provides visualized data for a variety of data types (Figure 3-2).

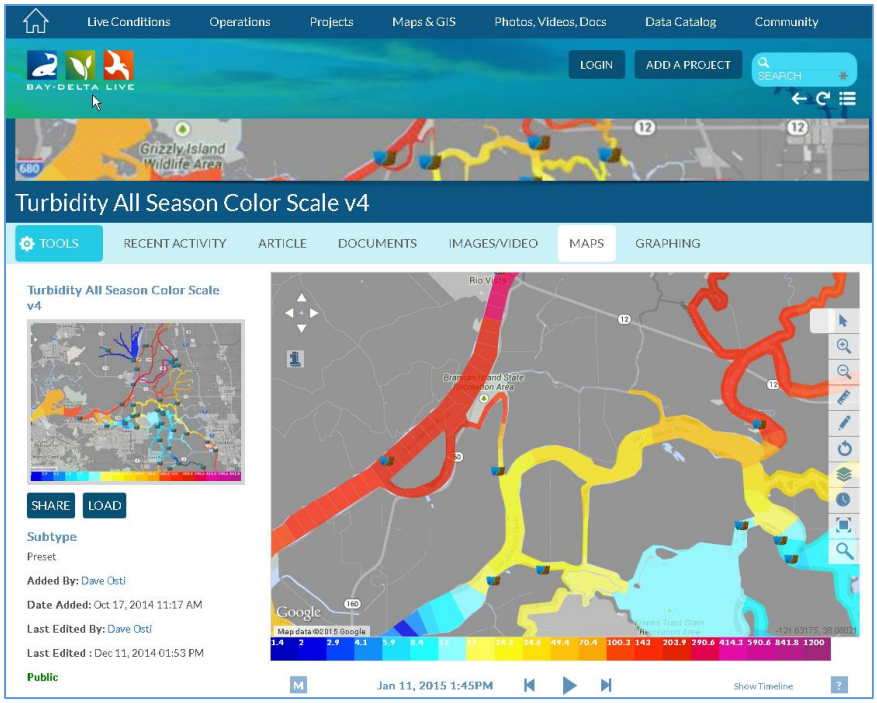

Figure 3-2: An example of visualized data for annualized turbidity in the Bay Delta.
Source: www.baydeltalive.com

Calculated values may represent indices or statistical analysis of data for a particular geography or time period. The output may include visual representations such as graphs, charts or maps of the desired data. Examples of calculated/visualized uses include:

- **Monitoring:** The ability to establish comparisons to baseline conditions and to evaluate the effectiveness of monitoring as well as the ability to inventory existing efforts to avoid unwanted duplication of effort, costs, etc.
- **Statistics/Trends**: Examining the statistical distribution of particular monitoring data for comparison with regulatory thresholds or restoration objectives may be required in certain situations were a single measurement is not sufficient to understand the status of a particular waterbody or the trends exhibited by changing values over time (i.e. is a given parameter increasing or decreasing over time?)
- **Best Management Practices (BMP)**: Assessing the effectiveness of BMP's often requires the synthesis of multiple data values to derive a defined index calculation or model to understand condition or change based upon a computed value.
- **Visualized data**: In many circumstances the desired output may not be reported as numeric values, but rather as charts, graphs or maps. Visualization of data in graphic formats often require computed or summary values as a means to generate the desired output. For example to represent monthly averages throughout a calendar year relative to a threshold daily values may require averaging by month.

Development of derived data product tools must follow upon the development of a core DMS that provides the underlying and comparable data from participating organizations such that these calculated or visualized products may be developed per system capabilities.

3.4.1: Hosting of Data from Data Generators

It is expected that some participating data generators will not have the capacity to develop and maintain their own web-accessible DMS. In these cases accommodation for their data to be hosted elsewhere may be necessary. Three recommended solutions may resolve such cases and should be assessed on a case-by-case basis:

Recommendation 13:

Develop a mechanism to host data for participants in the DMS who lack internal capacity to do so themselves.

1. Hosting of data by generators may be made a condition of participation in the federated system (the generator establishes their own capacity to host data);
2. If one or more participants are willing and able to store and manage data on behalf of others using their own, existing infrastructure hosting can be provided as an in-kind or fee for service; or
3. The same infrastructure used to host the web-interface to the federated data system could also be used to maintain a database for orphan data.

The third option while functional, is least desirable since it requires additional development and resources for long-term maintenance and support within the structure of the SD-IRWM.

3.4.2: Web Interface

The web interface through which the DMS operates and interacts with data generators' internal systems is the focal point of development for purposes of the SD-IRWM. It is recommended that the web interface for the DMS serve two primary functions.

1. Data discovery and retrieval;
2. System management.

> ***Recommendation 14:***
>
> *Provide access tools for users to discover and retrieve data, and for administrative management of the DMS system using a web interface.*

The data discovery/retrieval function provides a web-based database platform designed to serve as the means to access data within the DMS. The design should include a multiple pathways to data discovery, including keyword, watershed and map-based queries of the data. Once an end user has discovered the desired data the interface should provide options to download data in its original form (e.g. a structured data file for use in offline analysis) or as a derived data product (calculated values, data visualizations, etc.). Additionally, the system should allow for machine-to-machine interactions allowing desired subsets of data to be accessed through web services without direct end-user interaction through a service level application program interface (API).

Secondly, the web interface should serve as a management front end for data generators and system managers (Figure 3-4). For example, providing a data dashboard to show the current status of available data — an online indication of whether a participating agency has made their data accessible or not at the time of the data request. A dashboard may offer additional useful information and statistics about the DMS such as indicating when new data have been added or may be forthcoming, or to indicate if a particular participating organization's system is offline for maintenance and/or other reasons making a portion of the data unavailable. The management interface would also provide a means for data generators to update their registered data and metadata.

