Preserving Geospatial Data: The National Geospatial Digital Archive's Approach

Greg Janée
UC Santa Barbara

NGDA genesis

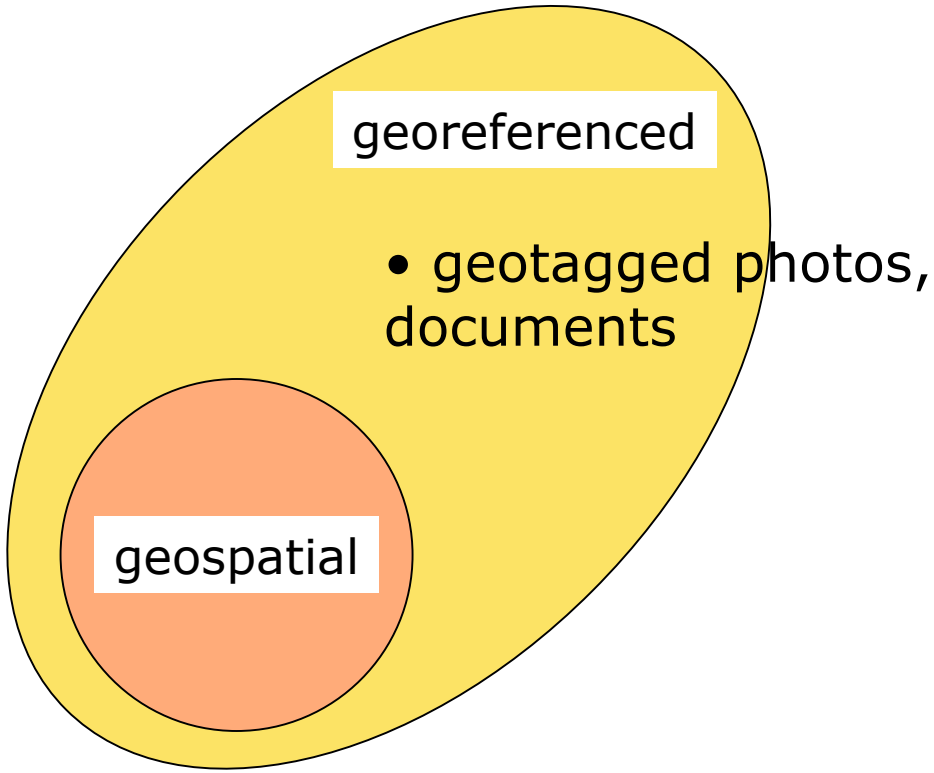
- One of eight initial NDIIPP partners
- Members
  - UCSB, Stanford, UT Knoxville, Vanderbilt
- Goal
  - How to preserve geospatial data, on a national scale, for future generations?

Three questions

- What's special about geospatial?
- Are there *any* design principles that can last a century?
- Can we define a useful, implementable, minimal level of preservation?

Geospatial data

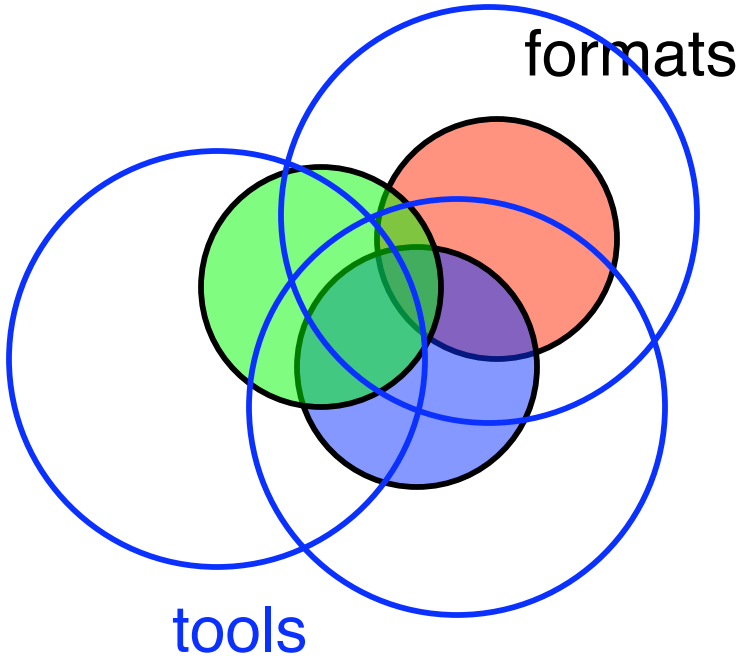
- Representations of Earth's surface
  - remote-sensing imagery
  - aerial photography
  - maps
  - sensor data
  - GIS data



Challenges

- No uniform data model
  - vector, raster, topological, discrete, continuous, …
- Proprietary formats

⇒ Many barriers to data mobility



Challenges (cont.)

