OGC Testbed-16

Analysis Ready Data Engineering Report

Publication Date: 2021-01-13

Approval Date: 2020-12-14

Submission Date: 2020-11-19

Reference number of this document: OGC 20-041

Reference URL for this document: http://www.opengis.net/doc/PER/t16-D018

Category: OGC Public Engineering Report

Editor: Joan Maso

Title: OGC Testbed-16: Analysis Ready Data Engineering Report

OGC Public Engineering Report

COPYRIGHT

WARNING

LICENSE AGREEMENT

Permission is hereby granted by the Open Geospatial Consortium, ("Licensor"), free of charge and subject to the terms set forth below, to any person obtaining a copy of this Intellectual Property and any associated documentation, to deal in the Intellectual Property without restriction (except as set forth below), including without limitation the rights to implement, use, copy, modify, merge, publish, distribute, and/or sublicense copies of the Intellectual Property, and to permit persons to whom the Intellectual Property is furnished to do so, provided that all copyright notices on the intellectual property are retained intact and that each person to whom the Intellectual Property is furnished agrees to the terms of this Agreement.

If you modify the Intellectual Property, all copies of the modified Intellectual Property must include, in addition to the above copyright notice, a notice that the Intellectual Property includes modifications that have not been approved or adopted by LICENSOR.

THIS LICENSE IS A COPYRIGHT LICENSE ONLY, AND DOES NOT CONVEY ANY RIGHTS UNDER ANY PATENTS THAT MAY BE IN FORCE ANYWHERE IN THE WORLD. THE INTELLECTUAL PROPERTY IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NONINFRINGEMENT OF THIRD PARTY RIGHTS. THE COPYRIGHT HOLDER OR HOLDERS INCLUDED IN THIS NOTICE DO NOT WARRANT THAT THE FUNCTIONS CONTAINED IN THE INTELLECTUAL PROPERTY WILL MEET YOUR REQUIREMENTS OR THAT THE OPERATION OF THE INTELLECTUAL PROPERTY WILL BE UNINTERRUPTED OR ERROR FREE. ANY USE OF THE INTELLECTUAL PROPERTY SHALL BE MADE ENTIRELY AT THE USER'S OWN RISK. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR ANY CONTRIBUTOR OF INTELLECTUAL PROPERTY RIGHTS TO THE INTELLECTUAL PROPERTY BE LIABLE FOR ANY CLAIM, OR ANY DIRECT, SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES, OR ANY DAMAGES WHATSOEVER RESULTING FROM ANY ALLEGED INFRINGEMENT OR ANY LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR UNDER ANY OTHER LEGAL THEORY, ARISING OUT OF OR IN CONNECTION WITH THE IMPLEMENTATION, USE, COMMERCIALIZATION OR PERFORMANCE OF THIS INTELLECTUAL PROPERTY.

This license is effective until terminated. You may terminate it at any time by destroying the Intellectual Property together with all copies in any form. The license will also terminate if you fail to comply with any term or condition of this Agreement. Except as provided in the following sentence, no such termination of this license shall require the termination of any third party end-user sublicense to the Intellectual Property which is in force as of the date of notice of such termination. In addition, should the Intellectual Property, or the operation of the Intellectual Property, infringe, or in LICENSOR's sole opinion be likely to infringe, any patent, copyright, trademark or other right of a third party, you agree that LICENSOR, in its sole discretion, may terminate this license without any compensation or liability to you, your licensees or any other party. You agree upon termination of any kind to destroy or cause to be destroyed the Intellectual Property together with all copies in any form, whether held by you or by any third party.

Except as contained in this notice, the name of LICENSOR or of any other holder of a copyright in all or part of the Intellectual Property shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Intellectual Property without prior written authorization of LICENSOR or such copyright holder. LICENSOR is and shall at all times be the sole entity that may authorize you or any third party to use certification marks, trademarks or other special designations to indicate compliance with any LICENSOR standards or specifications.

This Agreement is governed by the laws of the Commonwealth of Massachusetts. The application to this Agreement of the United Nations Convention on Contracts for the International Sale of Goods is hereby expressly excluded. In the event any provision of this Agreement shall be deemed unenforceable, void or invalid, such provision shall be modified so as to make it valid and enforceable, and as so modified the entire Agreement shall remain in full force and effect. No decision, action or inaction by LICENSOR shall be construed to be a waiver of any rights or remedies available to it.

None of the Intellectual Property or underlying information or technology may be downloaded or otherwise exported or reexported in violation of U.S. export laws and regulations. In addition, you are responsible for complying with any local laws in your jurisdiction which may impact your right to import, export or use the

Intellectual Property, and you represent that you have complied with any regulations or registration procedures required by applicable law to make this license enforceable.

Table of Contents

Chapter 1. Subject

The Committee on Earth Observation Satellites (CEOS) defines Analysis Ready Data (ARD) for Land (CARD4L) as "satellite data that have been processed to a 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".

This OGC Testbed 16 Engineering Report (ER) generalizes the ARD concept and studies its implications for the OGC Standards baseline. In particular, the ER analyses how modern federated data processing architectures applying data cubes and Docker packages can take advantage of the existence of ARD. Architectures for ARD should minimize *data transmission* and allow and favor *code transmission* and remote execution. This ER also considers a workflow in which new processes are triggered as soon as new data becomes available. This is part of the event driven discussion.

Chapter 2. Executive Summary

Users spend a majority of their time preparing data relative to doing data analysis. Being able to get data that is ready for analysis saves a lot of time and effort and permits fast results and interpretation. This makes the concept of Analysis Ready Data very attractive for the decision maker. This Engineering Report has conducted a survey of the different interpretations of the ARD concept that can be classified into *content readiness* and *technical readiness*, being both necessary to enable fast analysis.

CEOS is conducting the most rigorous approach to *content readiness* by focusing on stablishing the requirements for making some selected satellite products ready for analysis and formalizing *Product Family Specifications* (PFS). The main characteristics of the CEOS ARD for Land (CARD4L) can be easily extended other dataset sources including in-situ measurements. Distilling from the CARD4L concept, a dataset is ready for analysis when it represents one or more physical variables, it is georeferenced in a common CRS, it is homogeneous and comparable in time, it is flagged with quality, wrong or missing values tags, and the process of creation is fully documented.

This Engineering Report also contemplates *technical readiness* aspects, such as the availability of a cloud free, continuous data in space and evenly distributed in time, as well as infrastructures to provide on-demand products or to process ARD on the cloud. These technical readiness aspects are referred to also as ARD by other communities but in this document they are referred to as technologies to improve the usability of ARD.

This Engineering Report puts ARD in the context of OGC standard services and provides several examples of use cases where usability and technical readiness of ARD is enabled by adding OGC web services. Use cases presented are: detection of significant events from several sources; integration of diverse sources to enrich a time series for remote sensing phenology variable extraction; protection of Very High Resolution (VHR) data access while allowing for data processing, forest fire detection and monitoring in remote areas; and machine learning training with ARD and ARD trained model discovery.

Some previous initiatives to provide exploitation platforms for ARD rely on OGC services and transversal technologies. Despite the use of common technologies, most of them work in isolation. This Engineering Report explores how these and other transversal technologies can be used to define a federation of exploitation platforms that integrates several sources of ARD in a distributed computing environment. The federation uses the data cube metaphor (that can be described with the Coverage Implementation Schema) to deal with the heterogeneity of the data. The federation should consider ways to parallelize and distribute processing among different services minimizing the amount of data that is transmitted. A future interoperability experiment could go deeper into testing the approach by implementing some of the proposed use cases.

A chapter in this document explores how to use ARD in the context of Machine Learning. The concept of training ready data is introduced as training sets defined as annotation on top of ARD following a particular PFS. A catalogue of models trained to performing a particular task on a ARD that follows a particular PFS is also introduced.

Finally, a future collaboration between CEOS and OGC is proposed where the OGC can contribute to increase the usability of ARD by considering ARD in data discovery, access and processing web

services and to broaden the concept to other types of data. The collaboration with the OGC can also be beneficial in disseminating the ARD concept and benefits to its membership.

The Engineering Report identifies a concrete need for CEOS to include a formal indication of the physical variables that current and future CEOS PFS represent, preferably as a permanent URI in a definition service (e.g. the OGC Definitions Server). The Engineering Report also identifies the need for OGC services to support URIs to characterize the physical variables that data represents in data access services, as well as in data processing services inputs and outputs; in the same way that the SensorThings API is already doing with the ObservationProperty definition URI. We anticipate that the use of URI for physical variable will contribute to an automatic matching between data and processes that will, in turn increase the availability of derived ARD products.

2.1. Document contributor contact points

All questions regarding this document should be directed to the editor or the contributors:

Contacts

Name	Organization	Role
Joan Maso	UAB-CREAF	Editor
Alaitz Zabala	UAB-CREAF	Contributor
Alba Brobia	UAB-CREAF	Contributor

2.2. Foreword

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. The Open Geospatial Consortium shall not be held responsible for identifying any or all such patent rights.

Recipients of this document are requested to submit, with their comments, notification of any relevant patent claims or other intellectual property rights of which they may be aware that might be infringed by any implementation of the standard set forth in this document, and to provide supporting documentation.

Chapter 3. References

No normative references are required in this document. Some fundamental references are:

- CEOS Analysis Ready Data Strategy version 1, October 2019. http://ceos.org/ard/files/CEOS_ARD_Strategy_v1.0_1-Oct-2019.pdf
- Dwyer, J. L., Roy, D. P., Sauer, B., Jenkerson, C. B., Zhang, H. K., & Lymburner, L. (2018). Analysis ready data: enabling analysis of the Landsat archive. Remote Sensing, 10(9), 1363. https://www.mdpi.com/2072-4292/10/9/1363
- Gonçalves P. (2019) OGC Testbed-15: Federated Clouds Analytics Engineering Report. http://docs.opengeospatial.org/per/19-026.html
- Percivall G. (2020) Geospatial Coverages Data Cube Community Practice. https://portal.ogc.org/files/18-095r7

More non normative references can be found in the Bibliography at the end of this document.

Chapter 4. Terms and definitions

The following terms and definitions apply.

- **Analysis Ready Data**

  sensed data that have been processed to a minimum set of requirements and organized into a form that allows immediate analysis with a minimum of additional user effort for further interoperability both through time and with other datasets. (source: CEOS http://ceos.org/document_management/Meetings/Plenary/30/Documents/5.5_CEOS-CARD4L-Description_v.22.docx)

  NOTE The CEOS original definition uses the work "satellite" instead of "sensed".