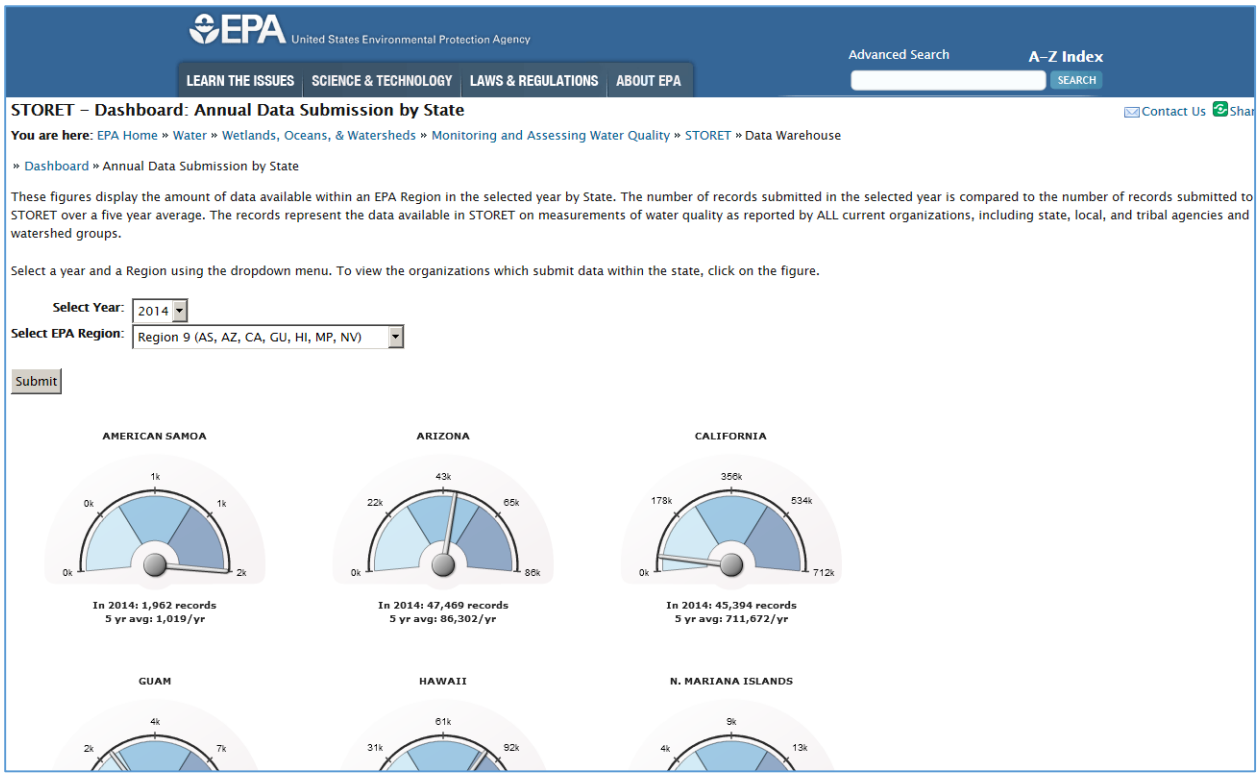

Figure 3-4: An example of a dashboard view provided by the EPA STORET system showing the number of data submissions by geography (state) for a given year and multi-year average. Other views provide details on the contributing organizations, stations, number of data records by data type, etc.

A centralized DMS will also typically require a web interface for data generators to submit their properly structured data on a regular basis (Figure 3-5). For centralized systems such as CEDEN, data submitted by a generator is checked and approved by a data moderator, a person who takes data submissions from the providers and loads submissions into the system. While having someone manually load data can be effective, and provides an additional QA/QC check, it does slow the process from the time of data generation to data availability.

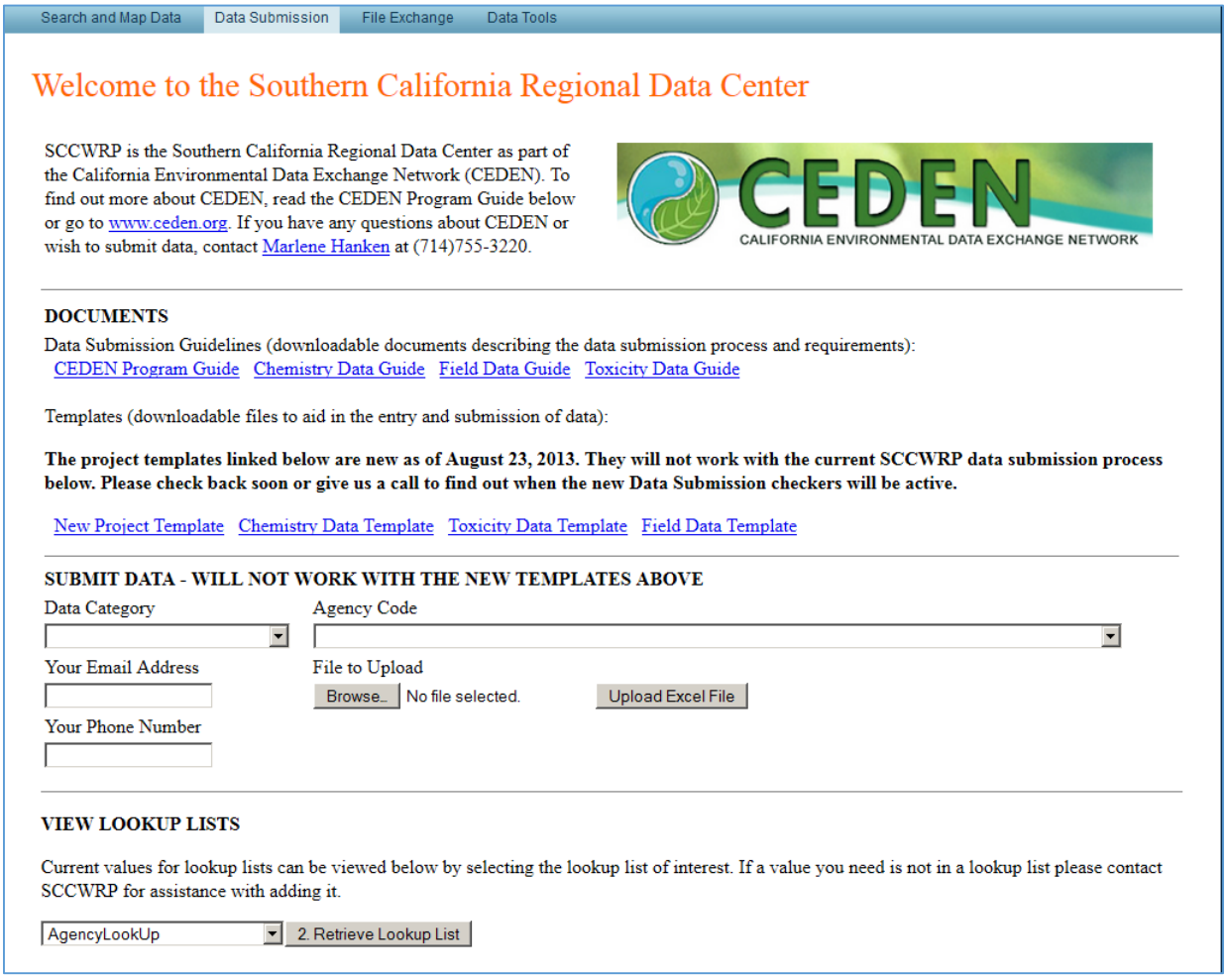

Figure 3-5: An example of a web-based data submission interface for CEDEN, through which the data generator uploads a properly formatted data file for review and loading by a moderator to the central data server for access via the public facing web query interface.

Federated systems replace the requirement for a data input mechanism. Once a given dataset has been registered to the catalog, metadata updates may be scheduled to automatically update the repository on a regular basis, to refresh the catalog with any changes are made by a data provider. These updates must be validated by the system before making them available to users

of the DMS. Data is delivered on demand from the individual data systems making up the federated architecture and data structure and formatting is handled by pre-configured programming code. This approach does require a degree of advance planning and configuration to handle various types of data required for use by participants in the region. Once it is in place there is no need for a data moderator or data librarian to handle each new data set collected by a data generator if their system has already been integrated into the DMS. This moves the burden of data formatting and submission from the data provider to the DMS making the data transfer process more efficient and cost effective than the centralized model.

A key consideration for any web-based data system is the development of an intuitive and easy-to-use interface that makes data discovery and download as easy as possible. Early web-based databases often provided complex interfaces with too many options, making the experience of identifying and accessing the needed data a frustrating process for users (Figure 3-6).

Recommendation 15:

Provide an intuitive and easy-to-use interface based on user input and testing to ensure an effective user experience.