- Multiple granule sizes
  - features
  - layers
  - databases
  - projects
  - cartographic end products
- Relational data
  - geodatabases

```
a0000004d.gdbindexes
a0000004d.gdbtable
a0000004d.gdbtablx
a0000004e.blk_key_index.atx
a0000004e.col_index.atx
a0000004e.gdbindexes
a0000004e.gdbtable
a0000004e.gdbtablx
a0000004e.row_index.atx
a0000004f.gdbindexes
a0000004f.gdbtable
a0000004f.gdbtablx
a00000050.gdbtable
a00000050.gdbtable.sdc
a00000050.gdbtable.sdc.prj
a00000050.gdbtable.sdi
...
```

Challenges (cont.)

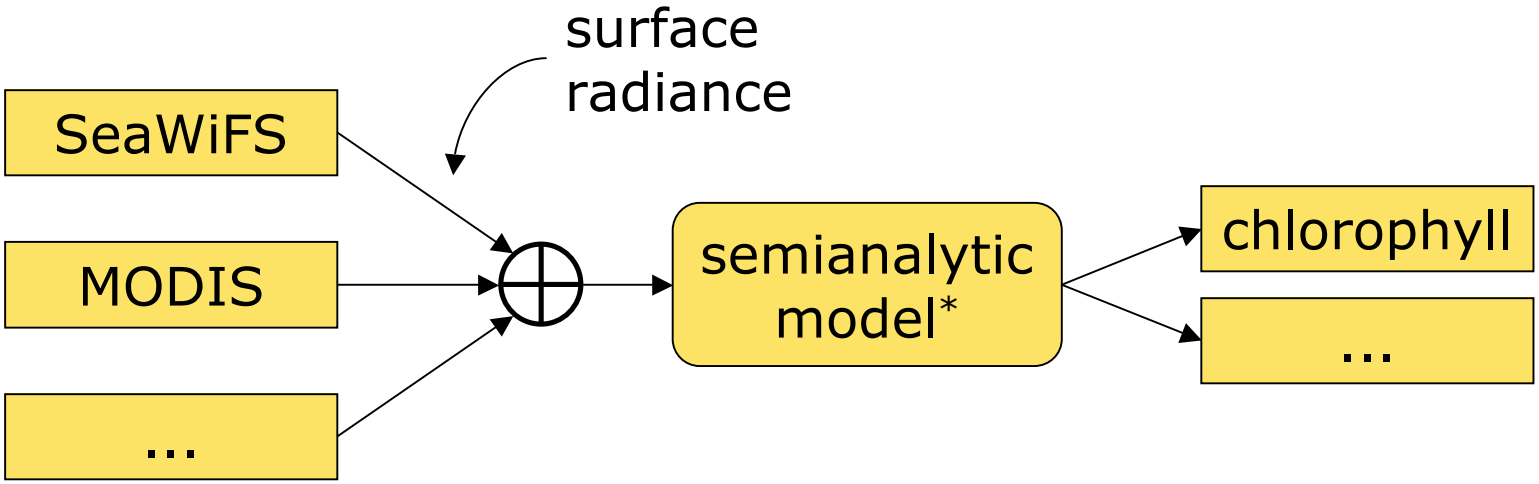
- Large extent
  - storage
  - time
- Extensive context
- Implicit context
- Dynamic data

Visit the USGS Landsat website for important information regarding:

- ground station facts,
- Landsat calibration parameter file details,
- satellite ephemeris information,
- satellite anomaly investigations,
- data acquisition information,
- image processing particulars,
- data product guidance,
- SLC-off data product details,
- and sample data products.

