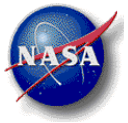


The Virtual Climate Data Server (vCDS): An iRODS-Based Data Management Software Appliance Supporting Climate Data Services and Virtualization-as-a-Service in the NASA Center for Climate Simulation

John L. Schnase[1], Glenn S. Tamkin[2,3], W. David Ripley, III[2,3], Savannah Strong[2,3], Roger Gill[2,4], and Daniel Q. Duffy[2,1]

Office of Computational and Information Science and Technology[2]
NASA Center for Climate Simulation (NCCS)[3]
Computer Science Corporation (CSC),
[4] Innovim, LLC
NASA Goddard Space Flight Center
Greenbelt, MD 20771

The Data Management System Project

Part 1 – Background (5 Minutes)

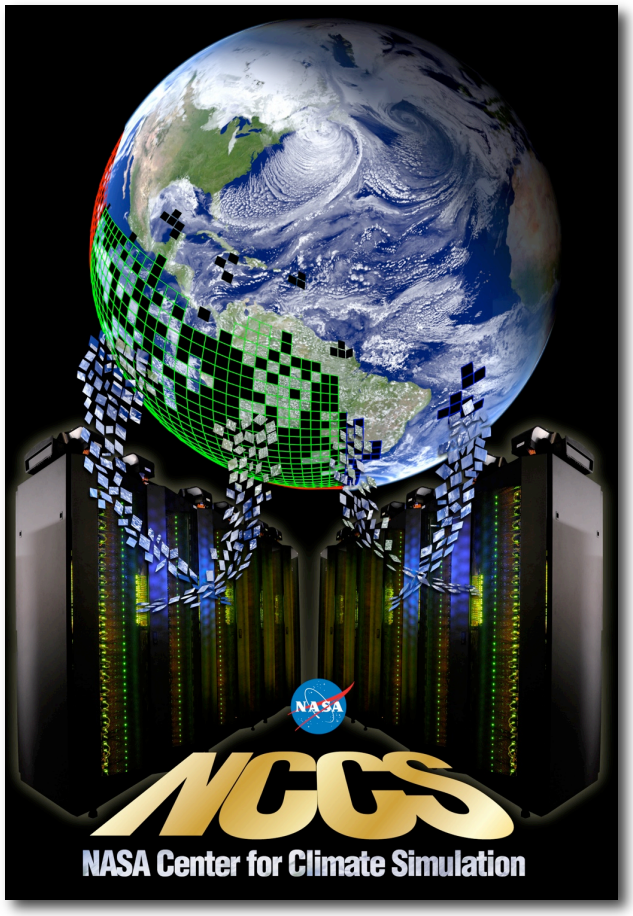
- vCDS Concept and Rationale
- vCDS 0.9 Anatomy / vCDS 0.9 Products

Part 2 – Where We Are Now (10 Minutes)

- NetCDF/IPCC Toolkit
- Adminstrative Extensions
- Repetitive Provisioning
- Operational Deployment
- Amazon Cloud vCDS-IPCC-ESG-v0.9

Part 3 – Wrap Up (5 Minutes +)

- Next Steps
- Discussion



Concept and Rationale

Scenario

A customer approaches the NCCS with a new dataset they want us to manage ...

Q. What technology is needed to quickly meet that customer's requirement under the follow constraints:

- The solution should be: simple, fast, and cheap;
- provide core capabilities to get started, but extendable to accommodate future needs;
- be flexible, with the ability to use, optimize, and change deployment configurations in response to resource availability;
- allow the new dataset to be integrated into an existing data collection; and
- come with a help desk and user support?

*The DMS Project has been looking at **iRODS data grid software** as a potential solution ...*

Definitions

Customer – an individual scientist, a lab, project, or mission.

Dataset – may be products generated by a GCM, may be observational data or a subset thereof, reanalysis data, or specialized products of value to an individual scientist or lab.

Manage – may refer to short-term file storage, long-term archival preservation; data may be used online by a person or application.

Examples

Abound:

- IPCC AR5 data for ESG.
- MODIS Atmospheres data for CMIP5.
- MERRA downscaled meteorological and environmental data.
- AgMIP, CERES, SMOS, Laboratory for Atmospheres, the Snowfake project ...

We begin the next phase by working with IPCC AR5 products ...

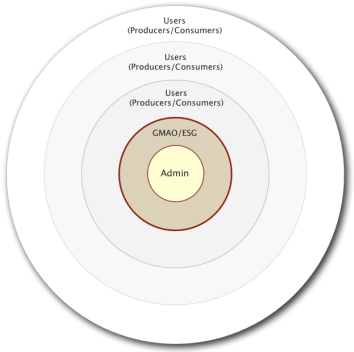
FY12 Q1/Q2

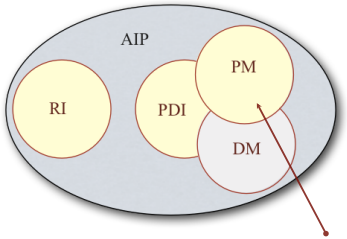
Managed Collections: (1) Publication Datasets

•GISS Collection – Served to ESG in Amazon

- *Ingest IPCC AR5 data into CDS 0.9*
- *Operationally harden CDS 0.9 to CDS 1.0 (TRL 9)*
- *Expose collection to ESG publication system*
- *Develop Collection Administrator's interface*

A production system in Nebula that mirrors the NCCS's current capabilities ...

