BIGDATA, littledata and EvErYtHiNg in Between
Strategies for Scientific Data Management

John Helly

La Jolla, California 92093

hellyj@ucsd.edu

April 22, 2014

Outline

1 What's Unique About Scientific Data Management?

2 The Scientific Method And Reproducibility

3 Digital Library Framework

Outline

Create account Log in

WIKIPEDIA
The Free Encyclopedia

Article Talk | Read View source View history | Search

Science

From Wikipedia, the free encyclopedia

This article is about the general term, particularly as it refers to experimental sciences. For the specific topics of study by scientists, see Natural science. For other uses, see Science (disambiguation).

Science (from Latin *scientia*, meaning "knowledge") is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the universe.[1] In an older and closely related meaning (found, for example, in Aristotle), "science" refers to the body of reliable knowledge itself, of the type that can be logically and rationally explained (*see History and philosophy below*).[2] Since classical antiquity science as a type of knowledge was closely linked to philosophy. In the early modern era the words "science" and "philosophy" were sometimes used interchangeably in the English language.[*citation needed*] By the 17th century, natural philosophy (which is today called "natural science") was considered a separate branch of philosophy.[3] However, "science" continued to be used in a broad sense denoting reliable knowledge about a topic, in the same way it is still used in modern terms such as library science or political science.

In modern use, "science" more often refers to a way of pursuing knowledge, not only the knowledge itself. It is "often treated as synonymous with 'natural and physical science', and thus restricted to those branches of study that relate to the phenomena of the material universe and their laws, sometimes with implied exclusion of pure mathematics. This is now the dominant sense in ordinary use."[4] This narrower sense of "science" developed as scientists such as Johannes Kepler, Galileo Galilei and Isaac Newton began formulating *laws of nature* such as Newton's laws of motion. In this period it became more common to refer to natural philosophy as "natural science". Over the course of the 19th century, the word "science" became increasingly associated with the scientific method, a disciplined way to study the natural world, including physics, chemistry, geology and biology. It is in the 19th century also that the term *scientist* was created by the naturalist-theologian William Whewell to distinguish those who sought knowledge on nature from those who sought knowledge on other disciplines. The *Oxford English Dictionary* dates the origin of the word "scientist" to 1834. This sometimes left the study of human thought and society in a linguistic limbo, which was resolved by classifying these areas of academic study as social science. Similarly, several other major areas of disciplined study and knowledge exist today under the general rubric of "science", such as formal science and applied science.

Montage of some highly influential scientists from a variety of scientific fields. From left to right:
Top row: Archimedes, Aristotle, Ibn al-Haytham, Leonardo da Vinci, Galileo Galilei, Antonie van Leeuwenhoek;
Second row: Isaac Newton, James Hutton, Antoine Lavoisier, John Dalton, Charles Darwin, Gregor Mendel;
Third row: Louis Pasteur, James Clerk Maxwell, Henri Poincaré, Sigmund Freud, Nikola Tesla, Max Planck;
Fourth row: Ernest Rutherford, Marie Curie, Albert Einstein, Niels Bohr, Erwin Schrödinger, Enrico Fermi;
Bottom row: J. Robert Oppenheimer, Alan Turing, Richard Feynman, E. O. Wilson, Jane Goodall, Stephen Hawking

Contents [hide]

Part of a series on
Science

Formal sciences	[show]
Physical sciences	[show]
Life sciences	[show]
Social sciences	[show]
Applied sciences	[show]
Interdisciplinarity	[show]
Philosophy and history of science	[show]

Science portal · Category

V • T • E

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Languages
Afrikaans
አማርኛ
Ænglisc
العربية
Aragonés
অসমীয়া
Asturianu
Avañe'ẽ
Aymar aru
Azərbaycanca
Bamanankan
বাংলা
Bahasa Banjar
Bân-lâm-gú
Basa Banyumasan
Башкортса
Беларуская
Беларуская (тарашкевіца)
Български
Boarisch
བོད་ཡིག
Bosanski
Brezhoneg
Català
Чăвашла
Cebuano
Česky
Cymraeg

Scientific Method According to Feynman

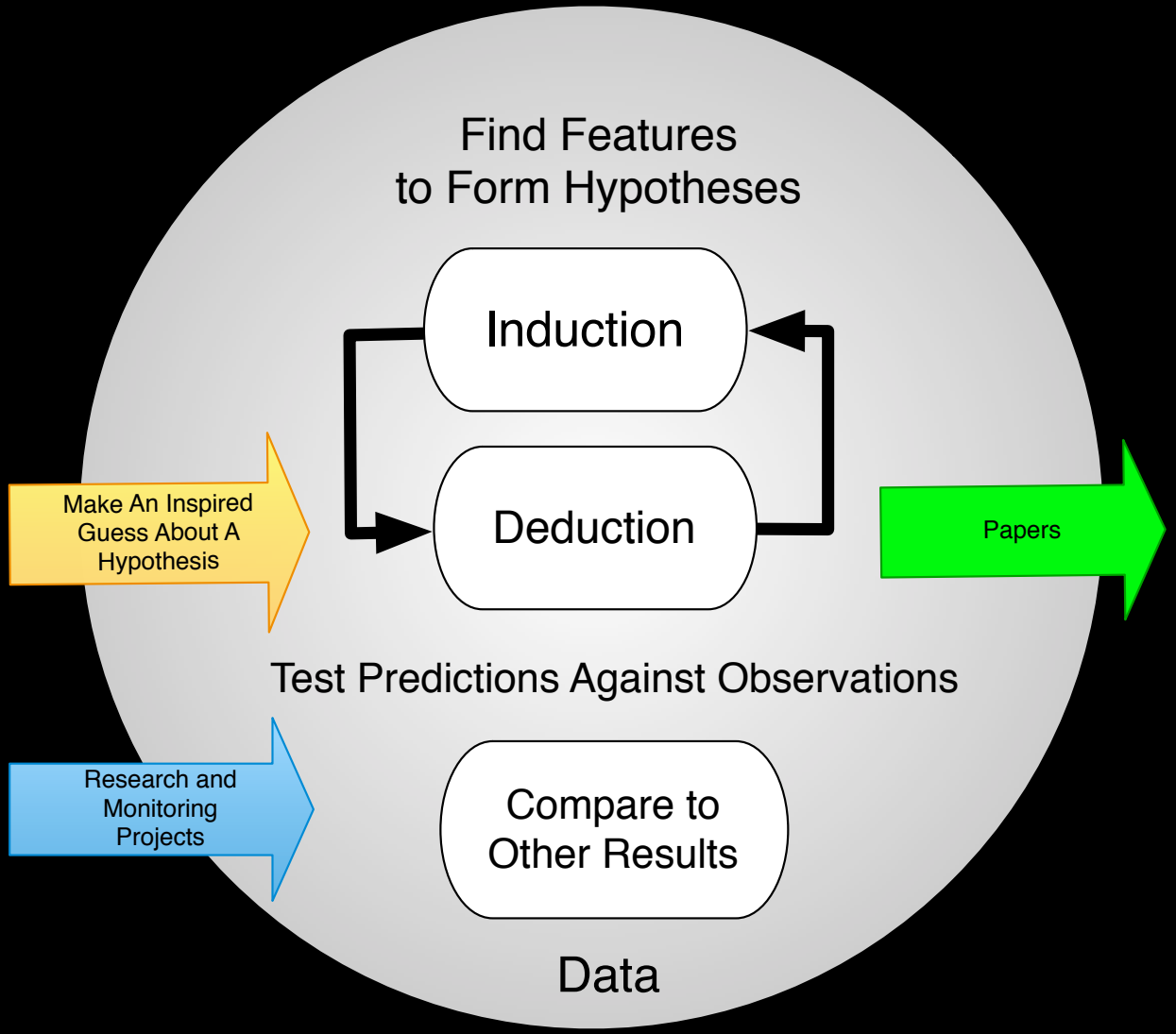
http://www.youtube.com/watch?v=EYPapE-3FRw

