CEOS Future Data Access & Analysis Architectures Study

Version 0.8, 6 July 2016

**This document is undergoing a major refactor. All previous contributions have been removed as the outline is changed (a back up has been made). Content will be refactored and replaced in the new structure over the coming days. Initial notes in achieving the restructure may not be understandable outside of the FDA Tiger Team (or to start with Rob Woodcock’s head).**

# Executive Summary

TBD

## Summary Recommendations

# 

# 

# 1.Introduction

## Overview

With each passing year, new generations of Earth observation (EO) satellites are creating increasingly significant volumes of data with such comprehensive global coverage that for many important applications, a ‘lack of data’ no longer becomes the limiting factor.

Extensive research and development activity has resulted in new applications that offer significant potential to deliver great impact to important environmental, economic and social challenges, including at the regional and global scales necessary to tackle ‘the big issues’. Such applications highlight the profile of EO to Ministers and other key people. However, for EO to make the most of this enormous potential, the gap between data and application needs to be bridged. Currently, many applications fail to successfully scale up from small-scale research to global or regional operations because of a lack of suitable data infrastructure. Even today, much of the archived EO satellite data sit under-utilized on tapes. Significant application potential remains consigned to prototypes, exemplars and test-beds.

It would not be technically feasible or financially affordable to consider traditional processing and data distribution methods to address this ‘scaling’ challenge, as the size of the data and complexities in preparation, handling, storage, analysis and basic processing remain significant obstacles in many countries, including as they support key GEO/CEOS initiatives such as the Global Forest Observations Initiative (GFOI), Disasters, Water Resources and the GEO Global Agricultural Monitoring initiative (GEOGLAM).

Addressing of this problem by individual users has not thus far resulted in an optimal solution and misses the opportunities offered through collaborative environments where both data providers and users can work together across domains and across geographic boundaries. However, the data management and analysis challenges arising from the explosion in free and open data volumes can be overcome with the opportunities offered by new, high-performance Information and Communications Technologies (ICT) infrastructure and architectures aimed at improving data management for providers and removing obstacles to data uptake by users.

From a historical perspective, the main highlights were the leaps from film/tape to digital, the continued evolution to make digital imagery available in near real time (days to hours), the policy decision to make moderate resolution EO data free and open (initially by INPE regionally and then by the USGS globally in 2008 and the European Union globally in 2014), Developments by NASA and universities to produce higher order products (e.g. MODIS and WELD) to include essential climate variables and climate data records. This evolutionary process continues with defining and agreeing to some standards related to land surface imaging analysis ready data (ARD), the concept of data cubes as organized structures to run applications, and Cloud Computing or Supercomputer hosting just to name a few.

## Purpose

The CEOS Future Data Analysis and Applications Architectures Ad-hoc team (FDA-AHT) has been tasked by the CEOS Chair team to assess the potential of new technologies and approaches, identify key issues and opportunities, and propose a plan of action for consideration by CEOS.

This report has:

1. Reviewed an inventory of relevant initiatives and plans being undertaken by CEOS and related agencies;
2. Reviewed lessons learned from the early prototypes currently underway with the governments of Kenya and Colombia;
3. Identified key issues and opportunities resulting from the trend towards Big Data, Analysis Ready Data, etc;
4. Made recommendations for the way forward for CEOS and its agencies, including in relation to standardisation, interoperability etc, and how the current CEOS priorities might benefit from the proposed activities.

This study is anticipated to be of value both to CEOS Agencies as data providers and to existing and prospective users of EO satellite data. The full potential of EO satellite data will not be realised with the obstacles that users face in current data handling and analysis approaches. Global initiatives such as GFOI and GEOGLAM exemplify the difficulties that countries without developed national spatial data infrastructures face in terms of lack of capacity in their ability to handle EO satellite data. This capacity gap is a major hindrance to the uptake of EO data in global initiatives. Moreover, even many developed countries are struggling to determine how best to capitalise on ‘big space data’ and would appreciate guidance on both best practice and more streamlined approaches to maximise value from different satellites.

CEOS investigations into next generation data systems must consider innovative solutions to the ‘last mile’ problem, where technological solutions have tended to fail. It should consider phased solutions that can help many countries in the near-term by working with CEOS capacity building partners, as we also work toward long-term solutions.

CEOS initiatives in areas such as disaster management and forest monitoring have identified that the obstacles to the uptake of EO satellite data are not only technical. Issues relating to user and intermediary awareness, understanding and capacity to exploit data are just as significant. The proposed studies would ideally include substantial engagement with external stakeholders including typical user groups, UN agencies and financing bodies such as the World Bank to ensure their perspectives are fully understood and reflected as we plan the way forward. These bodies are where we hope the benefits will ultimately be realised, and they should be engaged early and fully.

## Context

**CEOS Strategic Guidance (November 2013)**

Challenge: Engage Stakeholders to Optimize Relevance

Opportunity: Build Capacity for Earth Observation Products

Opportunity: Identify Gaps and Promote Complementarity

Strategic Direction: Optimize the Societal Benefit of Space-based Earth Observation

Strategic Direction: Remain the Focal Point for International Coordination of Space-based Earth Observations

**Global data flow study for the GFOI report**

**CEOS Chair non-meteorological applications…**

**Google Earth Engine and related - probably not here, see later**

## Structure of the Report

This report initially reviews the current trends and developments in EO systems architecture and applications. In the creation of this report submissions were made by a number of CEOS members regarding current trends and their specific development responses in EO systems architectures and applications (see Appendix A). As each agency has different terminology, operational methods, language and business and policy drivers the submissions appear different in the detail. Careful analysis though shows common trends and responses that are particularly relevant to the CEOS mission.

