Defining a Data Management Strategy for USGS Chesapeake Bay Studies

Cassandra Ladino, USGS

September 11, 2013

U.S. Department of the Interior
U.S. Geological Survey

CDI Webinar Series 2013

Key Questions

- **What data management means to different audiences?**
- **What are our next steps?**
- **What near term challenges are up a head?**

What Does Data MGMT Mean?

- **Data management doesn't mean the same thing to every one.**

- **The use cases for data management change when looking at a project from different perspectives… Scientist, Regional Team, Public Viewpoint.**

- **Data integration is the key aspect of changing viewpoints.**

Viewpoints & Levels of Data Integration

- **Scientist / Local Level – databases (Access, MySQL) to store observation records, seamless, queries multiple data source that have been integrated into a single data base.**

- **Regional Team / Connected Project Level – Independent data files, database (ScienceBase) to store desperate data files for a regional project team, allows archiving, allows focused group search and sharing**

- **Public / Disconnected Project Level – Multiple projects with multiple independent data sets located nationally in a single publically available database (ScienceBase), Relationships waiting to be discovered, previously unknown science connections made through key words, search, location.**

Viewpoints & Levels of Data Integration

- **Scientist / Local**

Question/ Query:

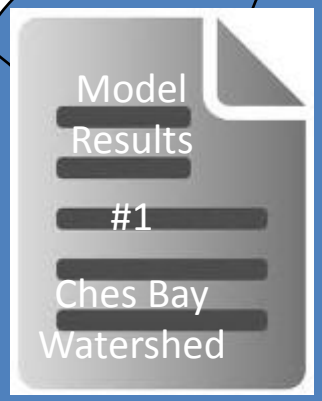
I have fish health information in the Difficult Run watershed, what are all the sample results from water quality and corresponding land use types?

Record Level Integration →

X,Y	Location Name	Lesions	Total Nitrogen	Total Phosphorous	Major Land use
xxx	xxxxxxxxxxxx	xxxxxxxxxx	xxxxxxxx	xxxxxxxxxx	xxxxxx

