The Comprehensive Large Array Stewardship System (CLASS) – Living with a multi- petabyte resource

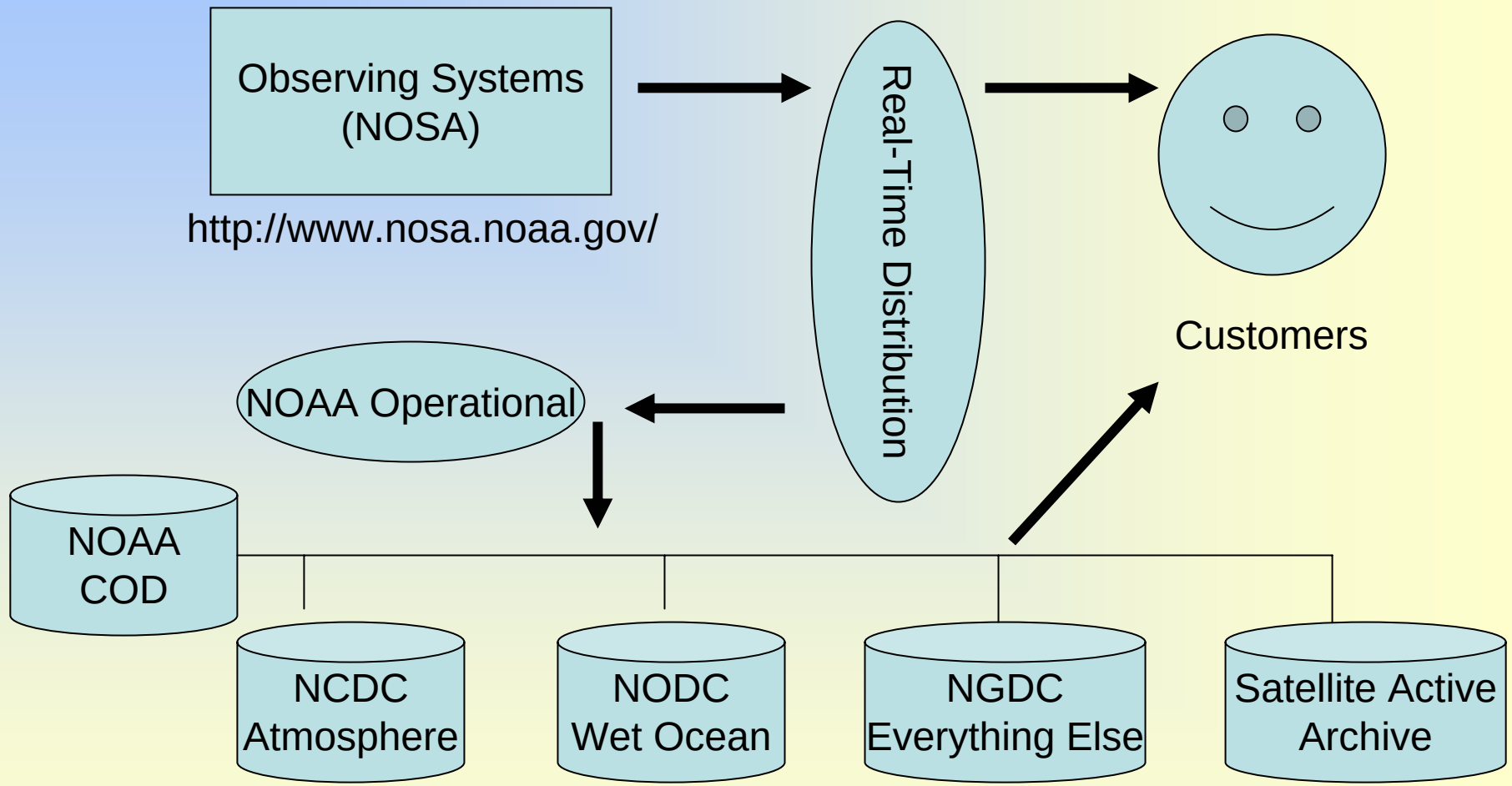
Outline

- What is CLASS?
- The role of CLASS in NOAA?
- The Future of CLASS
  - Hardware
  - API's
  - Relation to Data Centers

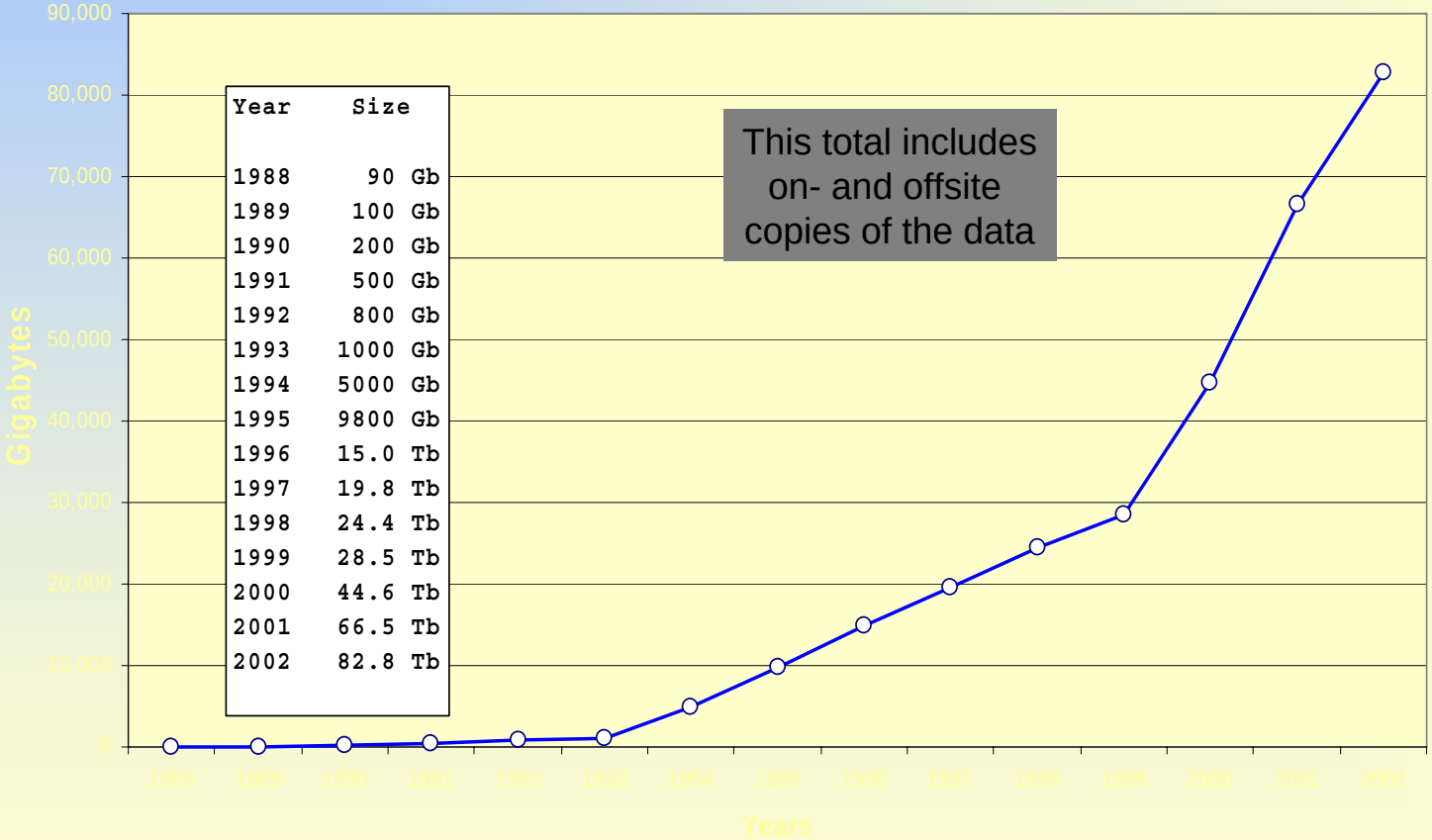
Why do I care?

- CLASS is a significant source of data for you and your users
- A CLASS API would allow independent usage of the data
- All NOAA data is being integrated either through GEOSS or CLASS
- CLASS is proto-typing a “hosted” data for simple storage

The data vision for NOAA



NGDC 14 Year Data Archive Growth

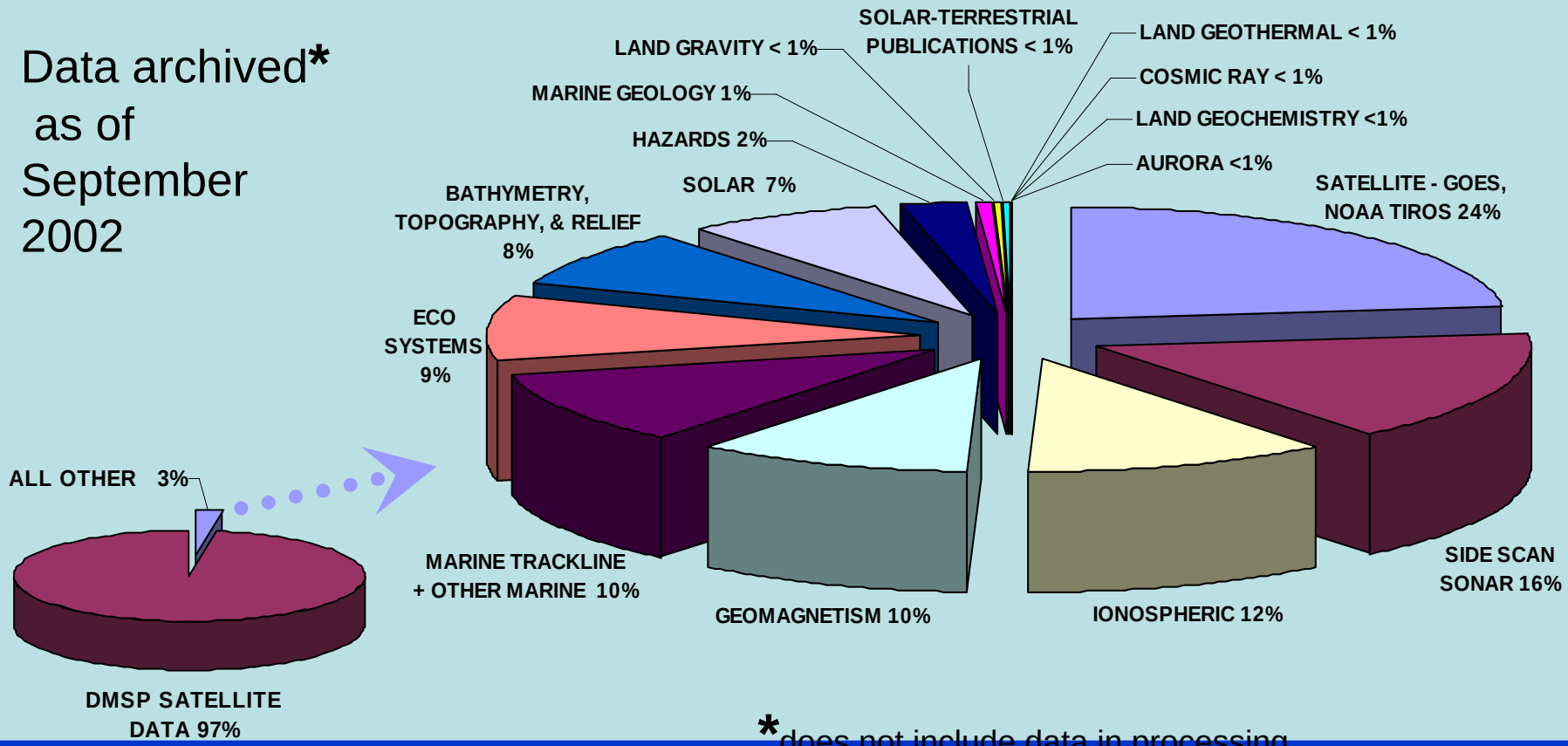


Year	Size
1988	90 Gb
1989	100 Gb
1990	200 Gb
1991	500 Gb
1992	800 Gb
1993	1000 Gb
1994	5000 Gb
1995	9800 Gb
1996	15.0 Tb
1997	19.8 Tb
1998	24.4 Tb
1999	28.5 Tb
2000	44.6 Tb
2001	66.5 Tb
2002	82.8 Tb

- NGDC stewards several hundred distinct types of data from multiple scientific disciplines
- DMSP data currently account for most of the volume displayed above
- Projected high volume data streams include CORS, side-scan, and multibeam

