OR 2010, The 5th International Conference on Open Repositories
Madrid, July 6-9, 2010

Curation Micro-Services
A Pipeline Metaphor for Repositories

Stephen Abrams *
Patricia Cruse *
John Kunze *
David Minor †

** UC Curation Center, California Digital Library*
† San Diego Supercomputer Center
University of California

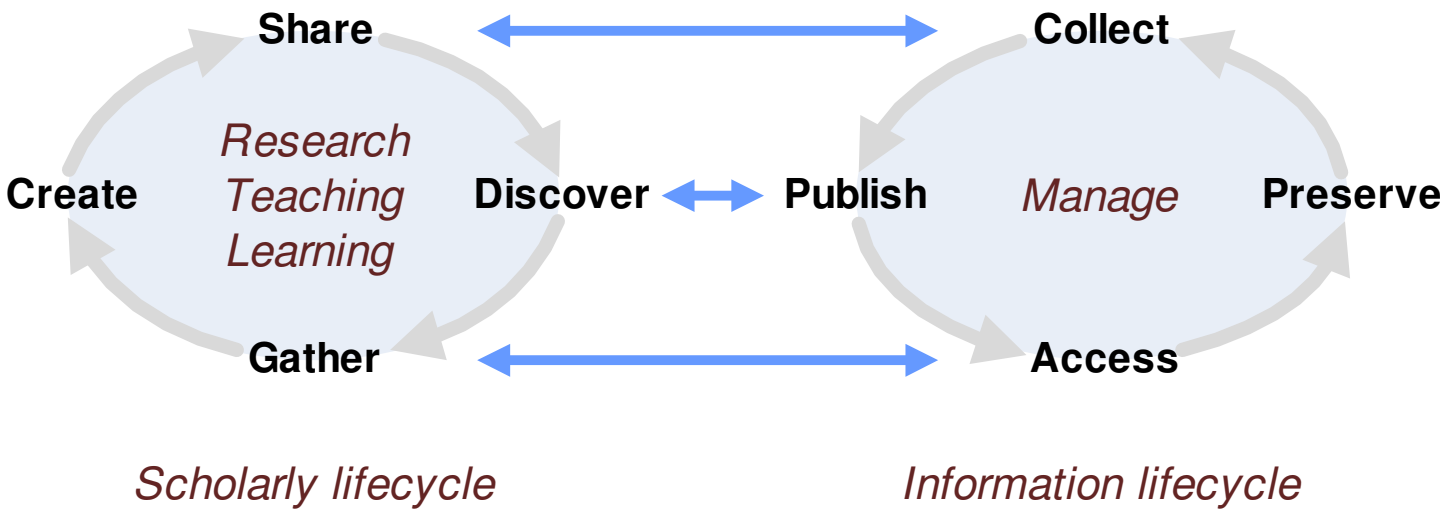
University of California Curation Center (UC3)

Creative partnership between the CDL, the 10 UC campuses, and peer institutions in the digital curation community

- An evolving community of shared concern and practice
- A means to pool and distribute diverse experience, expertise, and resources
- Robust solutions to counteract inevitable disruptive change

Digital curation

The set of policies and practices focused on *managing and adding value to a body of trusted digital content*, and facilitating the alignment of the scholarly and information lifecycles



UC Curation Center CDL California Digital Library SDSC SAN DIEGO SUPERCOMPUTER CENTER

Assumptions

Curated content gains

- Safety through redundancy *"Lots of copies keeps stuff safe"*
- Meaning through context *"Lots of description keeps stuff meaningful"*
- Utility through service *"Lots of services keeps stuff useful"*
- Value through use *"Lots of uses keeps stuff valuable"*