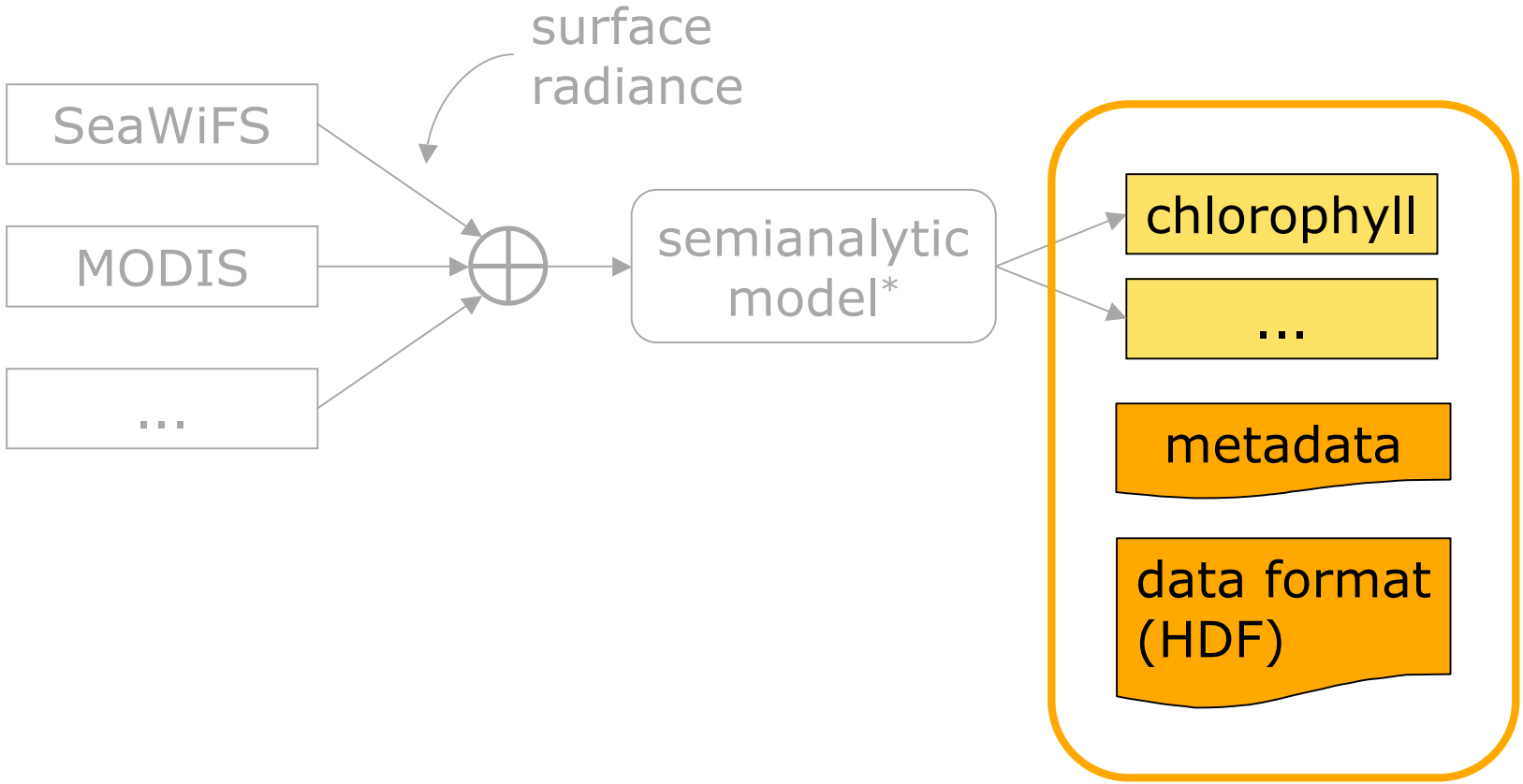
http://landsat.gsfc.nasa.gov/data/tech_details.html

