# CMMAP CIWG: Boulder, Winter 2013

John Helly, Chin-Hoh Moeng, Don Dazlich, Steve Krueger,
Hugh Morrison, Kelley Wittemeyer, Andy Heymsfield,
Marat Kharoutdinov

San Diego Supercomputer Center & Scripps Institution of Oceanography
University of California, San Diego
La Jolla, California 92093

*hellyj@ucsd.edu*

January 25, 2013

# Outline

# Outline

# Outline

# New XSEDE Allocations

XSEDE User Portal | Allocations/Usage

XSEDE User Portal | Allocations...

https://www.xsede.org/group/xup/allocations/usage

MY XSEDE | RESOURCES | DOCUMENTATION | ALLOCATIONS | TRAINING | USER FORUMS | HELP | ABOUT

Summary | Allocations/Usage | Accounts | Jobs | Profile | Publications | Tickets | Change Password | Add User | Community Accounts | SSH Terminal

## Projects

SHOW INACTIVE PROJECTS | SHOW EXPIRED/INACTIVE ALLOCATIONS

### Large Eddy Simulations (GigaLES) of Deep Convection Over Continental and Oceanic Domains Using the System for Atmospheric Modeling (SAM) Cloud-resolving Model (CRM)

**Project PI:** Helly, John

**Charge No.:** TG-ATM100027

| RESOURCE | AWARD | MY USAGE (SU) | VIEW USAGE | BURN RATE | END DATE [DAYS LEFT] | TYPE | STATE | USERS |
|---|---|---|---|---|---|---|---|---|
| ecss.xsede | 100% \| 3 SUs left<br>from 3 SU award | 0.0 | N/A | Burn Rate | 2013-12-31 [342d] | renewal | active | |
| Gordon Compute | 100% \| 1,700,000 SUs left<br>from 1,700,000 SU award | 0.0 | | Burn Rate | 2013-12-31 [342d] | renewal | active | |
| Kraken | 100% \| 1,600,000 SUs left<br>from 1,600,000 SU award | 0.0 | | Burn Rate | 2013-12-31 [342d] | renewal | active | |
| Stampede | 100% \| 500,000 SUs left<br>from 500,000 SU award | 0.0 | | Burn Rate | 2013-12-31 [342d] | renewal | active | |

Allocations for Supplements, Advances and Transfers appear in the portal before they appear in the accounting records at XSEDE sites. Please allow 24 hours after you receive an award notification for the allocation updates to appear in the accounting records of the XSEDE sites.

# Dazlich Benchmarks

## Benchmarking giga-LES

| platform | throughput - sim days/ wallclock days | allocation SU | simulation possible: days |
|---|---|---|---|
| kraken | 0.12 | 1,600,000 | 7.8 |
| stampede | 0.29 | 500,000 | 5.9 |

SAM6.10.3: 2048x2048x256, 2s timestep, 8byte floats, M2005 microphysics, RRTM radiation,1024 tasks

kraken: NICS Cray XT5

stampede: TACC Dell PowerEdge C8220 cluster, without MIC acceleration

For comparison first giga-LES had throughput of about 0.17 on 2048 BlueGene/L cores running 4byte floats, 1 moment microphysics and prescribed radiation.

# Dazlich Benchmarks

## Benchmarking giga-LES

| platform | throughput - sim days/ wallclock days | allocation SU | simulation possible: days |
|---|---|---|---|
| kraken | 0.12 | 1,600,000 | 7.8 |
| stampede | 0.29 | 500,000 | 5.9 |

SAM6.10.3: 2048x2048x256, 2s timestep, 8byte floats, M2005 microphysics, RRTM radiation,1024 tasks

kraken: NICS Cray XT5

stampede: TACC Dell PowerEdge C8220 cluster, without MIC acceleration

For comparison first giga-LES had throughput of about 0.17 on 2048 BlueGene/L cores running 4byte floats, 1 moment microphysics and prescribed radiation.