Procedure

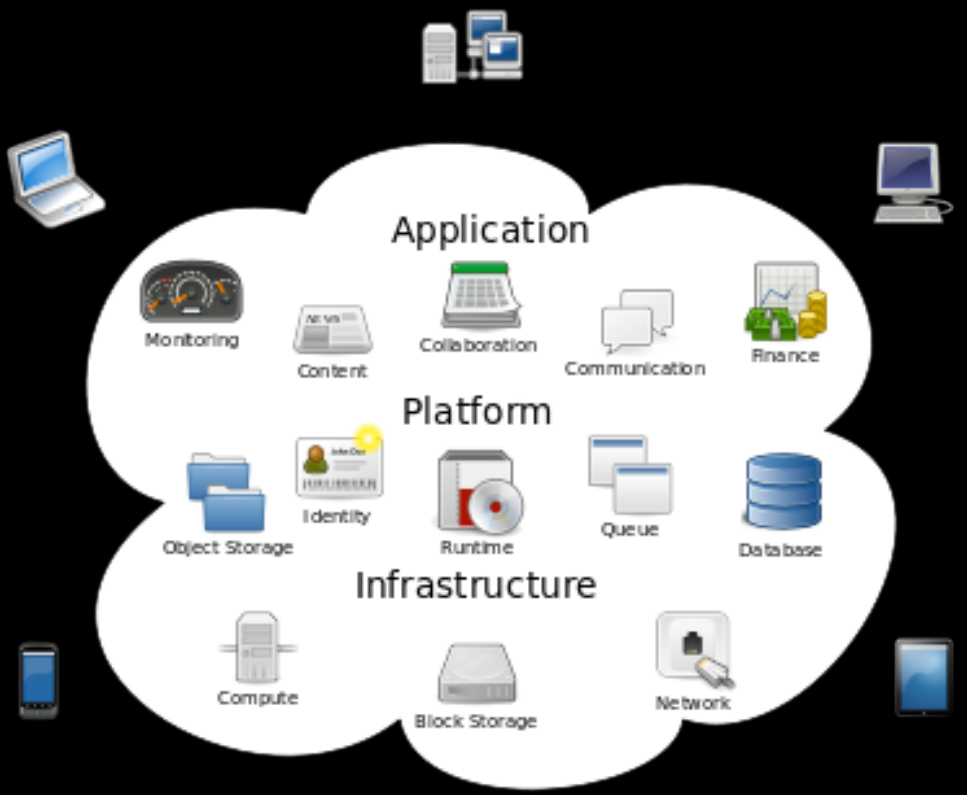
1. Come up with a theory and hypothesis
2. Compute consequences (i.e., make predictions)
3. Compare predictions to observations (nature, experiment, experience)
4. If predictions do not agree with observations, the theory and hypothesis are wrong

And we can never prove we are right, only that we are wrong

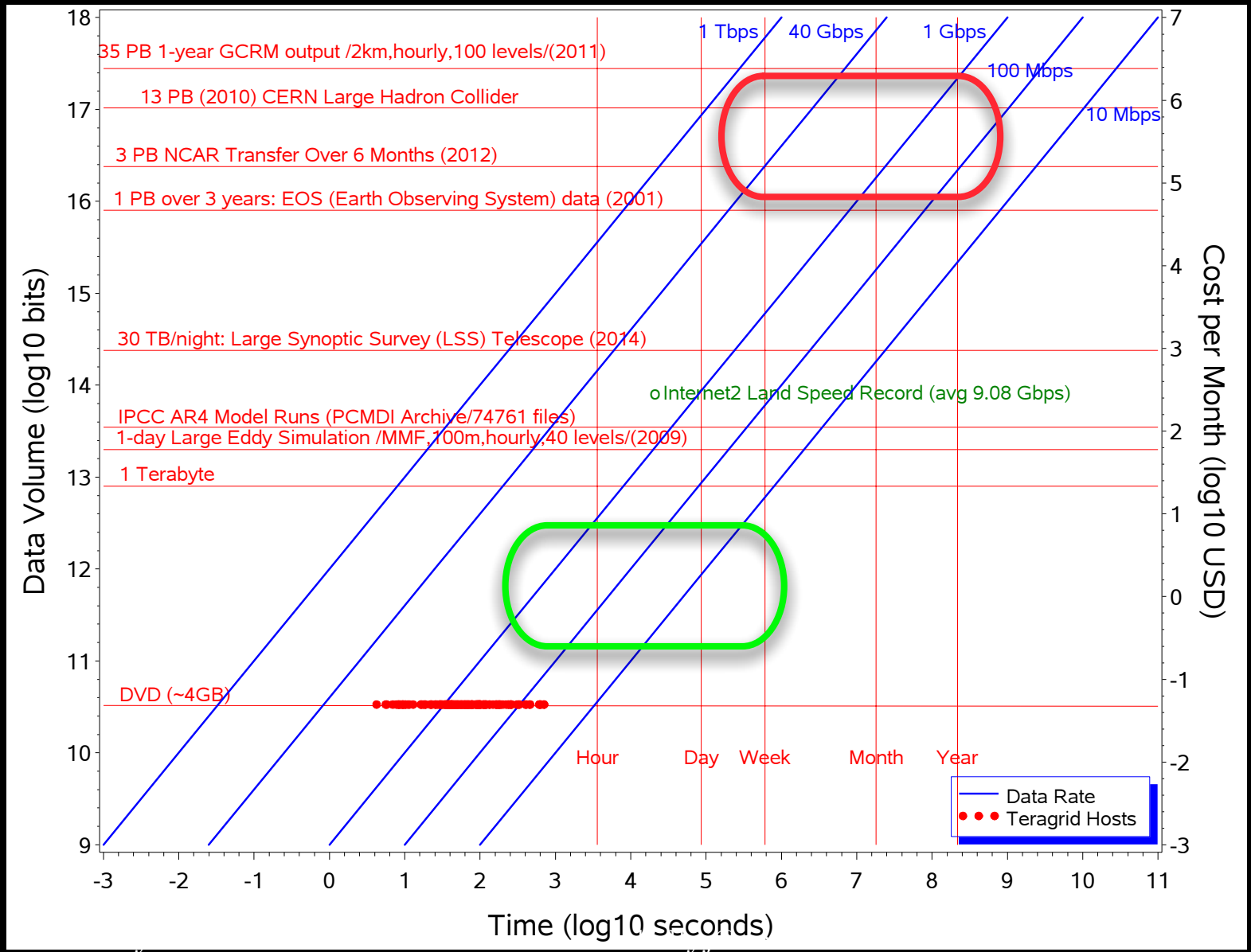
As a Data-intensive Iterative Procedure



With Constantly Changing Computing Technology Requiring Data Transportation and Migration



