CALIFORNIA WATER QUALITY MONITORING COUNCIL

Fact Sheet: Data Federation

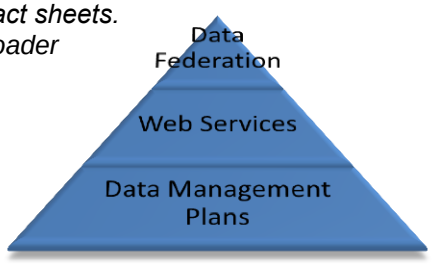
FALL | 2016
Version 1

Contact: Assistant Director Kristopher Jones, PhD
E-mail: Kristopher.jones@water.ca.gov
Phone: 916.376.9756

A COLLABORATION BETWEEN THE CALIFORNIA ENVIRONMENTAL PROTECTION AND NATURAL RESOURCES AGENCIES | www.MyWaterQuality.ca.gov

Leveraging Our Diversity of Environmental Data through Federation

In September 2015, the Delta Stewardship Council's white paper Enhancing the Vision for Managing California's Environmental Information offered a broadly shared vision for the advancement of environmental data sharing. Associated with the paper's recommendations are key concepts that we explain further in this series of fact sheets. The brief documents each address a different important mechanism within the envisioned broader data-sharing strategy: Data Management Plans, Web Services, and Data Federation.



We invite you to review the fact sheets, which range in sophistication from fairly basic (Data Management Plans) to moderately complicated (Web Services) to highly complex and coordinated (Data Federation). These subject technologies promise to modernize public data distribution, but a shared understanding of the underlying concepts, and of their respective value, is vital to realization of the vision.

What is federation?

- *Data federation* is a process where data is collected from distinct databases without copying the original data itself.
- A *federated database system is* a type of meta-database management system (DBMS), which transparently maps multiple autonomous database systems into a single federated database. The constituent databases are interconnected via a computer network and may be widely distributed geographically.

What are the benefits of data federation?

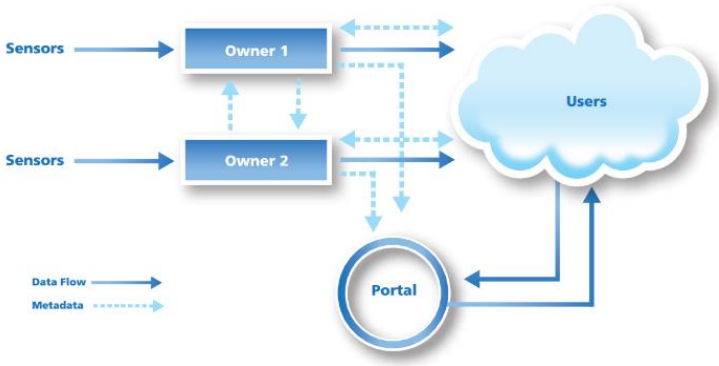
Understanding and management of California's natural resources are best supported through a coordinated approach to management of California's diverse, distributed and heterogeneous environmental datasets. As part of this approach, data federation will improve analyses though better integration and provide savings by better coordinating data collection.

How would federation help promote better decision-making?

Federation of California's environmental data will improve access to the necessary information for decision-making, ensuring that all available observations are used in assessment and planning efforts.

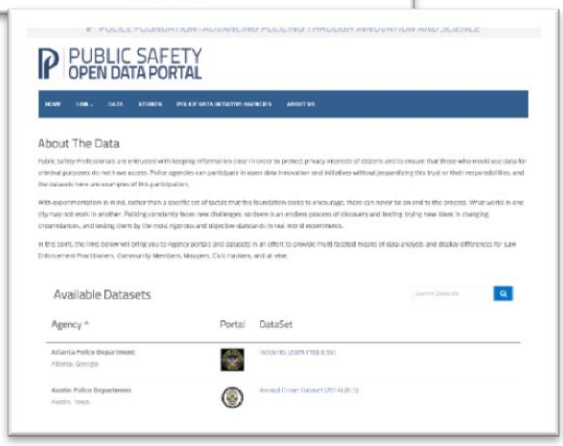
How will federation work?

- California's many environmental datasets ***continue*** to be stored in disparate, heterogeneous repositories;
- A federating body is established where data owners are ***expected*** to register their repositories;
- The federating body and data owners ***collaborate*** to implement protocols to support automated harvesting of descriptive metadata from repositories;
- Data owners are ***encouraged*** to make data available for ready, automated access via web services, and to include service links in their harvested metadata;
- The federating body maintains a searchable ***catalog*** based on the harvested metadata;
- The federating body or other organizations provide data discovery (search) and, where available, data access from singular data portals or gateways.



What are some examples of federation?

- **Water Quality Portal.** The US Geological Survey (USGS), U.S. Department of Agriculture (USDA), and the U.S. Environmental Protection Agency (USEPA) have collaborated on the Water Quality Portal (WQP) (http://www.waterqualitydata.us/) for federated discovery and access to water quality data. The underlying datasets – the USGS National Water Information System, the USEPA Storage and Retrieval Data Warehouse, and the USDA ARS Sustaining The Earth's Watersheds - Agricultural Research Database System – are very different in format and access methods as well as nomenclature. WQP provides uniform, transparent access to data in these repositories and bridges the differences in parameter naming, but data are accessed from a centralized copy of the datasets – potentially resulting in data synchronization issues.
- **Public Safety Open Data Portal.** This portal (http://publicsafetydataportal.org/) is intended to provide a single national portal for open datasets on law enforcement and public safety. By improving transparency, accountability and legitimacy in public safety operations, the portal enhances law enforcement agency relations with the communities served and protected. Unified, consistent access is provided to over 130 datasets from 57 participating agencies, and example stories are presented to encourage through illustration innovative analysis and visualization of the data.



What are the biggest obstacles to federation?

The biggest obstacles to federation include:

- **Lack of understanding.** Data owners and data users are often not familiar with the term or the concept,
- **Fear of costs.** Federation is perceived as an expensive undertaking.
- **Protectiveness of data ownership**. Data owners are unwilling to make data freely available for a variety of reasons, not all of which are reasonable.
- **Uncertainty about implementation.** Data owners and users don't see a clear path forward to implementing a federated system encompassing California's environmental data.

As a decision-maker, what can I do to support it?

Decision-makers can best support the advancement of federation through:

- **Educating**. You can help your peers and your organizations better understand the concept of federation by knowing where to direct them for more information.
- **Advocating for Scalable Pilots**. Federation is innately scalable, and any initial proof of concept can be based on connecting a single key well-managed dataset to a catalog and data access. The pilot effort can be built around key use cases and workflows identified by California's environmental information stakeholders and designed for leveraged scalability in order to support anticipated incremental implementation.
- **Encouraging use of standards**. Implementation depends on adherence by all participating organizations to agreed-on standards for interoperability. Your organization may participate in the identification and selection of standards, and must agree to use the standards.

Virtuous Cycle

Successful federation is highly dependent on broad participation by data providers, but as those providers contribute, they will appreciate the positive feedback loop of greater use of their own repositories. As data owners increasingly publicize and publish data holdings in support of federation, we also envision the spontaneous development of unanticipated, novel tools from unexpected parties that deliver new insights.