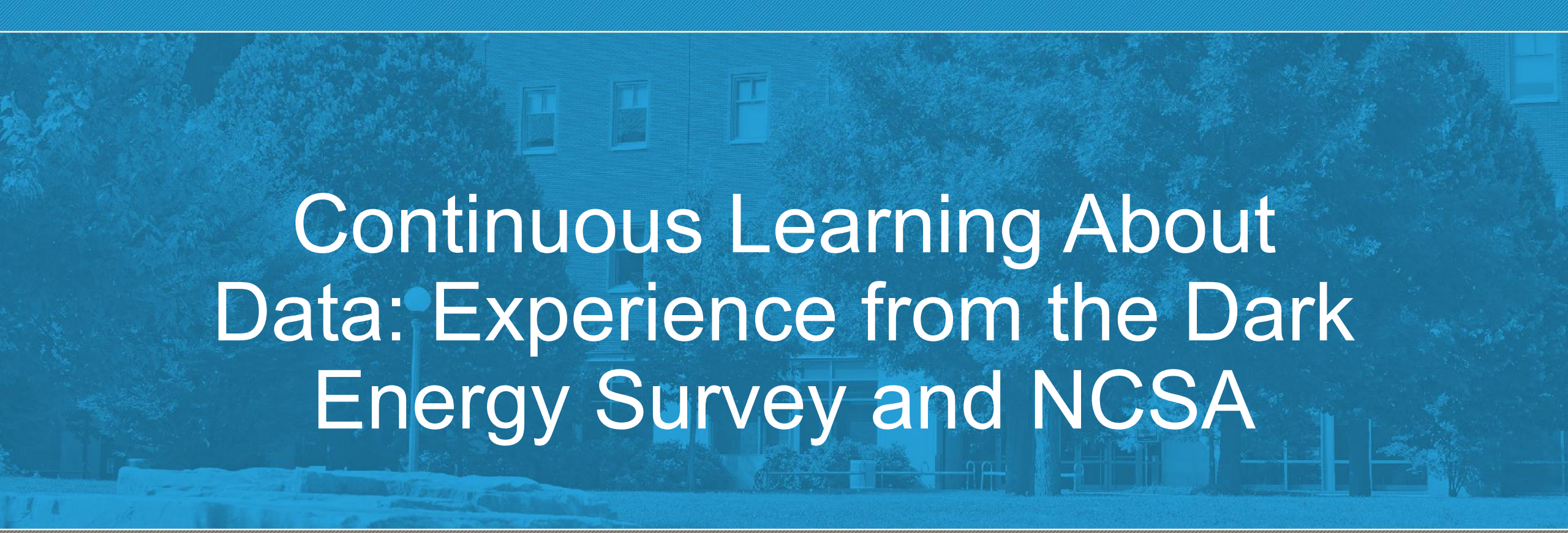


Continuous Learning About Data: Experience from the Dark Energy Survey and NCSA

Margaret Johnson and Don Petravick

MAGIC Monthly Meeting, April 3, 2019

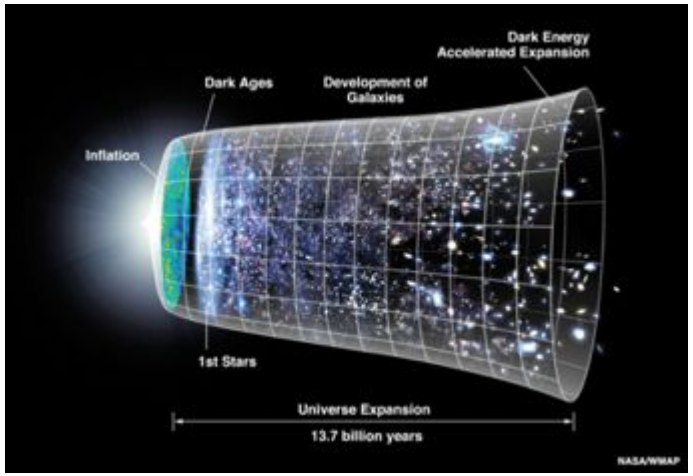
ILLINOIS
NCSA | National Center for Supercomputing Applications

