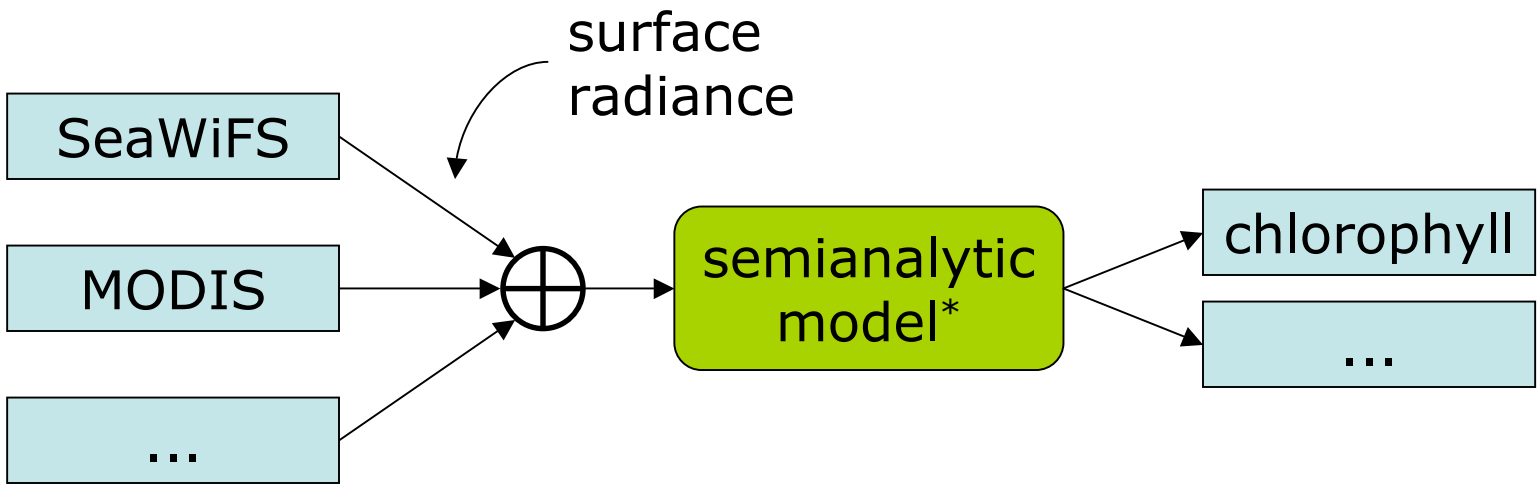
Preserving the Context of Science Data

GREG JANÉE & JAMES FREW
University of California, Santa Barbara

Outline

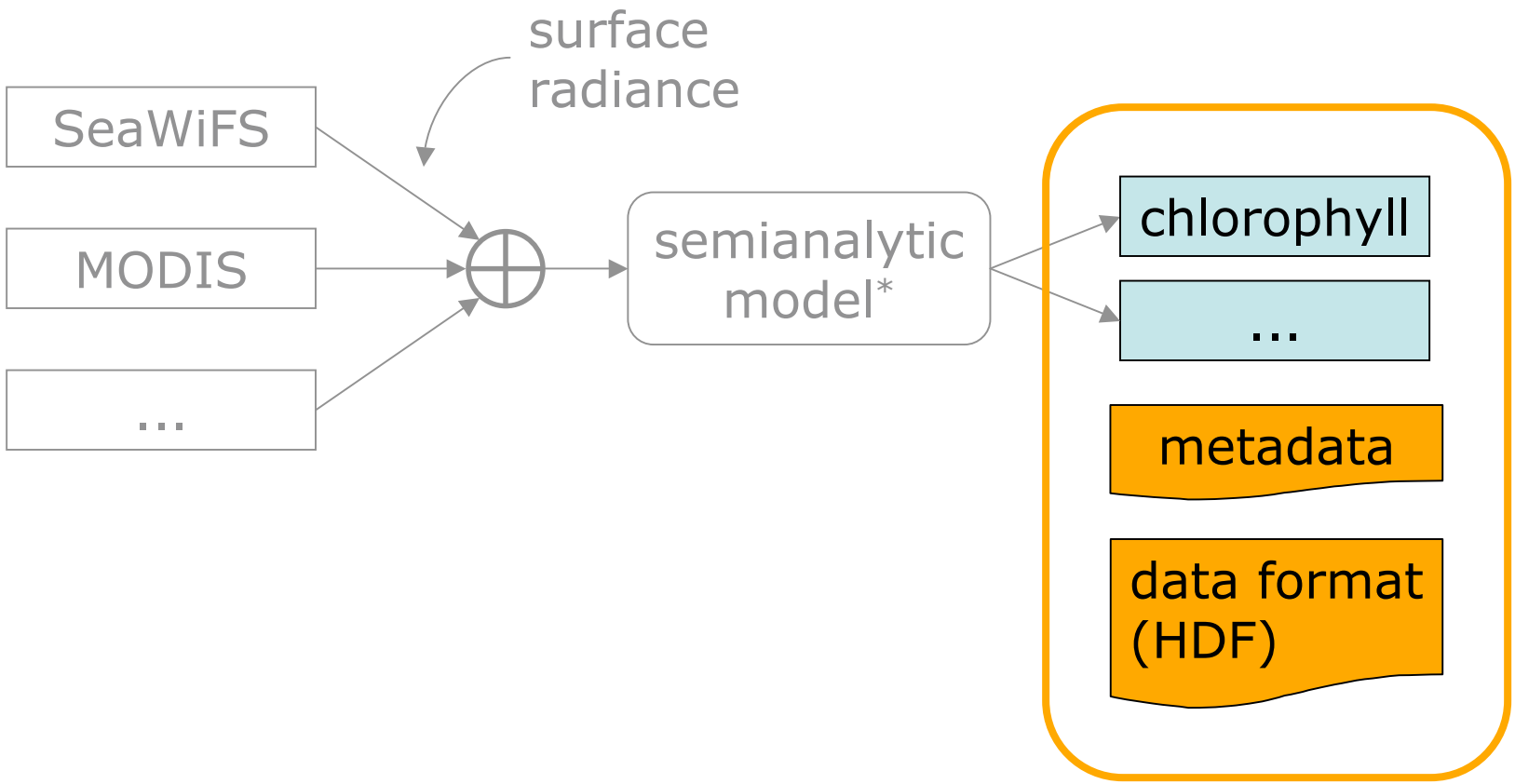
- The problem of context
  - a motivating example
- Preserving it
  - unified data model for data archives & format registries
- Capturing it in the first place
  - integrating wikis and repositories

