NOAA'S Future Data Activities:
Petabyte Archives, Metadata and Systems Integration

David Clark
NOAA/NESDIS/ National Geophysical Data Center
20th International CODATA Conference
Beijing, P. R. China

What is the future?

- Petabyte Archives
  - Comprehensive Large Array-data Stewardship System (CLASS)
- Metadata
  - Systems interoperability
- Integrated NOAA Observing systems
  - Global Earth Observation Integrated Data Environment (GEO IDE)

"More information has been produced in the last 30 years than in the last 5000" Pritchett, 1999

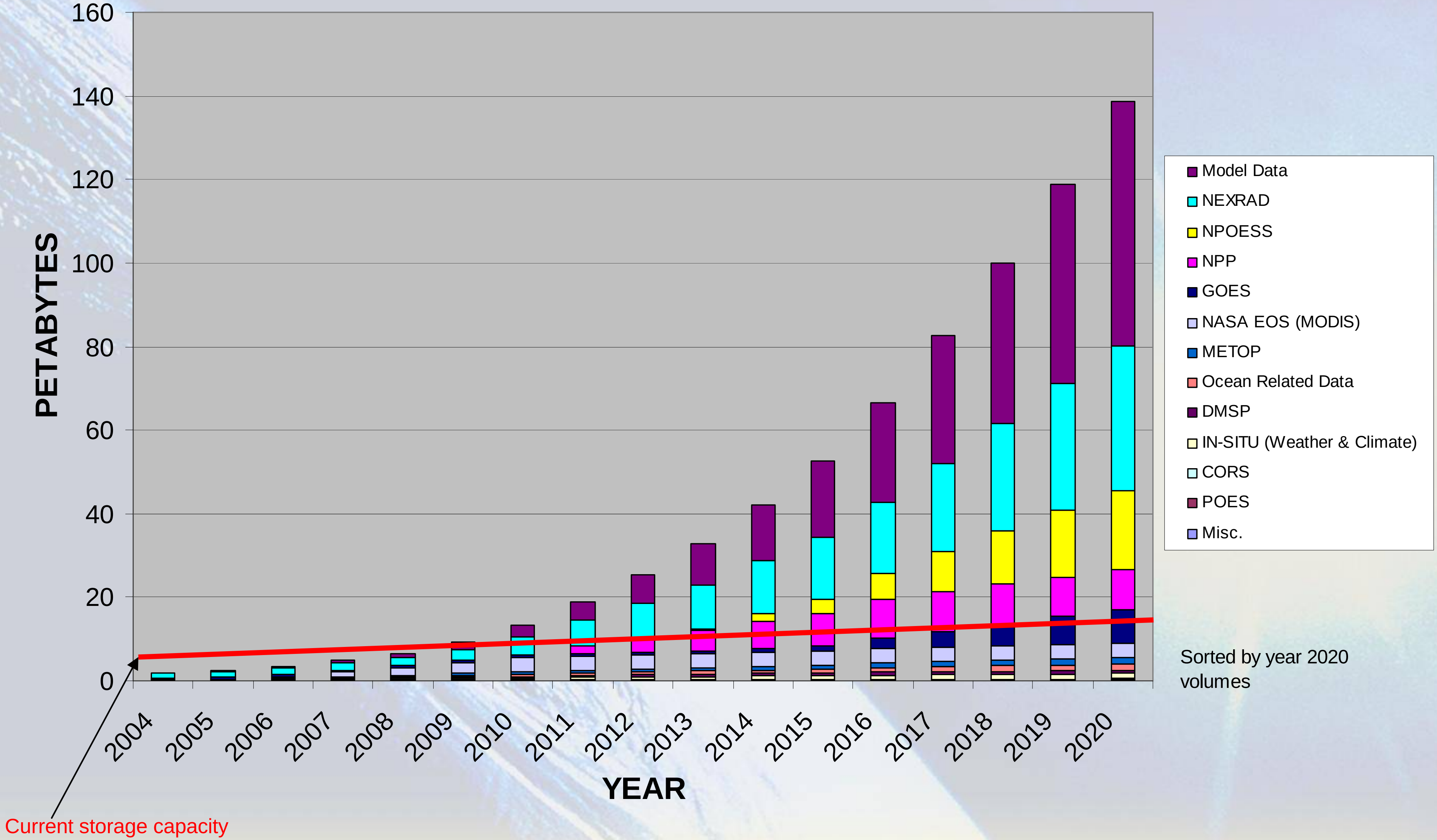
"Data is everyone's second highest priority"

Bretherton, *circa* 1988

A Petabyte Equals

- 1,000 Terabytes
- 1 million Gigabytes
- 500 billion ASCII pages
- 32,000 mile-high stack of paper
- 5 Billion pounds of paper
- 42.5 million pulp trees
- 12,000 football fields of file cabinets
- 5,500 years to download at 56 kbps

NOAA Data Archive Volume Projections



Comprehensive Large Array-data Stewardship System (CLASS)

Mission Statement

"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 NOAA's current and future environmental data sets will reside. CLASS provides permanent, secure storage, and safe, efficient data discovery and access between the Data Centers and the customers."

CLASS Goals

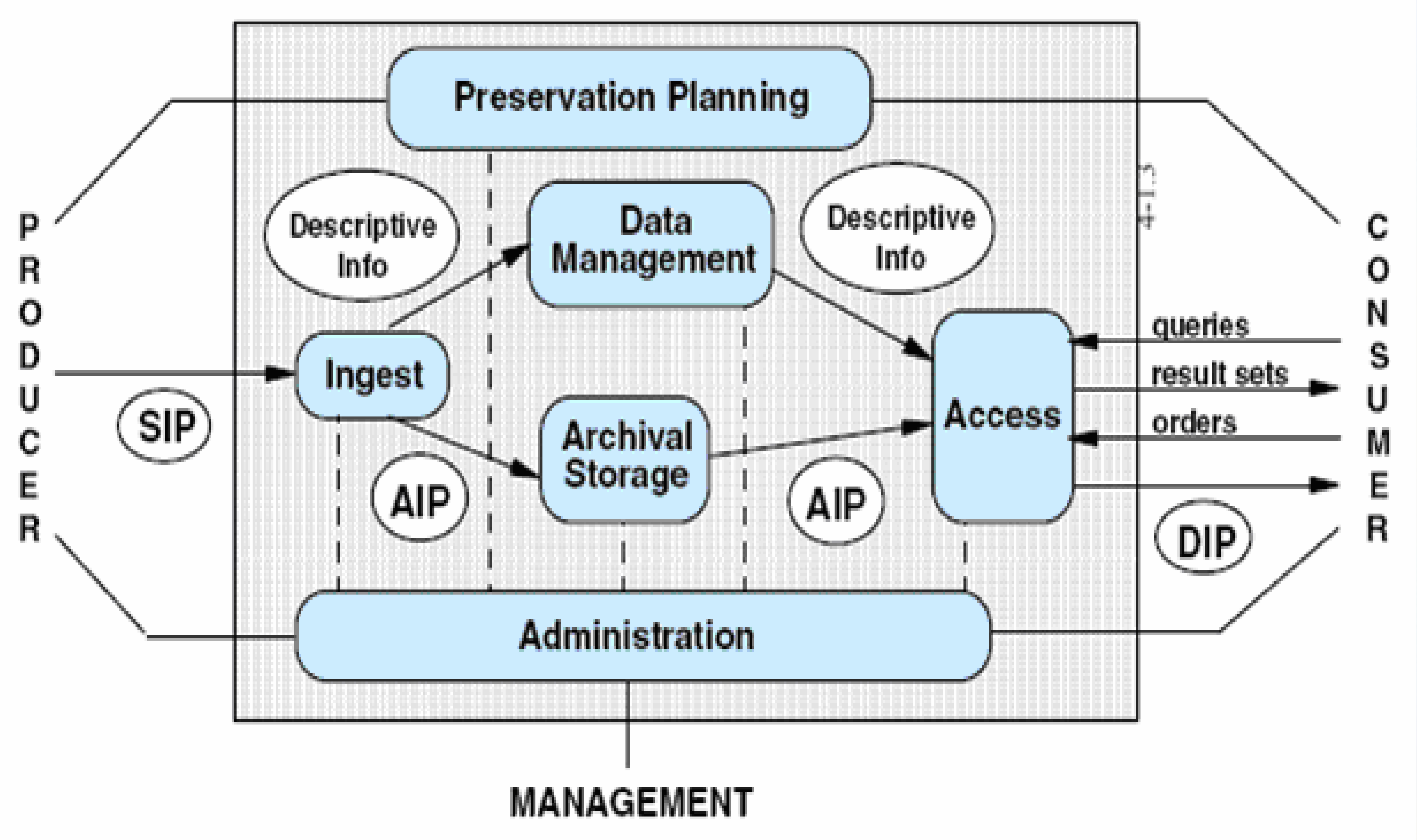
- Provide one-stop shopping and access capability for NOAA environmental data and products
- Provide a common look and feel for accessing NOAA environmental data and products
- Provide an efficient architecture for archiving and distribution of NOAA environmental data and products
- Reduce implementation costs by using reengineering, evolutionary effort
- Allow NOAA to fulfill its requirements regarding archive, access, and distribution of data from NOAA and other observing systems

