Relay-supporting Archives: Requirements and Progress

GREG JANÉE & JAMES FREW
University of California, Santa Barbara
TERRY MOORE
University of Tennessee, Knoxville

Summary

- Long-term preservation of digital content is an extended **relay** over time
- Repeated **handoffs** will occur at the
  - **physical** layer: bits ↔ storage systems
  - **logical** layer: objects ↔ repositories
  - **administrative** layer: collections ↔ archives
- Current archiving technologies could use some help making these handoffs easier
- A **relay-supporting** archival infrastructure will mitigate the fundamental risk of information loss
- Let's find out: AIHT-TNG

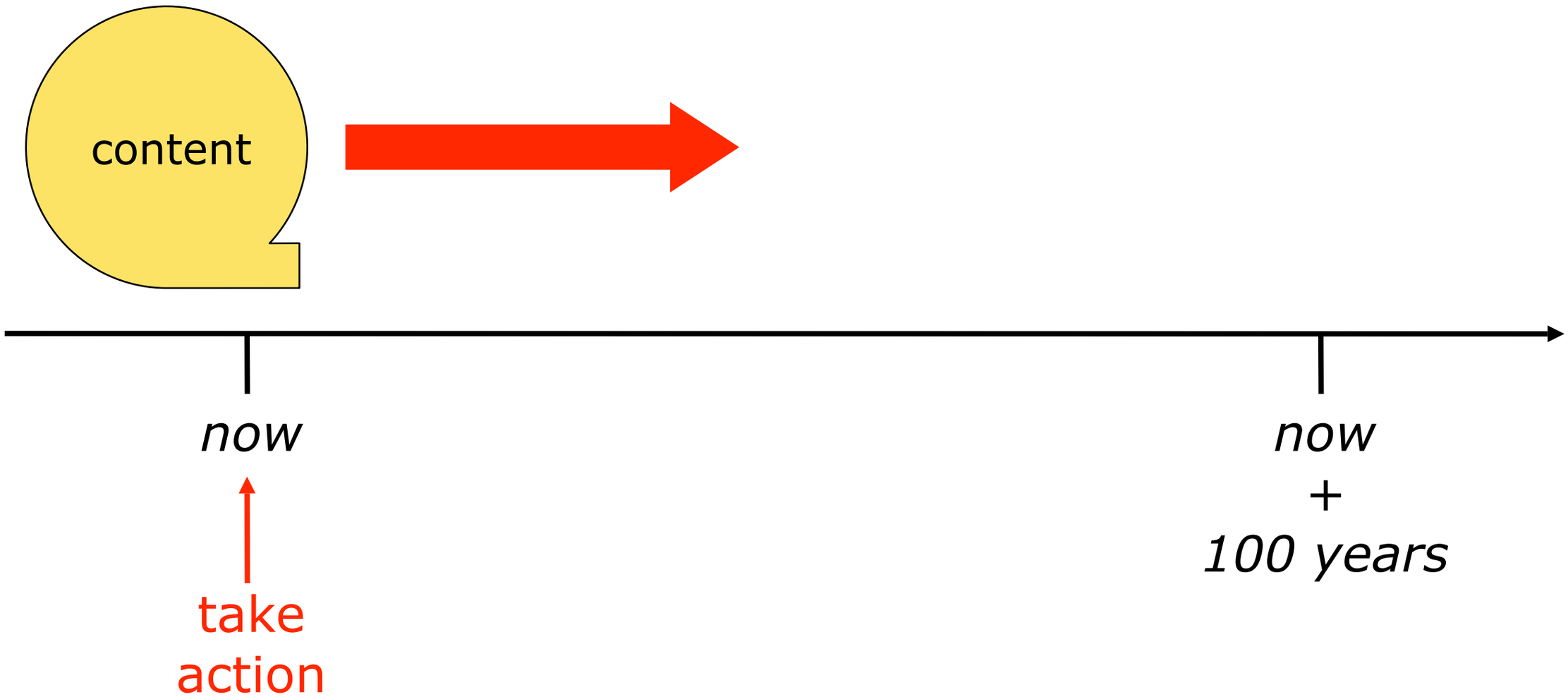
Outline

- Preservation is a relay
- Archive layers
- Core assumptions
- Archive migration
- Conclusions
- Afterword: Geospatial data curation