Ocean color example

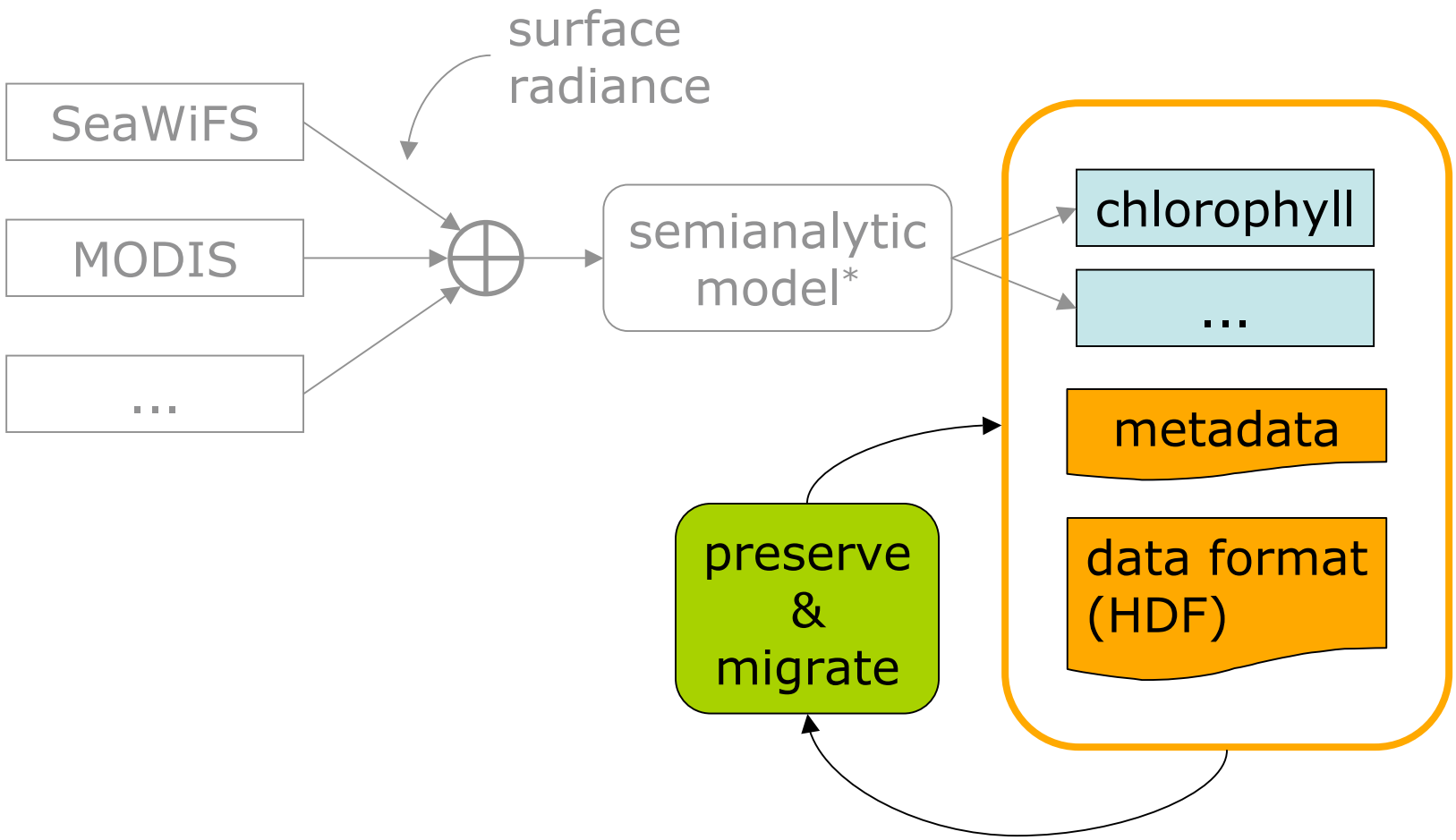


*S. Maritorena, D. Siegel (2005), Consistent merging of satellite ocean color data sets using a bio-optical model, *Remote Sens. Env.* **94**(4):429–440, doi:10.1016/j.rse.2004.08.014

User's view


surface
radiance
SeaWiFS
MODIS
...
semianalytic
model*
chlorophyll
...
metadata
data format
(HDF)

Preservation of use (only)