Viewpoints & Levels of Data Integration

- **Regional Team / Connected Project**

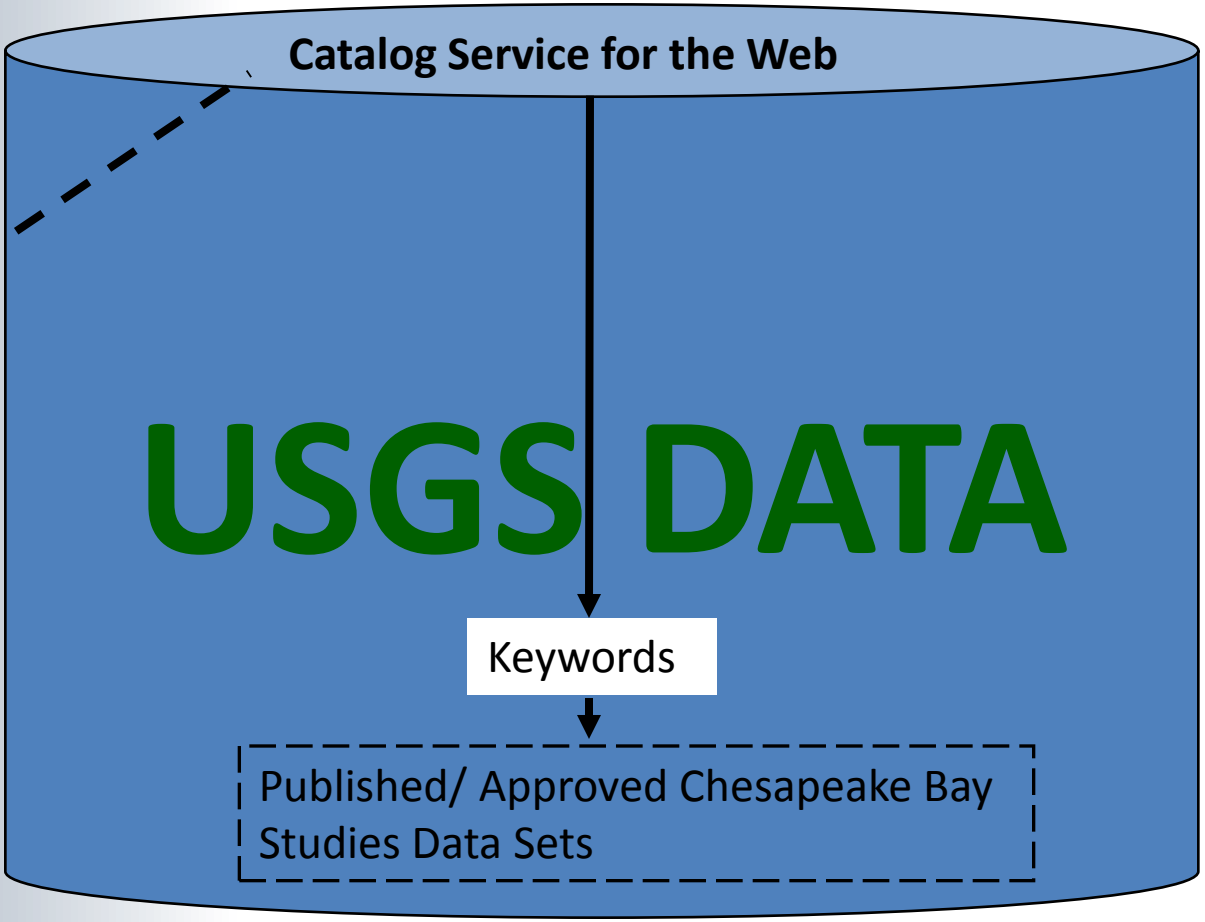


Chesapeake Bay Studies Scientific Results

USGS Network accessible storage location, data sets used for regional synthesis

Viewpoints & Levels of Data Integration

- **Public / Disconnected Project**



Where We Are and Where We Need to Go?

- **Currently practicing data management for public groups and partner agencies.**
  - **ScienceBase**
  - **Data Delivery**
  - **Web Service Provider**

- **FY13 & 14 set up plans and conceptual frameworks to implement data management practices based on using ScienceBase for regional teams.**

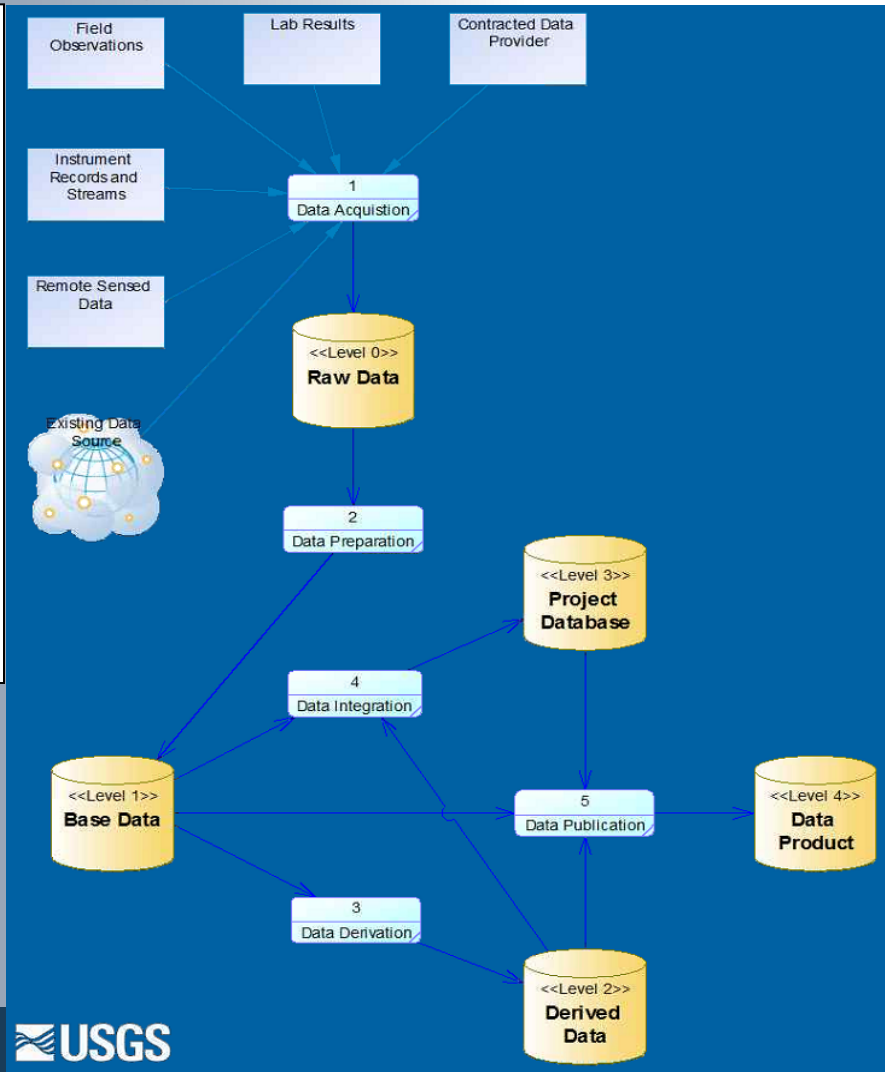
- **Future years develop a local infrastructure that allows record level integration for scientists.**