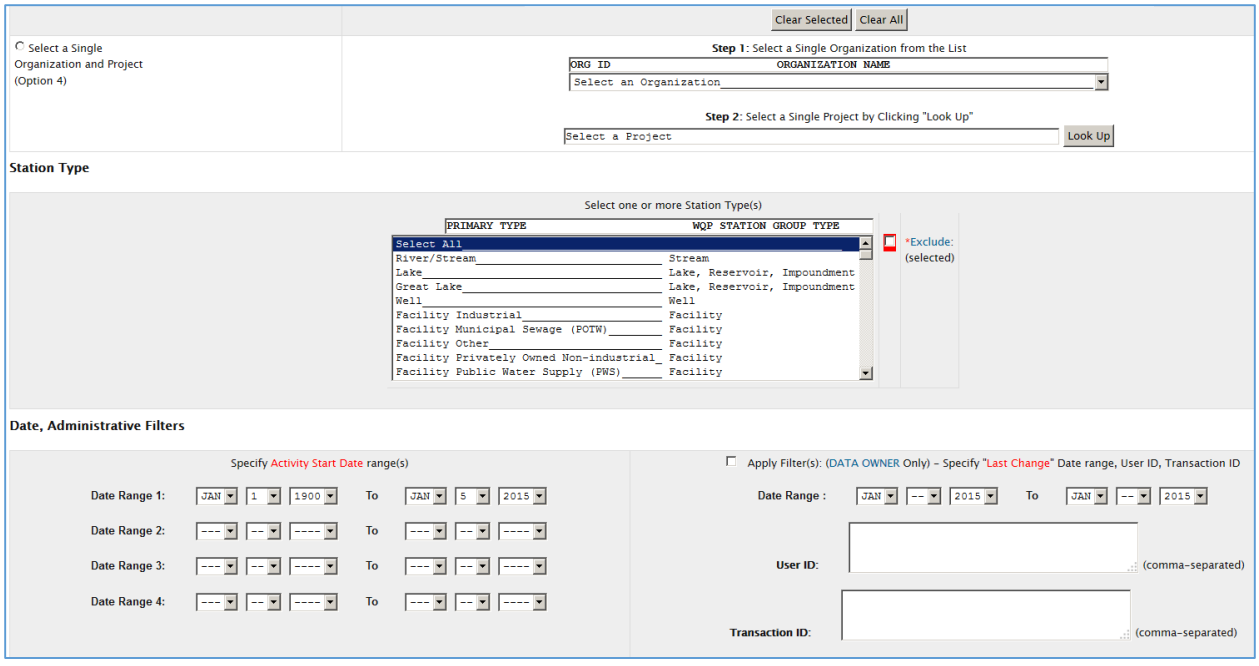

Figure 3-6: The interface provided to access data on the EPA WQX/Storet system runs over several screens with numerous menus and options that must be selected to narrow down the data search to what is desired. While the interface does achieve this goal, it does so with a non-intuitive interface that can frustrate end-users.

A system driven by high quality metadata can make the search experience much simpler, while still narrowing the data effectively. For example, map-based interfaces can allow the selection of a particular geography of interest and key word searches (Figure 3-7).

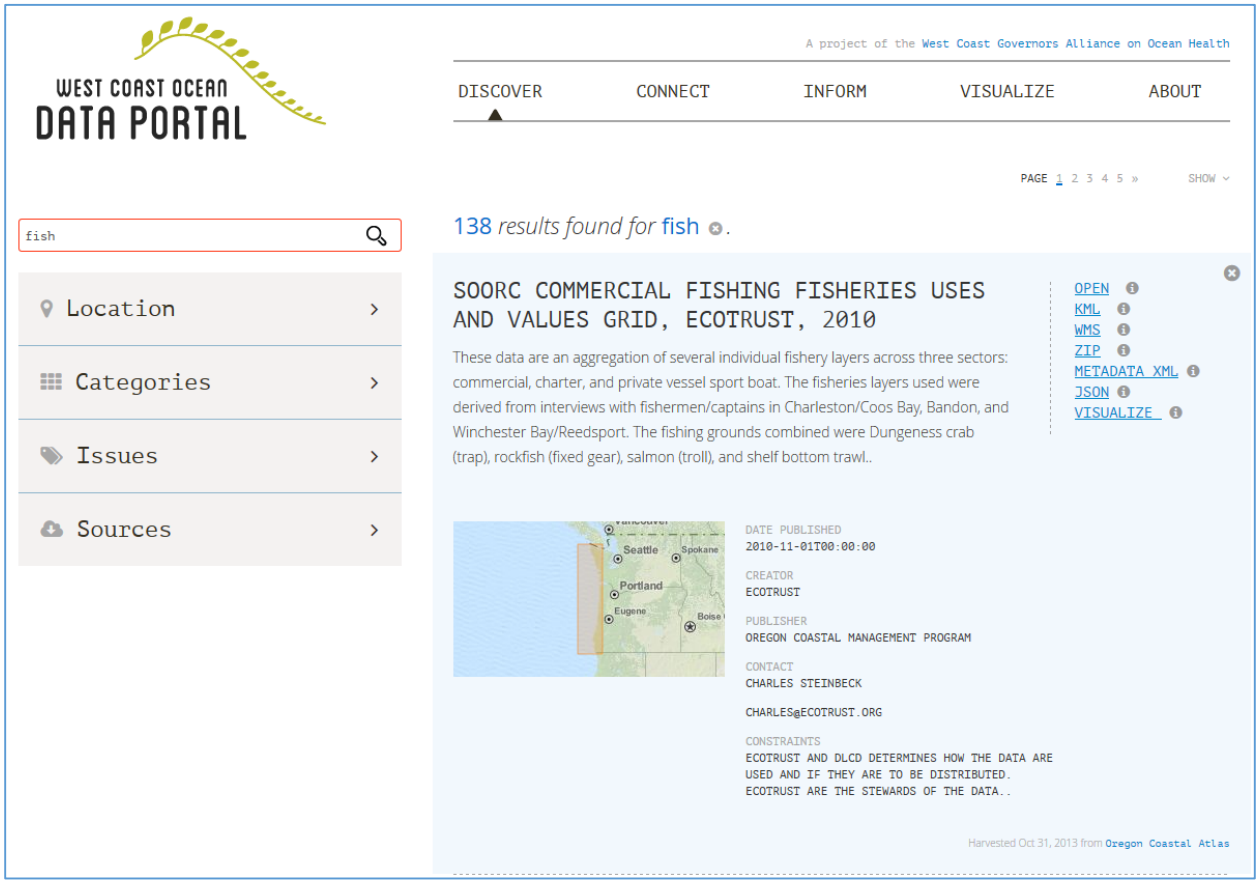

Figure 3-7: The WCGA Ocean Data portal provides a simple search interface allowing users to select data based on key word, geography, data category, topical (management) issues or data providers (left side of figure). Results are reported with brief information on screen along with a list of download formats available (right side of figure).

3.5: System Architecture

Viable software solutions are available from both open source and commercial software providers. Many federated systems are built using open source tools and the communities of support and technical knowledge of both agency staff and firms which may be contracted to develop the DMS are plentiful. A primary benefit of open source is that there are no license fees for the software

> ***Recommendation 16:***
>
> *Use open source software tools and standards as a basis for the federated data structure.*

itself and the non-proprietary platform may allow for more flexibility in selection of web-hosting services, scaling the system to accommodate additional use over time, and the option of upgrading or not-upgrading software versions as desired.

By contrast, commercial software is largely the opposite. Commercial software must be licensed, often based on the number of central processing units in the server. Thus scaling up can incur additional expense. Licenses are typically renewed on an annual basis, so maintaining currency requires significant ongoing funding. Additionally, forced upgrades to new versions may break the existing system and necessitate redesign and redevelopment every few years. Commercial software companies typically cease support for earlier versions at some point which may force upgrades and expenses to the system that are not otherwise necessary or desired. Commercial software typically provides less flexibility to customize because the underlying source code is not available, so changes and updates to features and functionality are limited to those implemented by the company who develops the software.

> ***Recommendation 17:***
>
> *Acknowledge, and plan for, integration with existing investments of partners' data systems. Existing investment and software products may already incorporate the required functionality to allow for integration with a federated data management system.*

The primary advantage of commercial solutions is that they typically offer technical support by phone or email and the software may be relatively easier to install and begin using "out-of-the-box" when compared to open source solutions. Additionally, some of the participating agencies and organizations have significant investments in commercial water data systems such as Kisters Hydstra <www.kisters.net/hydstra.html>. Among others, Kisters Hydstra provides the necessary functionality to allow for integration with a federated architecture allowing for integration of such existing agency systems with the regional DMS.

