A Data Model and Architecture for Long-term Preservation

Greg Janée, Justin Mathena, James Frew

University of California at Santa Barbara

Outline

- Project overview
- Character of geospatial data
- Observations on preservation
  - requirements
- Architecture
- Ongoing work

Project overview

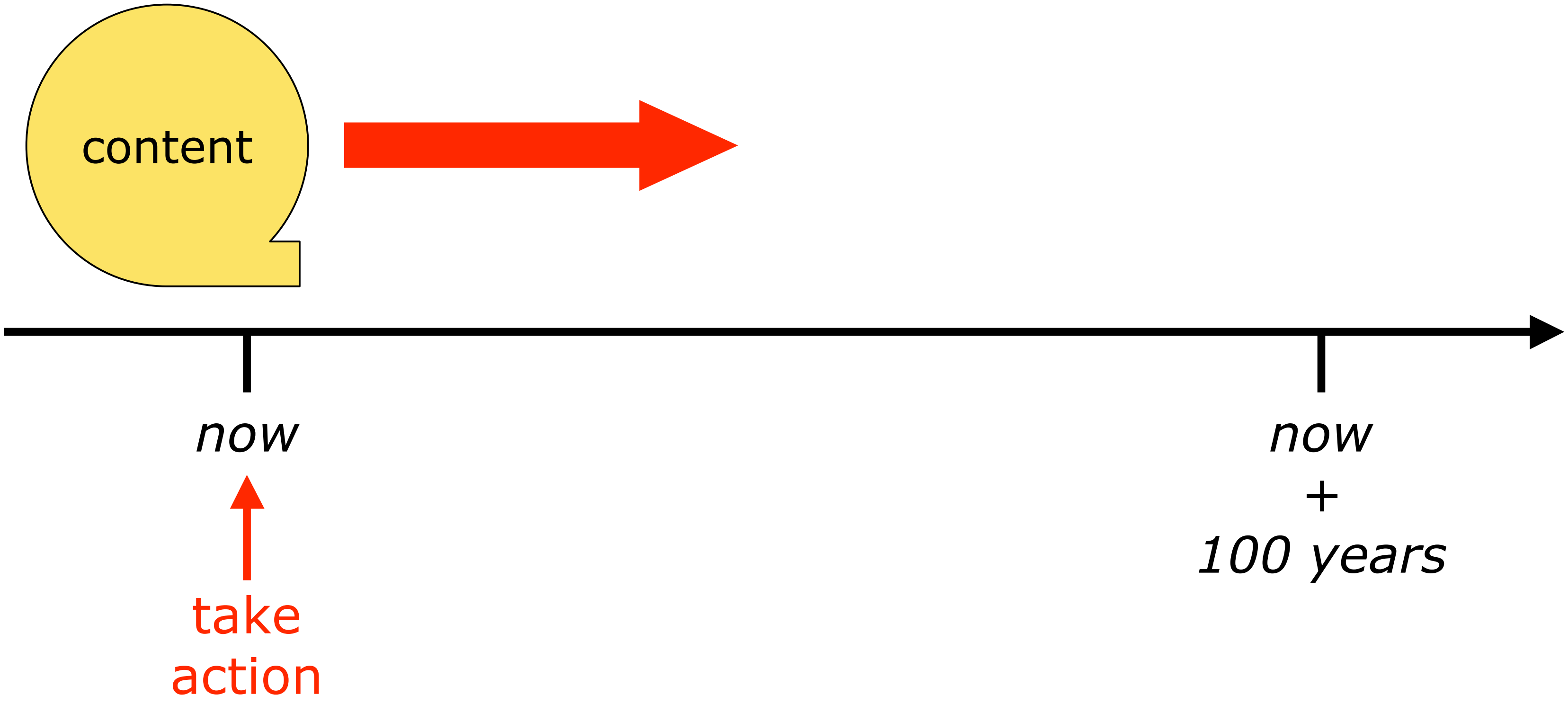
- National Geospatial Digital Archive (NGDA)
  - UCSB (Map & Imagery Laboratory)
  - Stanford (Branner Earth Sciences Library)
- Funded by Library of Congress's NDIIPP program

How to achieve long-term preservation of geospatial data on a national scale?

