A Data Warehouse Approach To Managing and Querying Environmental Data

Ben Shorr
Assessment & Restoration Division | Spatial Data Branch
NOAA | National Ocean Service | Office of Response and Restoration

National Environmental Monitoring Conference 2015
Chicago, IL

The information in this presentation reflects the views of the author, and does not necessarily reflect the official positions or policies of NOAA or the Department of Commerce.
Presentation Overview

Background on NRDA and Data Sources
Data sources, standards and needs

Data Warehouse and Common Data Models
Overview of flexible/scalable framework; Data Models and standards; related information/data. Information Management Portal and DIVER

Data Query and Delivery
Requirements that drive development of data discovery, query, reporting and export tools- supporting scientific analysis and reports for a Damage Assessment case
Natural Resource Damage Assessment (NRDA)

Preassessment Screen (exposure assessment)

Restoration Planning

- Field Studies
- Data Evaluation
- Modeling
- Injury
- Quantification

Restoration Implementation
Ecosystem Overview

ECOSYSTEM

Wetlands
Saltwater marshes provide spawning habitat for forage fish and other species, refuge area for juvenile fish and birds, nursery area for crab, shrimp, oysters.

Nearshore Benthos
Oyster beds, seagrass beds, and mudflats. Production area for crabs, shrimp, fish.

Photic Zone
Layer of Gulf waters that sunlight can penetrate ~ generally 200-300 feet from the surface.

Top Predators
Marine mammals, tuna, birds.

Deep Benthos
Cold seep organisms, deep water coral, contribute system-wide biodiversity.

POTENTIAL OIL IMPACT

Wetlands
Oiled, degraded or eroding marsh may reduce productivity.

Nearshore Benthos
Tar mats and oil in sediments may reduce benthic productivity and affect food web.

Photic Zone
Surface and dispersed oil may affect base of food web; larval fish are particularly sensitive to effects of oiling.

Top Predators
Predators may be affected by degradation of food web and by direct health effects from oiling.

Deep Benthos
Destruction of long-lived deep corals may reduce biodiversity and deep ocean productivity.

©2011 NOAA. Illustration by Kate Sweeney
Natural Resource Damage Assessment Data

• Vast amount of NRDA and Response data collected under different authorities, different formats, different destinations and management

• NOAA OR&R and partners were part of key NRDA and Response data streams early and adopted/adapted/created data flow processes.
File Collections
(aka NOAANRDA website)
## Signal to Noise

### Preliminary Measures and Dimensions; Priority of "Questions to Answer"; Data Sources that can be used to Answer Questions

<table>
<thead>
<tr>
<th>Priority of Measure Implementation</th>
<th>Measures (The Question to Answer)</th>
<th>Dimensions (Ways to slice the Question)</th>
<th>Current Data Warehouse Pulling Data From:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time</td>
<td>Spatial</td>
<td>Depth</td>
</tr>
<tr>
<td>1 Contaminant Lab Results</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2 Observation Data</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3 Additional Lab Data Results</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>4 Response Activities Count and Duration</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>5 Species Count</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>6 Instrument Results</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>TBD:</td>
<td></td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>- Photographs (Photologger)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Video Clips (TBD-&quot;Kaltura?&quot;)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Acoustic Clips (TBD-&quot;Kaltura?&quot;)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1.5 years into NRDA case
Data Warehouse Approach

Ingest Data
• Bring in data from different sources; flexible and scalable
• Adopt or adapt existing standards; develop and document new standards
• Manage **structured** and **unstructured** data/information
• Litigation quality
• Documented processes

Relate Information
• Examples: samples and observations; field data and photographs; laboratory packages
DIVER Overview

Data Integration, Visualization, Exploration and Reporting

DIVER is a data warehouse and query application developed by NOAA. The DIVER approach integrates and standardizes datasets so users can query across data holdings and download information and results.
Data Deliverables

Laboratory data example

• Full Laboratory packages
• Validation packages
• Either receive in template (Electronic Data Deliverable) specified by QAPP or coordinate with laboratory
• Sample cross-walk may be necessary to resolve Field/Chain of Custody/Laboratory identifiers
Common Data Model Examples (schemas)

