A Presentation to the California Estuary Monitoring Workgroup

Data Readiness Assessment

Identifying and removing obstacles to data sharing

Tony Hale, PhD
Co-Chair, Data Management Workgroup
Sept 13, 2017

Data Readiness Assessment:

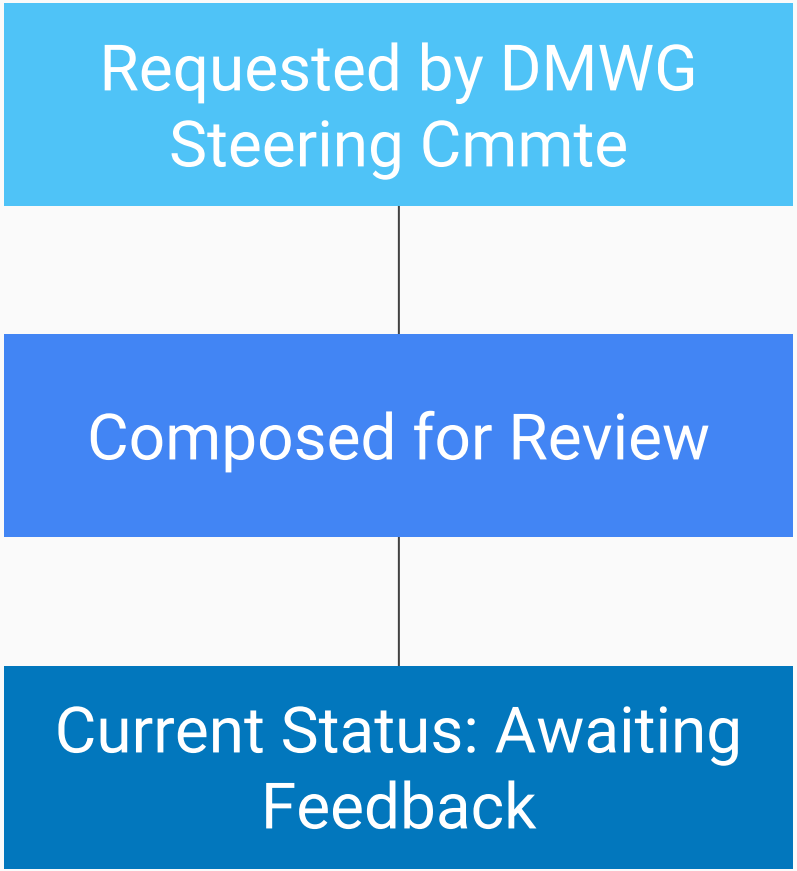
A tool used to assess the readiness of an organization to evaluate, design and implement an Open Data initiative.

(Adapted from http://opendatatoolkit.worldbank.org/en/odra.html)

Background

Adapted from guidelines for the development of an Open Data Handbook, composed by Tony Hale, Tad Slawecki, and Eric Maas of the Data Management Workgroup:

http://goo.gl/VgyL2b



Broader Purpose and Goals

- Assess current state of data openness for the department/program/agency
- Assess the organizational readiness (policy, organizational, resource capacity) to pursue an open data strategy
- Recommended procedures to advance the degree of openness for the organization’s data.