Curation is an outcome, not a place

- Focus on content, not the systems in which that content is managed

Curation stewardship is a relay

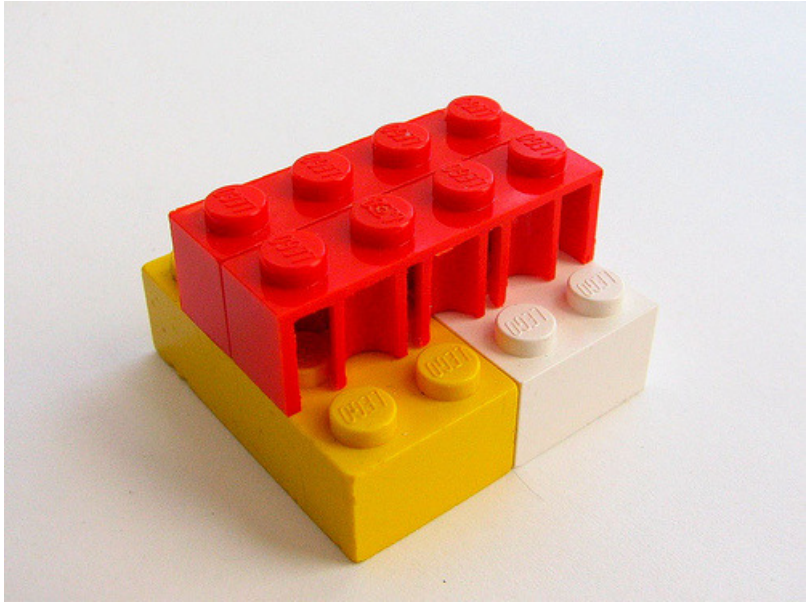
UC Curation Center
CDL California Digital Library
SDSC SAN DIEGO SUPERCOMPUTER CENTER

"A complex system that works is invariably found to have evolved from a simple system that worked"

– John Gall

"Entities must not be multiplied beyond necessity"

– William of Occam



Approach

Sound engineering principles (*c.f.* Occam, Gall, and Murphy) suggest

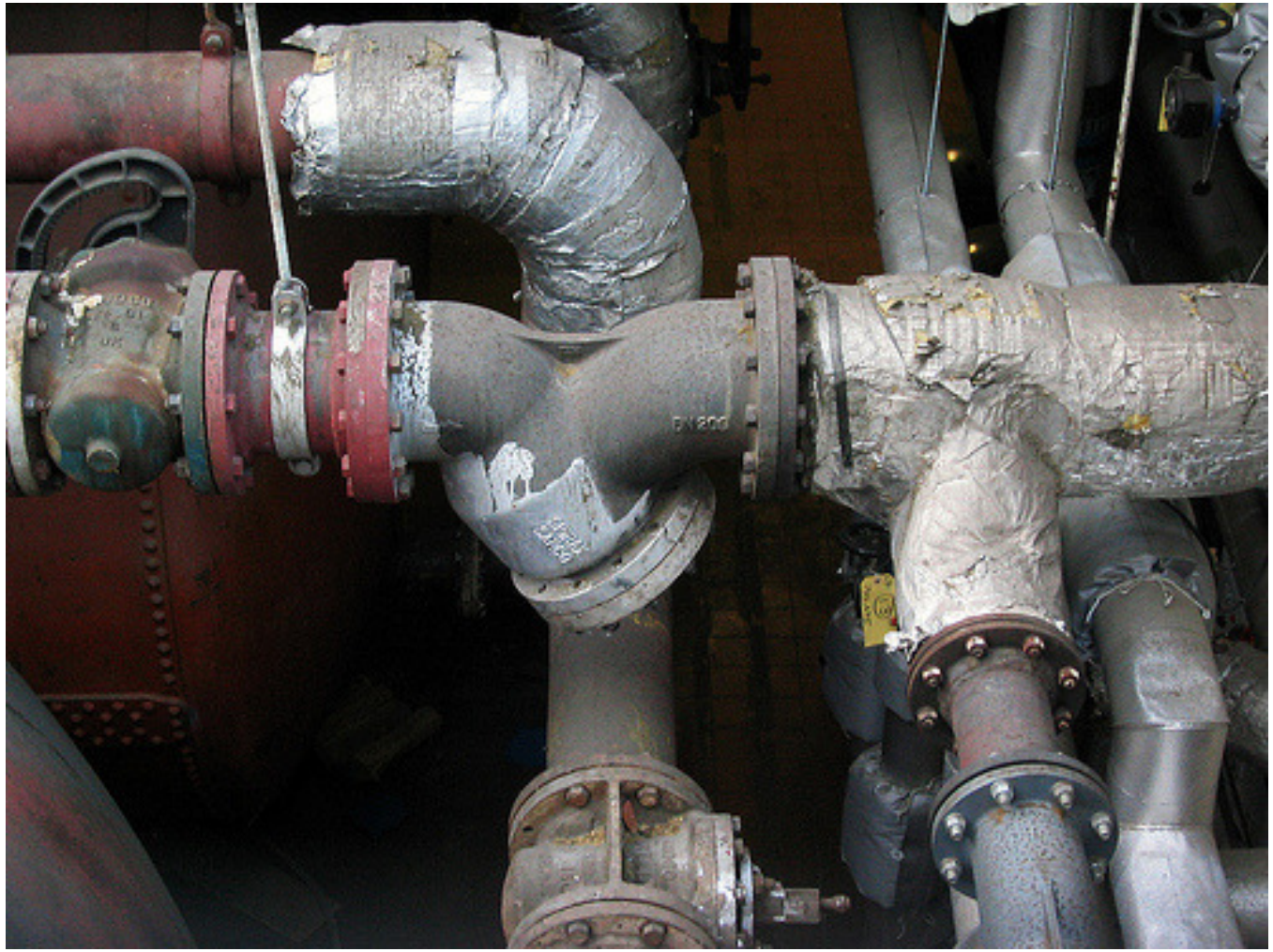
- Favor the small and simple over the large and complex
- Favor the minimally sufficient over the feature laden
- Favor the configurable over the prescribed
- Favor the proven over the (merely) novel

