Reining in Research Data

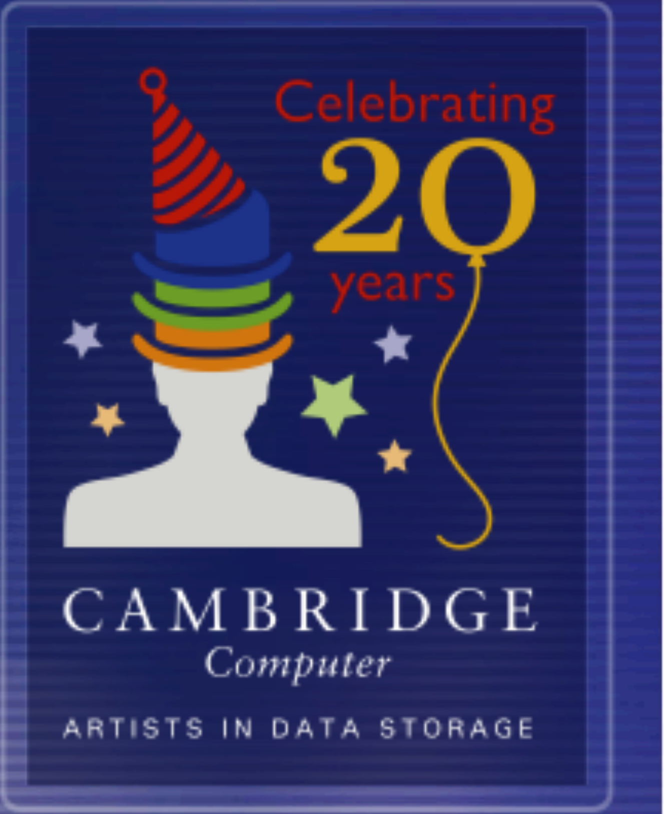
Adding Structure to Unstructured Data

Jacob Farmer, CTO
Cambridge Computer

David Bernick
Director of Engineering
Cambridge Computer

A Little Background On Cambridge Computer

- Founded in 1991 as a boutique integrator for backup and archive solutions
- Approximately 75 employees nationwide
- Clients of all shapes and sizes across all industries
  - Particularly strong in research and higher ed
- Industry-wide reputation for:
  - defining best practices for enterprise class data protection, and
  - for the early adoption of next generation storage solutions
- A unique business model that allows us to straddle the fence between academia and industry



Our Project – Defining Best Practices for File Management

- Inspiration for our project comes from SRB/IRODS
  - Bring parts of the SRB/IRODS vision to reality
    - Define a general purpose feature set
    - Intuitive user interface
    - Simplified API
- Inspiration also comes from numerous home grown solutions in our client base.
- What we have built
  - A prototype application that our clients can use to understand how they can take advantage of smarter file storage management.
  - A way of illustrating concepts to people who otherwise don't get it and don't want to invest the mental energy to figure out what we have all been talking about!
  - A file system reporting tool that can interact with the users themselves

Long Term High Level Goals

- Provide superior storage management housekeeping throughout the data life cycle
  - Backups, data integrity verification, storage tiering, disposition
- Facilitate the hand off between live research and preservation
  - Prime the pump with metadata that has been gathered throughout data life cycle
  - Provide a persistent file address for digital asset management systems, regardless of where the files are physically stored.
- Provide the framework for chargebacks and cost accounting.
- Enable scientist to better organize, share, and collaborate
  - Provide the framework for fulfilling data management plans

The Meaning of Words

- The word “archive” means different things to different people
  - Records management / retention
  - Immutability (WORM)
  - Migration of data to offline or near-line media
    - General belief is that files are now out of reach
  - Digital preservation
- The storage industry favors the term “*life cycle management*” which can mean all of these things.
  - This, of course, creates a different kind of confusion!

Challenges Specific To Managing Research Data

- There is a lot of data!
  - Hard to move, expensive to store, etc.
- The researchers work the way the want to work
  - Hard to get them to change behaviors without big carrots and big sticks
  - No matter what you try, some (many) will not comply
- Performance / latency concerns
  - We cannot introduce a performance hit
- Data formats become obsolete quickly
- Uncertainty over what to keep versus what to throw away
- Data accessibility v. privacy
  - The granting agencies want data to be shared
  - Researchers are afraid of getting “scooped”
  - HIPAA concerns, human subject data, etc.