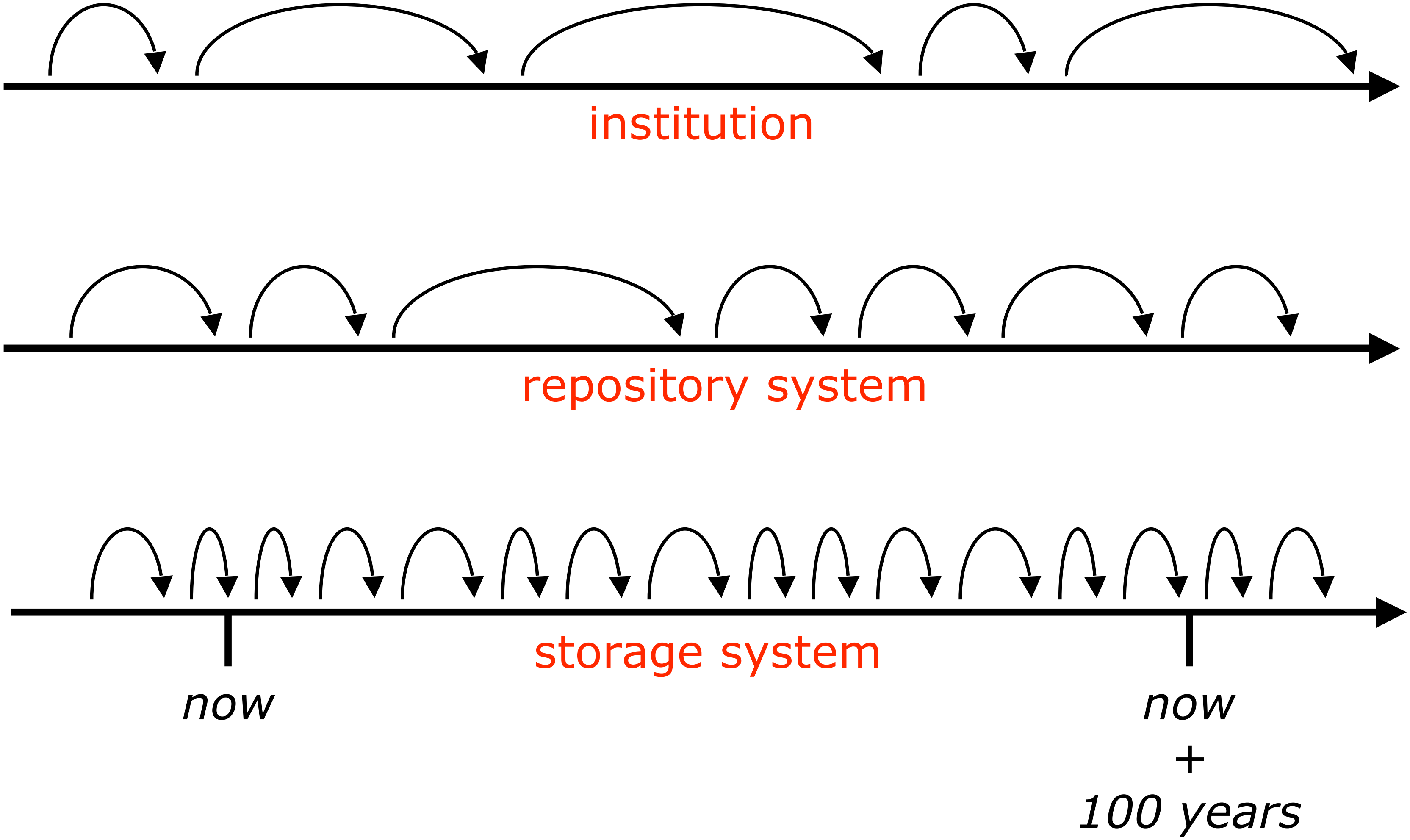
Geospatial data characteristics

- Voluminous
- Sensor platforms are long-lived
- Highly structured
  - support not ubiquitous
- Requires specialized interpretation
- Tied to Earth models