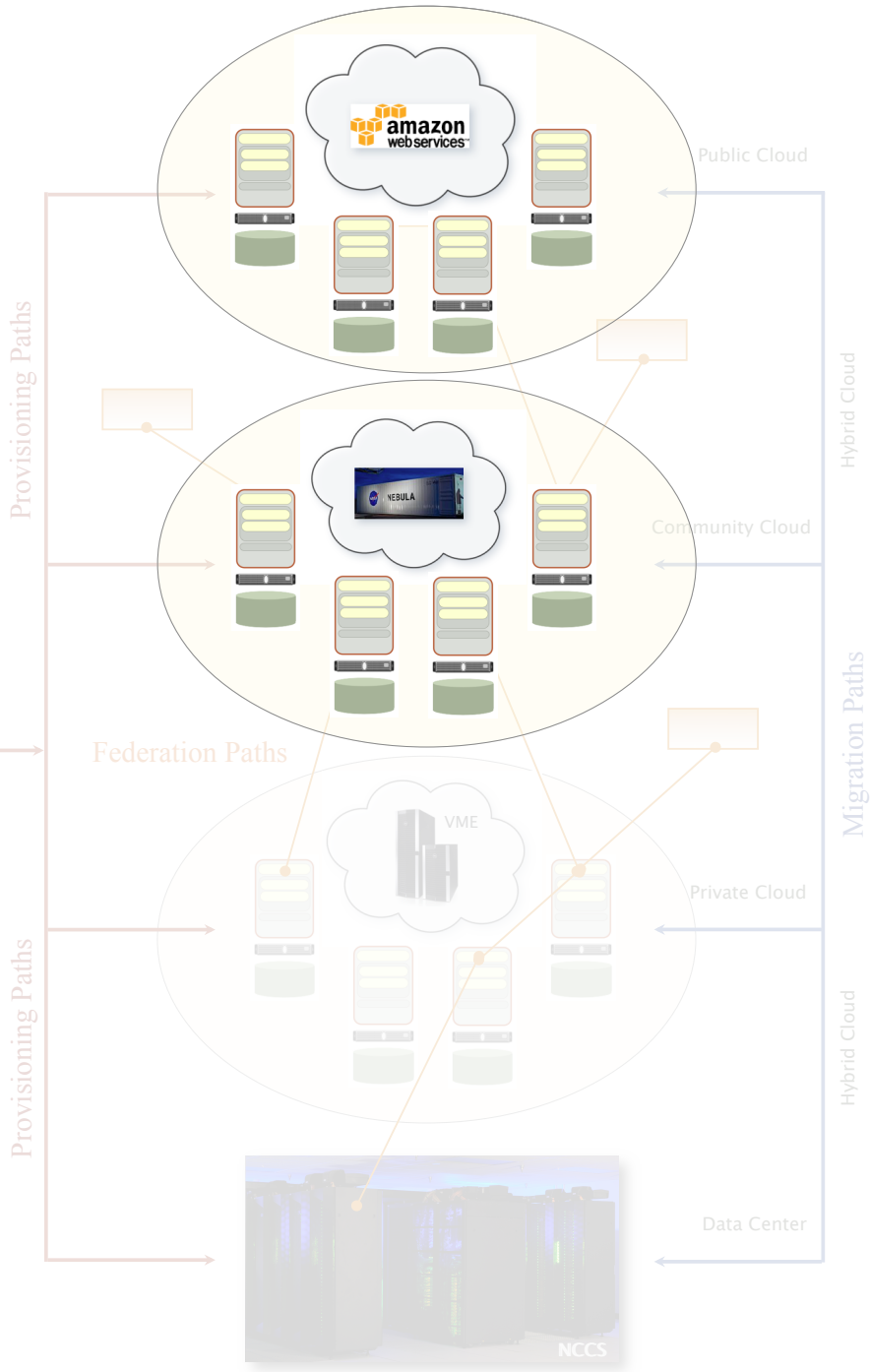




Develop requirements, implement Collection Administration policies and mechanisms, and specify OAIS Policy Metadata – all relatively easy with Publication Datasets.

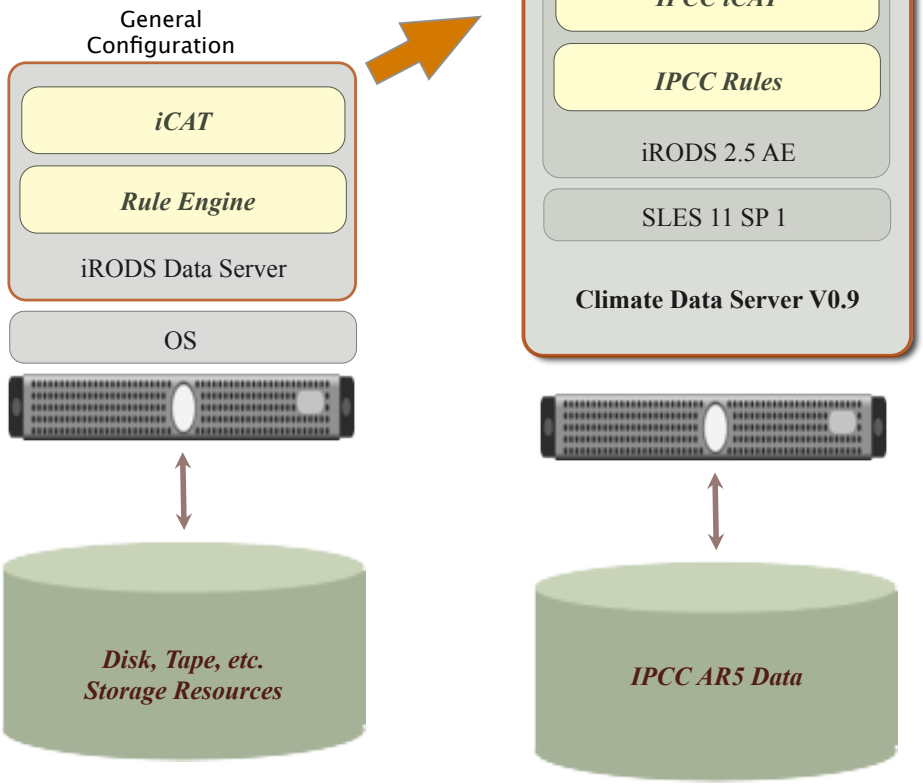
Published Collections	Estimated Current Size (TB)	Estimated Final Size (TB)
GISS IPCC	3.5	60
MERRA	0.5	0.5
CERES	0.1	0.1
AgMIP	0	0
GMAO IPCC	0	60
Total	**4.1**	**120.6**

Current estimates as of 9/1/2011. Total estimated size of IPCC AR5 200 TB.



vCDS V0.9 Anatomy

Our approach has been to build a core suite of general purpose scientific kits – such as NetCDF, HDF, and GeoTIF – that sit in the vertical stack above iRODS and below application-specific climate kits such as IPCC, MERRA, and SMOS ...



IPCC Kit + NetCDF Kit

CDS 0.9 Products

1.IPCC / NetCDF Module

iRODS microservices, rules, configuration settings, and software utilities required to implement canonical CRUD operations for IPCC/NetCDF system kernel ...

•Administrative Extensions

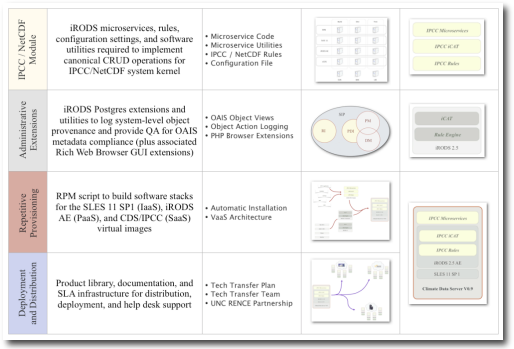
iRODS Postgres extensions and utilities to log system-level object provenance and provide QA for OAIS metadata compliance (plus associated Rich Web Browser GUI extensions) ...

