NASA's Earth Science Data Systems - Lessons Learned and Future Directions

Presented at

US Workshop on Roadmap for Data Preservation Interoperability Framework

H. K. “Rama” Ramapriyan
Earth Science Data and Information System Project
Rama.Ramapriyan@nasa.gov

March 29, 2010

Acknowledgements

- Materials for this presentation have evolved over many years of the author's work with NASA's Earth Science Data Systems and contain contributions from several people in the Earth Science Data and Information System Project, the Goddard Space Flight Center and NASA Headquarters
- Any opinions expressed here are those of the author and do not necessarily imply official NASA policy

Main Messages

- NASA’s Earth Science Data Systems
  - “System of systems” with a history of ~20 years
  - Serves many disciplines in Earth sciences – supporting Earth System Science
  - Archiving (a.k.a. preservation) and distribution are critical functions
- Interoperability is important
  - Needed for different purposes and at different levels
  - Search and Access across systems: Directory , Inventory, Data levels
  - Not all systems need to interoperate – need is driven by user community requirements
  - Standards facilitate interoperability – difficult to “mandate” standards – easier to adopt community accepted standards
- “Temporal” Interoperability
  - Maintaining readability and understandability over time
  - Enabled by media standards, migration policies, metadata and documentation standards

Outline

- Background
- Content
- Impact
- Access
- Management
- Best Practices
- Conclusions

NASA’s Earth Science Data Systems

- “Study Earth from space to advance scientific understanding and meet societal needs” -- *2006 NASA Strategic Plan*

- NASA’s Earth Science Data Systems directly support this objective by providing end-to-end capabilities to deliver data and information products to users

Core and Community Capabilities - Definition

- ‘Core’ data system elements reflect NASA’s responsibility for managing Earth science satellite mission data characterized by the continuity of research, access, and usability.
  - The core comprises all the hardware, software, physical infrastructure, and intellectual capital NASA recognizes as necessary for performing its tasks in Earth science data system management.

- ‘Community’ elements are those pieces or capabilities developed and deployed largely outside the NASA core elements and are characterized by their ‘evolvability’ and innovation.

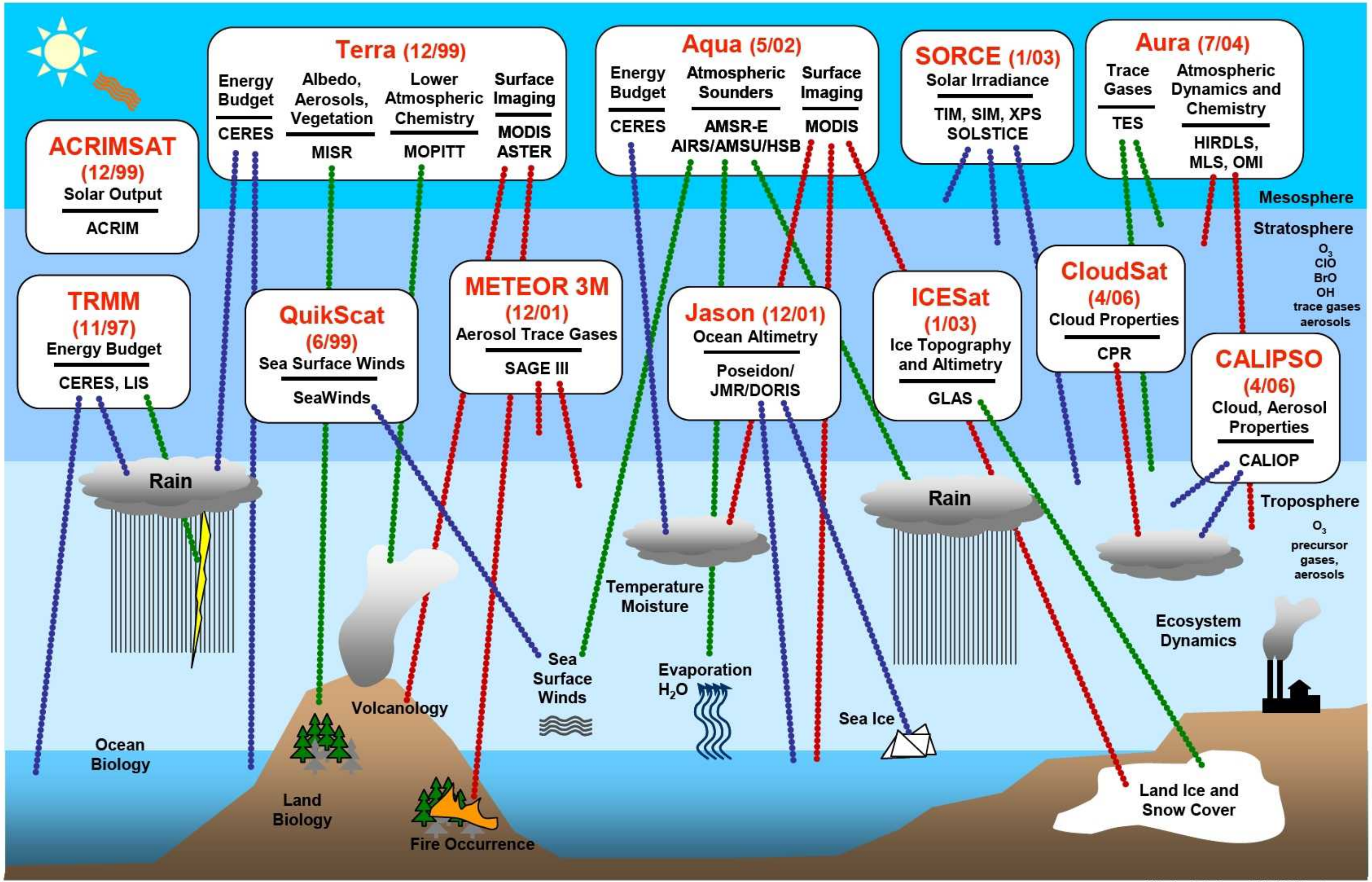
Core and Community Capabilities - Characteristics

CORE	COMMUNITY
Projects Subject to Programmatic Review	Projects Competitively Selected
Substantive NASA Oversight	'Light Touch' Oversight w/Significant Community Involvement
Tight Integration of Data System Tools, Services and Functions	Community-based Tools and Services Loosely-Coupled
Employ Well Established Information Technologies	Employ 'Edgy' or Emerging Technologies