In order to provide a degree of consistency over a complex set of interrelated issues and differing agency terminology the report has where possible categorised the various architectural concerns using The Open Group Architecture Framework (TOGAF). TOGAF defines four categories for an enterprise system architecture description, each of which provides a specific view into what is a single architecture:

1. **Business architecture**—Describes the business drivers and processes used to meet the business goals
2. **Application architecture**—Describes how specific applications are designed and how they interact with each other
3. **Data architecture**—Describes how the datastores are organized and accessed
4. **Technical architecture**—Describes the hardware and software infrastructure that supports applications and their interactions

In addition, the report has been limited to only those aspects of the EO systems architecture that are high priority or impact and directly relevant to the CEOS mission. By necessity this leaves some important architectural considerations outside of the report. However with these essential aspects identified a CEOS agency is free to implement and adjust other components according to their specific operational needs and the CEOS community can look to resolve community standards and interoperability concerns in future activities

The report is structured as follows:

**Chapter 2** consolidates contributions and identifies trends and priorities in EO systems architecture development across CEOS agencies. It serves as a baseline of current architecture and near future development responses.

**Chapter 3** discusses the challenges faced in EO system architecture design and development for the medium to long term future. It serves to identify the key challenges that must be addressed in future data architectures

**Chapter 4** describes key architectural responses that seek to resolve the challenges identified in Chapter 3. It is not a complete architectural description and focuses on the essential elements necessary for the CEOS mission leaving details to future projects or Agency developments.

**Chapter 5** summarises the outcomes of the report and presents recommendations on Future Data Architectures and activities for CEOS.

# 2. Current trends and developments in EO systems architecture and applications

Space agencies are faced with a number of trends which, taken together, are driving the need for change in the ways in which data are accessed, analysed, and distributed. The magnitude and speed of these environmental changes will determine the importance and urgency with which change is required in data architectures. They define ‘the challenge that Future Data Architectures must solve. In this section current and near future systems architecture developments have been consolidated from across a range of current activities under way in space agencies (see Appendix A for agency specific contributions).

## Notional Architecture Baseline

### Business Architecture

Key business drivers and requirements for existing EO systems:

#### Maximise the value of Earth observations

This is fundamental driver for all CEOS agencies and a key part of the CEOS Strategic Guidance. The business logic driving investment is the expectation that “publicly funded [EO] agencies should maximise the value returned to the country through the application of national data holdings”. As a fundamental driver most agency systems architectures have been designed to deliver calibrated observations and produce value added products for use by other Government agencies on predominantly National and Global societal, environmental and scientific problems.

In recent years there has been a steady trend across all agencies towards greater integration of EO data holdings with other data types held by more diverse Government agencies - “Comprehensive collection and integration of ... information independently controlled by governmental agencies should be promoted and such information should be disclosed appropriately to increase the convenience for users to access and handle such information.”

Increasingly EO data are valued not only for its scientific and technological value but as a potential field for economic growth through new commercial ventures and industry development. Agencies are being asked to promote and strengthen an EO industry whilst continuing to maintain the strong scientific and technological foundation necessary.

#### Open Data Policy

Data policy changes, especially in the United States and Europe, have been critical and are influential in leading to a significant trend across all agencies. The most important policy change was the USGS adoption of free and open Landsat data in 2008.[[1]](https://d.docs.live.net/0ab3be80ed1e9d36/CSIRO/EOITCP/CEOS/FDA/FutureDataArchitecturesReportOutlinev0-7Master.docx#_msocom_1) This allowed International Collaborators on the Landsat Missions to change business models, and to move from being image resellers to being data scientists and providers. Other agencies including INPE and JAXA also championed these changes, however the global reach of Landsat data meant that the US decision had the widest implications.

Within Australia, changes in government policy further supported the direction of free and open data. Agencies such as Geoscience Australia were able to support the development of simple but effective open licences under the Creative Commons framework, and to adopt and apply those licences to their Earth observation data distribution. Resources previously committed to licence management and manual distribution of products were able to be re-focussed on scientific exploitation of Landsat data.

NASA has been operating under such data policy since 1990; other organizations are moving towards such policies in recent years. We have seen the recent movement from Europe (e.g. Sentinel data) and Japan (e.g. mid to coarse resolution data) to provide free and open data. The most recent example is TerraSar-X releasing data for science use if <18 months old.

Fewer hurdles to data access implies broader use of data. This results in a much higher return on investment by organizations from their spaceborne and ground systems’ assets. However, this also means higher workloads for data systems. It can also be very difficult to track the resulting impact once it leaves the confines of the agency. There is considerable business model innovation occurring across many agencies and users of EO data as the impact and value of EO data now readily available is understood.

#### Open Source Software Policy

In addition to free and open data, it is also desirable to have free and open access to tools that facilitate use of data. Open source software policies help with this. A draft open source policy has just been released (March 2016) for public comment by the U.S. Federal Chief Information Officer (see<https://sourcecode.cio.gov/>). The CEOS Data Cube infrastructure depends on open source software for the ingestors (to build data cubes), the APIs and the user interface tools (to interact with data cubes). It is believed that open source software will stimulate application innovation and the increased use of satellite data.

#### Commercial interactions

NASA, NOAA and the USGS conduct a large part of their Earth observation activities through contracts with commercial entities. The European Space Agency and the Japanese Aerospace Exploration Agency (JAXA) conduct their activities through commercial entities as well, even though there are differences in the natures of contracts among the different countries. There is a distinction between commercial entities working under contracts with government agencies for development and operation of observing systems and data systems, and other commercial entities that apply the resulting information to some self-sustaining profit-generating activities. The Federation of Earth Science Information Partners (ESIP) is an example of both types of commercial entities collaborating with government and university organizations. From the point of view of data architecture, commercial interactions have an influence on standards for interoperability, among other things.

#### Distribution of scenes, granules as unit of storage

* + Files and file download is the established unit of distribution, but this is changing
  + Discovery of files. Quality metrics per file
  + Data Security/Custodianship
  + Economic growth from EO

### Application architecture

* + Application Types
  + Application development - who does what? Role of space agency, user, industry, Govt, Science agency
  + Development tools/methods
  + Interactions between applications
  + Quality Assurance and validation
  + Discovery

### Data Architecture

#### Calibration

Improved calibration science, and a measurement philosophy, became a focus for Geoscience Australia (GA) from 2008 as the agency re-focussed its EO efforts toward scientific exploitation of data, with surface reflectance as a basic measurement.