•Repetitive Provisioning

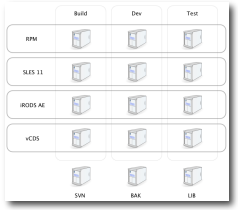
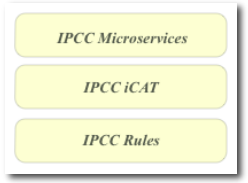
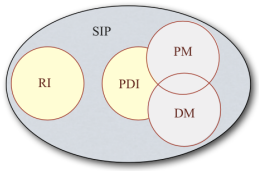
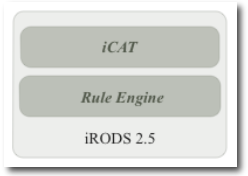
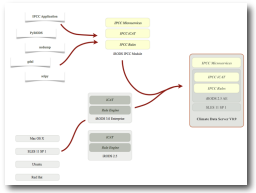
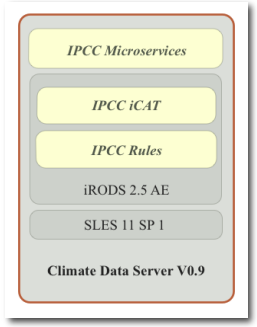
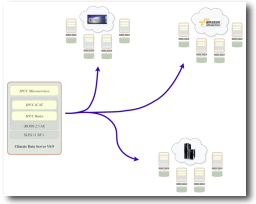
RPM scripts to build software stacks for the SLES 11 SP1 (IaaS), iRODS AE (PaaS), and CDS/IPCC (SaaS) virtual images ...

•Deployment and Distribution

Product library, documentation, and SLA infrastructure for distribution, deployment, and help desk support ...



vCDS V0.9 Products

IPCC / NetCDF Module	iRODS microservices, rules, configuration settings, and software utilities required to implement canonical CRUD operations for IPCC/NetCDF system kernel	• Microservice Code • Microservice Utilities • IPCC / NetCDF Rules • Configuration File		
Administrative Extensions	iRODS Postgres extensions and utilities to log system-level object provenance and provide QA for OAIS metadata compliance (plus associated Rich Web Browser GUI extensions)	• OAIS Object Views • Object Action Logging • PHP Browser Extensions		
Repetitive Provisioning	RPM script to build software stacks for the SLES 11 SP1 (IaaS), iRODS AE (PaaS), and CDS/IPCC (SaaS) virtual images	• Automatic Installation • VaaS Architecture		
Deployment and Distribution	Product library, documentation, and SLA infrastructure for distribution, deployment, and help desk support	• Tech Transfer Plan • Tech Transfer Team • UNC RENCE Partnership		

IPCC / NetCDF Module

IPCC NetCDF Metadata

Right now application-independent metadata appears in several places ...

A. Self-describing file header

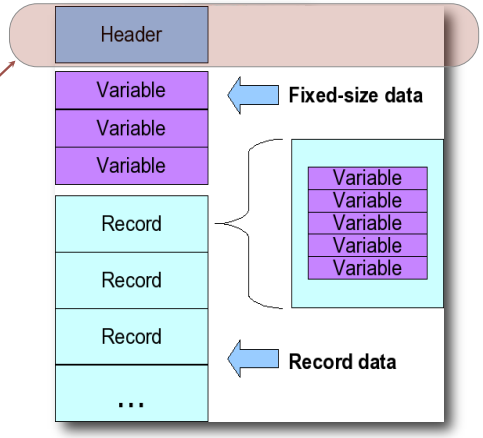
B. Filesystem path

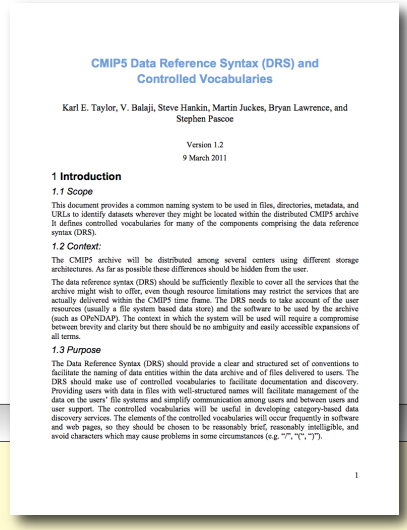
C. File attributes

D. File name

Metadata ownership / rules come from several places ...

- Producer / CMIP5 DRS
- Producer + Admin / Policy
- Operating system / Operating System
- Producer / CMIP5 DRS





CMIP5 Data Reference Syntax (DRS) and Controlled Vocabularies

Karl E. Taylor, V. Balaji, Steve Hankin, Martin Juckes, Bryan Lawrence, and Stephen Pascoe

Version 1.2
9 March 2011

1 Introduction

1.1 Scope

This document provides a common naming system to be used in files, directories, metadata, and URLs to identify datasets wherever they might be located within the distributed CMIP5 archive. It defines controlled vocabularies for many of the components comprising the data reference syntax (DRS).

1.2 Context:

The CMIP5 archive will be distributed among several centers using different storage architectures. As far as possible these differences should be hidden from the user.

The data reference syntax (DRS) should be sufficiently flexible to cover all the services that the archive might wish to offer, even though resource limitations may restrict the services that are actually delivered within the CMIP5 time frame. The DRS needs to take account of the user resources (usually a file system based data store) and the software to be used by the archive (such as OPeNDAP). The context in which the system will be used will require a compromise between brevity and clarity but there should be no ambiguity and easily accessible expansions of all terms.

1.3 Purpose

The Data Reference Syntax (DRS) should provide a clear and structured set of conventions to facilitate the naming of data entities within the data archive and of files delivered to users. The DRS should make use of controlled vocabularies to facilitate documentation and discovery. Providing users with data in files with well-structured names will facilitate management of the data on the users' file systems and simplify communication among users and between users and user support. The controlled vocabularies will be useful in developing category-based data discovery services. The elements of the controlled vocabularies will occur frequently in software and web pages, so they should be chosen to be reasonably brief, reasonably intelligible, and avoid characters which may cause problems in some circumstances (e.g. "/", "\", ":").

1

```
> pwd
/portal/GISS/AR5/piControl/E2-R_piControl_r1i1p1
> ls -al
-rw-r--r--  1 giss  admin  20828964  Jul  18  08:05  cSoil_Lmon_GISS-E2-R_piControl_r1i1p1_398101-400511.nc
-rw-r--r--  1 giss  admin  20828964  Jul  18  08:05  cSoil_Lmon_GISS-E2-R_piControl_r1i1p1_398101-400512.nc
>
```

CDS 0.9 Products

1. IPCC / NetCDF Module

iRODS microservices, rules, configuration settings, and software utilities required to implement canonical CRUD operations for IPCC/NetCDF system kernel ...