# Outline

# Issues

1. New small domain benchmarks
   1. Small domain (256 x 256 x 256), 800m grid cells / better benchmark for spinup of physics (Kraken, Stampede, Gordon)
   2. 1km resolution over May 2011 ARM to select specific time periods to run LES from there select time periods for GigaLES.
2. Have to wait for snow radiatively-active snow (Peter Blossey, Robert Pincus, Hugh Morrison, Steve Krueger)
3. Lagrangian tracers (Tak package)
4. Data
   1. 11 2-moment mPhysics variable (cloud water, ice, rain, snow, graupel, mass & number, water vapor mass)
   2. Lagrangian tracer(variables ???)
5. Steve K: recruit mission 5-min data from first GigaLES.

# Outline

# Ocean

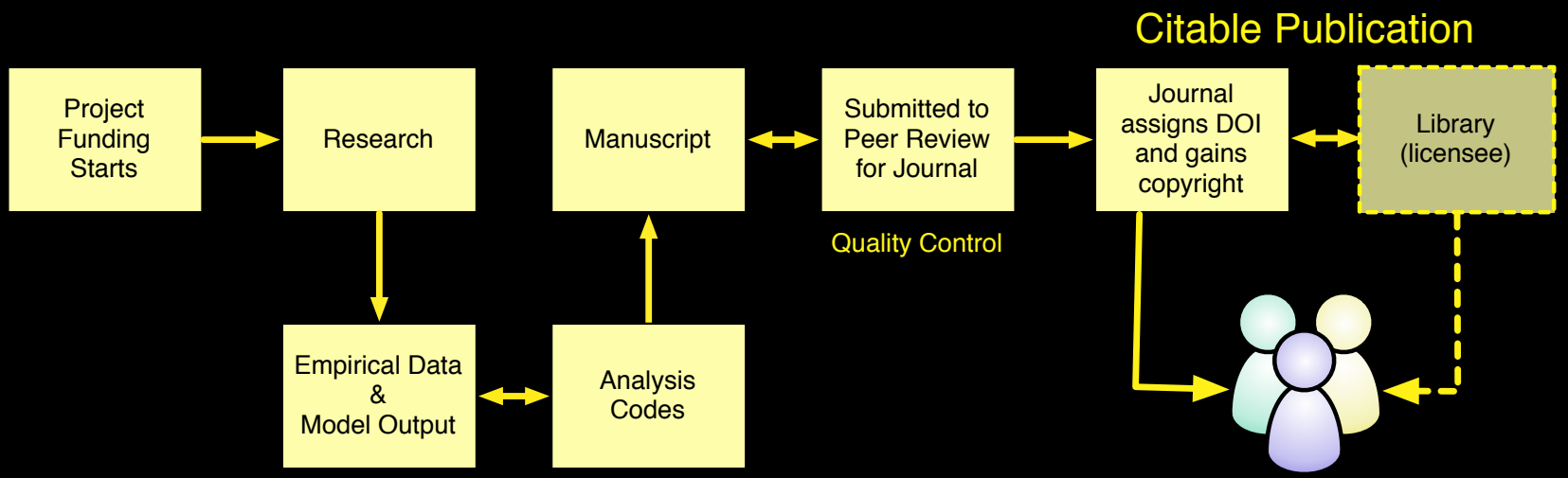
1. Tropical Western Pacific (periodic boundary condition?; is data good?) or GATE case
2. Tropical Warm Pool (Western Australia) http://acrf-campaign.arm.gov/twpice/
3. Use GCSS study case focus on Jan 18-25 (active period).