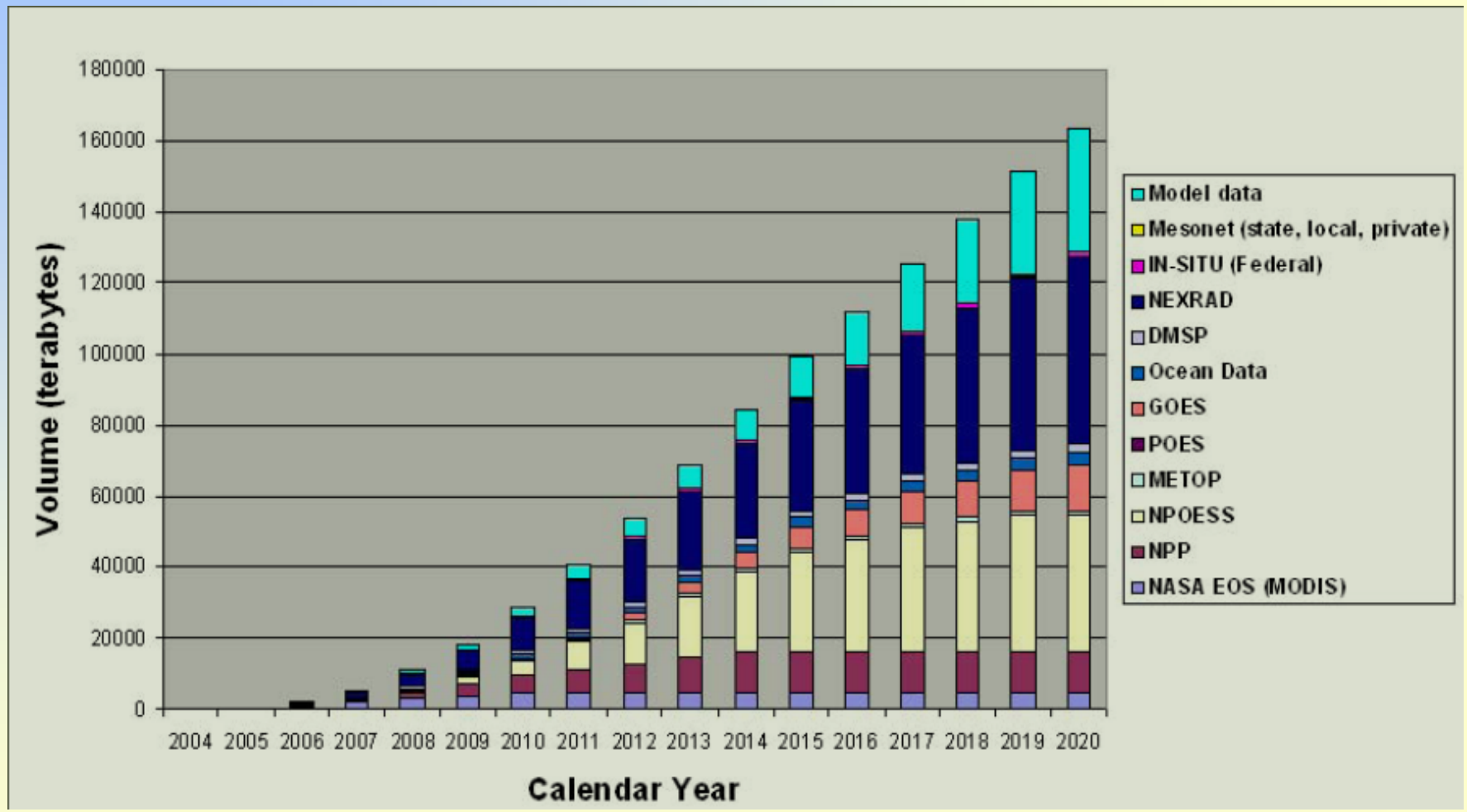
NGDC Holdings - % Mbytes by Data Type

Data archived* as of September 2002



*does not include data in processing

Whoa Nellie!


Volume (terabytes)
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
Calendar Year
Model data
Mesonet (state, local, private)
IN-SITU (Federal)
NEXRAD
DMSP
Ocean Data
GOES
POES
METOP
NPOESS
NPP
NASA EOS (MODIS)

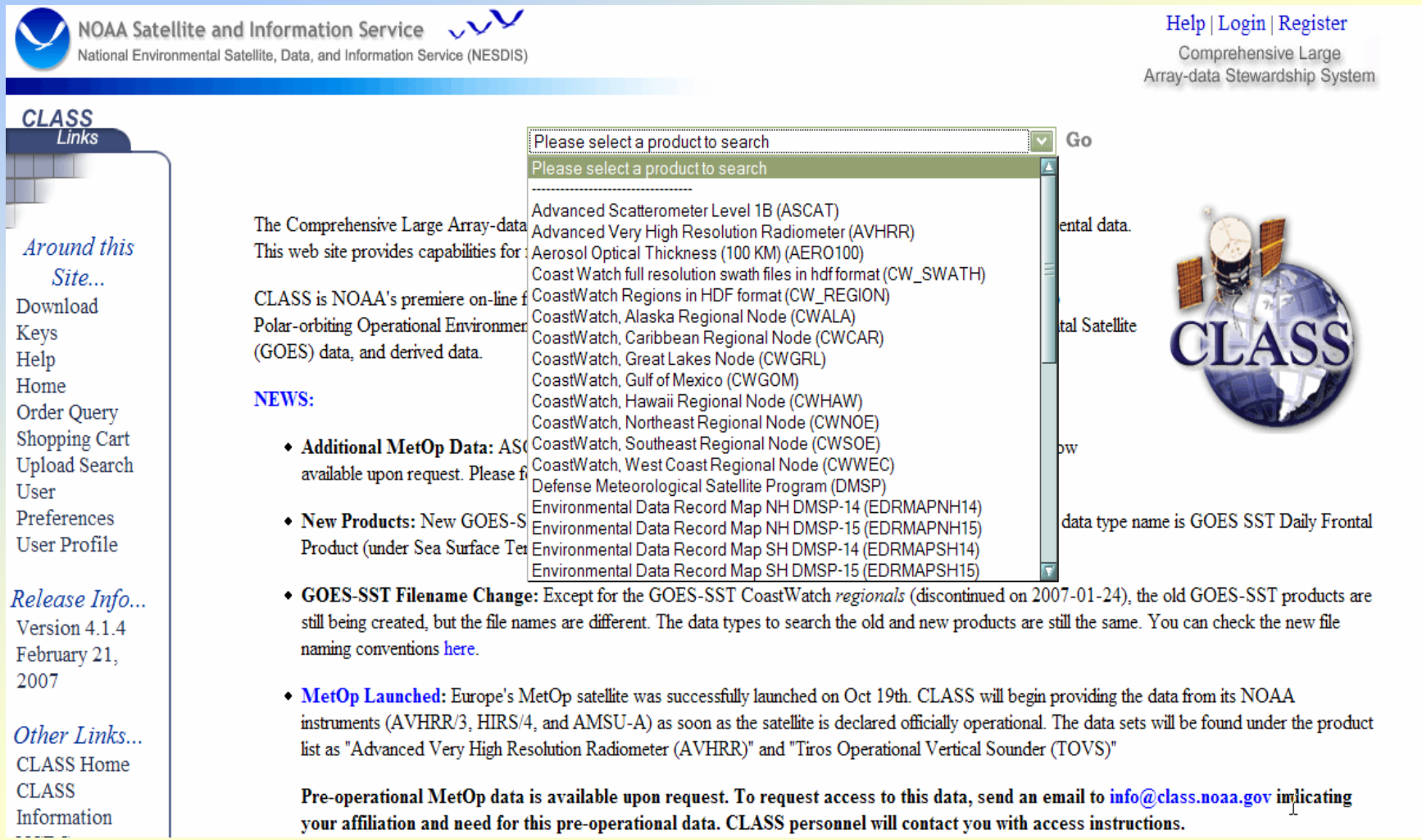
CLASS Vision - 2002

NOAA's National Data Centers and their world-wide clientele of customers look to CLASS as the sole NOAA IT infrastructure project in which all current and future large array environmental data sets will reside. CLASS provides permanent, secure storage and safe, efficient access between the Data Centers and the customers.

CLASS was selected from “best of breed”

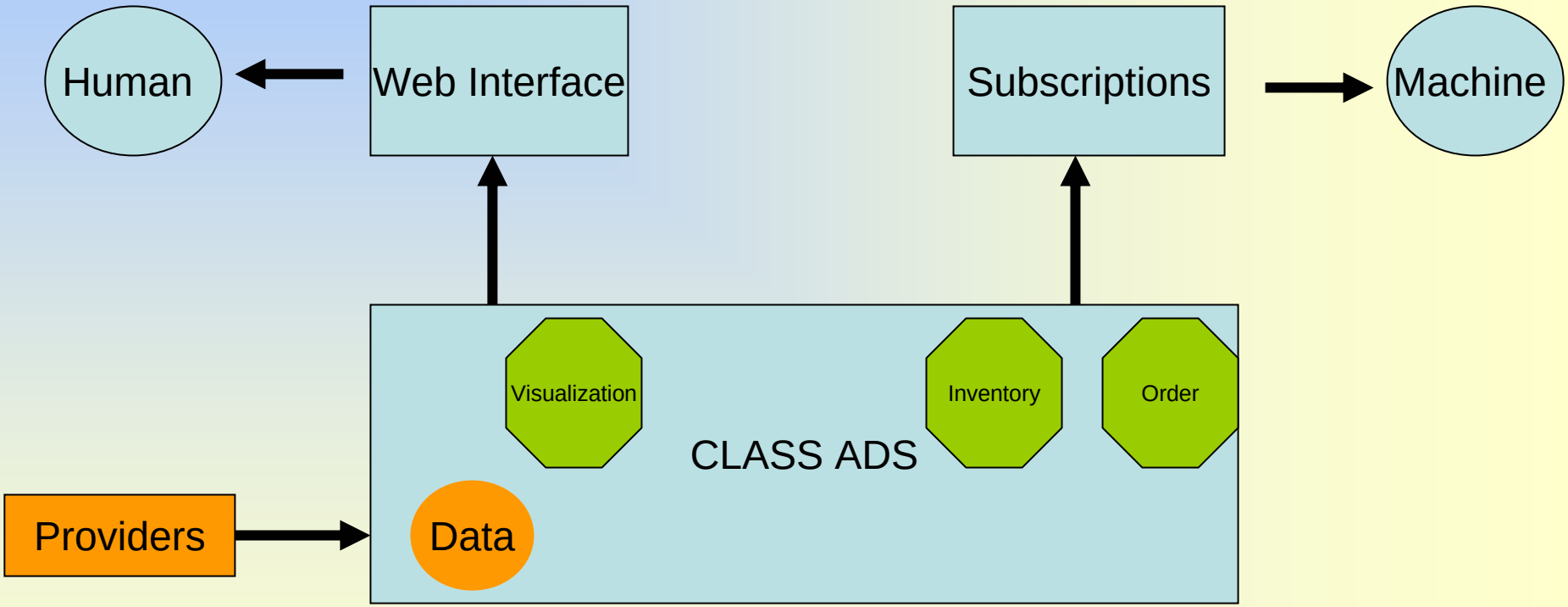
- Competition and selection between multiple existing systems
- SPIDR, SABR, SAA, etc..
- The SAA was chosen as the baseline system
- SAA has been evolving since 1994 and is organized around “campaigns”

CLASS Web



www.class.noaa.gov

CLASS Now


Human
Web Interface
Subscriptions
Machine
Visualization
Inventory
Order
CLASS ADS
Providers
Data

CLASS Boulder

- CLASS Boulder Facilities
  - 1,000 sq ft secure monitored space
  - 60 tons of cooling using two Liebert UH740C CRAC units
  - 80 KVA of Emergency Generator power from an APC 9215RM UPS for all CLASS systems

CLASS Boulder Installed IT Hardware

- 3 Dell 2850 PowerEdge servers running Red Hat Linux RHEL 4.0
- 1 Dell 850 PowerEdge server running MS Windows Server 2003
- 7 IBM p520 dual-cpu servers running IBM AIX 5.3
- 4 IBM p550 quad-cpu servers running IBM AIX 5.3
- APC ISX power management console
- IBM 7310-CR3 hardware management console (HMC)
- ADIC Scalar 100 LTO-2 tape library (backups)
- 2 SGI Origin 350 servers running IRIX OS
- CISCO Catalyst 48 port 6509E network switch w/ PIX firewall
- Data Direct Networks S2A3000 fiber channel SAN – 11 TB
- Brocade Silkworm 48000 4 Gbit SAN switch
- Data Direct Networks S2A9500 SATA SAN – 384 TB

CLASS Boulder Hardware not on site

- ADIC Scalar 10K LTO-2 tape library (CLASS data archive)
  - Due in Boulder in early April

Sprint/NOAA MPLS network OC3

- Installed, tested, and ready

Nearly Identical Node at Asheville, NC - NCDC

CLASS Boulder Server Room

- APC UPS, equipment racks, 30 ton CRAC unit

CLASS – Background

1990s

Satellite Active Archive (SAA)

Datasets:
POES
Coast Watch
DMSP Level 1b data & products
Radarsat SAR data

Capabilities:
Online delivery via ftp
Bulk order services
Subscription services
AVHRR browse
Web online services

Architecture:
Big Bird tapes

2006

CLASS Now (SAA Scaled)

Datasets:
GOES
POES
Coast Watch
DMSP data & products
Derived products

Capabilities:
Web online access
Online delivery (FTP, HTTP)
Off-line delivery
Subscription services (push/pull)
Bulk order services
Limited OpenDAP services
GOES & POES browse

Architecture:
Data online at Suitland on CLASS equipment,
Data online and accessible (robotic storage) via NCDC equipment,
Data archived at Asheville

CLASS Background Summary

- CLASS is a web-based data archive and distribution system for NOAA/NESDIS environmental data

- Archive … ingest, storage, metadata management, and data quality assurance
- Distribution … access, visualization, and data delivery

- CLASS is an extension of an 1995 operational system … SAA (Satellite Active Archive)
  - Transition to the CLASS architecture began in 2001

- CLASS currently supports POES and GOES data sets
  - GOES “campaign” is undergoing pre-operational testing

- CLASS will support additional campaigns, broader user base, new functionality currently being defined
  - CLASS must concurrently support ongoing operations and new requirements implementation

Observing Systems

- The first 7 sources of data for CLASS
  - NOAA and Department of Defense Polar-orbiting Operational Environmental Satellites (POES) and Defense Meteorological Satellite Program (DMSP) – **complete**
  - NOAA Geostationary-orbiting Operational Environmental Satellites (GOES) – **complete**
  - NOAA NEXT generation weather RADAR (NEXRAD) Program and future dual polarized and phased-array radars – **prototype in development**
  - National Aeronautics and Space Administration (NASA) Earth Observing System (EOS) Moderate-resolution Imaging Spectrometer (MODIS)
  - National Polar-orbiting Operational Environmental Satellite System (NPOESS) and NPOESS Preparatory Program (NPP)
  - EUMETSAT Meteorological Operational Satellite (MetOp) Program
  - National Centers for Environmental Prediction Model Datasets, including Reanalysis Products

Role of CLASS in NOAA

What is Scientific Data Stewardship? (SDS)

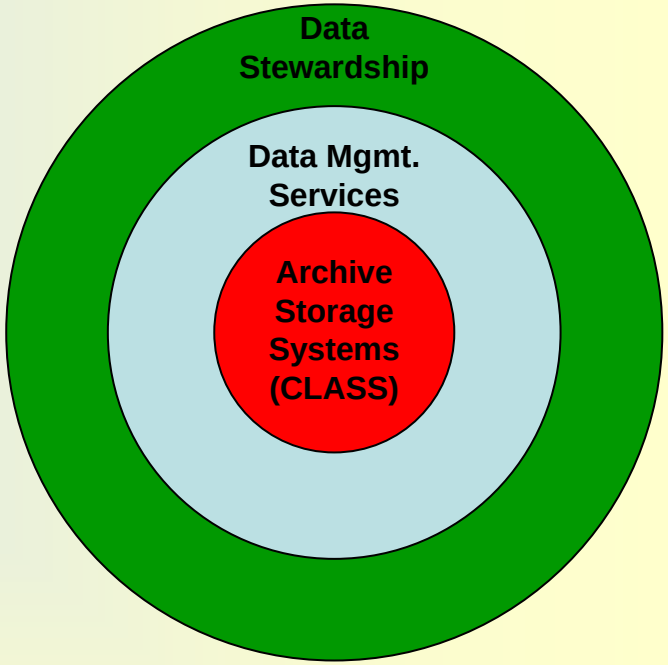
- “maintaining the science integrity and long term utility of scientific records”
- “ the actions which maximize the return on investment for archived scientific data”

NOAA Data Management

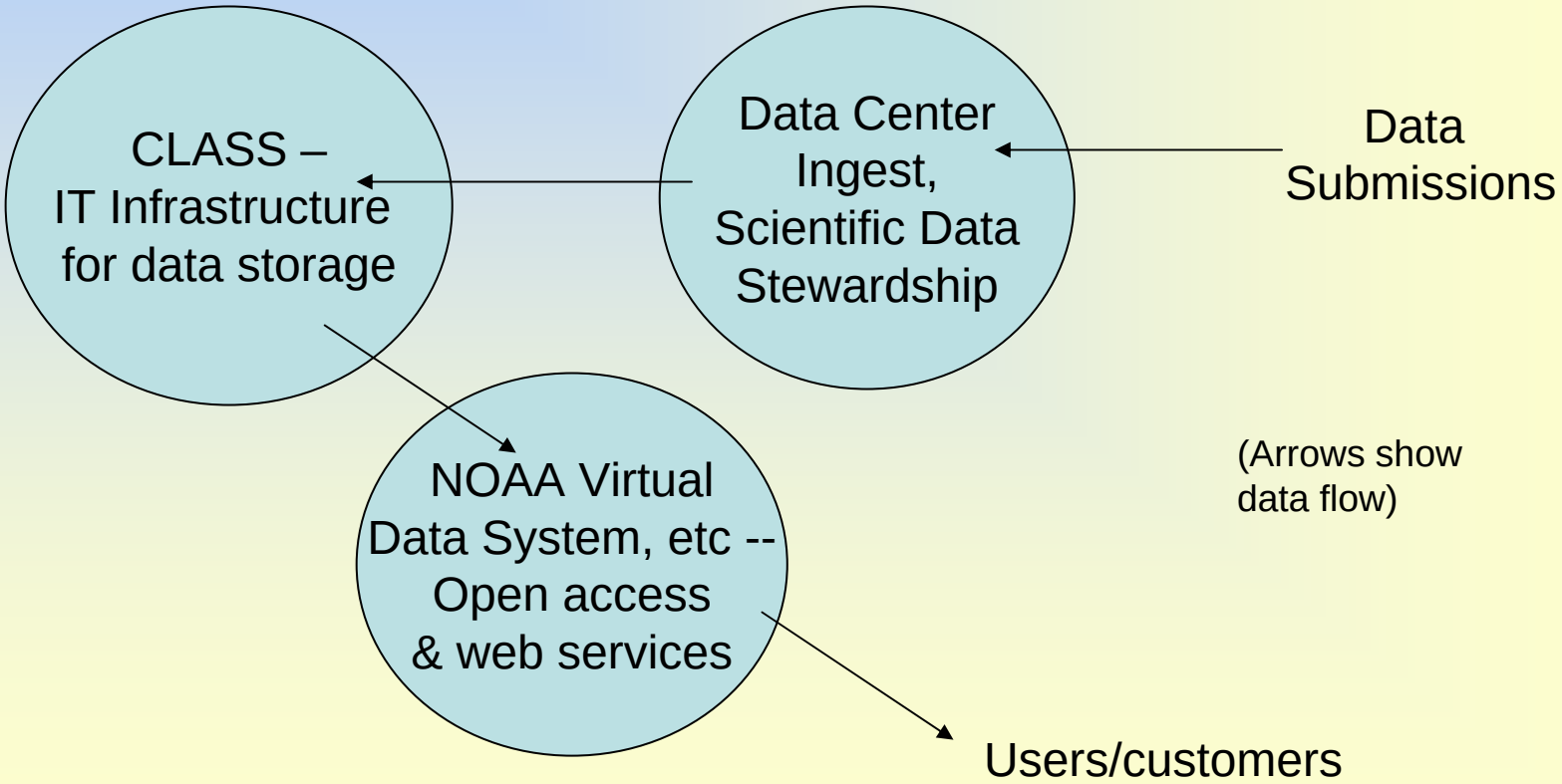
- ***The NOAA Strategic Information Technology Plan and Enterprise Architecture identify CLASS as the primary system to meet NOAA's data archiving requirements.***
- ***CLASS is: 1) a key component of NOAA Data Management Services, 2) an archive storage system and IT infrastructure for data storage***
- ***NOAA Data Management Service Mission Goals:***

  -- To be the most comprehensive and accessible source of quality climate, weather, oceanographic, biological, and geophysical related data and information services

  -- To assess the state of the atmospheric, oceanographic, terrestrial, solar and related geophysical environment



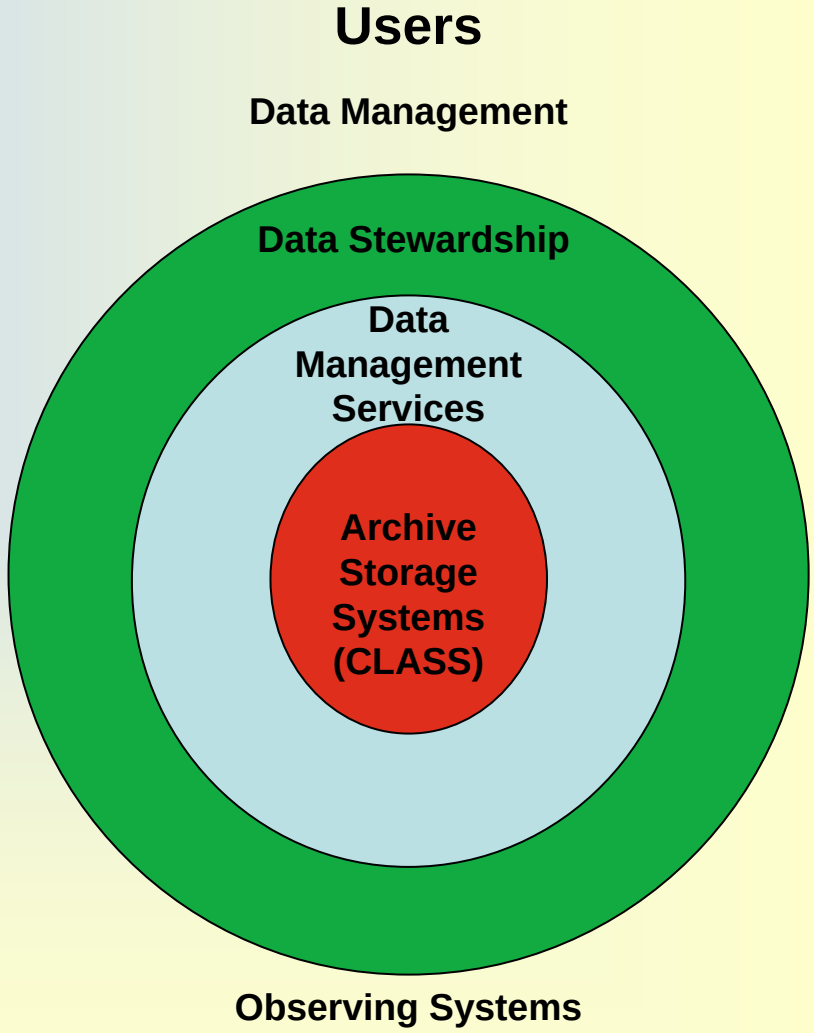
CLASS/Data Centers Concept of Operations Process Flow View



NOAA Data Management

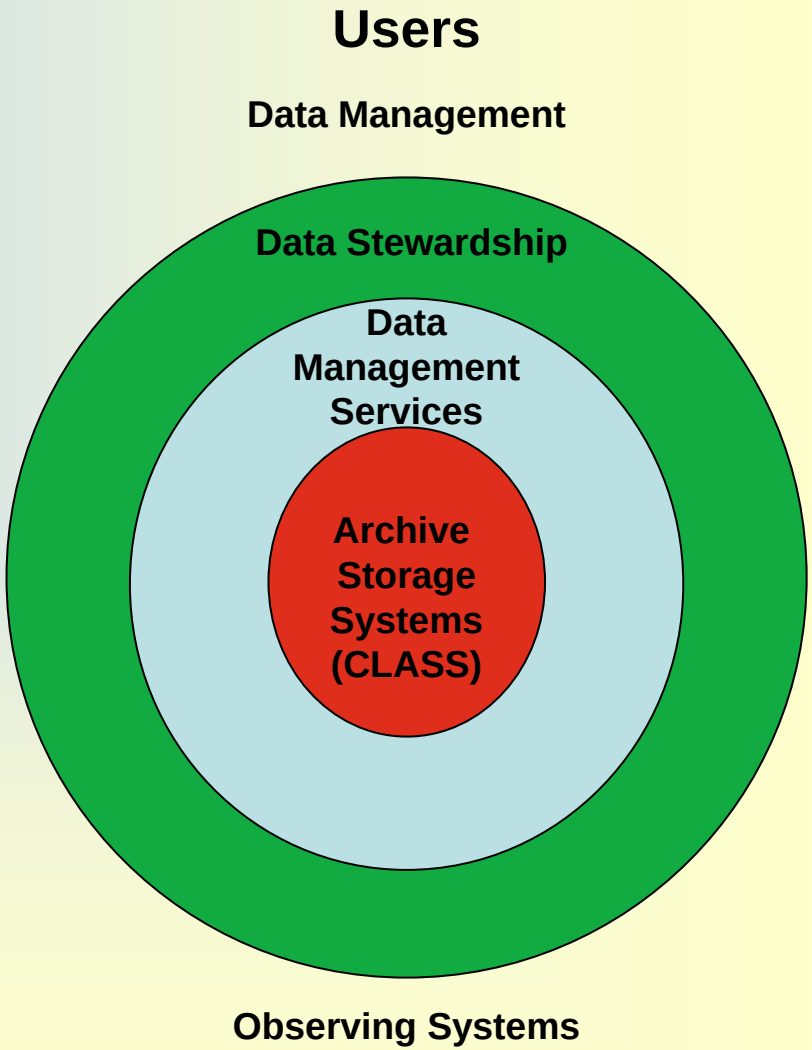
- **Data Management – Data stewardship and data management services – NOAA wide**

- **Data Stewardship**
  - Data Centers and Beyond
    - This includes documenting measurement practices and processing practices (metadata)
    - Providing feedback on observing system performance
    - Inter-comparison of data sets for validation
    - Reprocessing (incorporate new data, apply new algorithms, perform bias corrections, integrate/blend data sets from different sources or observing systems)
    - Recommending corrective action for errant or non-optimal operations.



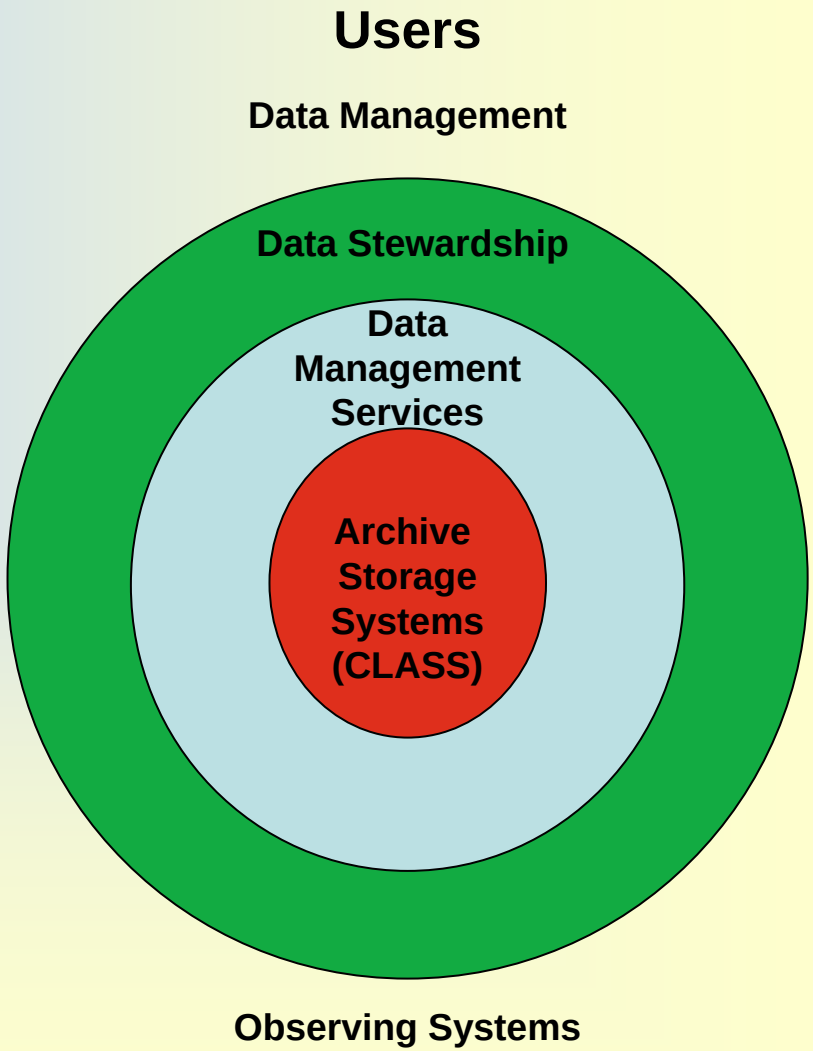
NOAA Data Management

- Data Management Services
  - **Data Centers and Centers of Data**
    - This includes adherence to agreed-upon standards
    - Ingesting data, developing collections, and creating products
    - Maintaining data bases; ensuring permanent, secure archival
    - Providing both user-friendly and machine-interoperable access; assisting users
    - Migrating services to emerging technologies
    - Responding to user feedback
    - Data management responsibilities – research, QC, archive, access
    - Fully mature and robust open access systems with E-commerce, free data, browse, FTP, data sub-setting, Web Services, GIS Services, OpenDAP, etc.
    - Includes NOAA/NESDIS Data Centers using the NOAA Virtual Data System (NVDS)

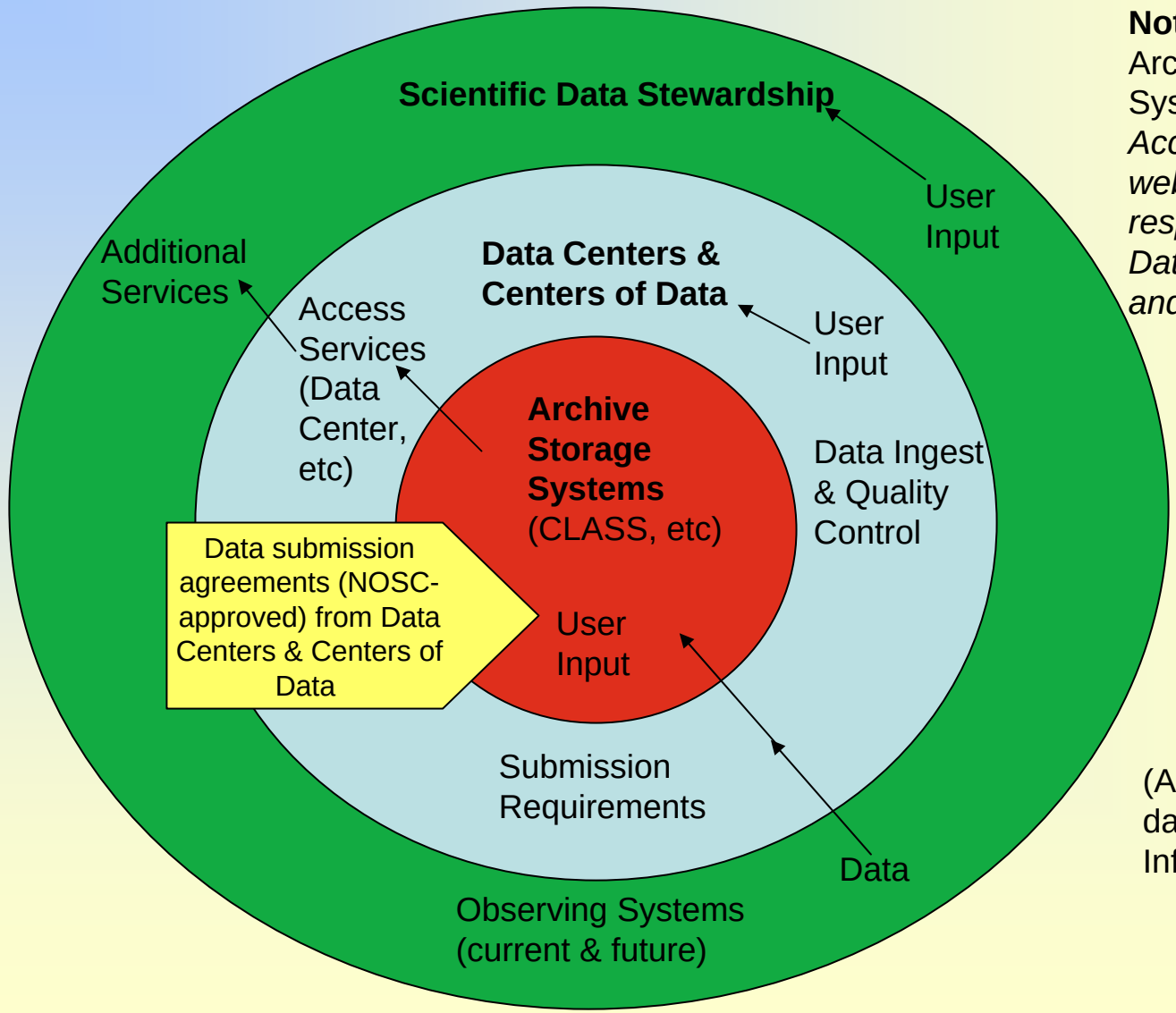


NOAA Data Management

- **Archival Storage Systems**
  - **CLASS** is a subset of Data Management Services
  - IT infrastructure to store NOAA's data (end-to-end hardware solution)
  - Data easily accessible via various (Data Center and Centers of Data) access systems – existing and future
  - Adherence to standards for dataset integration projects
  - Store data with unique formats



NOAA Data Management Concept of Operations Detailed View



Note: CLASS is an Archive Storage System.
Access services and web systems are the responsibility of Data Centers and Centers of Data.

(Arrows show data and Information flow)

NOAA Data Stewardship and Management Services Roles and Responsibilities: Lead and Shared

Open Archive Information System – Reference Model – 6 Requirements

Ingest (Data Centers and Centers of Data):

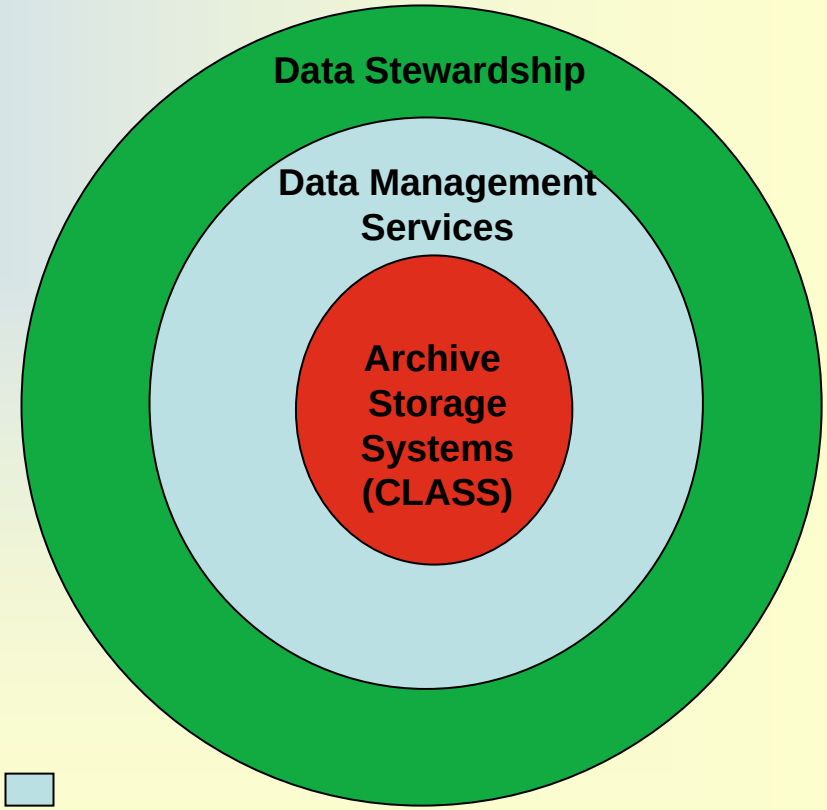
- Negotiate and accept information from information producers

Archive (CLASS in coordination with Data Centers):

- Obtain sufficient control to ensure long-term preservation
- Follow documented policies and procedures that ensure the information is preserved against all reasonable contingencies

Access (Data Centers and Centers of Data):

- Determine which communities (designated) need to be able to understand the preserved information
- Ensure the information to be preserved is independently understandable to the Designated Communities
- Make the preserved information available to the Designated Communities in forms understandable to those communities



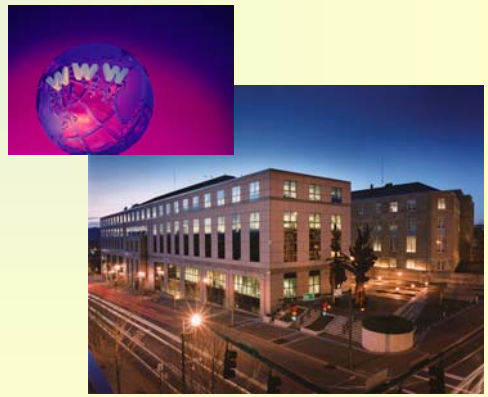
NOAA Data Management Vision

CLASS
Open Archive System Architecture

CLASS Capabilities:
- IT tape robotics for rapid retrieval
- High-speed disk access
- End-to-end hardware solution
- OAIS-compliant
- File structures and metadata to support existing & future Data Center systems

Data Center Responsibilities:
- CLASS operations, when CLASS development is complete
- NOAA Virtual Data System (NVDS)
- Service-Oriented Architecture
- Ingest and archive services
- Submission agreements
- Web interfaces for integrated data access (data portal)
- Web Services (SOAP, etc)
- OpenDAP & open access services
- GIS/OGC services
- Subscription services
- OAIS-compliant
- 24/7 operations support

Architecture:
- Dual site: Asheville and Boulder mirror sites
- Data stored on CLASS-funded, Data-Center approved storage systems

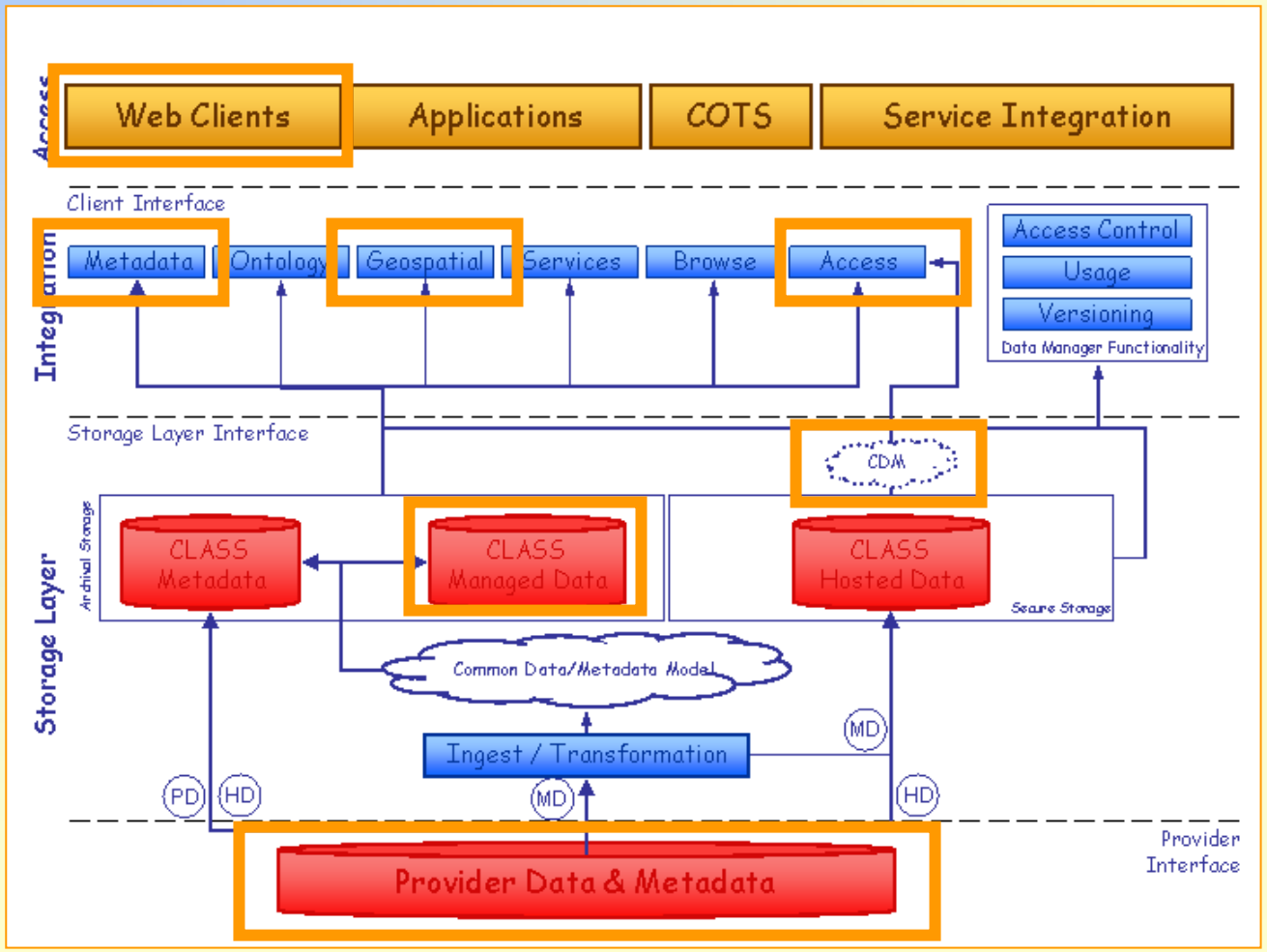


NOAA Data Management Issues

- Must have archive standards as determined by Data Centers—abide by OAIS & GEO-IDE principles
- Decision process/oversight for archival
- Metadata are critical
- Participatory design needed in all aspects of IT architecture and system design, so that Data Centers are fully engaged and approve of the final decisions
- Web systems and access services are the responsibility of the Data Centers and Centers of Data to develop and operate

CLASS Future

CLASS –NGDC Proto-type Scope



Goals

- First draft of a user focused web-services interface
- Demonstration of the concept of “fundamental separation” of archive and storage from access
- Interaction with and demonstration for users
- Technology discovery and evaluation of cutting edge tools for CLASS
- First integration of multiple data types through CLASS (time-series, grid, swath, etc..)
- Standards not technologies!

Technology Reviewed

- Thredds – The THREDDS Data Server (TDS) is a web server that provides metadata and data access for scientific datasets, building on and extending a number of existing technologies (metadata and data access for NetCDF, HDF5, GRIB, CDM, etc.)
- Rich Inventory – is a highly customizable metadata schema designed to minimize burden on the provider and maximize user search capabilities.
- NASA – EOS ClearingHOuse (ECHO) is a comprehensive data model based data search and access system.
- Unidata Common Data Model -The CDM is a unification of the data models of OpenDAP, netCDF, and HDF5
- OGSA-DAI The aim of the OGSA-DAI project is to develop middleware to assist with access and integration of data from separate sources via the grid
- Native XML Databases (xQuery) and Z39.50
- OGC webservices

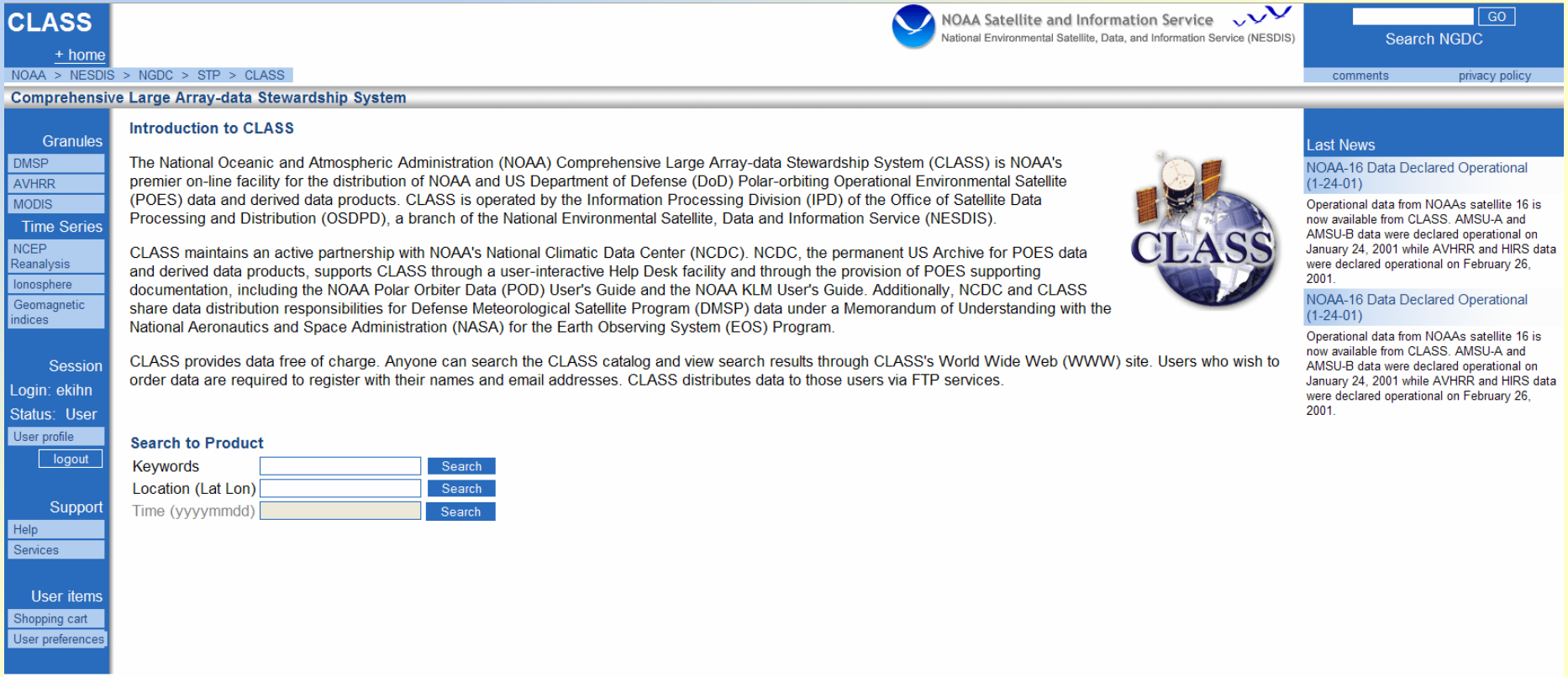
Web Services Advantages

- Interoperability
- Standards based
- Components loosely coupled
- Uses transports that are open e.g., HTTP
- Platform agnostic
- Language agnostic

CLASS Web-App

- Is a high-end web application built on the API
- Integrates managed, provider hosted data
- Integrates time-series, grids, swath data types
- Allows for load and performance testing

Catalog Metadata Search



Catalog Metadata Search

CLASS

+ home

NOAA > NESDIS > NGDC > STP > CLASS

NOAA Satellite and Information Service
National Environmental Satellite, Data, and Information Service (NESDIS)

Comprehensive Large Array-data Stewardship System

Granules
- AVHRR
- DMSP
- MODIS

Time Series
- NCEP Reanalysis
- Geomagnetic indices
- Ionosphere
- Sea Surface Temperature

Session
Login: ekihn
Status: User
- User profile
- logout

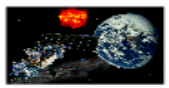
Support
- Help
- Services

User items
- Shopping cart
- User preferences

Search to product

Database	Region	Time Span

Defense Meteorological Satellite Program (DMSP)



The Earth Observation Group at NGDC is home to the Defense Meteorological Satellite Program (DMSP) Archive. In addition to maintaining the archive, the EOG performs research on the data as well as creating products.
The DMSP is a Department of Defense (DoD) program run by the Air Force Space and Missile Systems Center (SMC). The DMSP designs, builds, launches, and maintains satellites monitoring the meteorological, oceanographic, and solar-terrestrial physics environments.
Each DMSP satellite has a 101 minute, sun-synchronous near-polar orbit at an altitude of 830km above the surface of the earth. The visible and infrared sensors (OLS) collect images across a 3000km swath, providing global coverage twice per day.
The combination of day/night and dawn/dusk satellites allows monitoring of global information such as clouds every 6 hours. The microwave imager (MI) and sounders (T1, T2) cover one half the width of the visible and infrared swath.
These instruments cover polar regions at least twice and the equatorial region once per day. The space environment sensors (J4, M, IES) record along-track plasma densities, velocities, composition and drifts.
The data from the DMSP satellites are received and used at operational centers continuously.
The data are sent to the National Geophysical Data Center's Solar Terrestrial Physics Division Earth Observation Group (NGDC/STP/EOG) by the Air Force Weather Agency (AFWA) for creation of an archive.
Currently, data from 4 satellites (3 day/night, 1 dawn/dusk) are added to the archive each day.

90 / -180 / 180 / -90

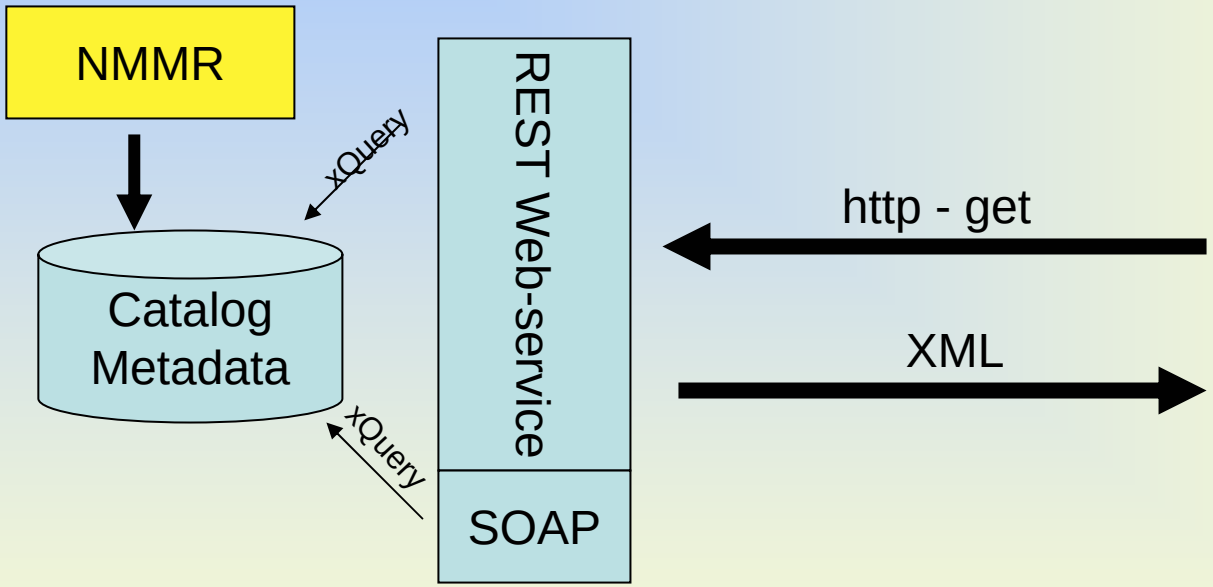
From: 1992-06-01T00:00:00Z00
To: 2006-12-31T23:59:00Z00

Go to product

Search to Product

Keywords [ ] Search
Location (Lat Lon) [ ] Search
Time (yyyymmdd) [ ] Search

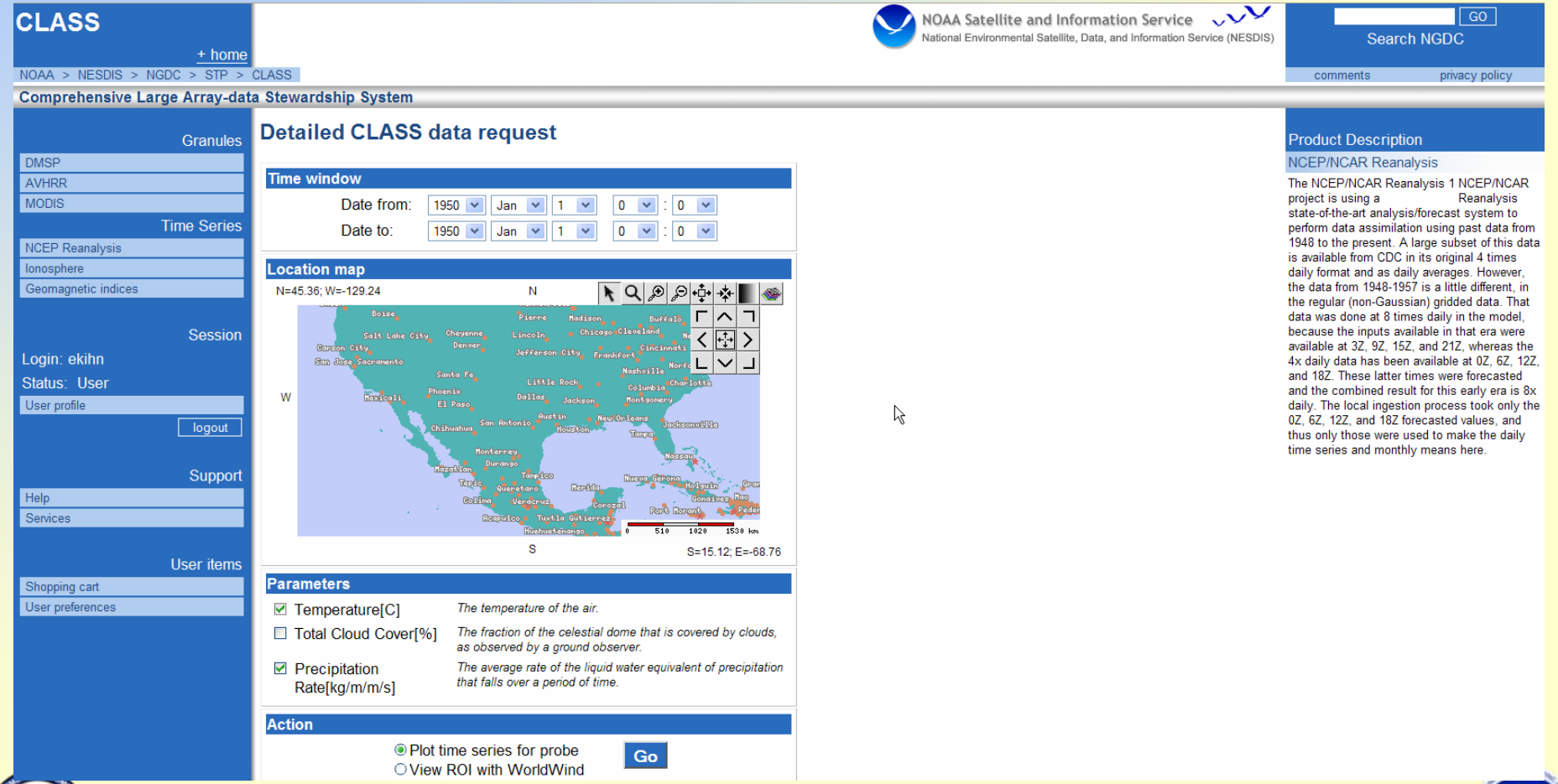
Catalog Metadata Search



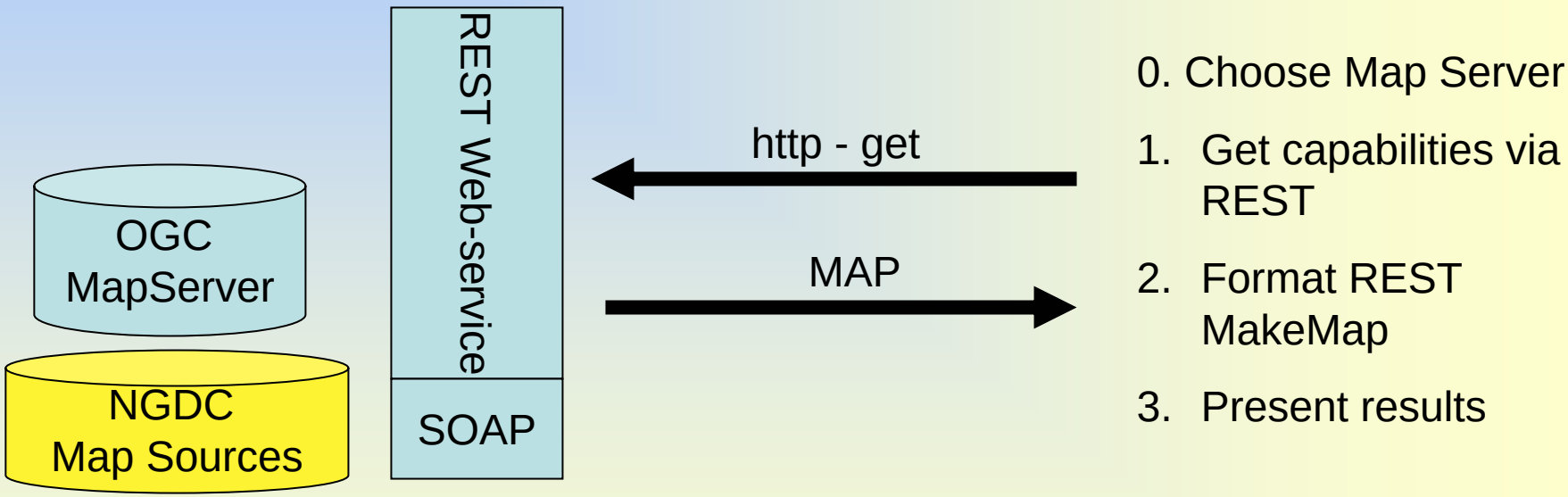
0. Choose collection level metadata set
1. Get capabilities via REST
2. Format REST Query
3. Parse results

- Store multiple metadata schema through native XML database
- Uses a standard protocol xQuery over SOAP/REST

CLASS Map Services



CLASS Map Services



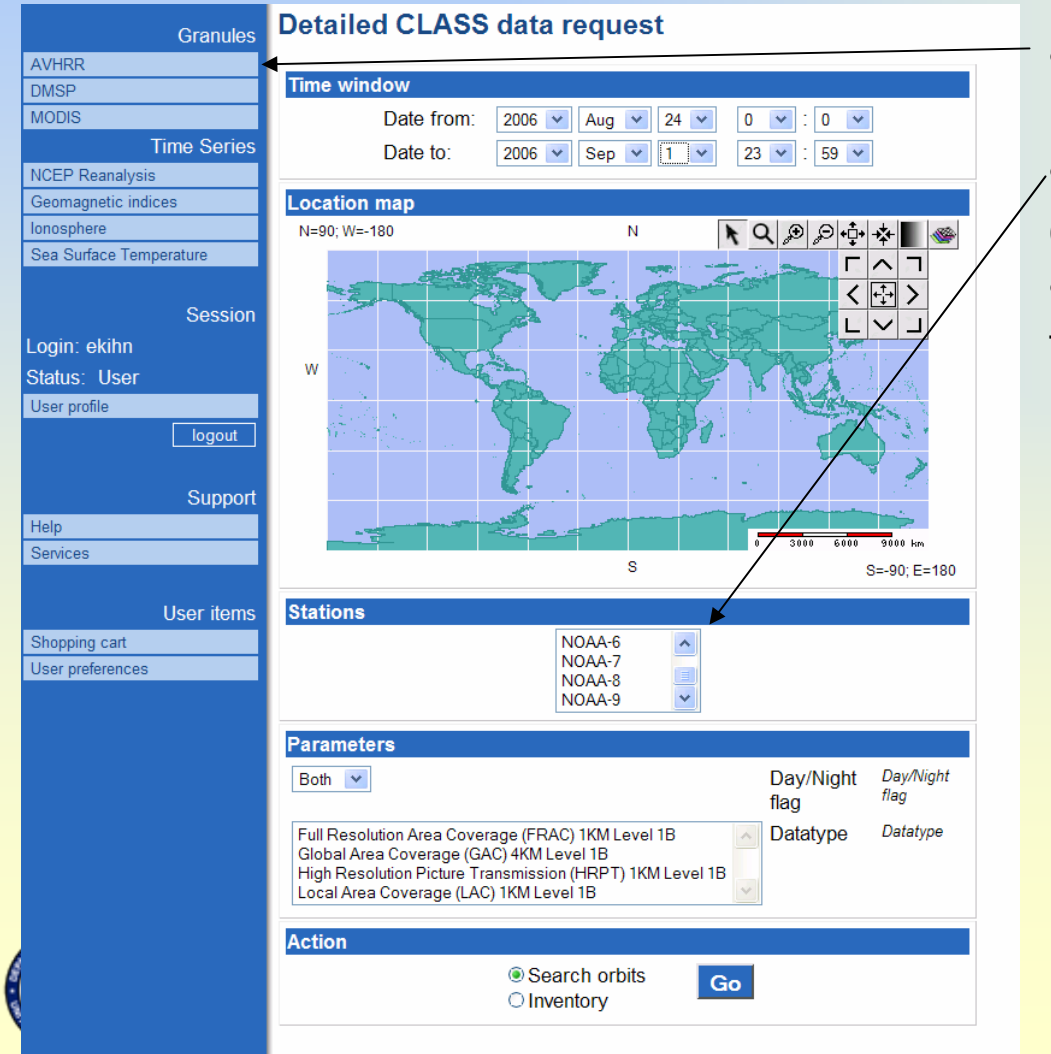
0. Choose Map Server
1. Get capabilities via REST
2. Format REST MakeMap
3. Present results

Very standard and accepted

Allows for cascading servers/services

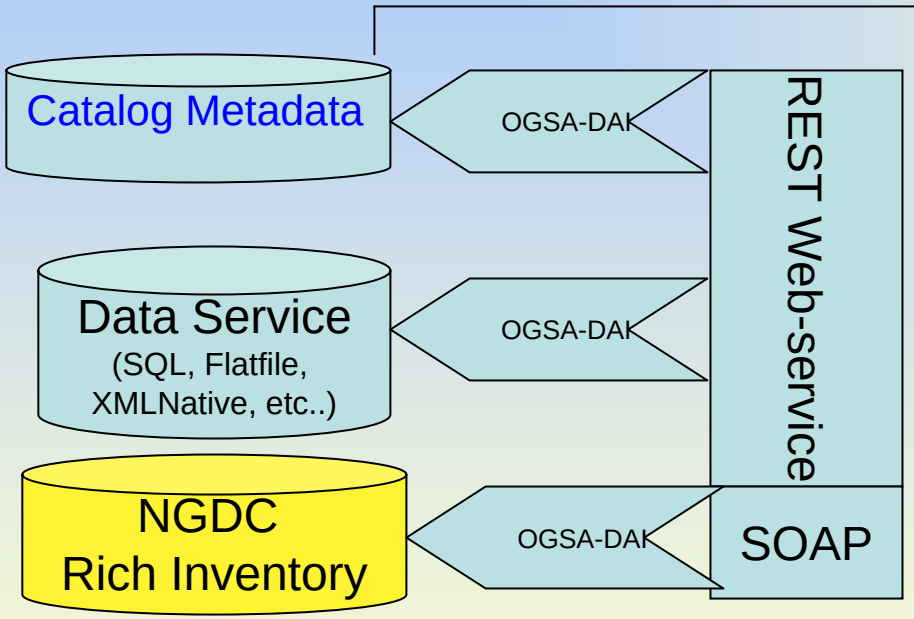
Provides standarized mapping services throughout CLASS

CLASS Granule Data



- Dynamically populated from metadata
- Order choices generated from catalog extension
- Order choices can include spatial, temporal, parameter based

CLASS Granule Data



0. Select data set
1. Get order info from catalog metadata
2. Format REST request to API
3. Parse list of entities
4. (optional) choose viewer and display
5. Select granules and pass to order service

CLASS Time-series

CLASS

+ home

NOAA > NESDIS > NGDC > STP > CLASS

NOAA Satellite and Information Service
National Environmental Satellite, Data, and Information Service (NESDIS)

Comprehensive Large Array-data Stewardship System

Granules
DMSP
MODIS
AVHRR
Time Series
NCEP Reanalysis
Geomagnetic indices
Ionosphere
Sea surface temperature
Session
Login: ekihn
Status: User
User profile
logout
Support
Help
Services
User items
Shopping cart
User preferences

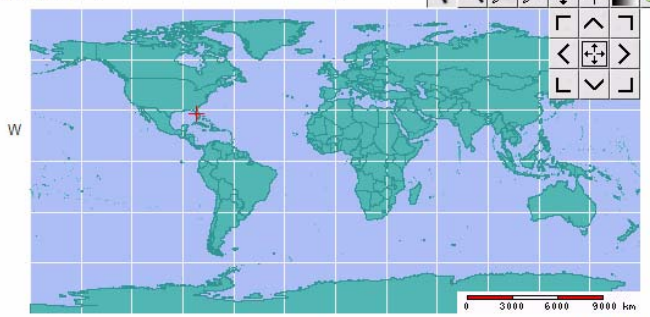
Detailed CLASS data request

Time window

Date from: 2001 Dec 24 0 : 0
Date to: 2001 Dec 31 23 : 59

Location map

N=90; W=-180 N



S S=-90; E=180

Parameters

- ☑ Temperature[C] — *The temperature of the air.*
- ☑ Total Cloud Cover[%] — *The fraction of the celestial dome that is covered by clouds, as observed by a ground observer.*
- ☑ Precipitation Rate[kg/m/m/s] — *The average rate of the liquid water equivalent of precipitation that falls over a period of time.*

Action

◉ View point
○ View ROI
○ Event search

Go

NOAA Satellite and Information Service
National Environmental Satellite, Data, and Information Service (NESDIS)

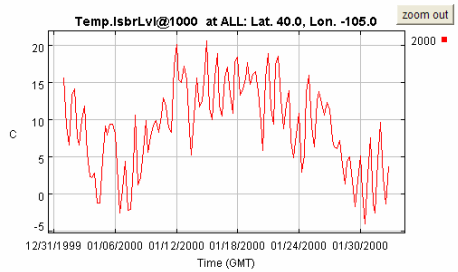
rdship System

Time Series Plots

Plot data vector is ready with size 2

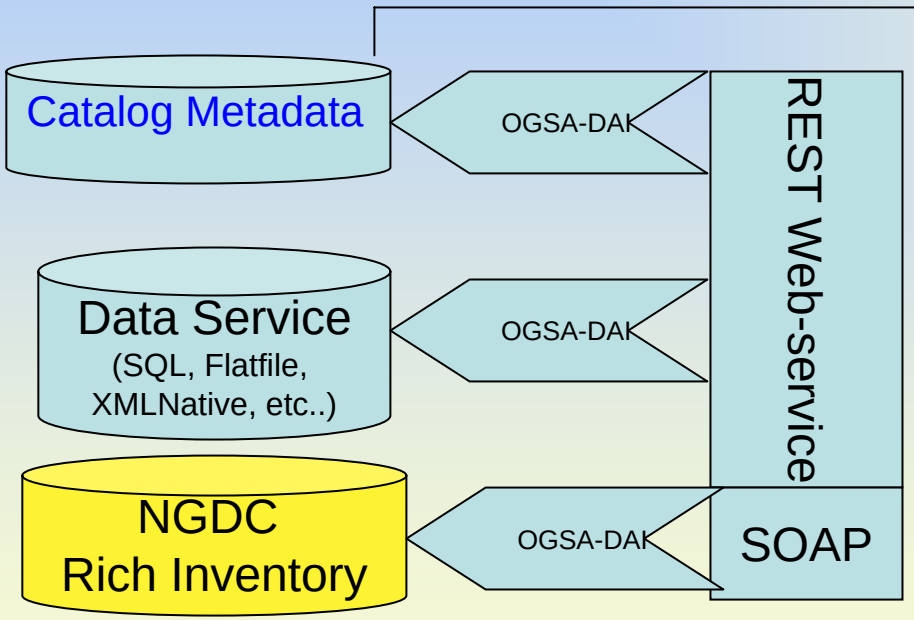
Temp.IsbrLvl@1000
PrecipRt.Surface@0
ALL: Lat. 40.0, Lon. -1

zoom out



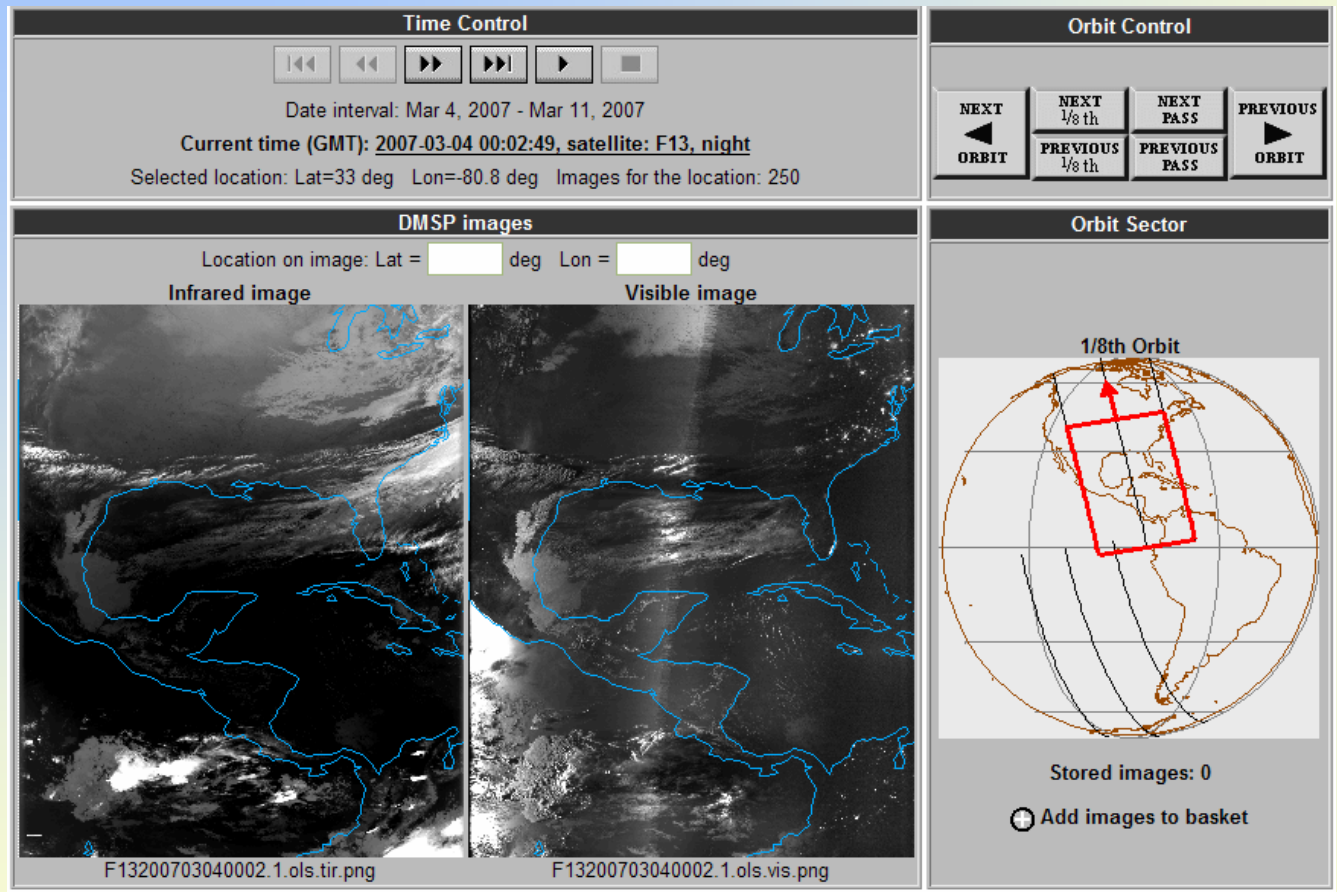
CLASS
A Component of NOAA Data Management

CLASS Time-series



0. Select data set
1. Get order info from catalog metadata
2. Format REST request to API
3. Results returned as a stream or file entity (URL)
4. Client acts directly on retrieved data

Visualization



- Catalog metadata provides service id
- Services described in WSDL
- Services for granules, time-series, grids

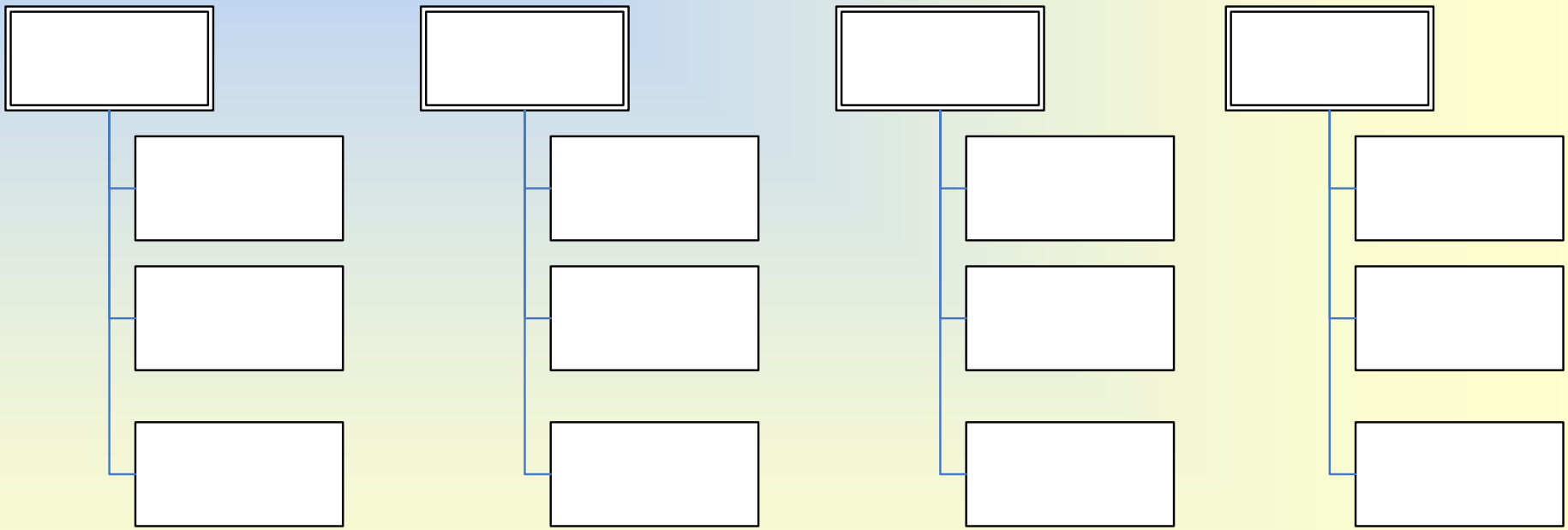
Services

- Services (such as models and transforms) fit the same general pattern as data ordering
- Each has a high level catalog entry describing output
- Each has a ordering extensions definign expected inputs and outputs

Delivery Options

- Time-series
  - Stream (XML)
  - NetCDF (Thredds)
  - File format (via service)
- Granule
  - File handle
  - Common Data Model (DMSP)
  - Subset/Transform (via service)

CLASS Pattern



Conclusion

- CLASS is a key component of NOAA's IT capability
- CLASS is an evolving operational system with long term - goals
- CLASS is a component of NOAA's data stewardship vision but only a piece
- CLASS issues will dominate NGDC's IT effort for the next several years

Backup Slides

Data Stewardship – Scientific Data Stewardship

- **(NOSC approved definition) Application of rigorous analyses and oversight to ensure that data sets meet the needs of users: Includes**
  - Documenting measurement practices and processing practices
  - Providing feedback on observing system performance
  - Inter-comparison of datasets for validation
  - Reprocessing (incorporate new data, apply new algorithms, perform bias corrections, integrate/blend data sets from different sources or observing systems)
  - Recommending corrective action for errant or non-optimal operations
- **Some unique attributes for each data set, but an integrated approach to data management**
- **Allowing opportunities to redirect the program based on advice and feedback**
- **Integrated suite of functions to preserve and exploit the full scientific value of NOAA's, and the world's, environmental data**
  - timely ingest
  - quality control processing
  - effective access to new and long-term records (data and metadata)
  - safeguarding of the climate records for future generations
- **Generation of authoritative long-term environmental records from multiple observing platforms**

Future Integration Plans

CLASS provide archive storage

- Data storage for all Data Center holdings & future growth
- Leading-edge technology
- High-performance hardware (disk and tape robotics)
- End-to-end hardware solution as approved by Data Centers
- OAIS-compliant
- File structures and metadata to support existing & future Data Center systems
- Support data mining and complex queries

Data Centers/NVDS provide ingest, archive, and access services

- Ingest, quality control, and archive services
- Development of integrated and blended (cross-platform) datasets
- Integrated, open data access
- Web interfaces for integrated data access
- Model, in-situ, radar, satellite data support
- Leverage existing infrastructure + new technology
- OAIS-compliant
- GIS services, Web services, OpenDAP, etc
- Partner locally (eg, AFCCC), regionally (eg, RENCI), and nationally (eg, CUAHSI, DLESE)

Outside World

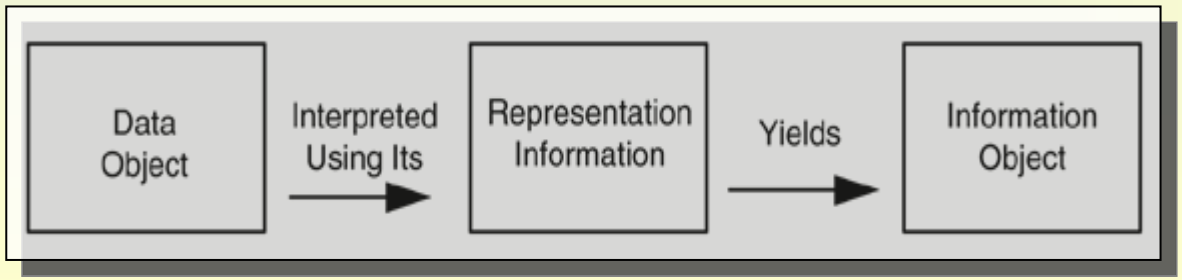
- **Users**
  - Government
  - Private
  - Public
  - Academia
- **Feedback (workshops, advisory groups, via customer service reps)**
- **Constituents**

Defining NOAA National Data Center Roles and Responsibilities: Archive Required Activities

- **To ensure information preservation an archive* must:**

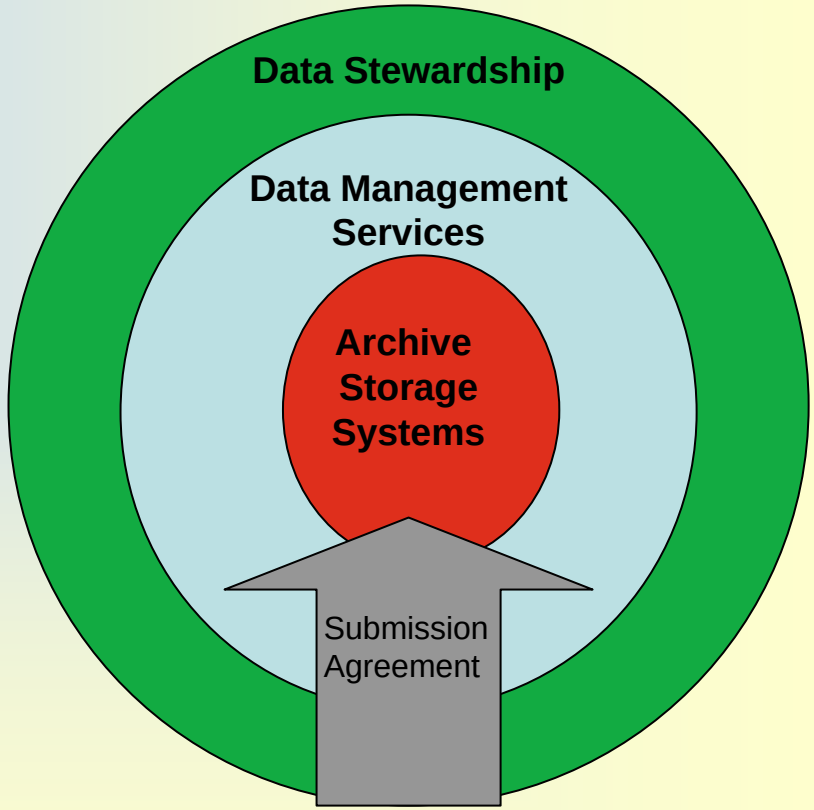
1. Negotiate and accept information from information producers
2. Obtain sufficient control to ensure long-term preservation
3. Determine which communities (designated) need to be able to understand the preserved information
4. Ensure the information to be preserved is independently understandable to the Designated Communities
5. Follow documented policies and procedures that ensure the information is preserved against all reasonable contingencies
6. Make the preserved information available to the Designated Communities in forms understandable to those communities

