Data Curation @ UCSB:
The Prequel

Greg Janée

August 8, 2012

Prologue

- *World Wide Web revolution*
  - invented
    → novel
      → widespread
        → commonplace
          → expected
            (→ a right?)

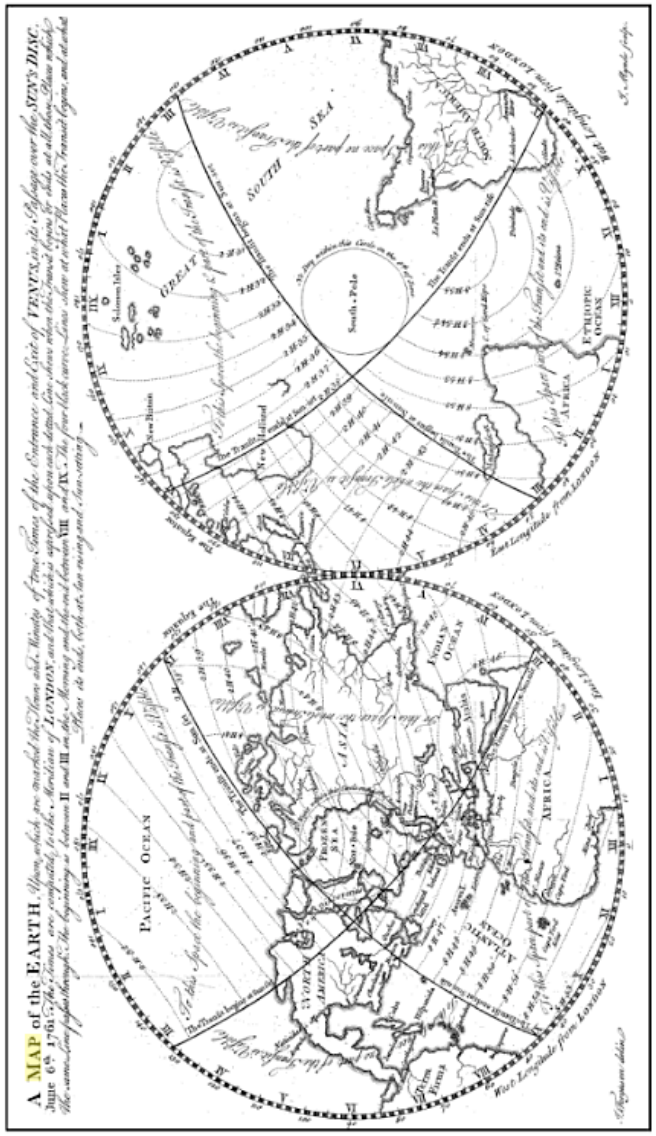

A Magazine Is an iPad That Does Not Work.m4v - YouT
http://www.youtube.com/watch?feature=player_embedded&v=aXV-yaFmQNk
Work
Reference
News
YouTube
A Magazine Is an iPad That Does Not Work.m4v
UserExperiencesWorks
Subscribe
3 videos
0:43 / 1:26

CHASING VENUS

THE RACE TO MEASURE the Heavens

ANDREA WULF

IN 1716, *British astronomer Edmond Halley published a ten-page essay which called upon scientists to unite in a project spanning the entire globe-one that would change the world of science forever. On 6 June 1761, Halley predicted, Venus would traverse the face of the sun-for a few hours the bright star would appear as a perfectly black circle against the burning...*



A mappemonde from 1770. Delisle's version would have had regions coloured to represent the visibility of the transit.

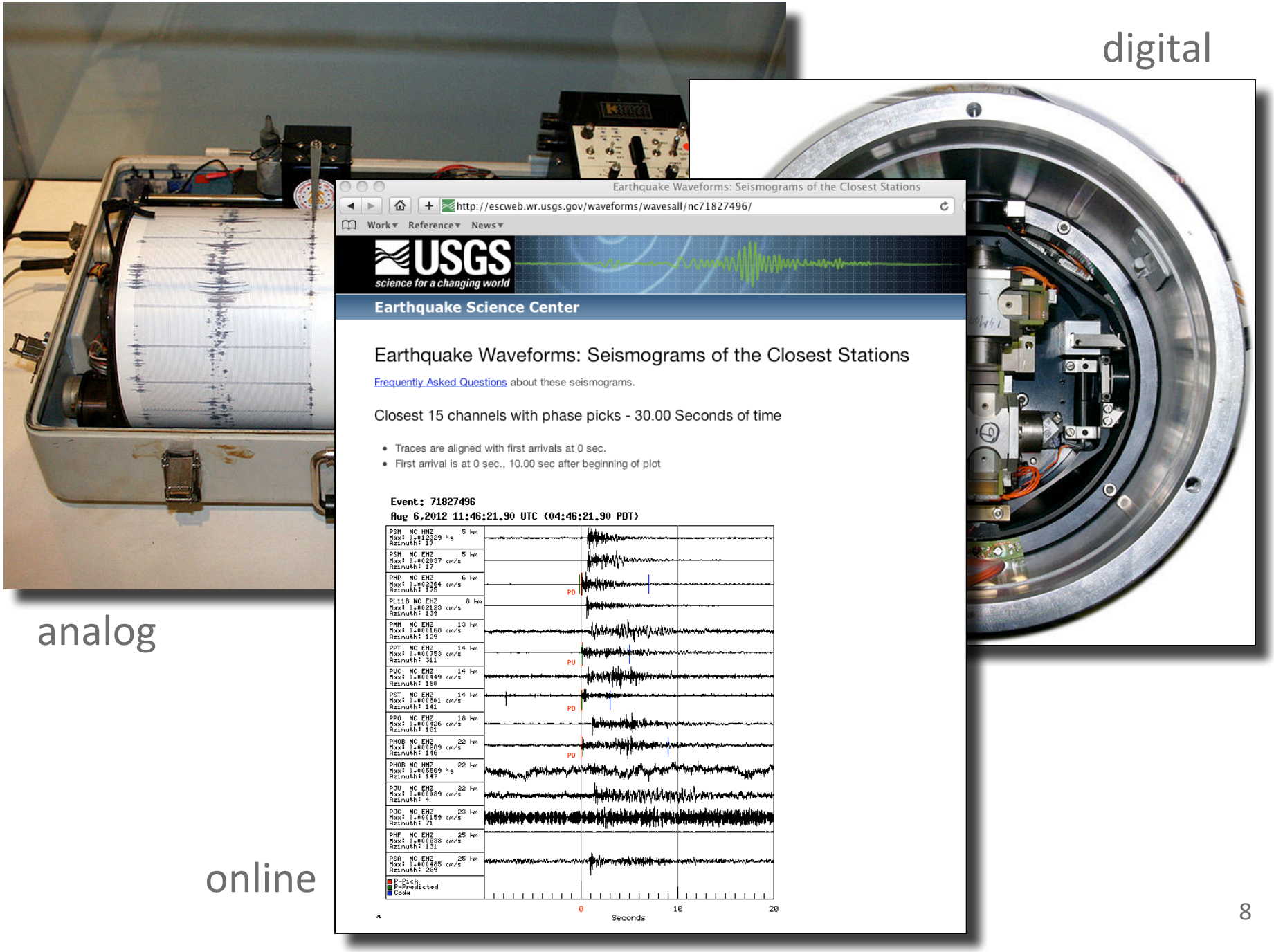
9


Advanced zoom – Mappemonde sur la quelle on a marqué les heures et les minutes d…passage sur cet astre le 6 Juin 1761, ces tems sont comptez au meridien de Paris
http://objects.library.uu.nl/reader/viewer.php?obj=1874-33170&pagenum=1&lan=en
Google
Work
Reference
News
Universiteit Utrecht
Item info - Help
| Mappemonde sur la quelle on a marqué les heures et les minutes du tems vrai de l'entrée et de la s....
EMONDE Sur la quelle on a marque
centre de VENUS sur le disque du SOLEIL
comptez au Meridien de PARIS . par Mr.
On a expliqué l'usage des nombres et des Couleurs qui sont
Cette Carte se trouve chez l'Auteur avec la Description imprimée au College Royal
Nouvelles
Découvertes
Nouvelle
Kebec
Louisbourg
Barton
© UBU

Science data revolution, part 1

- *Online revolution*
  - analog → digital → online
  - cutting edge → commonplace → expected
- *New expectations:*
  - online
  - instantly and forever available
  - (re)usable
  - discoverable
  - identified and citable
  - hyperlinked into fabric of scholarly communication