# Land

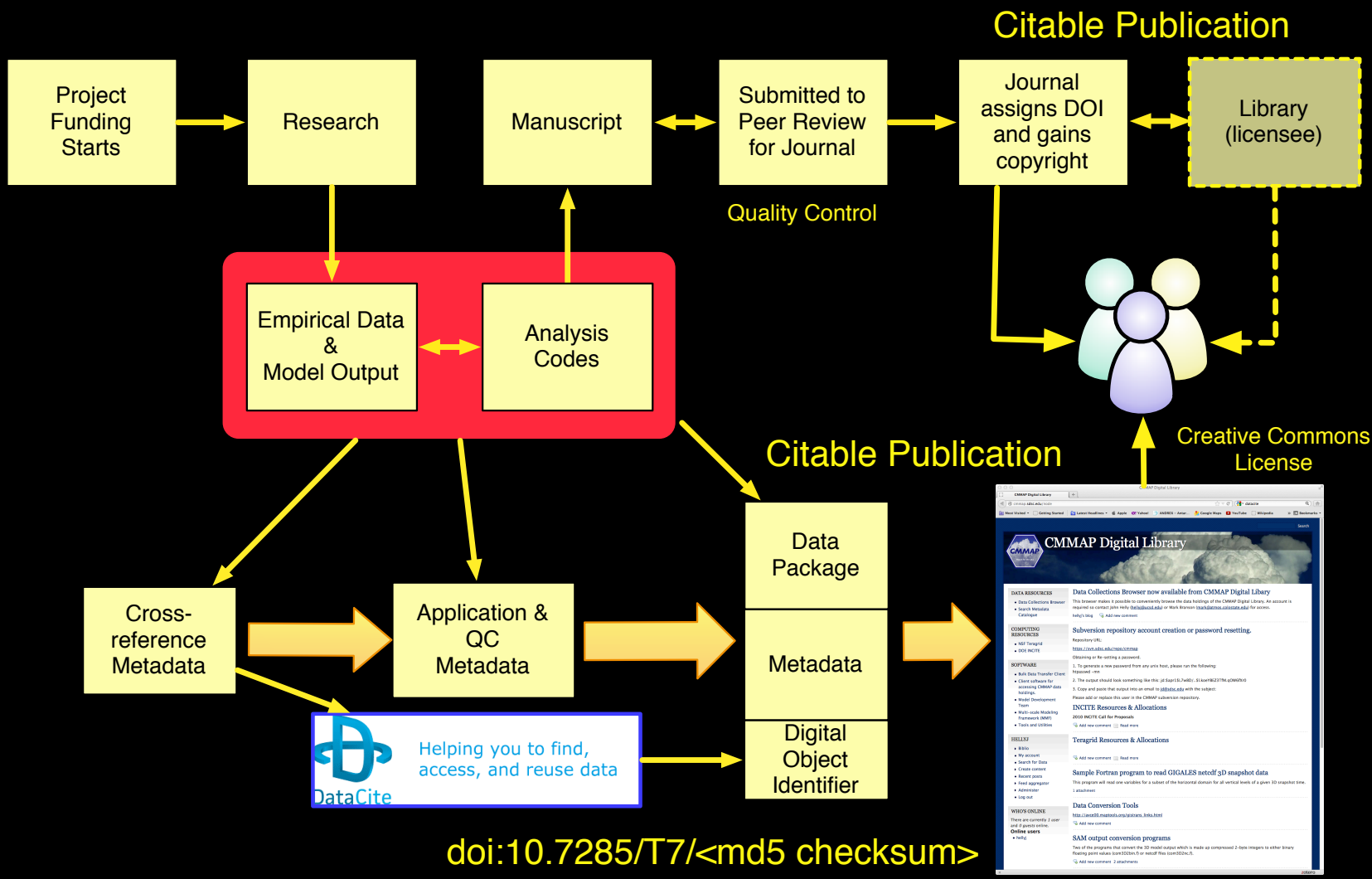
1. Mid-latitude C3E ARM Case Midlatitude Continental Convective Clouds Experiment (MC3E) (http://www.arm.gov/campaigns/sgp2011midlatcloud)
2. Which period is better for observations? (27-Apr (per Andy), May 10-11 or May 24) days based on review of radar imagery
3. Andy has aircraft data and particle data.
4. Need to set other parameters in mPhysics (hail vs graupel, aerosol properties)
5. Soil, land-cover?
6. Ask Steve K for soundings and external forcings.