- **Samples**: chemistry (NRDA, Response-EPA Scribe, historical), biological data, forensics...
- **Oceanographic**: cruise-collected sensor data
- **Observations**: shoreline, marsh, birds and mammals
- **Telemetry**: location tracking devices
- **Photographs**: keywords, location
- **Restoration data**: potential and implemented projects
Simplified Data Flow: Samples

Data Collection
- Samples
- Field Data

Data Management and Review
- Transcription
- Provisional
- QA/QC Validation
- Validated

Laboratory

Data Warehouse
- Samples
- Ocean Data
- Observations
- Photos

Data Access

Data Query

Visualization

Other Data Management

Data Sharing
Data Warehouse and Standardization

Collate Source Data → Apply Business Intelligence / ETL * Methods → DIVER Data Warehouse → DIVER Explorer

Steps include:
1. Define the common model
2. Accommodate additional data
3. STANDARDIZE
4. Incorporate QA/QC, Validation and Auditing

*Extract-Transform-Load

DIVER’S COMMON DATA MODELS:
- Samples
- Oceanographic
- Observations
- Telemetry
- Photos
- Related Information
- Restora0on
- Telemetry
- Photos
- Samples
- Oceanographic

Visualization (ERMA, GIS)

Visualization, Exploration, and Reporting

- Data for analysis
- Reports
- Technical Memos
- Publications
- Litigation
- Distribution
DIVER: Information Management Portal

• Centralized location for case/incident information management
  – Evolving to meet needs
  – Uses industry-standard tools including Open Source Liferay Portal Web Platform
    • Liferay Portal: http://www.liferay.com/products/liferay-portal/overview
    • Also used by NWS VLab (Virtual Lab Development): https://vlab.ncep.noaa.gov/. We are talking monthly.
DIVER: Information Management Portal

- “Wiki”-type communities for team collaboration
- Communicate updates and schedule
- Data submittal (laboratories, analysis, processed)
- Data query and reporting tools
- Help and training
DIVER Explorer Application

- Queries: Guided, Custom & Saved
- Download Data Packages
- Map & Legend
- Query by Shape
- Data Summary
- Data Tables
- Charts
- Photos
- Metadata
- Study Notes
- Export
DIVER Explorer: Guided Queries
DIVER Explorer: Dashboard Approach

Data Summary

Data Table

Charts

Metadata

Study Notes

Export
DIVER Explorer: Dashboard Approach

- Data Summary
- Data Table
- Charts
- Metadata
- Study Notes
- Export
DIVER Explorer: Dashboard Approach

- Data Summary
- Data Table
- Charts
- Metadata
- Study Notes
- Export
DIVER Explorer: Dashboard Approach

### Data Summary

The following file collections are available for sample trip id 1388.

<table>
<thead>
<tr>
<th>Source</th>
<th>Trip Date</th>
<th>Status</th>
<th>List of files</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samples</td>
<td>10-30-2010</td>
<td>Inserted</td>
<td>2010_1030_Richardson_Alec_FieldCollectionForms_DM.xls, 2010_1031_Richardson_</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FS in</td>
<td>2010_1031_Richardson_Alec_FieldCollectionForms_DM1.xls, 2010_1031_Richardson_</td>
</tr>
<tr>
<td></td>
<td></td>
<td>database</td>
<td>2010_1031_Richardson_Alec_Waybill_DM1.pdf, 2010_1031_Richardson_Alec_Final_COC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2010_1030_Richardson_Alec_FieldCollectionForms_DM1.pdf, 2010_1031_Richardson_</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2010_1031_Richardson_Alec_FieldCollectionForms_DM1.xls, 2010_1031_Richardson_</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2010_1031_Richardson_Alec_FieldCollectionForms_DM1.xls, 2010_1030_Richardson_</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2010_1031_Richardson_Alec_FieldCollectionForms_DM1.xls, 2010_1031_Richardson_</td>
</tr>
<tr>
<td>Visual</td>
<td>10-30-2010</td>
<td>Unvalified</td>
<td>2010_1031_Richardson_Alec_FieldDataForms_DM.xlsx, 2010_1031_Richardson_Alec_</td>
</tr>
<tr>
<td>Observations</td>
<td></td>
<td></td>
<td>2010_1031_Richardson_Alec_NonanalyticalCO2_DM1.pdf, 2010_1031_Richardson_Alec_</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2010_1031_Richardson_Alec_GPX_DM1.gpx, 2010_1031_Richardson_Alec_GDB_DM1.gdb,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2010_1031_Richardson_Alec_Photos_DM1.zip, 2010_1031_Richardson_Alec_PhotoLog</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2010_1031_Richardson_Alec_ReadMe_DM1.pdf, 2010_1031_Richardson_Alec_ReadMe_DM1</td>
</tr>
</tbody>
</table>
DIVER Explorer: Dashboard Approach
DIVER Explorer: Dashboard Approach