- Administrative Extensions

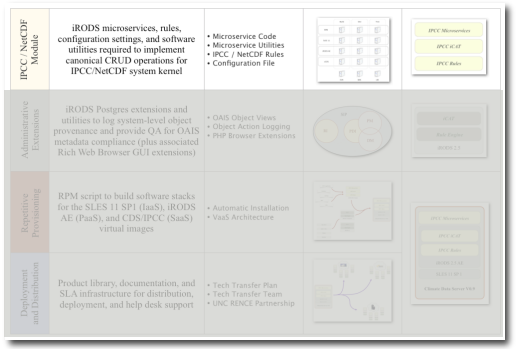
iRODS Postgres extensions and utilities to log system-level object provenance and provide QA for OAIS metadata compliance (plus associated Rich Web Browser GUI extensions) ...

- Repetitive Provisioning

RPM scripts to build software stacks for the SLES 11 SP1 (IaaS), iRODS AE (PaaS), and CDS/IPCC (SaaS) virtual images ...

- Deployment and Distribution

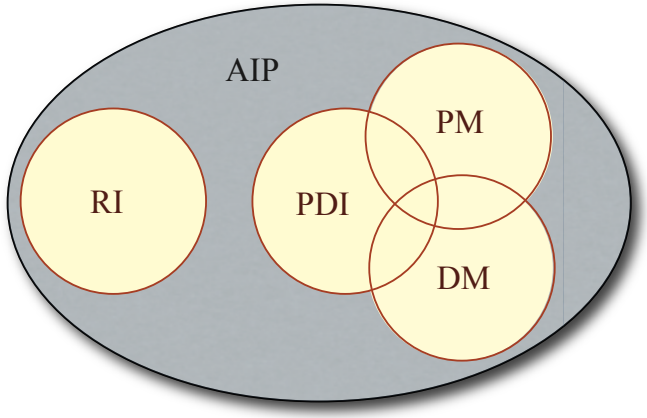
Product library, documentation, and SLA infrastructure for distribution, deployment, and help desk support ...



OAIS

Open Archival Information System (OAIS)

An OAIS is an archive, consisting of an organization of people and systems that has the responsibility to preserve information and make it available for a designated community ...

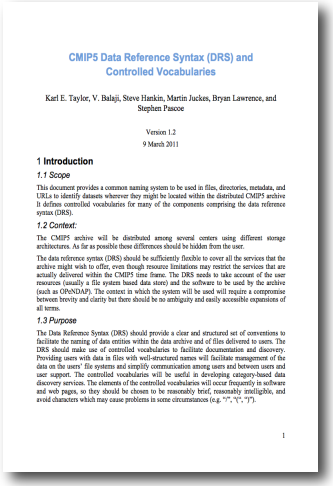


The reference model addresses a full range of archival information preservation functions including:

- ingest, data management, access, and dissemination;
- the migration of digital information to new media and forms;
- the data models used to represent the information, the role of software in information preservation, and the exchange of digital information among archives.

And it identifies both internal and external interfaces to the archive functions;

- it identifies a number of high-level services at these interfaces;
- it provides various illustrative examples and some 'best practice' recommendations;
- and it defines a minimal set of responsibilities for an archive to be called an OAIS.



CMIP5 Data Reference Syntax (DRS) and Controlled Vocabularies

Karl E. Taylor, V. Balaji, Steve Hankin, Martin Juckes, Bryan Lawrence, and Stephen Pascoe

Version 1.2
9 March 2011

1 Introduction

1.1 Scope

This document provides a common naming system to be used in files, directories, metadata, and URLs to identify datasets wherever they might be located within the distributed CMIP5 archive It defines controlled vocabularies for many of the components comprising the data reference syntax (DRS).

1.2 Context:

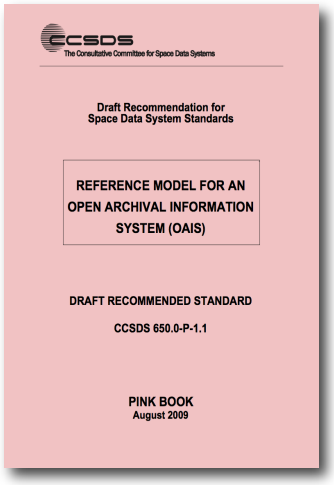
The CMIP5 archive will be distributed among several centers using different storage architectures. As far as possible these differences should be hidden from the user.

The data reference syntax (DRS) should be sufficiently flexible to cover all the services that the archive might wish to offer, even though resource limitations may restrict the services that are actually delivered within the CMIP5 time frame. The DRS needs to take account of the user resources (usually a file system based data store) and the software to be used by the archive (such as OPeNDAP). The context in which the system will be used will require a compromise between brevity and clarity but there should be no ambiguity and easily accessible expansions of all terms.

1.3 Purpose

The Data Reference Syntax (DRS) should provide a clear and structured set of conventions to facilitate the naming of data entities within the data archive and of files delivered to users. The DRS should make use of controlled vocabularies to facilitate documentation and discovery. Providing users with data in files with well-structured names will facilitate management of the data on the users' file systems and simplify communication among users and between users and user support. The controlled vocabularies will be useful in developing category-based data discovery services. The elements of the controlled vocabularies will occur frequently in software and web pages, so they should be chosen to be reasonably brief, reasonably intelligible, and avoid characters which may cause problems in some circumstances (e.g. "/", "\", ":").

1



CCSDS
The Consultative Committee for Space Data Systems

Draft Recommendation for Space Data System Standards

REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS)

DRAFT RECOMMENDED STANDARD

CCSDS 650.0-P-1.1

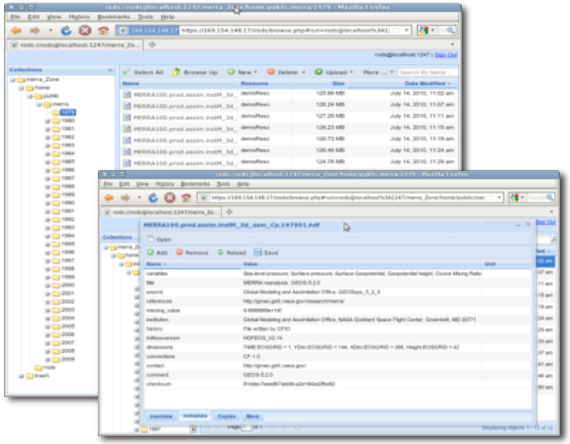
PINK BOOK
August 2009

Administrative Extensions

Major Functions

Basic system-level capabilities to log object provenance and provide OAIS package views of object metadata ...