Metadata – Competing Interests

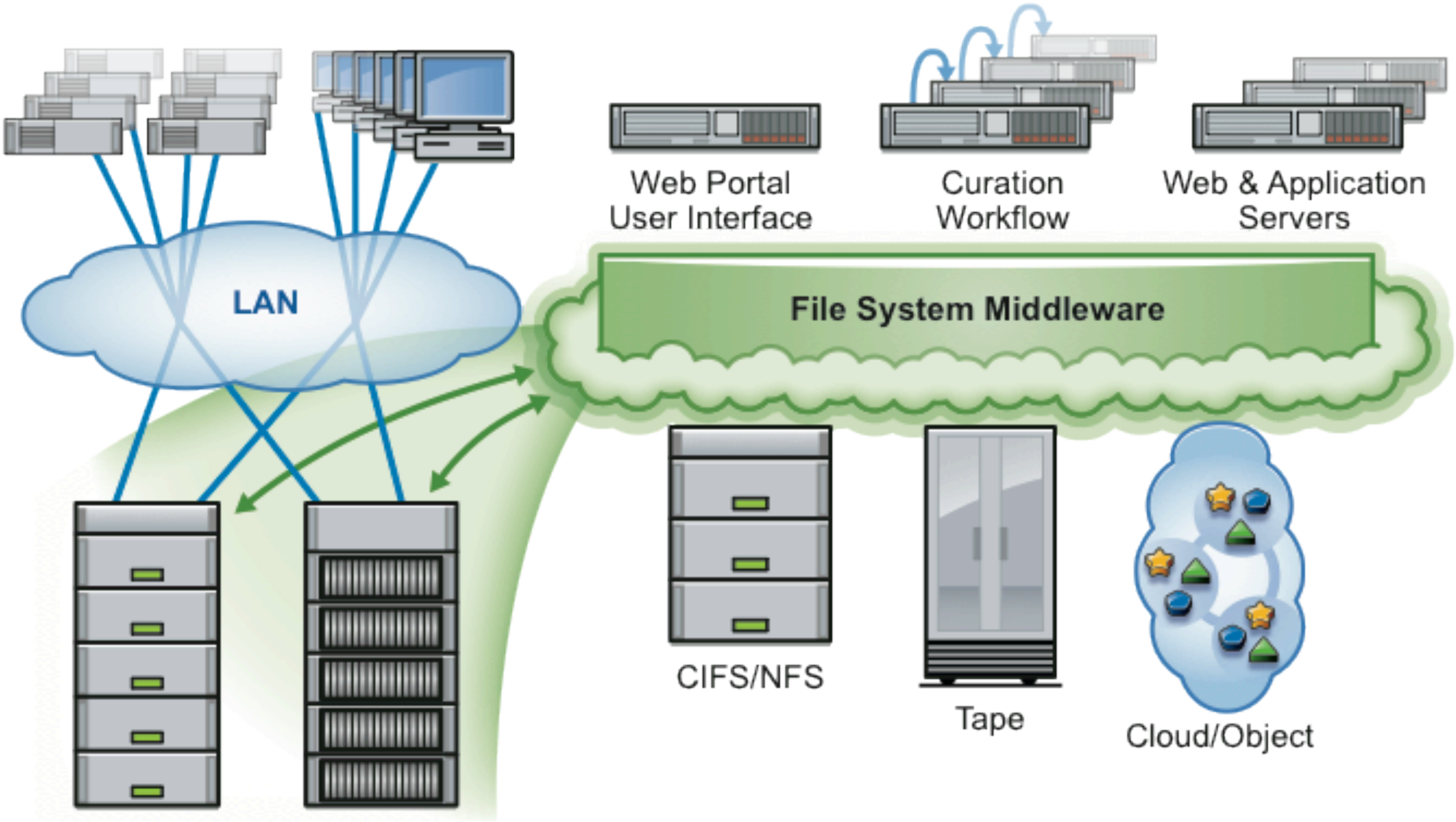
- Lots of people want metadata, but no one really wants to type it in!
  - Preservation systems will have to co-exist with the research group's own wants and needs
    - Research groups have their own idea of what metadata they want and how they want to use it
    - Storage administrators need certain metadata
    - Project managers and grant administrators need certain metadata
- Archival systems need to be interactive with live data
  - Preservation is a component of complete storage management.
  - Provenance – The metadata must have a way of keeping track of how archived data is used in future research.

The Real Trick – Getting the Metadata

- The Golden Rule of Data Preservation – "Preserve at the time of creation"
  - Translation: Capture metadata throughout the research pipeline
- Perhaps capture metadata when storage is provisioned
  - The presumes that there is a structured process for provisioning storage
- Capture metadata through an API
  - This requires a simple API that anyone can use
- Programmatically extract metadata from file headers, tags, and content
- Capture metadata through a GUI
  - Try to create incentives for users to key in metadata

The Mwah Hah Hah Plan to Conquer the World

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE


Web Portal User Interface
Curation Workflow
Web & Application Servers
LAN
File System Middleware
CIFS/NFS
Tape
Cloud/Object