Data Summary

Data Table

Charts

Metadata

Study Notes

Export
DIVER Explorer: Dashboard Approach

- Data Summary
- Data Table
- Charts
- Metadata
- Study Notes
- Export
DIVER Explorer: Dashboard Approach

Data Summary

Data Table

Charts

Metadata

Study Notes

Export
DIVER Explorer: Dashboard Approach

Data Summary
Data Table
Charts
Metadata
Study Notes
Export
DIVER Explorer: Query By Shape

Draw and Edit; Buffers; Standard Query Shapes
DIVER Explorer: Export

Export Packages:

• Includes full FGDC Metadata (Federal Geographic Data Committee)
• Spreadsheet; Shapefile (GIS); KML (Google Earth)
• Electronic field data (spreadsheets)
• Automated output of updated data to:

  Gulf Spill Restoration
  http://www.gulfspillrestoration.noaa.gov/

  ERMA Gulf Response:
  http://gomex.erna.noaa.gov/erma.html
Public DIVER for Deepwater Horizon

Making validated scientific data available to public, researchers and academia

Released May 4, 2015
https://dwhdiver.orr.noaa.gov/
DIVER Strategy for Data Management and Query Tools

National DIVER

• OR&R developing Regional DIVER sites, Contaminant Chemistry, Observational data, Photos, Restoration, Response & Restoration data models
• Flexible, Scalable, Adaptable
• Developing field data collection capability based on DWH techniques and tools- and ongoing OR&R response, assessment and restoration mission
DIVER Strategy for Data Management and Query Tools

National DIVER

- Regional Data Focus
- Templates for data ingestion (flexible; many data types)
- Data Transcription
- Automated Ingestion (ETL)
- Quick review and download (sharing)
- Connection to raw data and documents
Current Capabilities

File Collections
Ingest and manage data from many sources; structured and unstructured

Data Warehouse (ETL)
Integrate and standardize data from multiple sources (e.g. field-collected data, laboratory data, analysis)

Query and Export Tools
Santa Barbara: Refugio Beach Oil Spill

**NOAA DIVER Workspace**

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ephemeral Data</td>
<td>ED Team 1 Subtidal Sampling</td>
</tr>
<tr>
<td>Ephemeral Data</td>
<td>Bird and beach photo observations</td>
</tr>
<tr>
<td>Ephemeral Data</td>
<td>ED Team 1 Pore Water Sampling</td>
</tr>
<tr>
<td>Ephemeral Data</td>
<td>ED Team 2 Pore Water Sampling</td>
</tr>
<tr>
<td>Ephemeral Data</td>
<td>Field observations of the cormorant colony</td>
</tr>
<tr>
<td>Recreation Use</td>
<td>Recreation Use Site Photos</td>
</tr>
<tr>
<td>Ephemeral Data</td>
<td>ED Team 1 Water Sampling</td>
</tr>
<tr>
<td>Ephemeral Data</td>
<td>Bird and beach photo observations</td>
</tr>
<tr>
<td>Ephemeral Data</td>
<td>ED Team 2 Field Observations</td>
</tr>
<tr>
<td>Ephemeral Data</td>
<td>ED Team 5 Tarball Sampling</td>
</tr>
<tr>
<td>Ephemeral Data</td>
<td>ED Team 1 Tarball Sampling</td>
</tr>
<tr>
<td>Ephemeral Data</td>
<td>ED Team 3 Tarball Sampling</td>
</tr>
<tr>
<td>Ephemeral Data</td>
<td>ED Team 4 Tarball Sampling</td>
</tr>
<tr>
<td>Recreation Use</td>
<td>Recreation Use Site Photos</td>
</tr>
<tr>
<td>Ephemeral Data</td>
<td>ED Team 2 Tarball Sampling</td>
</tr>
<tr>
<td>Ephemeral Data</td>
<td>EPA Sea Urchin Fertilization Bioassay Results</td>
</tr>
</tbody>
</table>