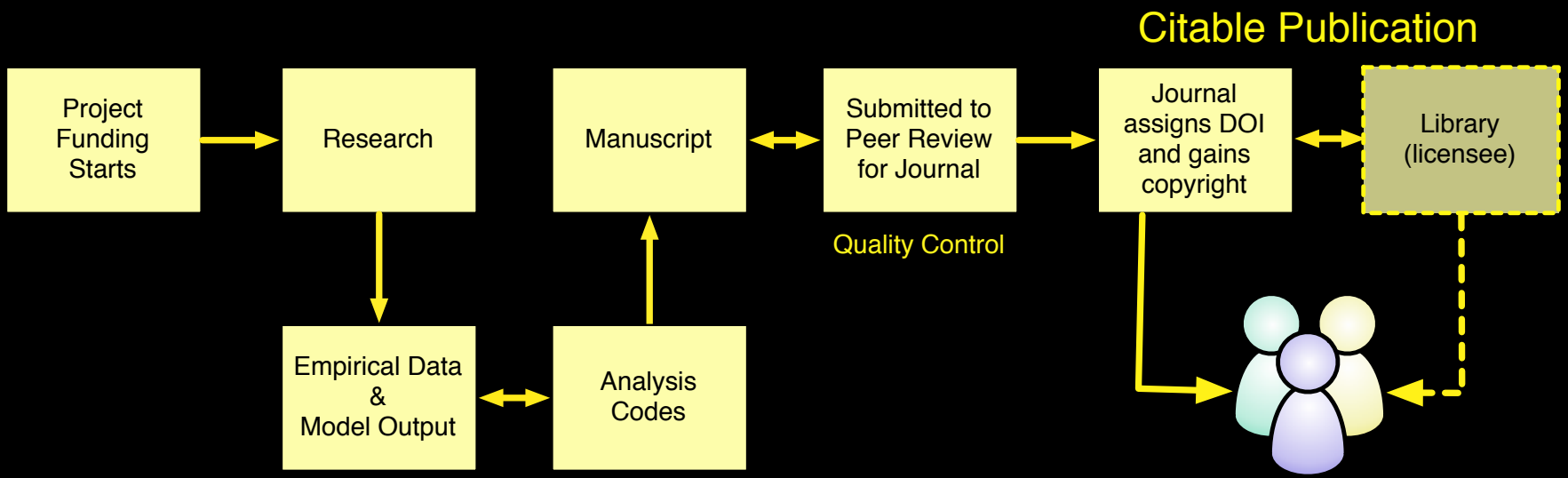
BIGDATA
A Moveable Feast



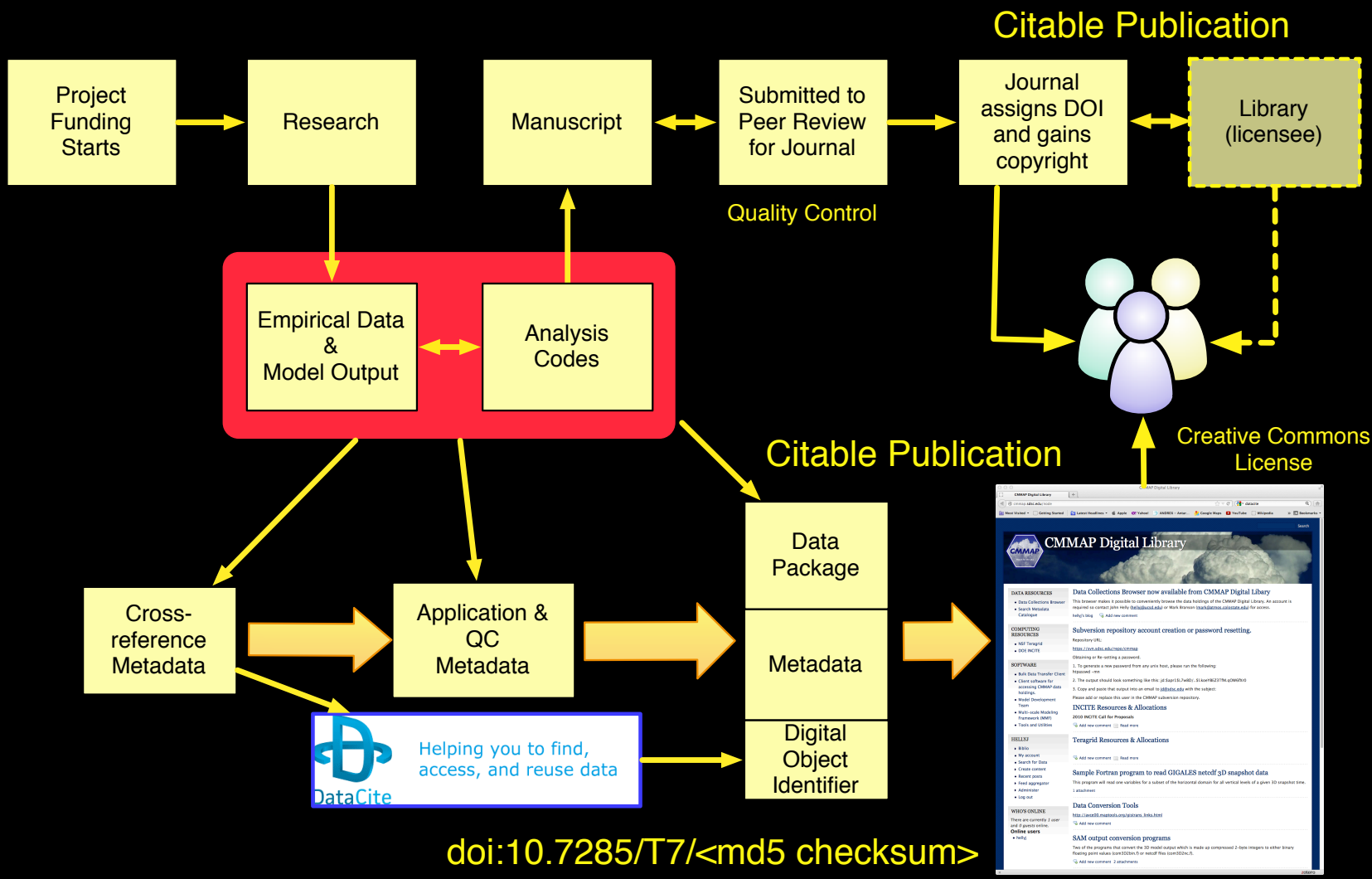
BIGDATA old news

- Key questions that never seem to change
  - What to keep, what to throw away?
  - How to keep the data you want (hardware, software, format)?
  - How do you back it up and how many copies do you need?
  - Can you actually restore it?
- Where do you put it?
- Who's in charge of it to make these decisions?
- What's the role of institutional libraries versus laboratories versus individual scientists?