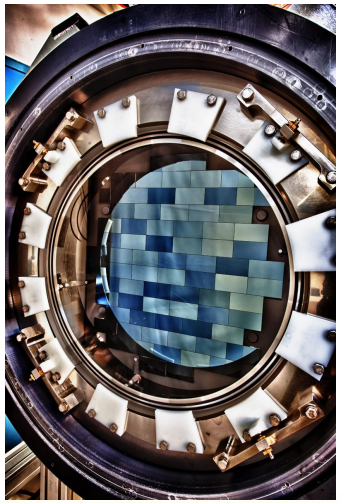
Outline

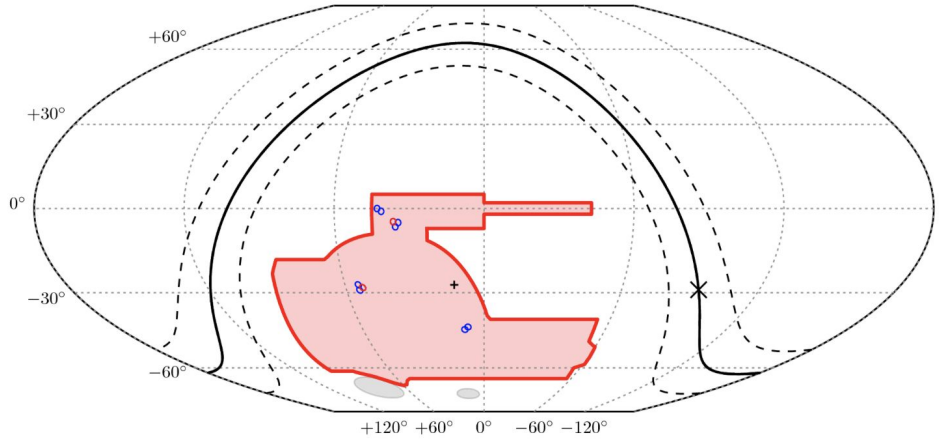
- The Dark Energy Survey
- DES Data Management
- DESDM provenance implementation
- Sustainable CI and the Clowder data management framework

The Dark Energy Survey (DES)

The Dark Energy Survey (DES) is a Stage III Dark Energy Task Force observational program, “designed to probe the origin of the accelerating universe and help uncover the nature of dark energy by measuring the 14-billion-year history of cosmic expansion with high precision.”