- **CEOS Analysis Ready Data for Land**

  products processed to a minimum set of requirements and organized into a form that allows immediate analysis with a minimum of additional user effort. These products would be resampled onto a common geometric grid (for a given product) and would provide baseline data for further interoperability both through time and with other datasets. The CARD4L products are intended to be flexible and accessible products suitable for a wide range of users for a wide variety of applications, including particularly time series analysis and multi-sensor application development. They are also intended to support rapid ingestion and exploitation via high-performance computing, cloud computing and other future data architectures. They may not be suitable for all purposes and are not intended as a "replacement" for other types of satellite products (source: CEOS PFS template. Example: http://ceos.org/ard/files/PFS/NRB/v5.0/CARD4L-PFS_Normalised_Radar_Backscatter-v5.0.pdf)

- **Interpretation Ready Data**

  geospatial data that has been submitted to some well documented common process to make it ready for direct human interpretation (probably with the help of some visualization tool) and eventual decision making. There is an expectation that result of an appropriate analysis that uses Analysis Ready Data as input will results on Interpretation Ready Data (definition by the authors based on several sources including https://www.geoaquawatch.org/wp-content/uploads/2020/05/ARD-GEO-AquaWatch-Discussion-paper.pdf)

- **Processing levels**

  a hierarchical list of numerical levels from 0 to 4 (sometimes with a letter A, B... after the number) that indicates the processing done to remote sensing satellite images before making them accessible. Level 0 refers to products are raw data at full instrument resolution (rarely made available), Level 1 refers to reconstructed at full resolution, time-referenced, and annotated with ancillary information, including some sort of radiometric and geometric calibration coefficients and georeferencing parameters that can be applied or not to the data itself. Level 4 are model outputs or results from analyses (that can be assimilated to physical measurements on the ground, such as temperature etc.) that uses satellite data as inputs. (definition by the authors based on several sources including https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy/data-levels)

NOTE	There has been historically some confusion on processing levels because they are not used consistently among agencies or even among different products in the same agency.

- **Product Family Specification**

  a specification of primary geophysical measurements that can be derived from CEOS satellite instruments as ARD. A PFS provides a list of requirements for a type of product (e.g. Surface Reflectance, Surface Temperature, Normalized Radar Backscatter, Polarimetric Radar, etc.) to be considered ARD. These lists of requirements are applicable to general metadata, per pixel metadata, geometric and radiometric corrections. For each requirement two levels of verification are provided: threshold and target (definition by the authors based on several sources including http://ceos.org/ard/files/PFS/SR/v5.0/CARD4L_Product_Family_Specification_Surface_Reflectance-v5.0.pdf)

NOTE	There is an assumption of a consensus process among agencies to define these requirements that should avoid the confusion created by the processing levels concept.

NOTE	CEOS developed the concept of "Product Families" as the second element of the CARD framework. CARD PFS is not prescriptive with regard to which data processing approach should be used. This recognizes that there are multiple processing approaches for producing ARD for a particular *Product Family* and that these will evolve through time. However, the data provider must document and disclose their methods as required by the PFS.

4.1. Abbreviated terms

AI	Artificial Intelligence
ARD	Analysis Ready Data
CARD	CEOS Analysis Ready Data
CARD4L	CEOS Analysis Ready Data for Land
CEOS	Committee on Earth Observation Satellites
IRD	Interpretation Ready Data
ML	Machine Learning
PFS	Product Family Specification

Chapter 5. Overview

Quite often - especially in data intense analysis - data preparation takes more time than the analysis itself. Data preparation involves a set of procedures that: 1.) Clean the data to remove artifacts or repetition, and 2.) Adapt the data to the requirements of the analytical tools. This preparation is a tedious and costly process that has to be completed by anyone that wants to use the data. This is particularly true for remote sensing data. This is due to the fact that satellite raw data needs to be corrected in several ways to get a product that is suitable to be used and combined with in-situ observations. The Committee on Earth Observation Satellites (CEOS) has promoted the concept of Analysis Ready Data (ARD) to simplify the use of these data. In ARD, the producer performs many of the common data preparation (pre-processing) steps, and carefully documents the provenance in the metadata. The prepared product is distributed with other offerings by several agencies thus allowing for an easier merge of products that share the same Product Family Specification (PFS). In CEOS, the PFS detail specific 'Threshold' and 'Target' requirements for the processed content.

This Testbed 16 Engineering Report (ER) is an attempt to consider ARD in relation to the current and emerging OGC Standards baseline.

Section ARD definition introduces the concept of Analysis Ready Data, provides definitions, and the possibility of extending the concept beyond remote sensing data.

Section Where to find ARD lists some current initiatives offering ARD products or ARD based services.

Section Architectures to provide ARD discusses the high-level architectures that can be used by the producers for creating and providing ARD.

Section Tools for using ARD discusses tools that help consumers to take advantage of ARD.

Section Federated architecture for ARD provides the description of a cloud based federated architecture to work with multiple sources of ARD and describes how some common use cases could perform in such architecture.

Section Applying ARD to Machine Learning discusses how ARD can be used in machine learning algorithms.

Section Recommendations includes some recommendations for future work, as well as recommendations for continuing the collaboration between CEOS and OGC are detailed.

Chapter 6. ARD definition

Raw satellite instrument data is the imagery and metadata as collected by the sensor and prior to any processing. Since there are some fundamental corrections that should be applied to the imagery before it is usable, most agencies will not distribute raw imagery. The raw satellite data are simply not ready to use.

There are geometric and radiometric corrections that can be applied in sequence to raw satellite data to make it more useful. There are many good tutorials explaining why these corrections are needed and how they can be applied. One example is the NRCan Geometric Distortion in Imagery [https://www.nrcan.gc.ca/maps-tools-publications/satellite-imagery-air-photos/remote-sensing-tutorials/satellites-sensors/geometric-distortion-imagery/9401)] web page. These corrections are required due to errors resulting from a variety of factors. The primary geometric corrections are required due to at least four factors:

- The orbit of the satellite platform is not parallel to a meridian, resulting in an image that is rotated in relation to the North-South axis.
- The surface of the Earth is not flat and there are changes in the "elevation" over sea level. Most of the time the sensor has a non-vertical perspective (the position of the Earth where the sensor is observing the Earth vertically is commonly called nadir). In these situations, the perspective combined with elevation changes distort the image.
- The Earth is not flat.
- The optical nature of the sensor may introduce additional distortions. Fortunately, most of these distortions obey geometric laws and the effects can be reverted by using the appropriate geometric correction algorithm(s).

Figure 1. Image over an area in Olot (North-East of Catalonia) with surface mountains due to an ancient volcanic activity.

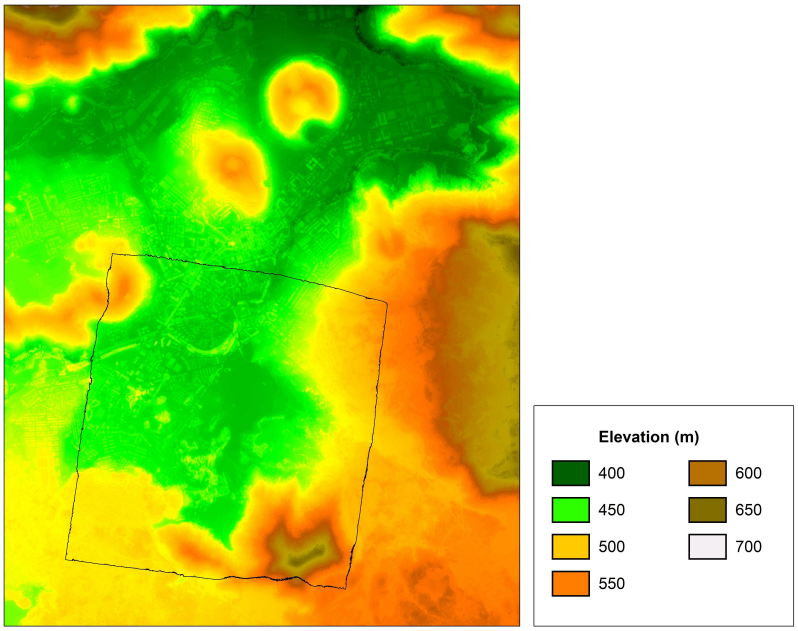

Figure 2. Coverage with elevation values including the same area in North-East of Catalonia.

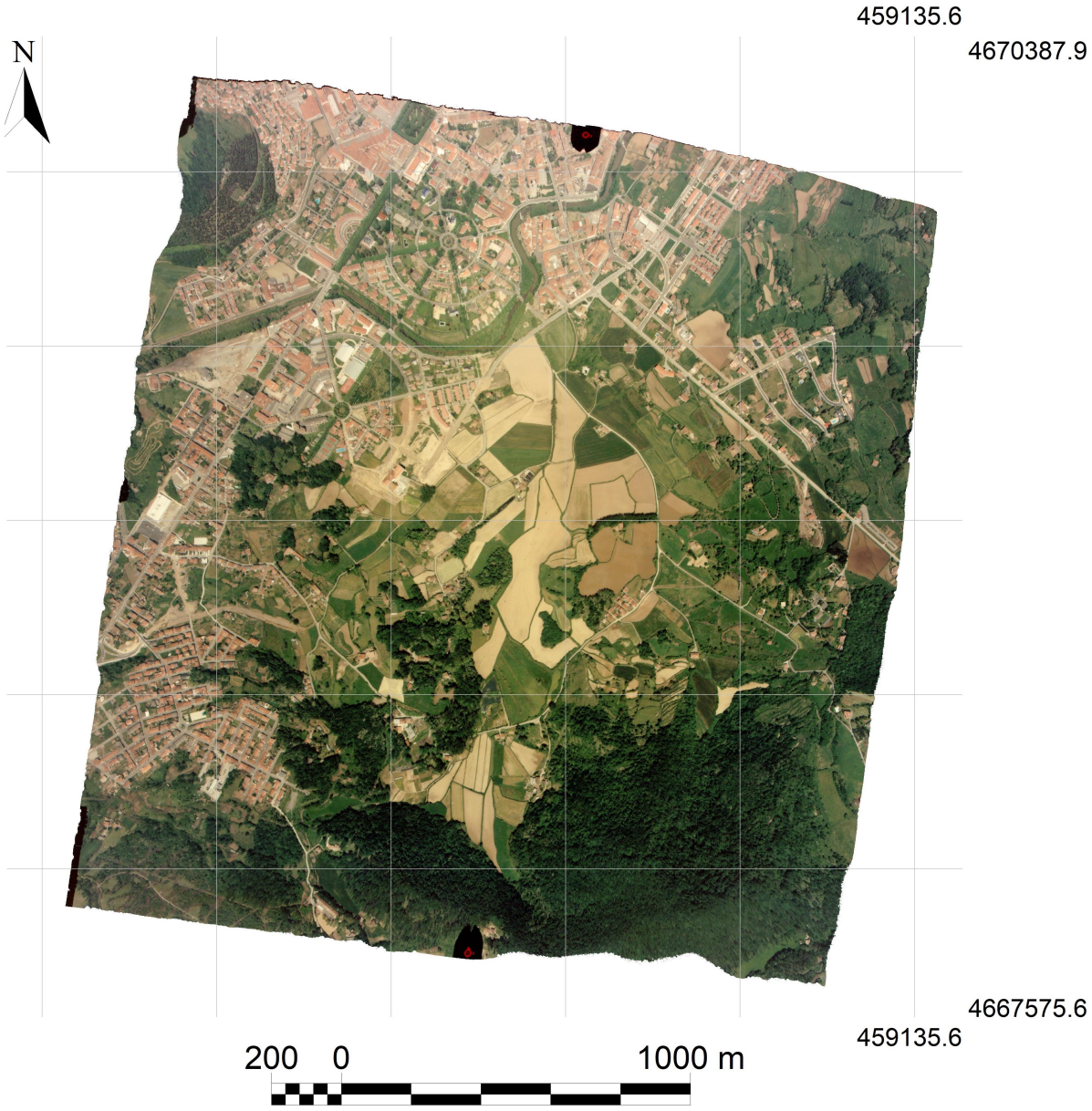

Figure 3. Result of applying a geometric correction to the image and at the same time transform it in a map over the UTM31N projection.