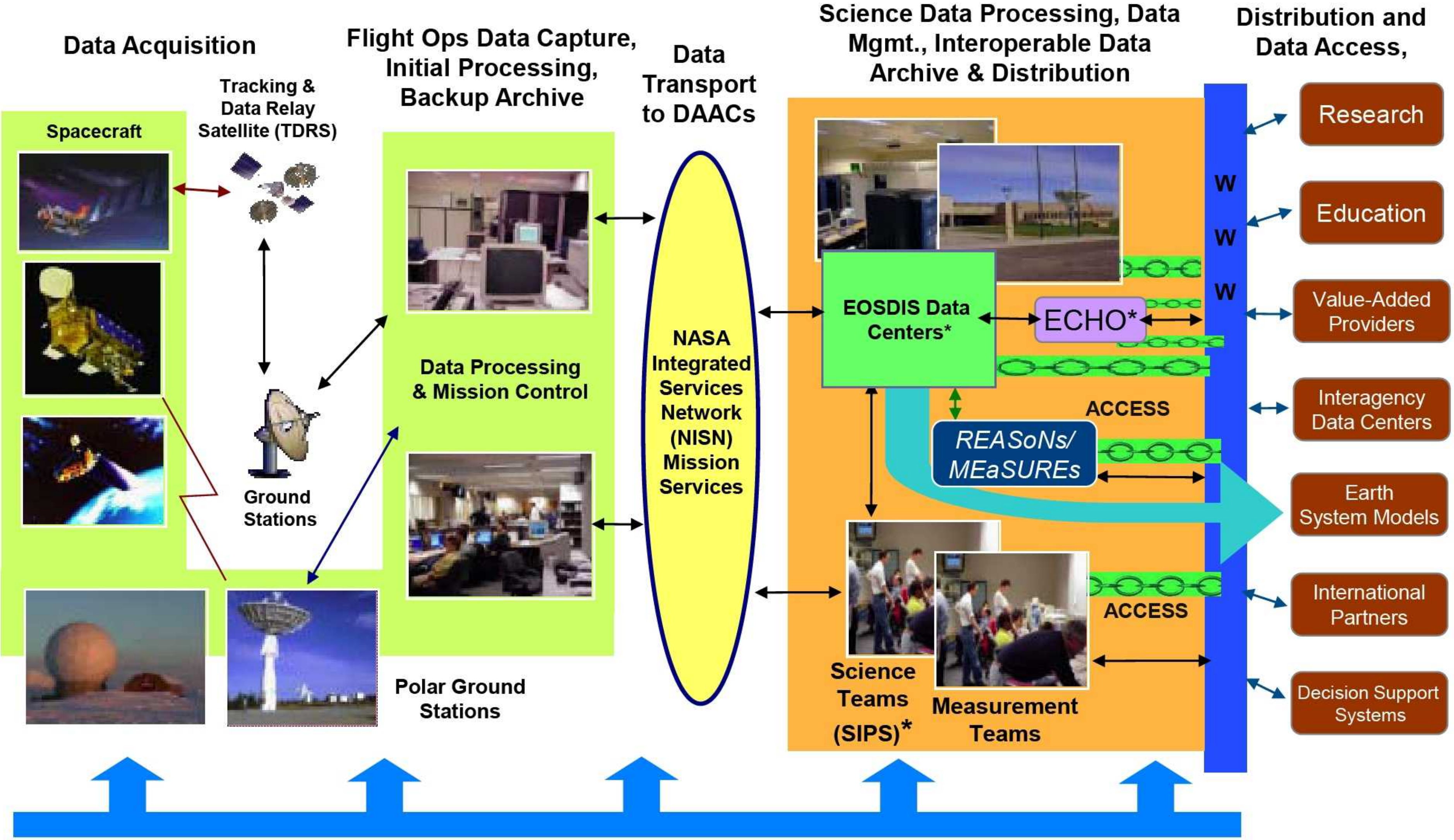
Core and Community Capabilities - Examples

- **Core – Earth Observing System (EOS) Data and Information System (EOSDIS)**
  - Operating since 1994, starting with "Version 0" managing heritage (pre-EOS) data at Distributed Active Archive Centers (DAACs) a.k.a. EOSDIS Data Centers and making them interoperate
  - Now managing all of EOS mission data and derived standard data products (in addition to pre-EOS data)
- **Community**
  - Research, Education and Applications Solutions Network (REASoN) Program – 42 five-year projects initiated in 2003/2004 – Output: technologies and data products
  - Advancing Collaborative Connections for Earth System Science (ACCESS) Program – 2-3 year projects starting 2005/2007/2009 – Output: technologies
  - Making Earth System data records for Use in Research Environments (MEaSUREs) Program – 30 projects initiated in 2007/ 2008 (Some completed REASoN Projects are continuing under this program) – Output: long time series of measurements - digital data products
  - Data products from community projects will be archived in EOSDIS after appropriate scientific reviews

EOSDIS Manages Data For All EOS Measurements



Earth Science Data Systems Context



Data Acquisition

Spacecraft

Tracking & Data Relay Satellite (TDRS)