How Do We Get There?

- It helps to think of data in terms of the stages it goes through in a project cycle to set up infrastructure.
- Then we develop Data Management Plan frameworks to ensure consistency and allow integration.
- Create more scientist incentive and “buy in”.

Data Stages

- **Raw Data** – field data, lab results, unmodified
- **Base Data** - cleaned up and quality assured
- **Project Data** – Hybrid products, specialized multi dataset integration for special project needs
- **Derived Data** – Analysis derivatives (WQ Trends)
- **Data Product** – Final data set referenced in report (ScienceBase)



DMPf Part III: Research Data Management Data Maturity Model

Graphic courtesy of the USGS DMPf Implementation Working Group (Steve Tessler, TX WSC, AK WSC, CDI)

Data Stages at The Regional Team Level

- **Major Goal: preserve and share data amongst team members at each data stage using ScienceBase.**
- **Data Folders:**
  - **Raw Data: 1 snap shot of raw data before any manipulation**
  - **Base Data: 1 snap shot of approved raw data for project use**
  - **Project Data: many snap shots of special, one off, exploratory analysis**
  - **Derived Data: many snap shots of final data products associated with one project as they become approved and published**
  - **Data Products: not needed (public level only)**

Data Management Plan for Local Data Integration

- After we develop a functioning system at the Regional Team level, we will focus more on record level integration.
- This involves creating Data Management Plans based on work by the USGS DMPf Implementation Working Group (Steve Tessler, TX WSC, AK WSC, CDI).
- We will need to adapt the DMPF's to each type of USGS Chesapeake Bay studies project.
- The final plans will be used at the inception of every project.

Summary of Challenges

- Educate and encourage regional level data management in workflows and/or identify a data “gate keeper”.
- Create DMPf’s for each of the sciences (WQ, Land Cover Change, Fish Health, Sediment…).
- Explore logistics and cost of implementing a local data management system to support local level project work and data analysis.
  - Location
  - Infrastructure
  - Schema
  - Accessibility
- Educate and encourage scientists on how to publish data as an independent product in unison with report publishing.
- Evaluate business model options (distributed vs. central)

ScienceBase as a Regional Level Tool

- **Positives**
  - **Accessible by all**
  - **User Interface**
  - **Backup and Storage**
  - **Easy folder setup**
  - **Easy permissions control**
- **Difficulties**
  - **Reinforcing use over local network storage**
  - **Checking in/ Checking out data (versioning)**
  - **Providing web services for all data layers**
  - **Uploads of many large data sets**

Questions / Suggestions?

- **Contact Email:**
  - **Cassandra Ladino: ccladino@usgs.gov**

- **Web URL:**
  - **Ches Bay ScienceBase**
    - **https://www.sciencebase.gov/catalog/?community=USGS+Chesapeake+Bay**
  - **Defining a Data MGMT Strategy OFR**
    - **http://pubs.er.usgs.gov/publication/ofr20131005**
  - **Ches Bay CDI Wiki Page**
    - **https://my.usgs.gov/confluence/display/cdi/Chesapeake+Data+Management+System+Greenfield**