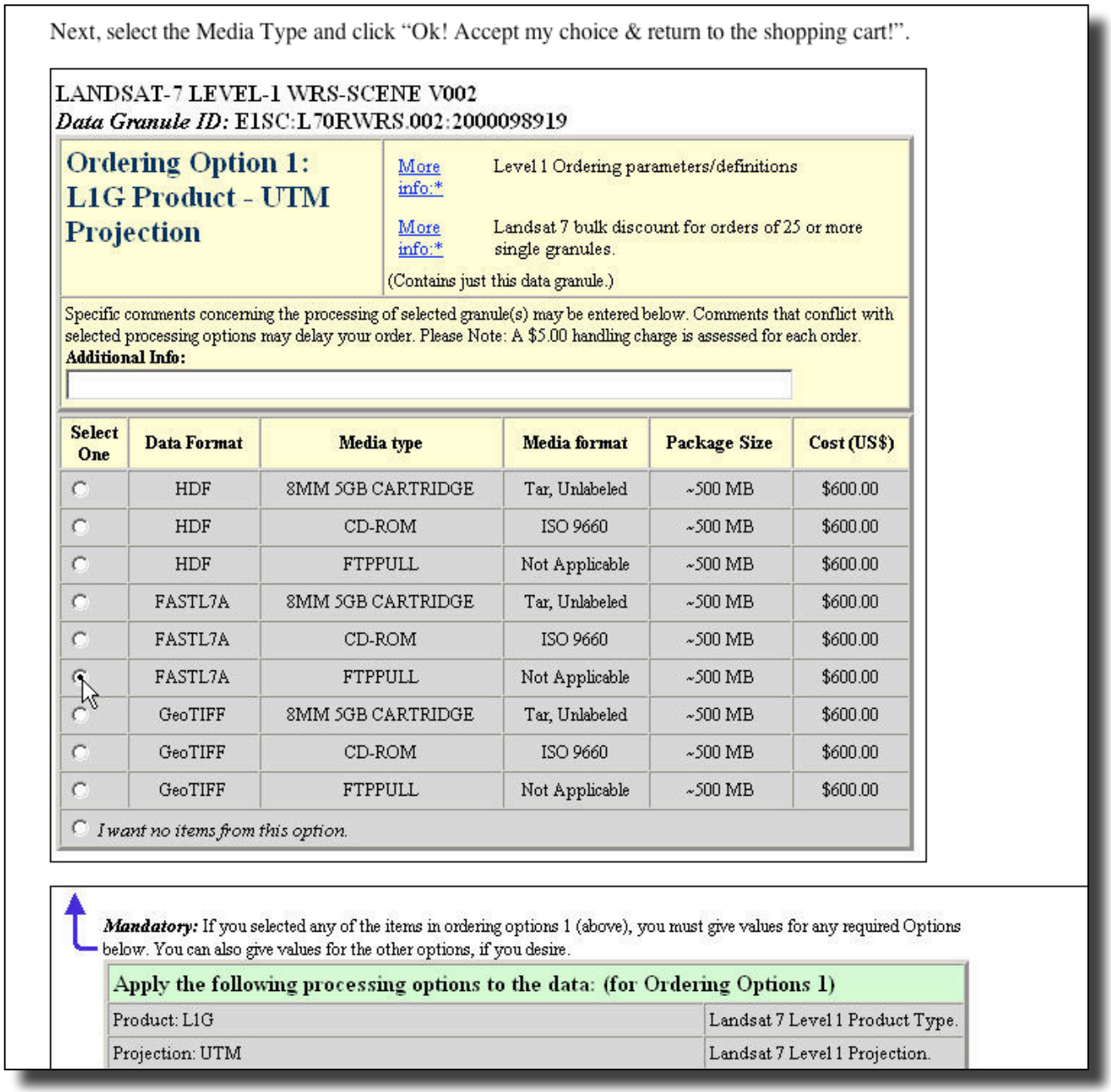
digital



analog

online

~2006: offline access



Next, select the Media Type and click "Ok! Accept my choice & return to the shopping cart!".

LANDSAT-7 LEVEL-1 WRS-SCENE V002
Data Granule ID: E1SC:L70RWRS.002:2000098919

Ordering Option 1: L1G Product - UTM Projection

More info:* Level 1 Ordering parameters/definitions

More info:* Landsat 7 bulk discount for orders of 25 or more single granules.

(Contains just this data granule.)

Specific comments concerning the processing of selected granule(s) may be entered below. Comments that conflict with selected processing options may delay your order. Please Note: A $5.00 handling charge is assessed for each order.
Additional Info:

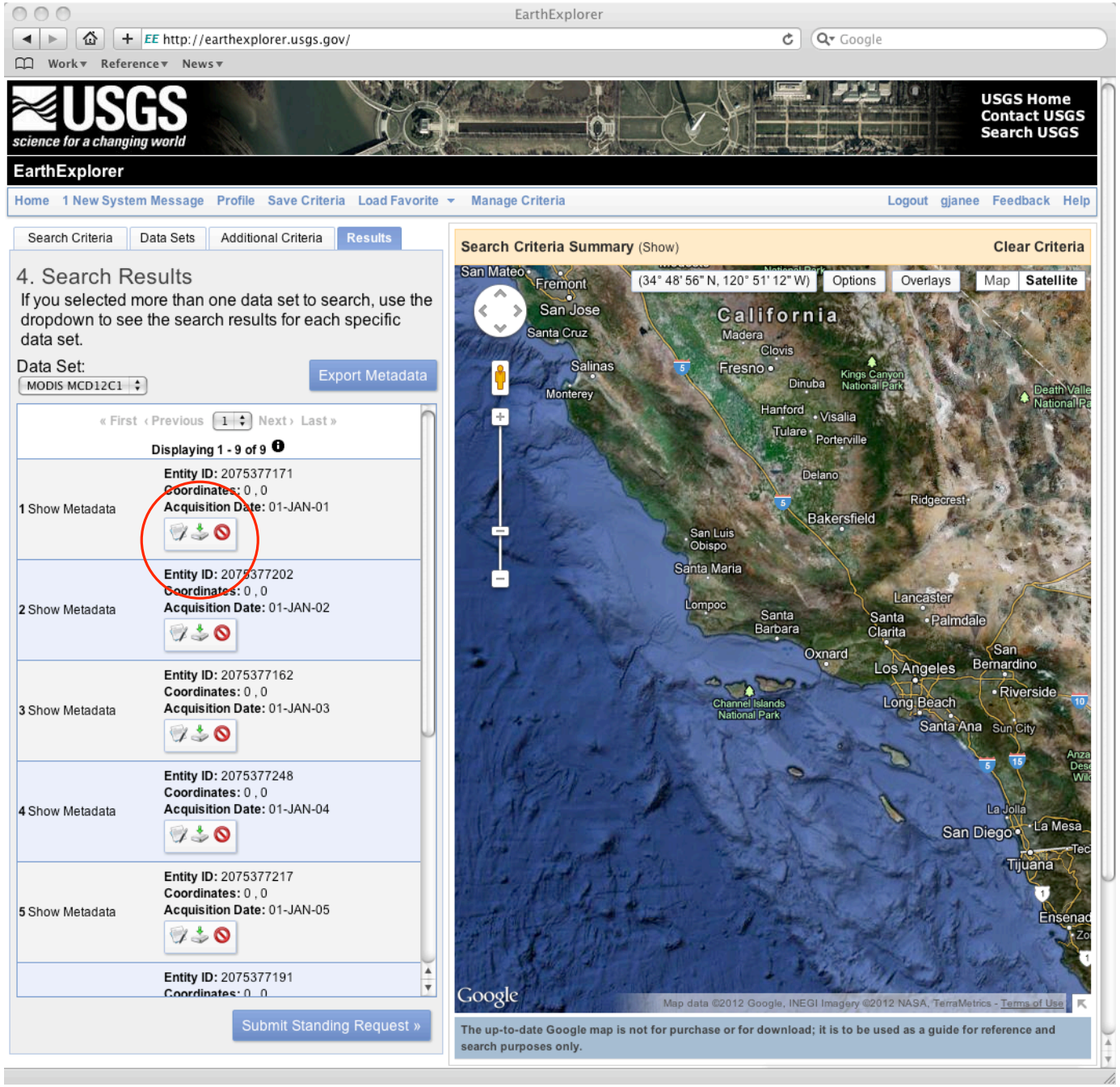
Select One	Data Format	Media type	Media format	Package Size	Cost (US$)
○	HDF	8MM 5GB CARTRIDGE	Tar, Unlabeled	~500 MB	$600.00
○	HDF	CD-ROM	ISO 9660	~500 MB	$600.00
○	HDF	FTPPULL	Not Applicable	~500 MB	$600.00
○	FASTL7A	8MM 5GB CARTRIDGE	Tar, Unlabeled	~500 MB	$600.00
○	FASTL7A	CD-ROM	ISO 9660	~500 MB	$600.00
○	FASTL7A	FTPPULL	Not Applicable	~500 MB	$600.00
○	GeoTIFF	8MM 5GB CARTRIDGE	Tar, Unlabeled	~500 MB	$600.00
○	GeoTIFF	CD-ROM	ISO 9660	~500 MB	$600.00
○	GeoTIFF	FTPPULL	Not Applicable	~500 MB	$600.00

○ *I want no items from this option.*

Mandatory: If you selected any of the items in ordering options 1 (above), you must give values for any required Options below. You can also give values for the other options, if you desire.

Apply the following processing options to the data: (for Ordering Options 1)	
Product: L1G	Landsat 7 Level 1 Product Type.
Projection: UTM	Landsat 7 Level 1 Projection.