***Open Archive Information System – Reference Model**



1. Negotiate and accept information from information producers – Submission Agreement => Data Centers Lead

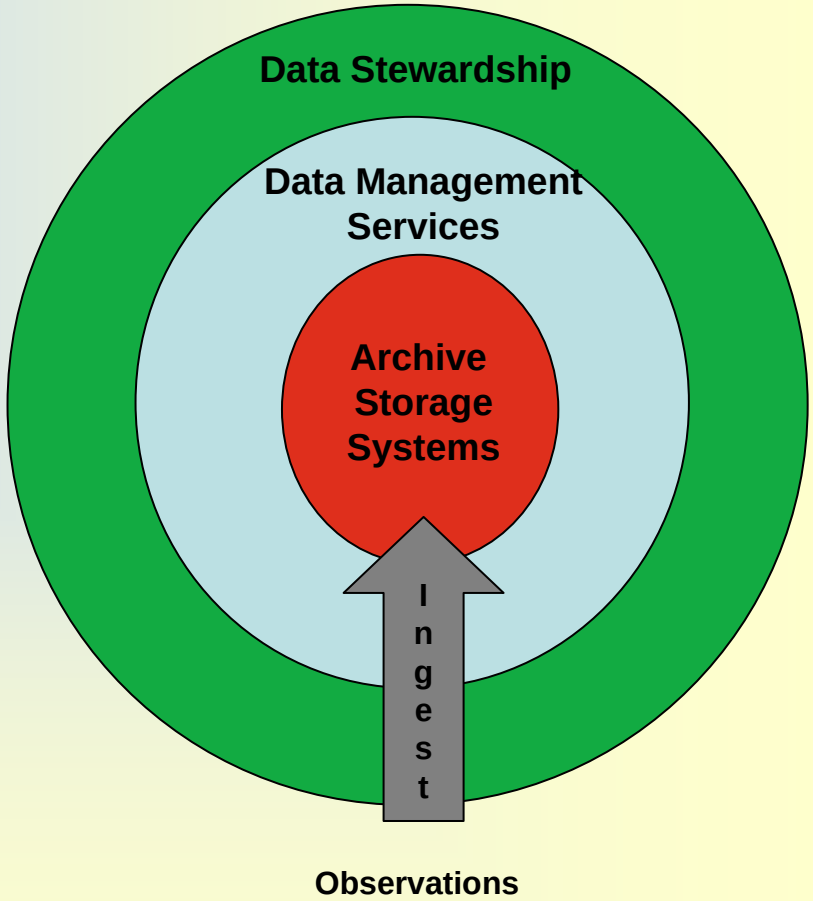
- Ensures that science requirements and other user applications are clearly defined with respect to NOAA's archive, access, processing, and reprocessing stewardship activities
- Provides requirements and prioritization for preserving and maintaining the basic storage of and access to critical data sets and derived products and their documentation, including verifying their quality and compliance with federal standards
- Assists in establishing requirements for the IT aspects (including security) for implementation of scientific data stewardship
- Science advisory panel/group input



Current and Future Observations Consolidated Observation Requirements List (CORL)

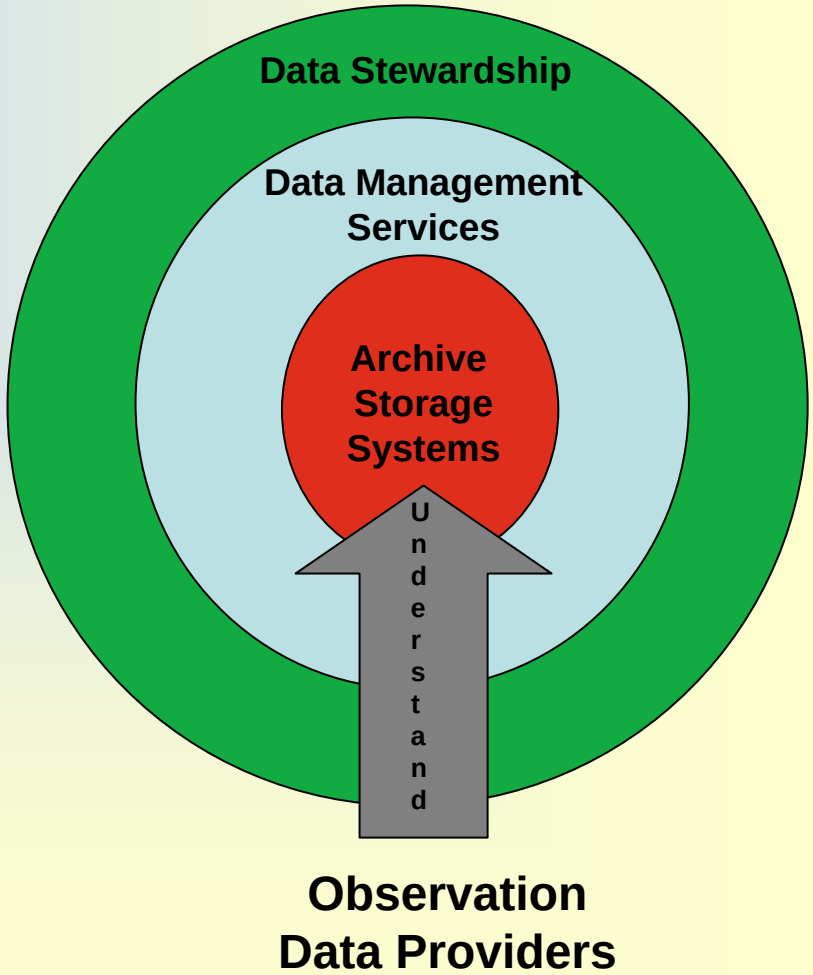
2. Obtain sufficient control to ensure long-term preservation => CLASS leads

- The submission agreement will specify how data are acquired including delivery schedule and any copyright and redistribution arrangements



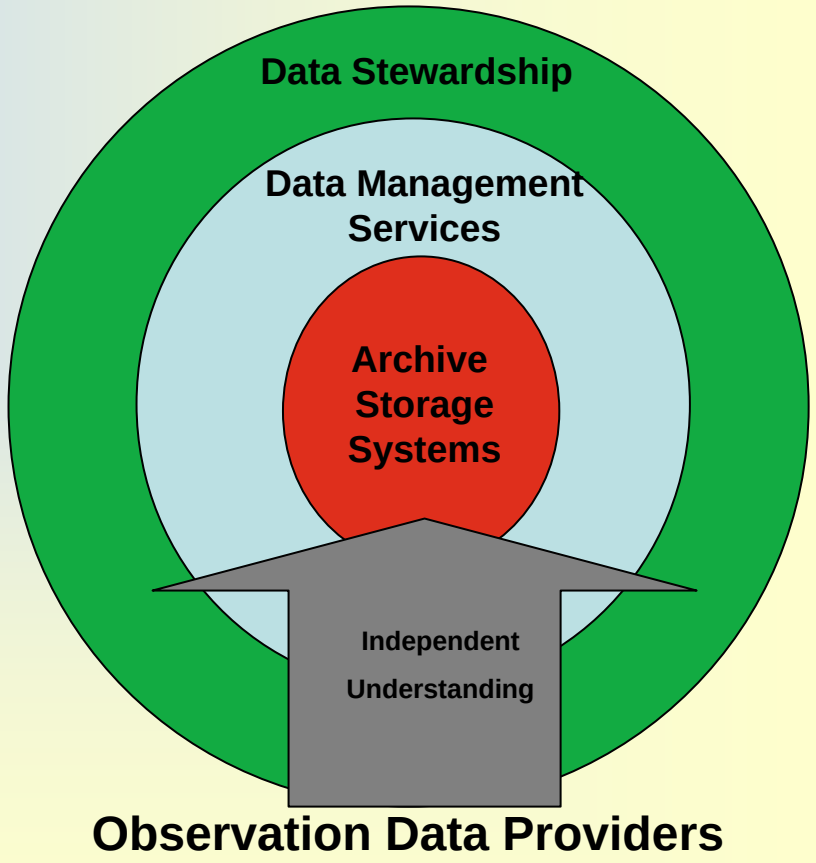
3. Determine which communities (designated) need to be able to understand the preserved information => Data Centers+ Lead

- Data management services determine what is required to preserve fundamental (raw) records
- Data stewardship determines what is required to preserve thematic (derived product) records
- Roles and responsibilities must be defined as part of the submission agreement



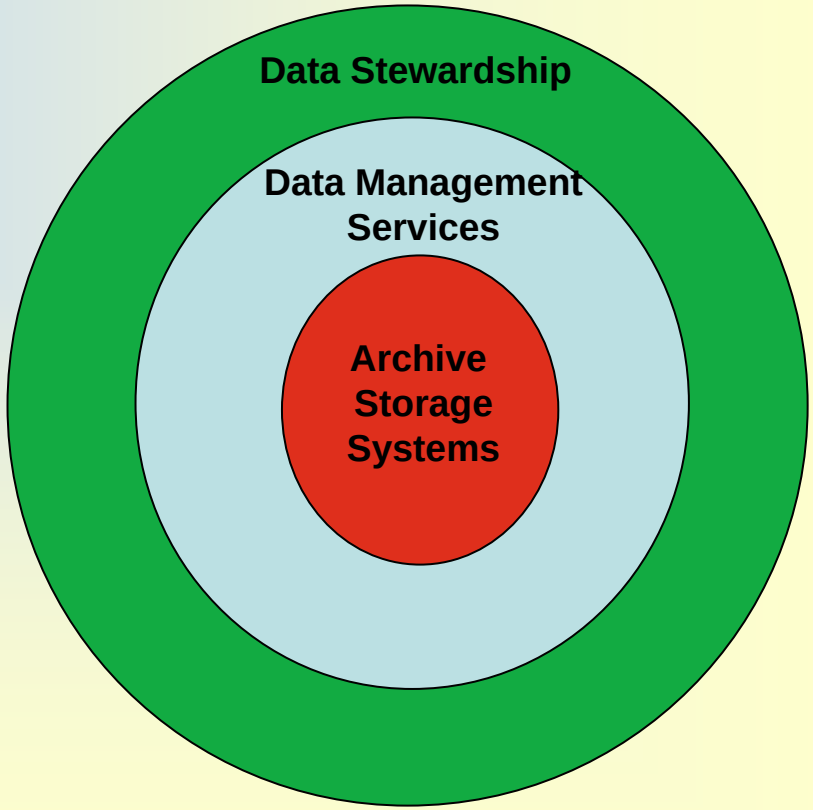
4. Ensure the information to be preserved is independently understandable to theDesignated Communities – Data Centers+ Lead

- Data Centers and Centers of Data ensure capture of enough information from data providers to ensure designated user communities can independently understand data
- This focuses on 4 metadata types – reference, context, provenance, fixity (data integrity)
- Issues



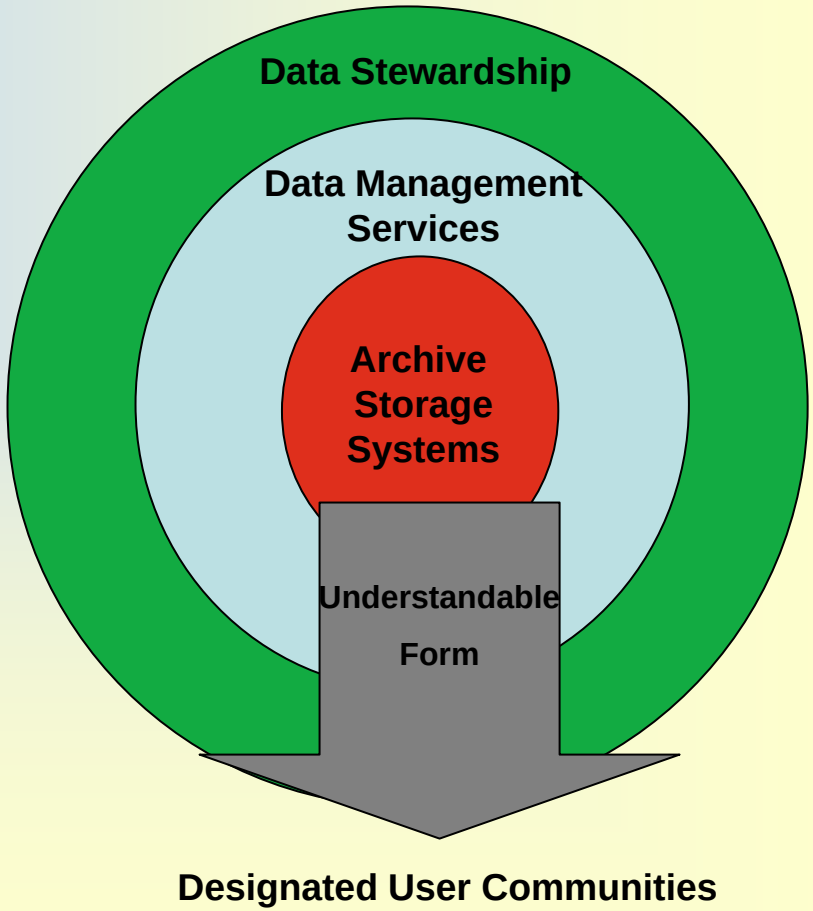
5. Follow documented policies and procedures that ensure the information is preserved against all reasonable contingencies

- CLASS ensures IT hardware and software
- Data Centers and Centers of Data ensure needed metadata on raw data and observing systems
- Data Stewards (data centers and beyond) ensure metadata on provenance and all higher products



6. Make the preserved information available to the Designated Communities in forms understandable to those communities

- Users require data and metadata be handed to them in easy to use forms
- This varies by user community
- Issues
  - Service Oriented Architecture

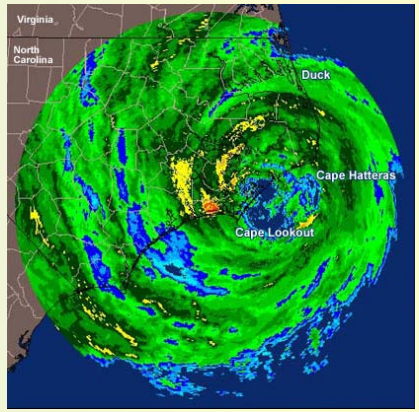


Data Centers

NOAA Virtual Data System (NVDS) Components

Data Access

- "HDSS Access System" (large volume data—satellite, radar)
- "Climate Data Online" (in-situ data)
- Geophysical and Oceanographic Data
- "NOMADS" model data access and CLASS satellite data access to be integrated with NVDS



Online Products and Services

- GIS Services
- Web Services
- Data Visualization (NEXRAD, etc)
- NetCDF, OpenDAP, etc

Data Ordering and E-business

- NESDIS E-government System (Online Store, off-line orders, financial transactions, business statistics, etc)

NVDS Integration with CLASS -- Successes

- NVDS provides GOES access for CLASS (work completed in 2003, continued enhancements)

- NVDS/NES provides satellite data ordering capabilities (off-line media) for CLASS (work completed in 2006)