Radiometric corrections try to correct three aspects:

- The intensity captured by the sensor needs to be converted into the amount of solar radiation received by the sensor. This is done by calibrating the sensor with a source of known intensity.
- There are effects produced by the atmosphere between the sensor and the Earth's surface that absorbs parts of the solar radiation in different ways depending on the frequency and on the atmospheric conditions of that day. Being able to remove these effects results in an image with an intensity of colors equivalent to a picture taken at low altitude by a drone or a camera on a high pole/platform. Once the correction is applied, the image is called Bottom Of the Atmosphere (BOA). This correction requires knowledge of atmospheric conditions present during the image acquisition time frame.
- The slope and orientation of the terrain changes the incident light angle and changes the intensity received by the sensor. In extreme cases there are shadows that return only a small fraction of light. A common solution for the latter case is to flag the shadows as invalid data, but the result is not acceptable for mountain areas where shadows are abundant.

Figure 4. Original image with the radiometry as received by the sensor (Top of the Atmosphere; ToA). Sentinel-2A color natural. R137 granule SPB (Huelva)

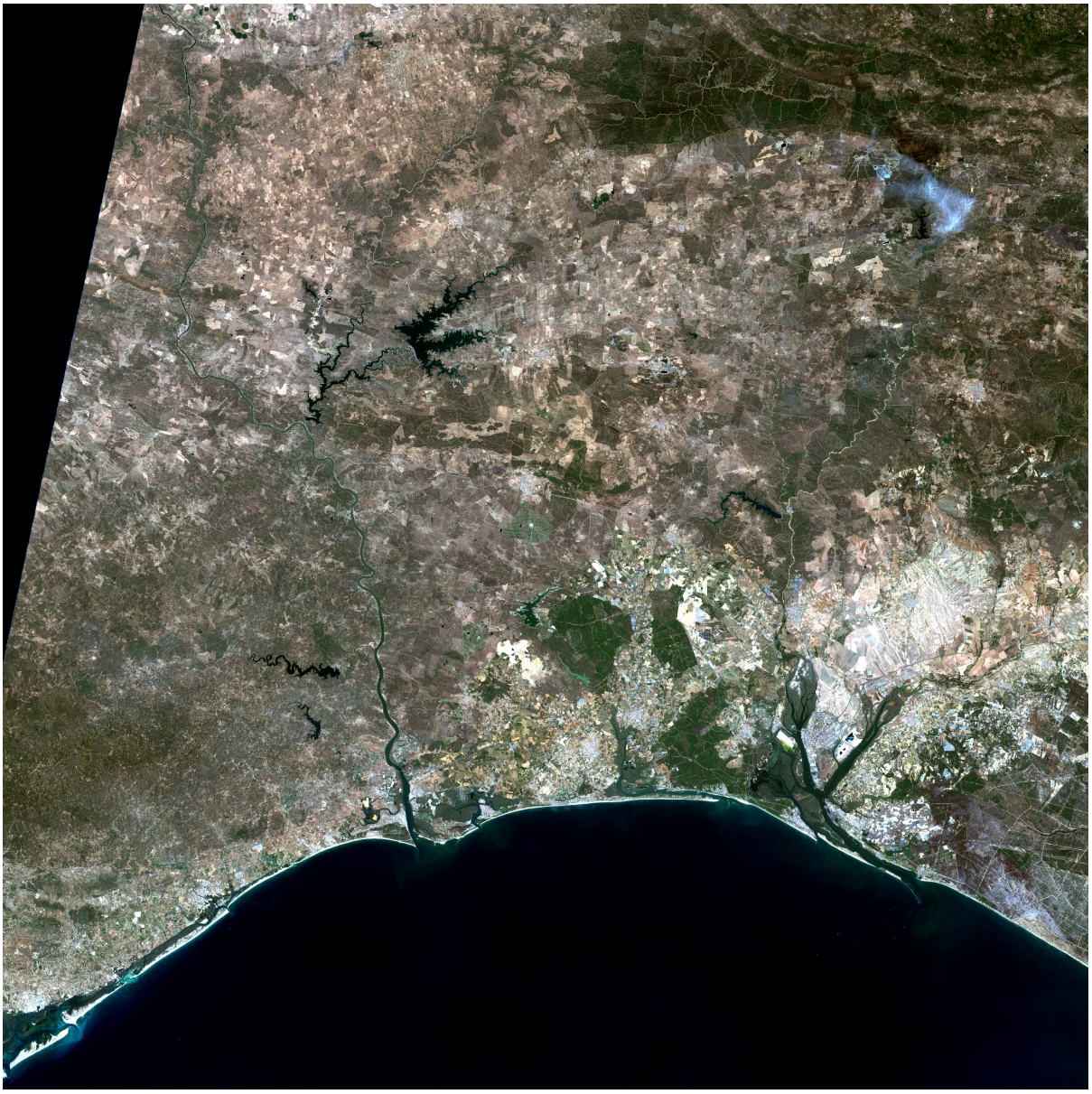

Figure 5. Result of applying radiometric corrections to the image and get radiation as if there was no Earth atmosphere (Bottom of the Atmosphere; BoA).

After many years, remote sensing has reached a level of maturity and the algorithms needed to apply to perform the corrections are well known. Each correction applied generates a new dataset that has an increased "Level" number. In general, Level 1 corrects detector variations within the sensor as well as geometric distortions due to the curved geometry of the Earth surface and the optical nature of the sensor. In Level 2, data is mapped into a cartographic projection and atmospheric effects are compensated for.

The abundance of similar sensor platforms should make possible the combination of compatible products together to increase the global temporal resolution. Combining similar products in common series is sometimes referred to as a Virtual Constellation. Unfortunately, the peculiarities of the different missions resulted in similar but different processing chains creating consistency problems in the exact definition of the processing "Levels". Mixing products from different missions still requires specialized intercalibration processes.

A new approach has been proposed by CEOS that is based on the level of readiness the data have for a particular domain. Instead of defining levels based on corrections, the readiness level tries to provide data ready for a particular analytical domain of application. The data will then be analyzed

and later interpreted based on two concepts: Analysis Ready Data (ARD) and Interpretation Ready Data (IRD).

Currently CEOS has defined ARD for land (CARD4L) and is working on ARD for aquatic media and is considering ARD for atmospheric products. For each of these domains, products are specified depending on the nature of the sensor used.

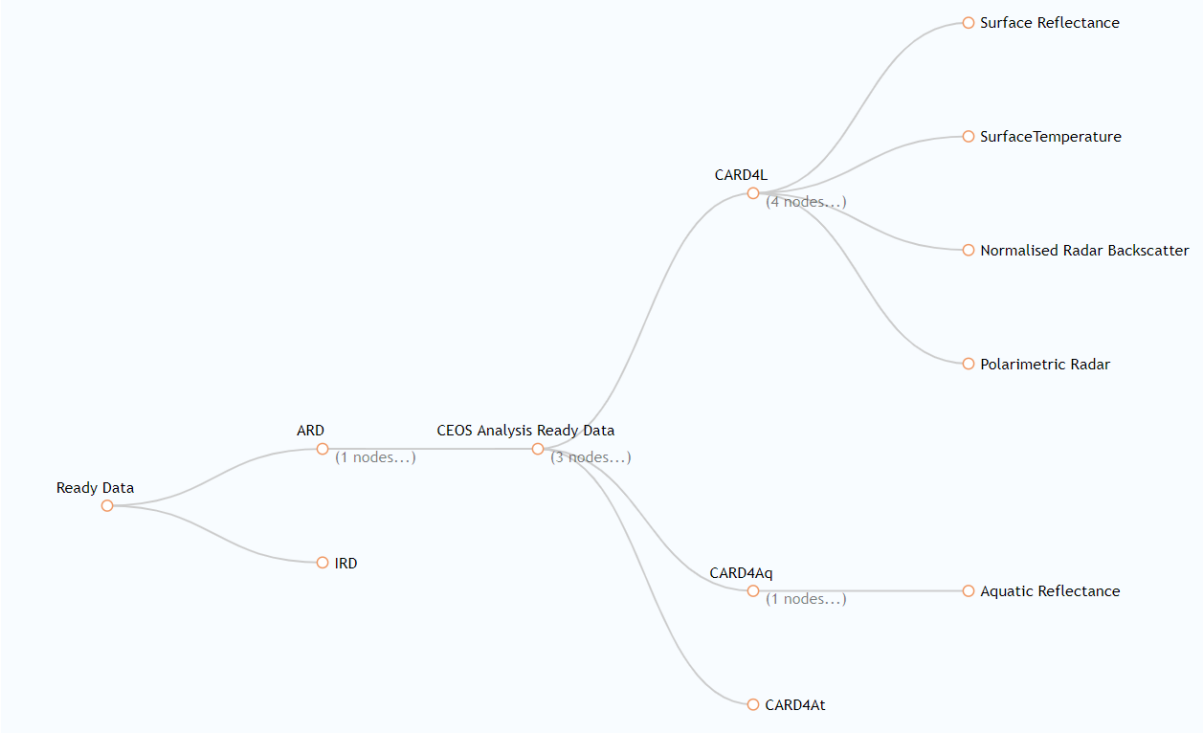

Figure 6. Analysis Ready Data definitions and their characteristics.

In practice the concept can be generalized to any kind of data. There is always a process of data organization and documentation done prior analysis. For example, drone imagery has similar problems to satellite data and requires geometrical corrections and mosaicking before it can be consumed.

6.1. Analysis Ready data and Interpretation Ready Data

Decision makers want to have data that help them to make informed decisions. Decision makers require data ready for direct interpretation (Interpretation Ready Data: IRD). This requirement dictates that somebody transforms the raw data into IRD before being provided to the decision makers. Analysis Ready Data is an effort to define a better way to divide the processing and transformation efforts between the producer and the user based on a common agreement on a solution independent form the domain of application. Producers will make available ARD that will be organized by the user (or by common processing facilities such as the ESA DIAS) into data cubes, regular time intervals data cubes, or another optimal data organization before executing a specific analysis such as a Land cover classification or a phenological study. These data will be analyzed by the customer organization and the results will be presented as IRD to the decision maker.