Scientific Workflow With Quality Control, Attribution and Dissemination of Results



... Augmented by Data Authorship and Publication



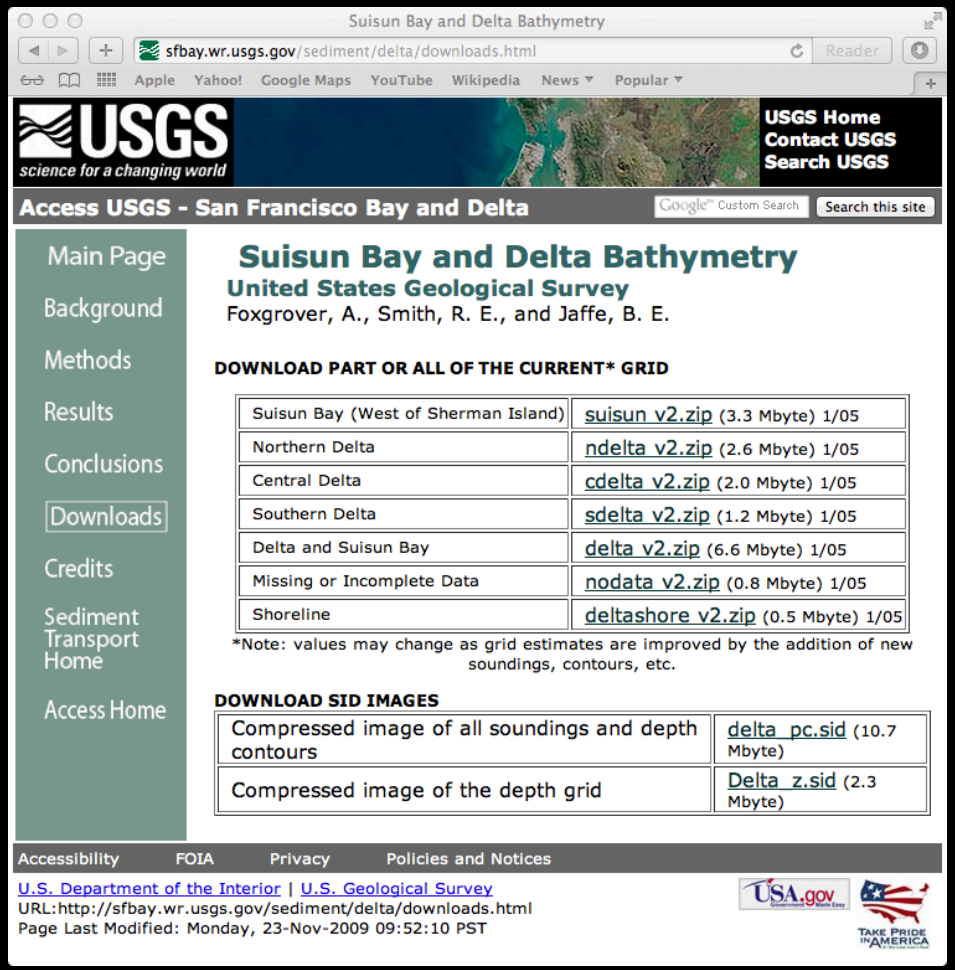
Outline

Reproducibility

(paraphrased from Feynman's Cargo-cult Science Speech Caltech 1974)

- One of the students told me she wanted to do an experiment that went something like this ... under certain circumstances, X, rats did something A. She was curious as to whether, if she changed the circumstances to Y, they would still do A.
- I explained to her that it was necessary first to repeat the experiment of the other person; to do it under condition X to see if she could also get result A, and then change X to Y and see if A changed.
- Then she would know that the real difference was the thing she thought she had under control.

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



- Data that is updated *in situ* without version control does not comply with the scientific method
- Streaming data has ambiguous provenance
- Web-services are a form of streaming data
- What about *the cloud*?

Caveat Emptor: No Archival *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

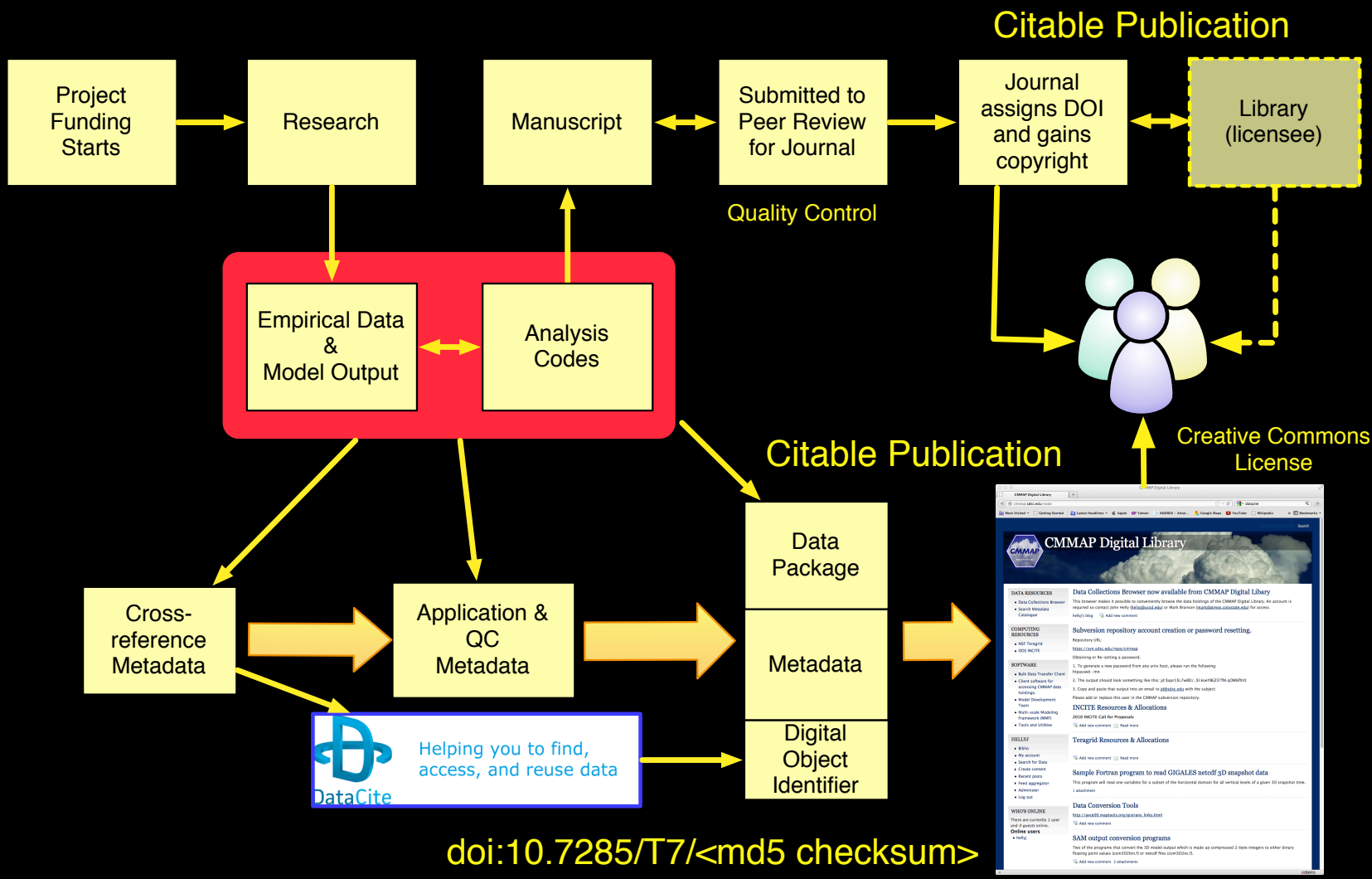
What Are the Implications Of the Reproducibility Requirement For Scientific Data Management?