Ground Stations

Polar Ground Stations

Flight Ops Data Capture, Initial Processing, Backup Archive

Data Processing & Mission Control

Data Transport to DAACs

NASA Integrated Services Network (NISN) Mission Services

Science Data Processing, Data Mgmt., Interoperable Data Archive & Distribution

EOSDIS Data Centers*

ECHO*

REASoNs/ MEaSUREs

ACCESS

Science Teams (SIPS)*

Measurement Teams

ACCESS

W W W

Distribution and Data Access,

Research

Education

Value-Added Providers

Interagency Data Centers

Earth System Models

International Partners

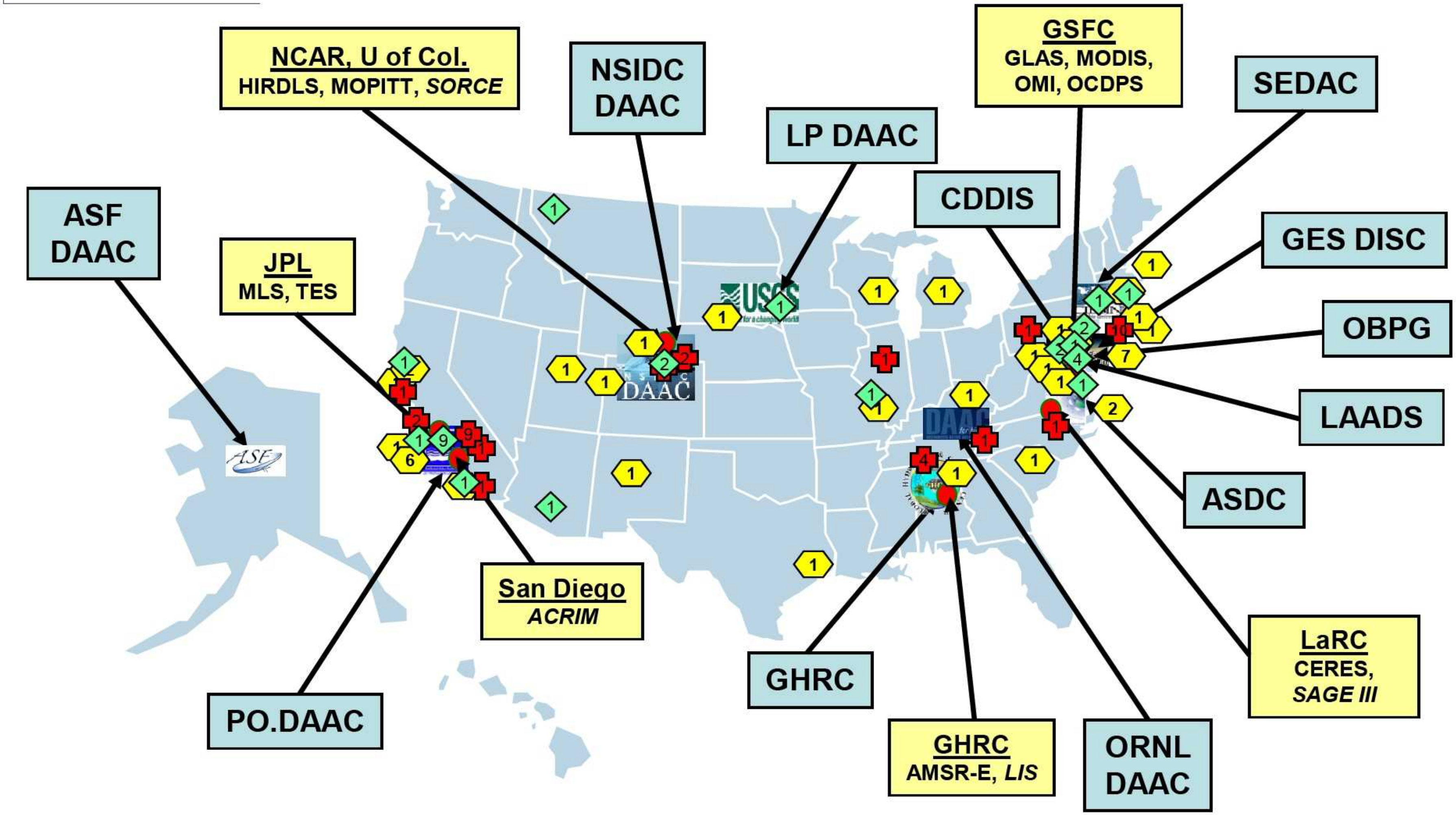
Decision Support Systems

Technology Infusion

(IT Currency, Standards, Reuse, Interoperability)

*EOSDIS Elements

Earth Science Data Systems (Core and Community)



NCAR, U of Col.
HIRDLS, MOPITT, *SORCE*

NSIDC DAAC

LP DAAC

GSFC
GLAS, MODIS, OMI, OCDPS

SEDAC

CDDIS

GES DISC

OBPG

LAADS

ASDC

ASF DAAC

JPL
MLS, TES

San Diego
ACRIM

PO.DAAC

GHRC

GHRC
AMSR-E, *LIS*

ORNL DAAC

LaRC
CERES, *SAGE III*

CORE

EOSDIS Data Centers

Science Investigator-led Processing Systems (SIPSs) ●

42 REASoN

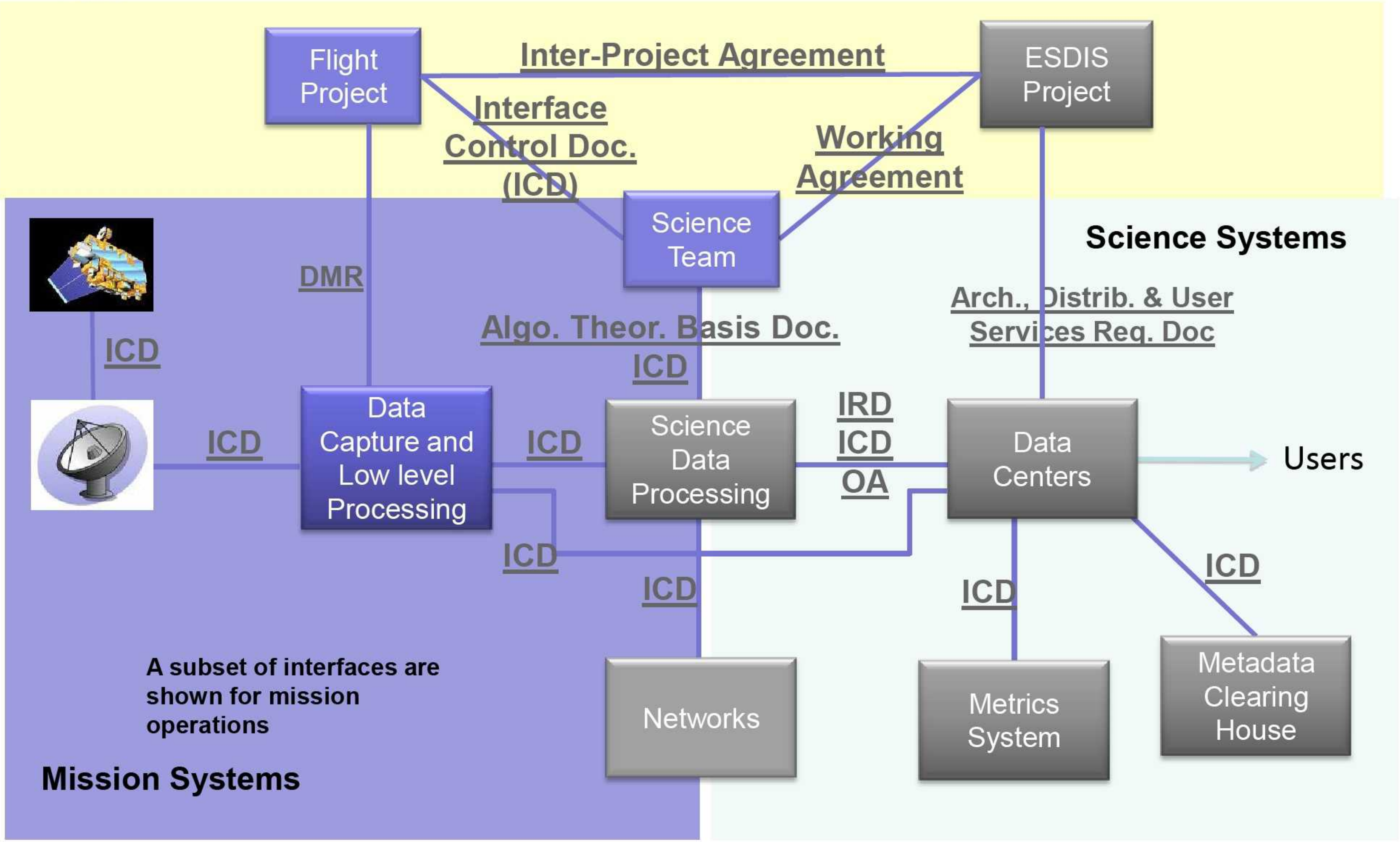
39 ACCESS

30 MEaSUREs

Complete History

COMMUNITY

Organizational/Functional Architecture and Interfaces



A subset of interfaces are shown for mission operations

Outline

- Background
- Content
- Impact
- Access
- Management
- Best Practices
- Conclusions