CLASS Performance Requirements

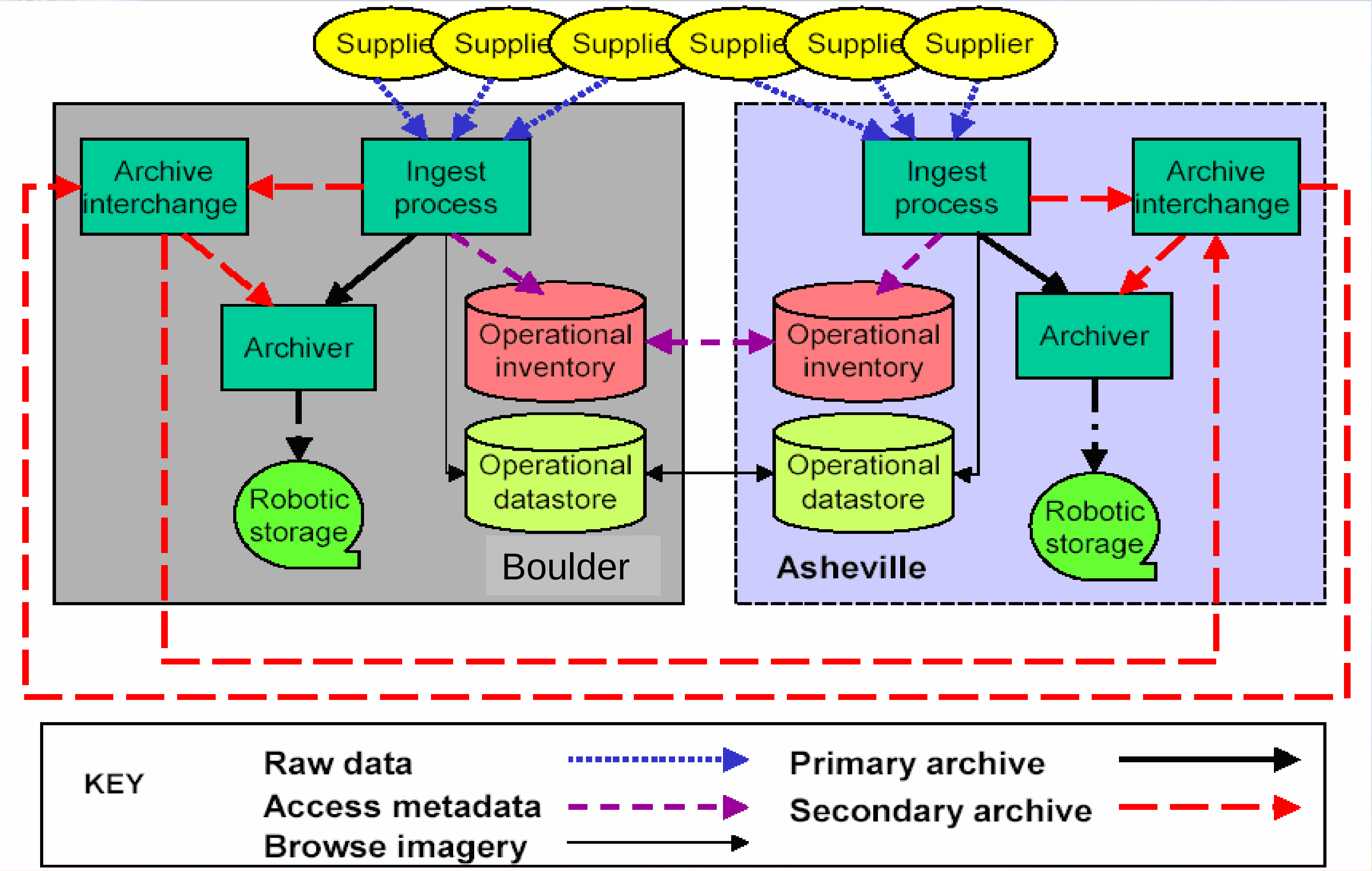
– Core Requirements
  - ingest, secure storage, and access to baseline large-array data
  - information pertaining to processing data, including documentation, processing algorithms and procedures
  - provide human and machine-to-machine interfaces to store, maintain, and provide access to data, information, and metadata
  - initiate pilot programs with the GEO IDE to support risk reducing development and phased integration of standards for metadata, machine-to-machine interfaces, and archive

CLASS Architecture

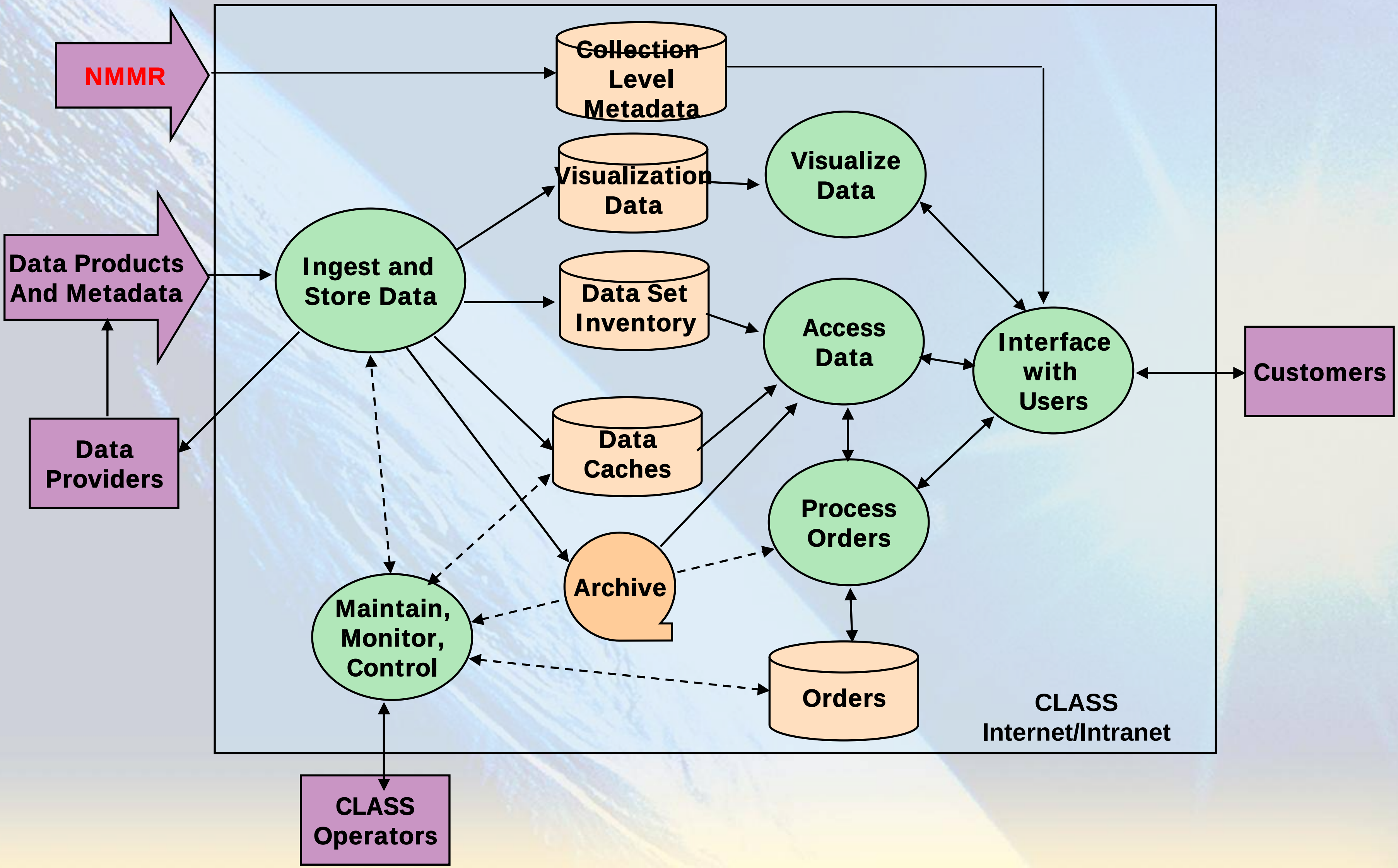
OAIS Functional Entities Ingest, Archive , Access & Data Management