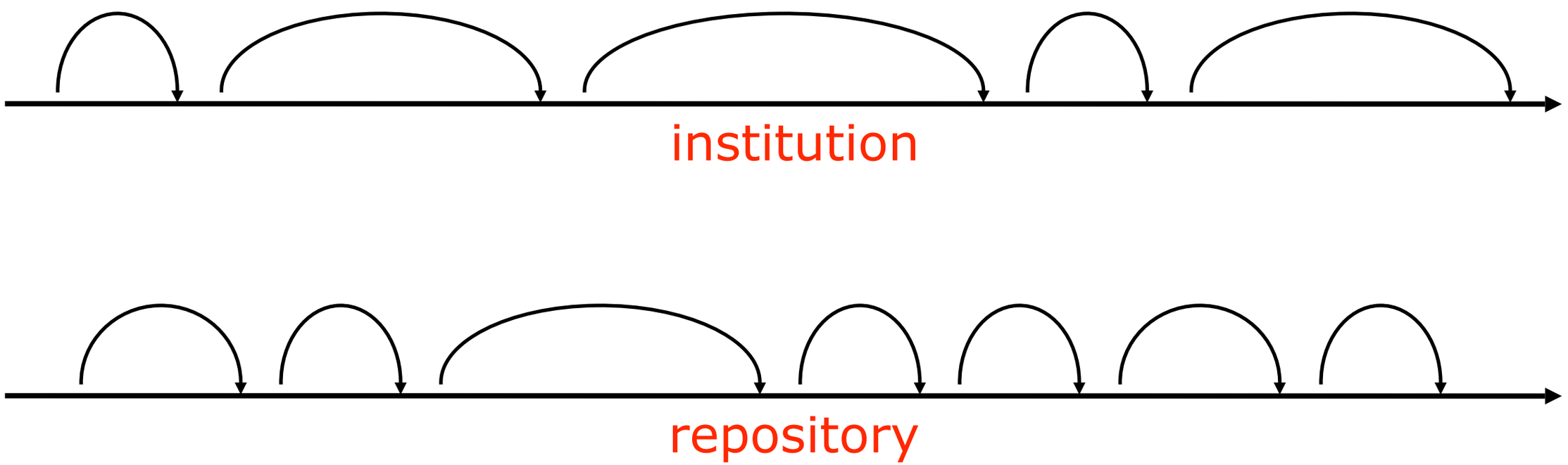
Outline

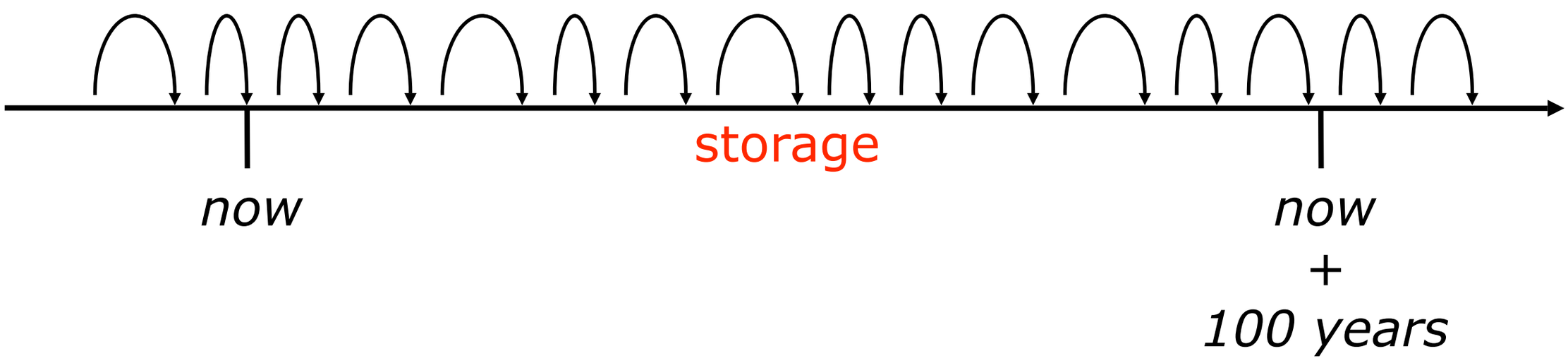
- Preservation is a relay
- Archive layers
- Core assumptions
- Archive migration
- Archive relay
- Conclusions
- Afterword: Geospatial data curation