1. Input and output data must be preserved over time and versioned if changes (i.e., updates, anomalies) are acquired.
2. Processing scripts and codes must be preserved *in association* with the input and output data.
3. Computing platform must be documented (e.g., analogize bug-tracking and fixing in open-source projects; reproduce the conditions that produce the bug) in association the data and the processing.

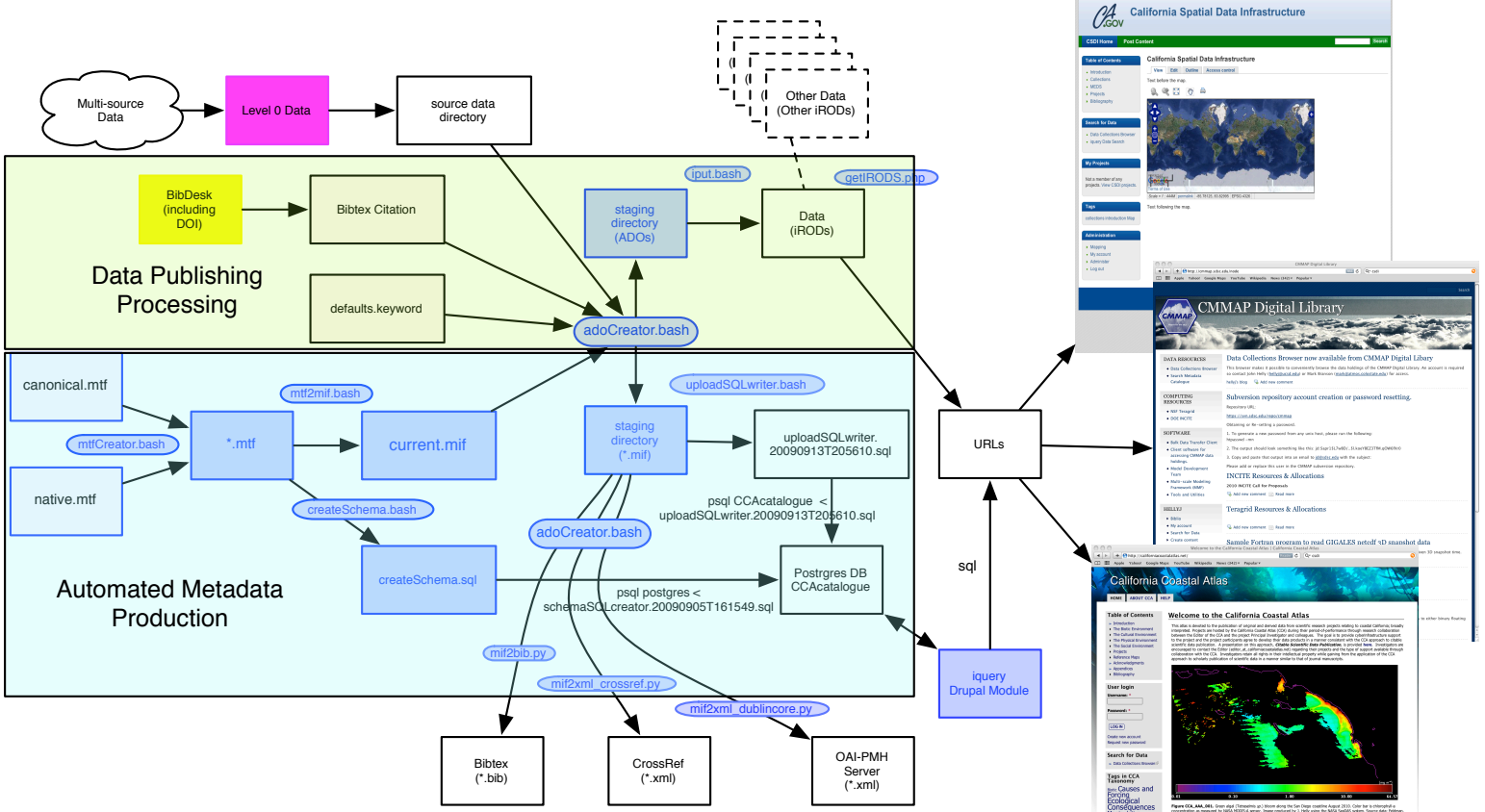
How do we handle this?

Outline

Through Data Publication and Citation Long Overdue



Digital Library Framework: Highly Automated Data Publication Workflow



Multi-lateral Metadata Interfaces

Digital Libraries Work

PNAS PNAS PNAS

Spatial characterization of the meltwater field from icebergs in the Weddell Sea

John J. Helly[a,b,1], Ronald S. Kaufmann[c], Maria Vernet[b], and Gordon R. Stephenson[b]

[a]San Diego Supercomputer Center and [b]Scripps Institution of Oceanography, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093; and [c]Marine Science and Environmental Studies Department, University of San Diego, 5998 Alcalá Park, San Diego, CA 92110

Edited by Michael Goodchild, University of California, Santa Barbara, CA, and approved November 22, 2010 (received for review August 20, 2009)

We describe the results from a spatial cyberinfrastructure developed to characterize the meltwater field around individual icebergs and integrate the results with regional- and global-scale data. During the course of the cyberinfrastructure development, it became clear that we were also building an integrated sampling planning capability across multidisciplinary teams that provided greater agility in allocating expedition resources resulting in new scientific insights. The cyberinfrastructure-enabled method is a complement to the conventional methods of hydrographic sampling in which the ship provides a static platform on a station-by-station basis. We adapted a sea-floor mapping method to more rapidly characterize the sea surface geophysically and biologically. By jointly analyzing the multisource, continuously sampled biological, chemical, and physical parameters, using Global Positioning System time as the data fusion key, this *surface-mapping* method enables us to examine the relationship between the meltwater field of the iceberg to the larger-scale marine ecosystem of the Southern Ocean. Through geospatial data fusion, we are able to combine very fine-scale maps of dynamic processes with more synoptic but lower-resolution data from satellite systems. Our results illustrate the importance of spatial cyberinfrastructure in the overall scientific enterprise and identify key interfaces and sources of error that require improved controls for the development of future Earth observing systems as we move into an era of peta- and exascale, data-intensive computing.