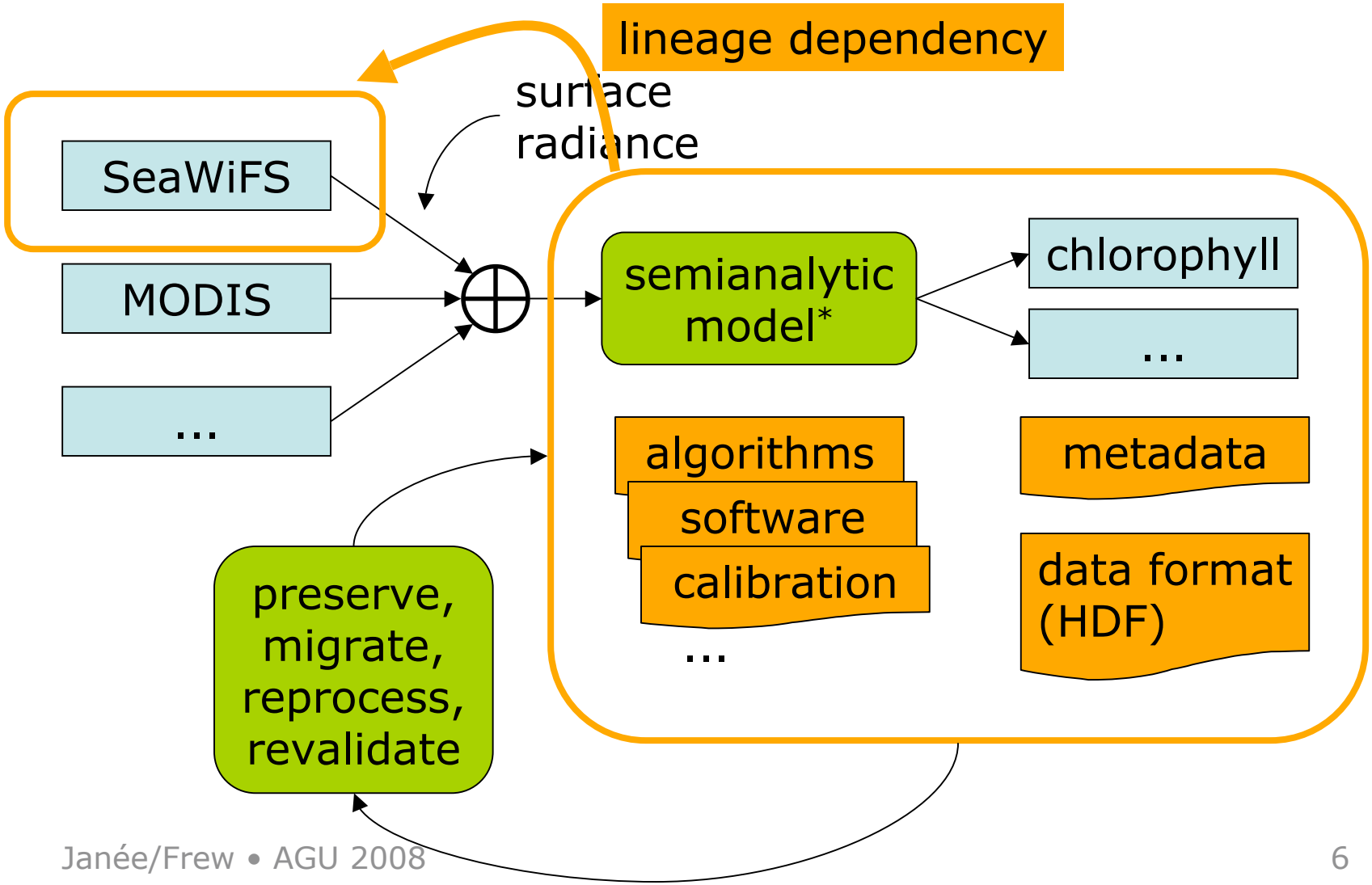
The curse of reprocessing

- SeaWiFS*
  - Reprocessing 5.2 - Completed July 12, 2007
  - Reprocessing 5.1 - Completed July 5, 2005
  - Reprocessing 5 - Completed March 18, 2005
  - Reprocessing [illegible] - Completed May 24, 2004
  - Reprocessing [illegible] 2002
  - Reprocess[illegible]000
    - Calibrat[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/