Build complexity through composition, not addition

Approach sufficiency through a series of incrementally necessary steps

UC Curation Center
CDL California Digital Library
SDSC SAN DIEGO SUPERCOMPUTER CENTER

The pipeline metaphor



The pipeline metaphor

The pipeline concept grew out of Doug McIlroy's articulation of the "Unix philosophy"

- "Make each program do one thing well"
- "Expect the output of every program to become the input of another, as yet unknown, program"
- "Design and build software … to be tried early"

McIlroy *et al*., "Unix time-sharing system forward," *Bell System Technical Journal* 57:6.2 (1978): 1902

UC Curation Center CDL California Digital Library SDSC SAN DIEGO SUPERCOMPUTER CENTER

Curation micro-services

Devolve curation function into a granular set of independent, but interoperable micro-services

- Since each is small and self-contained, they are collectively easier to develop, maintain, and deploy
- Since the level of investment in any given service is small, they are easier to replace when they have outlived their usefulness
- The scope of each service is limited, but complex behavior can *emerge* from the strategic composition of individual atomistic services
- All user/service (and service/service) interaction through public interfaces

UC Curation Center CDL California Digital Library SDSC SAN DIEGO SUPERCOMPUTER CENTER

Curation micro-services

	Value	**Annotation**	*of content by consumers*
		Notification	*of new content availability*
	Service	**Transformation**	*to create derivatives*
		Search	*of content and metadata*
		Index	*to enable fast search*
Curation		**Ingest**	*of content for curation*
Preservation	*Context*	**Characterization**	*to extract content properties*
		Inventory	*of curated content*
	State	**Replication**	*for safety*
		Fixity	*to verify bit-level integrity*
		Storage	*for long-term retention*
		Identity	*for long-term reference*

UC Curation Center CDL California Digital Library SDSC SAN DIEGO SUPERCOMPUTER CENTER

Design goals

Policy-neutral, protocol and platform independent

Linked data

The file system is the database

- All content and metadata fully expressed in the file system
- Some subset of metadata replicated in databases as an optimization for fast query

Multiple interface modalities

- RESTful HTTP
- Command line
- Procedural (with various language bindings)

