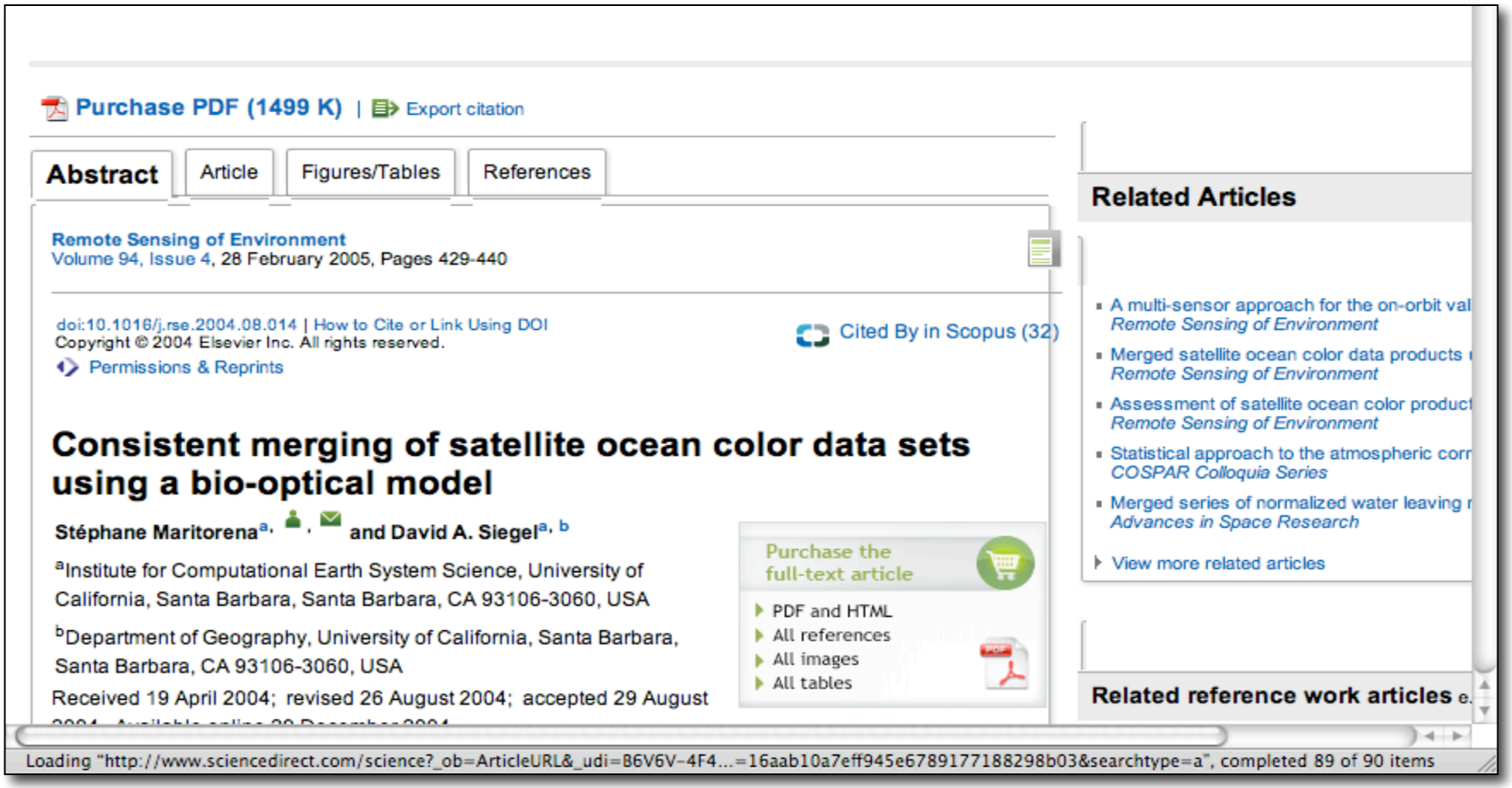
Earth science data: second class citizen in the scholarly record

GREG JANÉE
University of California, Santa Barbara; and
California Digital Library