- iRODS Postgres extensions
- iRODS microservice extensions
- Rich Web Browser extensions



CDS 0.9 Products

To appear in iRODS 3.0 release ...

1.IPCC / NetCDF Module

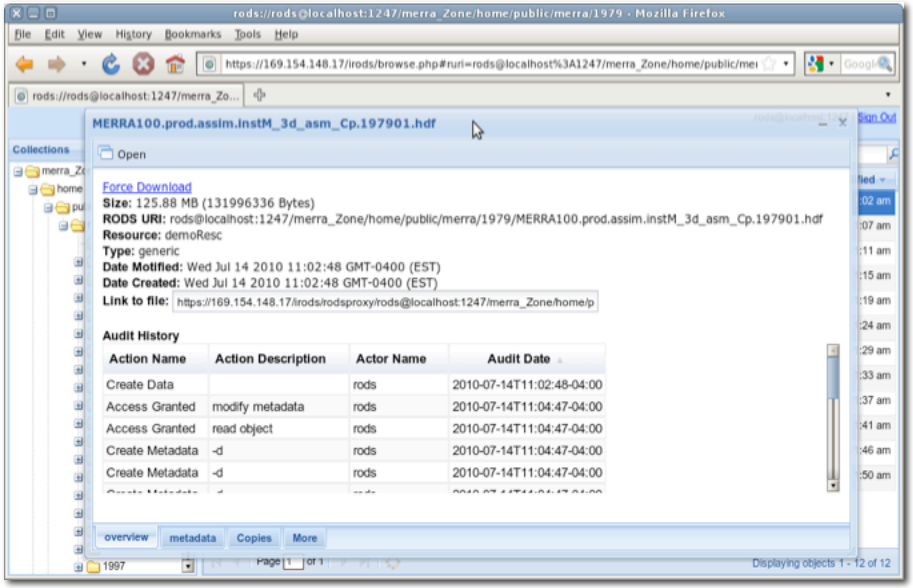
iRODS microservices, rules, configuration settings, and software utilities required to implement canonical CRUD operations for IPCC/NetCDF system kernel ...

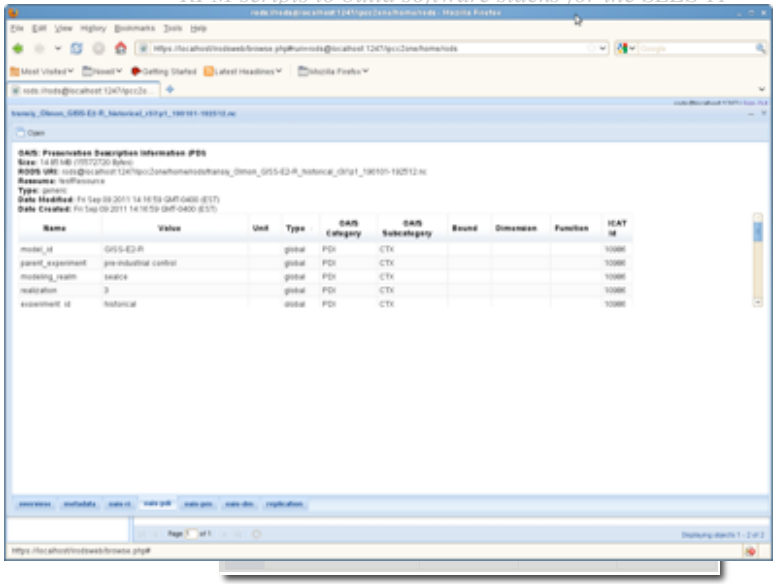
- Administrative Extensions

iRODS Postgres extensions and utilities to log system-level object provenance and provide QA for OAIS metadata compliance (plus associated Rich Web Browser GUI extensions) ...

- Repetitive Provisioning

RPM scripts to build software stacks for the SLES 11

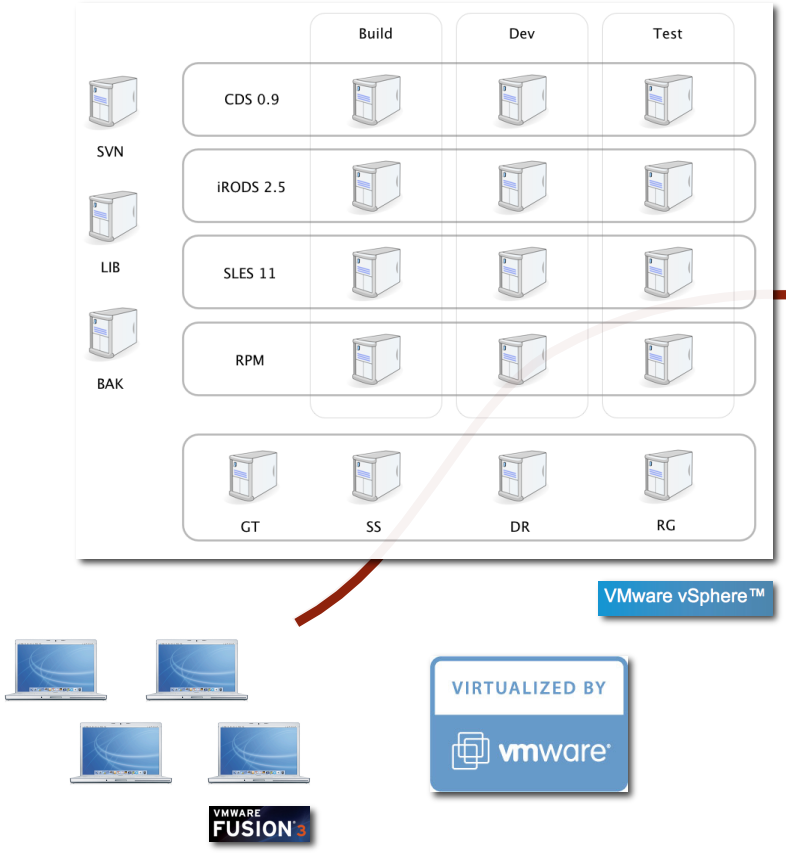




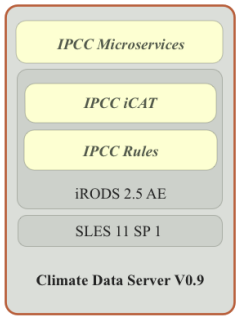
Repetitive Provisioning

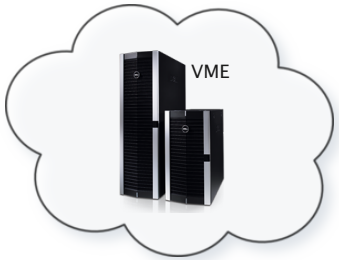
First, about our development environment ...