Code to interfaces

Principle of least surprise

Appropriate benchmark for user experience is Flickr

Implementation strategies

Merritt project

–Consolidate management of 140 TB of existing content

–Support centrally-hosted as well as locally-deployed solutions

–Partner with campus data centers for virtualized and cloud provisioning

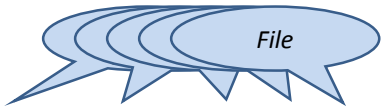
- *UC Berkeley, SDSC, TACC*

–Fine-grained application of modularity and orthogonality

–Open source work products (BSD)

–Exploit agile methods

–Stable URL references



http://example-store/state/default/1234/3/xyz

ANVL
ARK
Checkm
CAN
Dflat
EZID
Namaste
Pairtree
ReDD
http://www.cdlib.org/uc3/curation

4store
http://4store.org/

ERC/Dublin Kernel
http://dublincore.org/groups/kerne

ORE resource map
http://www.openarchives.org/ore

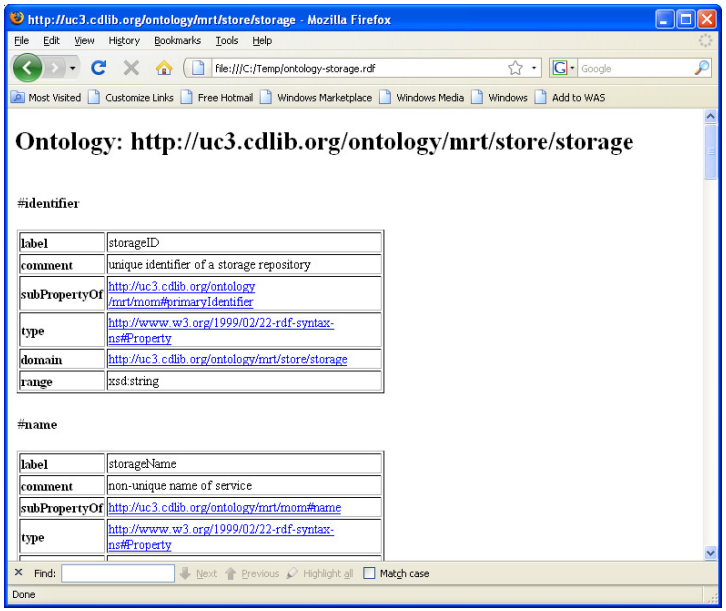
Zookeeper
http://hadoop.apache.org/zookeeper

UC Curation Center
CDL California Digital Library
SDSC SAN DIEGO SUPERCOMPUTER CENTER

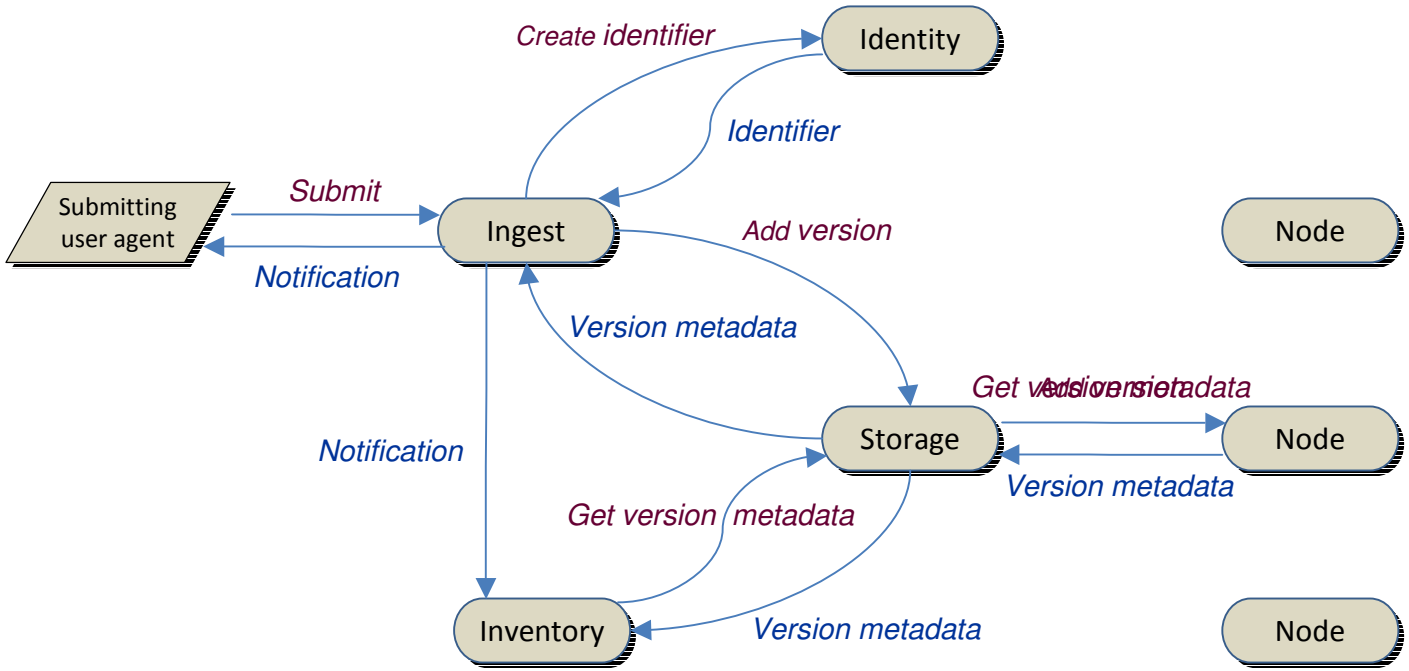
Demonstration

A few caveats … still a work in progress!

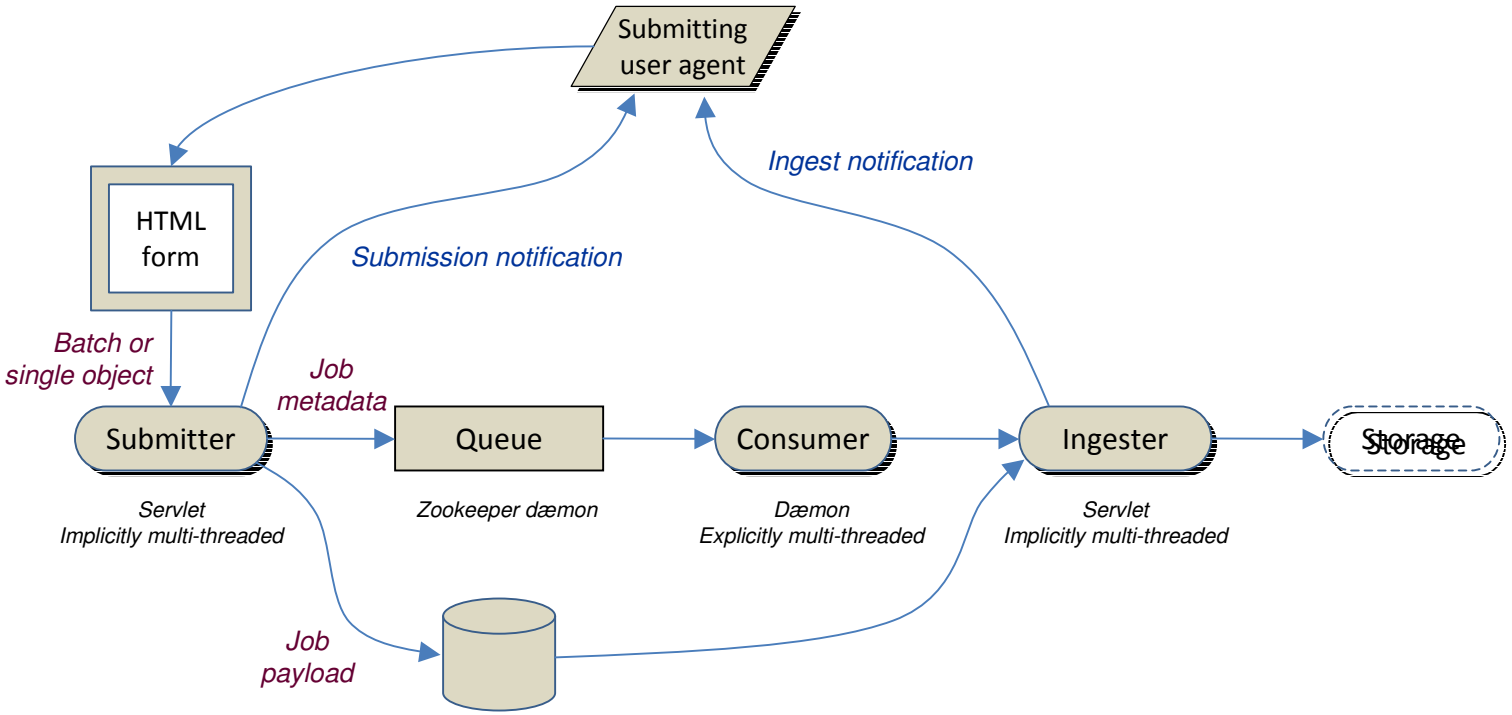
- The final interface style sheets are not yet applied
- Inventory service still under development
- More details available in a recent webinar
  https://confluence.ucop.edu/display/Curation/Home



Ingest process flow



Ingest implementation



UC Curation Center CDL California Digital Library SDSC SAN DIEGO SUPERCOMPUTER CENTER

Development roadmap

First wave	*Second wave* ✓	*Third wave*	*Fourth wave* ✓	*Fifth wave*	*Sixth wave* ✓
Identity	Inventory	Index	Search	Notification	Annotation
Storage	Ingest	Fixity	Replication	Characterization	Transformation
Object / collection modeling			Metadata standards		
Authentication / authorization			Semantic interoperability		
Policy / business model development					

UC Curation Center CDL California Digital Library SDSC SAN DIEGO SUPERCOMPUTER CENTER

Partnerships and collections

UC3 Digital Preservation Repository (DPR)
http://www.cdlib.org/uc3/dpr.html

- California Digital Newspaper Collection
- CDL eScholarship publishing
- Media Vault Program
- Minnesota Historical Society
- Open Context
- UCTV
- Water Resource Center Archive

and many others

UC3 Web Archiving Service (WAS)
http://was.cdlib.org/

DataCite
http://datacite.org/

DataONE
https://dataone.org/

Early community reaction

Collaborative development and integration projects with UC3 partners

Independent implementation of key Merritt specifications

–Georg-August-Universität Göttingen
–HathiTrust / University of Michigan
–Oxford University
– New York University
– University of North Texas

Digital curation group
http://groups.google.com/group/digital-curation

Curation micro-services Barcamp, Berkeley, August 16-17
http://groups.google.com/group/digital-curation/web/curation-technology-sig

–Possible follow-up events at iPRES (Vienna, September 19-24) and DLF (Palo Alto, November 1-3)

Curation BOF session, Wednesday, 15:45-16:45, Reino Unido A
http://or10.crowdvine.com/pages/bof

Summary

The pipeline concept provides a useful metaphor for curation micro-services embodying the following principles

- Modularity
- Orthogonality
- Parsimony
- Evolution
- Emergence

Key activity pattern: *define, decompose, recurse*

Build complexity by composition, not addition

Significantly lower levels of development effort, and higher quality solutions

For more information

Curation micro-services BOF, Wednesday, 15:45-16:45, Reino Unido A
http://or10.crowdvine.com/pages/bof

UC Curation Center (UC3)
http://www.cdlib.org/uc3

Micro-service specifications
https://confluence.ucop.edu/display/Curation

Digital curation group and Barcamp
http://groups.google.com/group/digital-curation
http://groups.google.com/group/digital-curation/web/curation-technology-sig

UC3 / CDL

Stephen Abrams Erik Hetzner Margaret Low Mark Reyes Perry Willett
Patricia Cruse Greg Janée David Loy Tracy Seneca
Scott Fisher John Kunze Isaac Rabinovitch Marisa Strong

UC3 / SDSC