# Outline

# What Is Data Authorship and Publication?



# What Is Data Authorship and Publication?

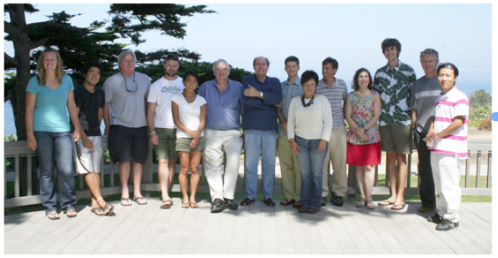


# Also Works for Short-lived Projects

## Scientific Research Projects

Coastal Hypoxia and Ocean Acidification

**UCMexus**

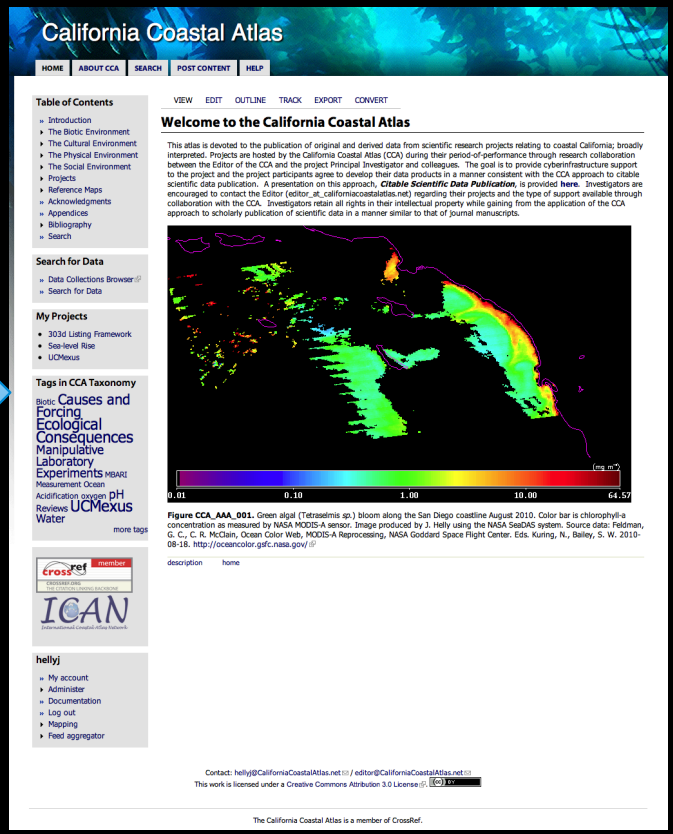



## Quality Review

Scientific Editorial Quality Control

Scientific Editorial Quality Control

## Sound Published Data, Citable & Cross-referenced



# Outline

# Brief History: Literature and Basic Functions

| | |
|---|---|
| Data Acquisition | Data is acquired through contribution and submissions, along with at least a minimal set of metadata. This initiates the automatic creation of a unique name for the ADO and a transportable metadata file bundled within the ADO. |
| Search and Retrieval | A search system provides for spatial, temporal, and thematic (such as keyword) queries based on metadata content. |
| Deletion Control | The ability to delete an ADO is tightly controlled to prevent the arbitrary deletion of data copied by users. In a manner analogous to journal articles, no one should be able to unpublish data. Errata can be accommodated by publishing a revision of the data. An important special case to consider is the editorial peer-review process requiring confidentiality and the ability to remove an ADO if not accepted for peer-reviewed publication. A looser deletion policy might allow deletion of data if it had never been copied. |
| Assignment of Persistent Names | The persistent name, or accession number of an ADO, as in Figure 1, is used in the data repository to access the ADO, monitor updates of previously published ADOs, identify the retrieval of ADOs by users, notify users of anomalies or issues related to an ADO, establish precedence by publication date, and enable citation in other publications. |
| Quality Control and Quality Assurance Policy and Methods | This function can exist (or not exist) to varying degrees, exemplified by peer review and non-peer review, as well as by anomaly detection and reporting, though it must be stated explicitly. Some investigation is beginning on how to semiautomate QA/QC for specific types of data. |
| Access Control | Access control enables data contributors to specify a password only they know and that may be provided to other users to access the contributed ADO. This approach enables data submitters to independently control access to their own published data. Any user attempting to retrieve a password-protected ADO from the system needs to obtain that password from the data's contributor. |
| Traceability of Data Heritage | A mechanism for establishing the heritage of data contained within an ADO informs users of the data's measured, derived, or computed nature. This approach is also essential to preserving intellectual property rights analogous to claims of copyright or trademark. |