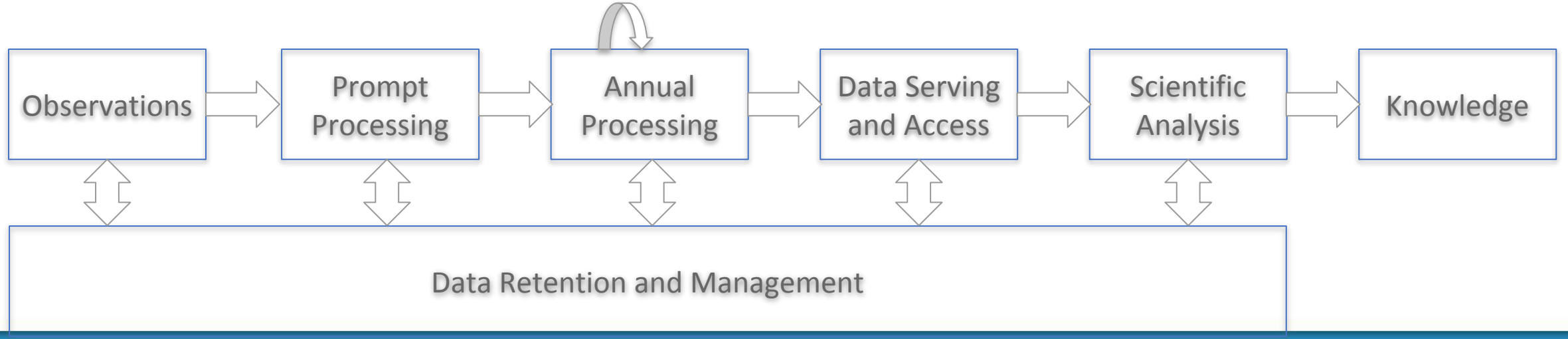






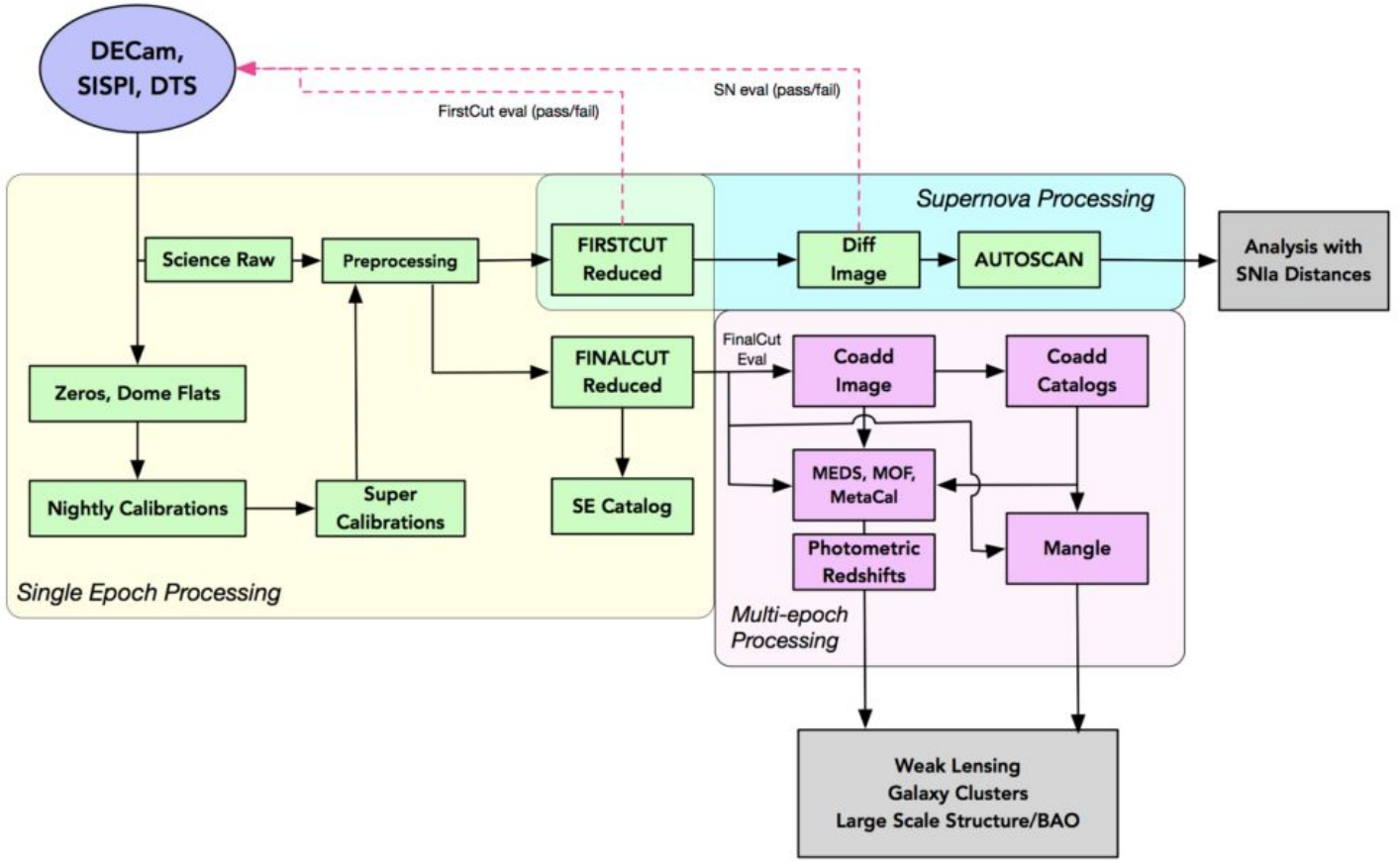
DES: a statistical sky survey

- Need large, high-quality, uniform datasets for statistical analysis of objects, with systematics well-characterized.
- DES is a 5.5-year survey covering 5000 deg^2 (~⅛) of southern sky
- Each section of the footprint observed multiple times in each filter; certain fields revisited every 10 nights for supernova search.
- Data processing and archiving is managed by NCSA.
- Data products are released to the collaboration for analysis.



DESDM: Data Processing

- Collect over 18,000 images per night = 1 TB
- Data are processed in two cadences:
  - Promptly (within 24 hours) for supernova follow-up and feedback to observing
  - Annually, reprocessing all data uniformly for best calibration and deep catalogs



Nightly Cadence

- Nightly production included:
  - SNe Difference Imaging (SNDIFF)
    - SNe candidate detection
    - Up-or-down assessment of SNE exposure set
  - FIRSTCUT processing
    - Up-or-down quality assessment of wide-field observations
    - SNE images are re-processed here to obtain first-look catalogs and quality assessments

Annual Cadence

- Annual release processing includes:
  - Review of calibrations.
  - Filtering the collection of raw single epoch images based on their FIRSTCUT quality assessment. One criterion is inclusion for co-addition.
  - Production of “super” bias and flat-field frames
  - FINALCUT reduction of single-epoch images and catalogs
  - Global Photometric calibration
  - Production of co-added images and catalogs
  - Production of mangle masks
  - Curating the final data products before release
- Summer activities for SNE included:
  - Re-processing of previous years’ data as algorithmic improvements warrant