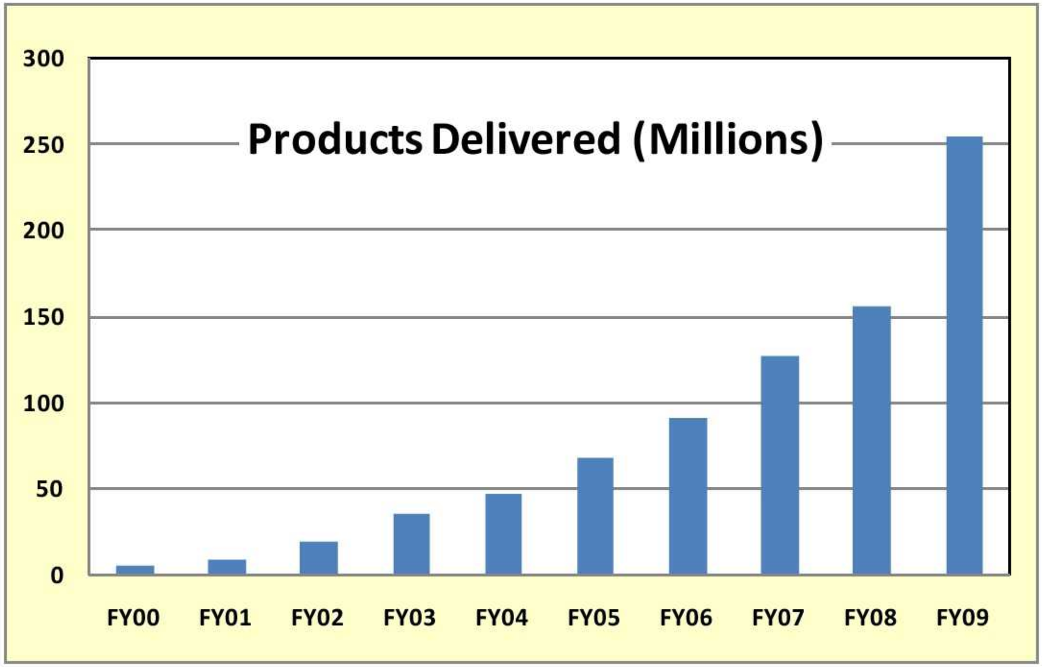
EOSDIS Key Metrics

EOSDIS Metrics (Oct 1, 08 to Sept 30, 09)

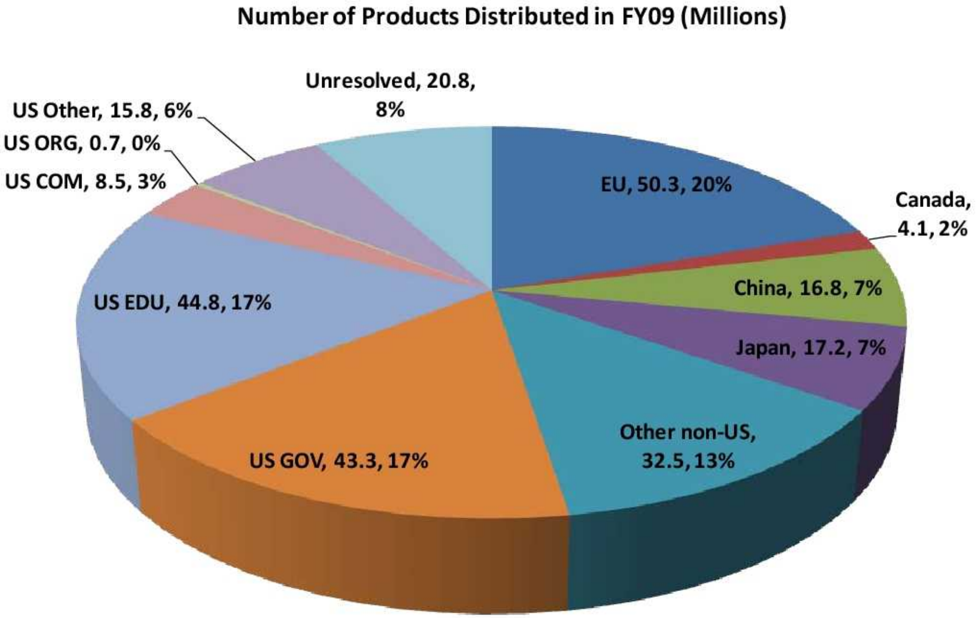
Unique Data Products	> 4000
Distinct Users of EOSDIS Data and Services	> 910K
Web Site Visits of 1 Minute or more	> 1M
Average Daily Archive Growth	1.8 TB/day
Total Archive Volume	4.2 PB
End User Distribution Products	> 254M
End User Average Daily Distribution Volume	6.7 TB/day

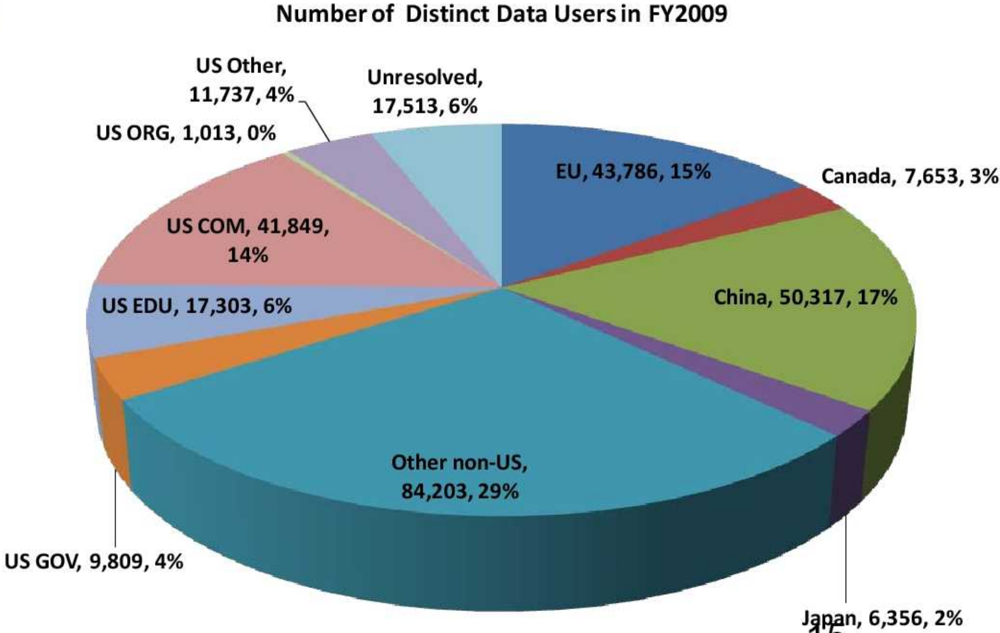
ESDIS Project Supports

Science System Elements	Data Centers	12
	SIPS	14
Interfaces	Interface Control Documents	32
Partnerships	US	8
	International	13
Missions	Science Data Processing	10
	Archiving and Distribution	38
	Instruments Supported	87



EOSDIS Data Distribution In FY2009





Metadata Standards

- EOSDIS Core System (ECS) Metadata Model – developed in mid-1990s
- Has had influence on FGDC metadata content standards' extensions to remotely Sensed data – see Content Standard for Digital Geospatial Metadata: Extensions for Remote Sensing Metadata, FGDC-STD-012-2002
- ECHO Data Model – specific to facilitating metadata searches on EOS Clearing House – close to ECS metadata model
- ISO 19115 – more recent standard; differences being examined between this and our legacy standards
- Data providers' compliance has been facilitated since the beginning of project by providing software toolkits