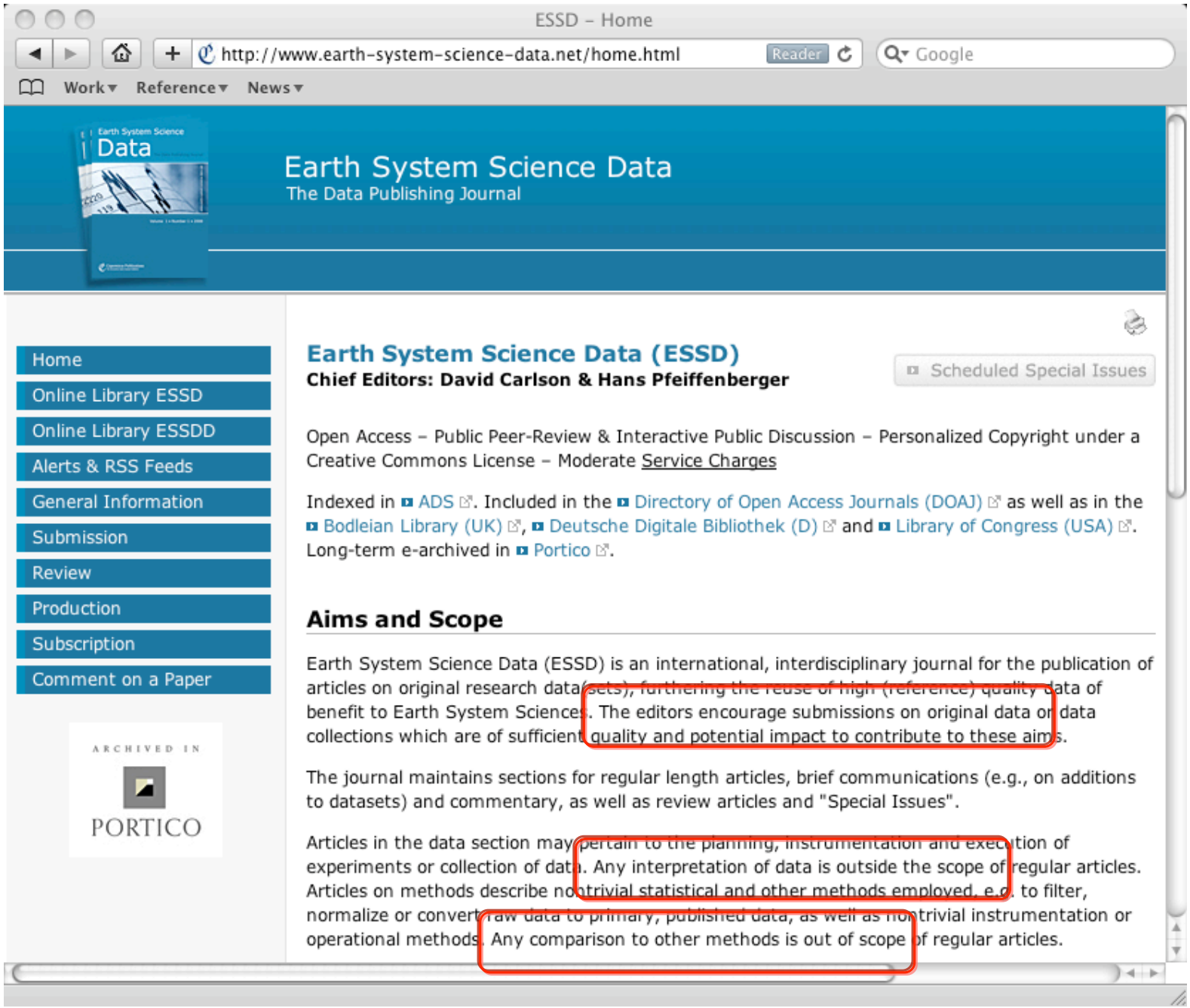
Today: online, immediate access

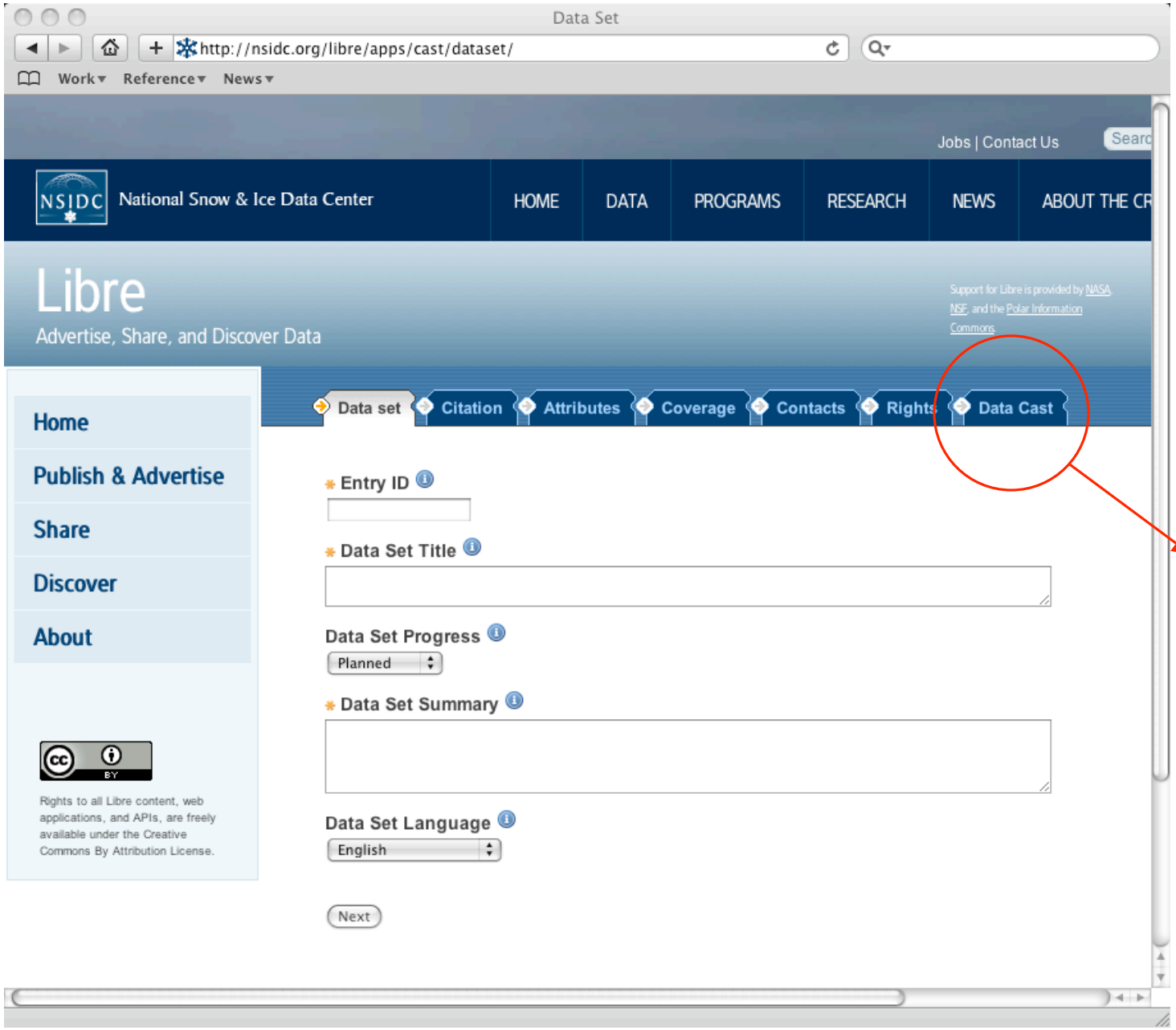


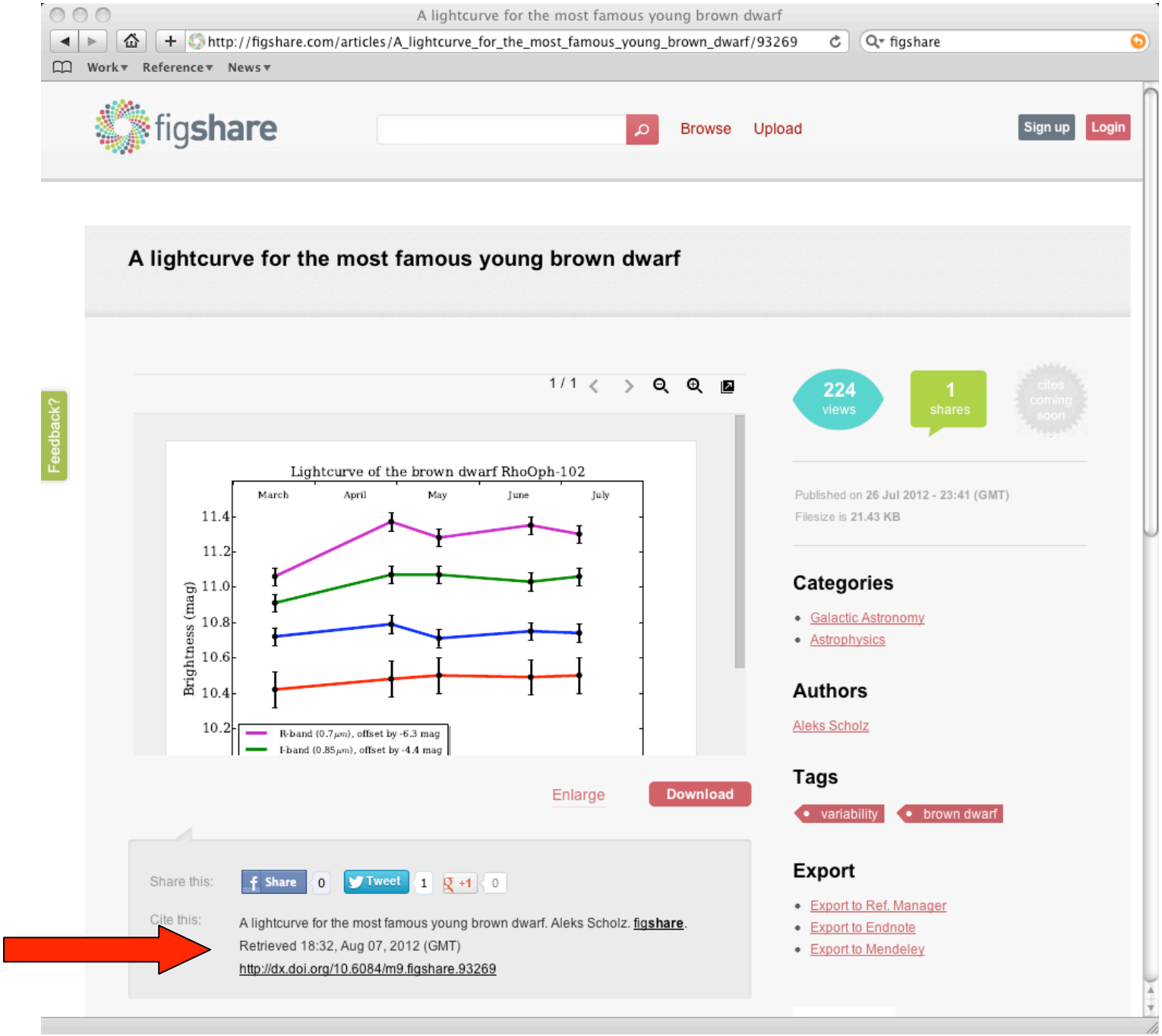
Identified,
citable,
linked,
<u>expected</u>

Science data revolution, part 2

- *Datasets increasingly seen as publishable works*
  - scholarly works in and of themselves
  - datasets are “published” and “cited”
- *Motivations*
  - give credit to data creators
  - track dataset impact (citation metrics…)
  - accountability
  - aid reproducibility
- ⇒ *New publication mechanisms*
  - ranging from formal/traditional
  - to less formal/innovative/revolutionary


ESSD – Home
http://www.earth-system-science-data.net/home.html
Reader
Google
Work
Reference
News
Earth System Science Data
The Data Publishing Journal
Home
Online Library ESSD
Online Library ESSDD
Alerts & RSS Feeds
General Information
Submission
Review
Production
Subscription
Comment on a Paper
ARCHIVED IN
PORTICO
Earth System Science Data (ESSD)
Chief Editors: David Carlson & Hans Pfeiffenberger
Scheduled Special Issues
Open Access – Public Peer-Review & Interactive Public Discussion – Personalized Copyright under a Creative Commons License – Moderate Service Charges
Indexed in ADS. Included in the Directory of Open Access Journals (DOAJ) as well as in the Bodleian Library (UK), Deutsche Digitale Bibliothek (D) and Library of Congress (USA). Long-term e-archived in Portico.
Aims and Scope
Earth System Science Data (ESSD) is an international, interdisciplinary journal for the publication of articles on original research data(sets), furthering the reuse of high (reference) quality data of benefit to Earth System Sciences. The editors encourage submissions on original data or data collections which are of sufficient quality and potential impact to contribute to these aims.
The journal maintains sections for regular length articles, brief communications (e.g., on additions to datasets) and commentary, as well as review articles and "Special Issues".
Articles in the data section may pertain to the planning, instrumentation and execution of experiments or collection of data. Any interpretation of data is outside the scope of regular articles. Articles on methods describe nontrivial statistical and other methods employed, e.g. to filter, normalize or convert raw data to primary, published data, as well as nontrivial instrumentation or operational methods. Any comparison to other methods is out of scope of regular articles.


Data Set
http://nsidc.org/libre/apps/cast/dataset/
Work
Reference
News
Jobs | Contact Us
NSIDC
National Snow & Ice Data Center
HOME
DATA
PROGRAMS
RESEARCH
NEWS
Libre
Advertise, Share, and Discover Data
Support for Libre is provided by NASA, NSF, and the Polar Information Commons.