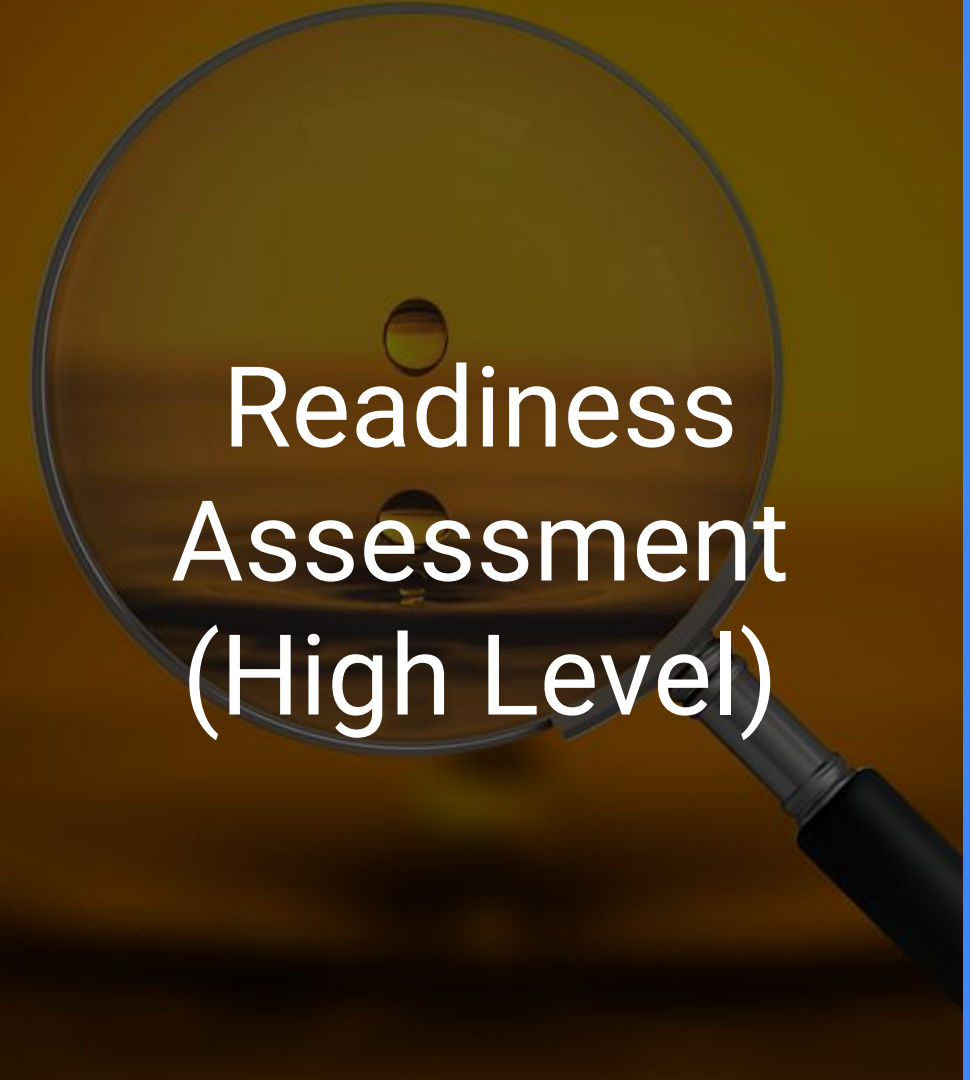
Key Concepts

Openness, when it comes to “open data,” is a spectrum, not a binary

State has legal frameworks and mandates that affirm transparency

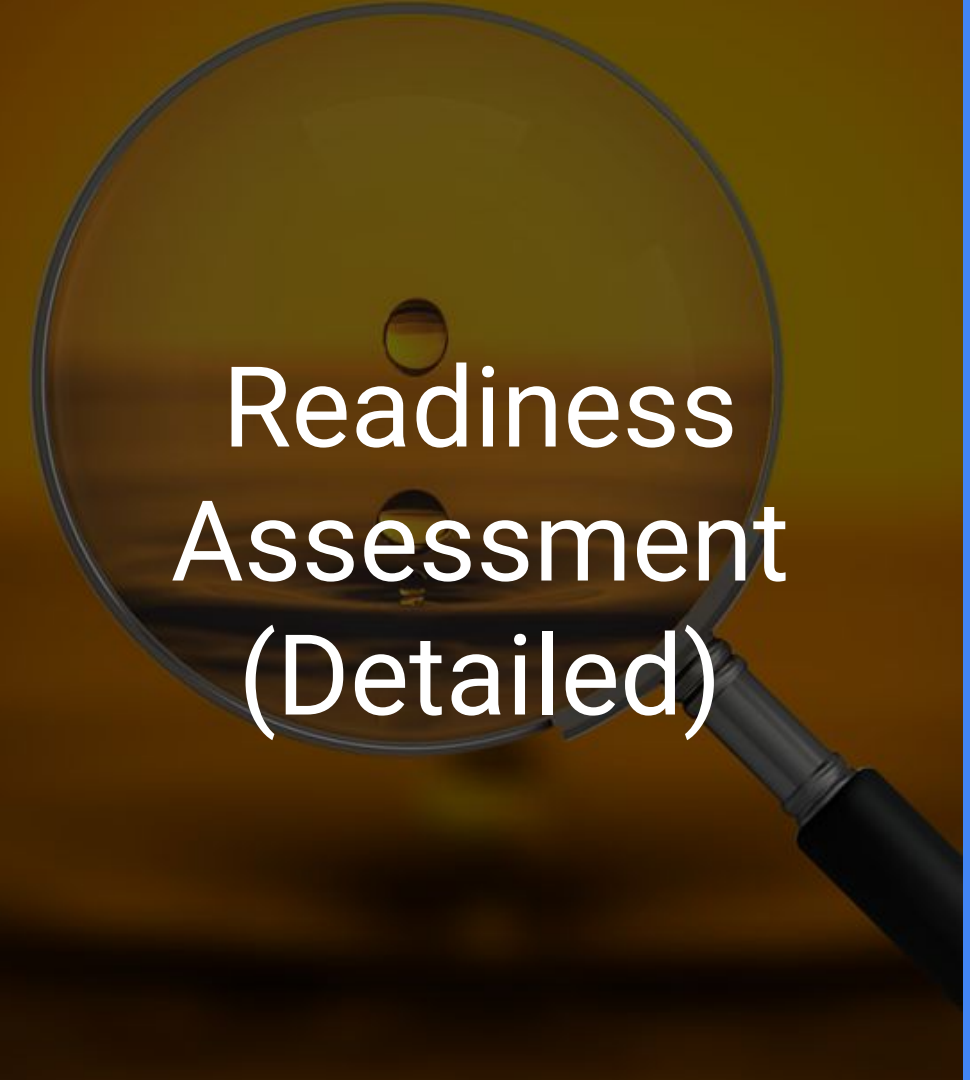
Science favors openness and sharing

An organization can be a program, agency, or dept



Readiness Assessment (High Level)

