CEOS Interoperability Terminology

CEOS - WGISS
Interoperability and Use Interest Group

Doc. Ref.: CEOS/WGISS/IUIG/CIT
Date: September 2020
Issue: Version 1.0

Contents

Introduction

In the context of Earth remote sensing, the terms *Analysis Ready Data (ARD), interoperability*, and *harmonization* are often used and, to a large extent, used inconsistently. This is particularly problematic in areas where *interoperability* is increasingly important but not well defined like the CEOS Analysis Ready Data and Future Data Access and Analysis fields.

The objective of this document is to define a set of terms to be used in exchanges across CEOS agencies and activities and eventually contribute to broader discussion in Earth Observation (EO) communities including commercial entities and standards bodies. Terms for interoperability for Data Discovery and Access are well defined as this field has matured considerably and industry standards, like those defined by the Open Geospatial Consortium, are in broad use both across CEOS and more broadly.

This document will focus on terminology for two key areas active in the CEOS community and projects where inconsistent use is problematic.

1. Analysis Ready Data (ARD) - concerned with the *content* of EO data and products
2. Analysis and Access of ARD – concerned with the *technology, use and analysis* of EO data and products

It is expected that further revision to terminology will occur as language matures and this document will be revised periodically to reflect the growing community consensus.

Analysis Ready Data

Five terms are defined in this document:

1. Analysis Ready Data (ARD)
2. CEOS ARD for Land (CARD4L) Products
3. Interoperable Products
4. Harmonized Products
5. Fused Products

The terms proposed in this document are based on the idea that *interoperability* refers to a continuum of data product *compatibility* – from completely different datasets on one end, to fully integrated products on the other. In addition, these terms do *not* prescribe storage format or other technology characteristics of ARD.

Analysis Ready Data (ARD)

An *Analysis Ready Data (ARD)* product is generated from raw data and processed so that it can be used without the need for further processing to be applied by users.

This definition is *intentionally* broad to cover a large range of processing levels and possibilities. There is, however, a minimum processing requirement to be an ARD-compliant product: the data must be processed to a geo-referenced projection to enable the position identification within the data product. Beyond this minimum requirement, additional levels of geometric and radiometric processing may be applied to further prepare the data for analysis, reducing the amount of pre-processing for an end user.

CEOS ARD for Land (CARD4L) Products

CEOS Analysis Ready Data for Land (CARD4L) Products are a subset of all ARD products and have been processed to a specified minimum set of requirements and organized into a form that allows immediate analysis with a minimum of additional user effort and interoperability both through time and with other datasets. As a standard, CARD4L represents the initial state from which basic dataset harmonization (see below) can be implemented for increasingly interoperable products.

To be CARD4L-compliant, products must meet either threshold or target levels of requirements based on a series of CARD4L Product Family Specifications (PFSs), which have been developed and can be found at the following website: http://www.ceos.org/ard/#slide3. PFS threshold and target level of requirements are organized into four primary categories:

1. General Metadata
2. Per-pixel Metadata
3. Radiometric and Atmospheric Corrections
4. Geometric Corrections

CARD4L-compliant products define a set of general and per-pixel metadata, radiometric corrections, and geometric corrections, which enables a basic level of common interoperability for the content (e.g. Two optical geophysical measurements are comparable, metadata uses comparable terms like band names or units of measure). This implies that the geographical and sensor characteristics represented in the product's metadata allow products from different sensors to be superimposed, compared, and generally worked with in a common environment. While necessary, CARD4L-compliance is not sufficient for products to be completely interoperable. For instance, CARD4L-compliance does not require products to use the same geodetic reference or have the same geodetic accuracy – only that it is reported in the metadata – whereas fully harmonized products will use the same geodetic reference to enable direct comparison without geospatial transformation.

The CARD4L PFSs do not define specific processing approaches or reference datasets to be used in processing; however, they can be considered a starting point for interoperable products. The data layers have undergone a sufficient level of processing to normalize the radiometry to a geophysical parameter (e.g., surface reflectance, surface temperature, etc.) and to a recognized geometric projection. Additionally, metadata shall be included, which fully describes the radiometric and geometric processing applied, including characterized accuracies.

The CARD4L framework enables data providers to conduct self-assessments on their compliance to the CARD4L PFS threshold and target levels. Data provider self-assessments are subsequently independently validated by the CEOS Working Group on Calibration and Validation (WGCV) prior to being officially recognized as CARD4L-compliant on the CEOS website referenced above.

Interoperable Products

Interoperable Products refers to a set of two or more ARD products which are sufficiently documented to enable processing across a continuum of geometric and/or radiometric standards to permit direct quantitative comparison.

This definition is intentionally broad and covers a large range of processing levels and possibilities but is constrained at one extreme by a minimum processing requirement for an ARD product: that the data is geometrically processed to a geo-referenced projection to enable the identification of the position of the acquired data.

Example 1: Imagery from two different satellite missions have been processed to two different, but documented, geographic projections with sufficient information to describe their cartographic and accuracy attributes. These two products are geometrically *interoperable* as they have sufficient information to be referenced to a common geometric model, allowing direct comparison.

Example 2: The same two images from Example 1 also have sufficient radiometric information (potentially to include ancillary data) to normalize their pixel values to a common and comparable radiometric correction (i.e., surface reflectance, surface reflectance with BRDF correction, spectral response, etc.), thus having the *potential* to geometrically reference/align pixels that are radiometrically normalized.

Harmonized Products

Harmonized Products refers to a set of two or more interoperable ARD products, which have been processed to common geometric and/or radiometric levels to enable direct comparison between the products.

Example 1: Two products, which are *interoperable* due to their known – but different – geographic projections and GSDs are re-projected to a common projection and GSD, resulting in a geographically *harmonized* product. If these same products support a sufficient level of radiometric interoperability, they might be further processed to a surface reflectance value (i.e., a geophysical parameter) thus improving the harmonization to include a compatible radiometric standard as well.

Example 2: Given that the two products in Example 1 also include sufficient spectral interoperability and documentation, they could be processed to be functionally and statistically common products, e.g. the NASA Harmonized Landsat and Sentinel-2 (HLS).

Fused Product

A *Fused Product* is a derived data product produced by merging two or more fully interoperable products. The derived data product contains values created from the merged data into a new single data product. While the input data may be provided with the fused product for reference, the input data are no longer independent data products in the new fused product. Therefore, it is not possible to go "backward" to recover the initial data products using the fused product.

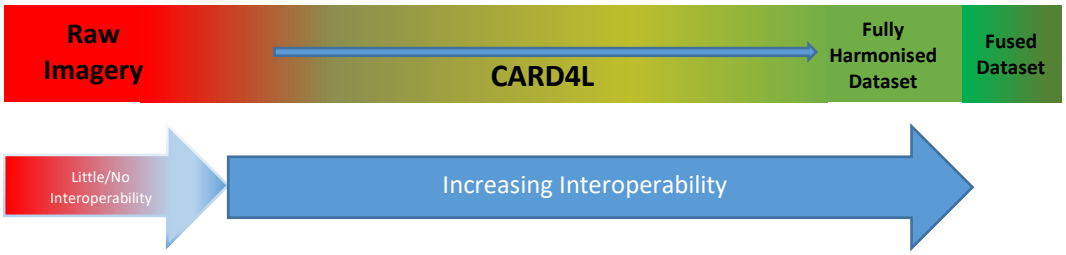
Example 1: Two separate streams of products, one at higher and the other at lower spatial resolution, are combined/fused to generate in output a single stream at the higher spatial resolution, for the common spectral bands and for the dates of the input products. ESA sen2like processor implements this processing for Landsat-8 and Sentinel-2, generating an output a stream of Sentinel-2 like products for all the dates of the input products.

Example 2: Two HLS products over a common geographic area (e.g., a Sentinel-2 tile) and different dates may be *fused* to create a synthetic interpolated image representation of a date between the two HLS images. Using numerous HLS images over an extended time-period, a time-series is produced on a daily interval basis from interpolated observations.

Interoperability Continuum

Interoperability represents a continuum of potential compatibility for products to work in numerous information technology systems and with other like-prepared data products. The ultimate product

interoperability (i.e., a harmonized or a fused dataset) is achieved when products have a fully consistent spectral, radiometric, geometric, metadata, and file format implementation where applications can interact fully with the data interchangeably without modification.



Analysis, Access, and Analysis Ready Data

The previous section of this document focused on clarifying CEOS interoperability terminology with reference to EO data *products* – the focus was on the *content* and its compatibility. It is also necessary to clarify the meaning of *Analysis Ready Data* in the context of *use*. This has been an area of confusion since saying data is *analysis ready* implies it is ready for use, yet there is considerable variation in use and what it means to be useable. As CEOS agencies and other organizations make increasingly large amounts of ARD content available in the Cloud where users can analyze it in-situ there are additional considerations for interoperability.

The objective of this section of the document is to define a set of terms to be used in the exchanges between CEOS groups and other organizations when discussing interoperability in the context of the Analysis and Access (use) of ARD. The terms are not in one sense new since Analysis and Access interoperability has a significant body of work behind it and is quite apparent in CEOS (e.g. CEOS International Directory (IDN)). The focus here is on those areas where the disruption of ARD, particularly ARD in the Cloud is having impact as identified in the CEOS Future Data Access and Analysis Architectures initiative (CEOS FDA).

Cloud Data Formats and ARD in the Cloud

Traditionally, satellite remote sensing data have been stored in self-describing formats, determined by the data custodian such as GeoTIFF, HDF and netCDF. Such data were then downloaded by users who would convert it into their required format for ongoing analysis. While these formats excel at encoding complex data structures and rich metadata and thus support custodial requirements, support by tools was uneven at best and reading and converting the data imposes a steep learning curve on end users. The emergence of remote data access protocols such as OPeNDAP and Web Coverage Service introduced the ability to encode the data for transmission as well as storage, thus uncoupling, to a degree, the storage format used by the data stewards, and the format received and used by the end user. In connections between tools and servers using these remote protocols, the user may neither know nor care about either the storage or the transfer encodings; the underlying software transparently moves the required data from data server to tool memory.

Cloud computing, or more specifically the publication of EO data into the Cloud has disrupted this approach. Users can now have direct access to the stored data and can access it in place for analysis. The previous separation of “custodial format” and “user analysis format” is no longer in place. To be Analysis Ready in the Cloud thus has the connotation of ARD content being provided in a format suitable for direct user analysis.

In addition, Cloud Object storage systems differ significantly from the usual File systems in terms of cost, scalability, and access protocol. For performance and cost reasons it is useful to be able to find and request small chunks of data, rather than an entire data file. A detailed discussion on this is beyond the scope of the document, but there are three separate approaches defined here that are relevant to ARD in the Cloud:

1) **Cloud-friendly formats** – These are a more traditional file format (e.g. GeoTIFF) which are internally re-organized and stored to a specific configuration that is more efficient to use in the Cloud. For example, Cloud-Optimized GeoTIFF, store each file as an object. The object contains multiple chunks of data, say a small spatial region from the entire image. Each object stores information about the internal file structure in a header of predetermined size, allowing readers to read the header first, then use that information to request segments of the data file using the

"range-get" feature of HTTP. They may also store lower-resolution versions of the (usually) image for fast retrieval. The objective of these formats is to retain a high degree of backwards compatibility with existing software tools whilst supporting reasonable performance when stored on Cloud Object storage.

2) **Cloud-native formats** – These are entirely new storage formats specifically designed for Object storage and exploiting the different access and performance available (e.g. Zarr). Cloud-native formats store data chunks as individual objects in the Object Storage. This eliminates the requirement to access first to get the header and then again to get suitable chunks. Chunks can also be distributed across the Cloud servers providing multiple network paths and thus greater bandwidth when accessing data. Use of Cloud-native formats does require support in the software tools which is occurring but is very new.

3) **Cloud Data Access API** – This approach leverages Cloud computing scalability and Compute+Object Storage connectivity to effectively hide the underlying storage structure. It is very much like the remote data access protocols discussed in the introduction to this section that are normally used for EO data download. Data servers, such as OPeNDAP and the HDF Highly Scalable Data Server, can extend their abstraction to hide the Object Storage data layout complexity, while still providing performance on par with data on random-access spinning disk by utilizing either cloud native or cloud friendly technologies above. Since data is being manipulated during access it can be returned in any format at the cost of compute time.

Types of Analysis Interoperability

Analysis interoperability can be defined simply as the same result is obtained for a given set of inputs when performing the same analysis using different tools. For example, in the context of the ARD terminology we can say two different water quality algorithms intended to produce the same metric when fed the same CARD4L data should produce the same Interoperable Product (at least within expected tolerances).

In practice analysis interoperability is more complex because it also relates to portability of the tool being used or different algorithms used for the same analysis purpose (e.g. cloud detection and removal as a purpose has several implementation options). Clarification is required in terminology since ARD in the Cloud carries with it expectations of "Analysis moving to the Data".

Analysis interoperability can be thought of at three different levels of abstraction, the first two of which are tied to code portability.

1. **Executable code**: an opaque ("black box") code package or file can be executed on more than one platform. This has become more achievable in recent years with the rise of containerization. This form of interoperability underlies the Open Geospatial Consortium's Web Processing Service.

2. **Source code**: a source code package can be executed on more than one type of platform. The high source code interoperability of Python and R science packages across many platforms have been instrumental in growing those two respective ecosystems.

3. **Algorithm:** multiple code interpretations of the algorithm can be executed. This type of interoperability is typically exposed via an Application Programming Interface, such as the standard

set of UN SDG Water Quality calculation end points exposed by multiple Data Cube implementations.