In collaboration with eminent remote sensing scientists at the CSIRO, GA was able to develop physics-based approaches to the calibration of Landsat images to produce standardised measurements of surface reflectance. In this approach corrections are applied for atmosphere, illumination conditions and the reflectance characteristics of the land surface. Corrections for terrain illumination were not possible at that time due to limitations in the quality of digital elevation models (particularly the spatial accuracy of DEMS when dealing with the first spatial derivatives of elevation).

GA also focussed on ensuring the absolute geographic accuracy of its data. With the support of JAXA’s ALOS-1 PRISM data stream and new methods to georegister entire image swaths (reference), Australia was able in 2010 to develop the Australian Geographic Reference Image (AGRI) (reference). AGRI has since provided a consistent, high precision reference frame of known accuracy. Working with the USA and ESA, GA has been able to use AGRI to improve the ground registration in Australia for Landsat and Sentinel missions.

#### Automated processing and standardised products

Automation of processing and a movement from bespoke to standardised image products was a key direction for GA from about 2009-2011. This work was supported by the Environment department which required continental coverage of imagery for forest extent monitoring to support Australia’s national carbon accounts. High costs of manual processing and empirical image-balancing to produce mosaics for carbon accounting were reduced through the bulk supply of calibrated images.

Ancillary data are used in the calibration (and later validation) phases of data preparation. With automation and improvement of the calibration and application science has come the need to ensure that ancillary data are also available from the many sources required in a timely manner to support product creation. Moreover, automation has allowed greater iteration and validation leading to improvements to applications and products.

* + Data - Observations, Products, Ancillary data (for cal/val etc)
  + Data Management
  + Cal/Val process - hmm, not sure what architecture category this needs to be in, brain fade, look at later.
    - ●**Role of Cal/Val FDAs** - Typical flow of data and information that has been used in NASA projects is: 1. Science teams develop calibration software, and calibration parameter files (e.g., Radiometric Look-up Tables) and deliver to processing teams; 2. Processing teams use software to generate products and send to science teams for Quality Assessment (QA); 3. QA metadata are included in products and delivered to archive data centers. In some cases, subsets specifically designed to support validation, covering ground truth sites are maintained in the archives for use by validation teams (e.g., The Oak Ridge National Laboratory (ORNL) Distributed Active Archive Center has subsets of MODIS data covering global validation sites).
  + Diversity and Volume of Data
  + Access mechanisms - data download, web browsing, open data

### Technical Architecture

* Services, Interfaces, Components (not how they are implemented which varies per agency operations but the building blocks that must exist (and do) for an Agency and for CEOS Community interoperability - again restricted to those bits relevant to this report):
  + Standards - standards might be best covered in each component section. E.g. discovery - metadata standards, data management - file formats
  + Computational Platforms
  + Networks

## Trends in User Requirements

* User demand trends
  + Ease of access and use - on-demand, online, any scale, affordable (at desired scale) - result of Data Giants changing user expectations Google MAps, Earth Engine, Cloud community (commodity petascale computing)
    - ○ JAXA WDTMi-R&D is an activity to develop technologies enabling WDTMi-Core and WDTMi-DS. In this activity, JAXA has conducted research and development to realize the common infrastructure considering user convenience beyond conventional means of delivery in terms of satellite data.
  + Diversity of applications - new science, new algorithms, new scales of processing (e.g. WoFS) -> new applications
  + Integration of sensor types and other data types (in situ geospatial, economic and demographic variables, climate variables, etc)
  + Real-time applications
  + Societal benefit - EO role in the digital economy, Govt useful products - continental scale, contribute to global
    - ○ JAXA WDTMi-DS is a supporting project creating businesses or services (downstream industries) utilizing marine-related information and data. This project is paired with WDTMi-Core and WDTMi-DS utilizes information and data provided by WDTMi-Core. WDTMi-DS supports development and advancement of conventional businesses and services as well as activities for value-added operators which create/provide novel services.
  + Responses to application changes - forest monitoring, food, disaster management have all identified issues relating to user and intermediary awareness, understanding and capacity to exploit data.
  + Open source tools and open data

## Trends in the EO systems architecture

* EO system architecture trends
  + Data VVV
  + Many more missions
  + More sensor types, (radar, hyperspectral\_
    - ●**Multi-sensor/ multi-temporal data** - MEaSUREs ESDRs and NOAA CDRs examples of long time series; also refer to ECVs - CEOS Database; multisensor datasets can arise from the same platform or multiple platforms; A-Train and other constellation examples;
  + Higher acquisition rates
  + Higher spatial resolution
  + HPC and Cloud
  + ARD

Partnerships

Australia’s development of the Australian Geoscience Data Cube from 2013 was through a consortium of agencies including the CSIRO and the NCI as well as Geoscience Australia. The principles behind the partnership were … (AL to fill in)

JAXA ○ WDTMi-Hub represents international collaboration activities (e.g. collaboration with similar program such as Copernicus in Europe) planned in WDTMi such as satellite data distribution including other satellites’ than JAXA’s.

High performance data and computing

High Performance Computing became available for the management and analysis of Australia’s Landsat data collections from 2011-2013 with funding to ‘Unlock the Landsat Archive’. This work applied the automated production systems to the entire Australian Landsat collection, which was moved to the Australian National University National Computational Infrastructure (NCI) to complete this work. Geoscience Australia became a partner in the NCI in 2012(?). The HPC environment allowed the data to be held on disc, rather than tape, directly attached to large computing resources.

#### Move to common infrastructure for all sensor and observation types

○ JAXA WDTMi-Core is a project to develop common infrastructure which integrates/manages/provides information/data of models, analytical results, and data through ocean related satellites and In-Situ observation. The goal of this project is to provide integrated data of both international ocean-observing satellites and In-Situ observations. The following is an example of sea surface temperature as satellite data which is planned to be provided through WDTMi-Core.