Ad-Hoc Campaigns

- Examples of ad-hoc campaigns include:
  - Parameter sweeps
  - Studies
  - Rare calibrations
  - One-off data products needed to understand standard data products
- Ad-hoc campaigns are done in the same framework and therefore with the same rigour

DES Data Management System - Provenance and Metadata Aspects

- DESDM generates many data products on many cadences.
- Processing is supported by a unified framework that is used for all campaigns.
- The framework is able to run community codes, while maintaining
  - a single metadata system
  - a single provenance framework
- This allows
  - Seamless specification of the pipeline inputs and outputs
  - Uniform provenance traces of data from raw to the highest level products

How DESDM Developed - Original Concept

- All files should be accompanied by rich metadata produced at the time data was produced.
- Collect provenance; however, representation varied some in the various schemas in a database, and some in files.
- While the information was present, it was not usable.
- Moreover, files were mutable.
- There was no mechanism to add characterization as more was learned about the data.

Revised Approach to Provenance

- The DESDM provenance model is now based on aspects of the Open Provenance Model (OPM).
- OPM provides a vocabulary and an ontology.
- Information elements are artifacts (immutable data) and processes (relating the immutable data)
- Uses tuples that describe parent-child relationships based on artifacts and processes.
  - WasGeneratedBy: an artifact and the process that generated it. (1:1)
  - Used: a process and all artifacts used by that process to obtain the same outcome. (1:many)
  - WasDerivedFrom: artifact to artifact relationship (1:many)

Implementation of OPM

- The DESDM workflow framework:
  - Records provenance in a file as it executes the sequence of programs in a pipeline.
  - Ingests the provenance tuples into the Oracle operational database.
- Tuples are queried using a recursive relational query.
- No energy was available to:
  - Fully understand tuple-based information framework.
  - Select or implement database-optimized for tuple-based queries, such as SPARQL.

Example QA process supported by the provenance system

- An image is subtracted from another reduced image of the same field. The resulting difference image shows a checkerboard pattern in the sky residual. Where did this pattern come from?
- Need:
  - What files and processing steps were used to generate the two reduced images.
  - What pipeline was run to create these images, and which steps were (not) run.
  - Which calibrations were applied, and where they are located.
  - Which version of a pipeline step was run and which parameter sets were applied.
  - What command line arguments were used for various steps/override of the default parameter files.

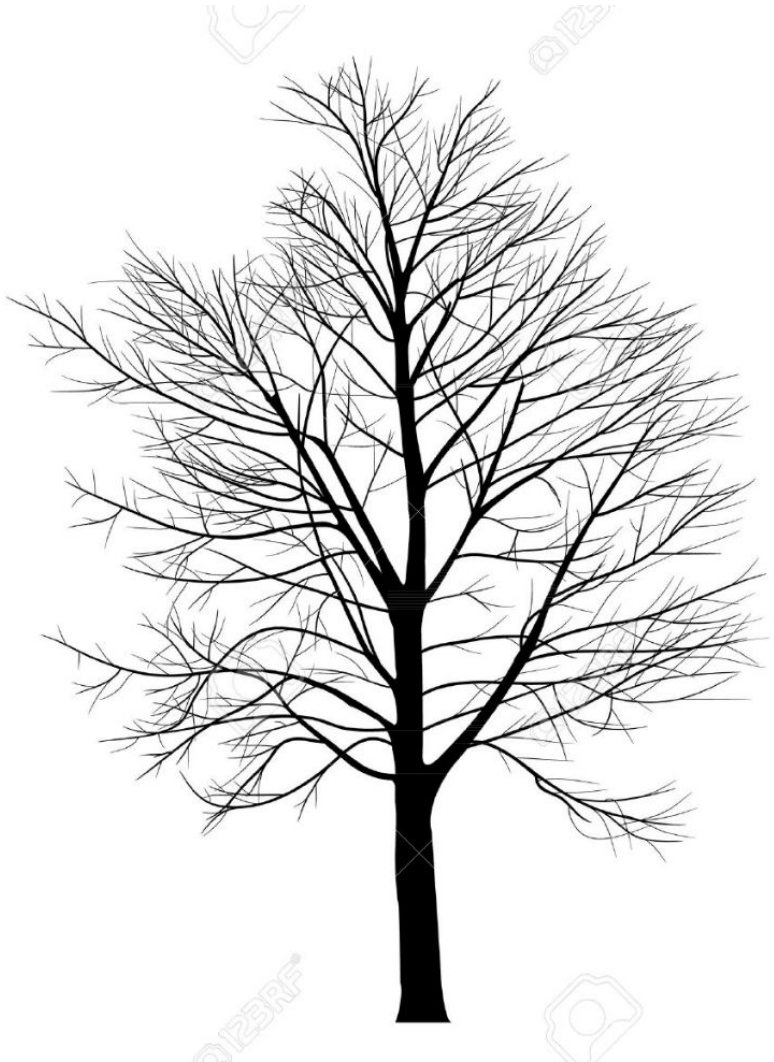
Need for additional information about data