Home
Publish & Advertise
Share
Discover
About
Data set
Citation
Attributes
Coverage
Contacts
Rights
Data Cast
Entry ID
Data Set Title
Data Set Progress
Planned
Data Set Summary
Data Set Language
English
Next
Rights to all Libre content, web applications, and APIs, are freely available under the Creative Commons By Attribution License.
Atom feed


A lightcurve for the most famous young brown dwarf
http://figshare.com/articles/A_lightcurve_for_the_most_famous_young_brown_dwarf/93269
figshare
Work
Reference
News
figshare
Browse
Upload
Sign up
Login
A lightcurve for the most famous young brown dwarf
Feedback?
1 / 1
Lightcurve of the brown dwarf RhoOph-102
March
April
May
June
July
Brightness (mag)
11.4
11.2
11.0
10.8
10.6
10.4
10.2
R-band (0.7μm), offset by -6.3 mag
I-band (0.85μm), offset by -4.4 mag
Enlarge
Download
224
views
1
shares
cites coming soon
Published on 26 Jul 2012 - 23:41 (GMT)
Filesize is 21.43 KB
Categories
Galactic Astronomy
Astrophysics
Authors
Aleks Scholz
Tags
variability
brown dwarf
Export
Export to Ref. Manager
Export to Endnote
Export to Mendeley
Share this:
Share
0
Tweet
1
+1
0
Cite this:
A lightcurve for the most famous young brown dwarf. Aleks Scholz. figshare.
Retrieved 18:32, Aug 07, 2012 (GMT)
http://dx.doi.org/10.6084/m9.figshare.93269

Refining data publication concepts

- *Parsons & Fox:*
  - "Data 'publication' is all the rage. Data authors and stewards rightfully seek recognition for the intellectual effort they invest in creating a good data set [...] As a result, people look to scholarly publication—a well-established, scientific process—as a possible analog for sharing data."
  - http://mp-datamatters.blogspot.com/2011/12/seeking-open-review-of-provocative-data.html

Science data revolution, part 3

- *Science increasingly:*
  - data-driven
  - cross-disciplinary
- *Facilitated by existence of online, well-curated data*

Data-driven science

- *Tenopir et al (doi:10.1371/journal.pone.0021101)*
  - survey of 1329 scientists
  - “Scientific research in the 21st century is more data intensive and collaborative than in the past.”

Summary so far

- *New set of expectations for science data*
  - online, available, citable, linked
- *New demands*
  - data as new kind of publication
  - data-driven science

Summary so far

- *New set of expectations for science data*
  - online, available, citable, linked
- *New demands*
  - data as new kind of publication
  - data-driven science
- ⇒ *Driving need for data curation:*
  - preserving data
  - ensuring its integrity
  - supporting identification, discoverability
  - maintaining meaningful, useful access
  - such that the next steward can do the same
- ⇒ *But data curation is not easy…*