3.5.1: System Specifications

At this stage it would be inappropriate to provide exact specifications for the development of the DMS. Though exact specifications cannot be made prior to decisions regarding data to be made available and the number of web services to be integrated into a federated system, the following considerations are provided to guide decisions about the structure of the DMS. General specifications for the DMS should include:

- Software architecture
- Hardware/Hosting platform
- Bandwidth
- Processing and availability
- System Storage Capacity
- Mirroring

3.5.2: Software architecture

The web interface and data transfer services should be developed using available open-source software. A number of appropriate open source platforms exist at present, and potentially new or improved platforms may become available by the time development of the DMS commences. Therefore, we recommend that an assessment of the best currently available technology be made when the SD-IRWM is ready to embark on development. This should be done either as a precursor to or component of preparing a Request for Proposal for development of the system. A good example of such an assessment of several open source data registry systems is the *WCGA RDF Data Registry Design Assessment* (2013) available at: www.westcoastoceans.org/media/data_network_act/wcga_rdf_data_registry_design_assessment_2013.pdf

3.5.3: Hardware/Hosting platform

Given the rapid evolution of computing technology, it is recommended that the hardware allocated to the DMS be designed to adjust elastically as needs evolve. In cloud computing, elasticity is defined as the degree to which a system is able to adapt to workload changes by provisioning and de-provisioning resources automatically, such that at each point in time the available resources match the current demand as closely as possible. Hosting environments such

as Amazon EC2 or Microsoft Azure provide elastic platforms which can be adjusted to meet demand in real-time.

Web Links: aws.amazon.com/ec2
azure.microsoft.com

In general, an elastic cloud application or process has three elasticity dimensions: Cost, Quality, and Resources, enabling it to increase and decrease its cost, quality, or available resources to accommodate specific requirements. The primary advantage of this approach, whether obtained through a commercial provider, a University or a State Agency such as the California Technology Agency Cloud, is that when there is little activity on the DMS, resources are de-allocated to reduce costs, and when use is heavy, additional resources are immediately brought to bear to ensure a high degree of system performance.

Recommendation 18:

Use elastic cloud computing at the host organization to provide for adaptive infrastructure as the system and workload demands necessitate.

3.5.4: Bandwidth

The bandwidth describes how quickly data can be transmitted to and from the server. More simply put, on the "information highway," it is analogous to how many lanes there are and what the speed limit is. Bandwidth costs are typically incurred as a monthly fee. For the website serving as a point of access for the system, it will be essential to obtain bandwidth sufficient to handle the anticipated data load. Using a hosting platform with elasticity can help to ensure appropriate bandwidth is available for the connection between the web interface and the user.

However, because federated data systems also rely on the connectivity of each participating data generator's system, it will be important to assess the connection speeds of participating providers. Providers who contribute data that is in high demand and of large size may require faster connections, or if this is not feasible, to arrange for caching of these data on a faster node. Getting a proper amount of bandwidth is necessary for the system input/output capabilities. The

process of searching metadata to identify which data meet a user's needs and should be pulled from the various generators requires relatively low bandwidth since it is essentially a text search and no data needs to be moved. When the actual data is requested demands will be higher, though distributed to each of the contributing nodes on the network. In general this facilitates rapid data transfers. However, for very large data requests it may be desirable to package the data offline. Then when the data are ready to be shared, the system emails the user a link to download the data package from a File Transfer Protocol (FTP) site.

3.5.5: Processing and Availability

In a broad sense, processing refers to the speed in which data requests are handled and the resulting data are delivered to the user. The processing speed of the DMS can be its greatest attribute or its greatest weakness. Delays due to processing will cost users time and cause frustration if the processing time is too slow. Elastic computing is a good solution to address processor power since resources can be dynamically allocated as necessary. A key benefit of a federated model is that data is available and updated as soon as the generator makes it available on their web-server. There is no need to submit data to a central location and/or for new or updated data to be handled by a third party. For data generators who cannot host their own systems, there may be a lag time associated with their delivery of data to a secondary host, whether the host is another partner in the SD-IRWM or the SD-IRWM DMS itself.

For the data users, most of the availability issues are invisible, with the most current data being provided at the time of the data request. One very important exception to this rule is when the connection to a data generator is offline. In such a circumstance, the DMS should report that a portion of the data is not currently available, offer an anticipated time of availability (e.g. if the site is temporarily down for maintenance) and provide a means to contact the data generator directly for more information.

Recommendation 19:

The DMS should provide end users and managers with data status information and provide data system managers the ability to indicate anticipated time until available.

3.5.6: System Storage Capacity

In relative terms, storage space is one of the most inexpensive components of a DMS. Storage is even less important for the federated data system model since the actual data is not housed on the server, but rather with the generator. The possible exception to this would be if the DMS also serves as a data node to the system. As noted earlier, if some data generators do not have the capability to host their own data, it may be appropriate for the DMS to offer the option of housing data as one of the services provided.

3.5.7: Mirroring

Data mirroring is a process by which an entire data system is duplicated on more than one host or server in real time. Mirroring provides an additional level of system security since there are typically two or more mirrors, each independently capable of providing all features of the primary system. In a federated architecture, mirroring may provide an alternative point of access should a primary server be off-line for maintenance or due to power or system failures. Mirroring requires a second host server, preferably in a separate location. This can add to the cost of a commercial host server, but may be inexpensive or even free if an SD-IRWM partner organization has the capacity to host the mirror in its own data center.

3.6: Phased Development

Development of a complex DMS can be an overwhelming process if attempted as a single undertaking. Therefore, we recommend that components of the development process be parsed into manageable subcomponents based on the time and cost associated with their development. As priorities and resources become available, these subcomponents may be developed individually, but with an understanding of how each will contribute to the overall DMS design and implementation. Table C-1 (Appendix C) provides an illustrative example of how prioritization might be

> ***Recommendation 20:***
>
> *Parse development of the DMS into manageable subcomponents based on the time and cost associated with each aspect of the data management system.*

accomplished relative to the complexity and cost of various components of a comprehensive DMS. Given the realities of available funding and resources to initiate development, use of a prioritization table such as in this example can help to focus resources on a limited set of goals to ensure success.

In this example we categorized major components for development into three general phases: First from a time-based planning perspective representing: Near-Term (~1-3 years), Mid-Term (~3-5 years), and Long-Term (5+ years) and secondly, from a financial planning perspective as: low, medium, or high cost. These categories are provided as general guidance and may change based on available technologies, skill and infrastructure at such time as development of a regional DMS is actually undertaken. At the present time, the SD-IRWM has not provided an indication of either time or budget for development of a DMS. Given the rapid development of computing technology, the cost and time to implementation for a regional DMS, as envisioned by the stakeholders and the AWG may change significantly, although as a general rule, it is anticipated that both would decrease somewhat over time.

Chapter 3: Recommendations

6. Use a federated data system structure for the SD-IRWM data management system.
7. Subject historical data to the same metadata requirements as all other data in the system.
8. Adopt existing metadata standards to maximize compatibility with existing State and Federal data systems.
9. Adopt open data standards to ensure interoperability among federated systems.
10. Develop a governance structure to oversee data sharing concerns, including data privacy and quality control protocols.
11. Prioritize data types based on: (1) those that address watershed health and sustainability, (2) those that are readily available, and (3) those that are of high quality.
12. Develop simple, informative output tools that address planning and management priorities of the stakeholders. Complex output and analysis capabilities should only be developed after the core data is successfully registered to the global DMS catalog.
13. Develop a mechanism to host data for participants in the DMS who lack internal capacity to do so themselves.
14. Provide access tools for users to discover and retrieve data, and for administrative management of the DMS system using a web interface.
15. Provide an intuitive and easy-to-use interface based on user input and testing to ensure an effective user experience.
16. Use open source software tools and standards as a basis for the federated data structure.

17. Acknowledge, and plan for, integration with existing investments of partners' data systems. Existing investment and software products may already incorporate the required functionality to allow for integration with a federated data management system.
18. Use elastic cloud computing at the host organization to provide for adaptive infrastructure as the system and workload demands necessitate.
19. The DMS should provide end users and data system managers with data status information and provide managers the ability to indicated anticipate time until available.
20. Parse development of the DMS into manageable subcomponents based on the time and cost associated with each aspect of the data management system.