- As data are processed, reviewed, and re-processed and re-reviewed, more is learned about the quality of the data.
- Need to record and recall additional information about logical collections of data elements or processes which have some kind of uniform quality. Examples:
  - All processing runs that are part of an annual release
  - A series of tests meant to confirm parameter for final processing
  - Blacklisting CCDs from input to a processing campaign
- Also need to build and recall historical records about prior successful and failed processing attempts and uses of an exposure.

Tags implementation

- A construct named “Tags” was invented and implemented in the operational database.
- Tags have:
  - Name
  - Purpose
  - Creator
- Tags can point to an arbitrary set of files.
- All files pointed to have the property specified by the tag.

Effect on processing campaigns

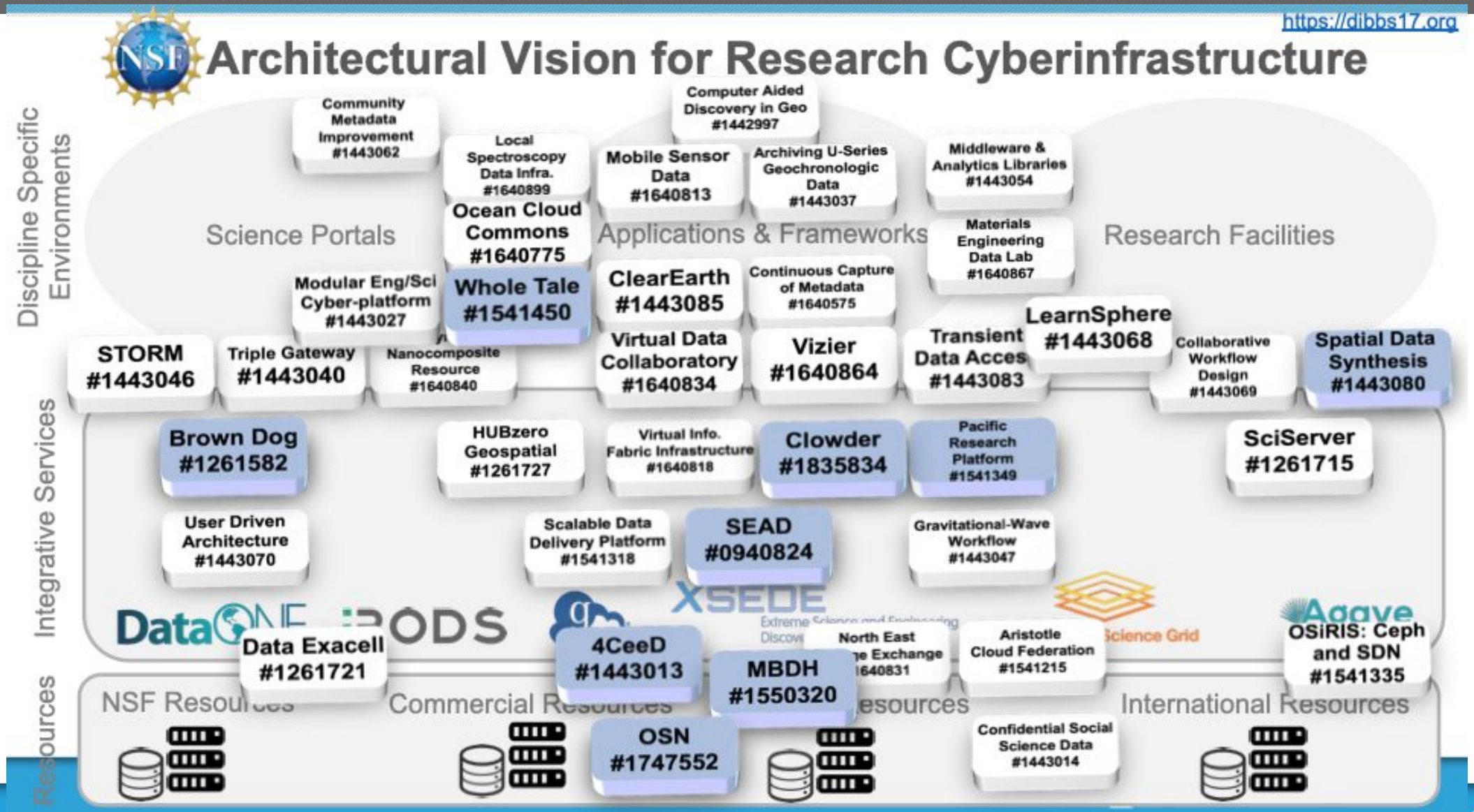


- Like a tree, data products become more diverse and tailored for specific science use. In DESDM, more specific products are seen as ad-hoc campaigns.
- In our view,
  - One measure of the success of a metadata and provenance system is that overall knowledge gained about the quality and packaging of data products is accessible to pipelines making increasingly tailored data products.
  - A principle is that at every fork in the tree relevant data be preserved in a way that is queryable and accessible.
  - For large systems, separating this data into its own data systems leads to success.
- Contrasting approaches include
  - Effectively embedding the information implicitly in run-time decisions of code
  - Materializing the data forests of file systems links
- In our view, these approaches are difficult to scale as higher level products are identified and produced, and important data is hidden.
- We consider it a success that DES has avoided these pitfalls.

DESDM result

- The DESDM provenance system was implemented as a result of experience and learning this set of primitives statisfies DESDM's needs.
- Suitable for a production system
  - that is operated by a small group.
  - with moderate requirements -- for example there is no provenance based automatic regeneration of data.

NSF DIBBs


https://dibbs17.org
Architectural Vision for Research Cyberinfrastructure
Discipline Specific Environments
Integrative Services
Resources
Science Portals
Applications & Frameworks
Research Facilities
Community Metadata Improvement #1443062
Computer Aided Discovery in Geo #1442997
Local Spectroscopy Data Infra. #1640899
Mobile Sensor Data #1640813
Archiving U-Series Geochronologic Data #1443037