The problem


Purchase PDF (1499 K) | Export citation
Abstract
Article
Figures/Tables
References
Remote Sensing of Environment
Volume 94, Issue 4, 28 February 2005, Pages 429-440
doi:10.1016/j.rse.2004.08.014 | How to Cite or Link Using DOI
Copyright © 2004 Elsevier Inc. All rights reserved.
Permissions & Reprints
Cited By in Scopus (32)
Consistent merging of satellite ocean color data sets using a bio-optical model
Stéphane Maritorena[a], and David A. Siegel[a, b]
[a]Institute for Computational Earth System Science, University of California, Santa Barbara, Santa Barbara, CA 93106-3060, USA
[b]Department of Geography, University of California, Santa Barbara, Santa Barbara, CA 93106-3060, USA
Received 19 April 2004; revised 26 August 2004; accepted 29 August
Purchase the full-text article
PDF and HTML
All references
All images
All tables
Related Articles
A multi-sensor approach for the on-orbit val
Remote Sensing of Environment
Merged satellite ocean color data products
Remote Sensing of Environment
Assessment of satellite ocean color product
Remote Sensing of Environment
Statistical approach to the atmospheric corr
COSPAR Colloquia Series
Merged series of normalized water leaving
Advances in Space Research
View more related articles
Related reference work articles
Loading "http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6V6V-4F4...=16aab10a7eff945e6789177188298b03&searchtype=a", completed 89 of 90 items

UCSB Ocean Color Group

http://www.icess.ucsb.edu/OCisD/

Work Reference News

ICESS Ocean Color Research Group

ICESS Institute for Computational Earth System Science

PEOPLE	PROJECTS	PRODUCTS/SERVICES
• Dave Siegel	• Bermuda Bio-Optics Project (BBOP)	• Ocean Color in-situ Database
• Stéphane Maritorena	• Plumes & Blooms	• Garver, Siegel, Maritorena Model (GSM-01) IDL code files
• Norm Nelson	• MEASURES Ocean Color Project	• MEASURES: Ocean Color Merged Data Sets
• David Court		
• Nathalie Guillocheau		
• Dave Menzies		
• Eric Fields		