Curation challenge 1: storage

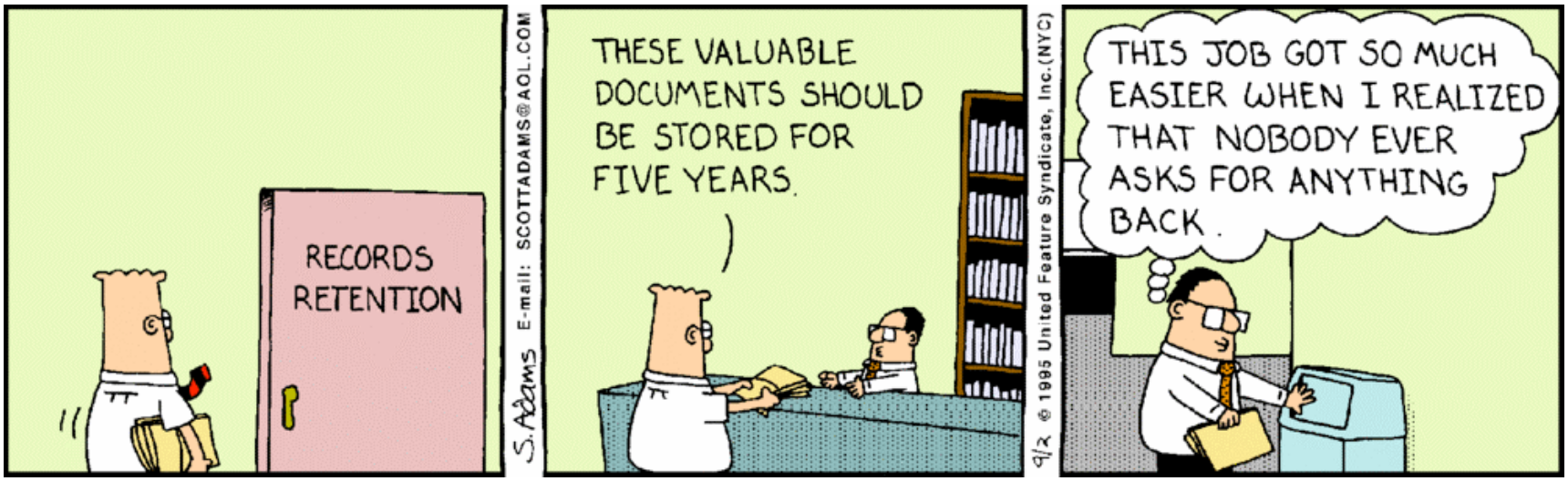
- *Bit storage problem is not solved!*
- *David Rosenthal's challenge: save a petabyte for a century with 50% probability of no bit loss*
  - "…we would need to improve the performance of current enterprise storage systems by a factor of at least 10^9"
  - problem: disk sizes (10^{13} bits) are matching bit error rate (10^{-14})

Storage

- *Not quantifiable yet*
- *State of the art recommendation:*
  - the more copies, the safer
  - the more independent copies, the safer
  - the more frequently the copies are audited, the safer
- *Implication for curation:*
  - preservation requires dedicated institutions, specialists

Auditing storage

- *Dilbert:*

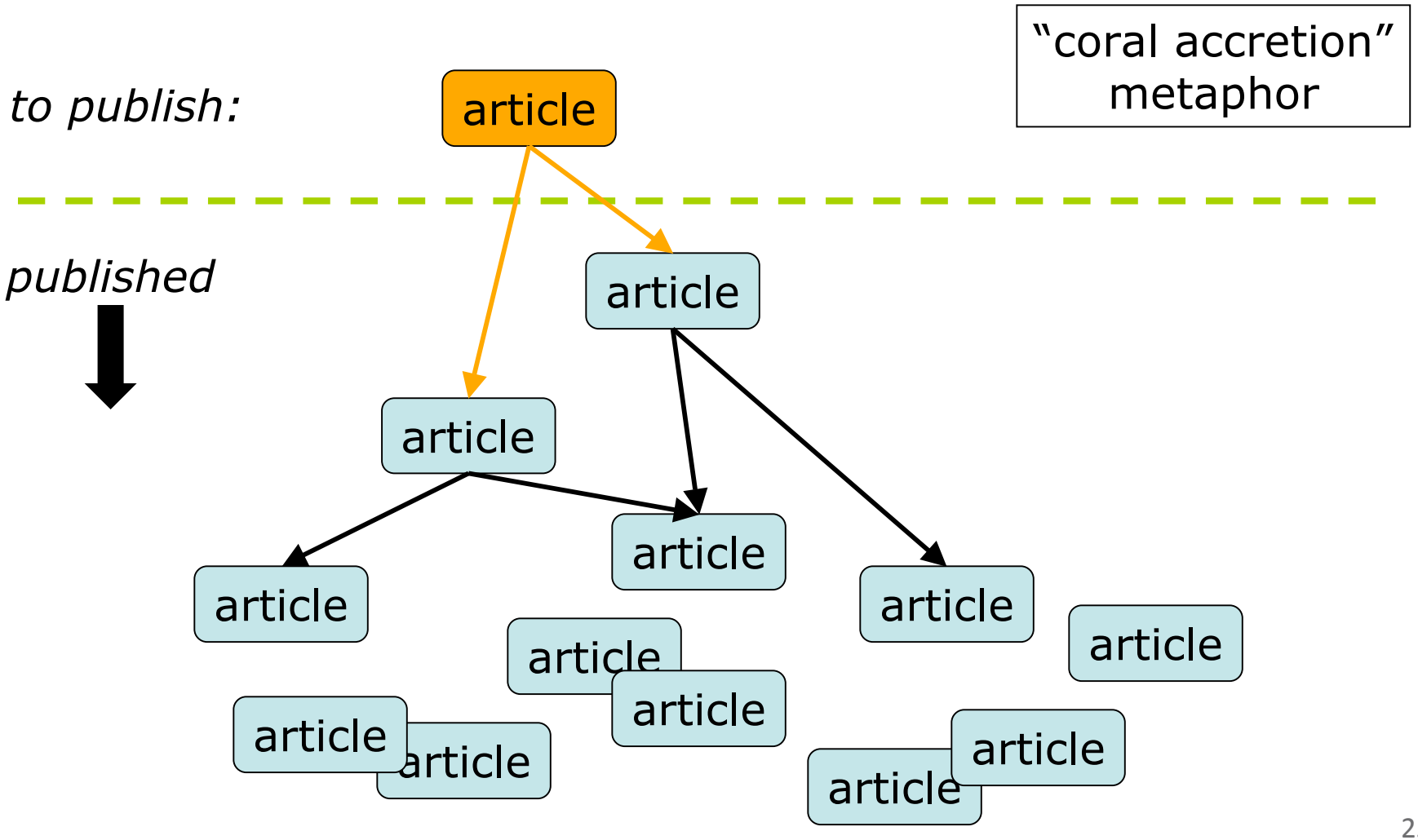


- *Greg's law of backups:*
  - if you haven't verified a backup, you don't know that you have a backup

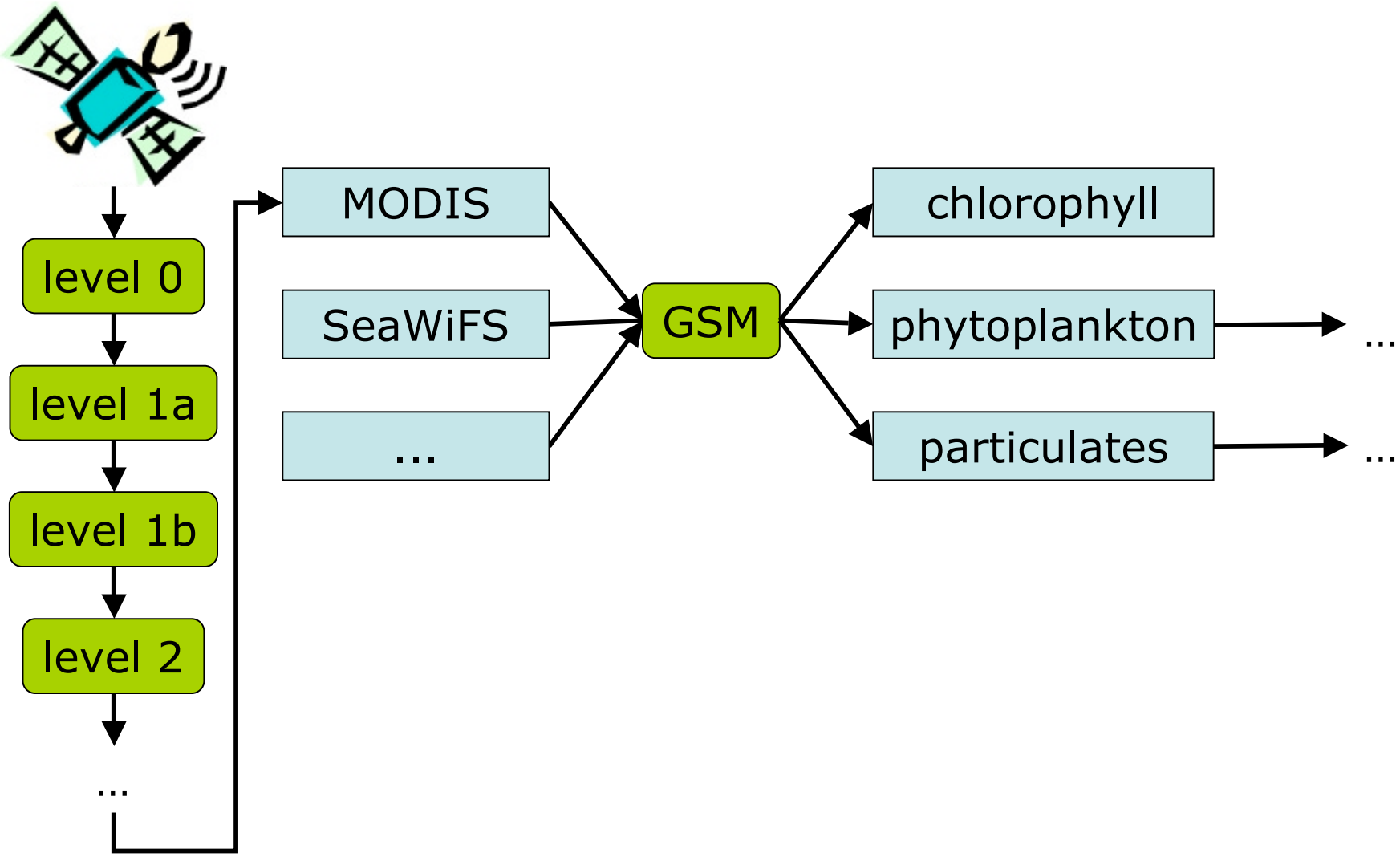
Curation challenge 2: complexity

- *Science data is more complex than publications*
  - larger scales
  - at both large and small scales
  - ergo, curating data is more difficult
- *Scholarly publication characteristics*
  - article is created, reviewed, accepted
  - one, well-defined "publish" event
  - creator, creation-related artifacts play no role after publishing
  - article (PDF file) is single unit of reuse; "use" = viewability
  - article remains unchanged… forever