- Review policies / legal frameworks for work of the organization
- Review data management practices
- Identify gaps between policies and practices



Readiness Assessment (Detailed)

- Review data management infrastructure
- Review individual datasets
- Assess alignment of reporting requirements and data salience

Data Maturity Framework

Center for Data Science & Public Policy
THE UNIVERSITY OF CHICAGO

Data Maturity Framework
Data and Tech Readiness Scorecard

Category	Area	Lagging	Basic	Advanced	Leading
How is Data Stored	Accessibility	Only accessible within the application where it is collected	Can be accessible outside the application but proprietary format, requiring specialized analysis software	All machine readable in standard open format (CSV, JSON, XML, database)	All machine readable in standard open format and available through an API
	Storage	Paper	PDFs or Images	Text Files	Databases
	Integration	Data sits in the source systems	Data is exported occasionally and integrated in ad hoc manner	Central data warehouse - realtime aggregation and linking (Automatic)	External data also integrated
What is Collected?	Relevance and Sufficiency	The data you are collecting on subjects of interest is irrelevant to the problem you want to solve: ie you want to do predict which students need extra support to graduate on-time but don't have data on graduation outcomes	Some of the data you have is relevant, but it is insufficient because key fields are missing, ie no data on academic behavior or attendance history, etc.	You have data that is helpful and relevant for solving the problem but not sufficient to solve it well. ie you have yearly academic and demographic information but are missing extra-curricular activities, or interventions they were targeted with	You have all the relevant data about all the entities being analyzed and it's sufficient to solve the problem you are tackling
	Quality	Missing rows (people/address level entities missing in the data)	Missing columns (variables missing)	No missing data but errors in data collection such as typos	No missing data and no errors in data collection
	Collection Frequency	Once and never again	yearly	frequently	realtime
	Granularity	City level aggregates	Zipcode/Block level aggregates	Individual level (person or address) level data	Incident/Event level data
	History	No History Kept - old data is deleted	Historical data is stored but updates overwrite existing data	Historical data is stored and new data gets appended with timestamp, preserving old values	All history is kept and new data schema gets mapped to old schema so older data can be used
Other	Privacy	No privacy policy in place	no PII can be used for anything	ad-hoc approval process in place that allows selected PII data to be used for selected/approved projects	Software defined/controlled privacy protection that allows analytics to be done while preserving privacy based on predefined policies
	Documentation	no digital documentation or metadata: data exists but field descriptions or coded variables are not documented	data dictionary exists (variables and categories defined)	data dictionary plus full metadata available (including conditions under which the data were captured)	data dictionary plus full metadata available including collection assumptions, what's not collected, and potential biases

Related categories:

- Data Utilization
- Data interoperability
- Timeliness
- Attribution

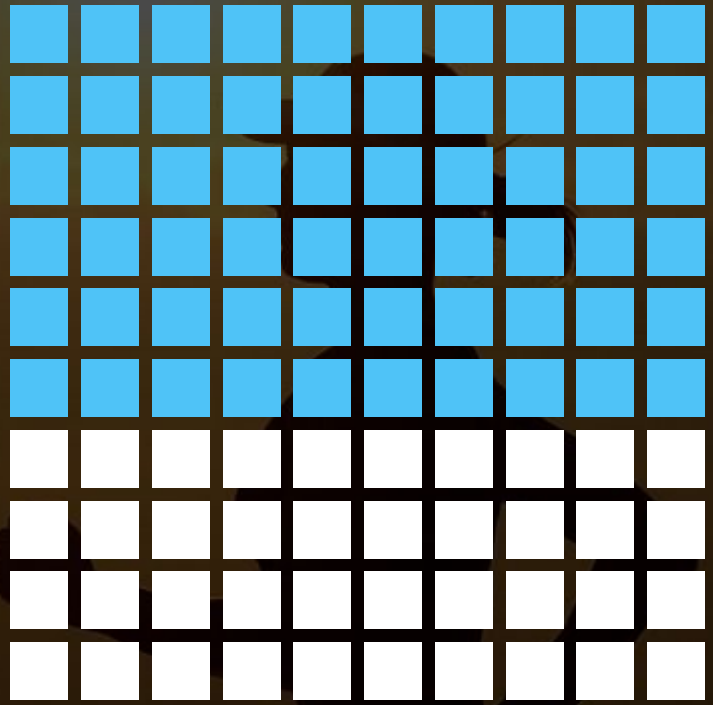
Likely Gaps to Fill

Identify the critical steps towards openness

- Data Documentation and Attribution
- Data Storage and Accessibility
- Data Utilization
- Data Interoperability
- Data Collection

Recommended Next Steps:

- Determine salience to CEMW
- Identify resources
- Solidify plan
- Execute plan (outreach, interviews, implementation for change)



Thank you.

Questions?

tonyh@sfei.org