# 3. The challenge of changing user expectations and increasing EO data volume, variety and velocity on EO systems architecture

* Pressure points in architectures due to Data VVV
  + Scale
    - Volume pressure
      * **large data volumes - volume metrics** - Archive growth in EOSDIS since 2000 is shown in the figure below. NISAR (launch year 2020) is expected to produce 26 PB/year. SWOT (launch year 2020) is expected to produce about 14 PB/year. Include data volume expectations from other missions here as well as NOAA and USGS data volumes. It would be useful to develop a summary chart that shows the aggregate data volumes from all CEOS members and projecting to the future. The large data volumes imply a need for very accurate search services so that users are not overwhelmed with the volumes they need to download.
    - Velocity pressure
    - Variety pressure
    - Summary of the total impact?
      * Data volumes to large to move to local analysis platform
      * Local analysis platform not large enough
      * Cloud allows affordable very large scale processing on-demand - dat a production vs data acccess(storage strategies)
  + Network distribution pressures - see Global data flows
  + Discovery
  + Quality
    - **Ensuring and conveying data quality** - Users are interested in knowing whether a particular dataset is suitable for their use. When there is a choice of similar datasets users need advice on which works best for them. Because data quality is subjective and dependent on the science or application focus, how do we provide a generic/cross cutting indicators of potential data quality? QA4EO has developed a set of principles regarding data quality. ISO 19157 deals with standards for data quality metadata. NASA’s Earth Science Data System Working Group (Data Quality Working Group) and the ESIP Federation’s Information Quality Cluster are analyzing a number of use cases and making recommendations on how best to capture, describe, and enable access to and use of the data quality information.
  + Analysis to Data rather than data to Compute
    - ● **Moving analysis to the data** - Given the large and growing amounts of data, it would be useful to move analysis software to where the data are held and transmit to users only reduced volumes of the results of analyses. This will help in reducing network loading as well as the storage resources needed at the user's’ facilities. Information Technology security challenges remain regarding how to make archives more robust and accommodate diverse types of analysis software from a large number of users. There are also challenges in prioritizing use of scarce resources (hardware, computing, etc) by a large number of users. Charging policies (as opposed free supply of resources) may be considered.

###### Visualisation in web time and many different types of devices

* + - **ease of access - visualization, near-archive services**. Because different users need different types of access with a wide variety of access services, a diverse set of access services needs to be available. Users with low-bandwidth internet access should be able to access, visualize, analyse and download data. NASA’s Global Imagery Browse System (GIBS) with its WorldView client is a good example of access and visualization of large quantities of data with fast response. NASA’s Giovanni (**G**eospatial **I**nteractive **O**nline **V**isualization **AN**d a**N**alysis **I**nfrastructure (<http://giovanni.gsfc.nasa.gov/>) is a good example of a service that provides analysis and visualization for a variety of datasets.
* The impact of changing user expectations
  + Ease of access
  + Delivery channels (see CEOS slide)
    - Cloud availability - large scale computation affordable
      * ●**Cloud service**s - NASA has been implementing a few use cases pertinent to Earth science data to understand issues and assess costs involved . Cloud services are beneficial when there are unpredictable or occasional peaks in computational and/or storage loads. An example of an occasional peak in computation is when a full mission dataset is preprocessed due to an improved algorithm. Cost of retrieval of large quantities of data from commercial cloud storage is a concern. The CEOS Data Cube project is collaborating with Amazon to test the AWS infrastructure and how it can best support the Data Cube plans.
    - Ubiquity of geospatial web applications
      * ●**Access mechanisms** - Browsing of images from archives in the past had been limited to thumbnail size, low resolution images. Recently, full resolution full resolution image browsing has been implemented, making it possible for users to look at specific details in their regions of interest before acquiring the data. An example of such an implementation is NASA’s Global Imagery Browse Services (GIBS), which provides a scalable, responsive, highly available, and community standards based set of imagery services. These services are designed with the goal of advancing user interactions with EOSDIS’ inter-disciplinary data through enhanced visual representation and discovery. It provides for display of multiple layers of data including a map overlay. See<https://earthdata.nasa.gov/about/science-system-description/eosdis-components/global-imagery-browse-services-gibs> for details.
    - Just my data - Global, Continent, Paddock, Backyard
      * **search strategies/search relevance** - users want to get everything relevant to them and nothing else (i.e., high precision and recall). NASA’s Earth Science Data System Working Group (Search Relevance WG) is examining various issues related to relevance such as heuristics, spatial relevance, temporal relevance, and dataset relationships. [GEO Data Access Broker](https://www.earthobservations.org/documents/meetings/201401_geo10_ec/GEO-DAB-(Nativi%20et%20al.).pptx) is an example of how metrics are used to rank search results from a widely distributed set of Earth science data sources. As noted above, increasing data volumes imply a need for very accurate search services. Convenient ways of accessing datasets related to other datasets of interest should be available. Theme based “bundles” or “virtual collections” would be useful.
    - On demand
    - Companies trying to meet “daily coverage” or more - temporal increase
      * ●**Companies trying to meet “daily coverage” or more** - In recent years several commercial organizations have been entering the market with constellations of small satellites providing coverage of the Earth daily or even more frequently. See<http://www.thespacereview.com/article/2534/1>.
    - My algorithm
    - ● Commercial value add demanding surety of data/service supply
    - ●**Data policies, copyrights, user registration, and privacy -** Clear understanding is needed regarding policies regarding data and information, copyrights, user registration and privacy. Can more agencies agree to stamp their data as “GEOSS Data Core”?
    - **greater emphasis on adding value - various kinds of users exist**. Some are providers of value added products to other users. Architecture should support users at all levels of a value chain.
    - Open Data and Open Source
      * Changing business models - e.g. google and others have a massive public goods value add and only monetize a very small portion of it (and still make a huge amount of money). Without the public goods they couldn’t make money.
      * ●**Increasing number of Open data sources and open APIs** - Relates to free and open data policies (see first bullet). Open data sources imply servers that allow independent client access. These may be configured as a loosely coupled federation with servers offering search services (using standards) on metadata and varied access to data files and a wide variety of independent clients (that implement the search standards) that can access the servers.
    - Data Formats
      * Most of the data are stored typically in forms convenient to producers. These forms are not necessarily convenient for users to access. Different users need different access mechanisms. Some want access via the traditional granule download access; some want granule level access after some subsetting (time, space, bands); some want pixel level access, i.e., without regard to granule boundaries. It is challenging to provide access to spatial subsets of long time series of data. How can we provide enough descriptive information to the users to enable the types of access they need?
  + Types of users - Govt, Global, Science, SMEs
    - Large increase in capable skills but not deep EO science skills particularly related to the economic (business building) growth opportunities being sought
    - ●**Coverage of wide range of disciplines** - Earth science by its very nature involves many disciplines. The data need to serve users in various disciplines individually as well as users working on interdisciplinary science and applications. This implies seamless interoperability to better serve users. Best practices for the interuse of data formats and agreed upon profiles or “flavors” of existing standards become very important. There is a greater need for tools that support multiple file formats seamlessly.
    - **Diversity or user community and disciplines to be covered by data systems** - NASA covers many Earth science disciplines through its discipline oriented DAACs (see<https://earthdata.nasa.gov/about/daacs>). World Data System - list of members would be a good source to list disciplines and show diversity challenge (see<https://www.icsu-wds.org/community/membership>). Also look at the membership list of the Federation of Earth Science Information Partners (ESIP) (see<http://www.esipfed.org/esip-member-list>). GEO Societal Benefit Areas are illustrative of the diversity of user communities. Do the diverse user communities imply an increasing need for interdisciplinary use of data collections? Interdisciplinary use of data collections imply a need to facilitate interoperability across data file formats.
  + Analysis for the masses - simplicity of use
  + Velocity opportunity
* Identified aspirations and open problems
  + to support computation localised with data for large data volumes to deal with data flow issues
  + Investigations in the use of Cloud and HPC environments
  + Economic growth expectations (regional business building) - EO role in the digital economy (● Increasing demand to see applications / jobs created etc)
  + Network bandwidth
  + **capacity building** - Data providers will need to support development of training packages to help in capacity building.
  + - **cost-effectiveness** - Hardware costs going down have helped us in handling ever increasing amounts of data, even though the data collection rates and users’ capability to ingest and analyze data also increase commensurately. Challenges are in other costs associated with data management. Evolutionary nature of data systems is essential to stay on top of this problem. Occasionally changes need to be revolutionary. For example, during 2005-2008, EOSDIS implemented a number of changes in an activity call Evolution of EOSDIS Elements. As a part of this effort, most of the data were migrated from robotic near-line tape-based archives into online RAID-based archives; software was re-architected to reduce the number of line of code significantly, and some of the functions were redistributed across the different elements of EOSDIS. This resulted in nearly 30% reduction in annual operating costs.
  + - **standardisation and interoperability** - Standards are key to interoperability. IT would be useful to list all types of entities for which standards should exist and be applied, and list current applicable standards . An example of such a list is available at<https://earthdata.nasa.gov/user-resources/standards-and-references>. It is preferable to have a small number of standards to facilitate search and access to data in an interoperable manner (search standards, controlled keywords, metadata). We also need a more diverse set of access standards to support actions such as subsetting, file format conversions, spatial projection conversions, tile access, and pixel level access. Questions to be answered: what is analysis ready data? can a definition/standard hold across disciplines for analysis ready data?
  + Single Sign On
  + Monitoring & reporting capabilities;

