

Enhancing the Vision for Managing California's Environmental Information

"We can still fight about water. But let's at least understand what we're really fighting about - and let's do it in our own California way, using the best data and interactive social technologies."

- Chacha Sikes, "For a New California Water Atlas," *Boom* (Fall 2013)

"The state leverages a secure and reliable technology infrastructure and shared services. This requires an information technology infrastructure that leverages the advantages of cloud computing, robust shared services, and reusable components."

- California Department of Technology, Goal 3 of *Strategic Plan 2014 Update*

The challenges facing California's Delta are not new, but they are underscored by the urgencies of the moment. Pressing matters such as climate change, an aging water infrastructure, and a deep drought force all Californians to recognize the fragile balance between our natural resources and their competing uses across the state but nowhere so poignantly as in the Delta. There, natural resource decisions must accommodate both the rate of environmental change and the fast pace of intervention efforts.

To be effective under these circumstances, agencies, organizations, and public interests must make use of today's innovations with an equally fast rate of adoption. And they must share the information and data produced through these innovations across conventional jurisdictions and agency boundaries. In short, the need for transparency and sharing of data demands an open community of science with interoperability standards, state-of-the-art data exchange and access tools. Decision makers, analysts, and public interests recognize the need for such a sharing initiative. With the world's most advanced technology resources located here in California, how do we apply our intellectual resources to this problem, to ensure that sharing information becomes the norm for natural resource management rather than the exception? And how do we best complement and build upon our past technology investments?

This document seeks to characterize the opportunities for contemporary data management and to envision how we might identify the obstacles before us while charting a path toward an open community of information sharing.

For this iteration of the document, we have crafted a deliberately open-ended scaffolding upon which we can build a shared vision: a road map, so to speak, for leveraging and sustaining existing efforts while pursuing common priorities. We invite Summit participants to “fill in the blanks” by responding to the questions posed in the tracks. Out of the Summit should emerge a collective and refined version that might help to guide future work to integrate, share, present, and analyze data concerning the California’s environmental data.

Definitions

Technology has become so interwoven into the fabric of our professional and personal lives that the terms we use to describe its many components and concepts have loosened to the point of ambiguity. Part of what we lack when we encourage greater data-sharing is a clear understanding of what “data-sharing” means, both in and of itself and its inherent benefits to its associated partners. For the purposes of this document, we use the following terms to describe the areas of data management most salient to the promotion of more effective data sharing, analysis, and reporting.

- **Data.** In the context of our discussion, data are any information derived from research, monitoring, observations, analysis, aggregation, complex models, scientific calculation or anything deriving from such material. They might be captured in various formats and housed in various places – ranging from images, tables, or charts stored in a computer to paper-based assessments stored in a filing cabinet.
- **Database.** The formal container for data that have been assembled together is called a “database.” Most databases today are known by the acronym RDBMS which is short for “relational database management system,” meaning that the data contained within each database are held together in a web of relationships in order to reduce redundancy and improve information access efficiency. Any given RDBMS has distinctive characteristics including schema, tables, and records. In today’s world of web services and data-sharing, the databases remain important but often located remotely from where the data might be eventually displayed on various web sites or presentation/visualization tools.
- **Data-sharing.** This refers to the practice of making data available for transfer to digital systems via commonly understood protocols. What qualified as data-sharing in the past – the manual exchange of paper-based or even digital files – no longer meets the minimum standard for effective data-sharing. We are challenged to keep pace with more dynamic and automated forms of sharing that are increasingly understood as its definition.
- **Business Model.** In our paper, we use the term “business model” to refer to the systems of funding and administrative support that ensure the continuity of technological innovations over time. Discontinuous funding under failed business models disrupts data systems development and leads to technological cul-de-sacs.
- **Data Integration.** The process of collecting data from disparate sources and making it meaningfully available to users is called data integration. Before data can be harmonized and placed into conversation with other data it must often be checked for quality, transformed, and converted into a common format. All of these practices account for phases work commonly associated with data integration.
- **Data Library.** A cataloging system that affords research access to data via its description or metadata is called a “data library.” This term includes its associated staff of technicians and librarians. A data catalog is usually a prominent feature of a data library.

Framework Questions

General Questions Addressed to Everyone

We begin by posing some general questions that apply to audiences irrespective of specialization. When we consider the “big picture” for fostering stronger data sharing across the region, we must address some overarching concerns.

- What are the primary obstacles to data coordination, sharing, and interoperability?
- What has been the history of data sharing from the perspective of your organization? To what extent do legacy techniques still influence the practice or expectations surrounding data sharing?
- What has your organization done to foster data sharing?
- In your experience, are things progressing in the area of data sharing? If things are progressing at different speeds across your domain, why is this?
- What are the biggest technical strides that have been instituted to foster greater sharing?
- How do we make data mean something more to decision makers?

Business Models

Tools to facilitate data collection, integration, presentation and sharing are not without cost. Devising successful business models to support sustainable technology measures becomes all the more important as we seek to reduce duplicative systems and pursue cost-sharing opportunities. How do we ensure that our approach is supported by sound economic drivers? We appeal to leaders in the business community, among non-governmental organizations, and within public agencies to share their successful techniques for sustainable business model development.

- What are some successful funding mechanisms?
- What is the true value of good data? How do you measure it?
- How do you foster stakeholder buy-in?
- How important/influential is the development of open source software to your planning and investment strategy?
- Are you involved in an open data movements, standards, or initiatives? On what scale are they occurring? How would you characterize the drivers for these movements?
- Given a very competitive technology environment, what role does collaboration play in ensuring success in today's marketplace of ideas?
- What process do you adopt to ensure that your fiscal priorities align with your information management needs?
- How do you encourage data sharing among your clients, constituents, or staff? Is this an important measurement of success for you?

