A LAYERED-STACK ARCHITECTURE PROPOSAL FOR PORTABLE AND SCALABLE LONG-TERM DIGITAL DATA ACCESS

Keith Johnson, Preservation Architect

Stanford University

keithj@stanford.edu

PERSPECTIVE

- Provisioning centralized long-term digital preservation services.

- *Economic* Approach

- Effectiveness = Real-World Results
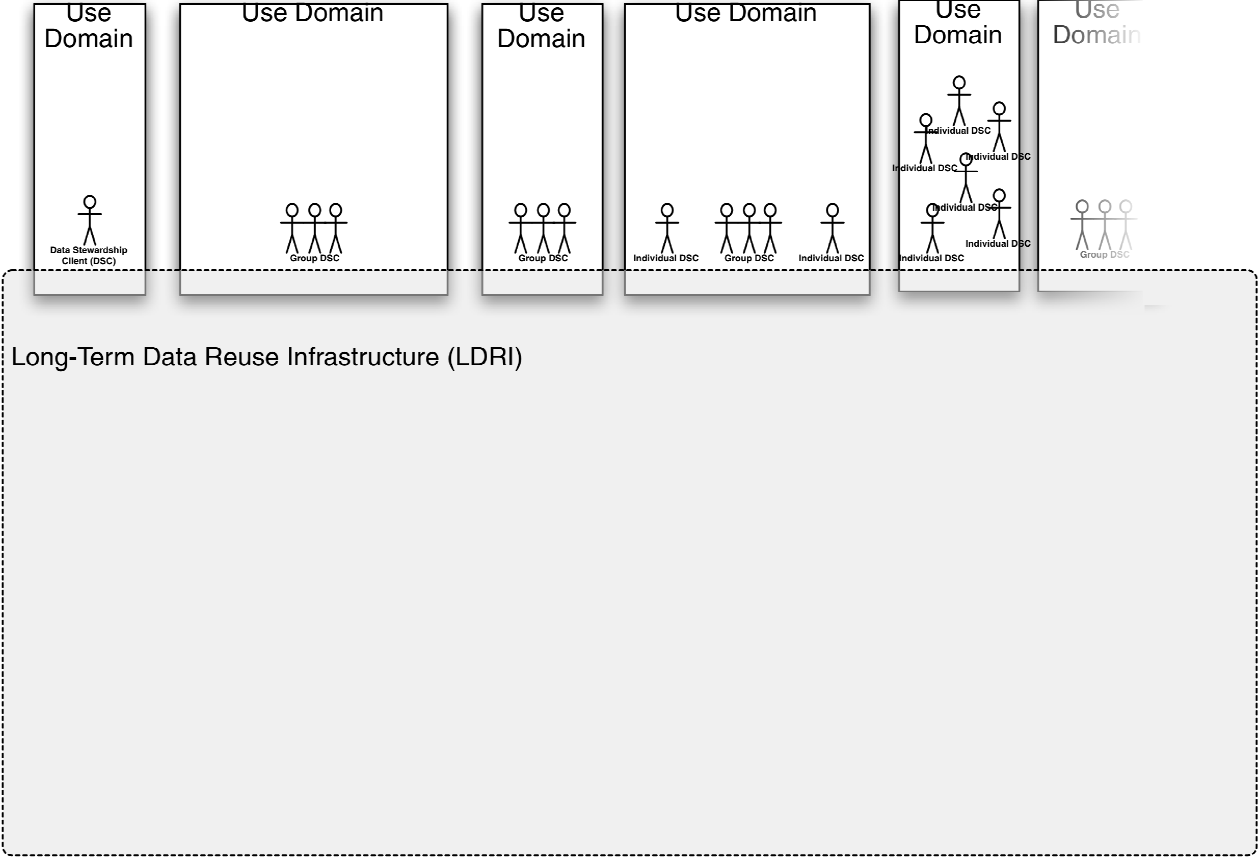
- *Risk Modeling* Methodology

WHAT CONSTITUTES EFFECTIVE DIGITAL PRESERVATION?

- …when Digital Information Resources are *used* in the future.

1. When a Resource—that *could* be valuable in a future activity—*is* valuable in that activity.
2. When previous stewardship decisions compromise its future Value the *least*.

(2a. When many use contexts are facilitated.)

CONCEPTUAL MODEL

Use Domain — Data Stewardship Client (DSC)

Use Domain — Group DSC

Use Domain — Group DSC

Use Domain — Individual DSC, Group DSC, Individual DSC

Use Domain — Individual DSC, Individual DSC, Individual DSC, Individual DSC, Individual DSC

Use Domain — Group DSC

Long-Term Data Reuse Infrastructure (LDRI)

DEPOSIT RISKS

Producer fails to deposit

Deposit does not prove Useful
(especially to Producer)

BIT-PRESERVATION RISKS

Corruption

Systemic Doubt

Deposit does not prove Useful
(Not "complete", expensive to
start working with.)

OVERARCHING RISKS

Failure to Scale

…predictably, linearly,
in both directions

Failure to adapt
to current resources

DISTURBING PHENOMENON

Agility in access—generally valuable

Immutability (aka corruption detection) in Preservation —generally valuable

Agility—drives up Preservation Costs

Immutability—drives up Access Costs

APPROACH

Strongly separate
"Preservation Access" systems
from "Use Access" systems

RISK MITIGATION STRATEGIES

Focus on "Data at Rest"

RISK MITIGATION STRATEGIES

Choose Native Filesystems as primary communication "channel" for Data

RISK MITIGATION STRATEGIES

Preservation through audit of de-correlated copies

RISK MITIGATION STRATEGIES

Buffer multiple transactions into streams

RISK MITIGATION STRATEGIES

Encapsulate and Summarize whenever possible

ARCHITECTURE STACK FOR CLIENT

- Application
  - Transaction interface
  - Automatic packaging
- Filesystem
  - Primary data channel
- Serialized form
  - SPOORE

SPOORE CONCEPT

(Self-describing Portable Objects Optimized for Reuse and Exchange)

- Self-describing bit-stream, so highly portable
- Offers flexibility in granularity, supports summarization

FILESYSTEM METAPHOR

- Logical Hierarchy with distributed ownership
- All references consist of Path and Timestamp
- All transactions are saved with Timestamp

STORAGE MODEL

- Based largely on LOCKSS and ACE
- Transaction-based (not file-based)
- Storage Level Services
- Storage Services
- Time-based Aggregates
  - Summarization Checksums

THANK YOU

Keith Johnson

Stanford University

keithj@stanford.edu

NOTABLE QUOTES

"Preservation is an outcome." –Clay Shirky

"It's all Triage." –Clay Shirky and Cliff Lynch

"Risk management is our compass"