The DMS Project has worked in a virtualized environment – including MacTops with VMware Fusion and a VMware vSphere dev/test server farm.



This environment has influenced the way we are thinking about building, distributing, and deploying CDS components ...





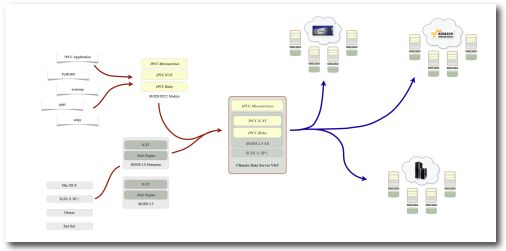
Rationale: Why create an RPM?

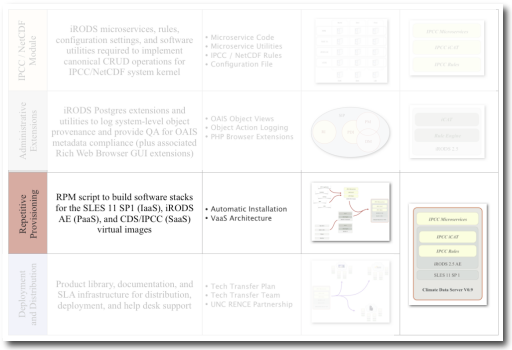
- Automate installation

Goal is to be able to conveniently install iRODS and our vCDS software stack in different environments and on different platforms ...

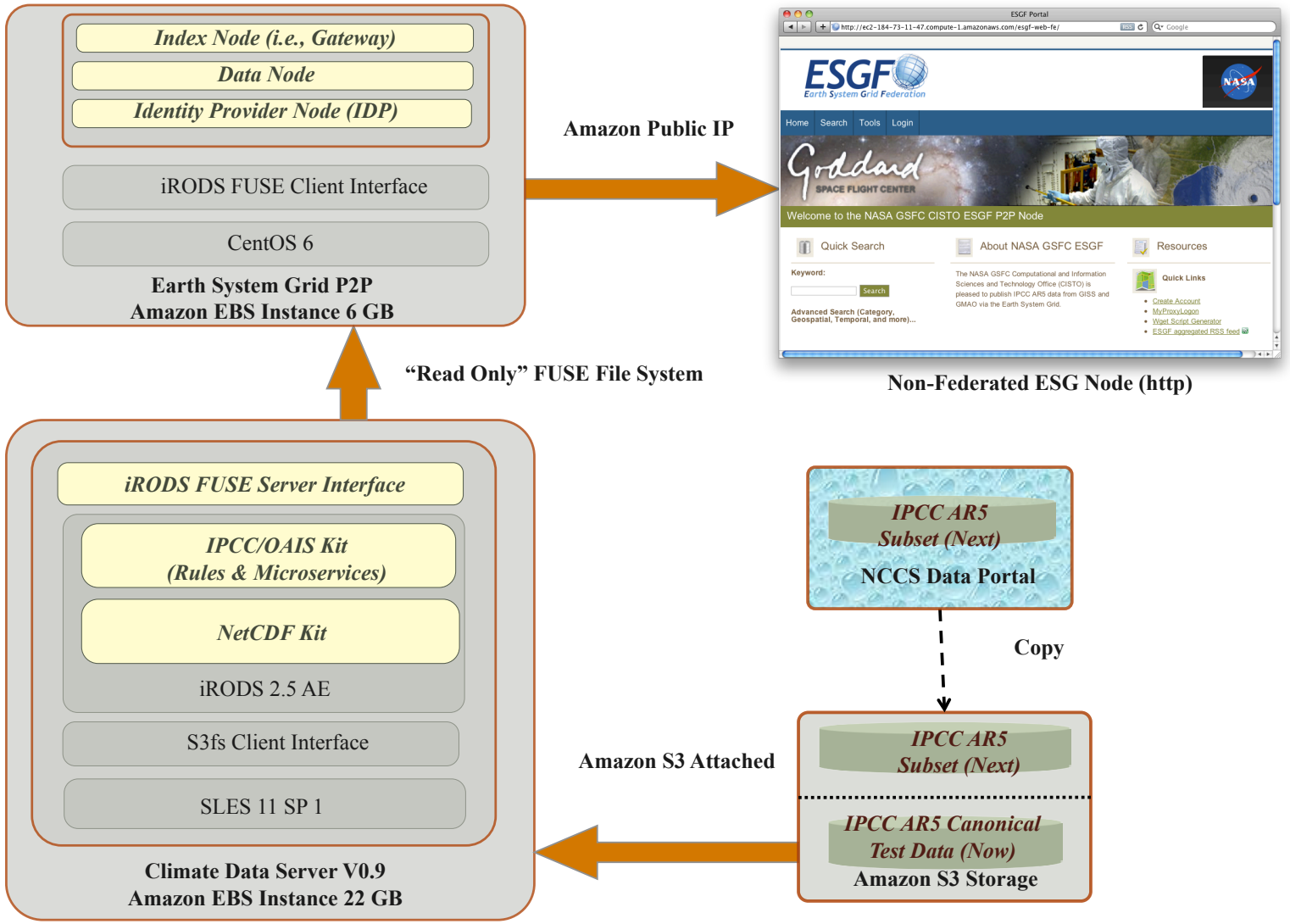
- Reduce installation errors and eliminate the user inteface

Installing iRODS "out of the box" is cumbersome, and manual installation leads to errors and unstable systems ...





Operational Deployment - Amazon Cloud vCDS-IPCC-ESG-v0.9



... then move to other primary and derived products ...