Data Format Standards

- **EOS products' primary format is HDF-EOS**
  - "Particularization" of the Hierarchical Data Format developed by NCSA in early 1990's and currently maintained by The HDF Group (THG)
  - HDF is a "formatting system" – provides specifications of structure as well as software tools
    - Considerable flexibility within structure – boon and bane
  - Selected in early days of EOSDIS after assessing several formatting systems in existence at the time
  - Other formats needed by community (e.g., NetCDF, GeoTIFF, binary) are accommodated by translation software
- **Current effort to generate XML maps of HDF file structures to facilitate readability of archived files independent of HDF software tools**
- **Standards Processes Group, one of NASA's community-based Earth Science Data System Working Groups, recommends standards for adoption by NASA – applies to metadata, data, access protocols, etc.**

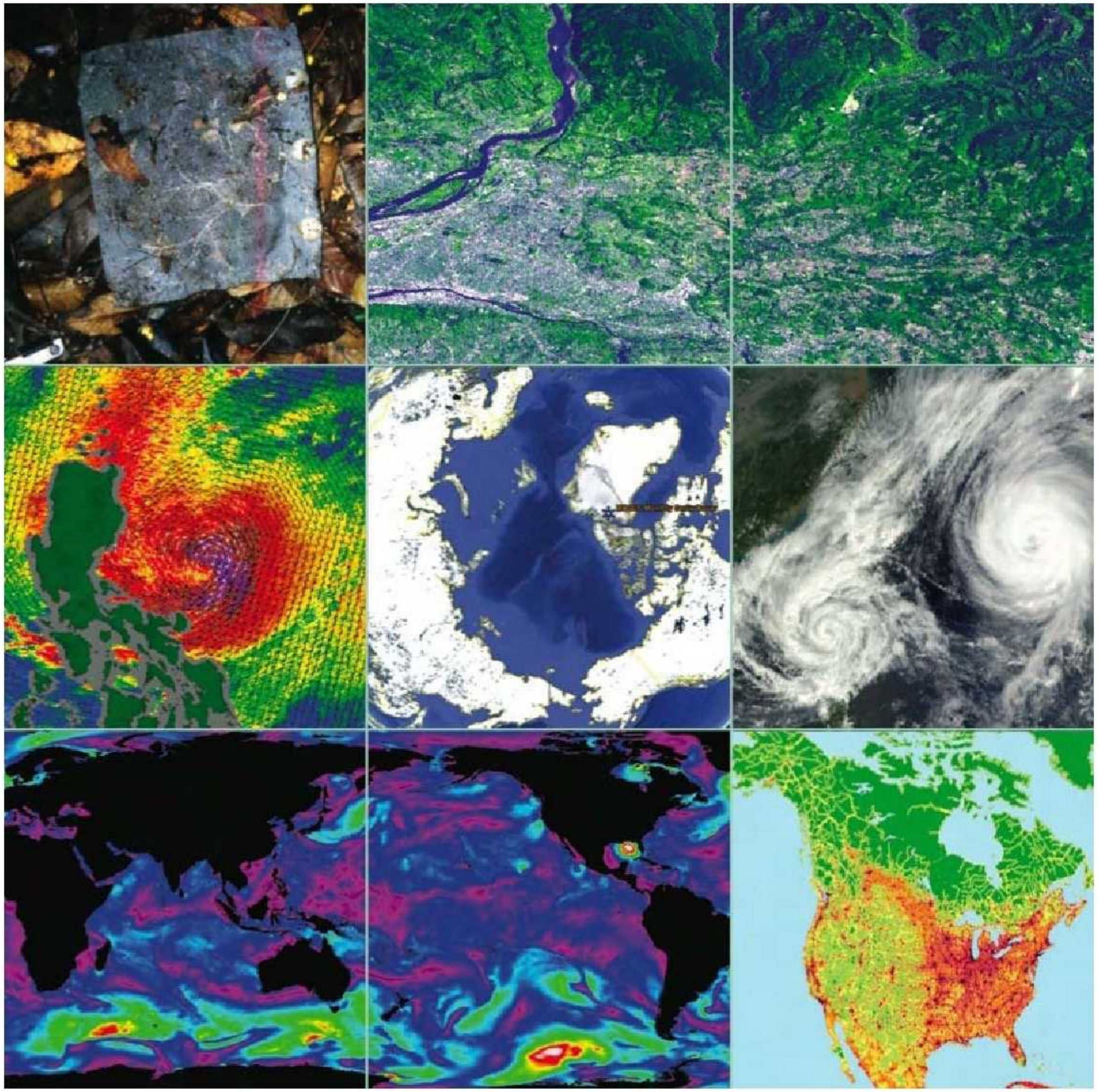
Outline

- Background
- Content
- Impact
- Access
- Management
- Best Practices
- Conclusions

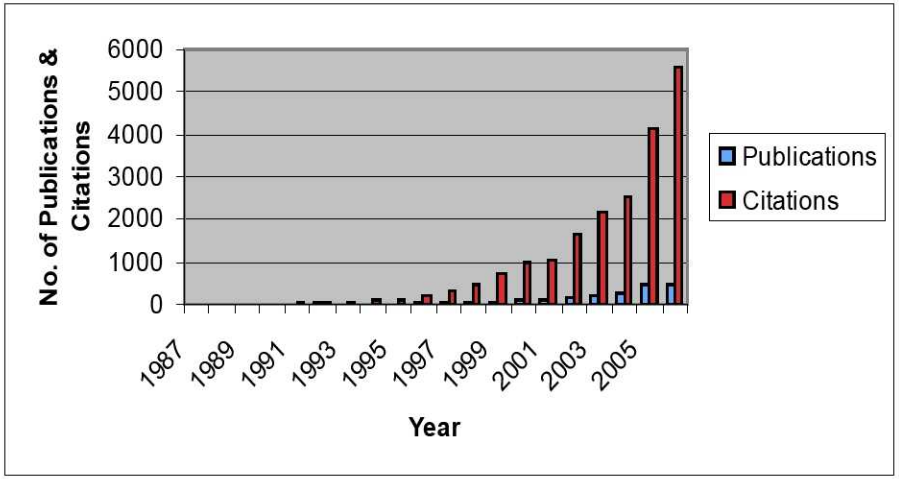
Impact of EOSDIS

- EOSDIS manages data from a large number of Earth observing instruments
  - All data and derived products are "born digital"
  - Archiving and distribution are important functions
- Data are used by a large number of scientific and applications users
  - e.g., in climate models, global change analyses, disaster response and impact analyses
  - Data integrity and verifiability are very important
- While NASA is not a "permanent archive" agency, it has to maintain a "research archive" for as long as data are used for scientific research and/or transition responsibility to permanent archives
  - Loss of data would have a negative impact on future verifiability of conclusions from global change analyses