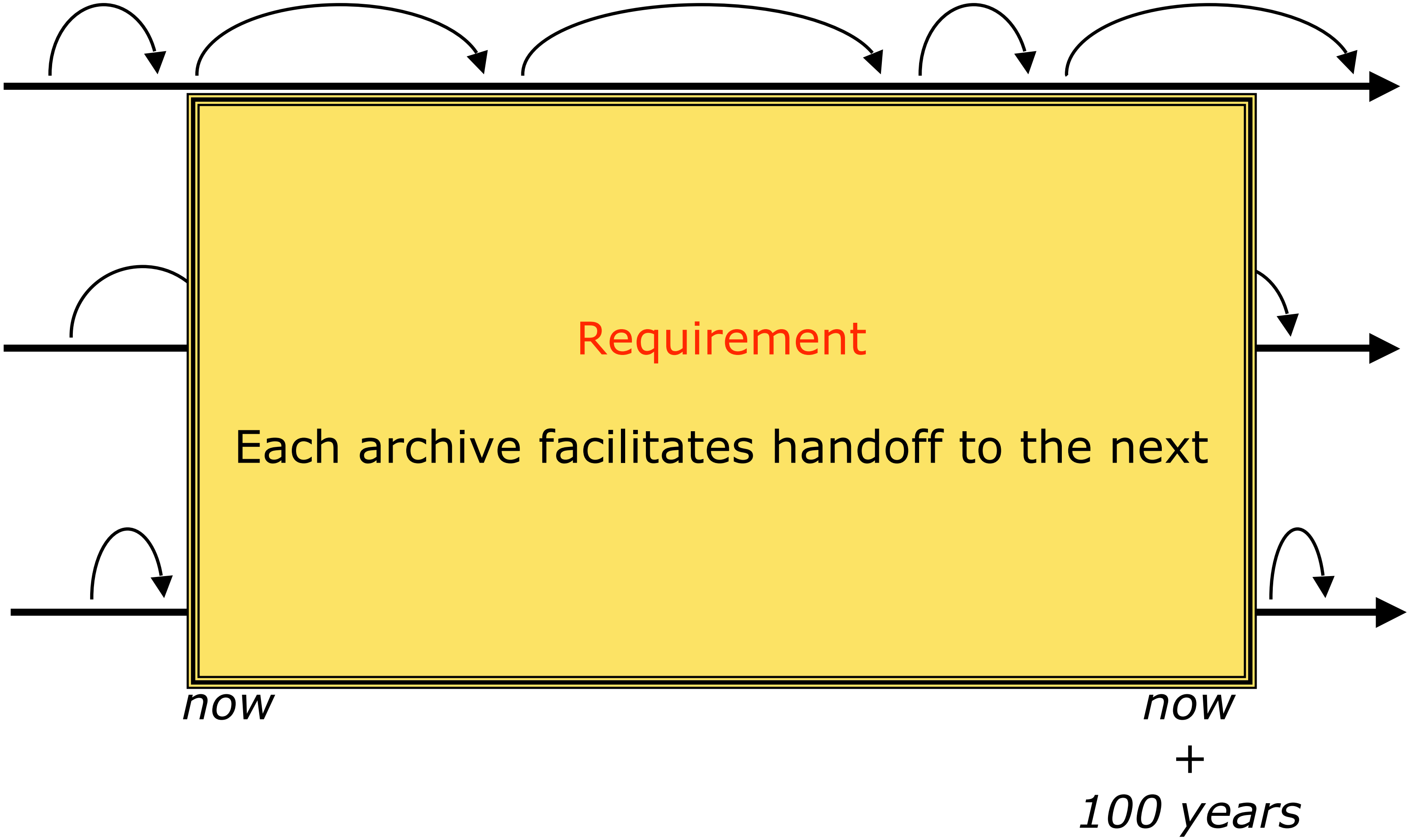
Starting point



Preservation: relay across time



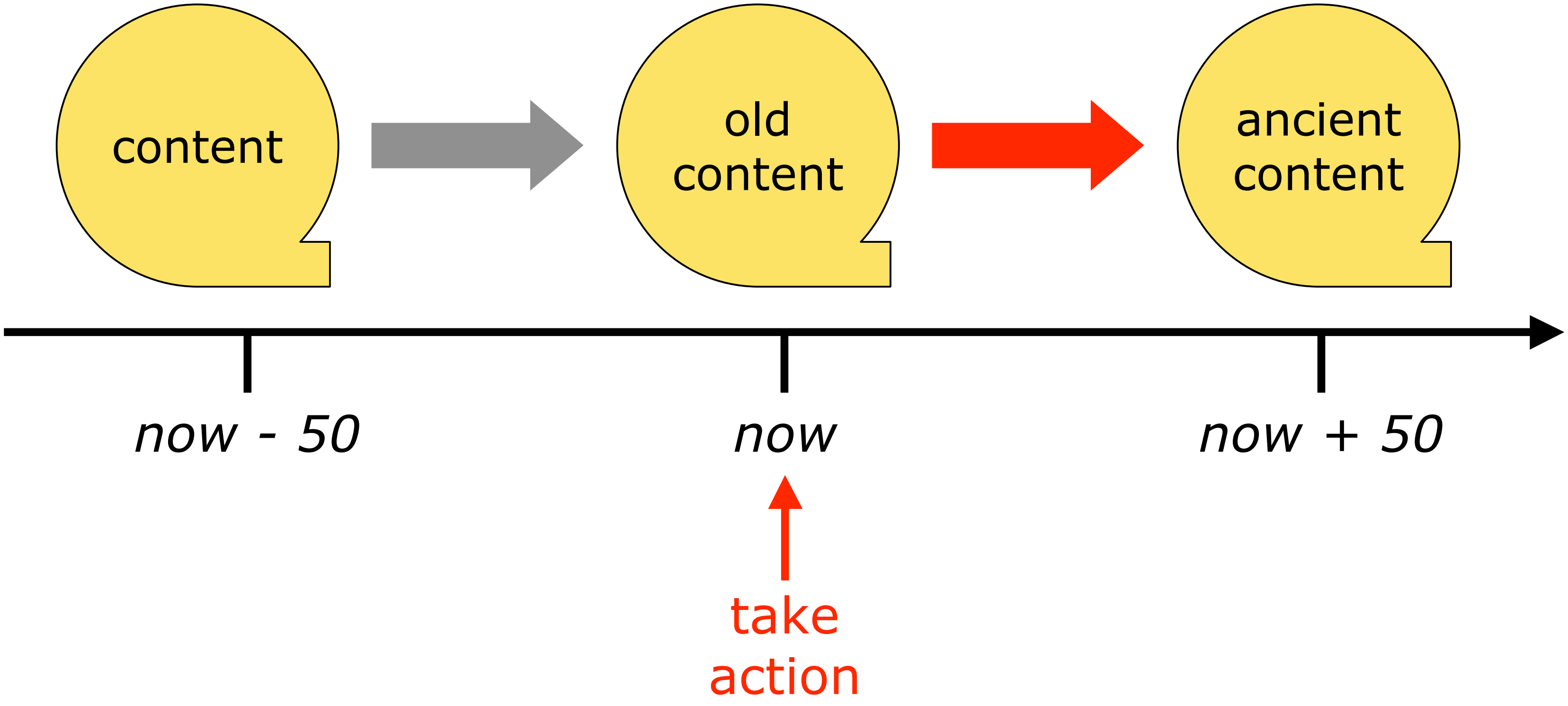
Preservation: relay across time



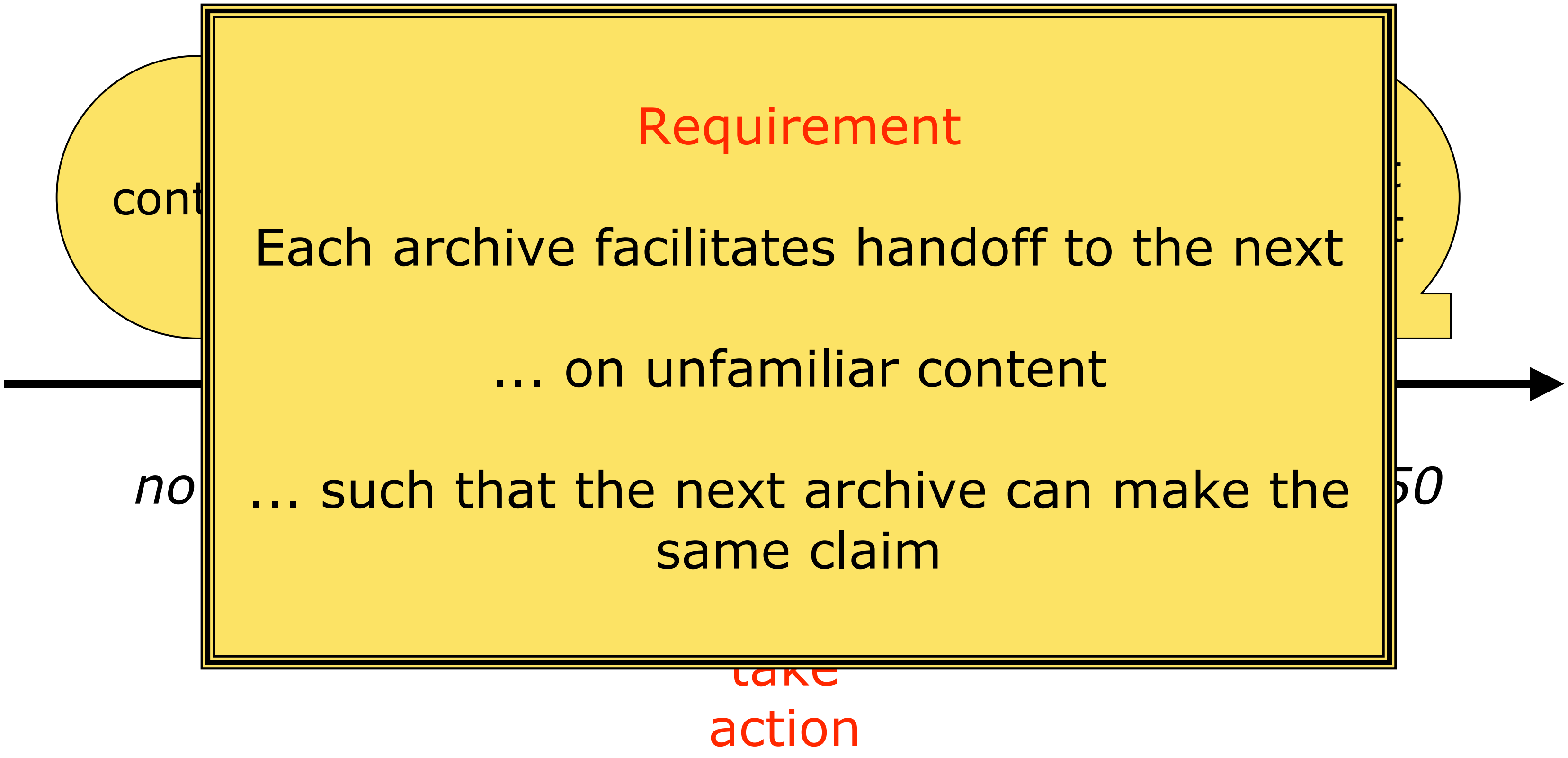
Requirement

Each archive facilitates handoff to the next

Mid-century perspective



Mid-century perspective



Requirement

Each archive facilitates handoff to the next

... on unfamiliar content

... such that the next archive can make the same claim

Preservation: mitigation of risk

- Preservation is an *outcome*
- Risk: insufficient resources and/or desire
- Risk: handoff
  - e.g., from failing institution
  - e.g., from unsupported repository system