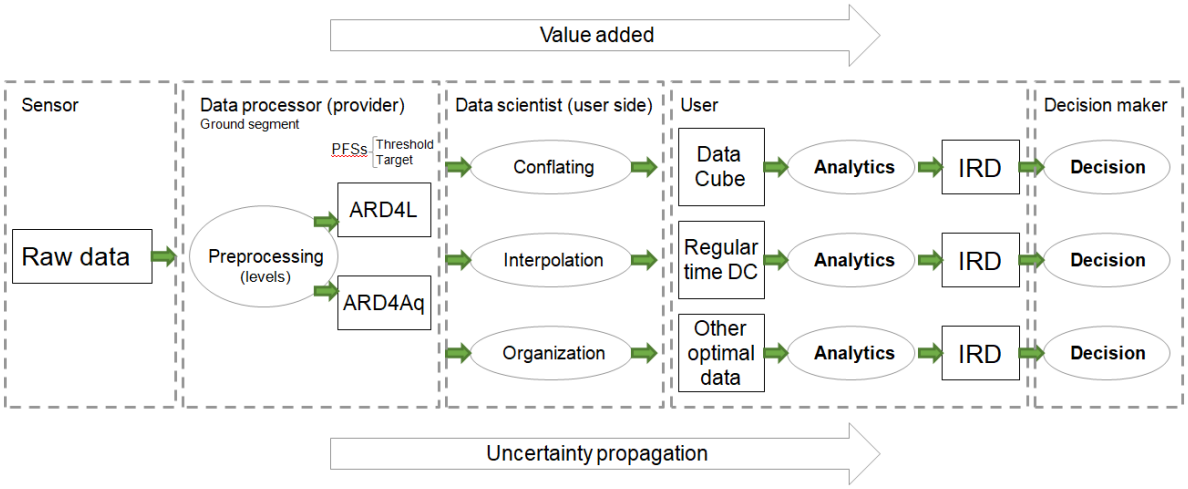

Figure 7. Distribution of processing between produced and consumer in ARD.

ARD is an abstract concept that can only be implemented in the context of a common domain of application. However, several definitions have been proposed and some of them are collected here:

- Time-series stacks of overhead imagery that are prepared for a user to analyze without having to pre-process the images (C. Holmes, [1]).
- The product of some processing that qualifies that product for direct knowledge generation and fact display (I. Simonis, [2]).
- EO data offered in a way better suitable for consumption especially by non-programmers and non-EO experts (P. Baumann, [3]).
- Automated identification and pre-processing of all data needed to execute an analytic task regardless or source, format, or location (Testbed 16 CFP, [4]).

From these definitions and the observation of the current ARD products, the Testbed participants extracted some of the common characteristics that data should have to be considered ARD:

- Data is abstracted from the limitations of the platform that originated it:
  - The product is created in a way that knowing some details of the mission and the sensor is no longer necessary.
  - Data is made available in a coherent, cross-calibrated time stack of data.
  - Data is georeferenced in a well-known Coordinate Reference System (CRS).
  - There is a common agreement on the variable being measured and its units of measure.
  - The processing chain targets a group of users that work on a common topic (e.g. land, water, atmosphere).
  - There is an assumption of a continuous flow of data and some revisiting period creating a time series.
- Data usage is simplified:
  - There is an effort made by the producer to reduce the need for pre-processing by integrating the necessary correction in its product production chain.
  - Data is made available in an easy-to-use format. This is a commonly known and well documented format that is supported in multiple software products.

- The product is well documented (metadata) at the product level as well as a pixel level.

Even the though the ARD concept originated in the remote sensing community, none of these definitions limits the concept to only satellite remote sensing data.

6.2. CARD4L definition

CEOS is adopting the following definition of ARD:

"Satellite data that have been processed to a 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" [5]

In principle, this definition is not limiting the use of the data to a particular domain. However, in practice this has been the case because the definition was initially formulated and adopted by the CEOS ARD for Land (CARD4L). This obviously assumes that ARD is created for the land community. However, the same definition is now being used in other emerging domains such as aquatic media or atmosphere.

During the processing required by CARD4L, the products are resampled onto a common geometric grid (for a given product) and provide baseline data for further interoperability both through time and with other datasets.

CARD4L products are intended to be suitable for time series analysis and multi-sensor application development. They are also intended to support rapid ingestion and exploitation via high-performance computing, cloud computing, and other future data architectures.

In practice, every type of sensor content could be processed to produce a type of ARD. This type of ARD represent a realization of the definition into a concrete product type. The definition of the characteristics required for theses product types are collected in Product Family Specifications.

Figure 8. CEOS Analysis Ready Data Logo

6.2.1. CARD4L Product Family Specifications

A Product Family Specification (PFS) is a document that describes a specific product that can be generated from a generic sensor type. PFS contains a list of requirements for general metadata, per-pixel metadata (e.g. quality flags for clouds, shadows, no-data values, etc.), radiometric and atmospheric corrections, and geometric corrections. These processes are described in a simple way in [6]. For each aspect, a minimum requirement is optionally provided (called the Threshold level), complemented by a recommendation to deal with the aspect in a better way (called the 'Target' level).

Products that meet all threshold requirements should be immediately useful for scientific analysis or decision-making. Products that meet target requirements will reduce the overall product uncertainties and enhance broad-scale applications.

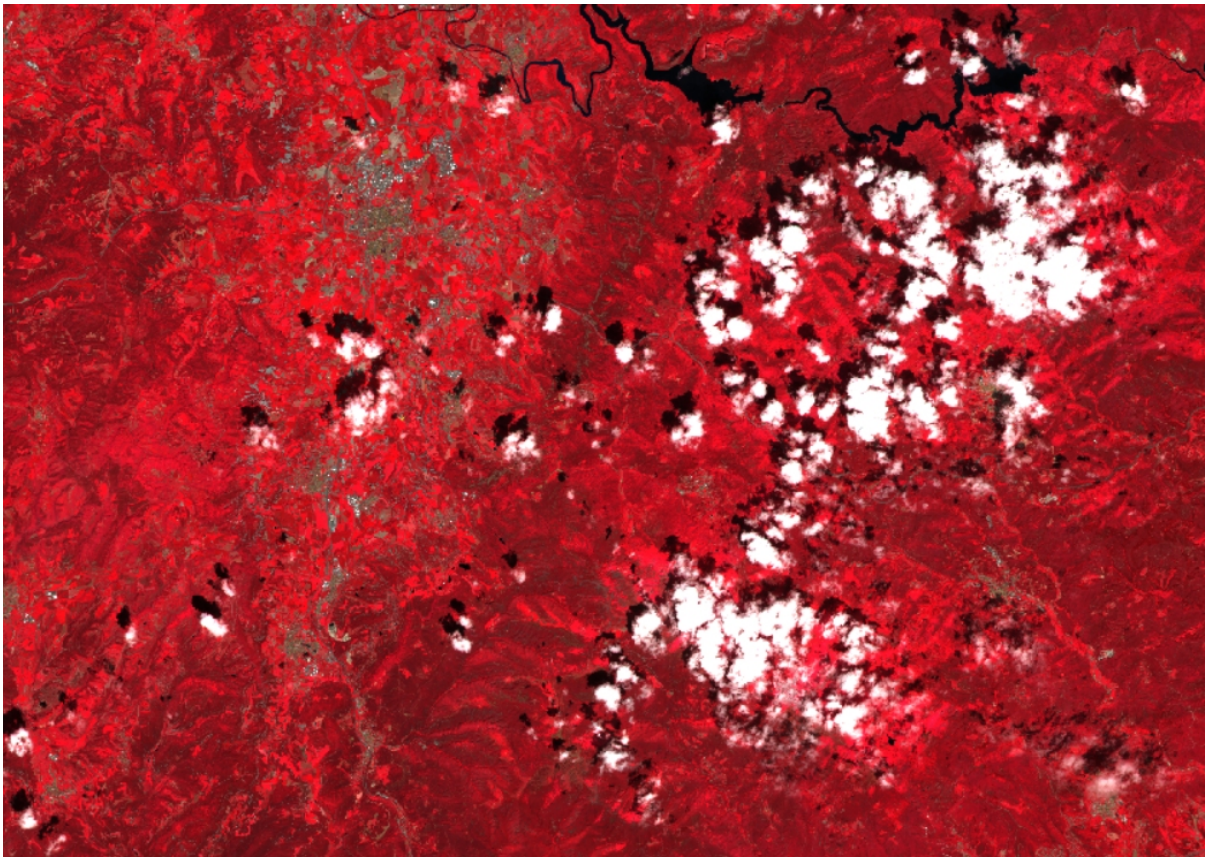

Figure 9. Sentinel-2 Level 2A False color image over an area in Vic (Central Catalonia) showing two artificial lakes, clouds and shadows.

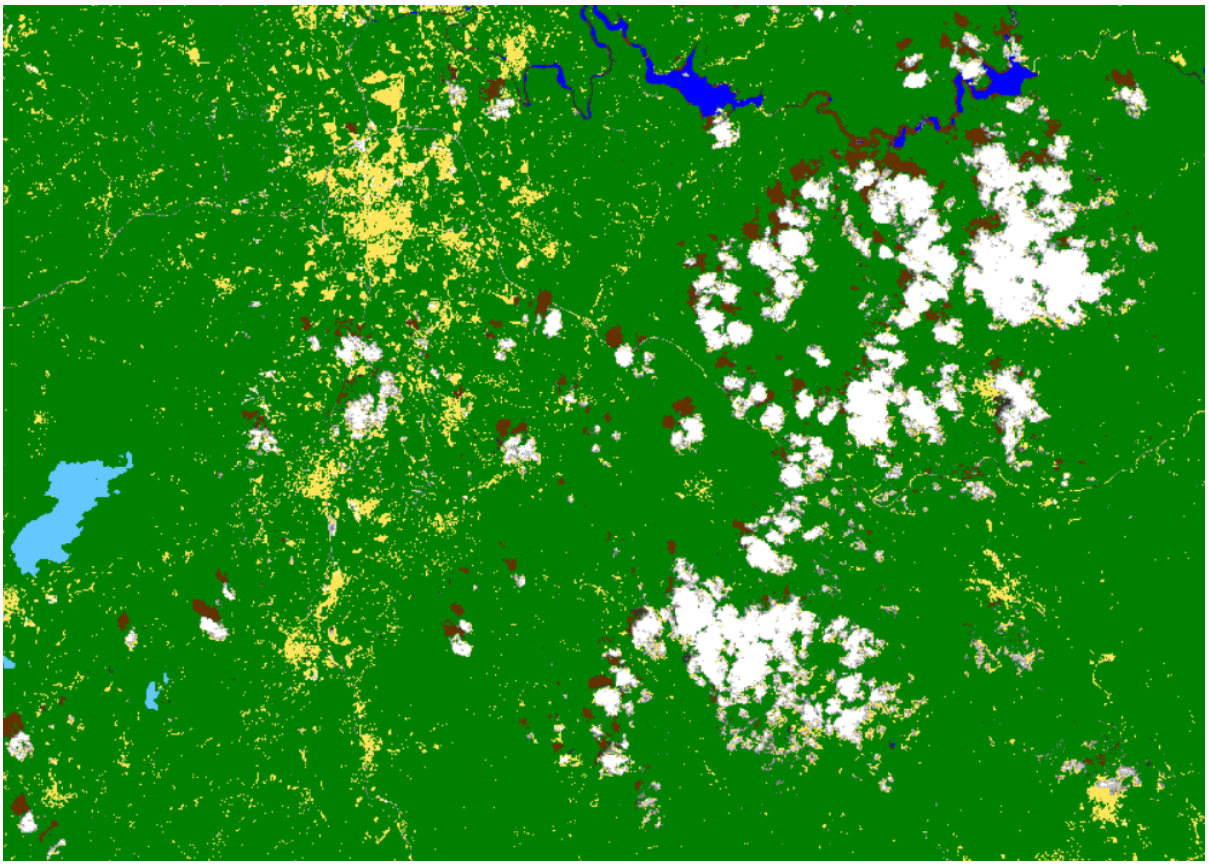

Figure 10. Sentinel-2 Level 2A per pixel data quality flags indicating vegetation, bare ground (mostly built environment), water, thin cirrus, clouds and shadows.