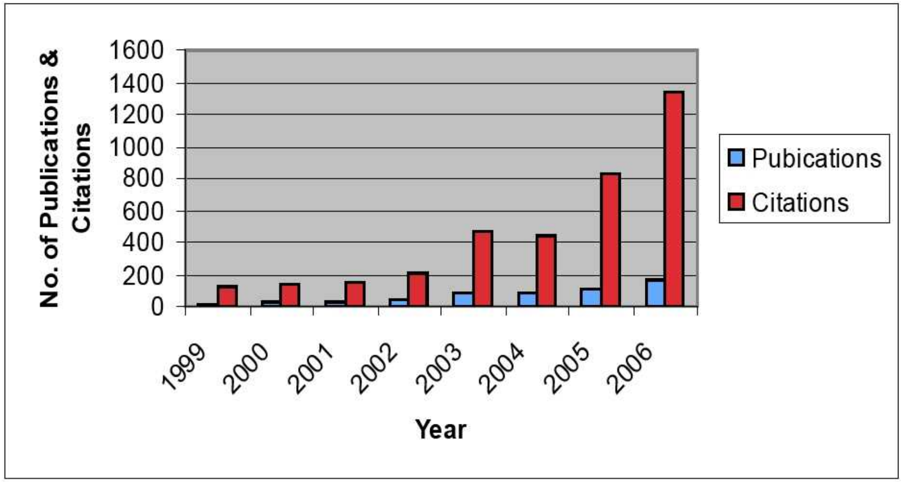
Impact of EOSDIS - Earth Science Research



Impact of EOSDIS - Scientific Productivity



Publications resulting from EOS Terra (12/99 launch) instruments and data



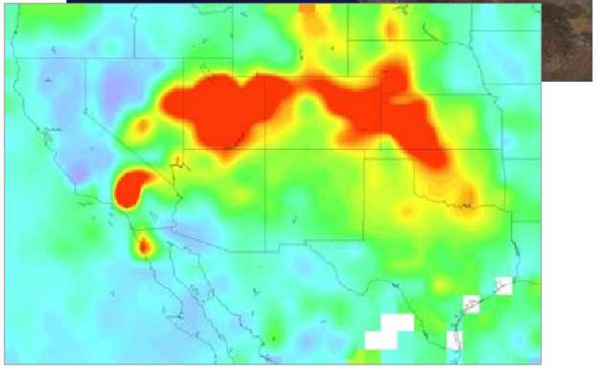
Publications resulting from EOS Aqua (05/02 launch) instruments and data

- Publications and citations shown here are a good indicator of scientific growth resulting from NASA's Terra and Aqua missions
- Pre-launch publications and citations are significant, but dramatic growth seen post-launch
- NASA's EOSDIS, through its well-established data management practices:
  - Produces and stores data and metadata in formats compliant with well-documented standards
  - Provides data, metadata and software tools promptly to a broad scientific community
- Data management is a key element in supporting scientific growth

- Terra metrics from Imhoff, M. L., S. C. Tsay, R. E. Wolfe, M. Hato, M. J. Abrams, B. A. Wielicki, D. J. Diner, V. V. Salomonson, J. R. Drummond, and J. C. Gille, 2007: Terra Senior Review Proposal, submitted to NASA Headquarters March 16, 2007
- Aqua metrics from Parkinson, C. L., S. E. Platnick, M. T. Chahine, V. V. Salomonson, A. Shibata, R. Spencer, B. Wielicki, J. Gainsborough, and S. M. Graham, 2007: Aqua Senior Review Proposal, submitted to NASA Headquarters March 16, 2007

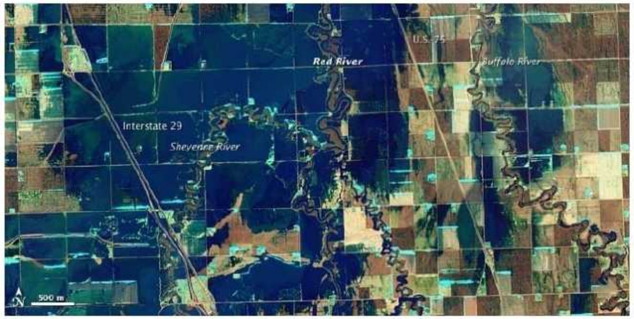
Impact of EOSDIS/Applications

Pyrocumulus clouds & smoke - Station Fire 2009

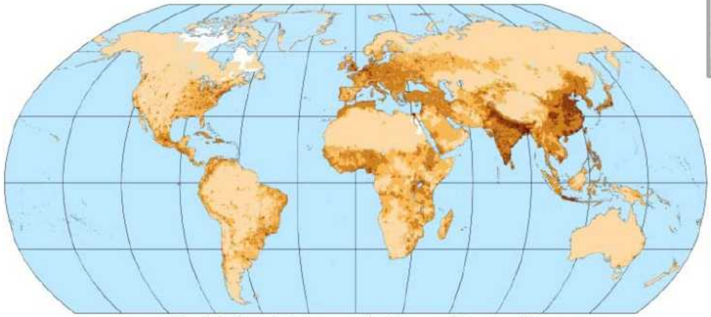


Carbon monoxide emissions - Station Fire 2009

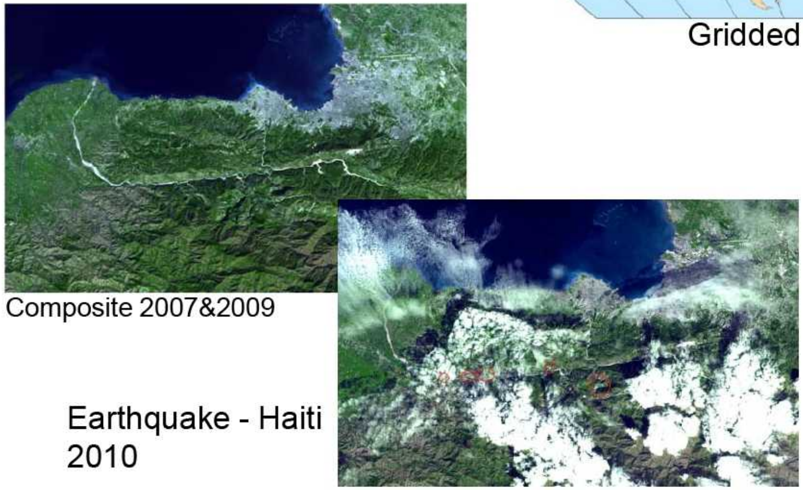
Flooding – North Dakota 2010



Snow - East Coast 2010

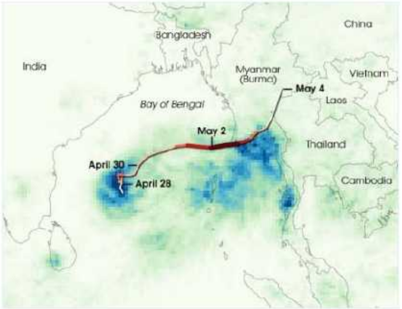


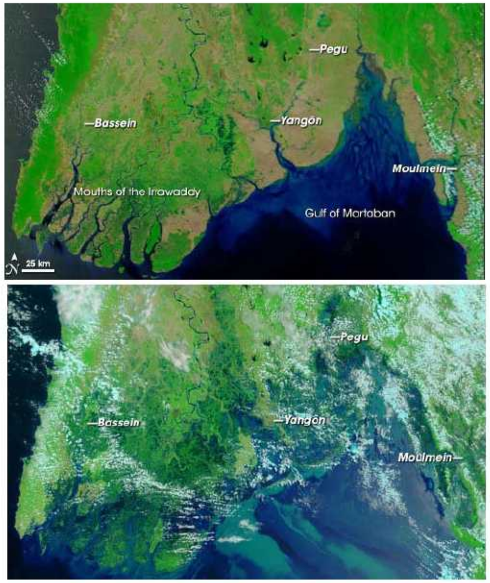
Gridded Population Density



Composite 2007&2009

Earthquake - Haiti 2010





Cyclone and Flooding – Myanmar 2008

Outline

- Background
- Content
- Impact
- Access
- Management
- Best Practices
- Conclusions