# Basic functions for the controlled publication of scientific data (updated)

| Function | Purpose |
|---|---|
| User Registration<br>Credentials & Authentication | A user ID and password are assigned to a given user while acquiring the user's email address and related contact information. The ID is used to audit data access and communication with users. |
| Data Acquisition | Data is acquired through contribution and submissions, along with at least a minimal set of metadata. This initiates the automatic creation of a unique name for the ADO and a transportable metadata file bundled within the ADO. |
| Search and Retrieval<br>Enabled by metadata | A search system provides for spatial, temporal, and thematic (such as keyword) queries based on metadata content. |
| Deletion Control<br>Version Control | The ability to delete an ADO is tightly controlled to prevent the arbitrary deletion of data copied by users. In a manner analogous to journal articles, no one should be able to unpublish data. Errata can be accommodated by publishing a revision of the data. An important special case to consider is the editorial peer-review process requiring confidentiality and the ability to remove an ADO if not accepted for peer-reviewed publication. A looser deletion policy might allow deletion of data if it had never been copied. |
| Assignment of Persistent Names<br>DOIs | The persistent name, or accession number of an ADO, as in Figure 1, is used in the data repository to access the ADO, monitor updates of previously published ADOs, identify the retrieval of ADOs by users, notify users of anomalies or issues related to an ADO, establish precedence by publication date, and enable citation in other publications. |
| Quality Control and Quality Assurance Policy and Methods | This function can exist (or not exist) to varying degrees, exemplified by peer review and non-peer review, as well as by anomaly detection and reporting, though it must be stated explicitly. Some investigation is beginning on how to semiautomate QA/QC for specific types of data. |
| Access Control<br>Authentication & New Open-ness | Access control enables data contributors to specify a password only they know and that may be provided to other users to access the contributed ADO. This approach enables data submitters to independently control access to their own published data. Any user attempting to retrieve a password-protected ADO from the system needs to obtain that password from the data's contributor. |
| Traceability of Data Heritage<br>Provenance | A mechanism for establishing the heritage of data contained within an ADO informs users of the data's measured, derived, or computed nature. This approach is also essential to preserving intellectual property rights analogous to claims of copyright or trademark. |

# Current Status

This PDF is available from The National Academies Press at http://www.nap.edu/catalog.php?record_id=13564

For Attribution -- Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop

ISBN
978-0-309-26728-1

236 pages
8 1/2 x 11
PAPERBACK (2012)

Paul E. Uhlir, Rapporteur; Board on Research Data and Information; Policy and Global Affairs; National Research Council

Add book to cart | Find similar titles | Share this PDF

# AGU's Position on Data Publication

## 2- Formal Publication of Data: An Idea Whose Time Has Come?

**Jean-Bernard Minster**[1]
*University of California at San Diego*

Every time I participate in a discussion on data citation and attribution or talk to colleagues who deal with a lot of data, the issue of data publication comes up. The point is that the whole idea of citation is difficult to discuss in the absence of the concept of publication. The idea of long-term data preservation, citation, and publication is a concept that is growing in the community. In my scientific union, the American Geophysical Union (AGU), there is a statement on data publication that reads:

> The cost of collecting, processing, validating, and submitting data to a recognized archive should be an integral part of research and operational programs. Such archives should be adequately supported with long-term funding. Organizations and individuals charged with coping with the explosive growth of Earth and space digital data sets should develop and offer tools to permit fast discovery and efficient extraction of online data, manually and automatically, thereby increasing their user base. The scientific community should recognize the professional value of such activities by endorsing the concept of publication of data, to be credited and cited like the products of any other scientific activity, and encouraging peer-review of such publications.[2]

# Current Status

## TG on Data Citation

**Presentation to CODATA**
**Taipei, Taiwan October 2012**

Enabling Science and Innovation
ICSTI
International Council for Scientific and Technical Information