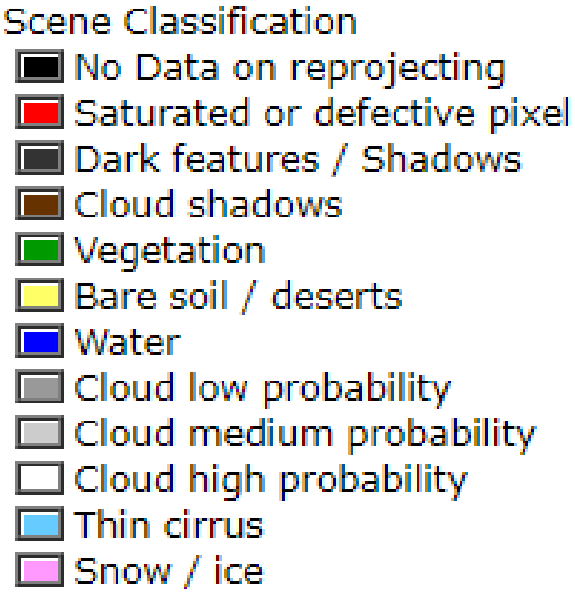

Figure 11. List of data quality flags used in the Sentinel-2 Level 2A distribution.

6.3. Current applications

Currently the CEOS CARD4L website offers the following PFS:

- Optical sensors
  - Surface Reflectance
  - Surface Temperature
- Synthetic Aperture Radar
  - Normalized Radar Backscatter
  - Polarimetric Radar

CEOS is also preparing a PFS for aquatic media:

- Aquatic Reflectance

CEOS is also considering extending the concept to the atmospheric domain.

6.4. Applicability beyond Remote Sensing

As stated above, the ARD concept originated in the remote sensing community. In the opinion of the authors of this ER, the CARD4L definition for remote sensing can be extended to address other domains in which geospatial data is created. Indeed, the characteristics of the ARD listed above are not exclusive to satellite data and can be expressed in the following general concepts:

- The data should represent one or more physical variables (the values should not be just digital numbers).
- Data should be georeferenced in a common CRS (indirect references or georeferenceable data should be avoided).
- Data values should be homogeneous and comparable in time (and become a time series).

- Data should be quality flagged, marking or removing wrong or missing values.
- Data should be fully documented, and processing done to prepare the data and semantics should be recorded in the metadata.

Every data type could have its own PFS that lists these requirements.

6.4.1. Role of the OGC in defining ARD types (and the OGC Definitions Server)

The creation of a workflow by combining processes is still a manual task that requires knowledge of the characteristics of the processes available and user experience. GIS analyists do this for a living. Today, Geographic Information Systems (GIS) tools do a poor job selecting the inputs compatible for a given process. Most of them only filter data inputs by file extension or format. A better automatic selection could be achieved by a better analysis of the metadata characterizing the data. In particular, metadata about the meaning of the values in the data could help. Attaching semantics associated with concepts about the meaning of the values of the data should improve the matching between processes and data inputs.

ARD PFS define a very concrete set of requirements for a product. Data complying with a PFS could be semantically tagged as such. Processes and workflows compatible with an ARD PFS could also have their inputs semantically tagged with the same tags. The OGC Definitions Server [http://defs-dev.opengis.net/def/ready-data] is a database of concepts that can be used for semantic tagging.

The participants propose having a concept for each of the products characterized by CEOS as a PFS. Each concept is associated with a URI that can be included in the metadata of the product. The usefulness of this URI is explained later.

For this Testbed, a Ready Data scheme was created in the OGC Definitions Server (developer instance).

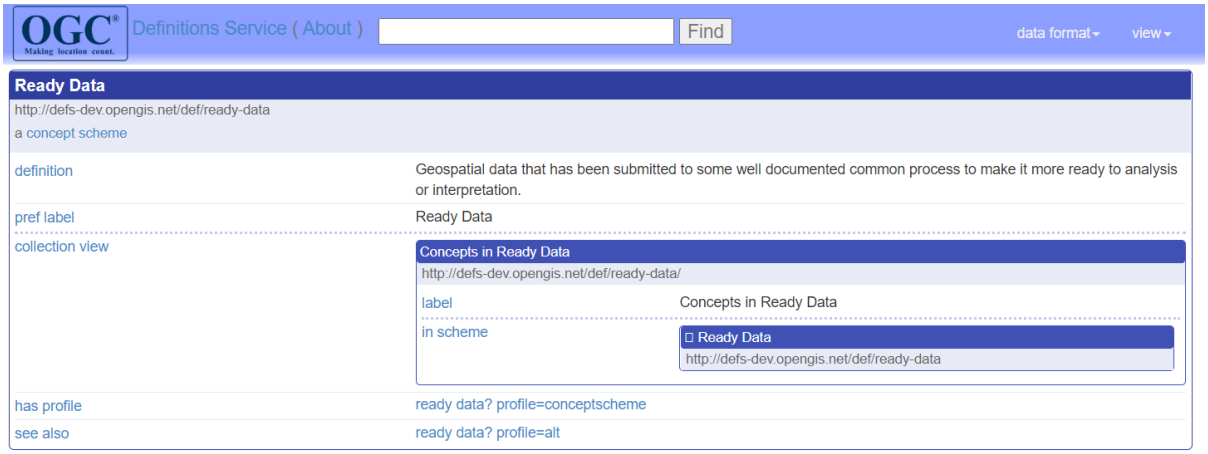

Figure 12. Definition Server Ready Data scheme.

This schema contains a definition and a URL for each of the current CARD4L PFS. For example, the Surface Reflectance PFS has the URL http://defs-dev.opengis.net/def/ready-data/pfs-surface-reflectance.

Figure 13. Definition Server response to the PFS Surface Reflectance URL

In the future, data sources claiming to be compatible with CARD4L Surface Reflectance PFS could include the http://defs-dev.opengis.net/def/ready-data/pfs-surface-reflectance in its metadata. A process compatible with this particular product will then be able to take it as an input automatically.

6.5. Technical Readiness

The CEOS definition of ARD focuses on the semantics aspects of the data but intentionally avoids mentioning technical details on how the data should be made available. The definition does not give indications about the data format or the data APIs that should be used to provide ARD. The CEOS definition focuses on *content readiness* instead of *technical readiness*. *Content readiness* is a necessary step for easy analysis that clarifies the semantics of the data. Some researchers are also sensitive to *technical readiness* and consider readiness part of what ARD should be. In this section we present two degrees of *technical readiness* that some actors, particularly in the private sector, consider as part of the requirements defining ARD. The implications of the *technical readiness* in the concept of CEOS ARD are analyzed in the CEOS Interoperability Terminology [7] document, on section *Analysis, Access, and Analysis Ready Data*

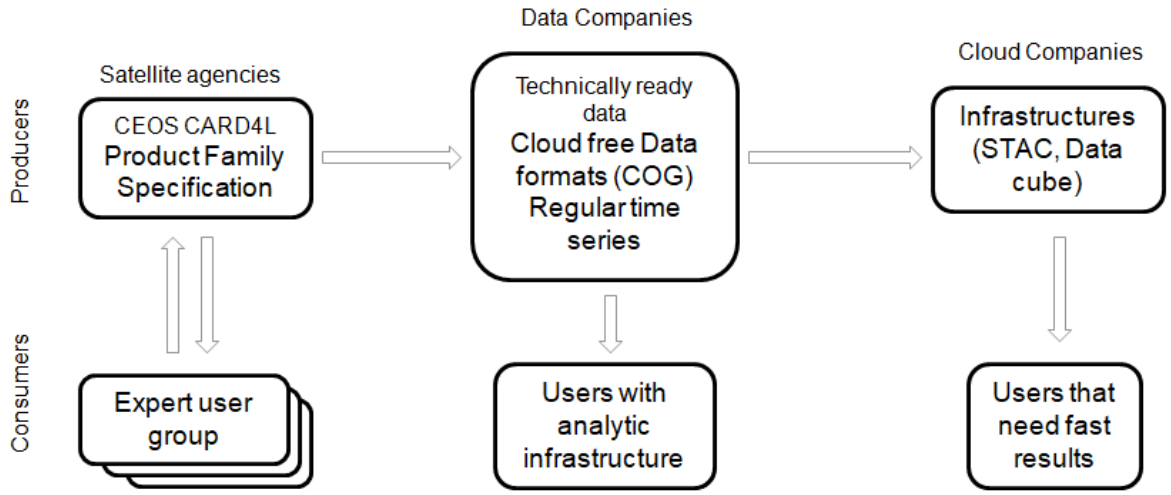

Figure 14. Three interpretations of the ARD concept depending on the target users. The last two interpretation deal with technical readiness.

6.5.1. Cloudless, mosaics and regular time series

The first step towards facilitating technical readiness is to provide a more continuous dataset. For optical data, one of the main issues is the presence of clouds. One common practice to solve this issue is to combine several images captured close in time and select the pixels that do not contain clouds and create a cloud free scene. Since the resulting product is a combination of multiple scenes, the product is commonly distributed as a periodic time series (e.g. every 16 days). The product is in a distribution pattern that covers the area without gaps organized in tiles that are independent of the satellite paths. Some authors consider these extra steps necessary for their products to be considered ARD (e.g. GLAD2020). To simplify access, these tiles are provided in an easy-to-use format such as Cloud Optimized GeoTIFF (COG).

6.5.2. Infrastructures and data cubes

A second step towards facilitating technical readiness involves moving the processing infrastructure for the consumer to the user. Some companies and space agencies are trying to make data easy to use by providing a cloud-based processing infrastructure that provides access to the ARD through APIs. Some providers call these processing services *ARD*. One example of this approach is explained in [8]. The article emphasizes this aspect by calling it "Cloud Ready Data (CRD)". The article describes ARD to:

- Efficiently support complex queries on huge datasets.
- Support highly parallel read and write.
- Store highly complex many-dimensional datasets.
- Ensure metadata and data relations are maintained.
- Integrate with high-level tools that allow working in the domain paradigm instead of the data paradigm.

The article summarizes these characteristics in this sentence: *"Analysis ready data is data made available with the tools, documentation and infrastructure to allow instant and easy analysis across the entire domain"*. The author complains that making data open in the *wrong* format is making

data open to a small niche of domain specialists that can afford to spend time understanding the format thereby excluding many organizations and people.

Even if there is agreement that *technical readiness* is very important for achieving good coupling with real implementations of processing algorithms, the participants in this OGC activity consider these aspects complementary to the ARD definition but not part of it. The ER section Tools for using ARD comes back to these issues.