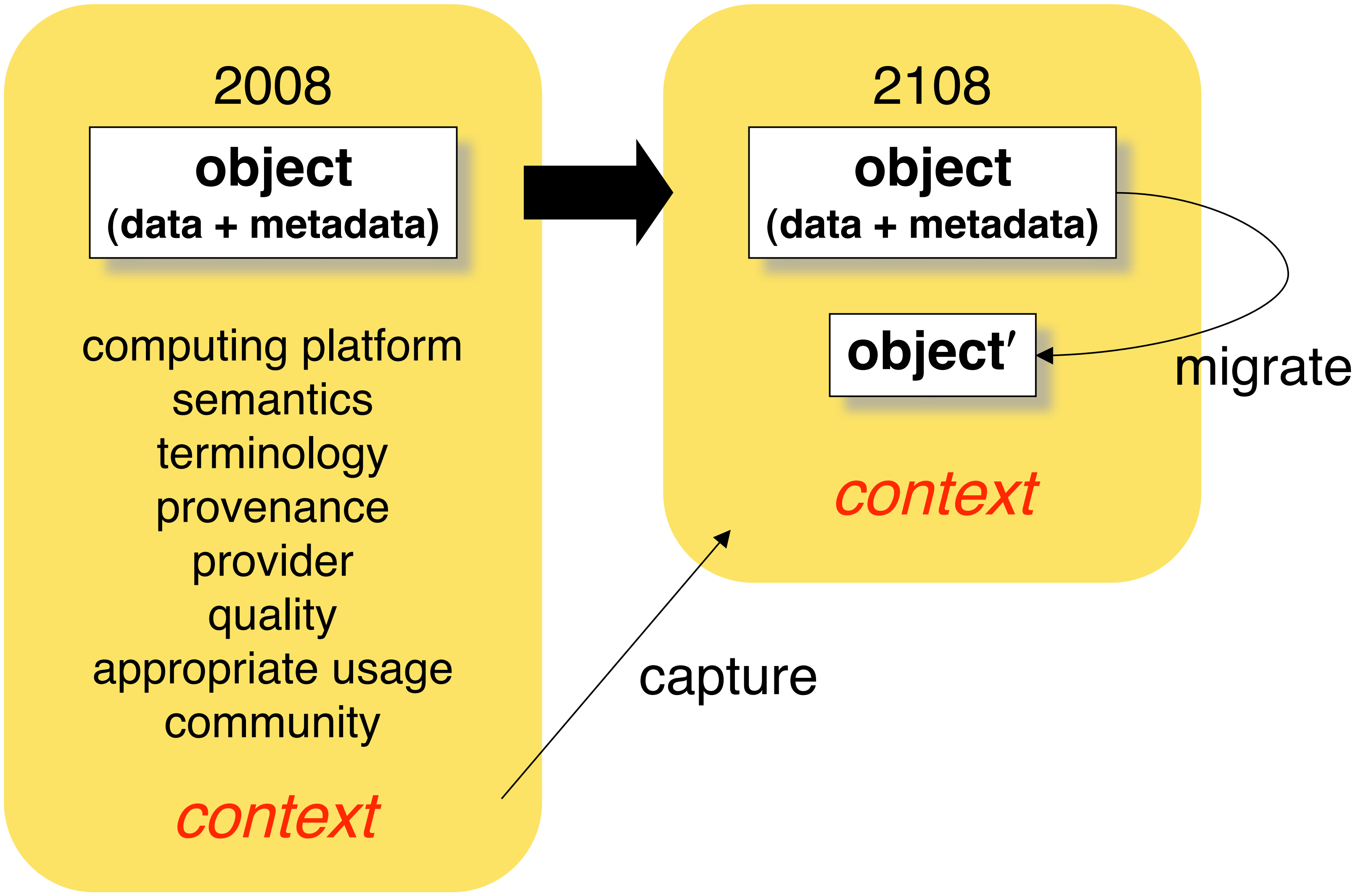
Preservation: mitigation of risk

- Preservation is an *outcome*
- Risk
- Risk
  - e
  - e

Requirement

Each archive supports a low-cost, robust "fallback" preservation mode

Preservation: context



2008

object

(data + metadata)

computing platform
semantics
terminology
provenance
provider
quality
appropriate usage
community

context

2108

object

(data + metadata)

object′

migrate

context

capture

Geospatial data context

- Complex
  - sensor, platform characteristics
- In practice, not handled as metadata
- Deep understanding of provenance required
  - to support reprocessing

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

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

Ozone reprocessing requirements

- xDRs
- Delivered IPs
- Engir
- Uploa
- Datal
- Softw
- Calib
  - d
  - a
  - ta
  - lo
  - n
  - ir
- All pr
- All sc
- All reports