Antarctica | remote sensing | hydrography | surface water

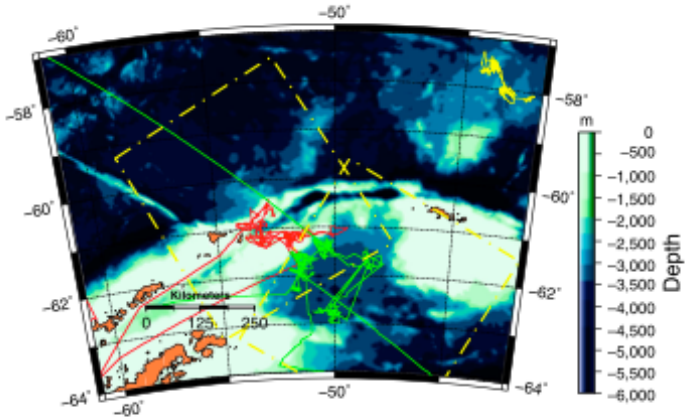
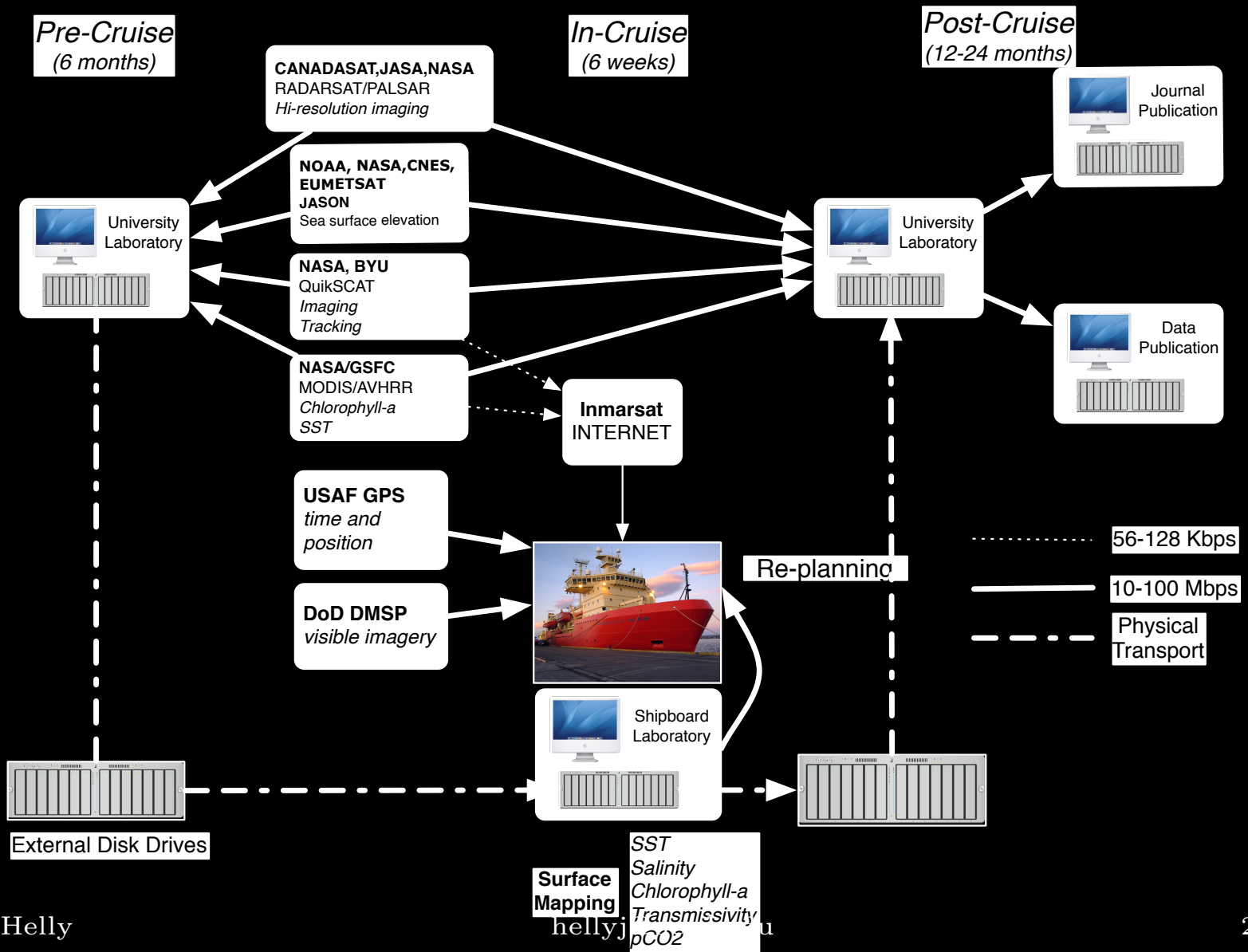


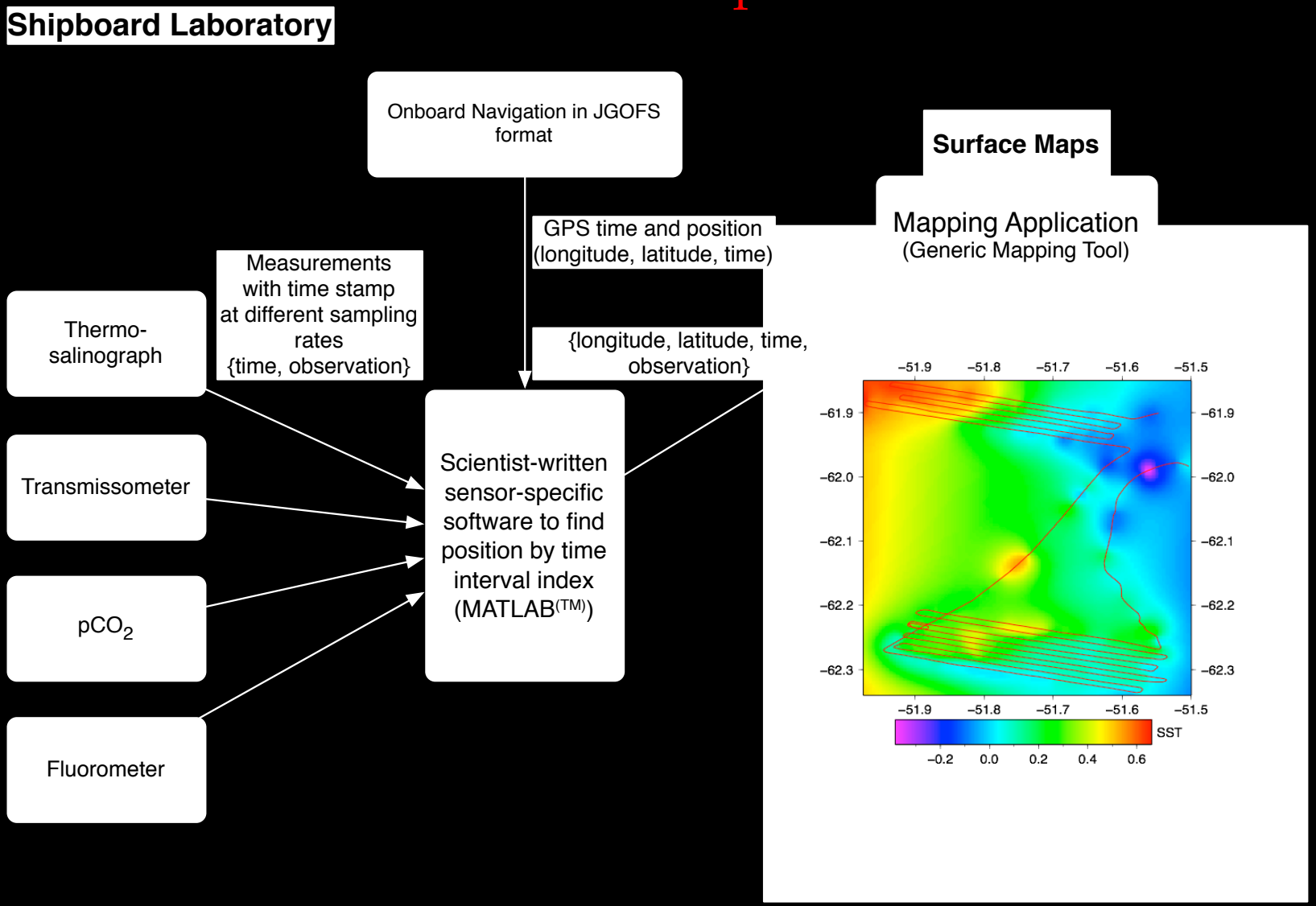
Fig. 1. Regional map of the study area in the Weddell Sea and Scotia Sea showing the GPS-based ship track for the three research cruises, LMG0514a (red), NBP0806 (yellow), NBP0902 (green), as well two representative RADARSAT-1 image frame boundaries (dashed yellow, wideband mode). Bathymetry is from the Global Topography 11.1 (1).