6.5.3. Pairing data to processes

From an OGC services point of view, data processing algorithms are exposed via a Web Processing Service (WPS). They are individual processes that are characterized by their inputs and outputs. Processes are analytical tools that can only produce results if the input data is available and ready. In that sense, each WPS process define the requirements for the data inputs it is able to handle. Each process defines its own Process Ready Data (PRD). Creating a reusable workflow to process data into useful information requires being able to combine processes efficiently in a process chain. In order for the chain to work, the previous process should produce a PRD compatible for the next process to run. From workflow point of view, ARD is the PRD of the first process of the workflow and IRD should be the result of the last process of the workflow.

The level of reusability of a workflow will depend on the constant availability of ARD for different geographical extents and data providers.

In this chapter the ARD concept and examples were presented. The next chapter presents how the producers are embracing the concept to provide ARD to the users.

Chapter 7. Where to find ARD

Satellite Space Agencies, research centers, and companies are receptive to the ARD concept and are starting to publicize their ARD offerings in different ways. Today, despite the observed popularity of the concept such as by The Group on Earth Observations (GEO), the number of offerings is still limited.

7.1. Satellite data providers

7.1.1. ARD for their satellites

Since 1972, the joint NASA/ U.S. Geological Survey Landsat series of Earth Observation satellites have continuously acquired images of the Earth's land surface, providing uninterrupted data. They were the first agencies that embraced the ARD concept. The long-lasting program is composed of a series of satellites with different characteristics. Fortunately, the USGS was able to reorganize its production process to create a homogeneous product called "Collection 1". The USGS clearly states that the Landsat Collection 1 is ARD in a Control Book [https://prd-wret.s3.us-west-2.amazonaws.com/assets/palladium/production/atoms/files/LSDS-1873_US-Landsat%20C1-ARD-DFCB-v7.pdf]. This ER clarifies that ARD originated from the "Level 2" (NOTE: "Level 2: processing refers to the generation of Top of Atmosphere (TOA) Reflectance, Surface Reflectance (SR), TOA Brightness Temperature (BT), Quality Assessment (QA), and Surface Temperature (ST) scenes as inputs to ARD". Landsat is preparing a second version of the homogeneous version that is called "Collection 2". This product is not publicized as ARD yet but the intention is to have a ARD distribution [https://prd-wret.s3.us-west-2.amazonaws.com/assets/palladium/production/atoms/files/Landsat-C1vsC2-2020-0602-lmws.pdf]. Since the USGS web content does not reflect the ARD component or the Collection 2, the OGC asked USGS directly and they confirmed that Collection 2 will be ARD.

ARD from the USGS can be downloaded using several methods [https://www.usgs.gov/core-science-systems/nli/landsat/landsat-data-access?qt-science_support_page_related_con=0#qt-science_support_page_related_con]. EarthExplorer [https://earthexplorer.usgs.gov/] is the most popular. The Landsat ARD is only available for the US territory.

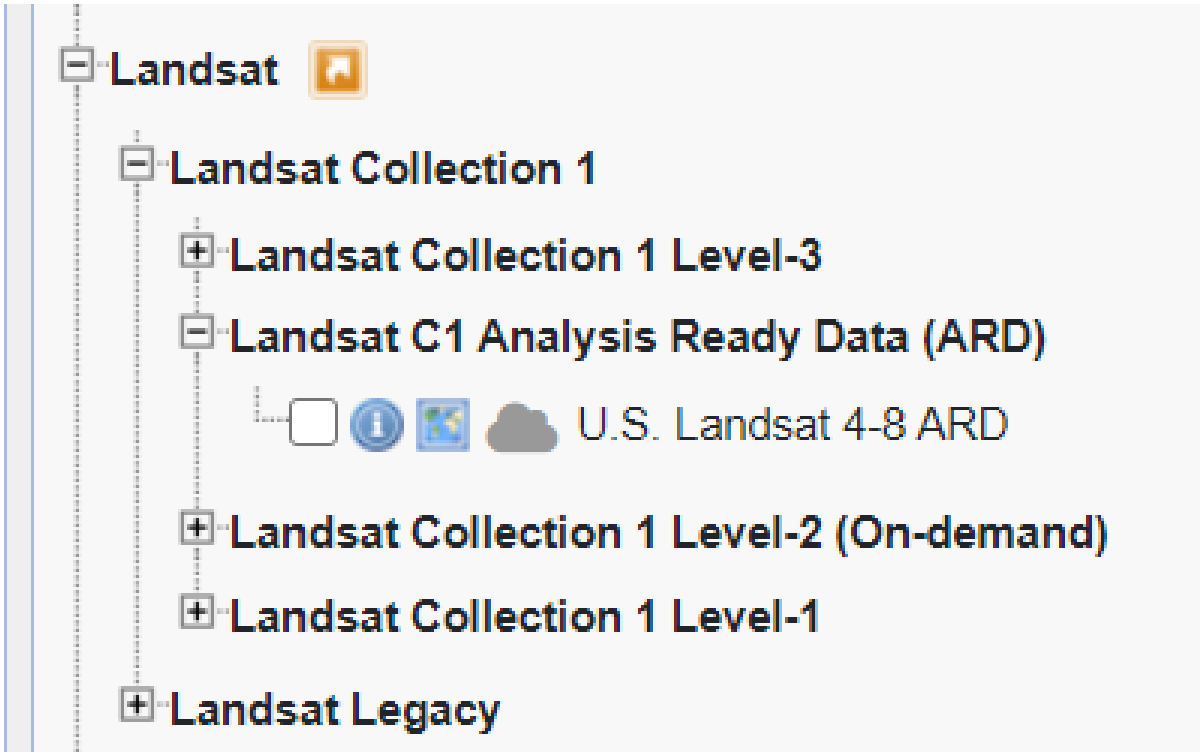

Figure 15. USGS ARD offered by EarthExplorer as one of the many products available.

The European Space Agency (ESA) manages the Copernicus Sentinel satellite constellation. The Sentinel-2 A and B satellites provide high resolution optical products. ESA produces a BOA product called Level 2A that is available in the Copernicus Open Access Hub [https://scihub.copernicus.eu/dhus/#/home]. This product is not advertised as ARD but it is very close to complying with all the CEOS PFS as stated in this presentation submitted to the ESA Living Planet Symposium 2019 [9].

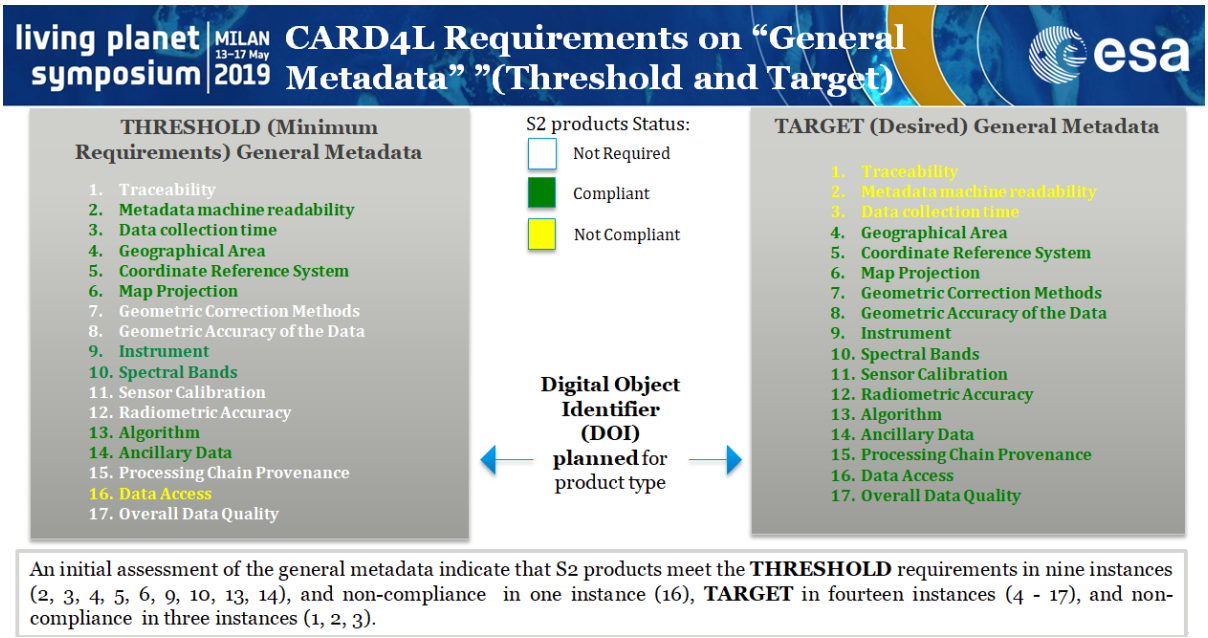

Figure 16. ESA Living Planet presentation slide showing how close Sentinel-2 Level 2A product is to the CEOS4L PFS. (Boccia V. et al. 2019)

The intention of ESA is to progress to CARD4L compliance and aim to have all their products conforming to CARD4L PFS.

The Sentinel-2 Global Mosaic (S2GM) service offered as a component of the Copernicus Global Land

Service provides composites from time series of Sentinel-2 surface reflectance observations. This product is advertised as Analysis Ready Data (ARD) for sustainable management of natural resources [https://land.copernicus.eu/imagery-in-situ/global-image-mosaics/node/16]. Once again, a homogeneous and regular time series is often considered a necessary feature of ARD even if the CARD4L PFS does not require or recommend that.

The need to produce CEOS ARD from Sentinel-2 is emphasized in a UK report from the Catapult Network [https://catapult.org.uk/].

> "If the anticipated industry growth in the UK is to be realized, the community must quickly establish access to reliable, operational data services for ARD. This project [Sentinel2 ARD] sought to support the UK moving towards this ideal [...]. Currently, a suitable tool for producing Sentinel-2 ARD is not openly available, nor a standard associated with this product type. Catapult and Defra EOCoE identified an opportunity to fund a piece of work to help facilitate UK exploitation of Sentinel-2, while simultaneously raising the profile of EO and UK expertise in this field." (https://media.sa.catapult.org.uk/wp-content/uploads/2017/09/14123619/Sentinel-2-ARD-Project-Summary_final.pdf).

Joint Nature Conservation Committee (JNCC) is taking the lead on filling this gap of Analysis Ready Data (ARD) products for the Sentinel-1 and Sentinel-2 platforms by pre-processing data to a proposed UK standard to be adopted in academia, industry and government. The generation of ARD ensures continued UK contribution to wider international community efforts to develop standards and methods of EO data access. JNCC is the public body that advises the UK Government and devolved administrations on UK-wide and international nature conservation. (https://jncc.gov.uk/our-work/analysis-ready-data-ard/).

The Japanese Space Agency (JAXA) offers SAR ARD [http://www.eorc.jaxa.jp/ALOS/en/palsar_fnf/fnf_index.htm]. ARD is offered for the following missions JERS-1/SAR (1996), ALOS/PALSAR (2007-2010), and ALOS-2/PALSAR-2 (2015-). The data is offered free of charge and it is available for the Annual summer data (June-September) and covers the whole world in granules of 1 x 1 deg. or 5 x 5 degrees. Data is comprised of gamma-zero image and forest/non-forest (FNF), ancillary data (local incidence, mask, etc.), spaced in 0.8 arcsec (approx. 25 m) and averaged in 100 m, ortho-rectified, and slope corrected. The geometric accuracy is 10 m [10], [11].

