CALIFORNIA WATER QUALITY MONITORING COUNCIL ▪ DATA MANAGEMENT WORKGROUP

Fact Sheet: Glossary

DECEMBER | 2016

Version 1

Contact: Tony Hale, PhD
E-mail: tonyh@sfei.org
Phone: 510-746-7381

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

Glossary for Information Management Discussions

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 partner agencies and the public. For the purpose of our discussions, 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.

Definitions

- **Business Model.** In our discussion, 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. Sustainability is key, for discontinuous funding under failed business models disrupts data systems development and leads to technological cul-de-sacs.
- **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.
- **Data Management Plans.** Data Management Plans are documents that capture essential information from researchers about their datasets to ensure alignment between scientific and data management goals.
- **Data Federation.** The process of combining data from disparate repositories, creating a unified whole that permits cross-repository querying, comparisons, integration, and re-distribution is called "data federation." It is among the highest forms of data virtualization and usually occurs across heterogeneous data owned by multiple organizations.
- **Data Integration.** The process of collecting data from disparate data 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.
- **Data Owners.** The data owner is the authority identified with the creation of the data. The owner may not be the data steward, or they may be one in the same. Often a repository might serve as a steward for purposes of distribution and visualization, but another organization retains ownership of the actual data.
- **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.

- **Data Steward.** A data steward ensures that the data it oversees is managed according to set governance standards. The steward can be a person or organization who would coordinate with the appropriate authority to apply the appropriate set of formatting, quality control, and routine management choices to the data in its charge.
- **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 RDMBS 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.
- **Governance.** In the context of data management, governance can take the form of written agreements such as Service Level Agreements (SLAs) or Memoranda of Understanding (MOUs) that establish expectations or mandates for service related to data management. However, governance also pertains to a workgroup or committee entrusted to make ongoing decisions with respect to data collection, data utilization, and data sharing according to a strategic framework.
- **Metadata.** In short, metadata is information about data. It is used to describe essential details about a dataset -- its origin, owner, purpose, format, provenance, allowed uses, quality characteristics, etc -- and as such it becomes all the more essential when disparate datasets are integrated into larger repositories. Metadata form the critical map to ensure data integrity and reliable meaning.
- **Open Data.** Open Data is a term used to describe both the *legal* and *technical* availability of information, used by anyone for any purpose, often while preserving attribution.
- **Web Services.** Web Services comprise technical tools that afford machine-to-machine communication using standards-based protocols for a cost-efficient transfer of information.

Prepared By:

Tony Hale, PhD
Co-Chair, Data Management Workgroup
Program Director, Environmental Informatics
San Francisco Estuary Institute