FTP site

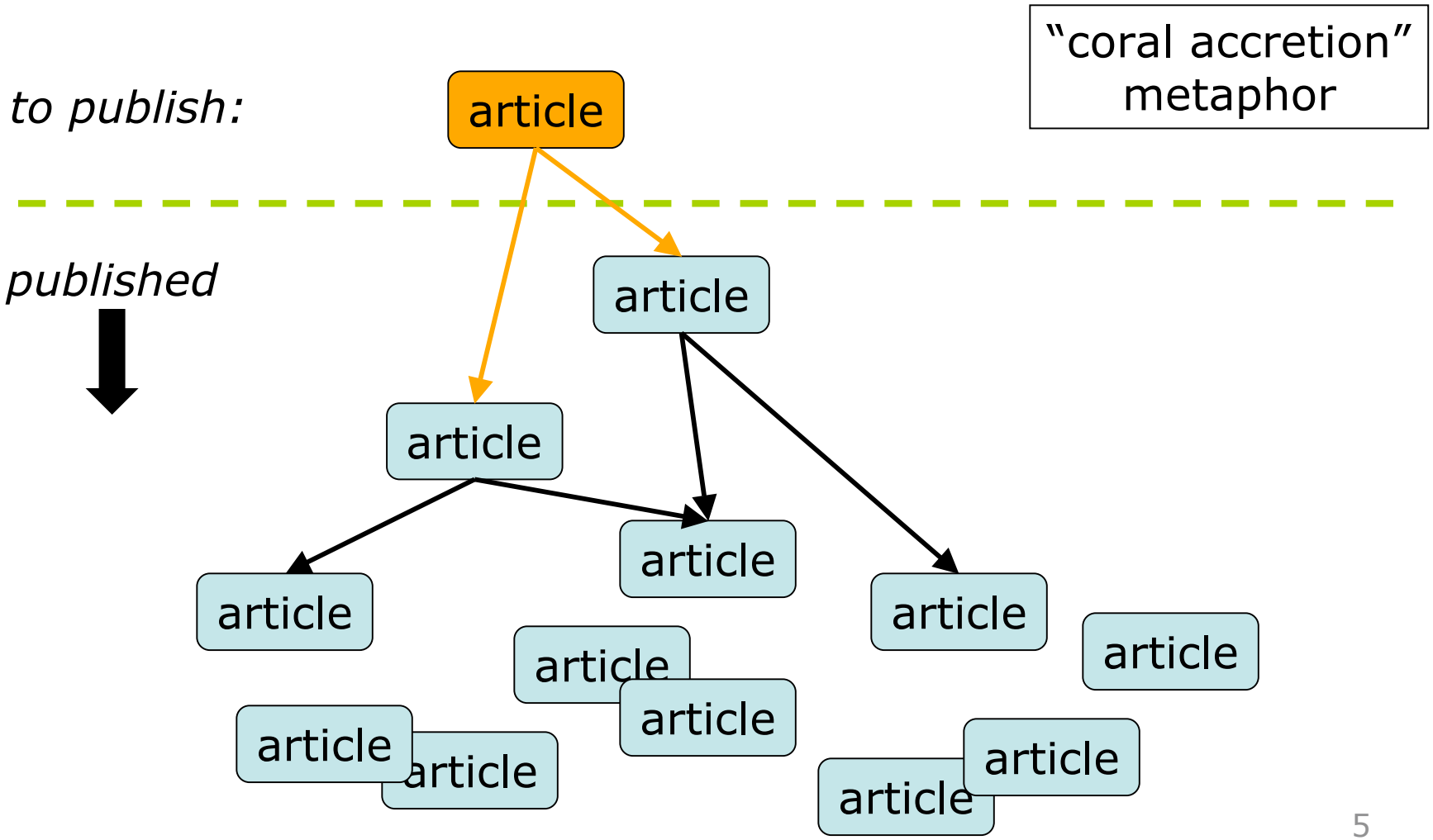
Outline

- What makes citing Earth science data difficult?
- Two efforts addressing Earth science data specifically:
  - ESIP data citation guidelines
  - versioned granule identifiers