## Constraints

* Data sovereignty, trust and economic growth
* Budget constraints - new business models (massive public good, monetisation, redistribution of costs, capturing and communicating metrics for CEOS agencies of use and value add)
* Capacity constraints - volume, computation, bandwidth
* Specialist skills (geometric and radiometric correction, etc)
* Current architectures are unlikely to respond to this and CEOS community needs to respond as a whole to ensure Global challenges can be addressed and the full potential of EO satellite data realised in Govt, Global and industry segments.

# 4. The Future of EO Data Architectures

* Recast the architecture in Chapter 2 describing where it has changed in order to address the challenges in Chapter 3

## Business Architecture

**i. Discovery and Access**

1. Machine Level Discovery and Access: All data are available for search and access with machine-callable APIs.

2. Cross-agency Discovery: Cross-agency data discovery is seamless.

3. Dataset Selection Guidance: Guidance is available on data selection based on fitness for purpose.

4. Metadata Naming Conventions: Key metadata follow standard naming conventions for Variables, Platforms, Instruments, Spatial Resolution, Temporal Resolution

5. Virtual Collections: Virtual collections can be organized / oriented around a science problem.

**ii. Usage**

1. Intelligent Tool Catalogs: Intelligent tool catalogs automatically suggest data analytics / visualization tools to work with the data.

2. Live Data Citation: Publications are linked to data and tools that allow interactions with the data.

3. Mobile Data and Processing: Data and processing move transparently as necessary to achieve optimal performance.

4. Quantitative Quality: All data have quantitative measures for data quality.

5. Reproducibility: Scientists can reproduce other scientists’ research results with high precision.

6. High-Quality Documentation: Concise, Comprehensive and Consistent documentation exists for all data variables.

7. Capacity Building: A rich set of capacity-building and translation mechanisms exists to facilitate leveraging data for use by people with limited literacy in science and advanced technology, and/or English.

8. Data Analysis at Scale: Users are able to analyze the entire data record for any data variable over any arbitrarily defined area.

9. Dataset Upgrading: High-value datasets are upgraded as necessary to fully support in the rich capabilities available in the data systems.