The National Remote Sensing Centre (NRSC) and Indian Space Research Organization (ISRO) have also embraced the CARD4L approach for their Resourcesat-2 and Resourcesat-2A satellites. They combine an atmospheric correction algorithm with a quality flag mask generation algorithm. The ARD is offered through an ordering platform. In the platform consumers can request the generation of the ARD products in an scalable way and on-demand [12].

7.1.2. Harmonized multiagency products

There are some emerging initiatives to generate ARD that combine two products in a way where they can be analyzed together. A common example of this are the efforts done to create a combined Sentinel-Landsat product. Landsat and Sentinel-2 data represent the most widely accessible moderate-to-high spatial resolution multispectral satellite imagery. Following the launch of the two Sentinel-2 satellites in 2015 and 2017, the potential for synergistic use of Landsat and Sentinel-2

data creates unprecedented opportunities for timely and accurate observation of Earth status and dynamics.

The Harmonized Landsat and Sentinel-2 (HLS) project is a NASA initiative aiming to produce a Virtual Constellation (VC) of surface reflectance (SR) data acquired by the Operational Land Imager (OLI) and Multi-Spectral Instrument (MSI) aboard Landsat 8 and Sentinel-2 remote sensing satellites, respectively. The HLS products are based on a set of algorithms to obtain seamless products from both sensors (OLI and MSI): atmospheric correction, cloud and cloud-shadow masking, spatial co-registration and common gridding, bidirectional reflectance distribution function normalization and spectral bandpass adjustment. Three products are derived from the HLS processing chain:

- S10: full resolution MSI SR at 10 m, 20 m and 60 m spatial resolutions;
- S30: a 30 m MSI Nadir BRDF (Bidirectional Reflectance Distribution Function)-Adjusted Reflectance (NBAR);
- L30: a 30 m OLI NBAR.

All three products are processed for every Level-1 input products from Landsat 8/OLI (L1T) and Sentinel-2/MSI (L1C) [13]. Data can be downloaded from here [https://hls.gsfc.nasa.gov/data/v1.4/].

ESA has developed an Analysis Ready Data OnDemand demonstrator service for Sentinel-2 and Landsat-8 implemented by ESA-Research and Service Support (RSS). The tool provides a web interface where users can select L1C data. The user can visualize the L1C footprint on a map. The user can also select one of the available Atmospheric Correction Algorithms, the output projection and format, and then submit a processing task On-Demand. An effort to harmonize the results to provide as much as possible a uniform atmospheric corrected product is done using the following four tools as an starting point: Sen2Cor, LaSRC, iCOR and MAJA [14]

7.2. Research Centers

Research centers can also provide ARD. Production of ARD can be the result of their pre-processing of data in preparation to environmental research projects such as forest status, land cover change etc. Since they already do that for themselves, they may decide to distribute the ARD for the benefit of others.

The Global Land Analysis and Discovery (GLAD) laboratory in the Department of Geographical Sciences at the University of Maryland produces another ARD product based on Landsat. The essence of the GLAD ARD approach is to convert individual Landsat images into a time-series of 16-day normalized surface reflectance composites with minimal atmospheric contamination. The global Landsat ARD product is provided as a set of 1x1 degree tiles. Each data granule is provided as a GeoTIFF file containing observation data collected for a single 16-day interval. There are 23 intervals per year. This product is not trying to infer BOA radiation using a model of the atmosphere (as the USGS ARD does). Instead GLAD is trying to harmonize the radiation using places on the Earth where radiation is supposed to be constant (pseudo-invariant areas). In addition to that, GRAD is providing a product that is regularly distributed in time.

ARD from GDAD can be downloaded here [https://glad.umd.edu/dataset/landsat_v1.1] (registration required).

7.3. Companies Adding Value

Mainly companies providing ARD are looking for a niche by facilitating usage of data by their costumers. Actually, the lack of a clear existing product for Sentinel-1 and Sentinel-2 that can be considered ARD created a niche for offering products that can be branded as ARD. The following paragraphs provide some examples.

EOX IT company is a partner in the DIAS MUNDI and is helping to develop more efficient ways how to disseminate the Sentinel-1 and Sentinel-2 data and information. EOX produce land use and agricultural applications such as crop classification and crop monitoring resulting in what EOX call CAP Analysis Ready Data [https://eox.at/2019/05/eoxcloudless-sentinel-1-and-sentinel-2-analysis-ready-data/#:~:text=CAP%20Analysis%20Ready%20Data]. More details can be found in the EO4Agri project report [https://eo4agri.eu/sites/eo4agri.eu/files/public/content-files/deliverables/EO4AGRI_D3.9-Guidelines-on-EO-ICT-support-improvement-to-stakeholders-v1_v1.2.pdf].

Terramonitor is another company offering ARD [https://feed.terramonitor.com/analysis-ready-use-cases/] Terramonitor describe their ARD offering as a *cloudless mosaic* of preprocessed satellite data with 10-band multi-spectral data. The images used for mosaic are atmospherically corrected and radiometrically calibrated. The product is suitable for different use cases such as analyzing and monitoring land areas, vegetation, forests, and calculating indices.

Instead of offering ARD, PCI Geomatica is creating and selling a set of tools to create ARD or improve existing ARD. These tools cover areas such as intercalibration of reflectance or radiance data, topographic normalization, geometrical correlation and data quality masks. PCI Geomatica is using Open Data Cube as a repository for ARD and provide tools to direct ingestion of the data in ODC [15].

Astraea is providing a service to simplify the process of using satellite information using a facility they call EarthOnDemand to access ARD and another service called EarthAI Workflow to process ARD and extract results. In this case, the offering is more focused on providing a platform to process ARD [https://astraea.earth/].

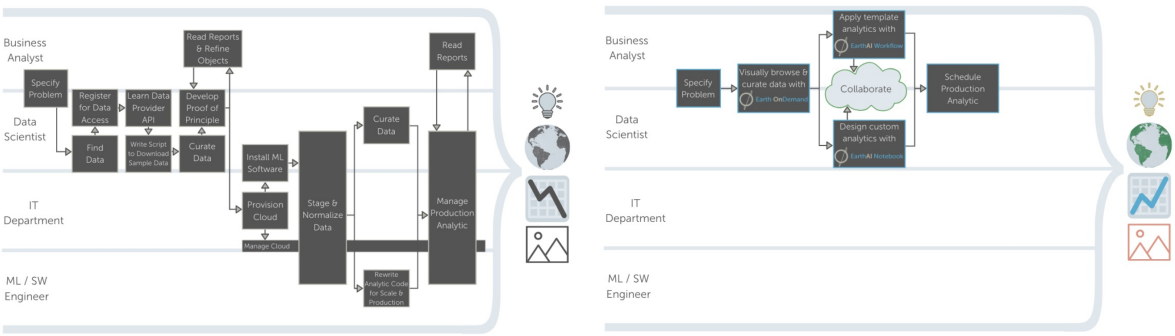

Figure 17. Astraea diagrams providing a graphical representation of a common workflow (left hand side) and their platform offering workflow (right hand side). This diagram is reproduced in this ER as the left-hand side illustrates common problems that users face in satellite data access even if the data is CARD4L conformant. Only the adoption of an extra architecture will help make the use of remote sensing data easier.

7.4. Continentally Wide Initiatives

The ARD phenomena has stimulated the creation of continent wide ARD platforms. These are the most well-known:

Digital Earth Africa (DE Africa) is a platform for the African nations. Users can access high-resolution satellite images from a single platform. DE Africa has made available a vast amount of data captured by ESA Sentinel-2 satellites in a format that makes the imagery accessible and suitable for general use. Satellite images captured by Sentinel-2 are particularly important for Africa because they offer 10 m resolution and are captured every 5 days, so land and water can be analyzed in unprecedented detail (https://www.digitalearthafrica.org/). DA Africa uses the Open Data Cube [https://www.opendatacube.org] open source solution.

Digital Earth Australia (DEA) is a platform that uses spatial data and images recorded by satellites to detect physical changes across Australia in unprecedented detail. DEA prepares these vast volumes of Earth observation data and makes it available to governments and industry for easy use (http://www.ga.gov.au/dea/about). Not surprisingly, DEA also uses the Open Data Cube [https://www.opendatacube.org], as the Digital Earth Australia is an active part of the Open Data Cube (ODC) development.

Euro Data Cube provides a one-stop-shop for EO and a platform to analyze an event or phenomena from different perspectives, providing multiple data sources in a way that several variables can be compared and correlated at the same time in a customized data pipeline. Euro Data Cube is a combination of several services in a single platform. (https://eurodatacube.com/documentation/about). Euro Data Cube is a proprietary solution developed by a partnership between industry leading companies. However, the solution shares some similarities with the ODC such as the use of similar python libraries and *XArrays* numeric libraries.

Chapter 8. Architectures to provide ARD

This chapter focuses on how the producer can offer ARD to the consumer. The chapter analyzes the different approaches that have been implemented or proposed by ARD producers to derive ARD from low level remote sensing data and how to expose that content to the consumer.