Chapter 4: Governance and Database Management Strategy

As a first step, development of a successful regional DMS will require a governance structure to ensure priorities are defined and data generators agree to meet requirements for data documentation, QA/QC and data availability maintenance and funding. Additionally, planning for the long-term support and maintenance of the system is essential. Core components of a governance structure are described below. Initially, the SD-IRWM will need to establish a structure or subcommittee tasked with accomplishing these charges on their behalf.

> ***Recommendation 21:***
>
> *As a first step, develop a DMS governance structure to define priorities and requirements for data formats, QA/QC, documentation, data availability maintenance and funding.*

4.1: Data Management System Governance

Participants in the data system should represent a range of stakeholders, including data generators, managers and users. Successful governance of a regional DMS requires a commitment of shared responsibility. Participating organizations must remain committed to maintaining their own shared data, including the commitment of appropriate funding and staff time necessary to maintain systems and data meeting requirements as defined by the governance structure. While there is a clear benefit to inclusion of individuals with a high degree of technical knowledge, it is equally important to consider input from representatives of stakeholders who use the data or data products to answer questions relating to watershed health and sustainability.

Specific roles of the governance body should include: prioritization of the data required in the initial development of the system, data standards, and data sharing agreements. Additionally, the governance body should explore approaches to ensure sustainable funding for the DMS. Potential models could include line item funding in the SD-IRWM budget, membership fees paid by participating organizations, or for specific development efforts, one-time grant funding opportunities. Identifying, and establishing a sustainable funding mechanism should be

undertaken prior to development of an RFP and contract to build the DMS. The governance body should define size, structure and parliamentary process as appropriate given the standards of the SD-IRWM.

As discussed, it is unreasonable to attempt to incorporate all potential data types from all generators at one time. Rather, data needs should be prioritized and resources allocated accordingly by the governance body. Data sharing and access conditions must be defined in advance of making any data available through the DMS. Establishing data sharing agreements/MOUs with participating organizations can help to ensure trust is built early in the process. Data sharing agreements should address topics such as:

- Which data will be made available?
- How quickly will data be made available following collection/lab processing?
- How frequently will new data be made available?
- Who may access the data? (if there is a concern regarding privacy or security)
- Data standards (QA/QC, metadata documentation, etc.)
- What long-term requirements and/or support will be available to maintain participating, federated data systems?
- Who is responsible for maintaining the global data catalog?

Additionally, the governance body should serve as a reviewer of system functionality and design specifications. This entity should ensure that the data included and data access interface appropriately meet the needs of users.

4.2: Web Development

Development: Development of the actual web-based DMS could potentially be accomplished by one of two parties. If the capability and capacity exist, development may be managed as a shared responsibility done "in-house" by one or more of the participating organizations. The advantage of this approach is that more detailed knowledge of the system design and management will remain within the SD-IRWM after development is completed. Alternatively, development may be outsourced to a third party with experience in development of federated systems, data management systems and web services. In either case it is advisable to build from an existing base wherever feasible. One advantage of using an open source solution is that the code is freely sharable. If there is a similar system that appears to nearly or fully meet the needs

of the SD-IRWM, the underlying software code can be requested and customized rather than building the DMS from the ground up. This could potentially save significant time and resources in developing the initial system.

Hosting: As described above for development, an assessment of hosting capacity within the SD-IRWM should be made. Hosting services and support for the system may be contracted out to an external provider. However, if there is capacity available to host the DMS, at least initially during the early development phase, time and resources may be saved. As the system matures and increased capacity is needed, an internally hosted system could be migrated to an alternate location with elastic computing capacity.

Output and Analysis Tools: The user interface provided for data and metadata access and query as well as the associated download tools will require review and refinement over time. A governance body may serve this role directly as reviewers of prototype systems prior to making the DMS available to a wider review through a public Beta test (drawing from SD-IRWM participants). As the system matures and there is sufficient data integrated, analysis or visualization tools to support answering questions may be developed, with decision-making and planning to follow. A review process for such tools is essential, not only to ensure usability, but also to test that such tools are properly computing derived values or other statistics.

Recommendation 22:

Designate a staff position with specific responsibility to conduct system maintenance and updates to the global data catalog.

System Maintenance: While development of the initial system will represent the primary expense and workload, there are several aspects of ongoing maintenance required to support a federated data system. Specifically, there must be someone responsible to host and maintain the global data catalog to ensure that participating organizations are able to properly register, add and update data provided through their systems. Hosting and maintenance do not necessarily need to be provided by a single organization. Hosting can be contracted to a third party provider or through resources supported by an agency or organization participating in the DMS. System support requires staff time and expertise, and should be defined as a specific responsibility of a

new or existing position. Additionally, new data types or data providers may be incorporated over time requiring staff support to integrate with the DMS. While a primary advantage of federated systems are that the underlying data and data management remain with the originating organization, there remains a need for someone designated to maintain the data catalog and website through which these federated data and web services are accessed.

Allocating one staff-person to support the system, (either via a new position, use of existing staff within the participating organizations and/or by contracting to a third party) should be planned throughout the life of the system. This individual would be responsible for basic system maintenance (e.g. updating scripts already in place, assisting the integration of new participants to provide their data through the federated DMS, system support and documentation). Funding a staff position for maintaining a federated system can present some difficulties since no single agency or organization is responsible. The SD-IRWM should identify options to support sustainable funding of a DMS manager/librarian beyond grant funding which may be used to initially develop the system.

4.3: Data Consistency

The governance body should also be responsible for managing and maintaining data format guidance, metadata standards and controlled vocabulary recommendations for each data type included in the system. For most data types likely to be included in the DMS, there are available standards from organizations such as the Open Geospatial Consortium (OGC), Federal Geographic Data Committee (FDGC) and/or International Standards Organization (ISO). These standards could be adopted, as opposed to developing new or unique standards. Adoption of existing standards will also help to ensure data included in the SD-IRWM DMS will be compatible with data available from comparable

Recommendation 23:

Develop written memorandums of understanding among participating organizations. These should address data and metadata standards and procedures should a member organization fall short of meeting expectations.

systems. Furthermore, the governance body should establish policies specifying responsible parties to set and enforce data standards if a particular data generator fails meet data requirements as agreed to in the MOU.

Data in a federated system relies upon established standards and practices. Any data generator failing to meet them would find that their data cannot be cataloged, discovered and accessed by the DMS. Therefore, it will be in the best interest of all participants to meet the agreed upon standards. Nonetheless, there may be situations in which a data generator has difficulty properly preparing their data or internal DMS web-services. Allocating time and resources to assist with this process, particularly in the early stages of DMS development will be essential to identify and work through the particular issues experienced by the initial participants.

4.3.1: Jurisdiction and Usage

Because data systems are readily accessible worldwide via the internet, there is a need to consider what, if any, regional borders are considered a part of the DMS. There may be minimal to no cost in allowing data generators to catalog their data with the DMS (once initially set up) However, this may not represent a desirable approach unless the data being made available meets the priorities of the SD-IRWM as documented in data agreements/MOUs. Therefore, there will be boundaries (jurisdictionally and/or geographically) at which the SD-IRWM may wish to limit the inclusion of data in the system.

4.3.2: Documentation

Documentation of the intrinsic details of the DMS will avoid set-backs brought on by personnel turnover at the management and/or technical levels. It is important that all aspects of the DMS are properly documented and cataloged. Optimally, such documentation would be included as a section of the website through which the DMS is accessed. It would provide information regarding both the technical and governance structures as well as the process for setting priorities, funding strategies, etc. Comprehensive documentation is vital in the long-term health and sustainability of the program.