CLASS Overview – Distributed Redundant Archive

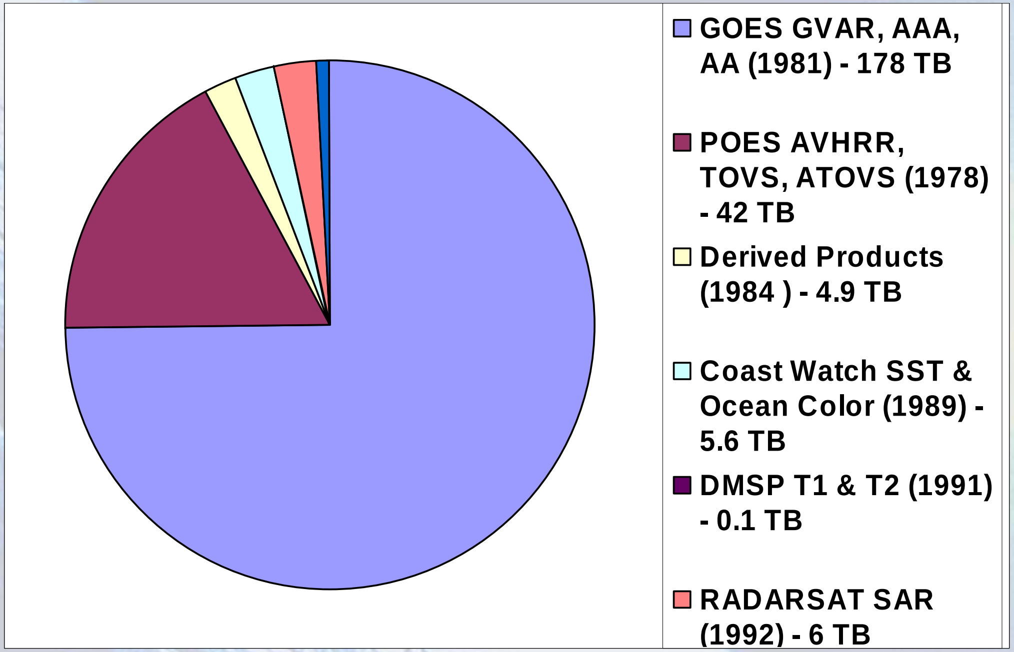


CLASS System Overview



Current Capability

CLASS maintains long-term, secure storage of and access to 238TB of environmental data growing at 0.78 TB/week



384 TB redundant Storage Area Network & 2 PB Tape Robotics

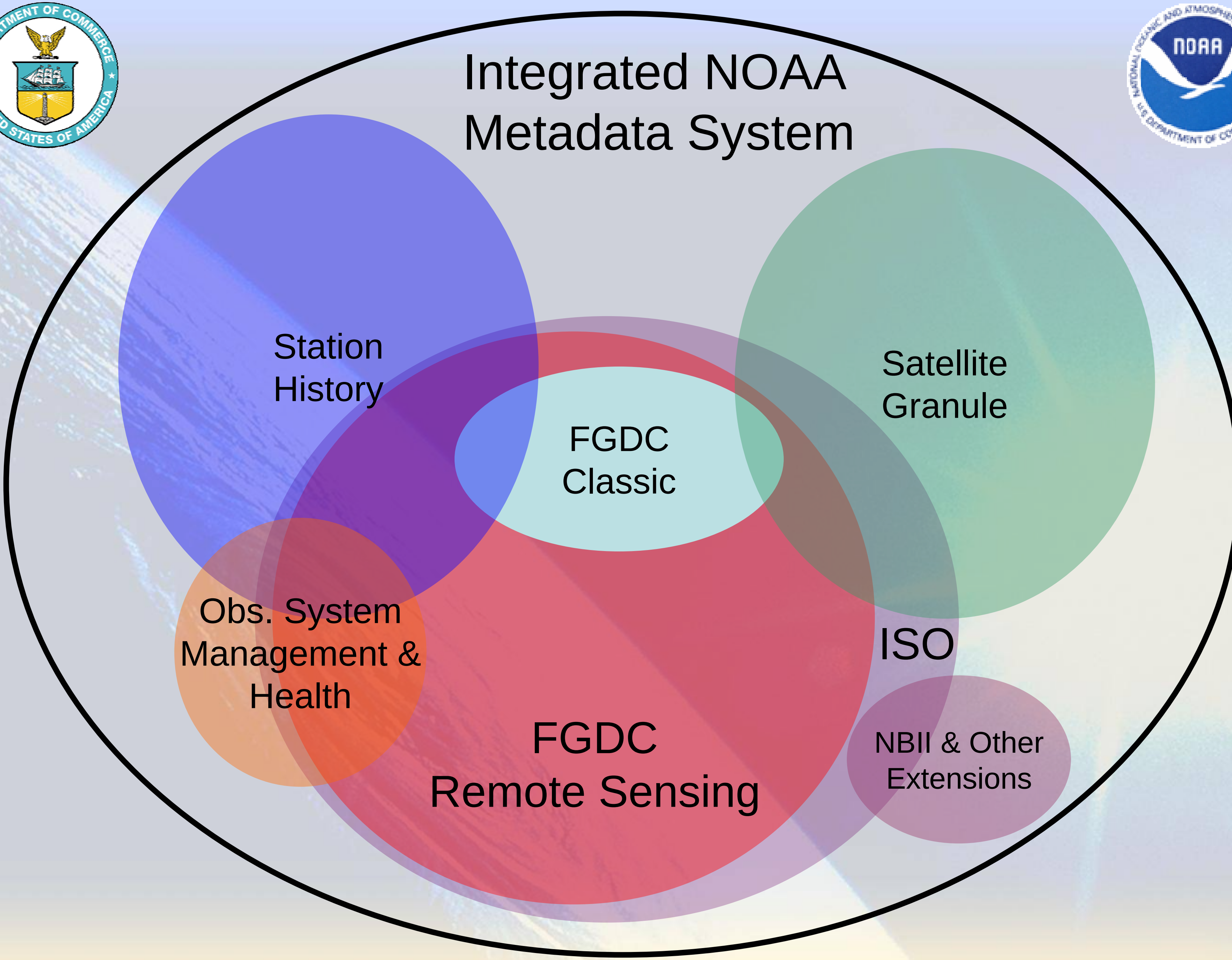
Metadata (Greek *meta* "after" and Latin *data* "information") are data that describe other data. Generally, a set of metadata describes a single set of data, called a resource.

from Wikipedia

NOAA Metadata Manager and Repository (NMMR)

- Supports multiple metadata standards
- Web, SOAP, and search interfaces
- Creation of metadata, with minimal understanding of FGDC standards
- Supports workflow with multiple states
- Collection/granule (parent/child) record sets
- Direct path to conforming to ISO 19115/19139

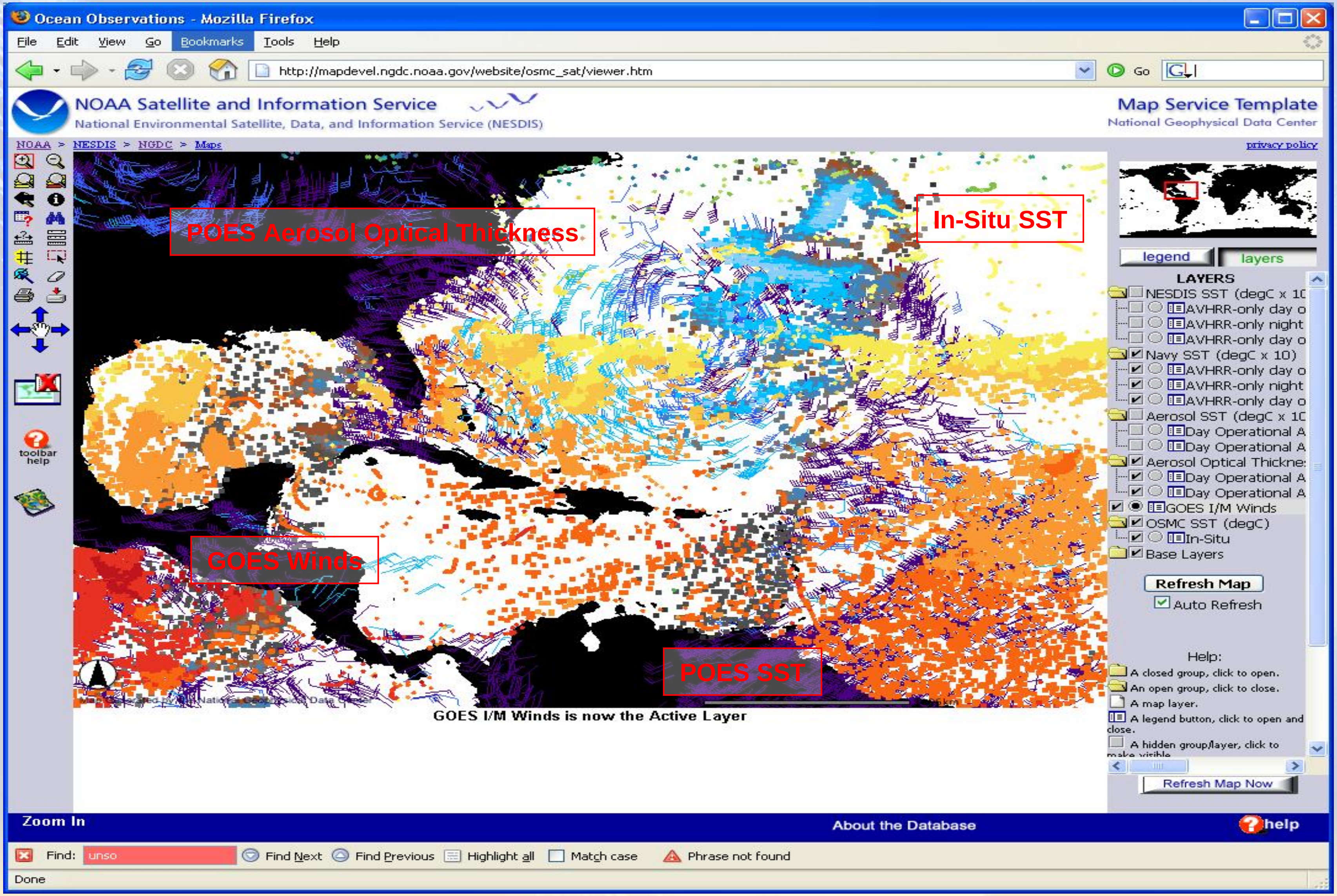
Integrated NOAA Metadata System



Why Metadata?

- Adherence to metadata standards
  - Leads to easier integration of data
  - More resources can be spent on development of data relationships than reformatting and manipulation of the data
  - Much more efficient archival and access to retrospective data
  - Leads to the integration of operational (real/near real-time data systems) and archive data systems.

Integrated Data Systems

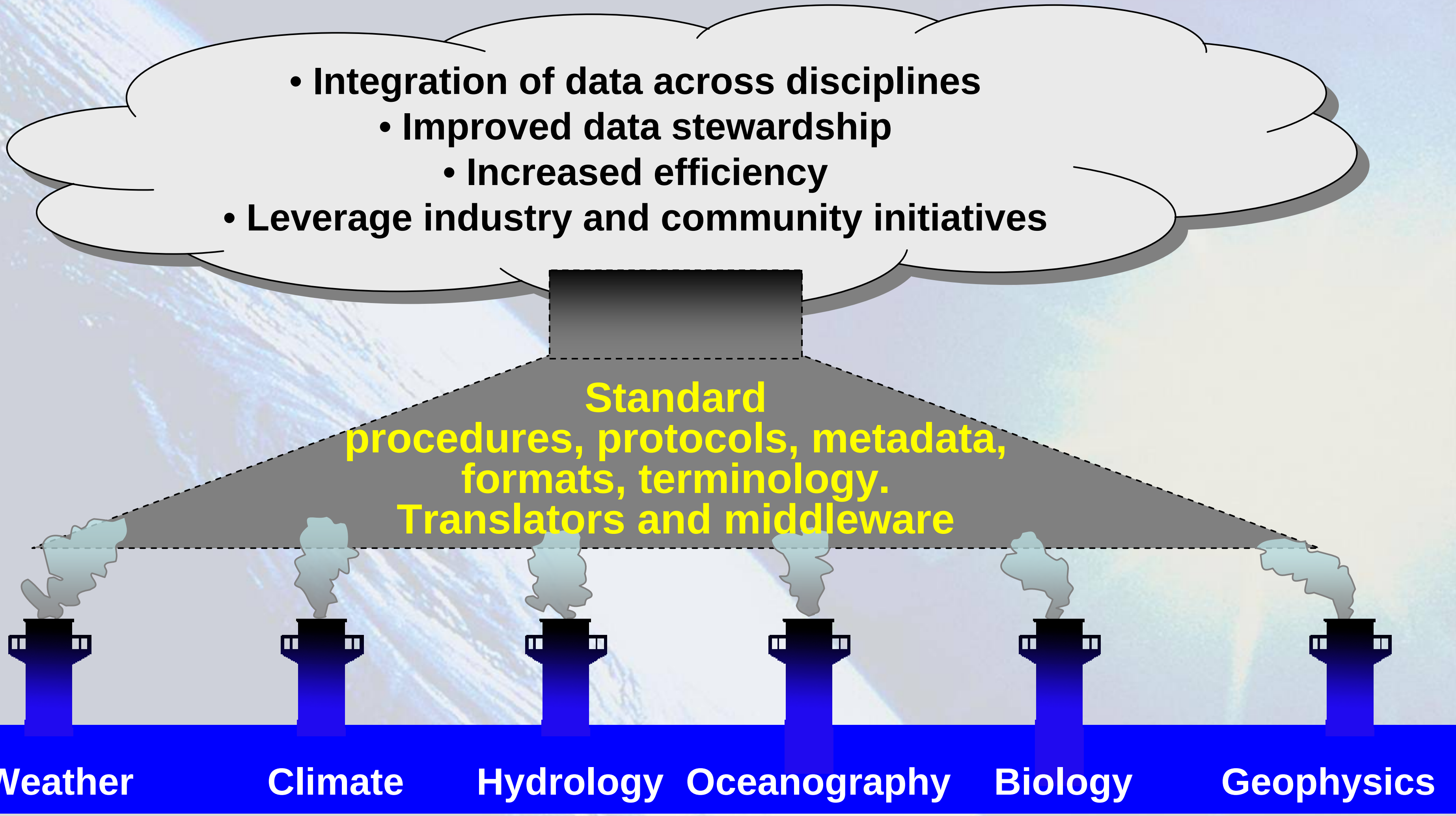


NOAA Encompasses a Challenging Diversity

- NOAA currently manages >90 environmental observing systems, some with hundreds of stations: including land-, sea-, air, and space-based observing platforms
- These systems gather >300 diverse environmental parameters (e.g. marine biological health, economic fisheries data, physical and chemical state of the atmosphere and ocean, paleoclimate proxy data, geodetic survey points, etc.)
- NOAA also requires other national, international and commercial data in its operations (some in real-time)
- NOAA data management systems include more than 50 significant stovepipe systems
- Future observing systems will produce vastly increased data volumes that will need to be archived and efficiently accessed by an expanding number of users
- NOAA is migrating from this current stovepipe environment to an information enterprise

Integrated Data Environment

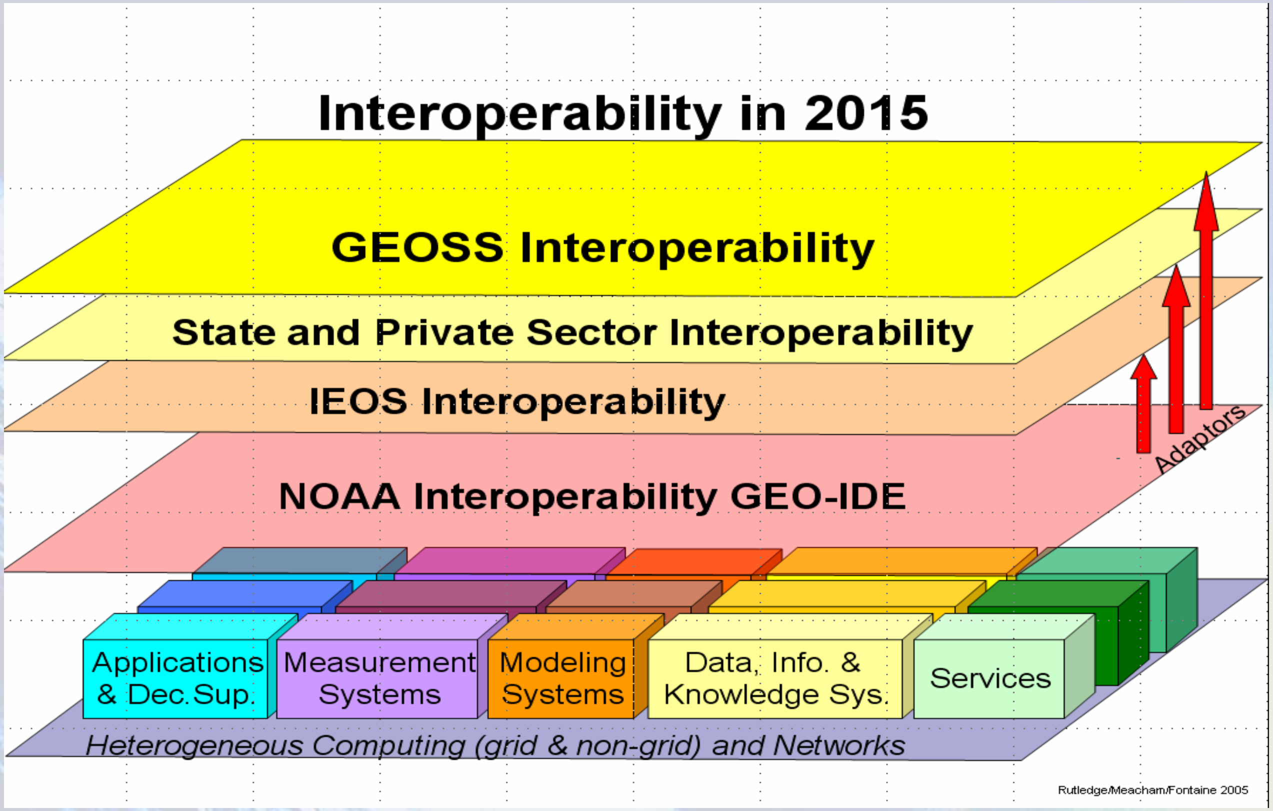
Bridging the gaps between stove-pipe systems



Response - NOAA's GEO-IDE

- **Scope** – NOAA-wide architecture development to integrate legacy systems and guide development of future NOAA environmental data management systems
- **Vision** – NOAA's GEO-IDE is envisioned as a "system of systems" – a framework that provides effective and efficient integration of NOAA's many quasi-independent systems
- **Foundation** – built upon agreed standards, principles and guidelines
- **Approach** – evolution of existing systems into a service-oriented architecture
- **Result** – a single system of systems (user perspective) to access the data sets needed to address significant societal questions

Vision



- “System of systems” – a framework to effectively and efficiently integrate NOAA’s many systems
- Minimize impact on legacy systems
- Utilize standards
- Work towards a service-oriented architecture

ArcIMS Map with ~100 Data Layers

National Ocean Service (10)

CCAP
CO-OPS
CORS
NCOP
NST
NST Mussel
NWLON
PORTS
SWMP
CREIOS

National Marine Fisheries Service(3)

National Observer Program
Habitat Assessment
MRFSS

National Weather Service(15)

ASOS
BOY
COOP
DART
FNP
HMISC
MAN
MDCRS
METXX
NERON
NEXRAD
Profiling Network
Rawinsonde
Region Networks
VOS

NOAA Research(50)

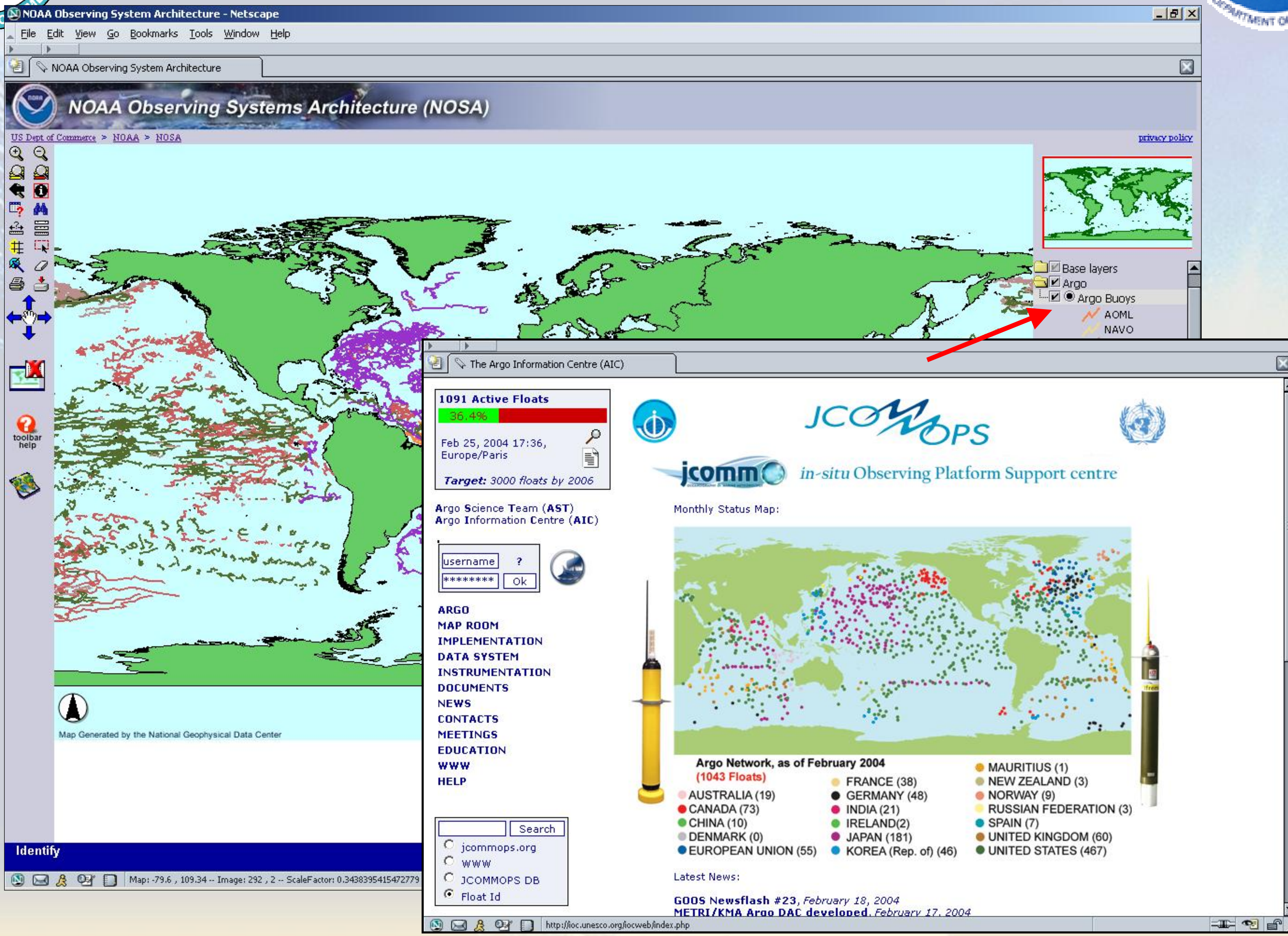
ISIS, SURFRAD, AIRMoN, ETOS, RAMAN, AERO, CCGG, DOBSON, HATS, STAR, AOC, BAO, GRIDS, HRDL, MOPA, OPAL, RASS, RADAR, TARS, SODAR, Teaco, GSLN, STRATUS, TAO, FOCI, Hyrdophones, Wind Profiler, Ships of Opportunity, Water Vapor Dial...

NESDIS(7) GOESWINDS, DMSP, IONOSONDE, MOBY, QUICKSCAT, USCRN...

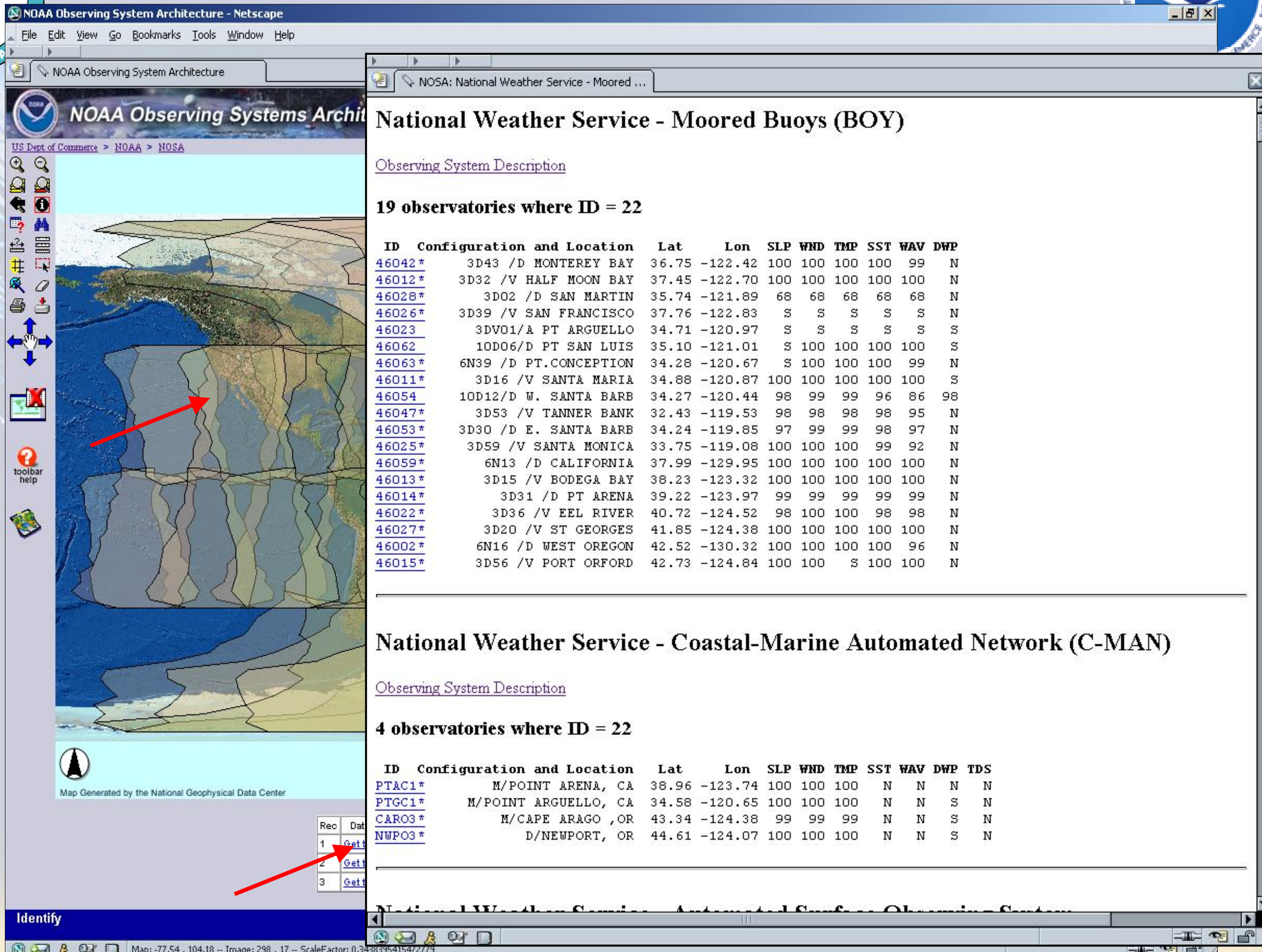
Other...(9)

GODAE, GHCN, GSN, GUAN, Fluxnet, AERONET, RAWS, WCRP-BSRN, WOUDC

ArcIMS Site and Metadata Links



Integrated Satellite and In-Situ Data Access



National Weather Service - Moored Buoys (BOY)

Observing System Description

19 observatories where ID = 22

ID	Configuration and Location	Lat	Lon	SLP	WND	TMP	SST	WAV	DWP
46042*	3D43 /D MONTEREY BAY	36.75	-122.42	100	100	100	100	99	N
46012*	3D32 /V HALF MOON BAY	37.45	-122.70	100	100	100	100	100	N
46028*	3D02 /D SAN MARTIN	35.74	-121.89	68	68	68	68	68	N
46026*	3D39 /V SAN FRANCISCO	37.76	-122.83	S	S	S	S	S	N
46023	3DV01/A PT ARGUELLO	34.71	-120.97	S	S	S	S	S	S
46062	10D06/D PT SAN LUIS	35.10	-121.01	S	100	100	100	100	S
46063*	6N39 /D PT.CONCEPTION	34.28	-120.67	S	100	100	100	99	N
46011*	3D16 /V SANTA MARIA	34.88	-120.87	100	100	100	100	100	S
46054	10D12/D W. SANTA BARB	34.27	-120.44	98	99	99	96	86	98
46047*	3D53 /V TANNER BANK	32.43	-119.53	98	98	98	98	95	N
46053*	3D30 /D E. SANTA BARB	34.24	-119.85	97	99	99	98	97	N
46025*	3D59 /V SANTA MONICA	33.75	-119.08	100	100	100	99	92	N
46059*	6N13 /D CALIFORNIA	37.99	-129.95	100	100	100	100	100	N
46013*	3D15 /V BODEGA BAY	38.23	-123.32	100	100	100	100	100	N
46014*	3D31 /D PT ARENA	39.22	-123.97	99	99	99	99	99	N
46022*	3D36 /V EEL RIVER	40.72	-124.52	98	100	100	98	98	N
46027*	3D20 /V ST GEORGES	41.85	-124.38	100	100	100	100	100	N
46002*	6N16 /D WEST OREGON	42.52	-130.32	100	100	100	100	96	N
46015*	3D56 /V PORT ORFORD	42.73	-124.84	100	100	S	100	100	N

