Project launched in 2011 as a prototyping exercise
- Inspired by IRODS
- Built on top of IRODS

V. 1 put into production in mid 2013
- Removal of all IRODS code

V. 2 released as production software in late 2014 with limited client base.

Early 2014 - Spun out of Cambridge Computer into its own company.
- Still in “stealth mode” but making a lot of noise anyway
- Currently in beta for v.3
Differences from IRODS

- Built for very large scale
  - Billions of files
  - Petabytes

- Flexible job scheduler rather than rules engine
  - If you need to do anything fancy, write it in Python, or Java or whatever, or we will write it for you.

- Metadata catalog is decoupled from the application
  - Change database technologies over time

- Closer to storage management than collection management
  - We get intimate with the storage infrastructure

- Currently Starfish does not federate user directories across institutions
  - We will, however, integrate with In Common . . . soon
The Big Vision

Library
- Open Links
- Curation Workflow
- Access Controls
- Fixity Checks
- Repository Retention

Research
- Annotation
- Collaboration
- Workflow
- Data Management Plans
- Open Access
- Institutional Retention

IT Operations
- Storage Management
  - Backup/Restore
  - Tiered Storage
  - Data Migration
  - Cloud Interface
- Reporting
  - Chargeback / Showback
  - General Reporting
  - IT Governance
The OOO Model For Data Life Cycle Management

Organization
- Namespace
- Metadata
- Content Index

Orchestration
- Policy Engine
- User-driven
- API

Operation
- Copy - Move - Delete
- Other Data Processes

Containers
- File Systems
- Object Stores
- Tape Archives
- Cloud Storage
Typical Content Management “Stack”

- **Application Metadata** (Database)
- **File System Metadata**
  - File Attributes and Permissions (UNC Path Address)
- **Storage Devices**
  - (Sequential Block Address)
- **Content Management**
- **File System**
- **Storage**
Data Management “Stack”

- Content Management Applications
- Tribal Knowledge
- File System
- File System(s)
- Storage Tier 1
- Storage Tier 2
- Storage Tier 3
Data Management Stack with Middleware

File System Middleware
Global File System & Life Cycle Management
(UNC Path and/or Object Address)

Policy Engine
Data Mover
Metadata

File System

File System(s)

Storage Tier 1

Storage Tier 2

Storage Tier 3
All in One: Backups, Archiving, Curation, Storage Management
A “Virtual” Global File System

Users

Compute Nodes

Data Acquisition

Primary NAS

Data Mover(s)

Low-cost NAS

Tape Archive

Object Store or Cloud Service

Unified Portal / Gateway for Primary, Secondary, and Tertiary Storage