**iii. Integration**

1. Data: NASA data can be easily compared, merged, fused and/or assimilated with (depending on the science case) data from other agencies, nations and other entities

2. Tools and Services: Tools and services within the community are easy to use in concert.

3. Sharing: Scientists are able to share all scientific resources (data, tools, results, workflows, contextual knowledge)

* + Move to SR products, ARD, higher level products for ocean, land, forest, food, ...
  + Consolidation of common architectural components across collections (same delivery mechanism/experience for all sensor types)
  + Increasing partnerships - across Govt agencies (users and custodians roles), with HPC/Cloud providers
  + Theme (e.g. ocean, food) based collections (virtual) rather than sensor based
    - ●**Theme based collections of data** - Collections of data addressing specific themes such as hurricanes, fires volcanic eruptions, algal blooms, sea-level rise, etc. are useful to the research and applications community. Typically data for a given theme are distributed across multiple archives. It is not necessary that the data themselves be gathered to a single location to meet these thematic needs, as long as they can be pointed to from a single location (e.g., from a file of URLs) . The NASA ESDSWG Virtual Collections Working Group has explored ideas about such “Virtual Collections”.
* Two views to consider (minimum) the impact on 1) a CEOS Agency and 2) the CEOS Community as a whole:
* Discovery
  + Of pixels not files/granules - seamless interaction patterns
    - Visualisation
* Quality
  + Assurance
  + Communication of quality
  + Filtering by quality of pixels
* Timeliness of data availability and support of realtime applications (bush fires, algal blooms)
* Multiple modalities - integration and harmonization of multiple satellite sensors and products (e.g. includes common ancillary data like DEMs)
* Dynamic products and Provenance
  + Virtual themed collections served from the same underlying physical collections
    - **Trustworthiness - provenance, reproducibility** - Architecturally this means being able to convey to users the full lineage of data, links to documentation, software, relevant publications, etc.
* International interoperability of service providers
* Utility Services
  + Format conversion (for legacy systems)
* Multi-sensor environment
* ARD
* Standardisation, interoperability and integratability
* ●**Moving to ARD (Analysis Ready Data)** - Definition of ARD from<http://www.datacube.org.au/wiki/Category:Analysis_Ready_Data>: Analysis Ready Data (ARD) is satellite data that have been processed and organized so users are not required to invest time and resources in specialized skills to apply corrections for: 1. instrument calibration (gains, offsets); 2. geolocation (spatial alignment); and 3. radiometry (solar illumination, incidence angle, topography, atmospheric interference). While a large number of users will benefit from ARD that are defined thus, we do need to examine whether the definition holds across science disciplines/applications. What is analysis ready does depend on the type of analysis involved. An initial ARD draft document was started by the SEO and is under review by the LSI-VC. The “high level” definition is “*Analysis Ready Data (ARD) are satellite data that have been processed to a minimum set of requirements and organized into a form that allows immediate analysis without additional user effort. “*
* Data Distribution
  + Includes network challenges for some countries and scale out strategy for CEOS agencies
  + Multiple channels for customer access (see CEOS ppt image)
  + Common web standards and protocols
* Hosted Cloud, HPC, and Local processing
  + Global Data Flows study?

# 5. Conclusions

## Benefits

Securing CEOS objectives (link back to Context strategic guidelines)

## Recommendations

Working with GEO (Applications)

Working with OGC (Service interfaces)

CEOS Working Groups - WGISS

Space Agencies

International Data hubs

CEOS Community

Capacity building? Aligns with strategic guidance

The immediate development of the AGDC is likely to focus in the areas of:

1. International collaboration

2. Continuing the preparation of data to ingest into the data cube, including working with CEOS agencies to progress Analysis Ready Data concepts and practice, and establishing ingestion approaches that deal with radar, geostationary, and new LEO-optical data streams

3. Science and research partnerships to calibrate data streams, and to generate new products from large time series of EO data

4. Increasingly systematic use of high performance computing environments, including large data holdings (6.5 PB by 2020), research and industry ‘sand-pit’ environments, and mechanisms to facilitate innovation and commercial spin-offs from the data cube - such as facilitating ‘mini-cubes’ in commercial cloud environments.

5. Robust collection management capturing the lineage of processing and quality and uncertainty assessments, to produce analysis ready data, and routine production of data cube products

6. Close engagement with users, initially agencies in the Australian Government to enable business improvements through application of the data cube.

7. Forward looking research and development in areas such as Discrete Global Gridding Systems, 3D data cubes, integration of point clouds with grids, and interfaces with models.

**From the SDGC Global Data flows - some overlap with expected recommendations on architecture so included for consistency whilst FDA builds up its final recommendations:**

CEOS will provide coordination between space agencies to improve the space data products and access to the products.

· Establish definitions of ARD products and provide guidelines for their use.

· Evolve ARD products to meet current requirements for scene-based and future requirements for Data Cubes.

· Support the development of methodologies to use Data Cubes.

· Increase coordination among space agencies for cross calibration of products.

· Increase coordination among space agencies to achieve consistency and compatibility of optical and radar ARD products.

· Improve access to data directly from space agencies and through external partners.

· Increase access to older data sets for time series analysis and baseline classifications.

· Improve discovery tools appropriate for multi-sensor searches and for discovery and access to ARD products.

----

## Revised Architecture guidelines

CEOS Agency

CEOS Community

# 

# 

# Appendix A - Current Agency EO Systems Architecture Projects

## EUMETSAT

EUMETSAT the European Organisation for the Exploitation of Meteorological Satellites executes within the period of 2016-2018 a set of internal Pathfinder projects based on new big data technologies for the upgrade of its EUMETSAT multi-mission ground segment infrastructures, focused on facilities providing Data Services. Contracts will be placed with industry to make the solutions future proof and being a precursor of the EUMETSAT multi mission ground segment infrastructure. The Pathfinder projects will enable EUMETSAT to deliver enhanced data services to its Member States from 2017/2018. Some features addressed by EUMETSAT as part of these Pathfinder projects will be of relevance for integrated ground segments in view of standardisation and interoperability.

The following Pathfinder projects are executed by EUMETSAT:

* PF-I: on line data access (OLDA);
* PF-II: Delivery of high volume of data via EUMETCast Terrestrial;
* PF-III: Development of Web Map Services;
* PF-IV: Format conversion toolbox;
* PF-V: Hosted processing;

Business Architecture

Application Architecture

Data Architecture

Technical Architecture

**PF-I: on line data access (OLDA)**

The Pathfinder on on-line data access (OLDA) will address a new service to access EUMETSAT data and products, between the established NRT dissemination and the long term archive ordering service.