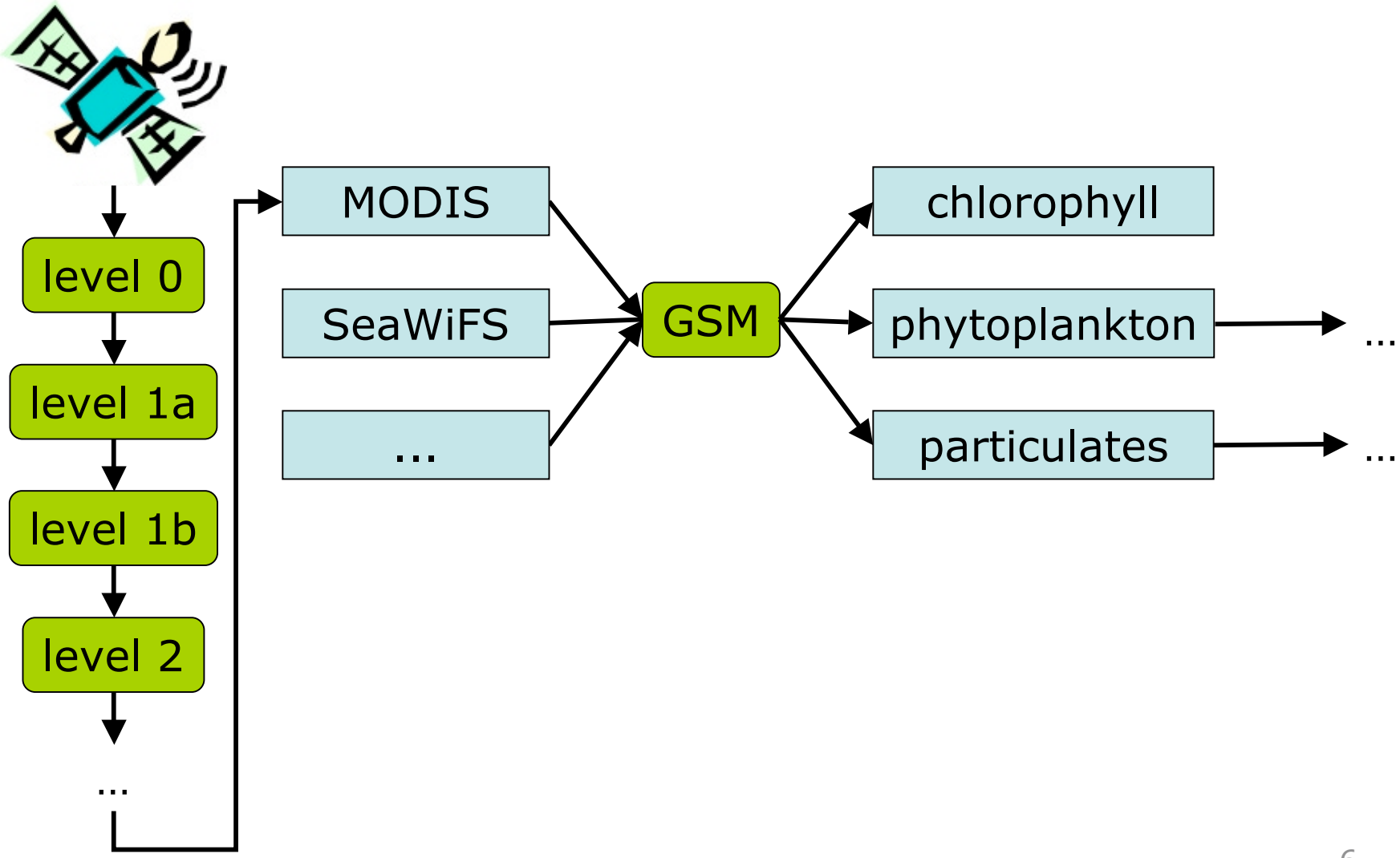
Scholarly literature characteristics

- One, well-defined "publish" point
- Author plays no role after publishing
- Article is single unit of reuse
- Article is unchanging… forever
- No concepts of versions or revision
  - instead: new articles respond to old
- Coarse-grained provenance
  - references to other scholarly articles (only)
- Multiple copies
  - equivalent; tightly controlled due to copyright

Scholarly literature publishing



Earth science data workflows



Earth science data — analysis

- Characterized by workflows
- Dynamic
  - time series may extend for decades
  - ergo, granules and granule aggregations

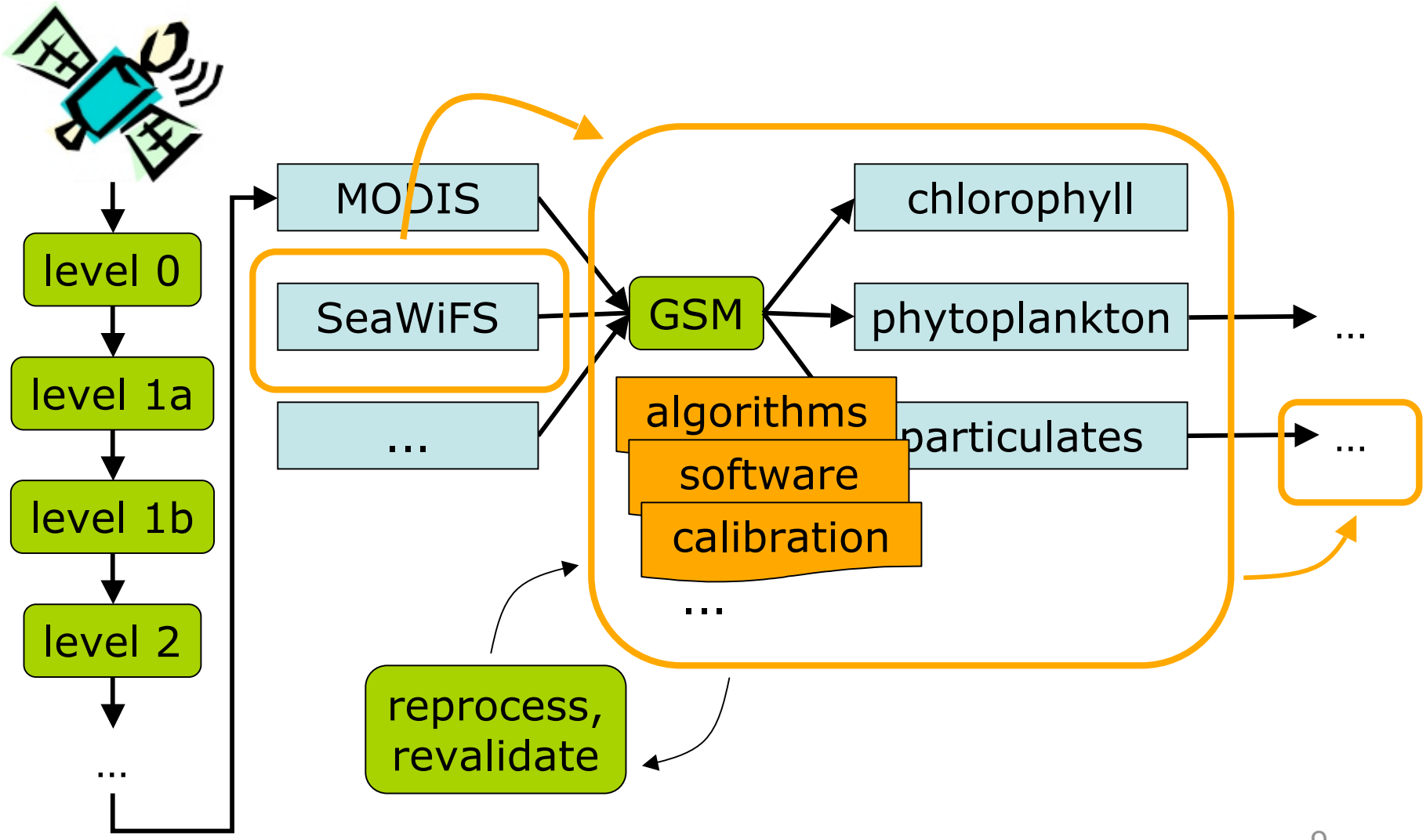
The curse of reprocessing

- SeaWiFS
  - Reprocessing 5.2 - Completed July 12, 2007
  - Reprocessing 5.1 - Completed July 5, 2005
  - Reprocessing 5 - Completed March 18, 2005
  - Reprocessing [illegible]1 - Completed May 24, 2004
  - Reprocessing [illegible] 2002
  - Reprocess[illegible]000
    - Calibra[illegible]
    - Calibration Update - April 10, 2001
  - Reprocessing 2 - August, 1998
  - Reprocessing 1 - January, 1998

new atmospheric, solar irradiance models

http://oceancolor.gsfc.nasa.gov/REPROCESSING/

Earth science data workflows



Earth science data — analysis

- Characterized by workflows
- Dynamic
  - time series extend for decades
  - ergo, granules and other aggregations
  - reprocessing
- Continued interaction between authors, archives

- Versioning important
  - strong incentive to move to new versions
- Fine-grained provenance
  - which versions of which inputs?
  - processed by which software?
- Multiple copies
  - less tightly controlled
  - hard to define and determine "scientific equivalence"

Implications for citation

- Different actors, different relationships
- Different concept of “publish”
- Provenance, versioning important
- Greater burdens on identifiers

ESIP data citation guidelines

- "The ESIP Preservation and Stewardship cluster has examined [DataCite] and other current approaches and has found that they are generally compatible and useful, but they do not entirely meet all the purposes of Earth science data citation."

http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines

ESIP data citation guidelines

- Differences in terminology reflect different worldviews

DataCite	ESIP
Publisher	Archive or Distributor
Publication year	Release date
Version optional	Required; major and minor versions

ESIP data citation guidelines

- Subset used ("micro-citation")
  - Doe, J. and R. Roe. 2001, updated daily. The FOO Gridded Time Series Data Set. Version 3.2. Oct. 2007- Sep. 2008, 84°N, 75°W; 44°N, 10°W. The FOO Data Center. doi:10.xxxx/notfoo.547983. Accessed 1 May 2011.

Versioned granule identifiers

- Need to identify:
  - datasets
  - granules
  - versions thereof
  - version-less versions
  - relationships between
- Would be nice:
  - scalable solution
  - easily managed
    - (requirement!)
  - fits with existing practices

- What we have:
  - persistent, locally-unique granule identifiers
  - simple, hierarchical organizations

http://www.alexandria.ucsb.edu/~gjanee/archive/2009/versioning.html

Example granule identifiers

- SPOT/Image
  - 553828104122518575 21J
- MODIS
  - MOD43A2.A1998365.h5.v8.001.1999001090020
- GSM
  - GSMchl.2003121.L3b_DAY.01.6

Example granule identifiers

product code — *tile identifier* — *production date/time*

- MOD43A2.A1998365.h5.v8.001.1999001090020

Julian acquisition date — *version*

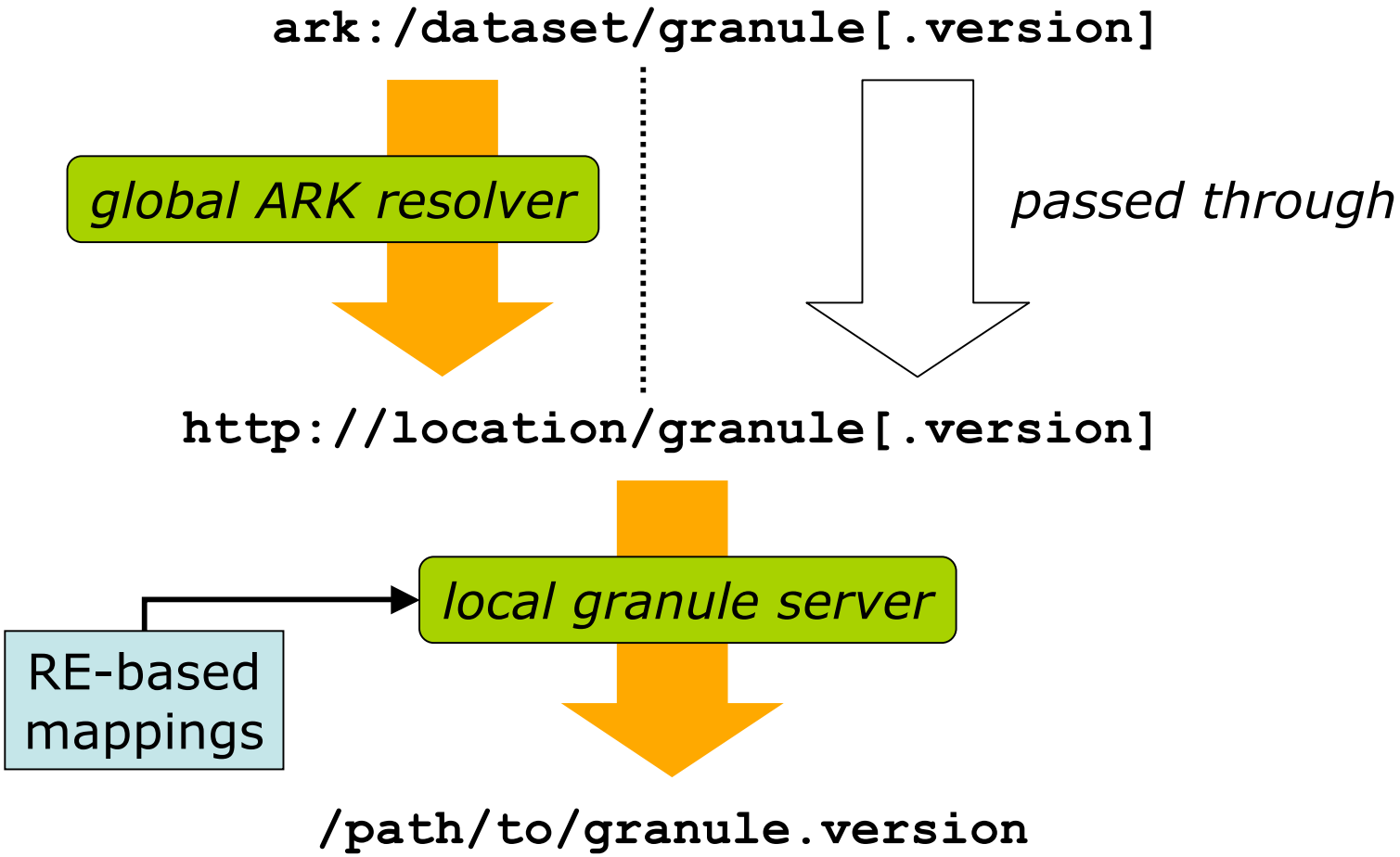
Granule identifiers — approach

Identifier	Semantics
`doi:dataset`	Entire dataset
`doi:dataset/version`	Specific version of dataset
`ark:/dataset`	Base identifier for granule identification
`ark:/dataset/granule.version`	Specific version of granule
`ark:/dataset/granule`	Latest version, or granule abstraction

Granule identifiers — approach

Identifier	
`doi:dataset`	Manually registered
`doi:dataset/version`	
`ark:/dataset`	Automatically supplied by EZID
`ark:/dataset/granule.version`	
`ark:/dataset/granule`	

Granule identifiers — architecture



Granule identifiers — results

Activity or risk	Action required
New granule	Nothing
New dataset version	Create (one) new DOI
Dataset moved	Update base identifier
Dataset reorganized	Update mappings
Dataset split	If on logical boundary, update mappings
Worst case	Turn off suffix passthrough; create individual granule identifiers

Summary

- Earth science data forms an "ecosystem" of related data products
  - preserving data ≡ preserving ability of data to function in that ecosystem
- New conventions and tools can address scalability and provenance challenges
- Version management, scientific equivalence remain thorny