Middleware & Analytics Libraries #1443054
Ocean Cloud Commons #1640775
Materials Engineering Data Lab #1640867
Modular Eng/Sci Cyber-platform #1443027
Whole Tale #1541450
ClearEarth #1443085
Continuous Capture of Metadata #1640575
LearnSphere #1443068
STORM #1443046
Triple Gateway #1443040
Nanocomposite Resource #1640840
Virtual Data Collaboratory #1640834
Vizier #1640864
Transient Data Acces #1443083
Collaborative Workflow Design #1443069
Spatial Data Synthesis #1443080
Brown Dog #1261582
HUBzero Geospatial #1261727
Virtual Info. Fabric Infrastructure #1640818
Clowder #1835834
Pacific Research Platform #1541349
SciServer #1261715
User Driven Architecture #1443070
Scalable Data Delivery Platform #1541318
SEAD #0940824
Gravitational-Wave Workflow #1443047
DataONE
PODS
XSEDE
Science Grid
Agave
Data Exacell #1261721
4CeeD #1443013
MBDH #1550320
North East Exchange
Aristotle Cloud Federation #1541215
OSiRIS: Ceph and SDN #1541335
NSF Resources
Commercial Resources
International Resources
OSN #1747552
Confidential Social Science Data #1443014

Clowder

- Clowder as a focused component in the NSF DIBBs program is integrated into ~100 different, mostly smaller systems, many of which support analysis as well the kind of production DESDM performs for DES.
- The Clowder metadata/provenance system is more fully featured and broader in applicability.
- https://clowder.ncsa.illinois.edu/

Supporting Scientific Research Data

- **Sharing**
  - Familiar user friendly web based “Dropbox”-like interface
  - Open source / deployable on available hardware
  - Customizable appearance/capabilities for community specific needs
  - Create separate “Spaces”, control who has access to what
- **Curation**
  - Social curation - tag/comment with members of the community, follow activities of others
  - Auto-curation - deploy “extractors” that examine the contents of files as they are uploaded, use machine learning and other tools to auto-“tag”
  - Custom previews - customizable lightweight visualizations to visually inspect data or collections of data
- **Publication**
  - Once at publication stage press button to find a domain relevant institutional archive, publish data, and generate citable DOI
- **Reuse**
  - Through “extractors” allow custom code to be run next to the data, hide complexities of execution tools and dependencies on HPC/Cloud resources

Active Curation

- **Active curation (AC) involves recording data and metadata as close to the source as practical and driving that acquisition through the deployment of capabilities that help data producers manage their research.**