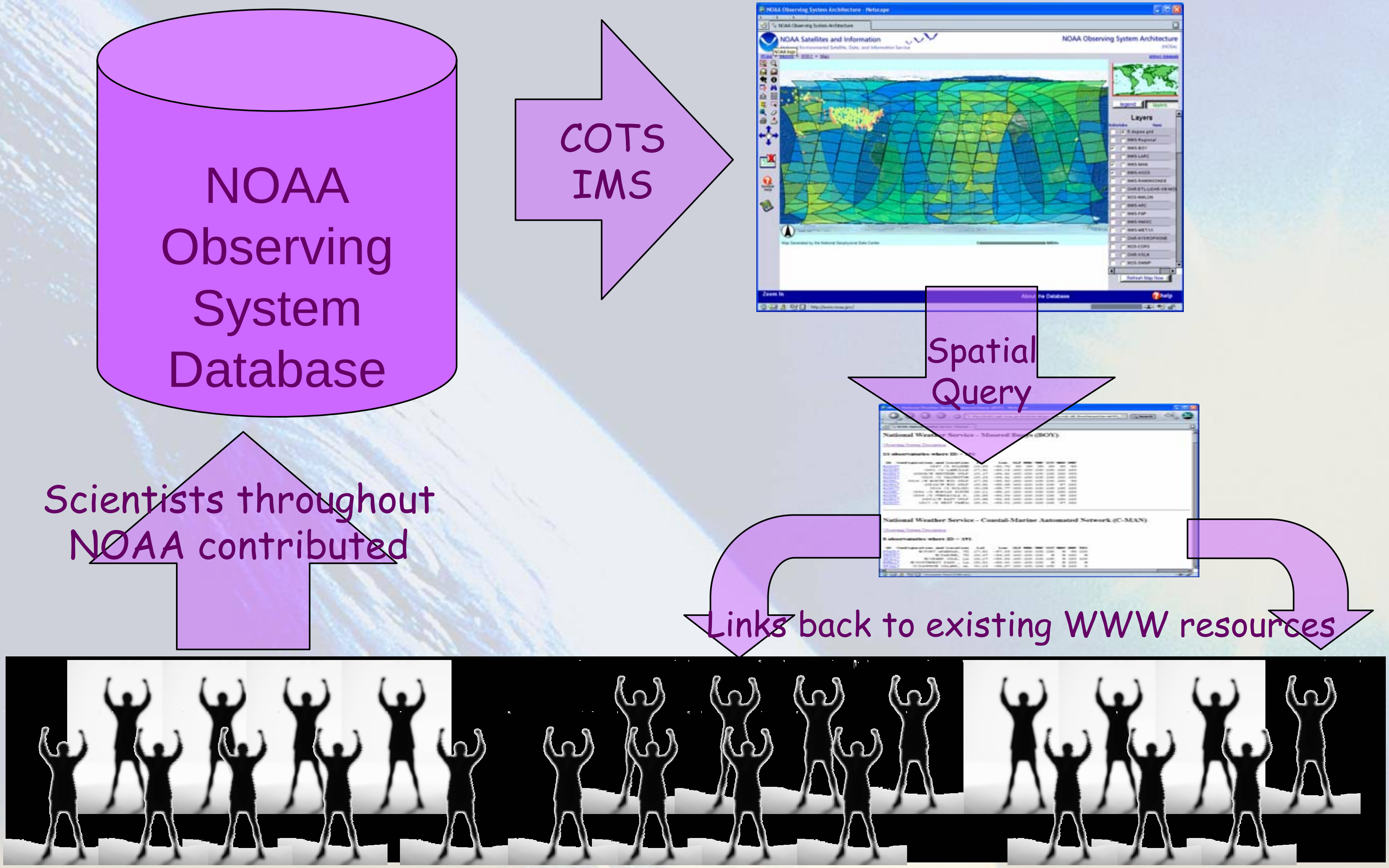
National Weather Service - Coastal-Marine Automated Network (C-MAN)