Preservation **relay**: Starting point

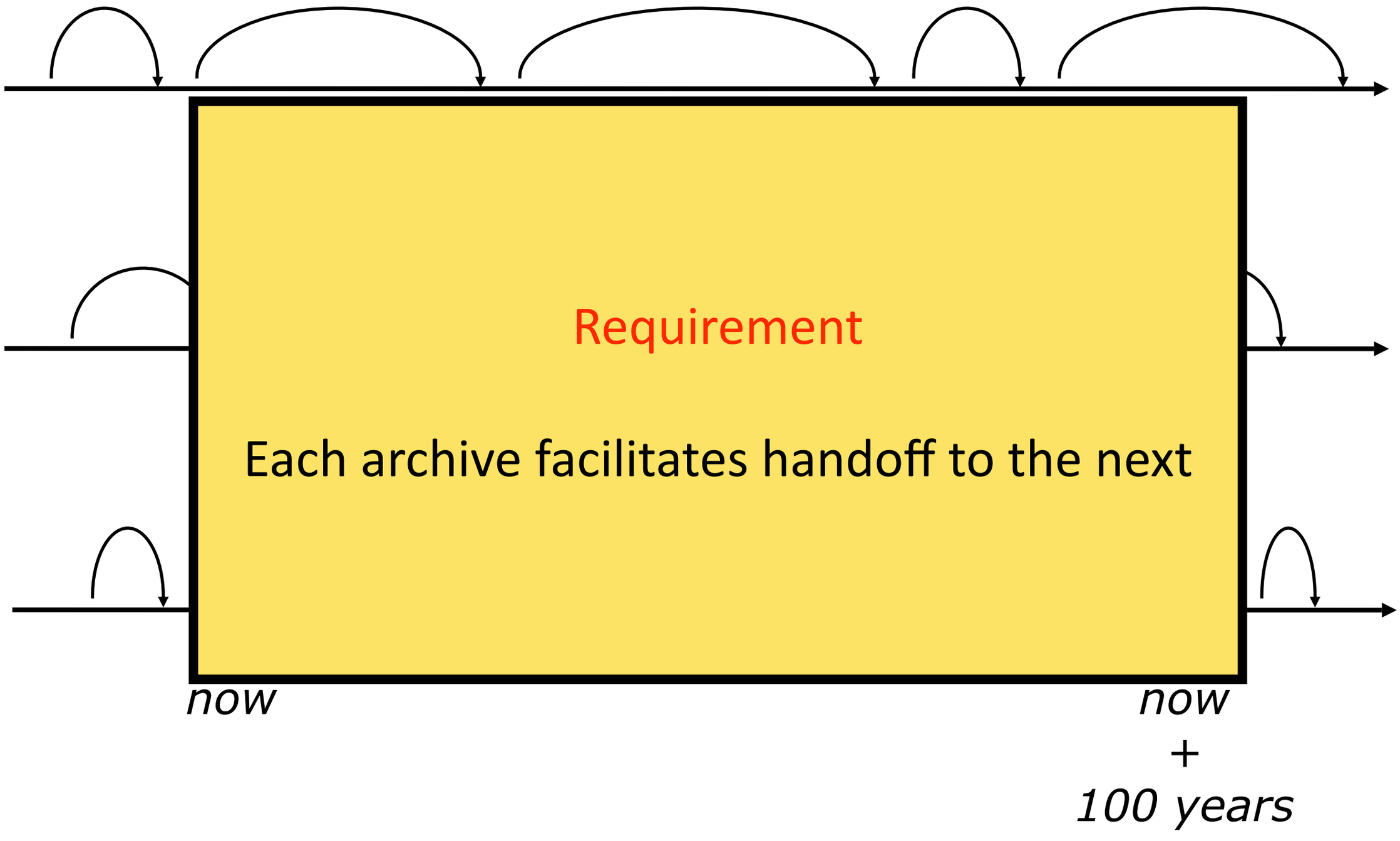


Preservation **relay**: across time

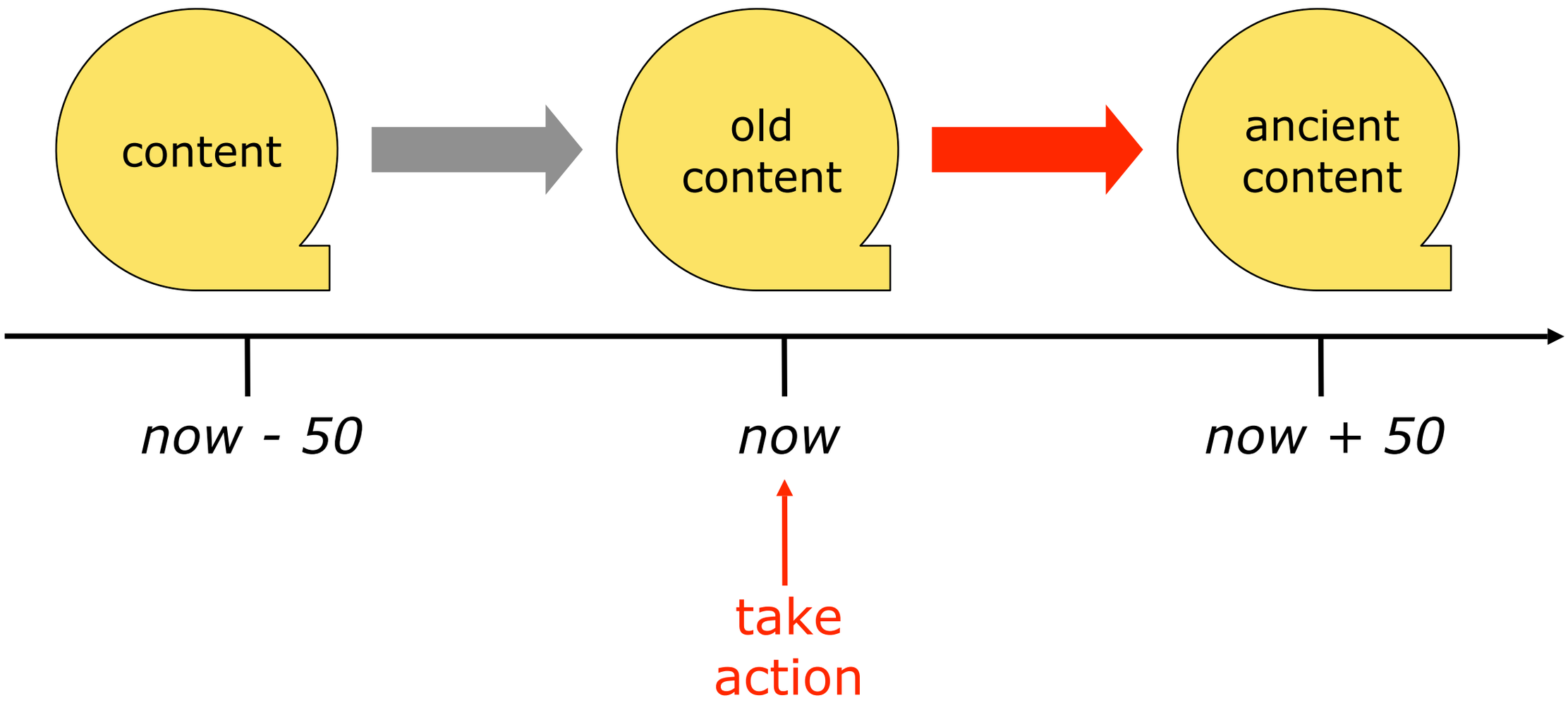




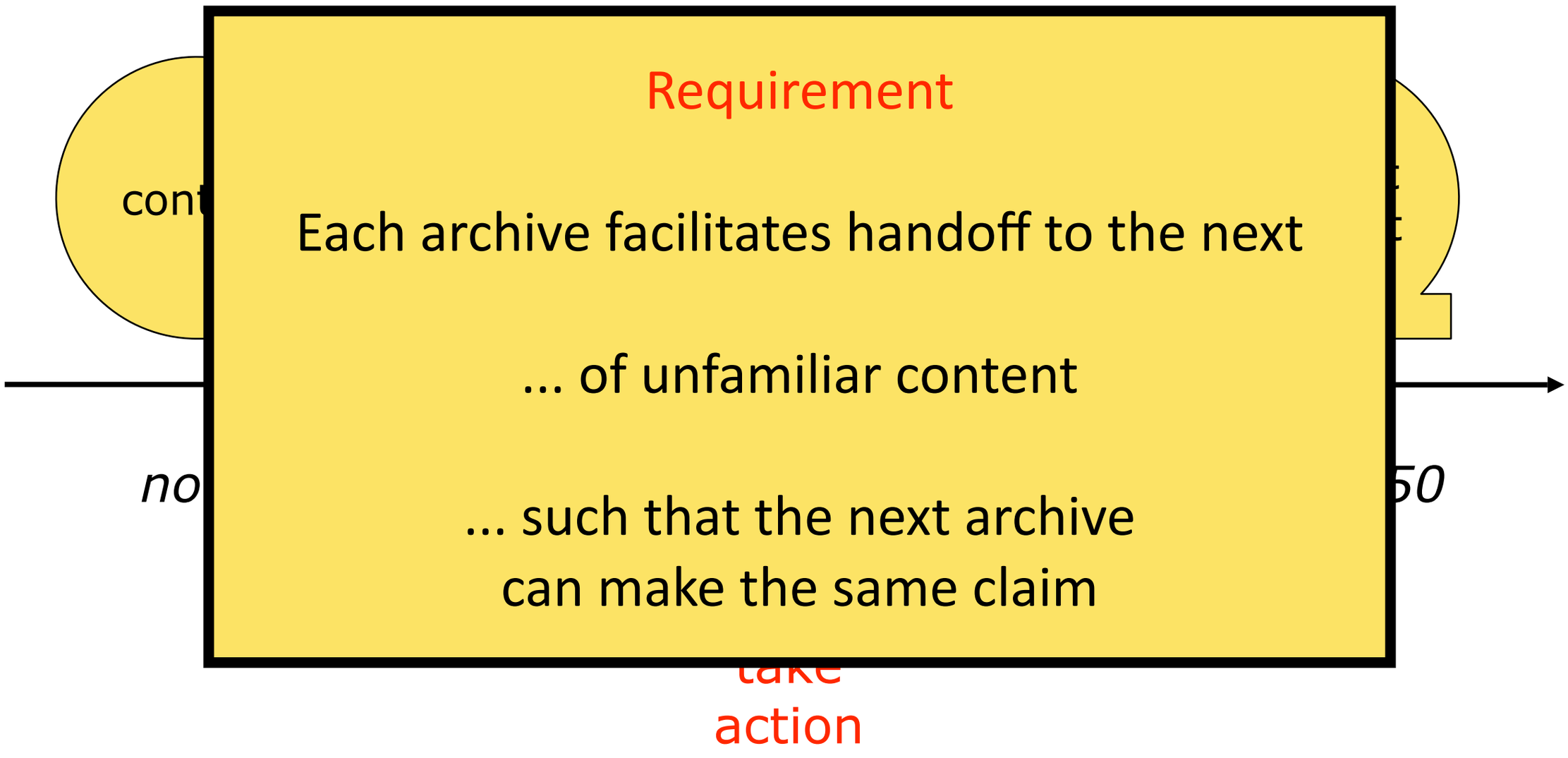
Preservation **relay**: across time



Mid-century perspective



Mid-century perspective



Mid-century perspective—
Repeated handoff/migrations:

- Across storage media / systems
  - past and future
- Across archive systems
  - transformation
  - reorganization
- Between institutions
  - policies

Challenge

- What kind of archive infrastructure can best facilitate the *ongoing* preservation of digital information, where a whole range of transitions at a variety of different levels will be required?

Outline

- Preservation is a relay
- Archive layers
- Core assumptions
- Archive migration
- Conclusions
- Afterword: Geospatial data curation

Archive Layers

- Physical
- Logical
- Administrative

Archive Layers: Physical

- Manages *bit sequences*
  - independent of interpretation
- Minimum requirement: read/write identified b.s.
  - e.g., any filesystem
- Archive problem =
  - guarantee storage reliability
  - arrange for copy-out when guarantee exprires
- Archive problem ≠