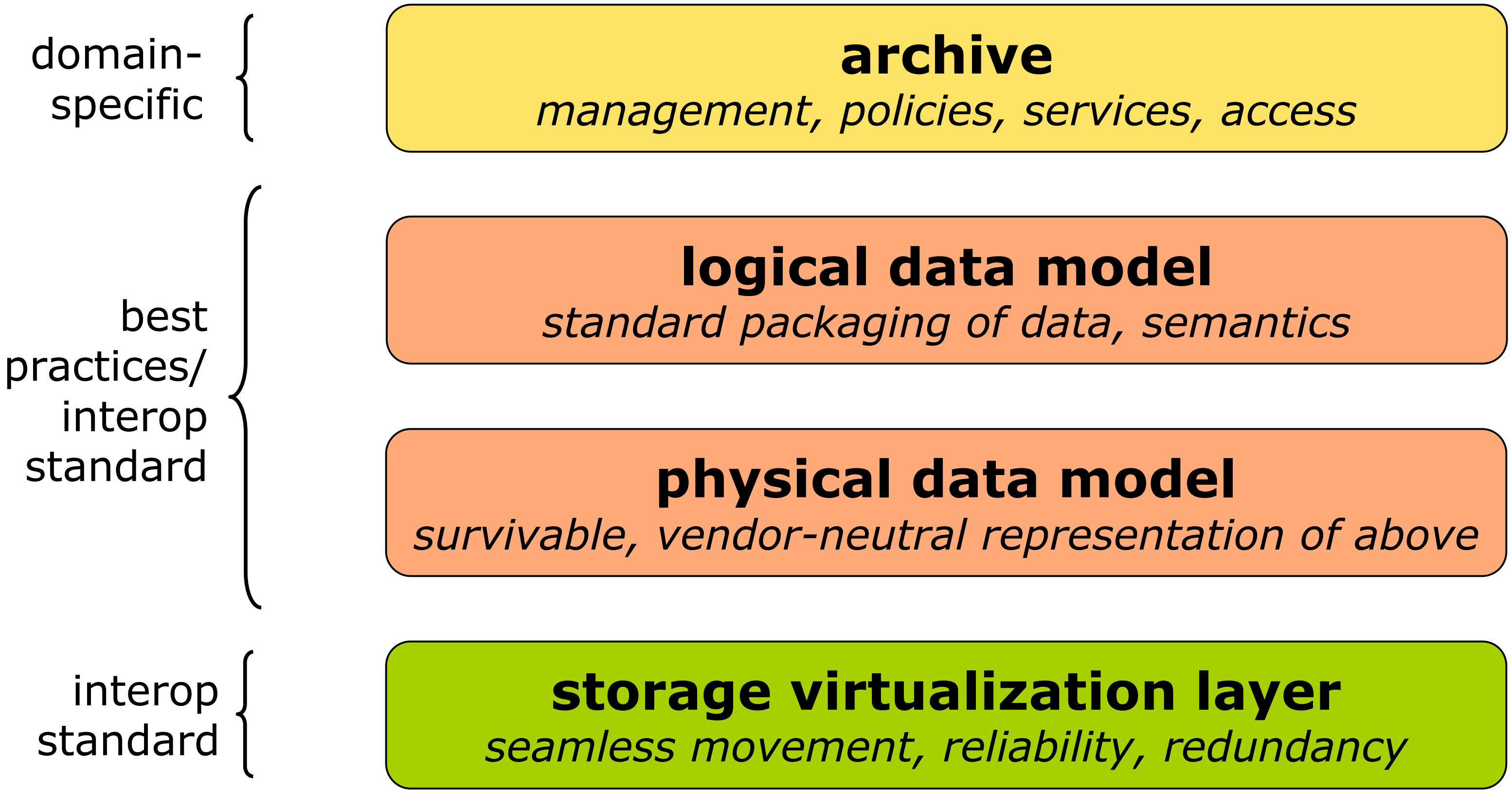
Requirement

Context must be preserved

... and context must accommodate complex networks of objects

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

Architecture



domain-specific

archive
management, policies, services, access

best practices/ interop standard

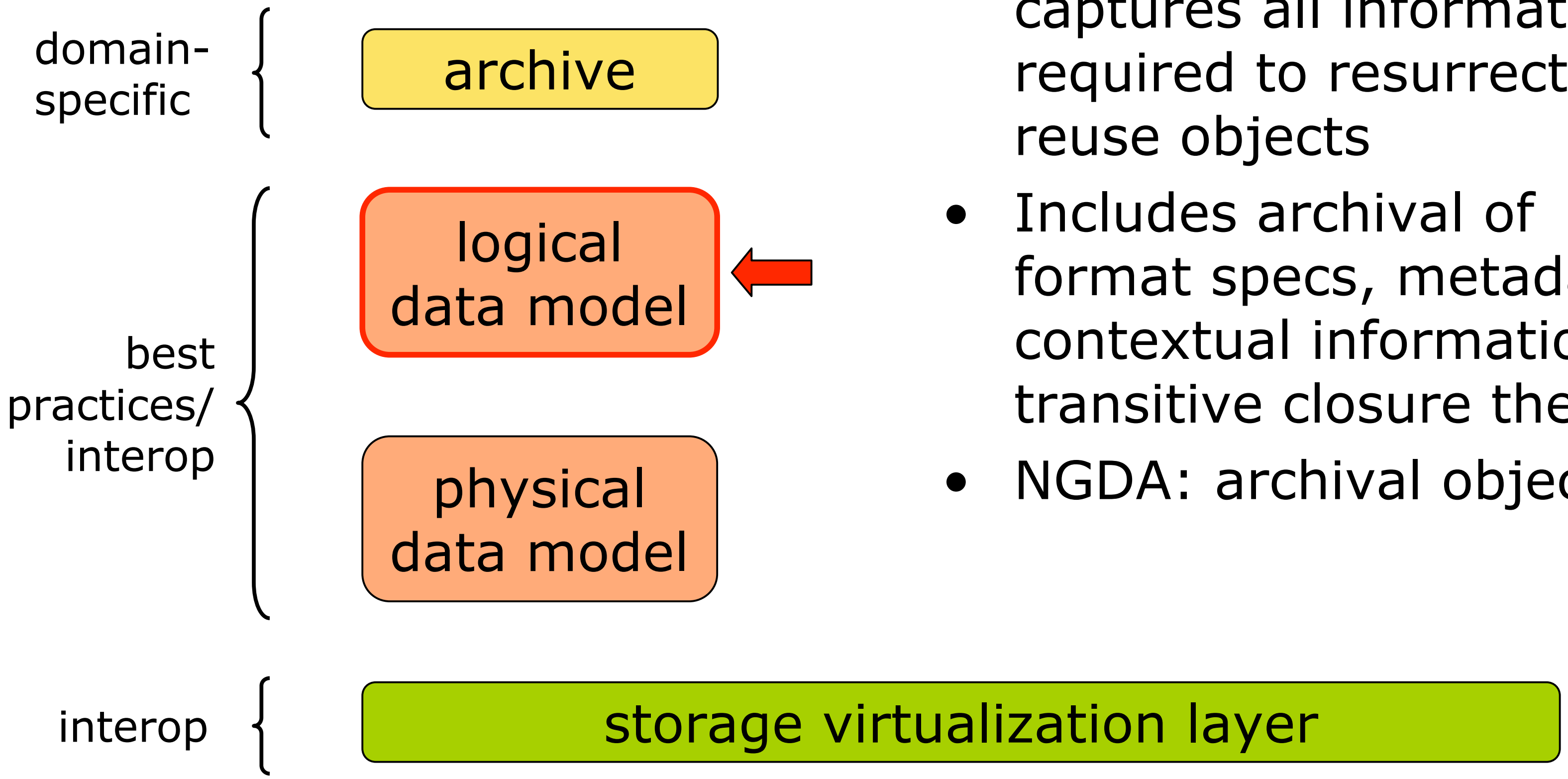
logical data model
standard packaging of data, semantics

physical data model
survivable, vendor-neutral representation of above

interop standard

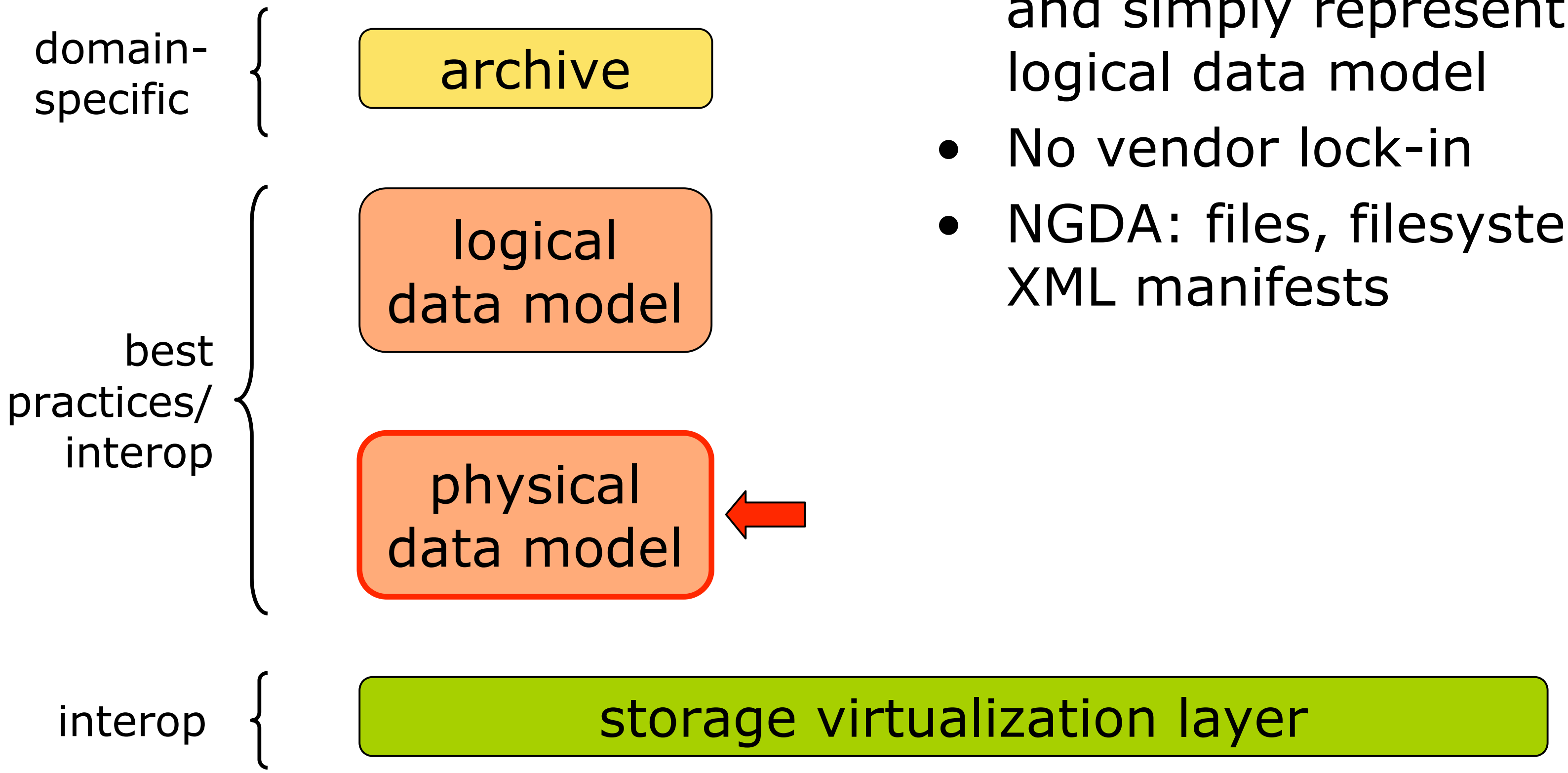
storage virtualization layer
seamless movement, reliability, redundancy

Architecture



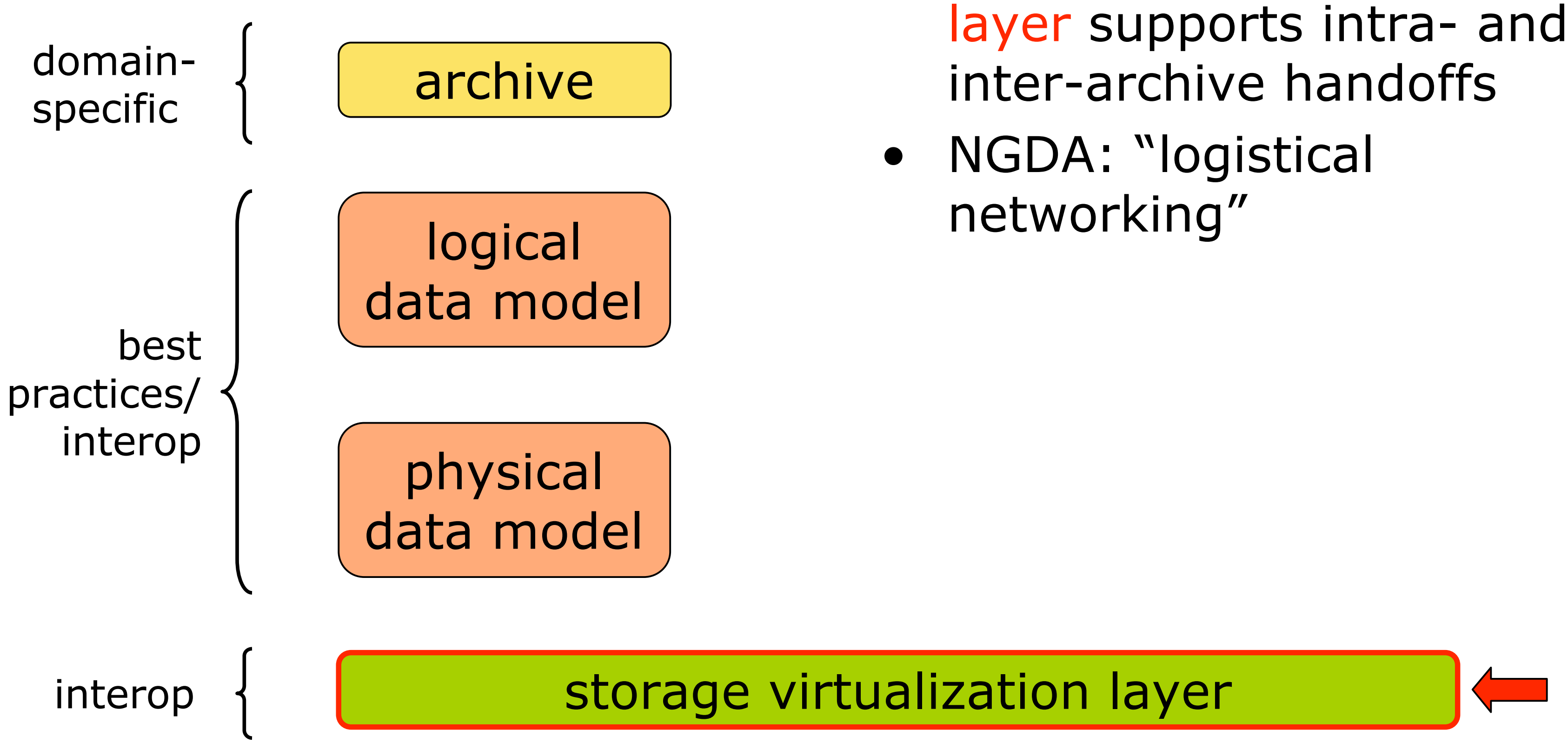
- Logical data model captures all information required to resurrect, reuse objects
- Includes archival of format specs, metadata, contextual information, transitive closure thereof
- NGDA: archival objects

Architecture



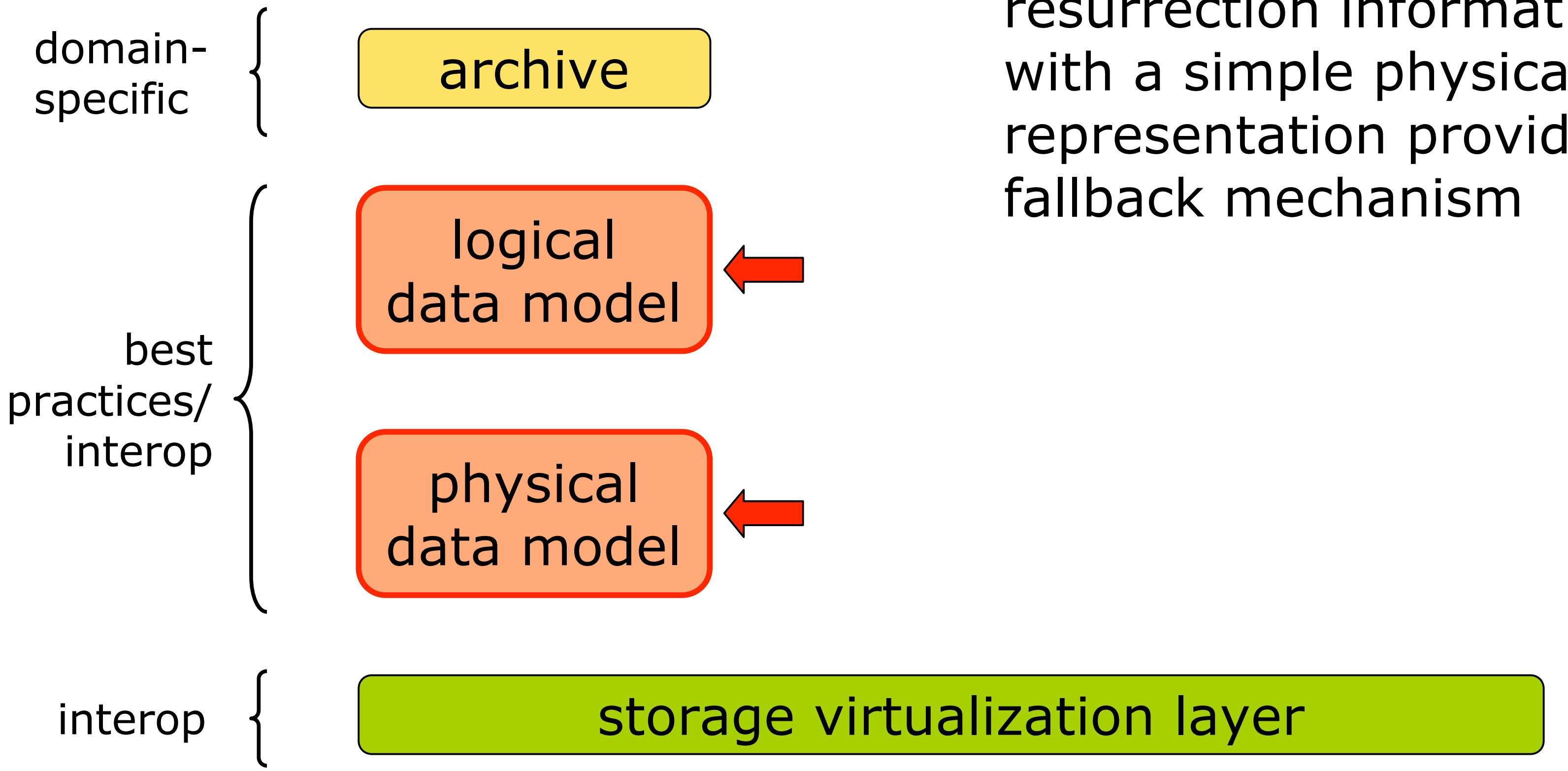
- Physical data model fully and simply represents logical data model
- No vendor lock-in
- NGDA: files, filesystems, XML manifests

Architecture



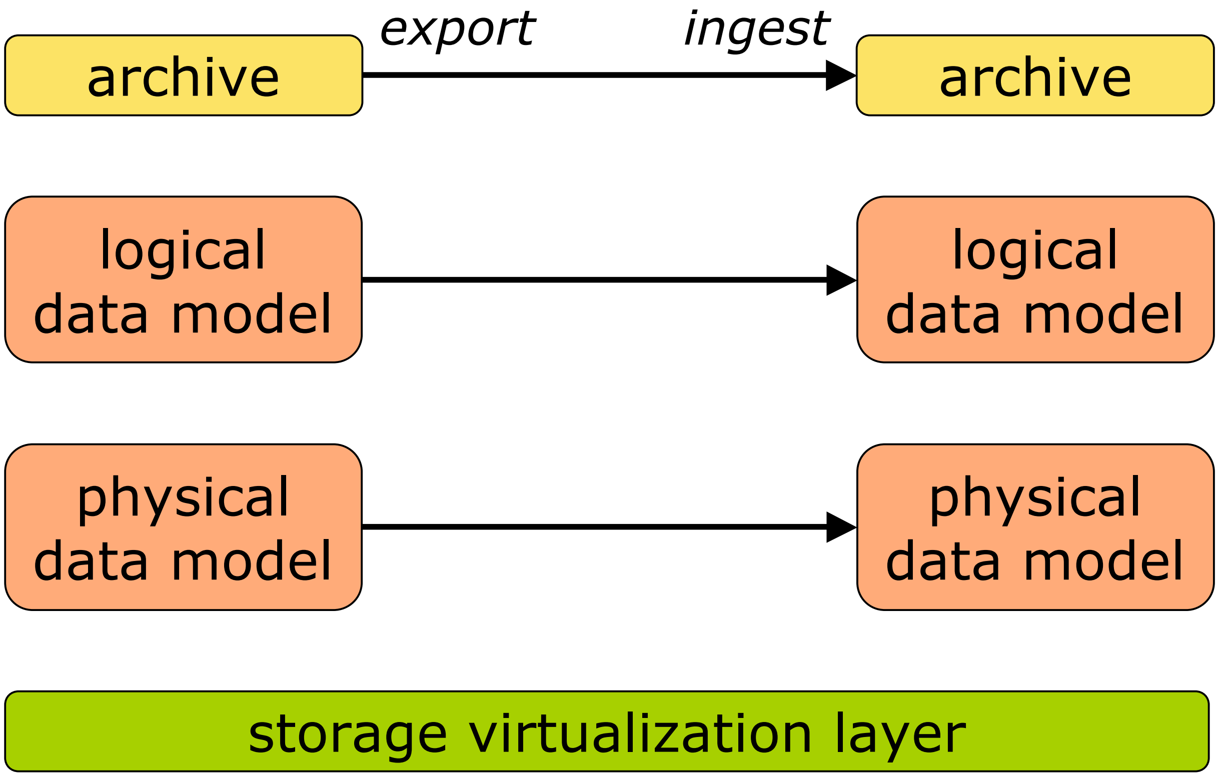
- Storage virtualization layer supports intra- and inter-archive handoffs
- NGDA: “logistical networking”

Architecture: fallback



- Combination of complete resurrection information with a simple physical representation provides fallback mechanism

Architecture: handoffs



export

ingest

archive

archive

logical data model

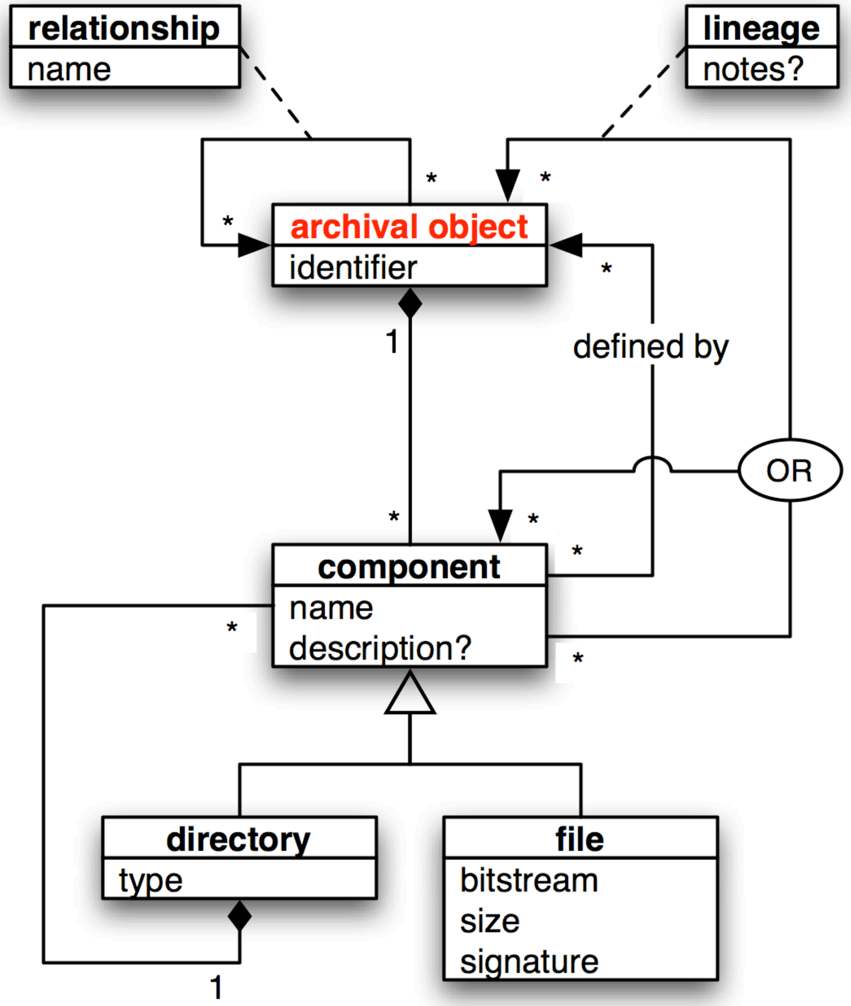
logical data model

physical data model

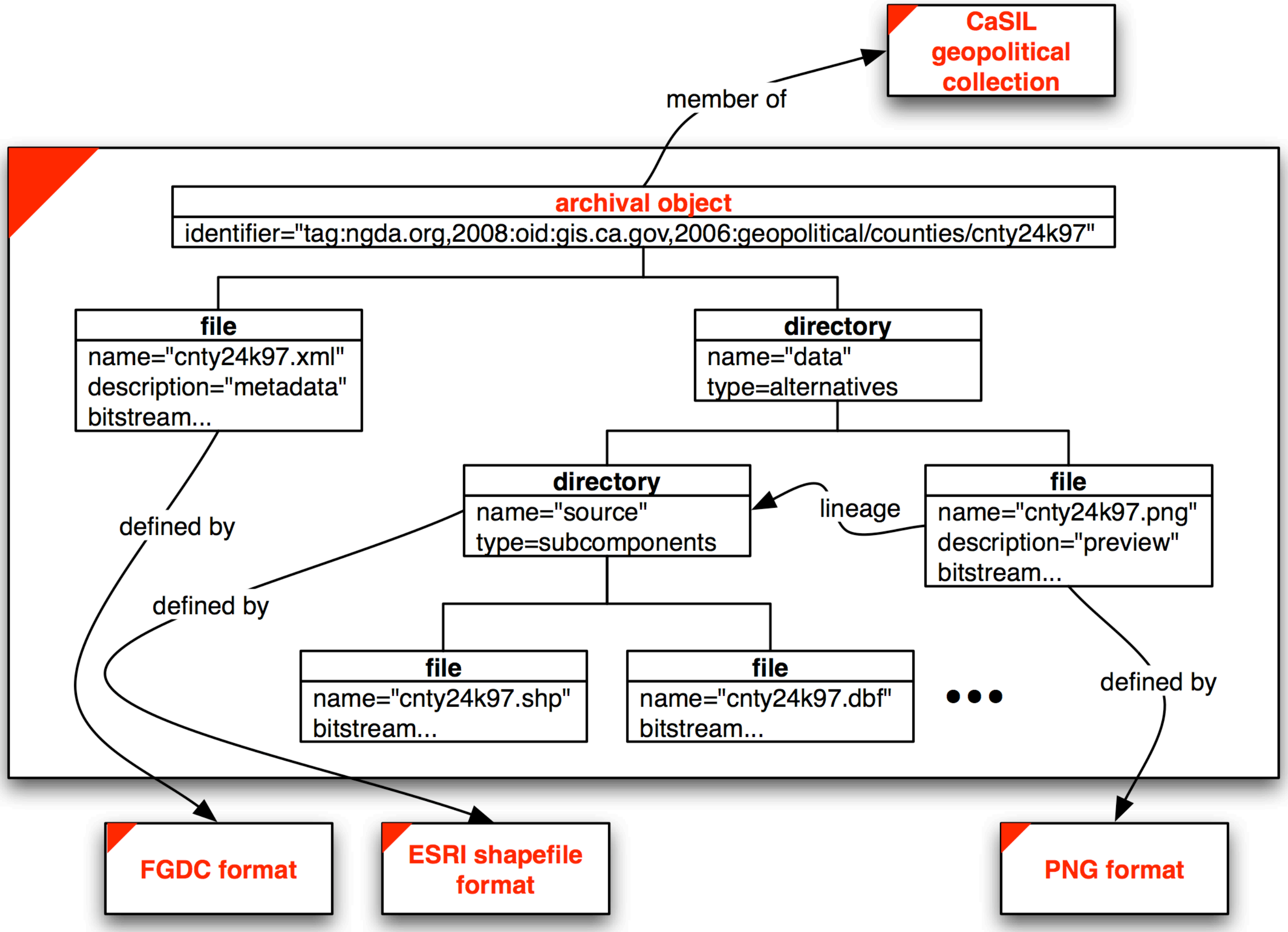
physical data model

storage virtualization layer

Logical data model



Example archival object



Physical data model

```
...identifier/
    manifest.xml
    cnty24k97.xml
    data/
        source/
            cnty24k97.shp
            cnty24k97.dbf
            ...
        cnty24k97.png
```

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

Storage abstraction

- Bitstreams
  - create, (delete), read, write
  - no modify
- Directories
  - create, (delete), list members
- Above identified by hierarchical pathnames

- Satisfied by filesystems, WebDAV, ...

Archive depencies

- Filesystem
- XML
- Character set(s)
- Identifier resolution mechanism(s)

Summary

- Architecture to facilitate handoffs, reduce risk, provide fallback
  - best practices
  - interoperability potential

- Ongoing work
  - “logistical networking” for storage virtualization
  - preservation profiles for other data models
  - format registries and other achive depencies
  - whole-archive descriptor
    - dependencies, policies

Questions?