FY12 Q3/Q4

Managed Collections: (1) Publication Datasets

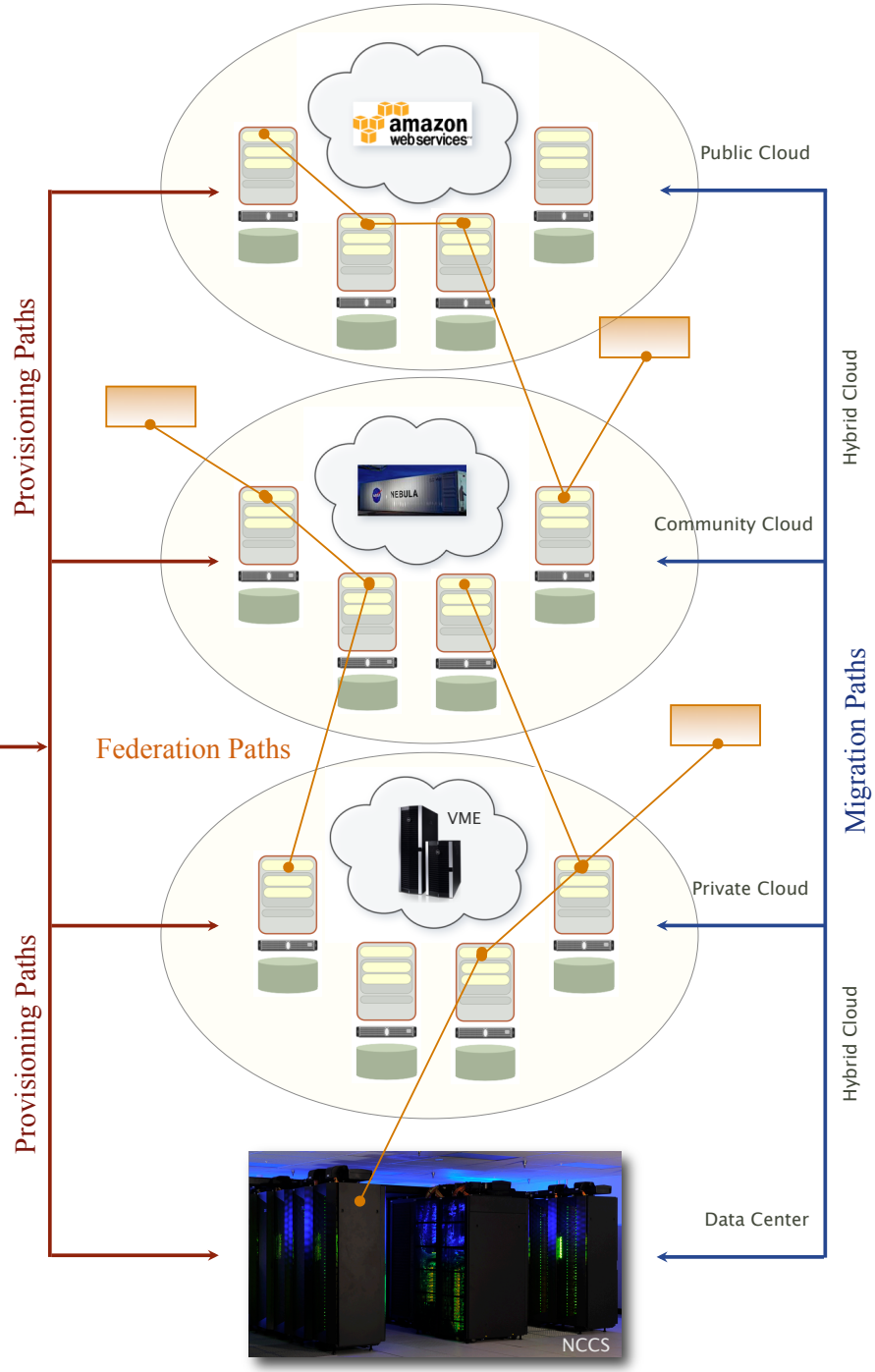
•MERRA Collection
- *Develop iRODS MERRA kit*
- *Develop iRODS HDFS, Swift, S3 drivers*
- *Create testbed HDFS, Swift, S3 repositories*
- *Expose collection to ESG publication system*

•CERES, AgMIP, GMAO Collections
Replicate/refine the Managed Collections processes as needed to accommodate customer interest, response, and needs – and budget, time, and political constraints ...

•NASA / RENCI jointly developed products:
1) CDS V1.0 Enterprise Edition
2) NetCDF, HDF science kits
3) IPCC, MERRA climate kits
4) HDFS, Swift, S3 drivers
5) iDROP Collection Administrators GUI

Published Collections	Estimated Current Size (TB)	Estimated Final Size (TB)
✓ GISS IPCC	3.5	60
MERRA	0.5	0.5
CERES	0.1	0.1
AgMIP	0	0
GMAO IPCC	0	60
Total	**4.1**	**120.6**

Current estimates as of 9/1/2011. Total estimated size of IPCC AR5 200 TB.



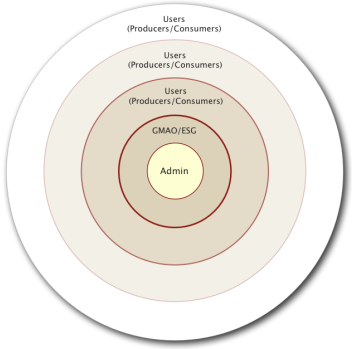
... then to the challenging task of active research collections.

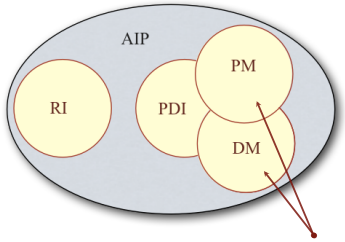
FY13 and beyond ...

Managed Collections: (2) Research Datasets

•Operational Research

- *Transition from "Archive" to "Managed Collections"*
- *Approach will be stepwise, incremental, logical*
- *Need established technology, a process, and an expert team*
- *And a conceptual model for how this is done ...*