  - semantics
  - storage system implementation

Archive Layers: Logical

- Manages *objects* and their *relationships*
  - e.g. document ↔ format
- Imposes *structure* and *semantics* on physical layer
- Mechanism not specified: could be:
  - filename (“document.pdf”)
  - “self-describing” format
  - link(document, format description)

Archive Layers: Administrative

- Manage
  - *Collections*
  - *Services*
    - e.g. content-specific search & presentation
  - *Policies*
    - e.g. selection, access, maintenance, ...
- Examples
  - stream(video(bits))
  - monitor(object(storage guarantee))

Outline

- Preservation is a relay
- Archive layers
- Core assumptions
- Archive migration
- Conclusions
- Afterword: Geospatial data curation

Core Assumptions

- Curation is horizontal
- Resurrection is more likely than immortality
- Interfaces should be minimal

Curation is Horizontal

- Individual digital archives won't last a century or more
  - must assume archives will fail
- Content migration interfaces must scale over time
- Solution: interface *stack*
  - analogous to communication networks
- Horizontal interoperability
  - single archive ↔ multiple physical layers
  - single physical/logical implementations ↔ multiple administrative policies and services

Resurrection Is More Likely Than Immortality

- Preservation must be cheap and easy
  - small burden on providers
  - flexible/adaptable infrastructure
  - level-of-effort accommodates variable resources
- Minimal option
  - e.g. Internet Archive (crawl → save → reflect)
    - captures *current* context

Interfaces Should Be Minimal

- Competing philosophies
  - "Core and extension"
    - minimum common functionality
    - negotiable extensions
  - "base and profile"
    - maximum common functionality
    - negotiable profiles (subsets)
- Archives should use core+extensions
  - guarantees fallback to core interfaces
  - simpler interfaces more likely to be supported by future archives
    - especially if considered "obsolete"

Outline

- Preservation is a relay
- Archive layers
- Core assumptions
- Archive migration
- Conclusions
- Afterword: Geospatial data curation

Archive Migration Today

- Physical: copy bits
  - stress networks, filesystems, bandwidth
  - opportunities for corruption
- Logical: match schemas
  - repository structures
  - content representations
- Administrative: negotiate treaties

Archive Relay: Physical Layer

- Global storage virtualization
  - archives share storage using open network protocol
    - e.g. LOCKSS, Logistical Networking (LN), ...
  - don't need to copy bits; just hand off storage references

Archive Relay: Logical Layer

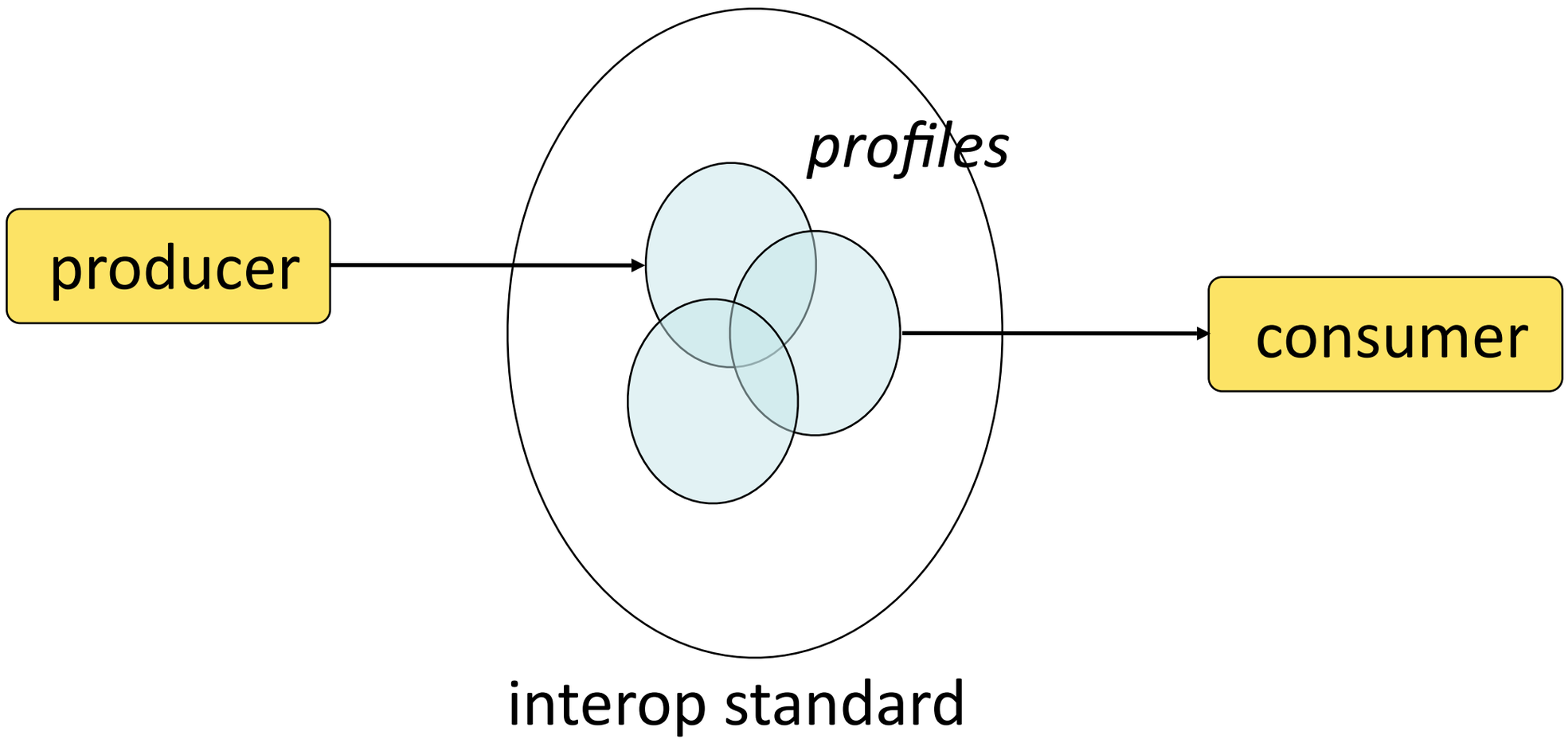
- *What* object is
  - boundary of object
  - integrity of components within boundary
- *How* to interpret object
  - internal metadata
  - “self-describing” files
  - persistent associations to format registries, other interpretation-defining objects

Archive Relay: Logical Layer

- Bitstreams
  - hold object content
- Objects
  - aggregative mechanism for bitstreams
- Identifiers
  - objects: persistent, universally unique
  - bitstreams: locally unique, for disambiguation
- Fixity metadata
  - support end-to-end integrity
- Inter-object relationships
  - define object, component semantics in transitive representation net
  - other ontological assertions

