**A FUTURE DATA ARCHITECTURES STRATEGY FOR CEOS**

**DISCUSSION DOCUMENT FOR 31st CEOS PLENARY**

Version 0.4, 27 Aug 2017

# Executive Summary

CEOS agencies have spent billions on space data acquisition and ground segment architectures to make data accessible to their users. This is necessary, but not sufficient, to realize the full societal benefits of these investments – we also need to ensure that Earth Observation (EO) data is provided to end-users with maximum efficiency.

A clear, shared vision of the Future Data Architectures (FDA) for EO is essential to boost the use of EO data for science, operational services, and commercial applications. Joint initiatives of CEOS agencies with a common understanding of FDA will generate significantly more impact by more quickly and efficiently reaching maximum user needs. The outcome of these initiatives should be a clear, complete, and well stewarded EO data offering, against which organizations can confidently invest their resources and leverage their channels to users in an effort to realize benefits beyond even the most traditional application areas.

CEOS agencies must be aware that:

1. The EO user base is expanding rapidly and increasingly includes non-experts whose main interest is the easy transformation of data into information.
2. Data volumes and diversity are also increasing rapidly, which requires alternatives to the traditional data download dissemination schemes.
   1. Alternatives include a coordinated effort in developing easy-to-use data analysis platforms and tools, such as Data Cubes, in which the user has access to processed and organized data for the transformation of Analysis Ready Data (ARD) into relevant information.
3. Data platforms can co-exist or interact with an increasing number of EO data service environments provided by the big commercial cloud providers with huge capacities for data storage and processing.
4. Only full commitment to user needs and expectations can assure that CEOS agency initiatives will be successful. Therefore, user acceptance and user behavior need to be rigorously measured in any FDA activity.
5. By providing ARD – optical imagery and others – CEOS agencies can facilitate the use of EO data. The same is true for applying Open Source principles for easy tailoring of software.
   1. Areas of high impact include coordinated and prioritized initiatives for standardization and interoperability between similar sensors

All activities proposed for CEOS agency implementation by the FDA Ad-hoc Team are a reflection of these insights and are summarized into five core initiative areas:

1. CEOS Analysis Ready Data (ARD)
2. Interoperable Open Source Tools
3. Data, Processing, and Architecture Interface Standards
4. Analytical Processing Capabilities
5. User Metrics

# Introduction

This document has been prepared for discussion and endorsement by the 31st CEOS Plenary in October 2017. It outlines a strategy for a Future Data Access and Analysis Architectures (FDA) approach for CEOS and its agencies. Consecutive CEOS Chair agencies (CSIRO in 2016, USGS in 2017, and EC in 2018) have identified this topic as a priority for space agencies and for CEOS. The FDA strategy for CEOS addresses both:

* the need for new paradigms in information systems to deal with *Big Data in Earth Observation (EO)* (i.e., with the increasing volume, velocity and variety of EO data);
* the opportunities provided by advances in high performance computing, and in cloud storage and processing systems that are changing user community expectations for accessing and using EO data, and in particular the need to integrate different satellite EO datasets;

Organizationally, this document will show that it is best to keep a FDA AHT in place for the foreseeable future as it works to integrate the activities of several key entities within the CEOS structure – specifically LSI-VC, SEO, WGISS, and WGCapD – and to keep the momentum that has been built.

Additionally, CEOS agency active engagement will be critical in achieving the FDA strategy outlined in this document. Without this, EO data and information will not have the desired impact because of its current limitations in utility, access, and analysis at scale by an increasingly diverse group of users.

# Context and Status

CEOS Principals are referred to the comprehensive report on FDA of October 2016, prepared by the CEOS Ad-hoc Team. It provides a thorough review of the context and the evolution of the user requirements and the technical possibilities. The report suggests that government-sponsored satellite Earth observations are at a key juncture – at which the observation techniques have advanced to have significant potential for addressing society’s grand challenges, and at which the number and diversity of users has increased dramatically. Users are increasingly non-technical, expect easy data access and use of the data, and require integrating EO data and information with other sources. Notwithstanding the progress made in recent years, a number of difficulties in using EO data is still a barrier to realizing their full potential and to properly harnessing “EO Big Data” for societal benefits.

The Ad-hoc Team has continued the analysis through 2017 to prepare final recommendations on the way forward for decision at the 31st CEOS Plenary. In parallel, a number of recommendations from the interim report have been progressed by CEOS groups:

1. SEO-led pilot projects to elicit real-world experience of novel architectures – including the CEOS Data Cube (CDC) – to provide input for further strategic decisions;
2. LSI-VC definition of CEOS standards for the promotion and application of Analysis Ready Data (ARD) using Analysis Ready Data for Land (CARD4L) as the pathfinder;
3. LSI-VC progress on the Moderate Resolution Sensor Interoperability (MRI) initiative, which both proposed an interoperability framework and conducted a case study utilizing a merged Landsat and Sentinel-2 dataset;
4. WGISS naturally continued its work on: discovery search engine optimization; access common standards for interoperability of product formats and application program interface (API) for analytics and data access services; exploration of emerging big data services including cloud computing.

The 2017 analysis by the Ad-hoc Team has focused on defining the substance for a collaborative strategy within CEOS, and how this would best be accomplished within the CEOS Work Plan and structure. This included a survey of CEOS agencies and groups to establish a common set of priorities from the membership. The most commonly suggested responses and desires to facilitate the user uptake for inter-agency collaboration through CEOS were:

* Simplify EO data uptake and application for many users
* Definition and automated delivery of interoperable ARD (e.g., CARD4L)
* Exploit cloud computing for storage and processing (e.g., exploitation platforms and data cubes)
* Algorithm portability
* Solutions that enable scalability from local (e.g., institutional) to global
* Open source tools alongside commercial – complementarity

# Proposed Strategy

The 2016 report proposed that the CEOS strategy achieve:

* de-risking and simplification of EO data for users;
* allowing users to make use of ALL available and relevant CEOS agency data;
* supporting CEOS ambitions in relation to its chosen grand challenges.

The strategy proposed to CEOS by the Ad-hoc Team centers around five core initiative areas:

1. CEOS Analysis Ready Data (ARD)
   * Implementation of CARD4L noting the critical need for progress in the land domain, which is less advanced than marine and atmosphere domains
   * A complementary comprehensive CEOS strategy for Analysis Ready Data, incorporating coastal, marine and atmospheric domains in a manner complementary to CGMS and WMO activities
2. Interoperable Open Source Tools
   * Provision of interoperable open source tools to exploit advances in technology and meet user demand, including increasing contributions to the CEOS Data Cube (CDC) initiative and utilization of data platform infrastructures
   * Ongoing program of ‘pilot projects’ for new technologies, undertaken at low-risk/low-cost, to ensure Principals have real world evidence about what suitable future projects might be
3. Data, Processing, and Architecture Interface Standards
   * Standards for portability of ‘algorithms’ that exploit satellite EO data between different exploitation platforms, whether they be based on the Open Data Cube technology, proprietary software, or Agency/country specific tools
4. Analytical Processing Capabilities
   * Implementation of modular and easy-to-use analytical processing capabilities in full computing environments for time series and other analysis
5. User Metrics
   * A data use metrics framework through which agencies can contribute to a ‘sector wide’ view of how EO data is actually being used, not just how much is being downloaded

# Initiative #1: CEOS Analysis Ready Data (ARD)

A comprehensive strategy for ARD standardization and production by CEOS agencies is a fundamental part of the proposed way forward for CEOS as ARD product generation becomes more of a core business for CEOS agencies and as technology and required ancillary data sources become more available. CEOS ARD seeks to address the need for removal of technical obstacles for users to access and apply data, and to ensure that *‘every available pixel counts’* – such that data from all space agencies, large and small, might benefit from this FDA strategy.