Scholarly literature publishing



Earth science data workflows



Earth science data — analysis

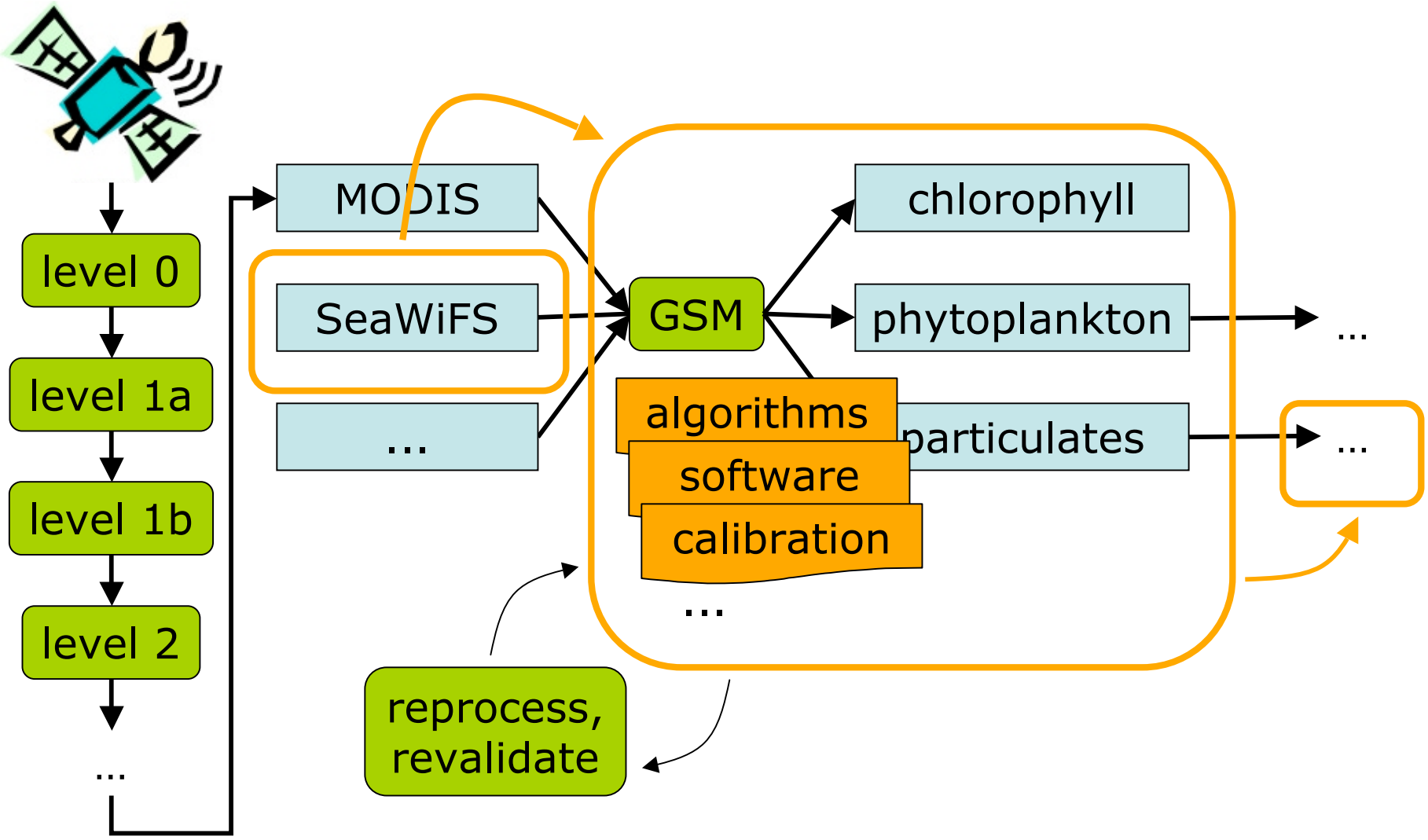
- *Characterized by workflows*
  - provenance is important
- *Dynamic*
  - time series may extend for decades
  - reprocessing

The curse of reprocessing

- *SeaWiFS*
  - Reprocessing 5.2 - Completed July 12, 2007
  - Reprocessing 5.1 - Completed July 5, 2005
  - Reprocessing 5 - Completed March 18, 2005
  - Reprocessing 4[illegible] - Completed May 24, 2004
  - Reprocess[illegible]02
  - Reprocess[illegible]00
    - Calibration Update - December 1, 2000
    - Calibration Update - April 10, 2001
  - Reprocessing 2 - August, 1998
  - Reprocessing 1 - January, 1998

new atmospheric, solar irradiance models

Earth science data workflows



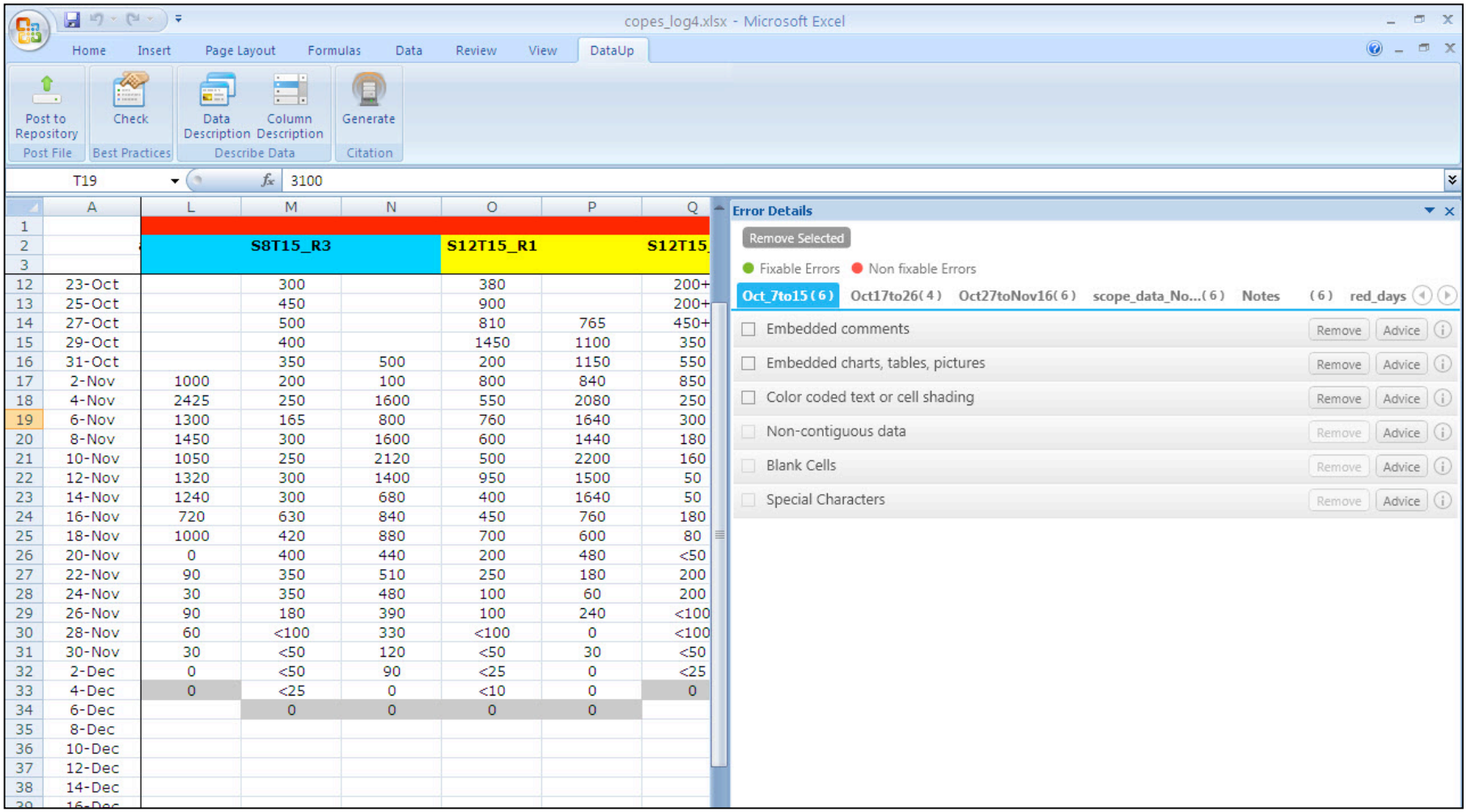
Earth science data — analysis

- *Characterized by workflows*
  - Provenance is important
- *Dynamic*
  - time series extend for decades
  - reprocessing
- *Versioning important*
  - strong incentive to move to new versions
- *Implications*
  - huge informatics complications
    - identification
    - provenance
    - management
  - curation involves partnership between authors, archives