Preservation of functionality



Ozone reprocessing requirements

- xDRs
- Delivered IPs
- Engineering data (incl. C3S data if not in RDRs)
- Upload files
- Databases
- Software (source code)
- Calibration artifacts
  - data
  - analysis tools
  - tables
  - logs
  - notebooks
  - instrument design
- All project documentation
- All scientific papers
- All reports

Mike Linda, "OMPS Aggregation and Packaging," 2006 CLASS Users' Workshop

Conclusions

- Science data exists in ecosystem of related data products
- Preserving data ≡ preserving ability of data to function in that ecosystem

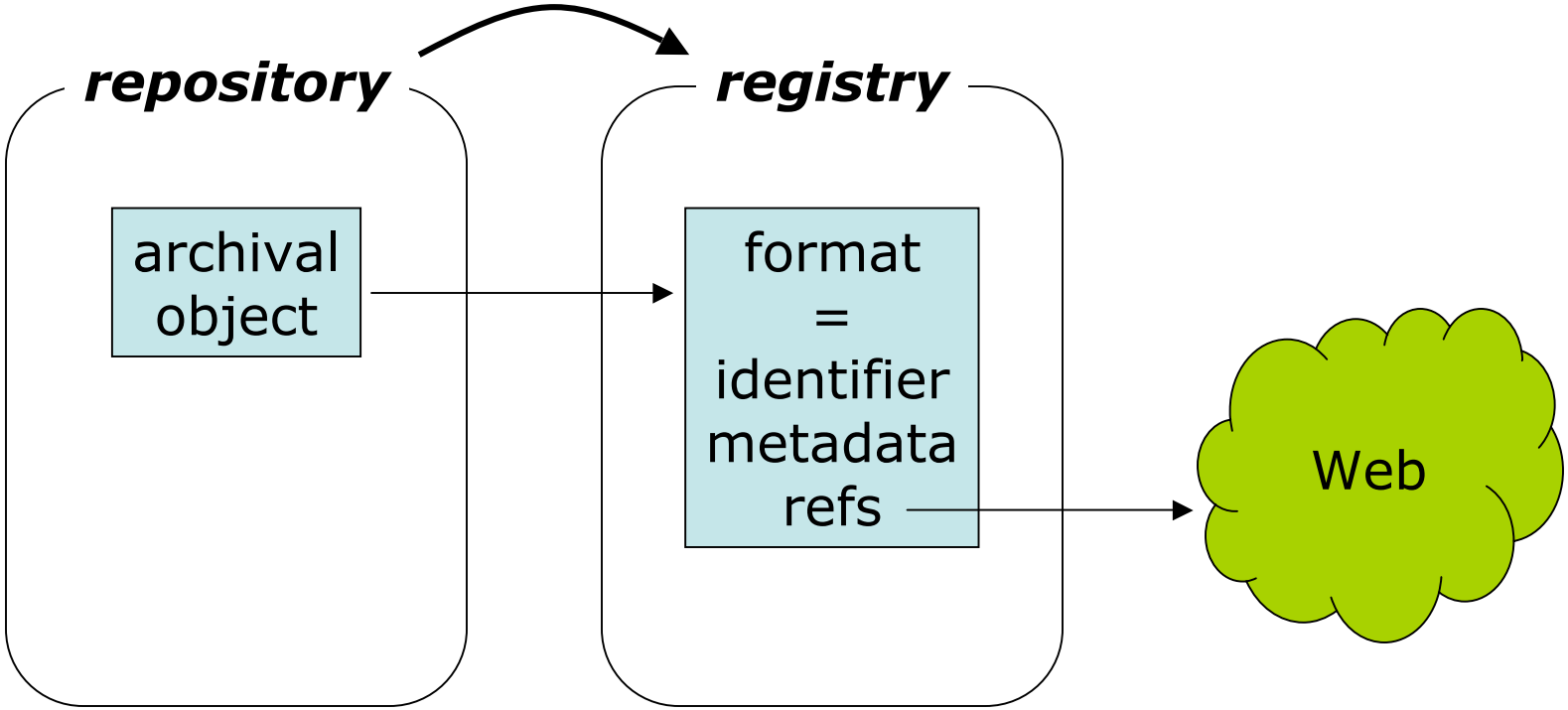
Outline

- The problem of context
  - a motivating example
- Preserving it
  - unified data model for data archives & format registries
- Capturing it in the first place
  - integrating wikis and repositories

NGDA

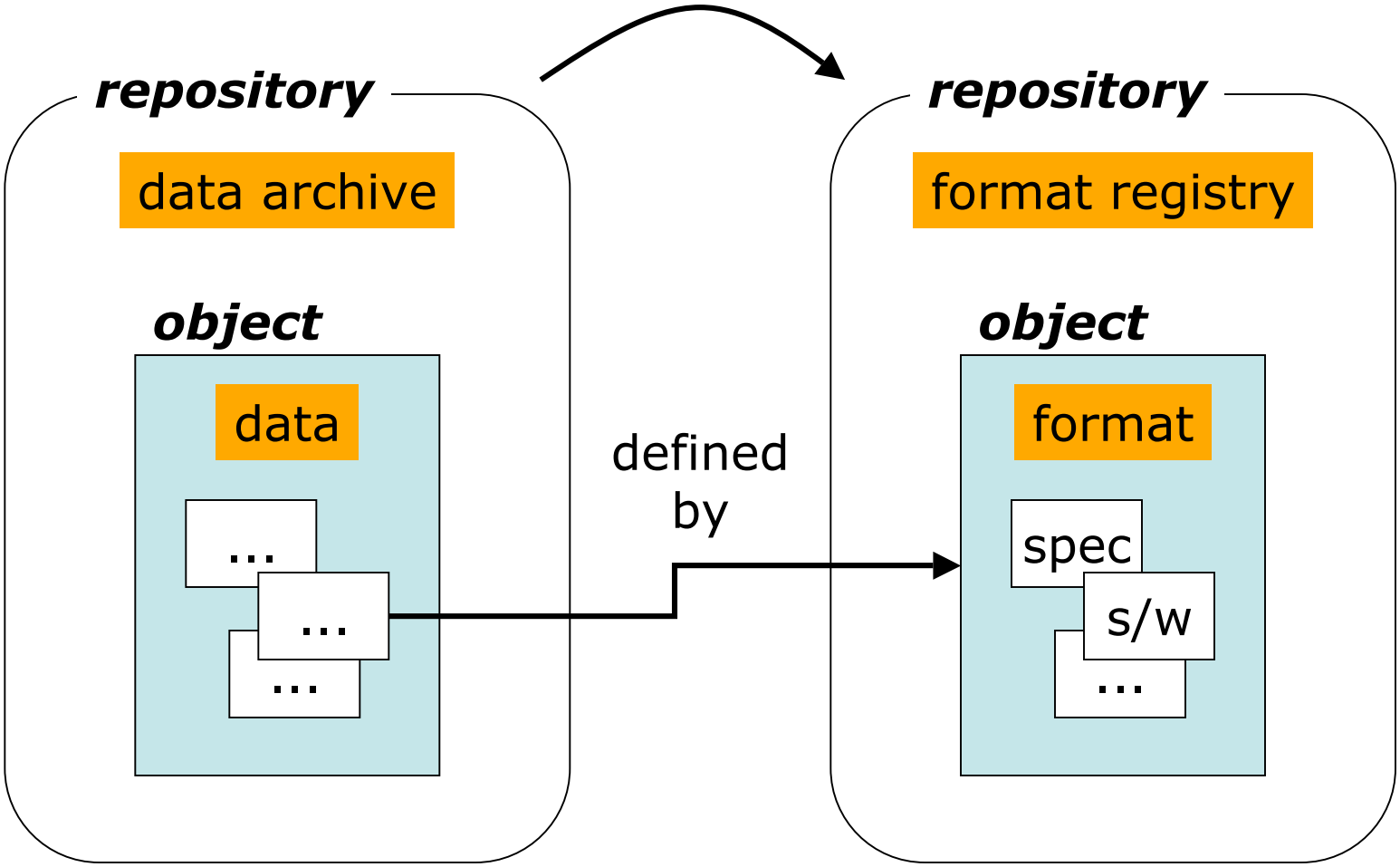
- National Geospatial Digital Archive
  - http://www.ngda.org/
- NDIIPP partner
- Researching long-term preservation
  - of geospatial data
  - on a national scale

Current format registries

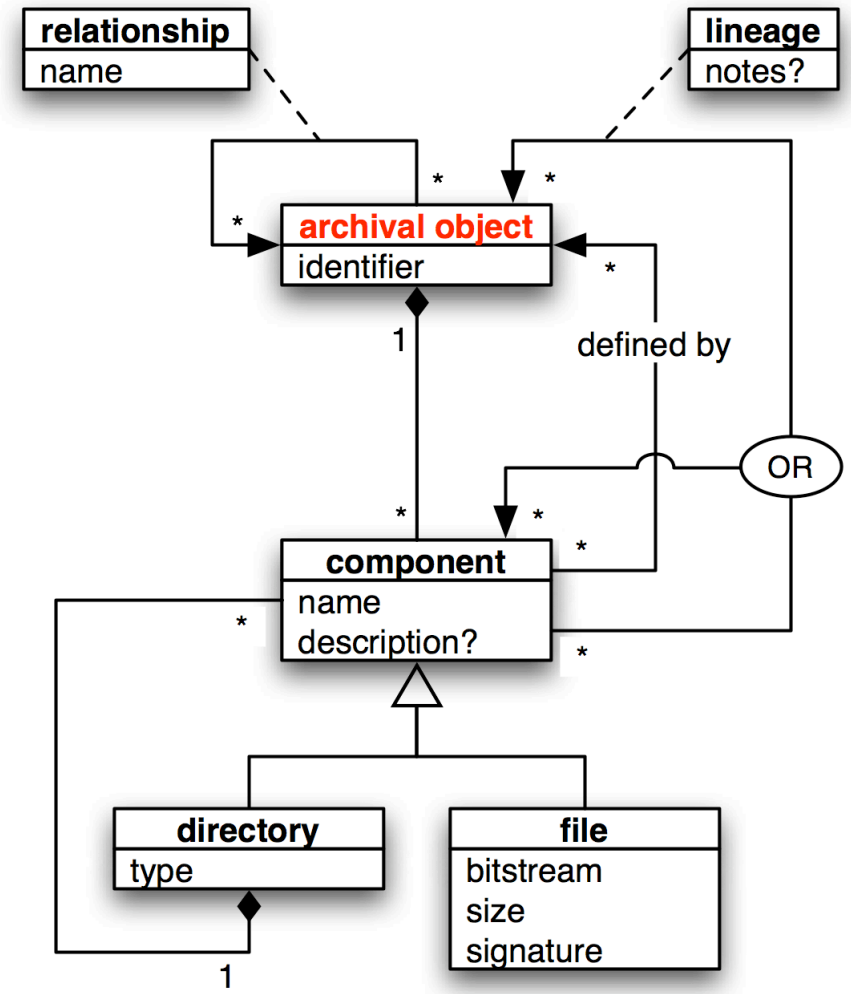


- Preservation of contextual information itself is largely unaddressed

NGDA data model



NGDA data model (UML)



Capturing context

- Community-related problems
  - distributed, implicit, inscrutable to outsiders
  - “known well to those that know it well”
- Semantic problems
  - formal semantics are too hard
  - multiple, conflicting, informal specifications
  - multiple software implementations
- Conclusion
  - context defined by **community of practice**

NGDA format registry