The characterization of the surface of the ocean with physical, chemical and biological data in the vicinity of a free-drifting iceberg poses a number of challenges. Besides being large, icebergs are affected by geophysical forces that alter their structure and movement: solar radiation, Earth's rotation, tides and currents, and winds and storms (5). As free-drifting, tabular icebergs pro-

Digital Libraries Work In University Labs and On Ships in Antarctica



... and Digital Libraries Work in Labs on Ships in Antarctica



Digital Libraries Work for All Kinds of Scientific Projects

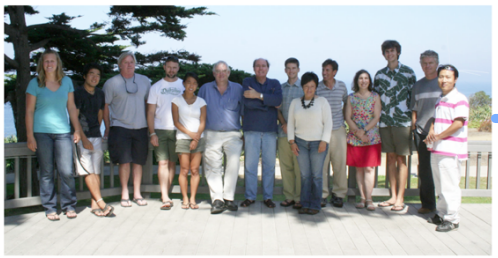
Scientific Research Projects

Quality Review

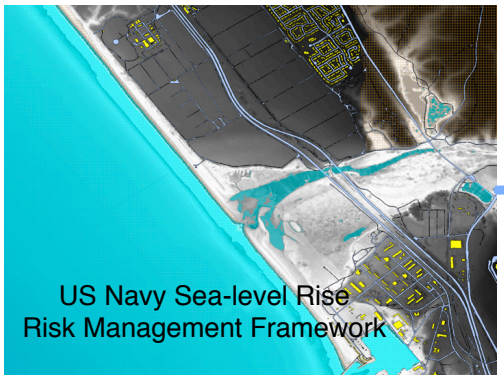
Sound Published Data, Citable & Cross-referenced

Coastal Hypoxia and Ocean Acidification

UCMexus

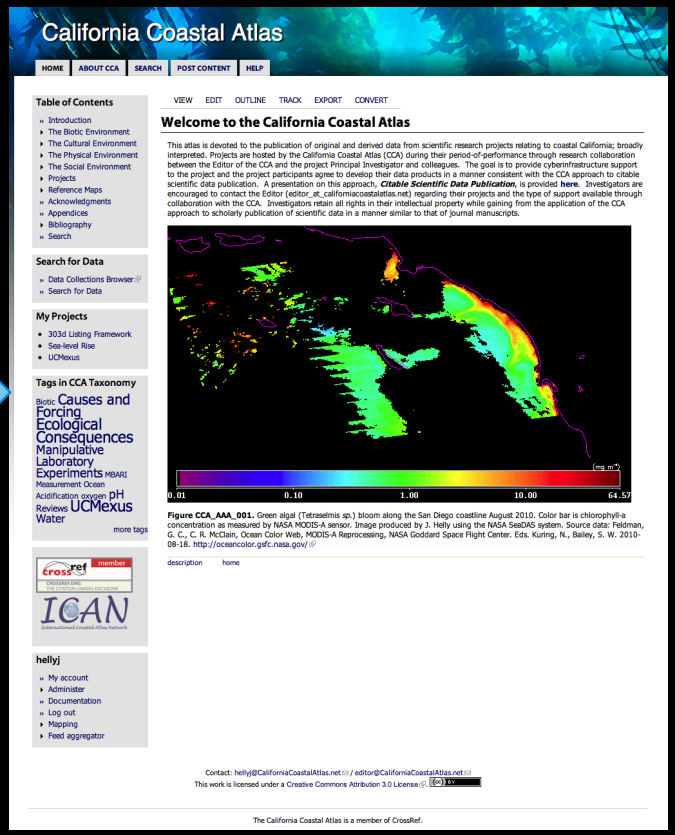


US Navy Sea-level Rise Risk Management Framework

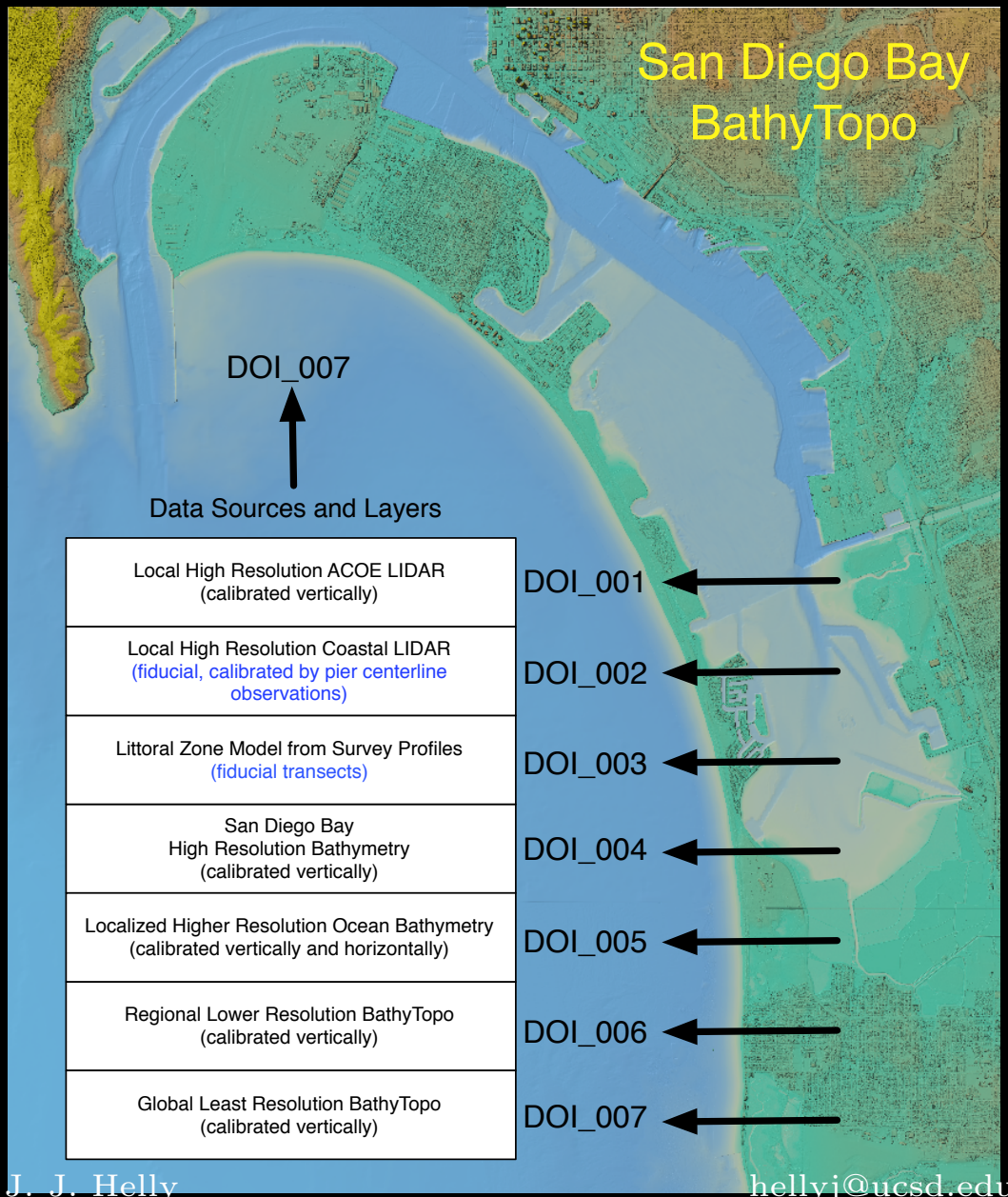


Scientific Editorial Quality Control

Scientific Editorial Quality Control

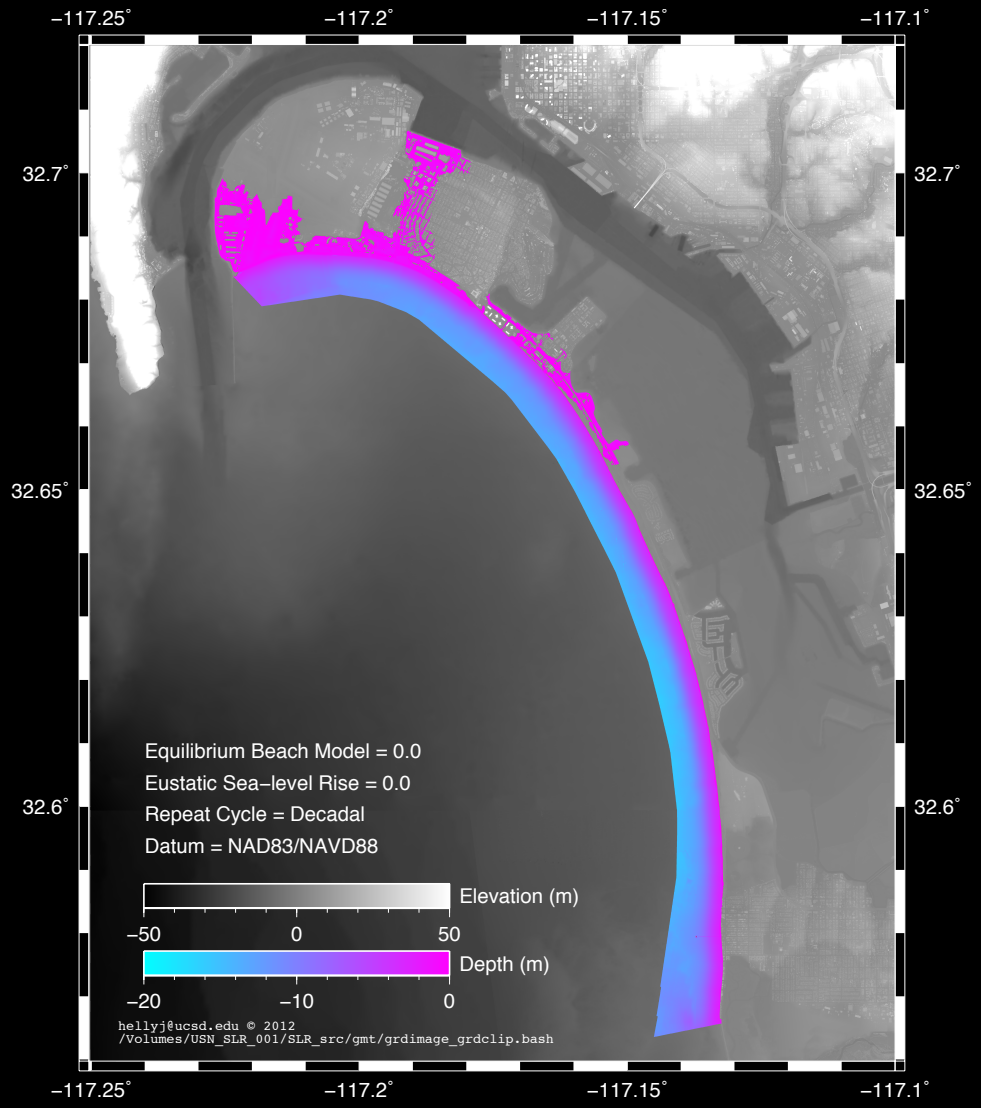


Multi-source Composite Example



- 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

Analyses Can Be Traced Back All Parts of the Processing



- Underlying basemap
- Model configuration
- Software used for post-processing

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*

Amazing New Opportunities Are Enabled Through Data Publication

- Data Fusion (conversely 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

Academic Senate
San Diego Division
UNIVERSITY OF CALIFORNIA, SAN DIEGO

Quick Links

Assembly | Committees | Staff | Location | About Us | Links

PROPOSED UC OPEN ACCESS POLICY

The University Committee on Library and Scholarly Communication (UCOLASC) presented a draft policy on Open Access Publishing to the Academic Council on July 25th. The policy has implications for faculty publishing, for broadening accessibility to scholarship, and for how the university is perceived. The UCSD Academic Senate Committee on the Library formulated an initial response to the policy draft and is now soliciting a campus-wide faculty response through Friday, November 9th.

The Policy

The proposed policy, which applies only to scholarly articles, has two key components:

1. The policy requires that a faculty member grant to the University a non-exclusive license to exercise any and all publishing rights allowed under copyright.
2. The policy requires that a faculty member deposit a version of their publication with either a University or another open access repository.

Books and other types of intellectual/creative work are not included in the current proposal. With some minor variations, UCSF, Harvard, Duke, Princeton, MIT, and more than 140 other institutions worldwide have already adopted this approach to Open Access. These adoptions are part of a wider trend toward Open Access begun by the NIH and the Wellcome Trust and as most recently exemplified by recent developments in the U.K.

Impact of the Policy

In simple terms, the upside of the proposed policy is that it would make scholarship available more broadly and permit the creation of new models for access to knowledge. The implication is that scholarly publishing would follow newspapers, music, and books in how the Internet has changed business models. What is clear, and has been for a very long time, is that current business models for scholarly publishing are unsustainable – the cost of closed access journals is increasing much faster that either the consumer or the higher education price indices, an unsustainable situation exacerbated by shrinking library budgets.

The downside of the proposed policy is, first, the burden it will place on faculty authors that does not exist today: namely uploading the manuscript to an archive or opting out. Second, the policy may have undesirable consequences for societies or other membership organizations that rely on closed-access publishing revenues to underwrite the cost of other services they 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

- Libraries have historically defined *the University*
- The role of university libraries is changing
- Scripps Institution of Oceanography Library closed this year
- The Water Resources collection at Berkeley went homeless within the last few years

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

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)

Outline

1 What's Unique About Scientific Data Management?

2 The Scientific Method And Reproducibility

3 Digital Library Framework