Small-scale complexity

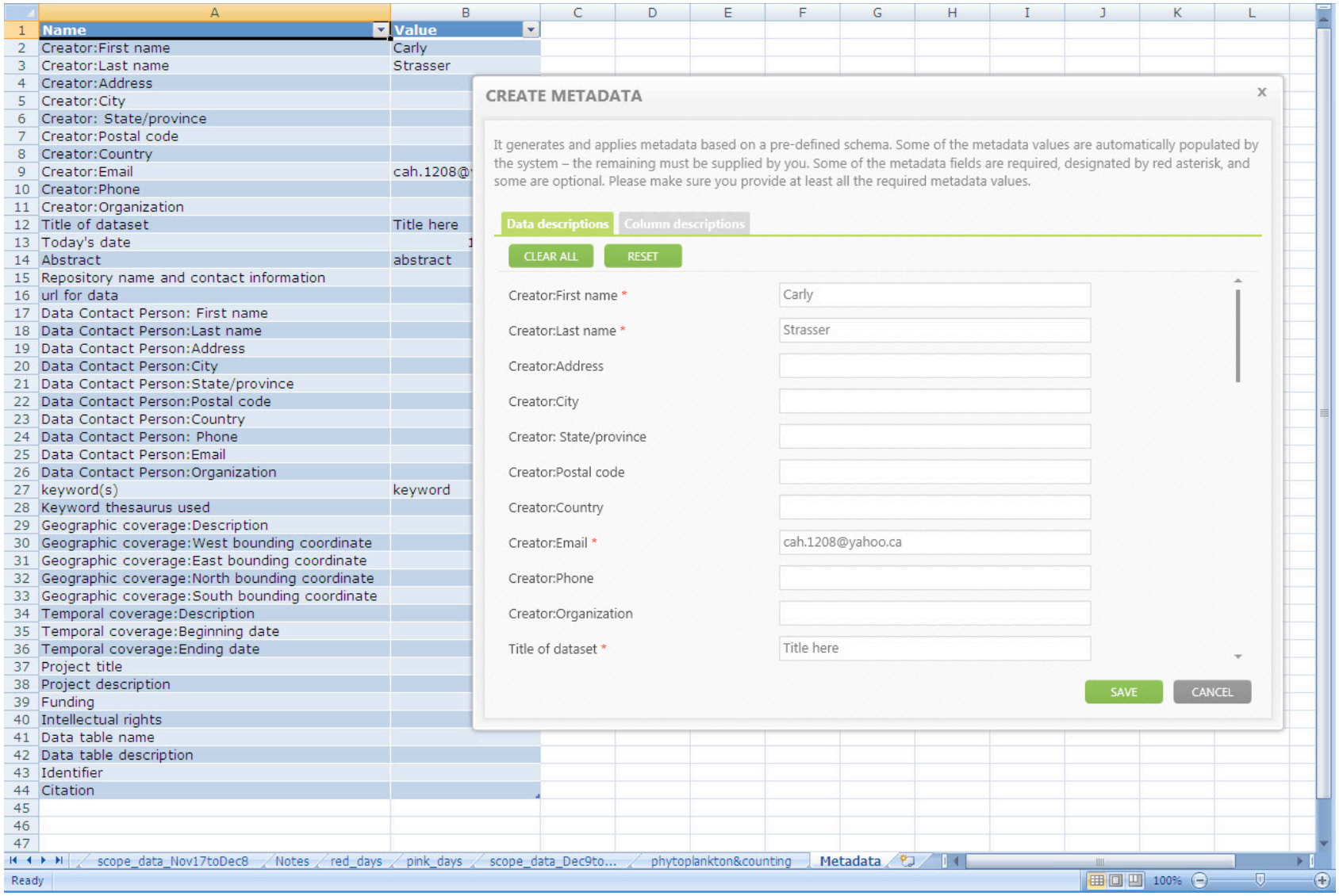
- *"Use" requires deeper knowledge, manipulation*
  - ergo, increased metadata, access requirements
- *CDL DataUp (née DCXL) project*
  - premise: spreadsheets are a mess
    - you know it's true
  - developing Excel plugin/online service to increase curate-ability

Best practices check



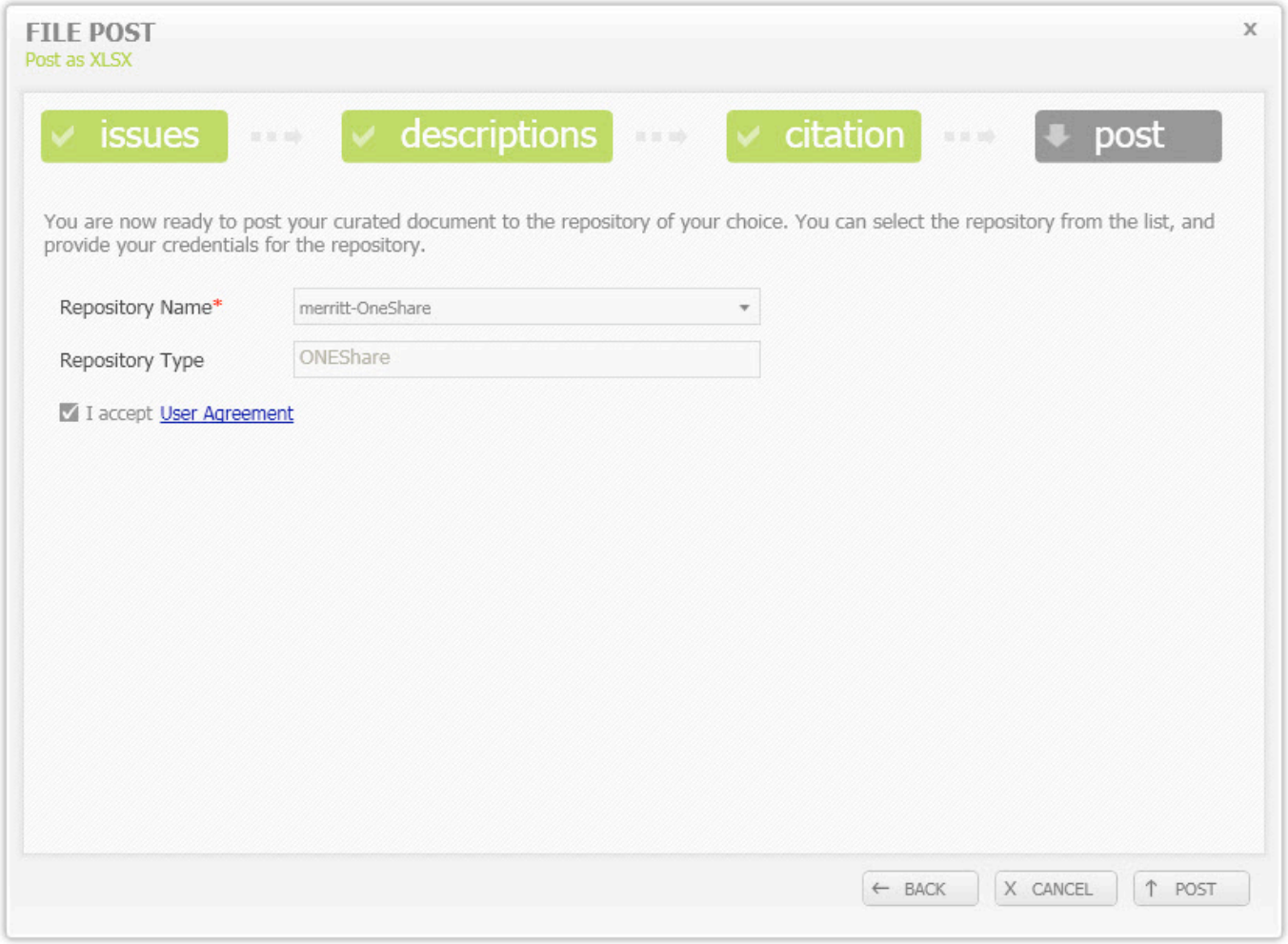
	A	L	M	N	O	P	Q
12	23-Oct		300		380		200+
13	25-Oct		450		900		200+
14	27-Oct		500		810	765	450+
15	29-Oct		400		1450	1100	350
16	31-Oct		350	500	200	1150	550
17	2-Nov	1000	200	100	800	840	850
18	4-Nov	2425	250	1600	550	2080	250
19	6-Nov	1300	165	800	760	1640	300
20	8-Nov	1450	300	1600	600	1440	180
21	10-Nov	1050	250	2120	500	2200	160
22	12-Nov	1320	300	1400	950	1500	50
23	14-Nov	1240	300	680	400	1640	50
24	16-Nov	720	630	840	450	760	180
25	18-Nov	1000	420	880	700	600	80
26	20-Nov	0	400	440	200	480	<50
27	22-Nov	90	350	510	250	180	200
28	24-Nov	30	350	480	100	60	200
29	26-Nov	90	180	390	100	240	<100
30	28-Nov	60	<100	330	<100	0	<100
31	30-Nov	30	<50	120	<50	30	<50
32	2-Dec	0	<50	90	<25	0	<25
33	4-Dec	0	<25	0	<10	0	0
34	6-Dec		0	0	0	0	
35	8-Dec						
36	10-Dec						
37	12-Dec						
38	14-Dec						