- **Social Curation (SC) drives this economic analysis further, looking at ways that crossgroup interactions can further motivate best practices.**

J. Myers and M. Hedstrom, "Active and Social Curation: Keys to Data Service Sustainability," NDS Consortium Planning Workshop, 2014
http://sead-data.net/sites/default/files/pubs/ActiveandSocialCurationKeystoDataServiceSustainability.pdf

Clowder “Tags”

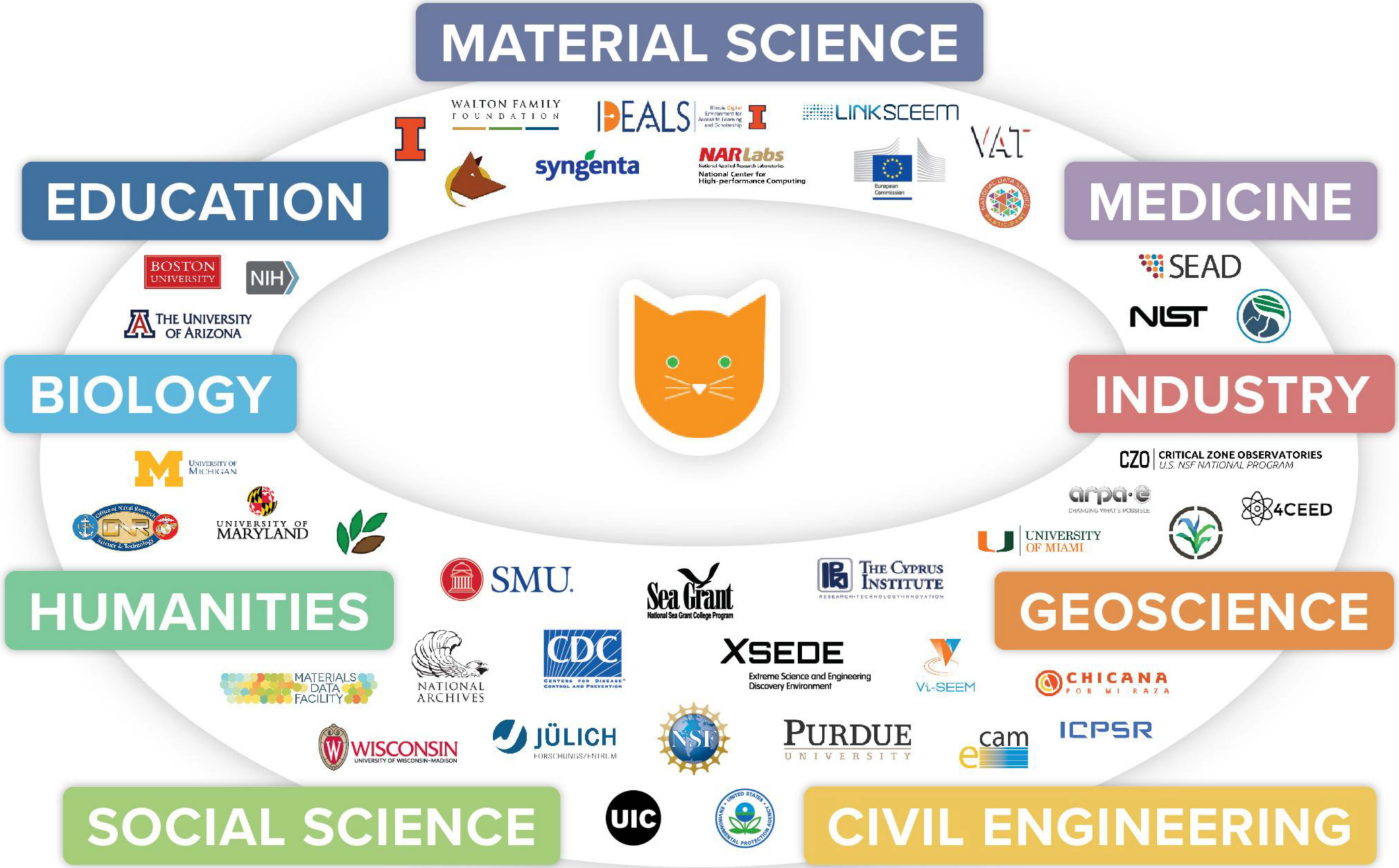
- Clowder allows tags, which are like Web 2.0 tags.
  - In Clowder these are just user defined labels.
  - Can be used without being mindful of subject-verb-predicate concepts.
  - There is no ontology relating labels to each other.
  - Consistency of labelling is up to users of a Clowder instance.
- This is similar to the DES implementation. In DES, when used for production, the ontology is understood among the small group of tag makers and users.