Access - Policy

- **NASA Earth Science Data Policy**
  - No period of exclusive access
  - Except where agreed upon with international partners, data and derived scientific products are available at no cost to all users
  - Any variation in access will result solely from user capability, equipment, and connectivity
  - All NASA-generated standard products are made available (upon request) along with the source code for algorithm software, coefficients, and ancillary data used to generate these products.
  - See Earth Science Reference Handbook (NASA, 2006) for full text of policy
- **Data are made available to all users promptly**
  - After an initial checkout period
  - Appropriate caveats about data quality are provided in product documentation

Access - Technical

- There are several ways to search for data of interest
  - Directory level information from Global Change Master Directory
  - Cross-Data Center searches through Warehouse Inventory Search Tool (WIST) – “inventory level interoperability” – uses EOS Clearing House (ECHO) metadata repository
  - Data Center-specific search tools
  - Tailored client software using ECHO metadata repository
- Almost all data in EOSDIS are held on-line and accessed via ftp
  - A small part, still held in near-line robotic tape archives are being migrated to on-line storage
- On-line services are available
  - e.g., subsetting, reprojection, mosaicing, format conversion
- Several data visualization and analysis tools are available at EOSDIS Data Centers

Outline

- Background
- Content
- Impact
- Access
- Management
- Best Practices
- Conclusions

Management

- NASA is not a "permanent archive" agency
  - It has to maintain a "research archive" for as long as data are used for scientific research and/or transition responsibility to permanent archives
  - Critical data are backed up off-site
- "Research archive" maintenance implies *continuing evolution*
  - keep up with technologies – hardware upgrades, data migration, upgrade of software and tools to "keep up with the times"
  - For example, all data were initially stored on near-line robotic archives; now they are on-line (RAID)
  - Data distribution was both on media and on-line; now it is only on-line (with very rare exceptions)

Major Types of Critical Data

- Science observations from the NASA mission/ instrument
  - The raw data records, the Level 1 data that can be used to develop refined Climate Data Records.[1]
  - Calibrated and geo-located radiance data. The definitive version of the EOS Level 1 data and any other data sets or products needed to interpret them.[2]
- Validation field campaign datasets and Inter-comparisons with other instruments
- Ancillary datasets from other agencies and projects
- Derived higher-level products, applications and research results

Footnotes:

[1] National Research Council. 2000. Ensuring the Climate Record from the NPP and NPOESS Meteorological Satellites, Committee on Earth Studies, Commission on Physical Sciences, Mathematics and Applications