4.3.3: Periodic Evaluation of the System in Place

> ***Recommendation 24:***
>
> *Plan regular system reviews every 3-5 years to evaluate the effectiveness and future priorities to guide maintenance and development of the DMS.*

Given the rapid pace of technology, it is essential to occasionally reevaluate the current systems in place. Periodic evaluations can help decrease unnecessary spending and reduce efforts designated to areas of devaluing importance. A decade ago, centralized databases with huge investments in hardware to store and process the data were the focus. Today federated solutions to these issues leverage connectivity and existing resources of multiple collaborators to build effective data sharing platforms. Given the recommendation of a phased development approach, there are natural opportunities to evaluate and assess progress at various stages of development. Realistically it will take 3-5 years to have a reasonably mature system in place that incorporates data from a broad range of generators and represents a variety of data types. System review should occur continually throughout the process and at strategic milestones along the way.

4.3.4: The Advisory Committee

Although the Advisory Workgroup was initially assembled to create this recommendation report, this group could potentially be maintained as a mechanism for review as the DMS is developed. This group represents organizations and individuals who are familiar with the policy and technical issues and have valuable insights. Members of the Advisory Workgroup already understand the scope of the task and have built a degree of trust and knowledge of the program. Members of the Advisory Workgroup, or a similar body, could provide effective leadership in taking the next steps to implement a regional DMS for the SD-IRWM.

4.4: Funding

Because any DMS developed will require initial funding to establish and ongoing funding to maintain and support the system, it is essential to plan for these needs prior to initiation of system development. An investment of time and effort early in the planning process will be

essential to the long term success of the system. An advisory committee should be established to explore models for financing both the construction and maintenance of the system. In particular costs associated with the hosting and staff support of the global data catalog, and website should be identified in advance. Such costs should be weighed in context of the benefits and value of the DMS in supporting participating organizations and end-users. For example, quantifying the staff time saved responding to requests for data. Or staff costs saved to obtain and organize disparate data sources on a case-by-case basis for each project or permit could be significantly reduced by the creation of a SD-IRWM DMS.

Chapter 4: Recommendations

21. As a first step, develop a DMS governance structure to define priorities and requirements for data formats, QA/QC, documentation, data availability maintenance and funding.
22. Designate a staff position with specific responsibility to conduct system maintenance and updates to the global data catalog.
23. Develop written memorandums of understanding among participating organizations. These should address data and metadata standards and procedures should a member organization fall short of meeting expectations.
24. Plan regular system reviews every 3-5 years to evaluate the effectiveness and future priorities to guide maintenance and development of the DMS.

Appendix A: Acknowledgments

A.1: Planning Team Members

Nancy Stalnaker, County of San Diego
Sheri McPherson, County of San Diego
Amber Rogers, County of San Diego
Dorian Fougeres, Center for Collaborative Policy
Meagan Wylie, Center for Collaborative Policy
Steve Steinberg, Southern California Coastal Water Research Project (SCCWRP)
Marlene Hanken, Southern California Coastal Water Research Project (SCCWRP)

A.2: Advisory Work Group Members

Sarah Agahi, County of San Diego
Rand Allan, County of San Diego
Tim Bailey, Santa Fe Irrigation District
Patrick Crais, California Landscape Contractors Association
Lesley Dobalian, San Diego County Water Authority
Jennifer Hazard, Alter Terra
Goldy Herbon, City of San Diego Public Utilities
Bob Leiter, University Of California San Diego
Peter Martin, City of San Diego Public Utilities
Kimberly O'Connell, University of California San Diego
Dawn Olson, City of San Diego Public Utilities
Keith Pezzoli, University of California San Diego
Bruce Posthumus, Regional Water Quality Control Board
Travis Pritchard, San Diego Coastkeeper
Oscar Romo, Alter Terra
Cor Schaffer, Santa Fe Irrigation District
Lan Wiborg, City of San Diego Public Utilities
Helen Yu, Regional Water Quality Control Board
Ilya Zaslavsky, University of California San Diego
Vicky Zhang, HDR Engineering

A.3: Stakeholder Group Members

Sara Agahi, County of San Diego
Rand Allan, County of San Diego
Khosro Aminpour, City of Chula Vista
Tim Bailey, Santa Fe Irrigation District
Anne Bamford, Industrial Environmental Association
Jason Batchelor, County of San Diego
Jack Bebee, Fallbrook Wastewater Authority
Brent Bowman, City of San Diego Public Utilities
Doug Campbell, City of San Diego Public Utilities
Roshan Christoph, AMEC
Mike Miskwish Connolly, Campo Band of Mission Indians
Jeff Crooks, Tijuana National Estuarine Research Reserve
Scott Daeschner, City of San Diego Department of Information Technology
Drew Decker, United States Geological Survey (USGS)
Lesley Dobalian, San Diego County Water Authority
Bryn Evans, Dudek
Valerie Fanning, University of California, San Diego
Jim Fisher, San Diego County Water Authority
Emily Fudge, U.S. Forest Service
Phil Gibbons, Port Authority
Doug Gibson, San Elijo Lagoon Conservancy
Richard Gilb, Airport Authority
Eileen Goff, GeomorphIS
Gladys Gonzalez, County of San Diego
Sophia Hanna, AMEC
Mark Hatcher, Sweetwater Authority
Jennifer Hazard, Alter Terra
Goldy Herbon, City of San Diego Public Utilities
Rob Hutsel, San Diego River Park Foundation

Joni Johnson, Rural Community Assistance Corporation (RCAC)
Keith Kezer, County of San Diego
Marlon King, County of San Diego
Mo Lahsaie, City of Oceanside
Pat Landrum, San Diego Association of Governments (SANDAG)
Eric Larson, Farm Bureau
Bob Leiter, University of California, San Diego
Brad Lind, San Diego Geographic Information Source (SanGIS)
Chris McKinney, City of Escondido
Judy Mitchell, Mission Resource Conservation District (MRCD)
Crystal Najera, City of Encinitas
Tim Nguyen, City of San Diego Public Utilities
Dawn Olson, City of San Diego Public Utilities
Emily Perkins, United States Geological Survey (USGS)
Travis Pritchard, San Diego Coastkeeper
John Quenzer, D-Max Engineering, Inc.
Oscar Romo, Alter Terra
Toby Roy, San Diego County Water Authority
Ken Schiff, Southern California Coastal Water Research Project (SCCWRP)
Rolf Schottle, AMEC
Corey Sheredy, AMEC
Steve Smullen, International Boundary and Water Commission
Andre Sonksen, City of San Diego Transportation & Storm Water
Marisa Soriano, City of Chula Vista
Alex Tardy, NOAA/National Weather Service
Mark Umphres, Helix Water District
Peter Vroom, City of San Diego Public Utilities
Lan Wiborg, City of San Diego Public Utilities
Michael Williams, City of San Diego Public Utilities
Joanna Wisniewska, County of San Diego
Darren Wright, Coastal Data Information Program (CDIP) and Southern California Coastal Ocean Observing System (SCCOOS)
Richard Wright, San Diego State University (emeritus)
Satomi Yonemasu, Weston Solutions
Helen Yu, Regional Water Quality Control Board
Ilya Zaslavsky, University of California, San Diego

Appendix B: Summary of Existing Systems

B.1: California Environmental Data Exchange Network

The California Environmental Data Exchange Network (CEDEN) was originally developed by the State Water Resources Control Board (SWRCB) with support for additional development and refinement by a team including private consultants and staff from Moss Landing Marine Laboratories (MLML). CEDEN's ambitious undertaking was to store numerous types of environmental data including water quality, bio-assessment, and sediment toxicity for both marine data and freshwater systems, stored in a single database. After over seven years in development CEDEN was recently taken back under the control of the SWRCB and is undergoing modifications. CEDEN is a prime example of a traditional data repository: a single server that stores mass amounts of data from various providers.