Data Integration

The last ten years have witnessed huge strides in data integration technology, challenging us to thoroughly reimagine how to bring data together from diverse data sources. With the advent of web services essentially disaggregating database from data, we can perform dynamic data mashups and analyses that were either prohibitively expensive or impossible when using yesterday's tools. The new opportunities lay before us. Where are the greatest hindrances to success in this critical area? What are the highest priorities to address as we proceed?

Often the devil is in the details. When we engage in broad-scale data integration, we must ask not only why the data integration is being done and what the objectives and deliverables should be, but from what systems will the data be sourced? Are all the data available to fulfill the requirements? What are the business rules? We must analyze the source systems. What are the options of extracting the data from the systems (update notification, incremental extracts, full extracts)? What is an ideal or required/available frequency of the extracts?

The challenges can sometimes seem great, but the opportunities are truly monumental.

- What technical measures have you instituted to foster sharing?
- What are the biggest challenges you've encountered in ensuring data interoperability?
- How do you ensure that a dataset is properly attributed to its source even as it is transformed, aggregated, or otherwise summarized?
- How have web services -- DaaS, web mapping services, etc -- changed the way you think about your data? What are some associated challenges with meeting demand?
- How has the concept of "data federation" influenced the ways that you think about data stewardship?

Data Library Development

The devil is often in the details when we are facing technical hurdles. Often what seems to be a small problem -- an old but still vital dataset locked in a legacy system, for instance -- can present an insurmountable obstacle for a resource manager who needs timely access to that information. Perhaps the data necessary to inform decisions is housed in repositories across various agencies and organizations. Data discovery and availability are primary initiatives being addressed by numerous parallel efforts. As we optimize our work in this regard, we must ensure that we serve the collective good.

- What are the biggest, most critical data gaps?
- In your opinion, what are the highest priority datasets to address today's ongoing and emerging concerns?

- How do you determine whether a “legacy” dataset should be made available to a target group? How do you facilitate access?
- What tools or methods do you use to locate data? How are data catalogs of value to you or your clients?
- Can you point to a case study of metadata standards advancing a project?

Data Analysis, Mining and Presentation Tools

In the field of information management, there is nothing sexier than a new, compelling data visualization tool. Today’s tools employ animations, powerful imagery, and insightful charts to convey information critical to decision-making. Often these data slice through otherwise opaque layers of tangential information to derive the kernel of salience. As we consider what these tools offer, however, do we see patterns governing their success in the context of environmental information management? Can we promote an approach or set of questions to influence the consideration of such tools, so that we are not led astray by the shiniest object? The visualization might be the tip of the data iceberg, but it can also indicate the wealth of information under the surface. Presentation tools often serve as the gateway to appreciate the many other deep and sustained investments that have delivered us to the present moment. As such we can consider how the presentation of data can be leveraged to demonstrate not only the efficacy of key analyses but also the value of data library development, data integration, and sound business models.

- What are some tools that represent significant strides in making meaning out of data?
- How have you found that tools can make data more understandable to your target audience(s)?
- What is the value of interactivity in managing and presenting data?
- How important is data integration to today’s tools?
- How has the advent and proliferation of web services altered the landscape of tool functionality and features?
- To what extent is good design an important factor in fostering information consumption?
- What is the next frontier in data presentation tools? What is the biggest surprise that our stakeholders will encounter?

Towards a Coherent Data-Sharing Framework

Looking to the next five years, we must ensure that California can realize its potential as a beacon of dynamic data integration in the service of environmental decision-making. Funding may be a challenge, but we must adopt successful business models to advance our efforts. The data sources, with their disparate standards and quality-control measures, may present obstacles to integration, but today’s tools afford us an unprecedented

opportunity to overcome these hurdles. Meanwhile, business intelligence technology can yield reliable insights, separating the wheat from the chaff, by showing us never before seen visual data analyses. In essence, environmental data management must fully respond to the exigencies of environmental resource management.

We foresee measurable benefits facilitated through a concerted effort to adopt consensus-based strategies, standards, and objectives. Broad participation across agencies, non-profit organizations, policy-makers, and private industry is critical to success. Working within a virtual framework that is transparent to all, the data-producing, data-stewardship and data-consuming communities must cooperate to produce a durable data-sharing framework.

Of course, if a framework is to be durable and transparent, it must also be adaptive and extensible, responsive to change, especially since it must keep pace with technologies many innovations. To that end, our virtual five-year roadmap will pursue the following goals:

- To establish reflexive change-control mechanisms to preserve the data-sharing framework's integrity while also encouraging its currency and salience;
- To define the acceptable minimum distribution obligations for all data produced with public funds, thereby ensuring that public data be made readily available in forms that can be consumed by widely used information systems;
- To establish data-format standards that facilitate data integration in ways that support the current needs of environmental information-sharing and decision-making;
- To establish data-quality standards that help to codify levels of reliability across various use cases and contexts;
- To promote business models and to open channels to funding opportunities that can offer sustainability across financial volatility or budget change requests;
- To ensure proper attribution of data sources – the provenance of the data – for the encouragement of data-sharing without concern about accidental or deliberate expropriation;
- To foster cooperation across agencies who share common interests in data sources, technology tools, or resource management outcomes.

This framework will balance its efforts to describe with the impulse to prescribe. It will seek to achieve a feasible, reasonable level of specificity that provides enough detail without excluding willing partners in the endeavor. To foster effective cooperation, our framework must be inclusive, open, and transparent in its definitions, scope, and purpose. By complementing current initiatives, policies, and data systems, the framework will gain maximum effect while also extending those established assets in new and necessary ways.