**File Collection**

<table>
<thead>
<tr>
<th>Type</th>
<th>File Name</th>
<th>File Size</th>
<th>Sharing Status</th>
<th>Review Status</th>
<th>Upload Date</th>
<th>Uploaded By</th>
</tr>
</thead>
<tbody>
<tr>
<td>Refugio DIVER Template</td>
<td>2015_0526_1103 - Samples</td>
<td>3.0 KB</td>
<td>Trustee Only</td>
<td>New</td>
<td>05/06/2015</td>
<td>Kamien Chiang</td>
</tr>
<tr>
<td>Refugio DIVER Template</td>
<td>2015_0526_2103 - Photographs</td>
<td>1.8 KB</td>
<td>Trustee Only</td>
<td>New</td>
<td>05/06/2015</td>
<td>Kamien Chiang</td>
</tr>
</tbody>
</table>

**Datasets for Integration into DIVER**

<table>
<thead>
<tr>
<th>Type</th>
<th>File Name</th>
<th>File Size</th>
<th>Sharing Status</th>
<th>Review Status</th>
<th>Upload Date</th>
<th>Uploaded By</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chain of Custody</td>
<td>2015_0526_COC_EOD.pdf</td>
<td>1.1 MB</td>
<td>Trustee Only</td>
<td>New</td>
<td>05/27/2015</td>
<td>Neil Ette</td>
</tr>
<tr>
<td>Field Sample Form</td>
<td>2015_0526_FSF_EOD.pdf</td>
<td>4.8 MB</td>
<td>Trustee Only</td>
<td>New</td>
<td>05/27/2015</td>
<td>Neil Ette</td>
</tr>
<tr>
<td>GPS File (gpr)</td>
<td>2015_0526_E03_GPS.gps</td>
<td>88.2 KB</td>
<td>Trustee Only</td>
<td>New</td>
<td>05/26/2015</td>
<td>Ben Shorn</td>
</tr>
<tr>
<td>Original Image Plus</td>
<td>2015_0526_EOD_Pictures.jpg</td>
<td>202.8 MB</td>
<td>Trustee Only</td>
<td>New</td>
<td>05/26/2015</td>
<td>Ben Shorn</td>
</tr>
</tbody>
</table>
Santa Barbara: Refugio Beach Oil Spill
DIVER Technology and Processes: Open Source Tools

- **ETL Process:** Pentaho Analytics
  - Schemas; PostgreSQL/PostGIS;
- **Portal:** Liferay
- **Explorer Query Tool**
  - Custom Java API and query engine
  - Mapserver/OpenLayers
  - Dojo Toolkit Javascript library
  - MDX query language – Pentaho Community Edition
  - Infobright (Hadoop integration)-fast column-oriented db
30 minutes goes quickly when you’re talking big data!

Technical Details:
- Amazon AWS; FedRAMP;
- NIST 800-53 security standards compliant
- Liferay Portal
- Pentaho Data Integration tools
- PostgreSQL/PostGIS
- Infobright (Hadoop integration)
- Mapserver/OpenLayers
- Dojo Toolkit Javascript library
- Custom Java API and query engine
- Agile development approach:
  (data management and tool development)
- IBM Domino (phasing out)

Senior Team (and co-authors):
- Ben Shorr (Spatial Data Branch)
- Dr. Amy Merten (Spatial Data Branch Chief)
- Dan Hudgens (IEc Inc.)
- Neal Etre (IEc, Inc.)
- Jim Anderton (Solea Consulting)
- Jerry Bower (Sirius Computer Solutions)