Ocean color example

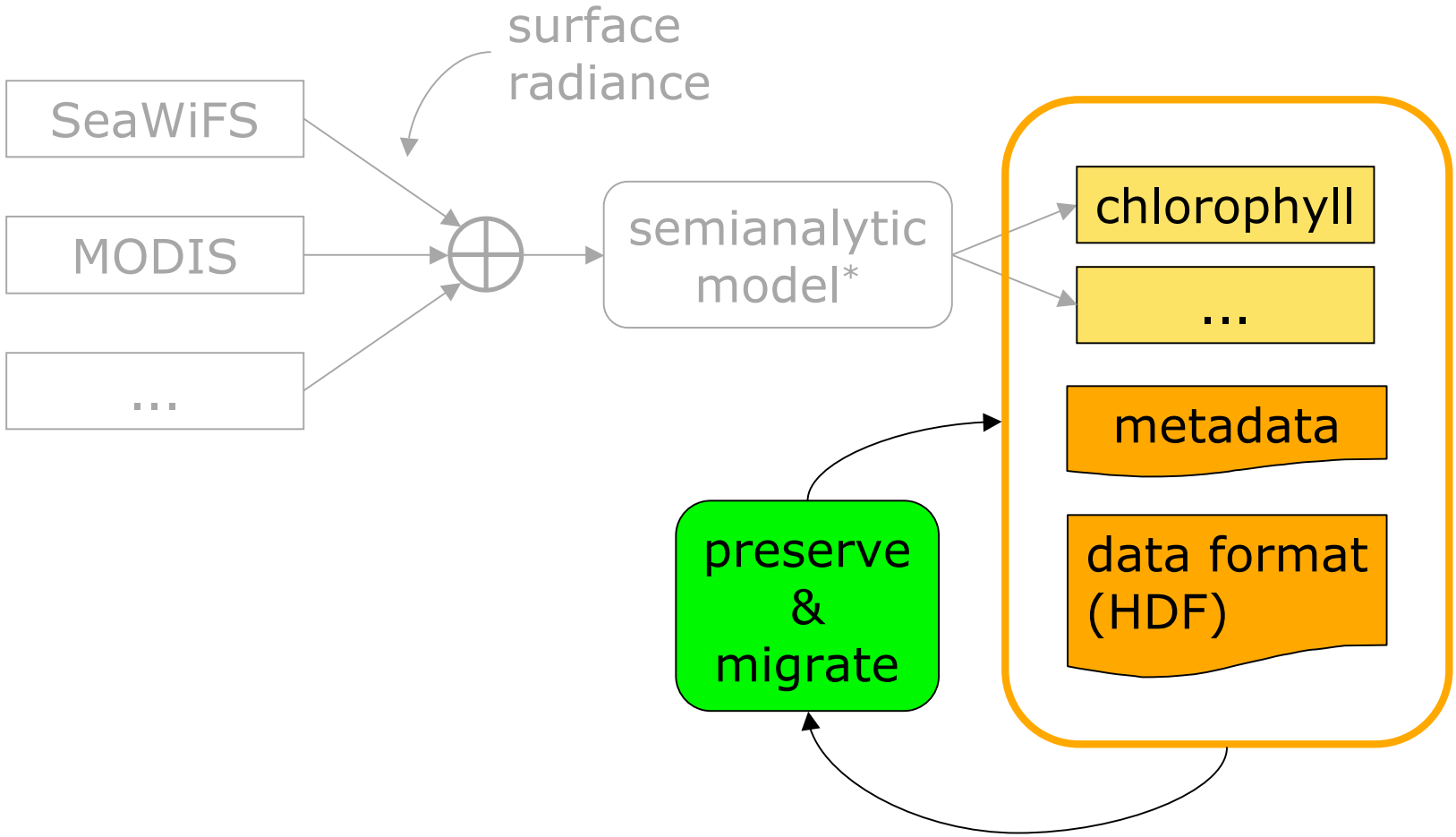


*S. Maritorena, D. Siegel (2005), Consistent merging of satellite ocean color data sets using a bio-optical model, *Remote Sens. Env.* **94**(4):429–440, doi:10.1016/j.rse.2004.08.014

User's view


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

Preservation of use (only)



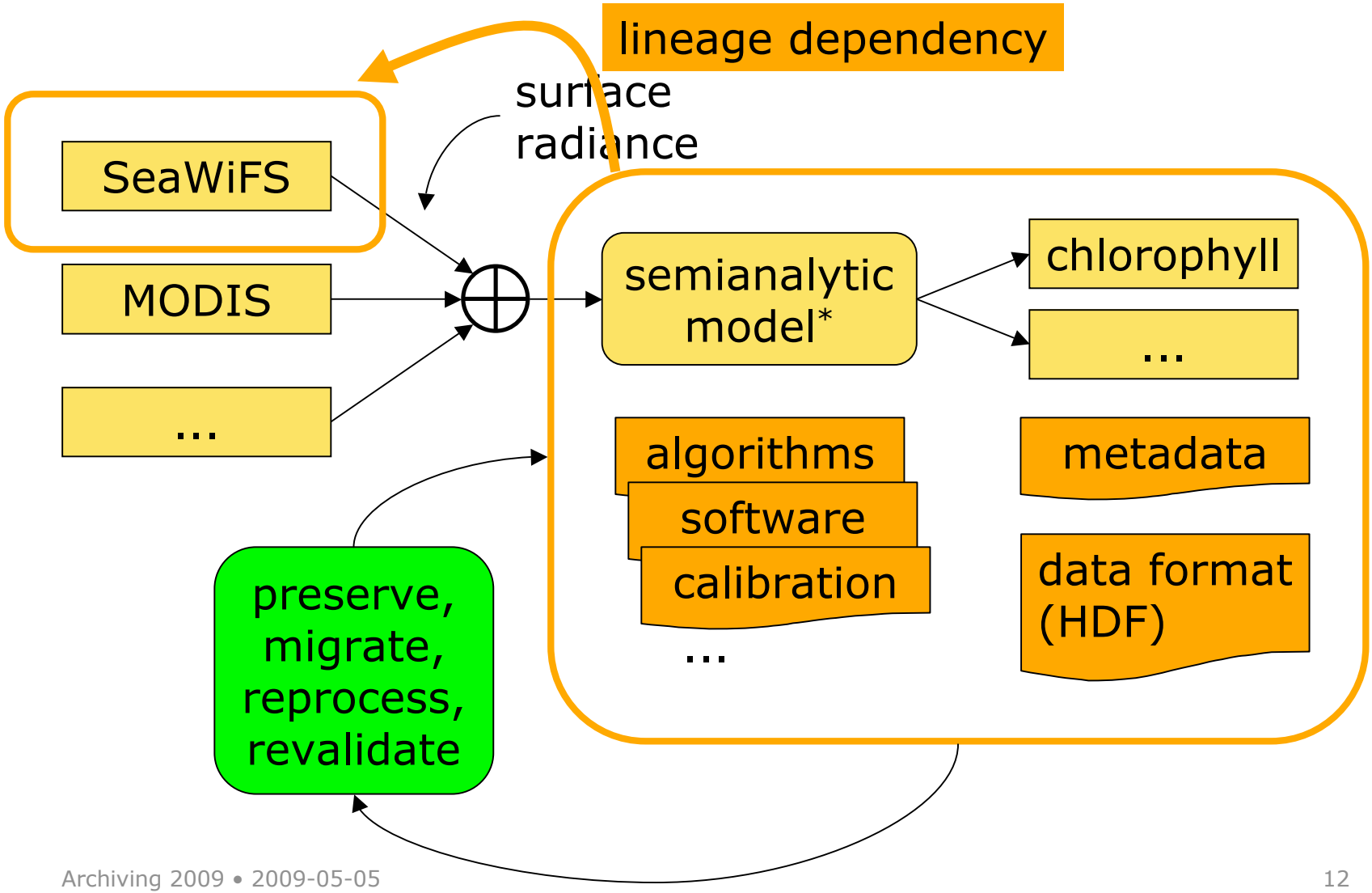
The curse of reprocessing

- SeaWiFS*
  - Reprocessing 5.2 - Completed July 12, 2007
  - Reprocessing 5.1 - Completed July 5, 2005
  - Reprocessing 5 - Completed March 18, 2005
  - Reprocessin 4.1 - Completed May 24, 2004
  - Reproce y 25, 2002
  - Reproce y 24, 2000
    - Calibration Update - December 1, 2000
    - 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