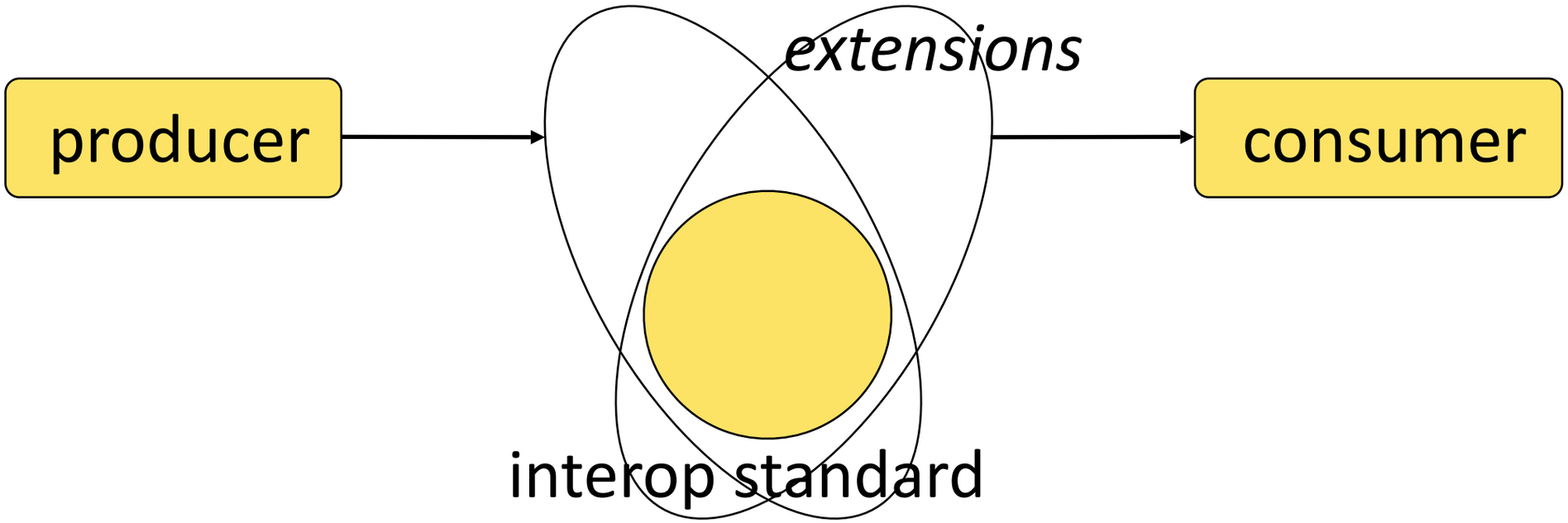
Too-large standards

- METS (GML, others...)
  - Easy for writers to produce
  - *but,* hard for readers to support in entirety