Inputting data can be difficult, as can retrieving data, especially if a user is seeking a specific data set based on stringent parameters. While CEDEN captures numerous data types and maintains rigorous quality control standards, its size and maintenance requirements are as extensive as the data it stores. Because CEDEN already possesses many of the capabilities and existing infrastructure for a regional DMS, it could serve as a potential platform for the storage and retrieval of many of the desired data types. However, in its current form, the complexity of data submission and slow time-to-availability make CEDEN an inappropriate option for the SD-IRWM to utilize as its data management system. However, given the future evolution of CEDEN including the potential addition of web-services, it could serve as a valuable component of a federated DMS for the region (Section 3.1). Therefore, CEDEN should be re-assessed as a potential component of the prospective regional DMS when actual development is initiated.

Web link: ceden.org

B.2: Surface Water Ambient Monitoring Program

The Surface Water Ambient Monitoring Program (SWAMP) is another data repository designed and built by MLML. This DMS captures ambient water quality data throughout the state and has the most rigorous data quality checks and controls in the state. The consistency of data that comes out of the SWAMP is unmatched by any system in the state. However, as a result of

these quality control checks, it takes as much as a year before submitted data may be retrieved for use. The underlying infrastructure (database design, templates and QA/QC protocols) provides some valuable insights for the development of a highly controlled, centralized DMS.

However, for purposes of a multi-user DMS as envisioned by the SD-IRWM, this is not likely to provide significant value. Because SWAMP data is distributed to the public via CEDEN, viewing the data is limited to those with direct access. The complex database structure is not user friendly for those without reasonably advanced database skills. The rigorously defined SWAMP database structure make it a poor choice as a model for a regional DMS and would require participating organizations to substantially modify their own data workflows to comply. Additionally, maintenance for such a system requires significant and ongoing staff oversight, commitment of time, and continued financial resources.

Web link: www.waterboards.ca.gov/water_issues/programs/swamp

B.3: California Integrated Water Quality System

The California Integrated Water Quality System (CIWQS) is used by the State and Regional Water Quality Control Boards to pinpoint areas of environmental interest, manage permits and other orders, track inspections, and document violations and enforcement activities. CIWQS also allows discharge permittees to submit information online (within certain programs) and makes data available to the public through reporting.

Individual National Pollutant Discharge Elimination System (NPDES) permit holders and enrollees under the statewide general sanitary sewer overflow (SSO) order submit data to CIWQS. Enrollees under the statewide general industrial stormwater permit can submit annual reports to the Stormwater Multi-Application, Reporting, and Tracking System (SMARTS). CIWQS replaced the Storm Water Annual Reporting Module (SWARM). The Water Boards have developed several reports to display CIWQS regulatory data. For those with access to the system, these can be accessed through the “reports” page of the website. Since CIWQS provides limited access for registered users, it is of limited direct value to those outside of the SWRCB. In the future, a subset of monitoring data submitted to CIWQS will be ported to CEDEN for public access.

Web link: www.waterboards.ca.gov/ciwqs

B.4: GeoTracker

GeoTracker is the Water Boards' DMS for managing sites that impact groundwater, especially those that require groundwater cleanup (e.g. underground storage tanks, Department of Defense, site cleanup program). Permitted facilities such as operating underground storage tanks and land disposal sites are also included. It provides both public and secure portals to retrieve records and view integrated data sets. Data sets from multiple State Water Board programs and other agencies can be viewed through an easy-to-use Google maps GIS interface. The interface allows users to view data in relation to streets/roads, satellite imagery, and terrain map views. Other sites that affect groundwater quality and wells along with other beneficial uses that may be affected can also be viewed. In this sense it represents a partially federated data system approach to data management.

GeoTracker reports help SWRCB and US EPA staff monitor data throughout the State. It provides most of the publically available data for given sites through its Document Manager Module enabling regulators within the State Water Board, Regional Water Boards, and local agencies to oversee and track project activities, compliance responses, milestone tracking, land use controls, and risk to water quality tracking. GeoTracker is the largest receiving system nationally for analytical and field data for cleanup sites. The web application is used for secure reporting of laboratory data, field measurement data, documents and reports. GeoTracker is a targeted system with a focus on groundwater monitoring and does not serve as a direct model for the envisioned DMS. However the approaches for providing public/private access control and a generalized map interface represent some features which may be emulated.

Web link: geotracker.waterboards.ca.gov/

B.5: Beachwatch

Beachwatch is a State Board funded repository that houses the coastal monitoring data required by AB411 for Beach Water Quality monitoring. For many years the system has been managed and maintained by SCCWRP using the open source, PostgreSQL database to store and manage data and reporting. Participating agencies submit data on a weekly or monthly basis via the Microsoft Access application used to manage data at the local agency level (participating

County Health Departments). The Access application also provides a variety of data reports for local use. Data are used internally by State Water Board staff and annually submitted to US EPA for statewide assessment. Public access to the data is through transfers to CEDEN and, in processed form, via the California Water Quality Monitoring Council's "Safe to Swim" portal. Beachwatch serves as a good example of a system designed for ease of use and submission, but lacks broad scope as required by the SD-IRWM. Like CEDEN, the system is designed as a centralized DMS to which all participating organizations submit their data using defined data templates. It represents a capability to capture a small and well defined set of data (primarily bacteria analytes) at pre-determined monitoring locations that remain consistent. While jurisdictions in the San Diego region currently contribute to this database, it provides little in the way of an appropriate example for a system required by the SD-IRWM uses.

Web links: www.sccwrp.org/ResearchAreas/DataManagement/BeachWatch.aspx
www.mywaterquality.ca.gov/safe_to_swim/

B.6: California Geoportal

The California Geoportal is an example of a federated DMS and provides easy and convenient ways to search, discover and use geospatial data resources. A primary goal of the California Geoportal is to improve access to California's geographic data portfolio, and expand the creative use of those data resources. Geoportal was developed using open data standards and technology to maximize flexibility and compatibility across organizations. Data represented in the Geoportal are not stored in a centralized DMS. Instead, the metadata for each dataset is captured while the underlying data are maintained by the originating organization. The role of the California Geoportal is to increase information transparency; to create an open environment for accessing important government derived geographic data. The resulting benefits will be to encourage information sharing, and to promote efficiency and effectiveness in providing timely and accurate geographic information for better and more informed decision making. While the portal is not specific to water data, a subsection titled the "California Coastal Geoportal" provides access to data and custom applications with the coastal community in mind. Software architecture and user interface design of the California Geoportal and the Ocean Data Portal provide useful

examples of how a federated SD-IRWM DMS might function effectively. Currently, the California Geoportal is managed and maintained within the State Infrastructure (currently the California Technology Agency). However, recent changes in staffing and planning for the future may lead to the infrastructure being moved to another agency. Without a fuller understanding of these evolving details, the State Geoportal is not recommended as a near-term option upon which to build.

Web link: http://gis.ca.gov/california-geoportal/

B.7: West Coast Governor's Alliance Ocean Data Portal

The West Coast Ocean Data Portal (ODP) is a project of the West Coast Governors Alliance on Ocean Health (WCGA) to increase discovery and connectivity of ocean and coastal data and people to better inform regional resource management, policy development, and ocean planning. The Portal will inform priority west coast ocean issues such as tracking sources and patterns of marine debris, adaptation to sea-level rise, understanding impacts of ocean acidification on our coasts, and marine planning.