The following high level requirements are addressed:

● Provide immediate direct download of data and products “as is”;

● Simple User Interface, including:

○ Single Sign On via EUMETSAT portal;

○ Search capabilities using filters (time of interest, spatial extent, mission);

● Monitoring & reporting capabilities;

● Standard based application programming interface based on HTTP/REST

● Service Level management for different user communities.

Interactions with other Pathfinder for data access such as the Tool Box, Web Services and Hosted Processing.

### PF-II: Delivery of High volume of data via EUMETCast Terrestrial

EUMETCast Terrestrial is a “push” multi-casting system (like EUMETCast Satellite) for near real time data dissemination that uses high bandwidth terrestrial networks rather than satellite transponders. It is suitable for the dissemination of high volumes of data but to a small number of users.

This service is planned to be operated in the future complementary to the push service using telecommunication satellites (EUMETCast satellite), suited for distribution of high volumes of data to many users.

The following high level requirements are addressed:

● Assess/confirm the relevance of the EUMETCast terrestrial concept:

○ For the delivery of high volume of data to international partners, on a best effort basis;

○ For the delivery of large or additional Data sets (e.g. level 1 data) that most users do not need, but a few require, in particular in the context of Copernicus/EU programmes (e.g. Level 1 Sentinel-5P data) ;

○ For the delivery of temporary, trial or project-related non-time critical datasets to NMSs or expert users (e.g. MTG study data, Climate data sets for evaluation, etc.)

● Demonstrate the capability of the system to serve as a backup to EUMETCast satellite for access to time-critical data (in particular Rapid Scan Service data) in case of loss of access to EUMETCast satellite (e.g. by a NMS during a critical weather event) and compare this capability to the capability of the OLDA;

● Provide a single and harmonized user interface across EUMETCast terrestrial and EUMETCast satellite exploited as a single-point managed service;

● Explore contractual arrangements where GEANT acts as a single point of contact for the involved terrestrial networks being used.

### PF-III: Web Services “EUMETView”

‘EUMETView’ established in 2015 provides today web imagery and visualised products through an Open Geospatial Consortium (OGC) standard interface (WMS), facilitating the overlay with other sources of geospatial data for EUMETSAT essential data.

This pathfinder is targeted to provide:

● More advanced, “licensed” Web Services providing non essential imagery in full temporal and spatial resolution to licensed users

● Enhancements with:

○ Download capabilities (Web Coverage Services)

○ Information services (Web Feature Services)

○ User interface and user support information

○ Scalability and meeting agreed service levels

○ Monitoring and reporting mechanisms.

### PF-IV: Format conversion toolbox

This Pathfinder will support internal and external users to work more easily with EUMETSAT provided data.

The objective of the tool box is to enable conversion of EUMETSAT proprietary formats to other community-desired formats and vice versa. This is planned to be done by provision of the actual software library or web based services for conversion and geographic sub-setting.

For the scope of the Pathfinder, the toolbox will include some of the most popular standard formats for the purpose of a representative demonstration.

● Conversion of EUMETSAT proprietary formats MSG Native, MTP Native, Metop Native and some BUFR, GRIB products to other community-desired formats (NetCDF, geoTiff) and vice versa , including the conversion into well known projections (WGS84);

● Implementation of the NetCDF format following the Climate and Forecast conventions for metadata;

● Inclusion of specific geographic sub-setting functions in synergy with the pathfinder on hosted processing;

● Verification of converted formats with a selected set of GIS tools;

● Provision of a web service supporting user access to the toolbox;

### PF-V: Hosted processing

The Pathfinder on Hosted processing will showcase the ability to apply simple processing software/algorithms to EUMETSAT data available and explore the related technical and service elements.

The objective is to reduce the volume of data transferred to users, to avoid unnecessary duplication of data on the user side and share expertise across EUMETSAT Member States on selected algorithms and processing software and maximise their use.

The Pathfinder will rely and make use of the other pathfinder concepts such as the Online Data Access and the Toolbox.

## 

## 

## NASA

### - NASA CMR/UMM

- Metadata is used in all aspects of NASA’s Earth Science data lifecycle from the initial measurement gathering to the access of data products by end-user. Missions use metadata in their science data products when describing information such as the instrument/sensor, operational plan, and geographic region. Acting as the curator of the data products, NASA Distributed Active Archive Centers (DAACs) employ metadata for preservation, access and manipulation of data.NASA has recently developed a single, shared, scalable Earth Science Metadata repository, the Common Metadata Repository (CMR), as a high-performance, high-quality metadata engine for next-generation Earth Observing System Data and Information System (EOSDIS). The CMR’s need for dealing with multiple metadata formats and underlying data models led to the development of the Unified Metadata Model (UMM), a common data model across metadata held in the CMR. The current version of the UMM identifies Science Collections, Science Granules, Science Services and Meta-Metadata concepts. To foster interoperability with other agencies and international partners, NASA has applied uniform international standards to Earth science metadata represented in the UMM using ISO 19115 Geographic Information Metadata format.

-

### - NASA - GIBS

- NASA developed the Global Imagery Browse Services (GIBS) system as a core EOSDIS component which provides a scalable, responsive, highly available, and community standards based set of imagery services. These services are designed with the goal of advancing user interactions with EOSDIS’ inter-disciplinary data through enhanced visual representation and discovery.

- These advancements are realized in the following ways:

- Improved Approachability & Extended Reach: Imagery greatly improves the usability of NASA Earth science data to new communities and improves cross-disciplinary data discovery through full-resolution, “no boundaries” (or “granule-free”) interaction patterns.

- Cohesive Approach to Imagery: As a core EOSDIS component, GIBS integrates with other core EOSDIS systems, components, and processes to provide a primary, authoritative source for EOSDIS imagery.

- Improved Cross-Discipline Research: GIBS leverages science expertise and interoperable standards to provide science-based products that enhance cross-discipline discovery and analysis.

## 

## 

## JAXA

WDTMi consists of 5 activities;

l Common-infrastructure (J-core) development project WDTMi-Core

l Downstream development project WDTMi-DS

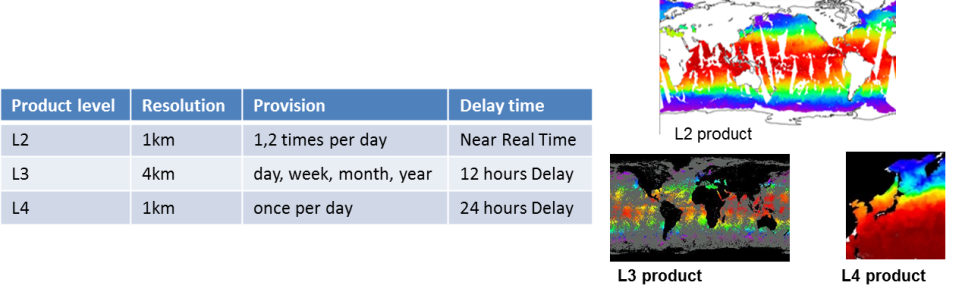
l Research and development project WDTMi-R&D

l User-community operation WDTMi-Com

l International collaboration WDTMi-Hub

WDTMi-Core:

* + WDTMi-Core is a project to develop common infrastructure which integrates/manages/provides information/data of models, analytical results, and data through ocean related satellites and In-Situ observation. The goal of this project is to provide integrated data of both international ocean-observing satellites and In-Situ observations. The following is an example of sea surface temperature as satellite data which is planned to be provided through WDTMi-Core.





WDTMi-DS:

* + WDTMi-DS is a supporting project creating businesses or services (downstream industries) utilizing marine-related information and data. This project is paired with WDTMi-Core and WDTMi-DS utilizes information and data provided by WDTMi-Core. WDTMi-DS supports development and advancement of conventional businesses and services as well as activities for value-added operators which create/provide novel services.

WDTMi-R&D:

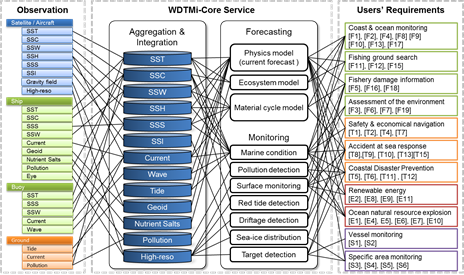
* + WDTMi-R&D is an activity to develop technologies enabling WDTMi-Core and WDTMi-DS. In this activity, JAXA has conducted research and development to realize the common infrastructure considering user convenience beyond conventional means of delivery in terms of satellite data.

WDTMi-Hub:

* + WDTMi-Hub represents international collaboration activities (e.g. collaboration with similar program such as Copernicus in Europe) planned in WDTMi such as satellite data distribution including other satellites’ than JAXA’s.

WDTMi-Com:

* + The WDTMi activities described above are thoroughly based on user’s needs. WDTMi-Com is an activity which creates a point of communication with user groups; entities and researchers related to various fields such as environment, fishery, marine safety, natural resources, energy, and security. JAXA has pursued initiative on WDTMi to meet the user’s needs by exchanging opinions with the groups.



## AGDC V2

AGDC Version 2

Continental-scale applications requiring a time-series of data were the final step toward an Australian data-cube. Most importantly, funding to complete a National Flood Risk Assessment (2012-15) required Geoscience Australia to conceive of the data structures, including regular tiling of ‘Analysis Ready’ data, that would be necessary to systematically apply algorithms to large time-series of imagery. This work produced the Water Observations from Space (WOfS) datasets (Mueller *et. al* 2016 doi: 10.1016/j.rse.2015.11.003). WoFS tested every Landsat pixel captured over Australia between 1987 and 2013 for the presence or absence of water, and compiled these into unique surface water maps. The land area of Australia is over 7 million km2., and the total number of observations used in the WoFS analysis is in the order of 10^13.

The code base, which has always been open, has improved in quality and has been shared with others including CEOS and the USGS. This shared code base is seen as a necessary step toward internationally coordinated approaches to the development of Data Cubes to fully exploit the potential of rapidly growing Earth observation data collections.

A key step in the developments from July 2015 has been to move from traditional geotiff file structures to the NetCDF-CF 1.6? Also: collection management; capture of provenance;

Future development directions for the AGDC include the incorporation of new data streams, incorporation of non-EO analysis ready data grids, continuing to support the progress of the Discrete Global Gridding System, and continuing cooperation and integration with the International Community.

# 

# 

# China

TBD

Everything beyond this point is not yet part of the report

# Contributions yet to find a location, be consolidated or identified as duplicate in the report

●**Crowdsourcing** - There is an increasing trend in utilizing volunteer services from large numbers of individuals in making observations and measurements. This may be called crowdsourcing or citizen science. Oxford Dictionary defines citizen science as “The collection and analysis of [data](http://www.oxforddictionaries.com/us/definition/american_english/datum#datum__4) relating to the natural world by members of the general public, typically as part of a [collaborative](http://www.oxforddictionaries.com/us/definition/american_english/collaborative#collaborative__2) project with professional scientists”. Examples of citizen science are: Community Collaborative Rain, Hail and Snow Network (CoCoRaHS), NOAA’s National Weather Service (NWS) Cooperative Observer Program (Coop), “OpenStreetMap”, etc. Some of the concerns regarding citizen science are, data and metadata standards, quality assurance, authenticity, certification, crediting participants, provenance, and preservation. Considerable amount of work is being done addressing such concerns by The Citizen Science Association (see<http://citizenscienceassociation.org/>). For example, The PPSR\_CORE Metadata Standard has been developed as a data-sharing protocol (See<http://citizenscienceassociation.org/2015/10/09/ppsr_core-metadata-standard/>).