This is an obvious standardization and coordination role for CEOS and is in response to what was the overwhelming priority for CEOS action suggested by the agencies responding to the FDA AHT survey. Multiple CEOS agencies have already initiated, or are on the cusp of defining, standards for ARD, and it is of paramount importance, to ensure optimal interoperability outcomes for the next generation of missions and information systems, that CEOS act decisively and quickly to establish CEOS ARD standards in the context of a broad ARD strategy. We anticipate strong support and participation for this strategy based on feedback from members to date. CEOS agency ARD product generation is intended to be additive and is not intended to preclude other individual agency product offerings.

Different philosophies may emerge for the routine production and provision of ARD in agency product pipelines including ‘on-the-fly’ product generation; but all will benefit from agreement on the minimum standards for fundamental corrections and formats so that space agencies of all sizes, and users of all kinds might participate in the provision or uptake of data to the common standard. This will greatly facilitate space agency contribution to the grand challenges CEOS seeks to support – such as the UN SDGs, disasters, and climate – and support the ambitions for mainstreaming EO into society among varied and non-technical users. Further adoption of this strategy facilitates integration of public good data, other CEOS agency data, and commercial data all working together.

The CEOS ARD strategy will build upon the work already underway within the LSI-VC and embed it within a broader strategy and schedule of relevance to all CEOS missions and user communities. The CARD4L (CEOS Analysis Ready Data for Land) standards drafted by LSI-VC will serve as a pathfinder with which to secure significant user feedback. The costs involved in agencies adapting their production pipelines to include CEOS ARD standard products are significant and will not be taken lightly, and certainly not without substantial evidence of the cost-benefit to each agency making that decision.

Consequently, the FDA AHT proposes the following activities as the basis for this CEOS ARD initiative:

1. Identify candidate CARD4L-compliant optical data product providers. Engage agencies to encourage routine optical product generation, and/or sustained provision of toolsets that enable generation of such products. (LSI-VC)
2. Identify candidate CARD4L-compliant SAR data product providers. Write to agencies encouraging routine SAR product generation. (LSI-VC)
3. Make ARD production or production capability available (e.g., on-the-fly and/or pre-processed). Develop a summary of ARD production status and implementation options. (LSI-VC)
4. Utilizing the LSI-VC CARD4L approach, engage ocean and atmosphere VCs to encourage similar CARD4O and CARD4A Definitions and Specifications. (LSI-VC)

To further this CEOS ARD initiative, CEOS agencies are asked to contribute in the following ways:

* Produce ARD data and/or provide ARD production tools
* Advance ARD for the ocean and atmosphere domains
* Participate in standard setting and review process (including user feedback loops for evolution)
* Engage through WGCV to define QA protocols and cross-validation projects across ARD products
* Reach out to commercial providers to obtain their participation in ARD (at their cost)
* Promote and enable discoverability of ARD datasets

# Initiative #2: Interoperable Open Source Tools

Each CEOS agency will continue to develop its data and computational infrastructures consistent with its capacity and user service mandates. CEOS has a role in identifying tools to support complementarity and interoperability across CEOS agencies in support of the FDA strategy objectives of:

* de-risking and simplification of EO data for users;
* allowing users to make use of ALL available and relevant CEOS agency data;
* supporting CEOS ambitions in relation to its chosen grand challenges.

Coordination among CEOS agencies has already led to the adoption of the first of these tools and to its promotion as an open source consortium – the Open Data Cube. The CEOS SEO stewards the CEOS implementation of this, as the CEOS Data Cube (CDC), supported by USGS, CSIRO, GA and others, as a tool for the realization of CEOS ambitions in relation to the grand challenges and GEO flagships. The CDC seeks to simplify the access and application of EO data for users, and to provide a platform whereby all space agencies conforming to the CEOS ARD standards can interoperably contribute their data and application algorithms.

The CDC is an important open source tool for development under this FDA strategy. A considerable suite of EO toolboxes, multi-mission tools for data processing, data product readers and writers as well as visualization and analysis tools have become available under open source conditions. Pilot implementations of the CDC are underway in support of GFOI, GEOGLAM, and other CEOS priorities. A 3-year Work Plan has already been defined and will continually be reviewed and revised to ensure consistency with this proposed CEOS FDA strategy. User feedback activities will be prominent in support of the user metrics activity. Additionally, individual (or groups of) CEOS agencies are anticipated to propose further tools as their respective plans emerge for evolution of their data and computing infrastructure and opportunities for collaboration emerge.

The FDA AHT proposes the following activities as the basis for this Interoperable Open Source Tools initiative:

1. Further develop and communicate CEOS Data Cube governance model. Seek broad awareness and acceptance. (SEO)
2. Develop strategies for local, national, regional, and global scale implementations, ensuring maximum interoperability. (SEO)
3. Prepare a long-term plan of CEOS Data Cube sustainment. (SEO)
4. Develop ‘self-help’ materials on CDC implementation and testing to help establish new Data Cubes that are consistent and interoperable. (SEO)
5. Demonstrate stakeholder promotion and adoption of ODC elements in global initiatives. (SEO)
6. Promote the adoption of open source tools, algorithms, and technologies in support of the goal of interoperability. (SEO)
7. Establish 'FDA Setup Support Scheme' whereby agencies can support projects that wish to get a data cube or some other future data architecture set up, to include some basic data, compute time, and storage. (SEO)

To further this Interoperable Open Source Tools initiative, CEOS agencies are asked to contribute in the following ways:

* Promote the use of interoperable open source tools and application algorithms to enhance the use and impact of CEOS satellite data
* Continue to support the CEOS Data Cube (CDC) initiative and progress the development of core components, supporting documentation, and training materials
* Support the initial deployment of the CDC architecture by providing training and capacity building to interested international users
* Contribute application algorithms to the CDC repository along with documentation and case studies

# Initiative #3: Data, Processing, and Architecture Interface Standards

As CEOS agencies are defining their processing and data dissemination standards, they seek to apply and follow international standards – including those generated by the CEOS Working Group on Information Systems and Services (WGISS). This does not only concern common standards in terms of catalogs, metadata, terminology, and semantics, but it also involves interoperability standards for data discovery and download, as well as common interface standards such as INSPIRE, OGC, W3C, and interoperability with other data access services (e.g., European Data Portal, international, GEOSS). One of the primary FDA concerns is the portability of products and algorithms between the various future data architectures, including the CEOS Data Cube and other agency-specific architectures. This ensures that users are not locked-in to a single platform as new architectures emerge.

The FDA AHT proposes the following activities as the basis for this Data, Processing, and Architecture Interface Standards initiative:

1. Review standards among different FDA approaches to promote interoperability of tools. Approaches might include Data Cube, standard API(s), as well as other FDA approaches. (WGISS)
2. Update CEOS data discovery and download standards moving from discovery of files to discovery and download of pixel level data across multiple sensors as user-defined ARD packages. (WGISS)
3. Promote awareness of new FDA paradigms and systems. (WGISS / WGCapD)

To further this Data, Processing, and Architecture Interface Standards initiative, CEOS agencies are asked to contribute in the following ways:

* Identify key data and metadata standards experts to join WGISS
* Support WGISS in the development of standards that ensure interoperability among one or more FDA platforms
* Support prototype testing of data and application standards to ensure successful implementation
* Identify specialists to join WGISS with expertise in their data holdings to provide guidance on how their holdings are changing to accommodate pixel level access
* Identify key System Engineers, Applications Liaisons, and Communication/Outreach Liaisons to engage with WGCapD in developing strategies to promote FDA paradigms and systems

# Initiative #4: Analytical Processing Capabilities

Future data exploitation environments should provide a variety of functionalities that allow analytical data processing including:

* a full computing environment allowing users to process data and information with high efficiency (i.e., CPU, memory, storage, bandwidth, access to other data)
* easy-to-use programming environments (based on open source software, if possible) with tools to perform generic operations (e.g., Python, R)
* implementation of modular service tools, allowing multiple users to invoke, use, add, and 'chain' the tools and functionalities as per user needs
* algorithm implementation
* time series analysis
* access to tool libraries from third party providers
* data generation service allowing to generate sub-sets of products per user specifications of area of interest and format
* projection tools, allowing users to convert data into different projection systems

The FDA AHT proposes the following activities as the basis for this Analytical Processing Capabilities initiative:

1. Develop and review a best practice white paper (i.e., reference architecture) for EO data analysis pipelines involving CEOS agencies and third-party systems (e.g., cloud, commercial providers) from a user analysis perspective with an emphasis on global agendas. (WGISS)
2. Develop and review a best practice white paper on interoperable EO data analysis APIs (i.e., how can a user-supplied algorithm be made portable between different EO data analysis platforms? How can such analysis packages be made discoverable and shareable?). (WGISS)
3. Develop a best practice single-sign on (SSO) authentication whitepaper to support machine-to-machine authentication for analysis services. (WGISS)

To further this Analytical Processing Capabilities initiative, CEOS agencies are asked to contribute in the following ways:

* Identify Data Analysts, System Engineers, and System Architects to join WGISS
* Provide agency computing resources for prototype testing of application algorithms to take advantage of locally stored data or to utilize web-based protocols (e.g., WCS, APIs) for data interaction
* Engage CEOS agencies who have implemented authentication systems to provide their lessons learned and best practices

# Initiative #5: User Metrics

All CEOS agencies are experiencing a shift in the number and nature of users seeking to benefit from their data and utilizing their information systems to do so. These users are increasingly coming from a diverse range of sectors of society, sometimes non-technical, and with expectations of ease of access, of use, and of integration of space agency data with other information. Each CEOS agency has its own strategy for managing this change in user base, and a number have identified the necessity of accruing and exchanging information among CEOS agencies on the reality of the evolution of the user base and how FDA implementations are impacting them.

The CEOS FDA strategy is being defined as space agency information systems start to respond to the new possibilities being provided by advances in computing, networking, and storage, and we can anticipate significant advances in the coming years. The proposed User Metrics initiative seeks to ensure planning and responsibilities are put in place for CEOS to leverage the experience being gained by individual agencies and to have an ongoing effort to collate available user metrics and to adapt the FDA strategy as these metrics are analyzed beyond what is simply being downloaded.

In terms of user metrics, ambitions should be focused on the utilization of increasingly sophisticated user management functionalities in the data access systems. Methods and tools applied in other data platforms environments (e.g., social platforms) should be evaluated and adapted to better characterize user behavior and means to catalyze EO data utilization.

The FDA AHT proposes the following activities as the basis for this User Metrics initiative:

1. Develop a standard set of user interview and user experience questions. Write to users, requesting feedback. Example users may include CEOS Data Cube, Open Data Cube, and other platforms and applications, though generic inputs should also be accommodated. (WGISS)
2. Develop a database/warehouse for storing user feedback. (WGISS)
3. Develop a capability to analyze and report user feedback and a standard set of statistics that CEOS Agency supported FDA platforms can report against, so that CEOS can provide summary statistics on how EO data is being used and what for. Promote results. (WGISS)
4. Analyze lessons from pilot activities, to scan the horizon, and to submit proposals for new pilots to Principals. (WGISS)

To further this User Metrics initiative, CEOS agencies are asked to contribute in the following ways:

* Engage CEOS representatives familiar with user needs analysis and metric reporting
* Support development of universal metric capturing
* Participate in open venues for discussion of abstract CEOS projects (webinars, conferences, meetings, etc.)