CODATA 23
Taipei 2012

Co-Chairs

Jan Brase, Germany
Bonnie C. Carroll, US (Presenter)
Sarah Callahan, UK

Support

Paul Uhlir, US

## Overview

- Objectives:
  - Examine key issues related to data identification, attribution citation a linking
  - Help Coordinate activities internationally
  - Promote common practices and standards
- Covers a diversity of interesting and geography
  - 3 Co-Chairs: UK, DE, US
  - +19 Members
  - 4 continents, 14 countries
  - 5 Consultants
- Funders
  - CODATA
  - Sloan Foundation
  - Institute for Museum and Library Services
  - Library of Congress
  - Microsoft Research

- Symposium and Workshop, Berkeley, CA August 2011: *For Attribution: Developing Data Attribution and Citation Practices and Standards*
- *3 Track session at CODATA 2012 on Data Publishing and Data Citation in Cooperation with the WDS*
- *Draft report on* *The State of Digital Citation* *(in process, 3/13)*

# Outline

# Why Is Data Publication Important for the Future of Scholarly Work?

- Provenance Can Be Unambiguously Established
  - Identification and verification of content (i.e., scholarly work product) can be done
  - Enables *chain-of-custody* to be determined
- Reproducibility of results is enabled
  - There is a published *copy-of-record* at time $t$ for the indefinite future
  - Version control is necessary to provide temporal record of changes to data
- Responsibility and Authority Can Be Correctly Assigned
  - Attribution and assignment of accomplishments and intellectual property rights
  - Anomaly correction and versioning of singleton and multi-component datasets can be better quality-controlled*

# *Caveat Emptor*: No *Copy-of-Record* Means No Reproducibility

*Note: values may change as grid estimates are improved by the addition of new soundings, contours, etc.

**DOWNLOAD SID IMAGES**

| | |
|---|---|
| Compressed image of all soundings and depth contours | delta_pc.sid (10.7 Mbyte) |
| Compressed image of the depth grid | Delta_z.sid (2.3 Mbyte) |