Observing System Description

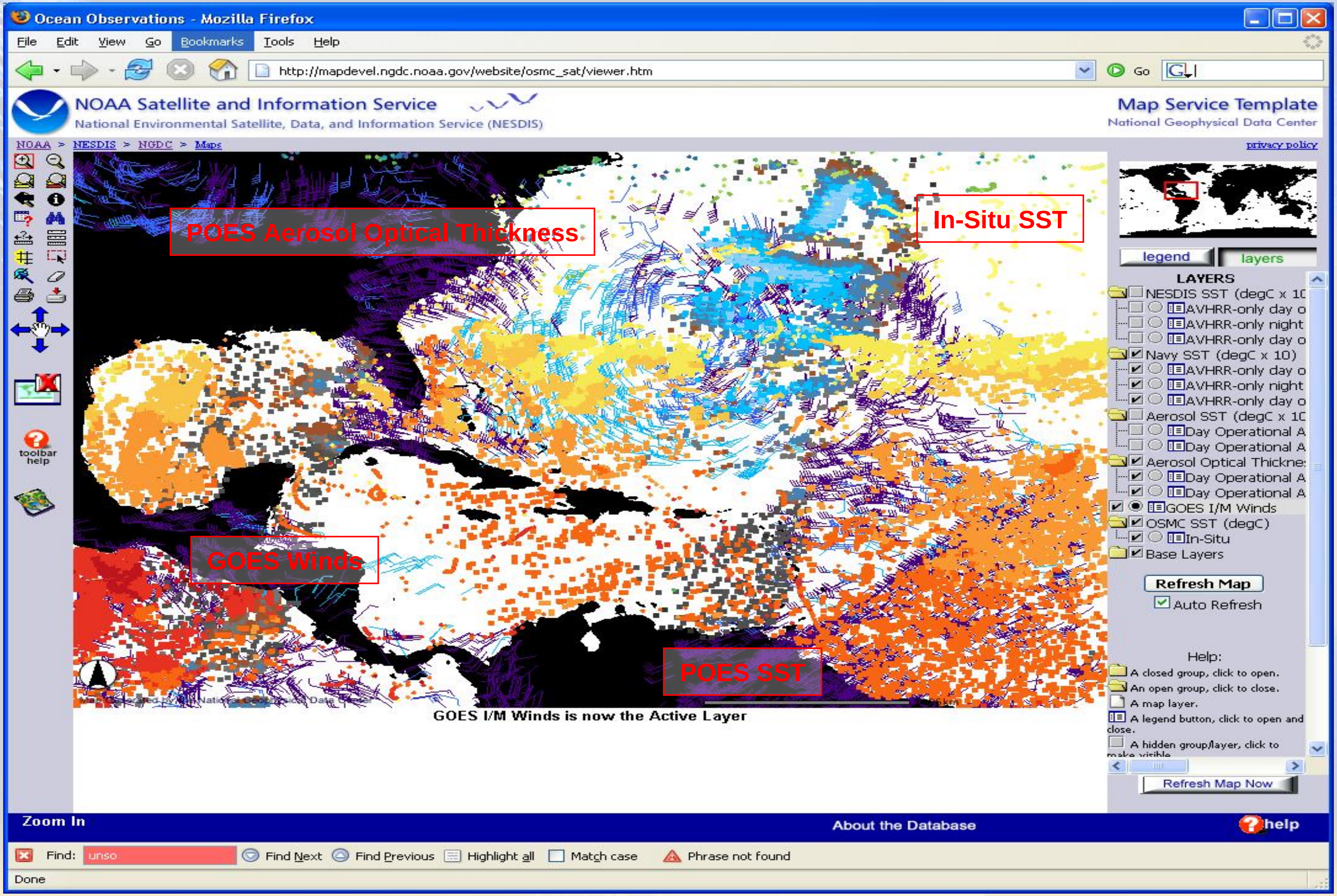
4 observatories where ID = 22

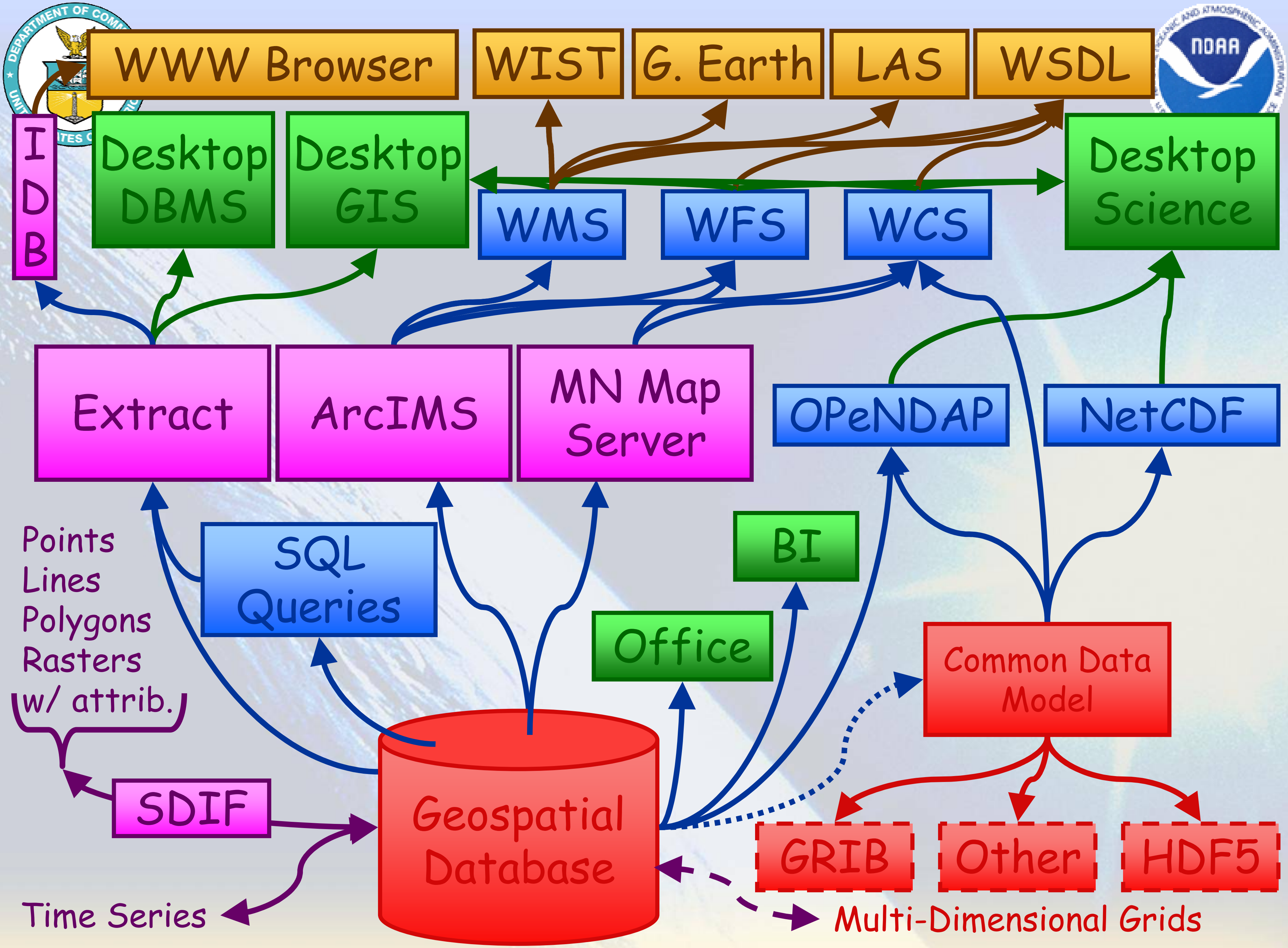
ID	Configuration and Location	Lat	Lon	SLP	WND	TMP	SST	WAV	DWP	TDS
PTAC1*	M/POINT ARENA, CA	38.96	-123.74	100	100	100	N	N	N	N
PTGC1*	M/POINT ARGUELLO, CA	34.58	-120.65	100	100	100	N	N	S	N
CARO3*	M/CAPE ARAGO ,OR	43.34	-124.38	99	99	99	N	N	S	N
NWPO3*	D/NEWPORT, OR	44.61	-124.07	100	100	100	N	N	S	N

What just happened?

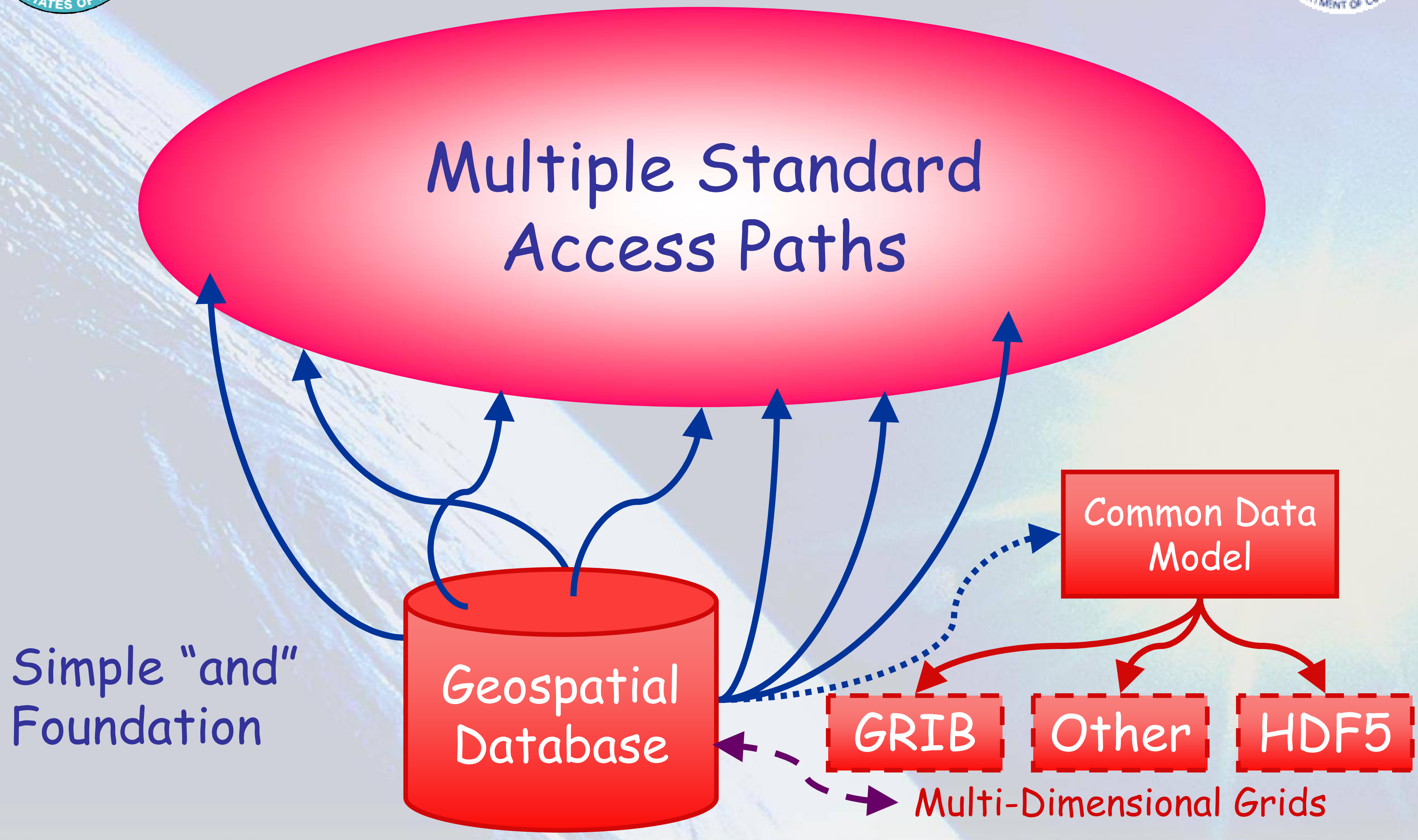


The Result: Integrated Data Systems!





WWW Browser
WIST
G. Earth
LAS
WSDL
IDB
Desktop DBMS
Desktop GIS
WMS
WFS
WCS
Desktop Science
Extract
ArcIMS
MN Map Server
OPeNDAP
NetCDF
Points
Lines
Polygons
Rasters
w/ attrib.
SQL Queries
BI
Office
Common Data Model
SDIF
Geospatial Database
GRIB
Other
HDF5
Time Series
Multi-Dimensional Grids


Multiple Standard
Access Paths
Common Data
Model
Simple "and"
Foundation
Geospatial
Database
GRIB
Other
HDF5
Multi-Dimensional Grids

Standards

- Standard names and terminology
- Metadata standards
  - e.g. FGDC and ISO 19115 w/ remote sensing extensions
- Standard formats for delivery of data/products
  - WMO, NetCDF, HDF, GeoTIF, JPEG, etc.
- Web Services Standards
  - World Wide Web Consortium
  - OGC (Features, Coverage, GML)
  - Community Standards: OPeNDAP (a REST service), Unidata's Common Data Model (CDM)
  - SOAP / UDDI / WSDL where appropriate

GEO-IDE - an essential component of environmental information management for NOAA

Integrated observing, data processing and information management systems

Connected by NOAA's Integrated Data Environment

Contributes to U.S. Global Earth Observation System (USGEO) and International Global Earth Observing System of Systems (GEOSS).

Important societal issues require data from many observation and data systems

	Discipline Specific View	Whole System View
Atmospheric Observations		
Land Surface Observations		
Ocean Observations		
Space Observations		
Data Systems	**Current systems are program specific, focused, individually efficient.** **But incompatible, not integrated, isolated from one another and from wider environmental community**	**Coordinated, efficient, integrated, interoperable**