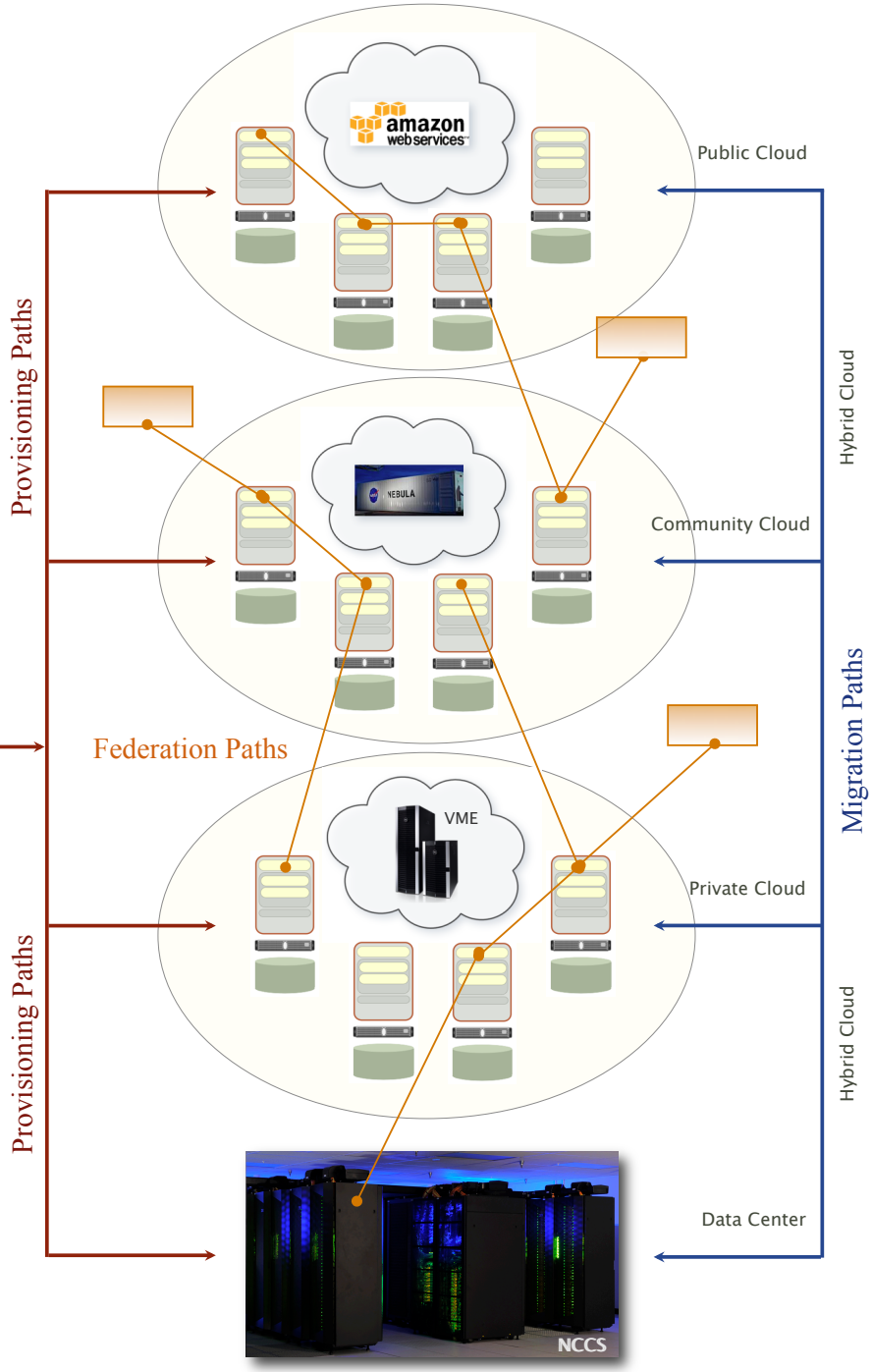




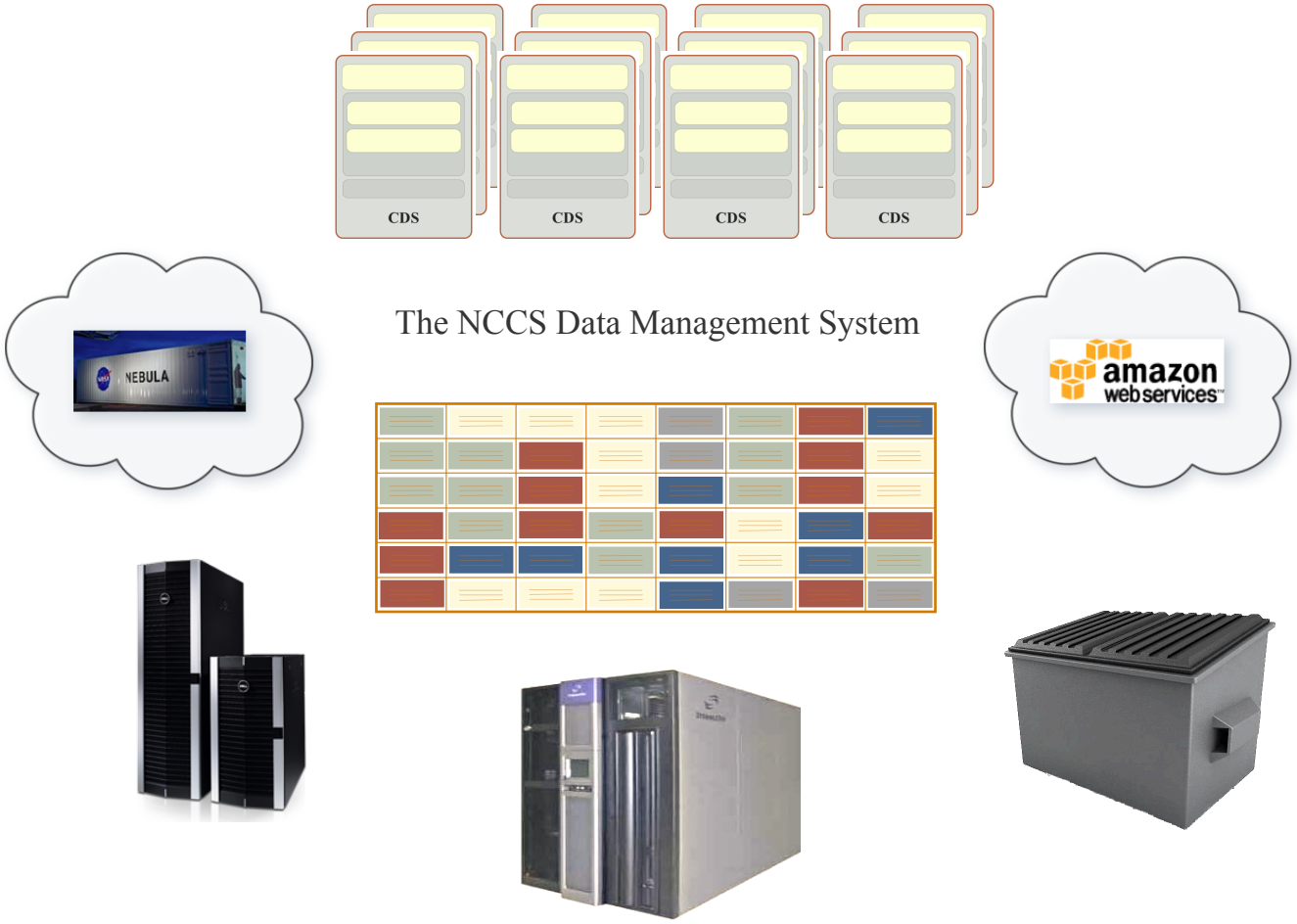
Develop requirements, implement user policies and mechanisms, and specify OAIS Policy Metadata and Discovered Metadata – this is where we add layers to the kernel.

Research Collections	Estimated Current Size (TB)	Estimated Final Size (TB)
Person	???.?	???.?
Person	???.?	???.?
Project	???.?	???.?
Project	???.?	???.?
Lab	???.?	???.?
Total	**???.?**	**???.?**

Circumscribed Datasets



Discussion


CDS
CDS
CDS
CDS
NEBULA
The NCCS Data Management System
amazon
web services