Metadata generation



	Name	Value
2	Creator:First name	Carly
3	Creator:Last name	Strasser
4	Creator:Address	
5	Creator:City	
6	Creator: State/province	
7	Creator:Postal code	
8	Creator:Country	
9	Creator:Email	cah.1208@
10	Creator:Phone	
11	Creator:Organization	
12	Title of dataset	Title here
13	Today's date	
14	Abstract	abstract
15	Repository name and contact information	
16	url for data	
17	Data Contact Person: First name	
18	Data Contact Person:Last name	
19	Data Contact Person:Address	
20	Data Contact Person:City	
21	Data Contact Person:State/province	
22	Data Contact Person:Postal code	
23	Data Contact Person:Country	
24	Data Contact Person: Phone	
25	Data Contact Person:Email	
26	Data Contact Person:Organization	
27	keyword(s)	keyword
28	Keyword thesaurus used	
29	Geographic coverage:Description	
30	Geographic coverage:West bounding coordinate	
31	Geographic coverage:East bounding coordinate	
32	Geographic coverage:North bounding coordinate	
33	Geographic coverage:South bounding coordinate	
34	Temporal coverage:Description	
35	Temporal coverage:Beginning date	
36	Temporal coverage:Ending date	
37	Project title	
38	Project description	
39	Funding	
40	Intellectual rights	
41	Data table name	
42	Data table description	
43	Identifier	
44	Citation	

Upload to repository



Curation challenge 3

- *Desirable knowledge*
  - authenticity, quality, uses (appropriate and historical), reputation, provenance
- *Easily obtained when creator = provider*
  - relatively, that is
- *But:*
  - after one change of stewardship?
  - after *two* changes of stewardship??

The New York Times – Breaking News, World News & Multimedia

http://www.nytimes.com/

Work ▾ Reference ▾ News ▾

HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | U.S. Edition ▾

Subscribe: Digital / Home Delivery | Log In | Register Now

The New York Times

Monday, August 6, 2012 Last Update: 8:18 PM ET

Search | ING DIRECT | Follow Us | Subscribe to Home Delivery | Personalize Your Weather

WORLD
U.S.
POLITICS
NEW YORK
BUSINESS
DEALBOOK
TECHNOLOGY
SPORTS
OLYMPICS
SCIENCE
HEALTH
ARTS
STYLE
OPINION

Autos
Blogs
Books
Cartoons
Classifieds
Crosswords
Dining & Wine
Education
Event Guide
Fashion & Style
Home & Garden
Jobs
Magazine
Movies

Hospital Chain Investigation Found Dubious Cardiac Work

By REED ABELSON and JULIE CRESWELL 6:33 PM ET

HCA, the largest for-profit hospital chain in the country, uncovered evidence of unnecessary — even dangerous — cardiac treatments at some of its medical centers in Florida.

Post a Comment | Read (90)

Suspect Is U.S. Army Veteran; Ties Seen to Racist Groups

By STEVEN YACCINO, JENNIFER PRESTON and SERGE F. KOVALESKI 6:14 PM ET

Officials said the gunman, shot and killed by the police, was Wade M. Page, who a rights watchdog said had white supremacist

LONDON 2012

NEWS | SCHEDULE | HIGHLIGHTS | PHOTOS | RESULTS

Doug Mills/The New York Times

1 2 3 4 5 6 7 8 9 10

Alex Morgan (United States) scored the winning goal.

In Final Seconds, U.S. Rallies Past Canada

By SAM BORDEN 24 minutes ago

The United States women's soccer team won, 4-3, in extra time and will face Japan in the gold-medal match.

- For Japan, Medal and Rematch 8:02 PM ET
- Post a Comment

OPINION »

Anxiety: Lost and Found

It was time for my son to go it alone on the streets of New York. What could go wrong?

- Keller: The Leak Police
- Edsall: Sheldon Adelson
- Editorial: Carrots and Sticks for School Systems
- Campaign Stops: Election Rules and Manipulation
- Taking Note: Early Voting in Ohio


The New York Times – Breaking News, World News & Multimedia
http://10.15.205.37/10.5079.6205.2398/x3a7q.html
Google
Work
HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | U.S. Edition
Subscribe: Digital / Home Delivery | Log In | Register Now
The power to do more
The New York Times
Monday, August 6, 2012 Last Update: 8:18 PM ET
Search
ING DIRECT
Follow Us | Subscribe to Home Delivery | Personalize Your Weather
WORLD
U.S.
POLITICS
NEW YORK
BUSINESS
DEALBOOK
TECHNOLOGY
SPORTS
OLYMPICS
SCIENCE
HEALTH
ARTS
STYLE
OPINION
Autos
Blogs
Books
Cartoons
Classifieds
Crosswords
Dining & Wine
Education
Event Guide
Fashion & Style
Home & Garden
Jobs
Magazine
Movies
Hospital Chain Investigation Found Dubious Cardiac Work
By REED ABELSON and JULIE CRESWELL 6:33 PM ET
HCA, the largest for-profit hospital chain in the country, uncovered evidence of unnecessary — even dangerous — cardiac treatments at some of its medical centers in Florida.
Post a Comment | Read (90)
Suspect Is U.S. Army Veteran; Ties Seen to Racist Groups
By STEVEN YACCINO, JENNIFER PRESTON and SERGE F. KOVALESKI 6:14 PM ET
Officials said the gunman, shot and killed by the police, was Wade M. Page, who a rights watchdog said had white supremacist
LONDON 2012
NEWS | SCHEDULE | HIGHLIGHTS | PHOTOS | RESULTS
NEXT
Doug Mills/The New York Times
1 2 3 4 5 6 7 8 9 10
Alex Morgan (United States) scored the winning goal.
In Final Seconds, U.S. Rallies Past Canada
By SAM BORDEN 24 minutes ago
The United States women's soccer team won, 4-3, in extra time and will face Japan in the gold-medal match.
For Japan, Medal and Rematch 8:02 PM ET
Post a Comment
OPINION »
Anxiety: Lost and Found
It was time for my son to go it alone on the streets of New York. What could go wrong?
Keller: The Leak Police
Edsall: Sheldon Adelson
Editorial: Carrots and Sticks for School Systems
Campaign Stops: Election Rules and Manipulation
Taking Note: Early Voting in Ohio
The power to do more

Recap

- *New set of expectations*
  - online, available, linked
- *New demands*
  - data as new kind of publication
  - data-driven science
- $\Rightarrow$ *Driving need for curation*
- *But data curation is hard:*
  - bit storage not solved at large scales
  - complex data, data lifecycles, workflows
  - determining authenticity, quality when original producer is gone

Unanswered questions

- *Who is responsible for curation? Who does what? Who pays?*
- *Data Curation @ UCSB proposal:*
  - scientists lack resources, expertise
  - "…curation does not necessarily directly or immediately enhance the science being performed now. As a consequence, it can be difficult for working researchers to allocate time and resources to address curation of their data since such allocation may come at the expense of obtaining more immediate results..."

Unanswered questions

- *Tenopir et al (doi:10.1371/journal.pone.0021101)*
  - Survey of 1329 scientists
  - "Scientific research in the 21st century is more data intensive and collaborative than in the past."
  - "Scientists do not make their data electronically available to others for various reasons, including insufficient time and lack of funding."
  - "Most respondents are satisfied with their current processes for the initial and short-term parts of the data or research lifecycle [...] but are not satisfied with long-term data preservation."
  - "Many organizations do not provide support to their researchers for data management both in the short- and long-term. If certain conditions are met (such as formal citation and sharing reprints) respondents agree they are willing to share their data."

Curation players

- *Discipline-specific repositories*
  - Protein Data Bank, Virtual Observatory
- *Government agency repositories*
  - NASA, NOAA
- *Consortia, federations*
  - DataONE, MetaArchive
- *Service providers*
  - CDL/UC3
- *Campus units*
  - ERI, MSI, NCEAS
- *Campus libraries*
  - You!

Who is responsible?
Who does what?
Who pays?

Data Curation @ UCSB project

- *Select case studies*
  - primarily in ERI, primarily in sciences
  - scientists, datasets, paradigms/workflows
- *Work through curation issues*
  - identification/citation
  - storage
  - access
  - funding
- *Examine*
  - help, services, expertise required
  - roles played by different organizations
  - Interactions between organizations, handoffs
- $\Rightarrow$ *Produce campus, library recommendations*

For more information

- *Greg Janée* *gjanee@eri.ucsb.edu*
- *James Frew* *frew@bren.ucsb.edu*