[2] Joint NASA-NOAA Workshop, USGCRP, LTA Workshop Report, 1998

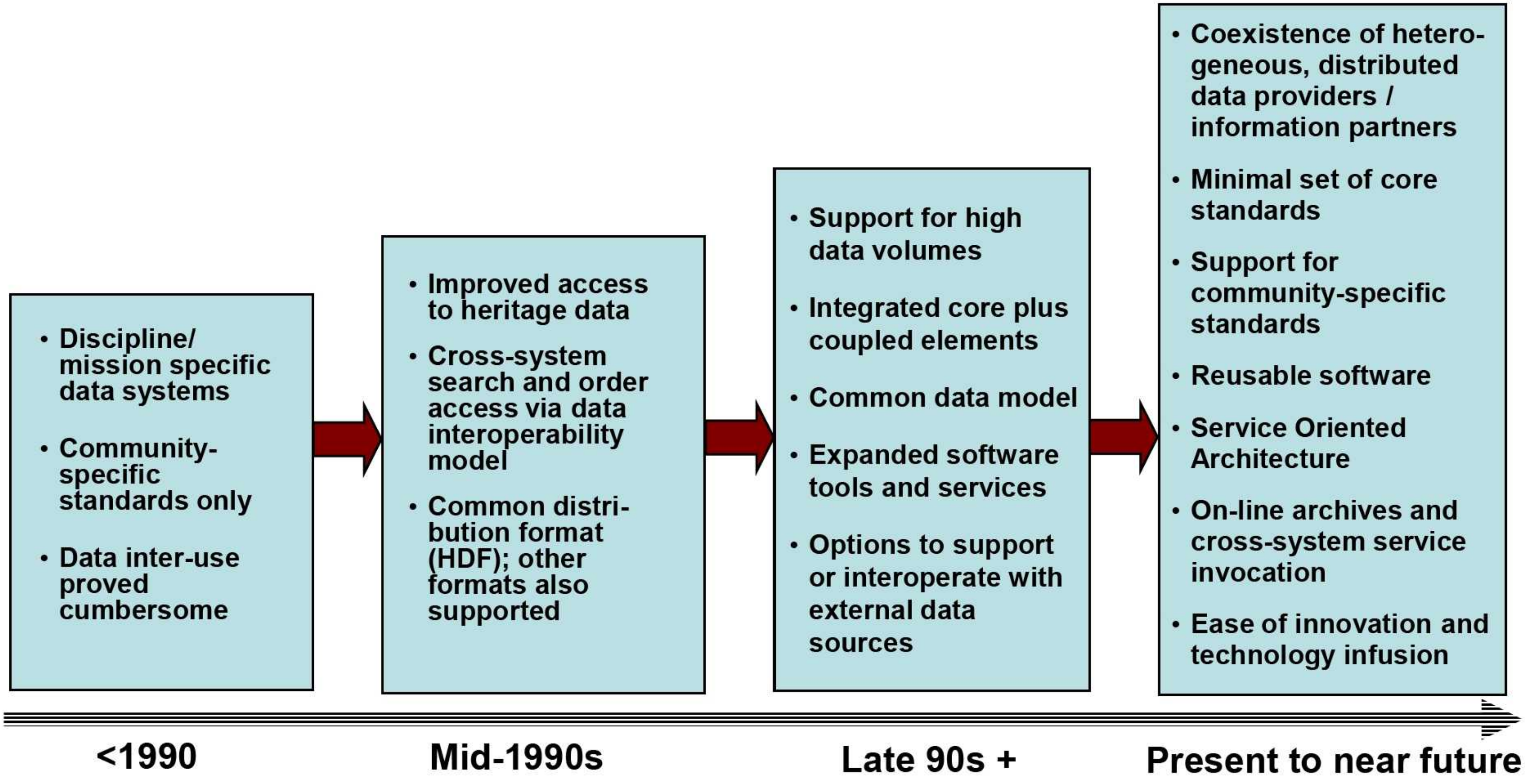
Major Types of Additional Information

1. *"Instrument/sensor characteristics including pre-flight or pre-operational performance measurements (e.g., spectral response, noise characteristics, etc.)*
2. *Instrument/sensor calibration data and method*
3. *Processing algorithms and their scientific basis, including complete description of any sampling or mapping algorithm used in creation of the product (e.g., contained in peer-reviewed papers, in some cases supplemented by thematic information introducing the data set or derived product)*
4. *Complete information on any ancillary data or other data sets used in generation or calibration of the data set or derived product*
5. *Processing history including versions of processing source code corresponding to versions of the data set or derived product held in the archive*
6. *Quality assessment information*
7. *Validation record, including identification of validation data sets*
8. *Data structure and format, with definition of all parameters and fields*
9. *In the case of earth based data, station location and any changes in location, instrumentation, controlling agency, surrounding land use and other factors which could influence the long-term record*
10. *A bibliography of pertinent Technical Notes and articles, including refereed publications reporting on research using the data set*
11. *Information received back from users of the data set or product"*[1]

Footnotes:

[1] Joint NASA-NOAA Workshop, USGCRP, LTA Workshop Report, 1998

Evolution of Data System Features



- Discipline/ mission specific data systems
- Community-specific standards only
- Data inter-use proved cumbersome

<1990

- Improved access to heritage data
- Cross-system search and order access via data interoperability model
- Common distribution format (HDF); other formats also supported

Mid-1990s

- Support for high data volumes
- Integrated core plus coupled elements
- Common data model
- Expanded software tools and services
- Options to support or interoperate with external data sources

Late 90s +

- Coexistence of heterogeneous, distributed data providers / information partners
- Minimal set of core standards
- Support for community-specific standards
- Reusable software
- Service Oriented Architecture
- On-line archives and cross-system service invocation
- Ease of innovation and technology infusion

Present to near future

Lessons learned and information technology advances coupled with user working group and advisory council advice and ideas supports a continuously evolving data system with growing capabilities for the user community

EOSDIS Evolution - 2015 Vision Tenets

Vision Tenet	Vision 2015 Goals*
Archive Management	▪ NASA will ensure safe stewardship of the data through its lifetime. ▪ The EOS archive holdings are regularly peer reviewed for scientific merit.
EOS Data Interoperability	▪ Multiple data and metadata streams can be seamlessly combined. ▪ Research and value added communities use EOS data interoperably with other relevant data and systems. ▪ Processing and data are mobile.
Future Data Access and Processing	▪ Data access latency is no longer an impediment. ▪ Physical location of data storage is irrelevant. ▪ Finding data is based on common search engines. ▪ Services invoked by machine-machine interfaces. ▪ Custom processing provides only the data needed, the way needed. ▪ Open interfaces and best practice standard protocols universally employed.
Data Pedigree	▪ Mechanisms to collect and preserve the pedigree of derived data products are readily available.
Cost Control	▪ Data systems evolve into components that allow a fine-grained control over cost drivers.
User Community Support	▪ Expert knowledge is readily accessible to enable researchers to understand and use the data. ▪ Community feedback directly to those responsible for a given system element.
IT Currency	▪ Access to all EOS data through services at least as rich as any contemporary science information system.

*Developed by EOSDIS Elements Evolution Study Team - 2005

Outline

- Background
- Content
- Impact
- Access
- Management
- Best Practices
- Conclusions

Lessons/Best Practices (1 of 5)

- Open Data Policy
  - NASA provides open access to data with no period of exclusive access
  - Most of the data are provided at no charge to any requesting user
- Both Core and Community Capabilities are essential to meet NASA's Earth Science program objectives
  - Core capabilities are needed for long-term stability and dependable capture, processing, and archiving of data and distribution of data to a broad and diverse communities of users, including value-added service providers
  - Community capabilities provide innovative, new scientific products as well as a path to technology infusion
    - NASA currently has four Earth Science Data System Working Groups (ESDSWG) – see http://esdswg.gsfc.nasa.gov/
      - Standards Processes Group
      - Technology Infusion Working Group
      - Reuse Working Group
      - Metrics Planning and Reporting Working Group
    - Working groups provide community-vetted recommendations to NASA to consider implementation
    - These recommendations as well as those from EOSDIS Data Centers, annual user feedback through surveys and at community conferences, interagency and international discussions influence NASA's programmatic direction
    - NASA needs to strengthen its effort in facilitating technology infusion from community to core systems

Lessons/Best Practices (2 of 5)

- Loosely coupled, heterogeneous systems can work together
  - Early development of EOSDIS (so-called Version 0) involved making heterogeneous systems interoperate in the “pre-WWW” era
  - Successful, with well-defined interfaces and a “thin” translation layer to spread queries to multiple databases and gather responses to present to users (“one-stop shopping”)
- Complex development of EOSDIS Core System (ECS) with “strongly coupled” components proved to be difficult
  - Eventually successful after reducing scope and allocating most of processing to Science Investigator-led Processing Systems
  - Version 0 Information Management System (IMS) was adopted for one-stop shopping across data centers
  - Managing standards and interfaces was key to success
  - Thorough interface tests and end-to-end testing was critical
- Community evolution of standards works better than top-down approach
  - Essential to provide flexibility to accommodate multiple standards and software tools to facilitate data use

Lessons/Best Practices (3 of 5)

- **Must plan for preservation**
  - Periodically refresh media including ‘touching’ all data
  - Budget for hardware refresh every three years
  - Metadata is a key cost driver
    - needs to be continually reconciled and updated
    - changes with each new data model
    - websites are useless without good metadata
  - Science discipline expertise is required for management of data

Lessons/Best Practices (4 of 5)

- One size does not fit all
  - Scientific disciplines have different ways of looking at the data and different vocabularies.
  - Need flexibility and tools to handle other data and metadata formats
  - Need some consistency to facilitate search and access across datasets
  - Enable/Facilitate development of different interfaces to support different communities

Lessons/Best Practices (5 of 5)

- **Data Systems must evolve over time**
  - In early 2005, NASA embarked on an EOSDIS Evolution Study
  - Addressed multi-faceted goals/issues:
    - Manage archive volume growth
    - Improve response and data access
    - Reduce recurring costs of operations and sustaining engineering
    - Update aging systems and components
    - Move towards more distributed environment
  - A vision for the 2015 timeframe was developed by the EOSDIS Elements Evolution Study Team
  - It is critical to manage transitions of an operational system that serves large numbers of users
    - Transitions are made incrementally
    - Each transition involves testing by interfacing systems' staff, and certification by affected users (or representatives)

Outline

- Background
- Content
- Impact
- Access
- Management
- Best Practices
- Conclusions

Conclusions

- NASA has significantly improved its Earth Science Data Systems over the last two decades
- Open data policy and inexpensive (or free) availability of data has promoted data usage by broad research and applications communities
- Flexibility, accommodation of diversity, evolvability, responsiveness to community feedback are key to success