The Portal links existing data systems together to provide an easy to use gateway to discover ocean and coastal data. Coastal decision-makers, researchers, and stakeholders use the Portal to access data and decision-support tools they need to understand and address high-priority regional issues. The Portal is funded through the National Oceanic and Atmospheric Administration's (NOAA) Regional Ocean Partnership grant. The Portal includes tools to help coastal managers track marine debris, prioritize clean ups, and advocate for policies to reduce the impact of trash on our beaches. Like the California Geoportal, the ODP is built on open technology standards as a federated data system. In fact the California Coastal Geoportal is one of many data catalogs included in the ODP catalog, making this an excellent example of how multiple, federated DMS can interact with one another. The technical infrastructure is maintained at the University of San Diego Supercomputer Center. However, as current grant funding expires in fall 2015, changes in staffing and planning for the future may lead to the infrastructure being moved to another location. Without a fuller understanding of these evolving details, the ODP is not recommended as a near-term option upon which to build. Nonetheless,

the underlying software infrastructure is open source and could be duplicated and modified for a SD-IRWM DMS. Doing so would require identification of a stable host and technical capabilities to install, modify and manage such a system.

Web link: portal.westcoastoceans.org

B.8: Water Quality Portal

The Water Quality Portal (WQP) is a cooperative service sponsored by the United States Geological Survey (USGS), the US EPA, and the National Water Quality Monitoring Council (NWQMC). It serves data collected by over 400 state, federal, tribal, and local agencies. Being a national system, the WQP is an excellent example of a large federated DMS. The portal could be leveraged as a resource for data already captured through national programs with nearly 200,000 records for the San Diego region already included. It is structurally similar to CEDEN (which eventually will pass much of its data to the WQP) and could prove a valuable source of supplementary federated data source for the system eventually developed by the SD-IRWM. As a Federal government system, little opportunity to influence future development of the WQP to meet SD-IRWM needs exists. Regional data can be submitted to the WQP, but the additional effort required to format and submit compliant into the WQP and to later locate these data when needed would result in an overly complex alternative. Developing a regional system specifically designed with regional users' requirements in mind has a much greater likelihood of success.

Web link: www.waterqualitydata.us

B.9: California Data Exchange Center

The California Data Exchange Center (CDEC) installs, maintains, and operates a statewide hydrologic data collection network, including automatic snow reporting gages for the California Cooperative Snow Surveys Program and precipitation and river stage sensors for the flood forecasting program. A centralized system provides access to store and process real-time hydrologic information gathered by various cooperators throughout the State. This information is disseminated to support forecasting and flood operations activities and to meet the data needs of collaborators, public and private agencies, news media, and the public. As a centralized system, aggregating data from multiple real-time sensors and sources, CDEC serves a functionally

different purpose than identified as priority needs for the SD-IRWM DMS. Thus the CDEC architecture, designed to manage real-time data is not an appropriate model for the SD-IRWM DMS. However, given additional tasks identified by stakeholders relating to flood management, drought and climate change would benefit from the integration of CDEC data, this is a system that should remain under consideration for inclusion in the federated system in the future.

Web link: cdec.water.ca.gov/

B.10: Integrated Water Resources Information System (IWRIS)

IWRIS is an integrative data management tool for water resources data. The system provides a web-based GIS application providing users the ability to access, integrate, query, and visualize multiple sets of data. The IWRIS integrates data from multiple statewide databases including the DWR Water Data Library (WDL), California Data Exchange Center (CDEC), USGS streamflow, Local Groundwater Assistance Grants (AB303), and data from local agencies. The intent for IWRIS is to provide a single point of access for state-wide water resources information by integrating multi-disciplinary data in support of Integrated Regional Water Management (IRWM). The system improves efficiency in data discovery, download and delivery through a flexible, expandable, and user customization interface. This system incorporates many of the desired features including web-based GIS functionalities, integration with statewide databases as well as local agencies and could serve as a good model for a regional system.

Web link: www.water.ca.gov/iwris

Appendix C: Development Matrix Example

Table C-1: An illustrative example of how categorization of major components proposed for development of a comprehensive DMS can be divided into phases based on the required time and cost to complete different aspects of the system. Not included in this example, is the identified value of any given component. For example, if the priorities of the system determine tools to assess watershed health and sustainability in light of drought and climate change (defined here as high cost, mid-term timeframe) are essential, this may be prioritized over less expensive or more rapid development options.

	Near-Term	**Mid-Term**	**Long-Term**
Low Cost	Data Inputs: • Inclusion of Metadata Data Processing & Transmission Data Outputs: • Raw data outputs • Simple, aggregated data summaries Other: • Training for input/output usage	Data Inputs: • Mobile apps Data Processing & Transmission Data Outputs	Data Inputs: • Innovative ways to get data into local agency databases • Inclusion of historical data Data Processing & Transmission Data Outputs: • Public education • Mobile apps
Medium Cost	Data Inputs: • Programming "data pulls" from participating agencies Data Processing & Transmission: • Web-interface for data requests Data Outputs: • Data catalogue	Data Inputs Data Processing & Transmission Data Outputs: • Statistics/Trends • BMP analysis	Data Inputs: • Field checking data Data Processing & Transmission: • Data marts to other data systems (e.g. WQX) Data Outputs
High Cost	Data Inputs: • Visiting and working with each agency for data acquisition setup Data Processing & Transmission: • Programming standard data formatting/operability between participant data sets Data Outputs: • Summarized data reports/assessments	Data Inputs Data Processing & Transmission Data Outputs: • Drought & Climate Change Output Tools	Data Inputs Data Processing & Transmission Data Outputs: • Project Status Dashboard

Appendix D: Stakeholder Prioritized DMS Design Features

Voted Essential	DESIGN FEATURE
27	Standardized, straightforward data formats and metrics
25	Appropriate confidentiality and security protocols
21	Organized around clear questions and intended user audiences
21	Assurance of data reliability through up-to-date QA/QC protocols for data
21	Includes metadata
20	Ease of use
19	Meets regulatory standards, as applicable
18	Ability to tier and filter data
17	Clear governance structure
16	Appropriate and explicit, purpose-driven sharing protocols
16	Includes local control over data
15	Accommodates research, analysis, and mapping
15	Has a user training component
15	Index of where one can “pull” collected/available data, and corresponding
14	Can encompass land, ocean, and atmospheric water data
14	Supports long-term trend analysis
14	Allows for geospatial linking
12	Facilitates data harvesting across portals
12	Supports necessary turnaround time
9	Automated ability to summarize data
9	Supports planning at multiple scales
7	Visualization tools for summary data
6	Public website portal for discrete purposes
5	Dashboard
3	Cloned management/API
2	Includes future projections, not only archiving of past information

Appendix E: List of Acronyms

API	Application Program Interface
BMP	Best Management Practice
CDIP	Coastal Data Information Program
CEDEN	California Environmental Data Exchange Network
CIWQS	California Integrated Water Quality System
DMS	Data Management System
EPA	Environmental Protection Agency
FDGC	Federal Geographic Data Committee
FTP	File Transfer Protocol
GIS	Geographic Information System
HTML	HyperText Markup Language
IRWM	Integrated Regional Water Management
IWRIS	Integrated Water Resources Information System
ISO	International Standards Organization
M2M	Machine-to-machine
MLML	Moss Landing Marine Laboratories
MOU	Memorandum of Understanding
MRCD	Mission Resource Conservation District
NOAA	National Oceanic and Atmospheric Administration
NPDES	National Pollutant Discharge Elimination System
NWQMC	National Water Quality Monitoring Council
ODP	Ocean Data Portal
QA/QC	Quality Assurance/Quality Control
RCAC	Rural Community Assistance Corporation
SANDAG	San Diego Association of Governments
SanGIS	San Diego Geographic Information Source
SCCOOS	Southern California Coastal Ocean Observing System
SCCWRP	Southern California Coastal Water Research Project
SD-IRWM	San Diego Integrated Regional Water Management
SMARTS	Stormwater Multi-Application, Reporting, and Tracking System
SSO	Sanitary Sewer Overflow
SWAMP	Surface Water Ambient Monitoring Program
SWARM	Storm Water Annual Reporting Module
SWRCB	State Water Resources Control Board
USGS	United States Geological Survey
WCGA	West Coast Governors Alliance on Ocean Health
WQP	Water Quality Portal
WQX	EPA Water Quality eXchange
XML	EXtensible Markup Language