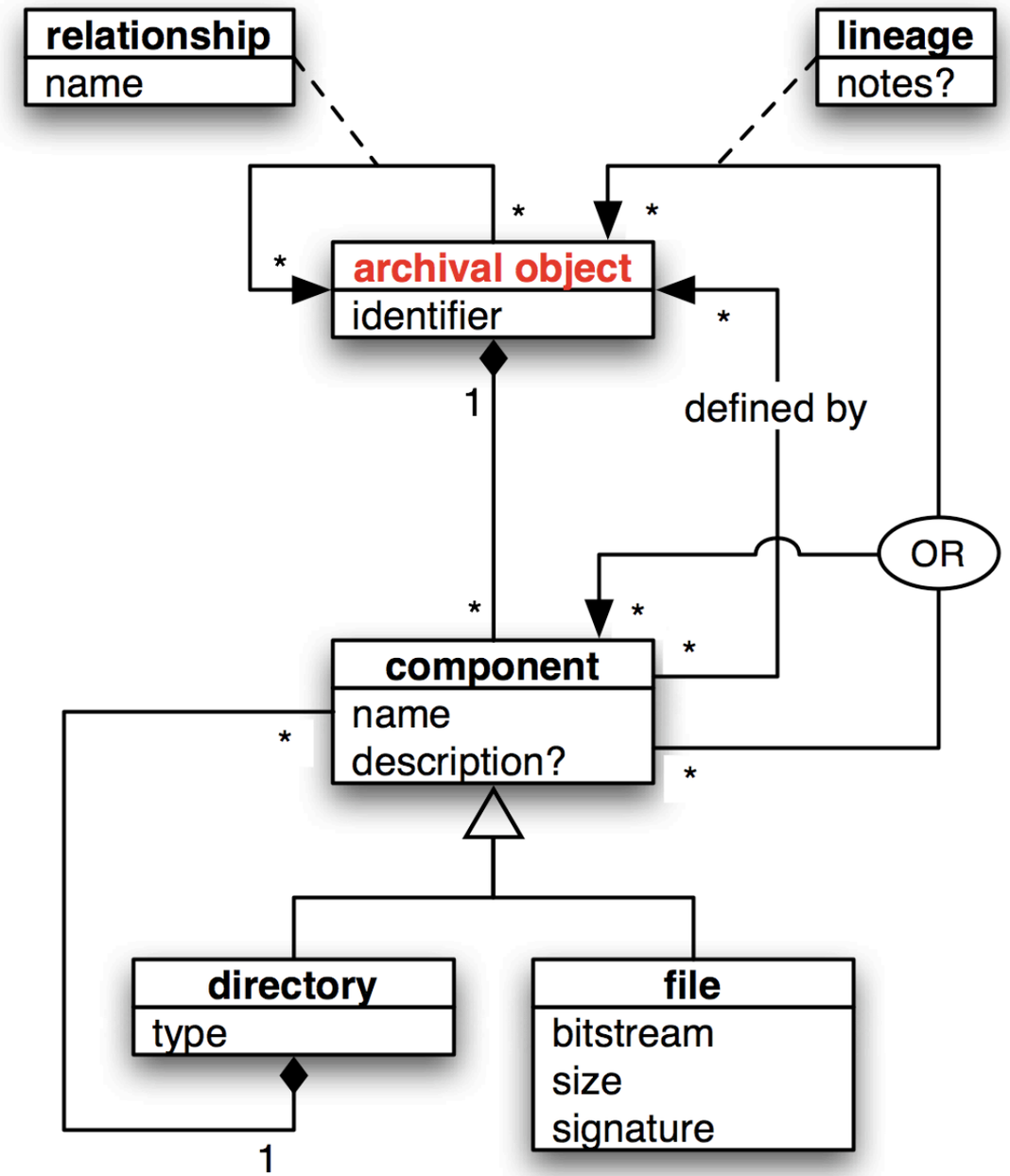
Too-small standards

- OAI-ORE
  - current focus of cross-repository interoperability
  - *but,* doesn't specify *standard* representations of required elements
    - fixity metadata, semantics, ...



Just right?

- NGDA data model
  - match preservation relay requirements
    - no more, no less



Archive Relay: Administrative Layer

- "Crawl point"
  - root object from which any other object in the archive may be (transitively) reached
    - analogous to "super block"
- Whole-archive dependency descriptor
  - archive's external dependencies
    - e.g. format registries, PID resolution systems, ...

Outline

- Preservation is a relay
- Archive layers
- Core assumptions
- Archive migration today
- Archive relay
- Conclusions
- Afterword: Geospatial data curation

Conclusion:
Relay = Baseline Interoperability

- Physical
  - common network storage abstraction
- Logical
  - standard uniform data model
- Administrative
  - standard crawl points
  - standard whole-archive dependency descriptors

Conclusion: Test Physical Relay

- Implement multiple repositories atop common storage substrate
- Should be able to
  - replicate data across dissimilar storage systems
  - migrate an archive without copying any content bits
  - change storage attributes via policy changes

Conclusion: Test Logical Relay

- Core data model supported by multiple archives
  - This will be (ahem) hard...
- Handoff variety of
  - object types
  - subject domains
- Specifically: handoff potentially incompatible content
  - e.g. streaming video → geospatial archive

Conclusion: Test Administrative Relay

- Mimic institutional handoffs
  - especially where source institution provides no support or guidance
- Receive & support content
  from archive that doesn't support it
  - e.g. streaming video ← geospatial archive

Conclusion: Time for a New AIHT

- Original was technology neutral
- New one should test a specific set of candidate interoperability technologies

Outline

- Preservation is a relay
- Archive layers
- Core assumptions
- Archive migration today
- Archive relay
- Conclusions
- Afterword: Geospatial data curation

Geospatial Information Semantics

- Time-dependence
  - repeated satellite orbits
- Vertical-dependence
  - elevation, pressure, …
- Massive data volumes
  - f(x, y, z, t, parameter)
- Unique data models
  - swath, vector field, …
- Unique data formats
  - HDF, HDF-EOS, netCDF, …
- Unique access interfaces
  - OpenDAP, OGC, …
- Raw values matter
  - not just colors…
- Multiple parameters
  - wavelength, polarization, …
- Version control
  - calibration, algorithms, …
- Multiple frames
  - projection, time, …
- Special values
  - Missing, out-of-range, …
- Long time series
  - decades ...

Example:
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
- Enginee
- Upload f
- Databas
- Software
- Calibrati
  - data
  - analys
  - tables
  - logs
  - noteb
  - instru
- All proje
- All scien
- 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

Questions?