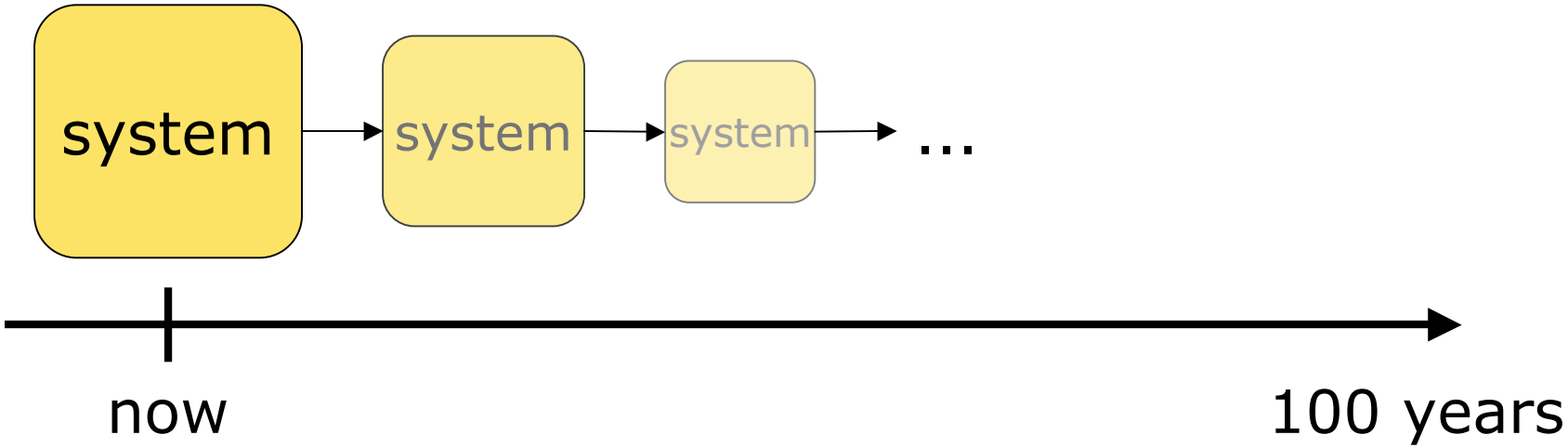
Mike Linda, "OMPS Aggregation and Packaging," 2006 CLASS Users' Workshop

Challenges— conclusion

- NGDA archive design requirements:
  - compound objects
  - aggregations and inter-object relationships
  - extensive context
  - equal treatment of data, context

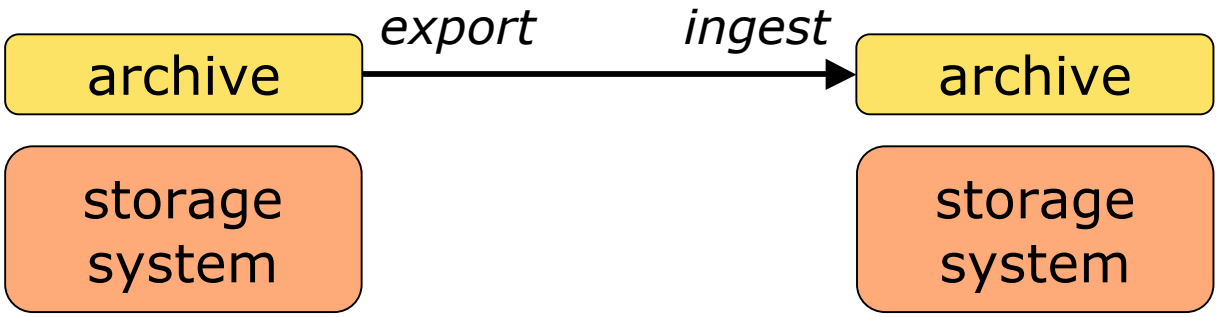
- Unmet challenges:
  - storage size
  - proprietary formats
  - relational data

Relay principle

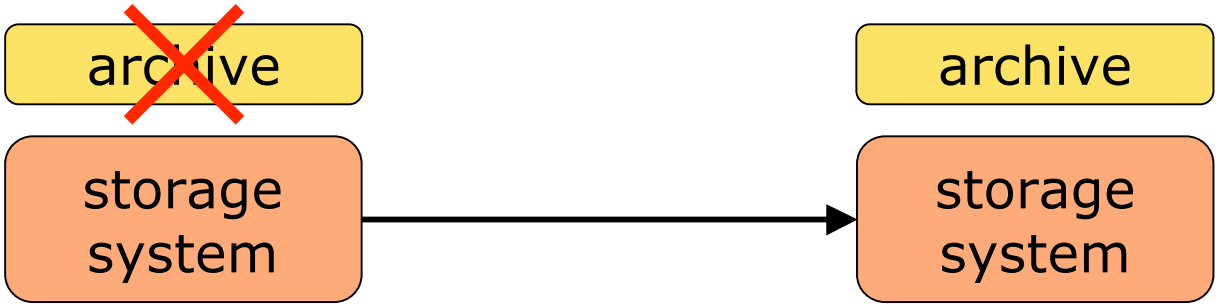


- A preservation system should support its own migration

Fallback principle



Fallback principle



- A preservation system should support some form of handoff of its content even if the system itself is no longer functional.