Clowder supports use of ontologies

- Support for an RDF-based ontology is part of the Clowder stack.
- Similar to DES, ontology tuples are generated as files are produced.
- The stack supports the use of ontology triples to record knowledge about the data files, by adding more predicates.
- RDF tools can be used to process “birth” metadata, provenance and metadata added by production processing due to the more uniform data representation based on RDF concepts.

Clowder supports gradual formalization of a provenance model

- The support for both tags and RDF in the Clowder stack supports a process for gradually formalizing how data is characterized.
- Sometimes standards provide a usable ontology.
  - DESDM was able to adapt the concept of tuples and the concepts of the Open Provenance Model in a straightforward way into its operational relational database.
- Clowder includes full support for an RDF-based ontology, and its tags implementation can be viewed as a gateway to walk into RDF for the specifics of an experiment.
  - Scientists work out the concepts with tags, and can promote mature concepts into a more formal RDF framework.


MATERIAL SCIENCE
EDUCATION
MEDICINE
BIOLOGY
INDUSTRY
HUMANITIES
GEOSCIENCE
SOCIAL SCIENCE
CIVIL ENGINEERING
WALTON FAMILY FOUNDATION
IDEALS
LINKSCEEM
VAT
syngenta
NARLabs
National Center for High-performance Computing
European Commission
BOSTON UNIVERSITY
NIH
THE UNIVERSITY OF ARIZONA
SEAD
NIST
CZO CRITICAL ZONE OBSERVATORIES U.S. NSF NATIONAL PROGRAM
arpa·e
4CEED
UNIVERSITY OF MICHIGAN
UNIVERSITY OF MARYLAND
UNIVERSITY OF MIAMI
SMU.
Sea Grant
National Sea Grant College Program
THE CYPRUS INSTITUTE
NATIONAL ARCHIVES
CDC
XSEDE
Extreme Science and Engineering Discovery Environment
Vi-SEEM
CHICANA POR MI RAZA
MATERIALS DATA FACILITY
WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON
JÜLICH
FORSCHUNGSZENTRUM
NSF
PURDUE UNIVERSITY
cam
ICPSR
UIC

Clowder Usage

		Biology	Civil Engineering	Comp. and Inf. Science	Education	Geoscience	Humanities	Industry	Materials Science	Medicine	Social Science
Brown Dog	NSF	X	X	X		X					X
BRACELET	NSF			X					X		
Chicanapormiraza	Univeristy of Michigan						X				
Countermeasures against Radiation	BARDA	X								X	
Crowd-Sourced Green Infrastructure	NSF		X								X
DataNet SEAD	NSF, NDS	X	X			X					
EarthCube GeoSemantics	NSF					X					
eCam	European Commission				X						
IML-CZO	NSF	X			X	X					
Immunomodulatory Effects of MSC	NIH									X	
Great Lakes Monitoring	Illinois-Indiana SeaGrant					X					
Great Lakes to Gulf	NGRECC, Walton Foundation					X					
Groupscope	NSF										X
KISTI	KISTI	X		X							
LinkSCEEM	European Commission					X	X				
NARA	NSF			X							
NIST	NIST								X		
PEcAn	NSF	X			X	X					
RIVEEL3D	Cyprus Institute						X				
SIMDAS	European Commission						X				
Southern Methodist University	SMU		X	X							
Syngenta	Syngenta	X						X			
T2-C2	NSF			X					X		
Taiwan NCHC	NCHC			X						X	
TERRA-REF	ARPA-E	X			X	X					
TRECC	ONR				X						
Vector-Borne Disease	CDC	X									
Vi-SEEM	European Commission						X				
XSEDE Decomposing Bodies	NSF			X			X				
XSEDE Image Analysis of Rural Photography	NSF			X			X				
XSEDE Large Scale Video Analytics	NSF			X			X				
XSEDE Real Stories of Bad Kids	NSF			X			X				X

Summary

- We've compared an experiment-focused effort and a reusable building block.
- For its era, DES managed to learn of some relevant standards, and incorporate elements of an ontology into a mid-life production system supported by a few developers focusing on the overall experiment and processing system.
- Clowder is a result of a long program of work, and is an element of the NSF DIBBs program. It has a broad and diverse use community, and more fully integrates standards and concepts of information science, including a set of features learning about data characteristics culminating in ontologies.

ILLINOIS

NCSA | National Center for
Supercomputing Applications

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov