An emergent approach to data curation
John Kunze, Stephen Abrams, Trisha Cruse, Perry Willett (CDL/UC3)

Complete curation emerging from small independent, micro-services
Create low barrier, low commitment tools instead of complex monolithic systems
Compose repository function from small, independent, interoperable micro-services – complexity by composition

A closer look at Storage
In Storage, most things are files.

The micro-services roster
Interoperation with value
Annotation of content by consumers of new content availability
Notification to create derivatives of content and metadata
Search to enable fast search
Ingest of content for curation
Application through service
Characterization to extract content properties of ingested content objects
Inventory for long-term retention
Protection of state
Replication for content for safety
Fixity to verify bit-level integrity
Storage for long-term retention
Identity for long-term reference

The wisdom of files
After 30 years, we’re good at modern filesystems
• files and directories (folders) are fast, plentiful, stable, and highly interoperable across platforms
• native OS tools will create, list, change, and backup
File-based micro-services will be easier…
• to develop, maintain, and to escape from
• to recombine in flexible ways
• to move upstream into use by content producers

What’s the thinnest smear we can add to our well-understood filesystems to create an object storage system?

Early success stories
Pairtree, written with help from the University of Michigan, underpins the HathiTrust holding all the Google books. Pairtree has been independently implemented in Perl, Java, and Python at a number of institutions.

BagIt, co-written with the Library of Congress (LC), describes a file package (“bag”) suitable for disk-based or fast network-based transfer of generic content. LC now receives most of its grant-funded partner content in bags.

Micro-services applied to scientific data curation
DataNet cyberinfrastructure
With leading roles in management, sustainability, and preservation for an NSF-funded DataNet project, UC3’s Ingest and Inventory services will receive and re-expose water archive, herbaria, and museum data, allowing it to become one of the first four DataONE member nodes.

CDL/UC3 joins DataCITE
The Identity micro-service supports scheme-agnostic identifier minting, binding, and resolving, core functions in encouraging data publication, citation, and archiving.

THUMP-DL data integration
THUMP “Data Lab” uses micro-service elements and REST-like design for very simple integration.