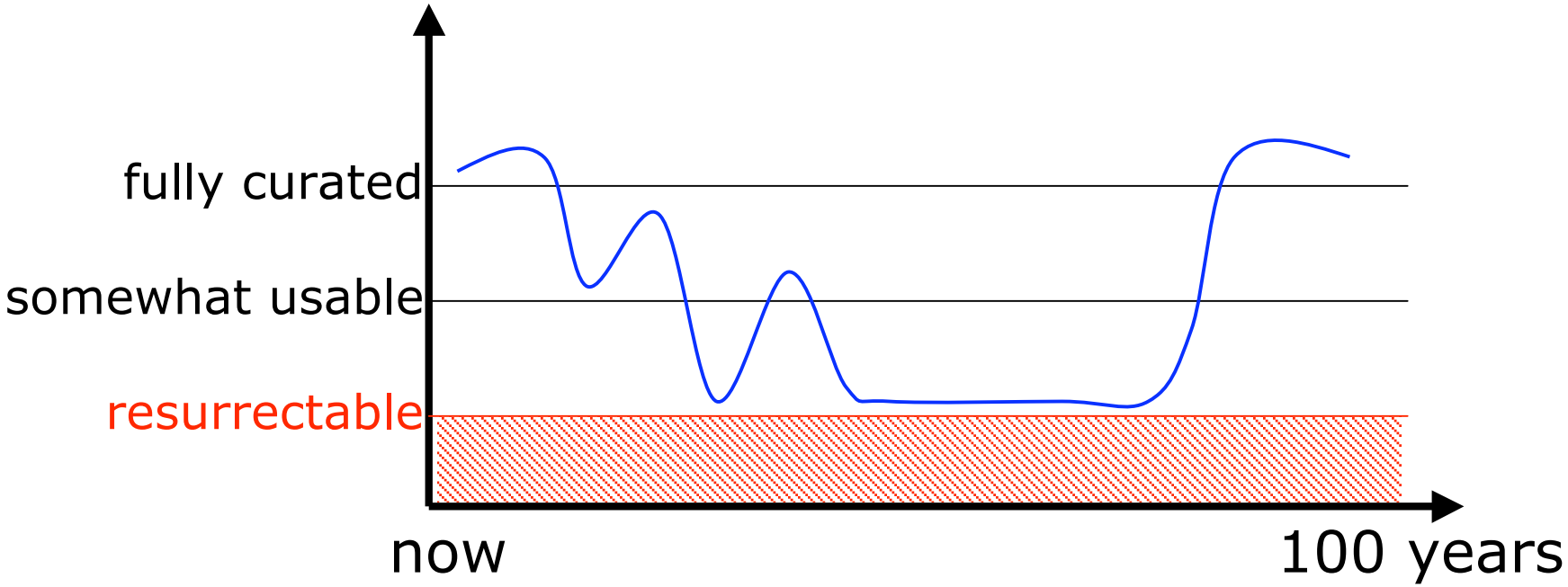
iPhoto example

iPhoto Library/

- 2008/
  - 11/
    - DSC_0035.jpg
    - DSC_0036.jpg
  - 12/
    - DSC_0042.jpg
    - …
- AlbumData.xml 
- Dir.data
- Library.data
- …

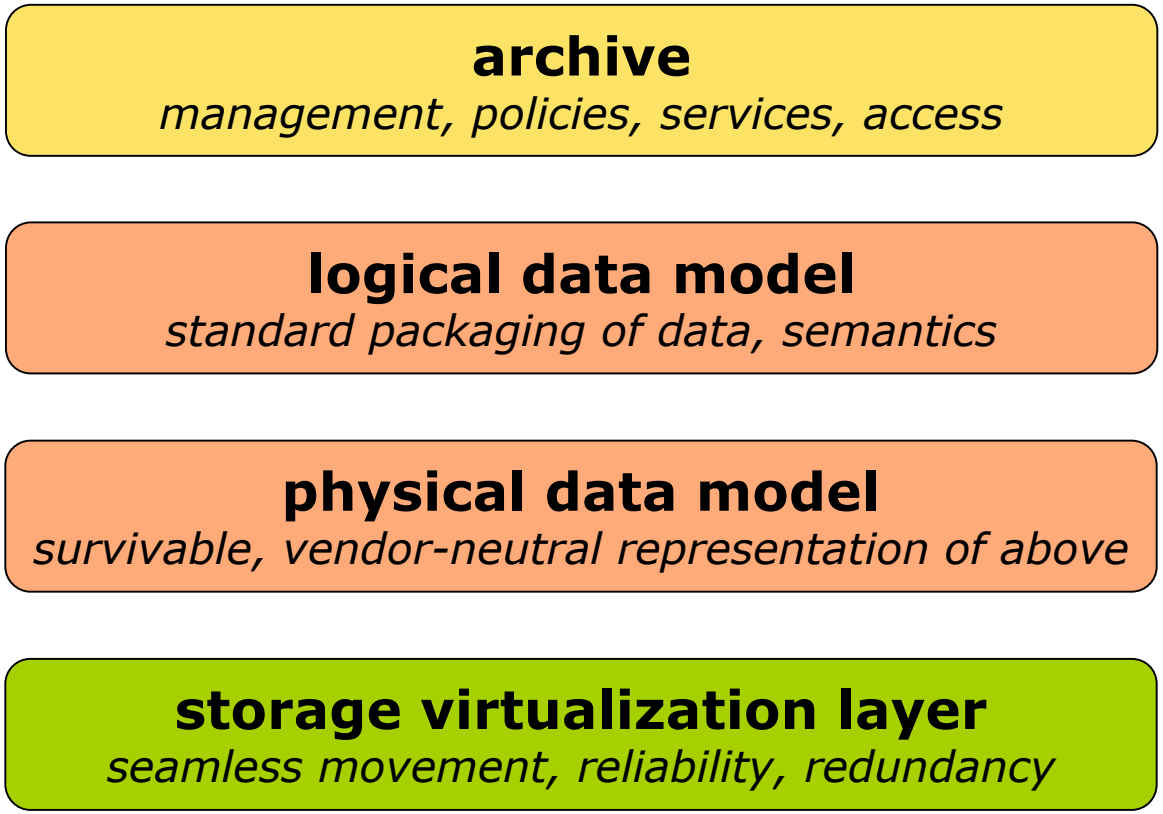
- all metadata
- self-describing schema

Resurrection principle



- A preservation system should allow archived information to lapse out of usability, but at all times should support future resurrection of full use of the information.

NGDA archive system



custom software

instantiation of OAIS

filesystems, files, XML

Logistical Networking

Physical data model

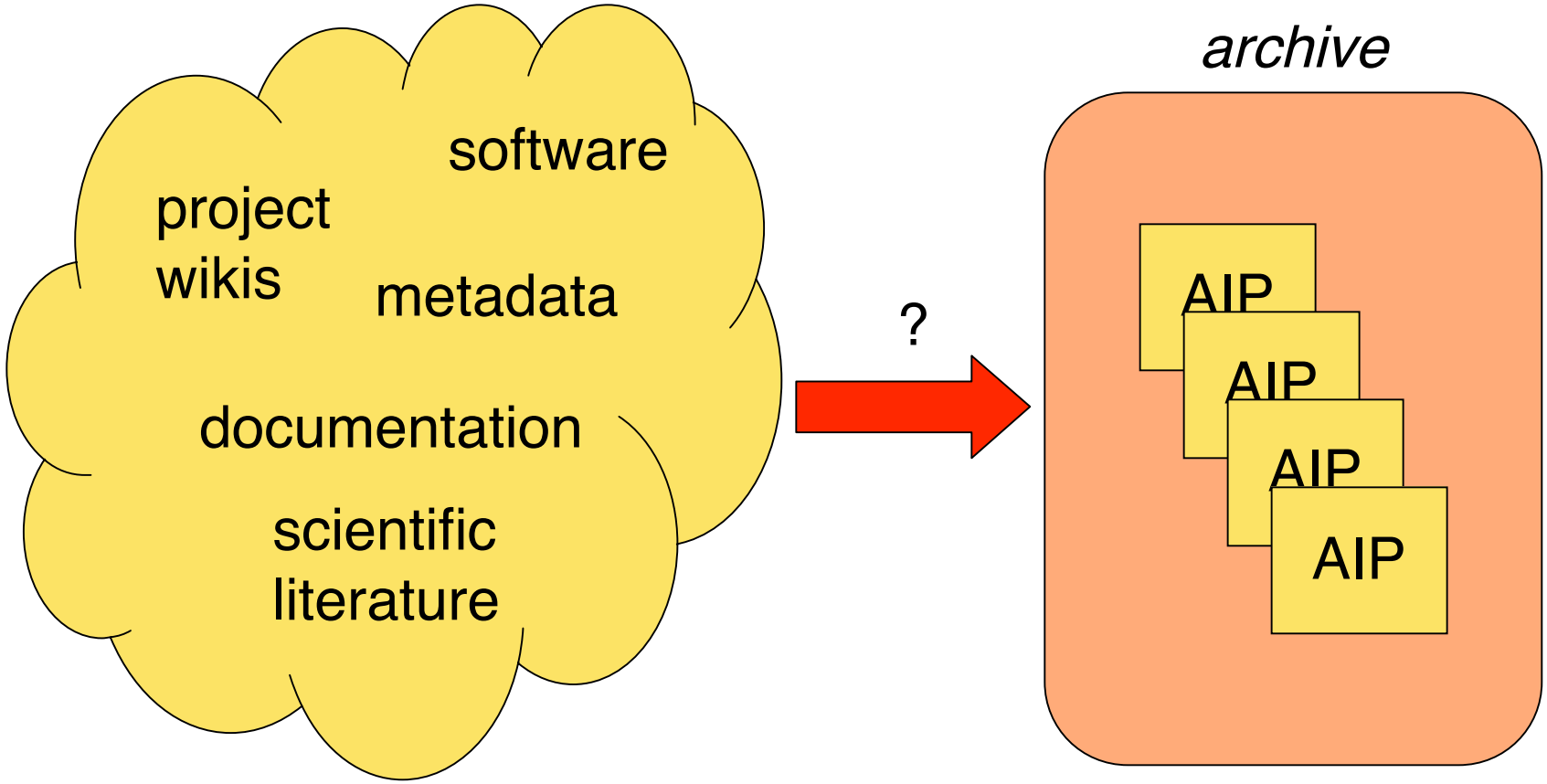
...pathname/
- manifest.xml
- cnty24k97.xml
- data/
  - source/
    - cnty24k97.shp
    - cnty24k97.dbf
    - ...
  - cnty24k97.png

- object structure
- fixity metadata
- inter- and intra-object relationships

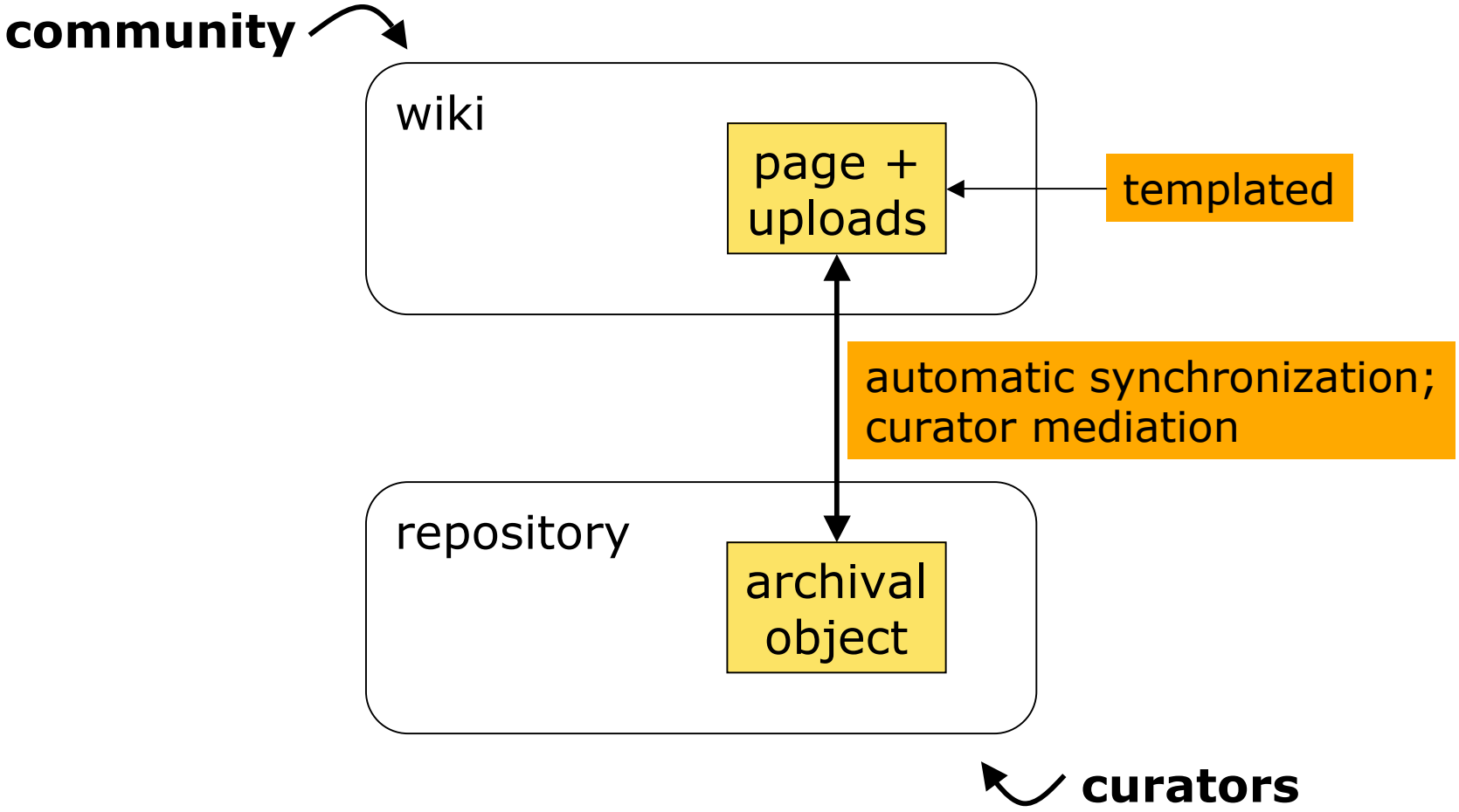
Defining context

- 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**

Capturing context



NGDA format registry



Acknowledgements

- UC Santa Barbara
  - James Frew
  - Catherine Masi
  - Justin Mathena
  - Adam Ross
- Stanford
  - Nancy Hoebelheinrich
  - Keith Johnson
  - Julie Sweetkind-Singer
- UT Knoxville
  - Micah Beck
  - Terry Moore
- NCSU
  - Steve Morris
- EDINA
  - Guy McGarva