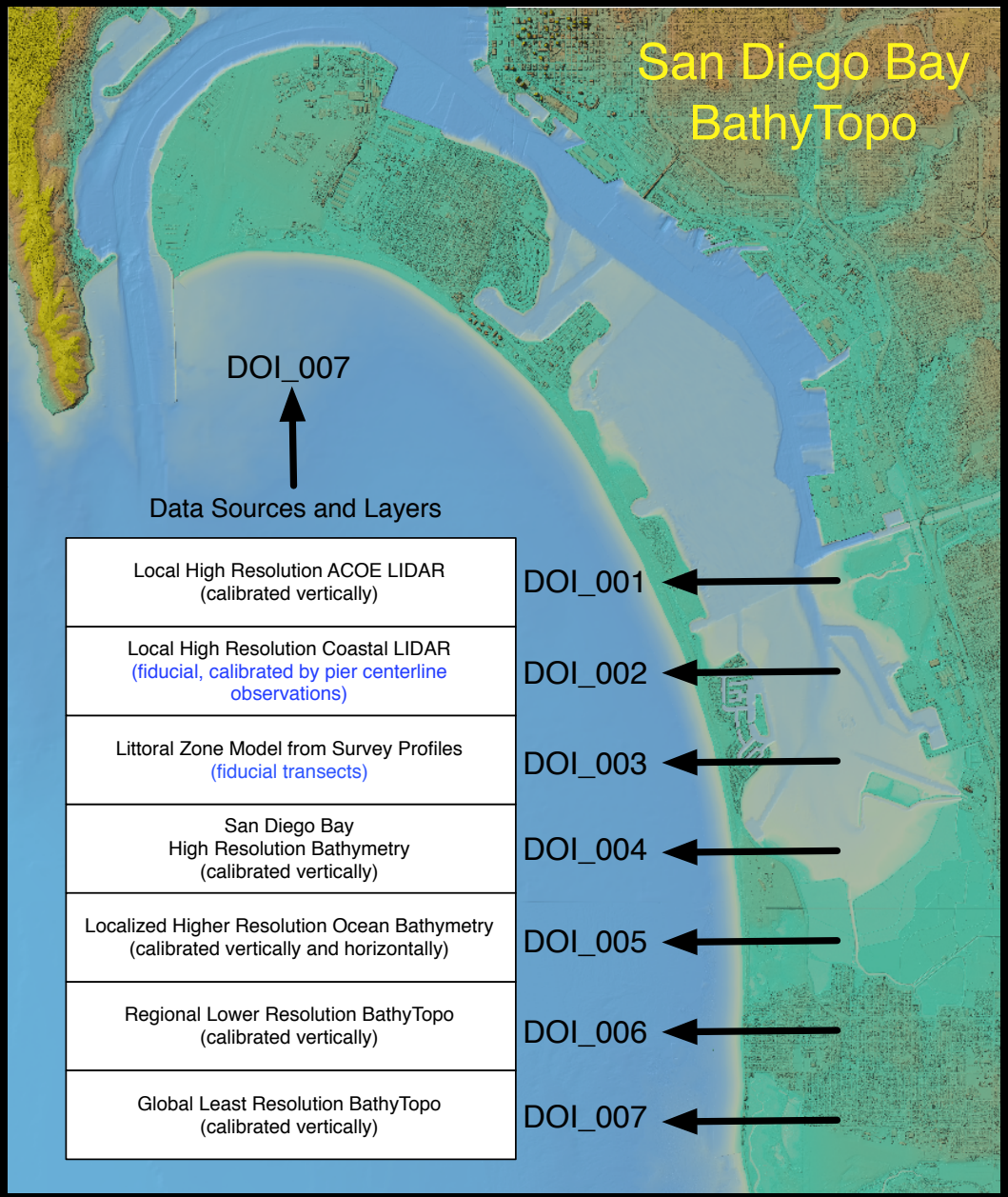
FOIA Privacy Policies and Notices

t of the Interior | U.S. Geological Survey
.wr.usgs.gov/sediment/delta/downloads.html
ed: Monday, 23-Nov-2009 09:52:10 PST

USA.gov Government Made Easy

TAKE PRIDE IN AMERICA

# Multi-source Composite Example




- Public source data (Level 0)
- Re-calibrated to common datums and error reduction (Level 1)
- Integrated into a composite dataset that is distinct and ready for hydrodynamic modeling

# Outline

# Amazing New Opportunities Are Enabled Through Data Publication

- Data Fusion (Decomposition) Can Be Increasingly Automated
  - Multi-source datasets can be accessed and integrated with higher-reliability using catalogue-level metadata
  - Data updating and versioning of singleton and multi-component datasets can be better quality-controlled through automated processing of large numbers of files of any size
- Applications Can Interoperate Reliably At New Levels of Scale and Complexity
  - Across disciplines and scales of space and time with an accurate, reproducible history of processing
  - New tools can be built to exploit the information from permutations and combinations of data components (we see this now in geospatial data)

# *New Developments*: This could be a game-changer

provide their members.

Finally, if the policy is adopted, there will likely need to be a gradual shift of library funds from traditional journal subscriptions to digital manuscript storage and access.

Associated with the policy are several important documents including a cover letter from Christopher Kelty (Chair, UCOLASC) to Robert Anderson (Chair, Academic Council), a presentation of the policy, and common questions and answers surrounding the policy. All can be found at: http://osc.universityofcalifornia.edu/openaccesspolicy/. The initial UCSD Library Committee response to the policy can be found here.

**More about Open Access and its implications can be viewed here.**

*** Discuss the Proposed UC Open Access Policy on the Academic Senate Forum.**

Back

Google Search | Site Map | Campus Directory

UCSD Official web page of the University of California, San Diego

Email Webmaster

Copyright ©2005 Regents of the University of California. All rights reserved.

# Help Make It Happen

- It took 10+ years to get this far
- Encourage your departments to recognize data citations in merit criteria
- Start using them in your manuscripts
- Find out what your institution is doing (or not)
- Teach your students and colleagues about it (most importantly the students)