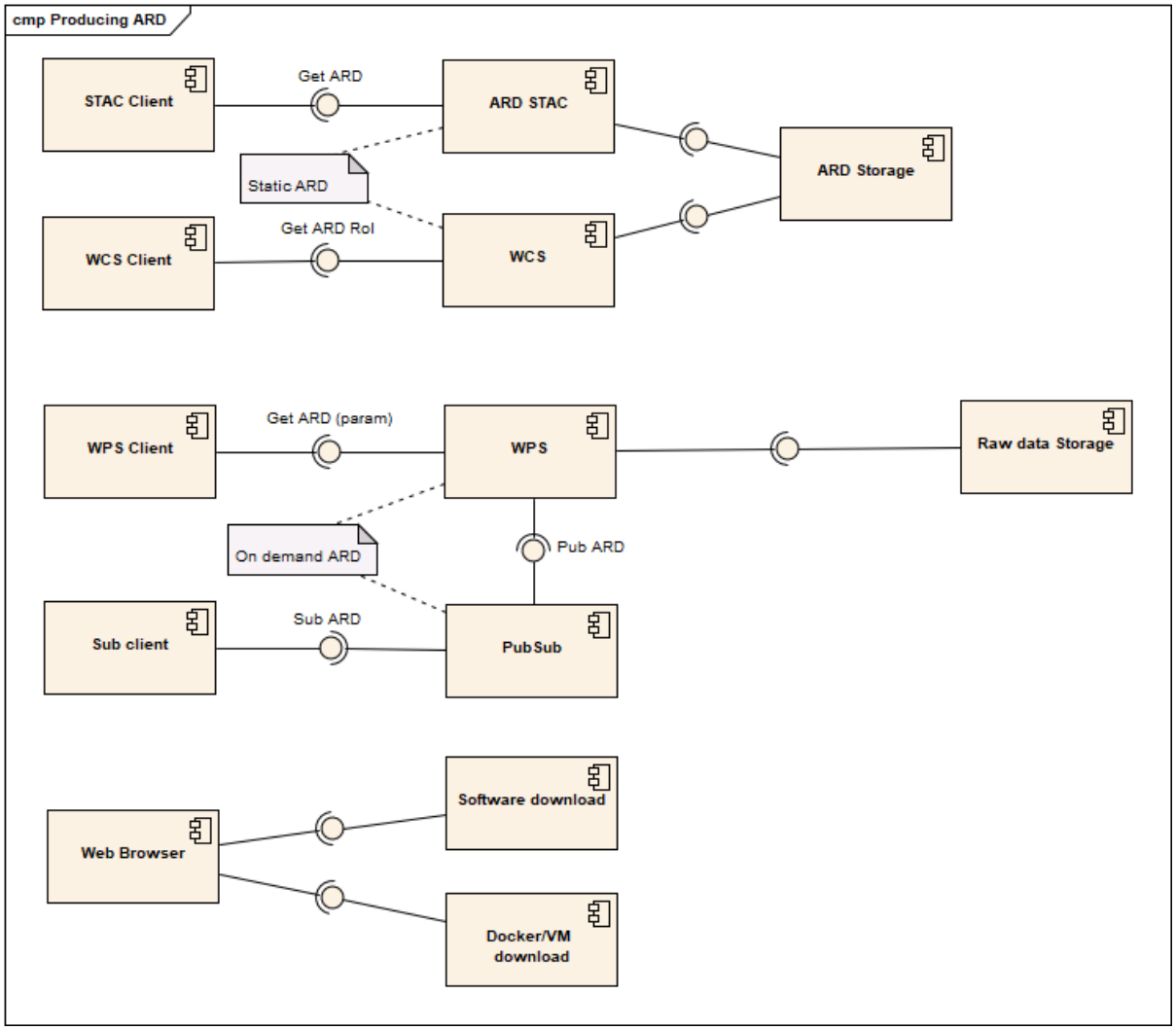

Figure 18. How producers offer ARD to consumers

8.1. Pre-prepare and download

Processing ARD takes time and effort. If all the areas of the data are considered equally interesting and there is an expectation that data is going to be consumed several times, an approach based on executing the processing chain to create ARD as soon as a new fragment of information is acquired is reasonable.

In this approach, the data is divided into a regular grid of big pieces (sometimes referend as granules) that are made available as files to be downloaded. The producer needs to have enough storage space to be able to offer all possible granules in an agile way.

To facilitate the process for occasional users, a data discovery facility could be setup. A web portal allows specifying the area of interest and the time period required and a list of links to access the individual granules that matched the conditions is created. Each link gives immediate access to the

data at the granule level.

For recurrent users, the provider should include an API that enables machines to automatically download the needed resources. A SpatioTemporal Asset Catalog (STAC) [https://stacspec.org/] can be used to implement a machine-to-machine download facility.

8.2. Region of Interest

Sometimes a granule might be too big compared to the consumer's region of interest or the region of interest could be accidentally in between four granules. In these cases, a facility that provides the capability to clip and extract the area that the users' need could be an interesting alternative to a pure file-based download system.

To facilitate the process for occasional users, a data discovery facility could be setup. A web portal allows specifying the area of interest and the time period required. Instead of a list of links to access the individual granules, an on-demand dataset is created with the data clipped to the area requested and the granules involved mosaicked.

This approach still requires preparing the ARD from the original sources and adds the need for some computational resources to process the area of interest

For recurrent users, the provider should include a web service accessed via an open API that allows machines to automatically download the needed resources. The OGC Web Coverage Service (WCS) or the emerging OGC API - Coverages can be used to implement a machine-to-machine download facility.

8.3. On demand generation

In this case, the ARD is not available directly and is created as part of the request process. This option addresses two possible cases: Lack of space for the provider to store the ARD and need to adjust some parameters before creating ARD.

To facilitate the process for occasional users, a data discovery facility could be setup. In this case, the web portal will allow for discovery of potential ARD products based on the existing raw data. There is the possibility of specifying parameters to condition the way the ARD will be generated (e.g. the possibility to provide a better DEM of the region of interest). The processing takes time and the actual products are delivered asynchronously. A common practice is to email the consumer when the products are ready to download.

Some asynchronous responses are difficult to automate (such as the email responses). For machine-to-machine communication it is better to use another form of notification such as an HTTP page that can be visited regularly until the products appear. This approach can be implemented as a WPS service. Other approaches may rely on a PubSub protocol such as MQTT [https://mqtt.org/].

This approach reduces the need for storage space but creates the need for scalable processing capabilities.

8.4. Toolbox for the user

Finally, the producer can delegate to the user the production of ARD. This way, the producer provides the necessary raw data as well as a software and auxiliary data to create ARD. The consumer will have to download the raw data and create the ARD in his/her computer infrastructure. The distribution of the software can be done as executable files that are specific to an operating system. When the setup process has several dependencies, the distribution can be encapsulated in a virtual machine for easy deployment or as a Docker container. The latter is useful for a deployment in the cloud when the consumer does not have enough inhouse computing capacity.

This scenario is ideal for immediate deployment of new types of ARD without forcing the producer needing to deploy new storage or computing facilities.

The sen2cor [https://step.esa.int/main/third-party-plugins-2/sen2cor/] module developed and distributed by ESA is an example of this kind of tools. Sen2cor is capable of taking a Sentinel-2 Level 1B and applying a transformation that generates ARD Surface Reflectance (Sentinel-2, Level 2A). This tool was initially distributed as part of the Sentinel Application Platform (SNAP) for anybody to download. With time, ESA was able to deploy the necessary storage capability to distribute it as a new product in its Sentinel Data Hub.

This chapter reviewed a set of approaches to provide analysis ready data to the users. In the next chapter, a set of tools to use ARD is presented.

Chapter 9. Tools for using ARD

In the CARD4L definition of ARD there is an assumption that the users have the access to a "rapid ingestion and exploitation via high-performance computing, cloud computing and other future data architectures" that will enable them to analyze the ARD. This chapter discusses elements and tools that users can use to take advantage of ARD. Some of these elements have been considered in other OGC Testbeds and OGC Pilots. More information can be found in the OGC Earth Observation Applications Pilot: Summary Engineering Report (OGC 20-073) [16].

9.1. Discovering ARD

Currently there are not many satellite products that can be considered ARD, so discovering these products is not difficult. In the future, an increase in the number of products available is expected. This trend includes the generalization of the PFS for many other data types. In the future, metadata catalogues will be needed to identify ARD products using a metadata element such as a keyword, a tag or a URI and also indicate which ARD type (i.e. PFS) the product complies with.

9.2. Downloading files with ARD

Satellite information (as well as remote sensing based ARD) is still distributed in large files. Imagery collected for entire satellite orbits and paths (considering the swaths they cover) are too large to be saved in a single file so they are cut into a series of scenes and granules that make the management of the files practical. Unfortunately, a pure file system is not prepared for geospatial data indexing. Therefore geospatial information in files is limited to scene identifiers and dates encoded in the file name (that is repeated in some internal metadata format). Still, the list of files necessary for the study of a region over time is very large and the names of the files are not easily queryable therefore becoming confusing and difficult to manage. This can be solved by adding an application that applies a spatial index, such as a R-Tree, on the file system. Some datacube implementation uses this solution.

9.3. Datacubes to Ingest ARD

Datacubes are a good alternative to a file system-based approach as they permit ingesting the files into a database that simplifies the access to the data. A data cube is a data structure that represents a multidimensional array together with metadata describing the semantics of the axes, coordinates, and cells. A data cube may have horizontal (i.e. x/y, lat/long), vertical spatial axes, temporal axes, or any other application dependent dimensions. These data structures as stored in database systems that is able to deliver data on demand. For Remote Sensing based ARD, the system will ingest the files and index the data in the multidimensional cube for fast access. Afterwards, the system is capable of handling requests for data that is inside a geospatial bounding box and a time interval and provide an agile response containing lists of elements (e.g. tiles) that comply to these constraints ([17]).

Each database system provides its own way of formulating queries and encodings to receive data. From the interoperability point of view, the Coverage Implementation Schema (CIS) [http://docs.opengeospatial.org/is/09-146r6/09-146r6.html] specified in the OGC Web Coverage Service WCS

2.0 Standard is a good interoperable tool to both describe the datacube dimensions and values and to also retrieve the content of a multidimensional coverage stored in the datacube.

As it is designed to support consistent time series that are ready for analysis, CARD4L compatible products are populating datacubes. Unfortunately, there are still important barriers for some studies: Continuity and regularity. Optical products are frequently affected by clouds. This has two effects: Some scenes are so covered by clouds that cannot be used, others present occultations and shadows that need to be masked. Other products are not sensitive to clouds but the geometry of the revisiting periods still becomes irregular. Datacubes should incorporate some tools to fill the gaps in images with occultations and are able to create artificial products that are completely regular in time by interpolating dates. Some authors ([18]) suggest that cloud free data and regular time series of cloud free data are two other types of higher level ARD.

9.4. Dockerized Environments to Process ARD

ARD reduces the user's effort by providing the data in a more friendly and consistent manner. Datacubes are a solution to organize and access the data. Unfortunately, they require a significant effort in terms of hardware infrastructure, data download, data ingestion, data management etc. Every user will be forced to repeat the same process. Datacubes provide tools to automate the process. However, instead of building their own datacube and ingesting the data, users should be able rely on someone else to do that for everyone and provide a datacube as a service over the Internet.

This is precisely the objective of the ESA and EC Data and Information Access Services (DIAS). Each DIAS provides a cloud processing capability in an environment where ARD has been already organized as datacubes accessible through APIs. To analyze the data, users have two alternatives: learn a processing language provided by the system or send the code directly to the processing infrastructure to execute it. The second alternative is more interoperable because the same code will work on different processing infrastructures with minimum modifications in the way the data is access. Docker containers are one of the most popular ways of sending code to a remote processing facilities. A Docker container is a standard unit of software that packages up code and all its dependencies so the application runs quickly and reliably from one computing environment to another. A Docker container image is a lightweight, standalone, executable package of software that includes everything needed to run an application: code, runtime, system tools, system libraries and settings. Container images become containers at runtime and in the case of Docker containers - images become containers when they run on Docker Engine. Containerized software will always run the same, regardless of the infrastructure. Containers isolate software from its environment and ensure that it works uniformly despite differences for instance between development and staging. Containers and virtual machines have similar resource isolation and allocation benefits, but function differently because containers virtualize the operating system instead of hardware.

Docker containers can be deployed in infrastructures that give access to datacubes of ARD. The code in the container can immediately start analyzing the ARD. The capacity of the analysis will depend of the type of the ARD available to the users in the infrastructure. Many, if not all DIAS, offer docker container execution. The different DIAS are trying to specialize their approach to find their markets. One specialization approach is to facilitate as many ARDs as possible to attract more users and algorithms.

Considering that the use of cloud processing environments (and Dockerized environments in particular) is not for free, the dilemma between the use of Dockerized environments versus owned datacube infrastructure is similar to the dilemma of renting versus owning a car. The decision as to which approach really depends on the intensity of usage. If an organization or user knows that they are going to use ARD from a particular area frequently for different studies, owning the infrastructure could be a good option. Conversely, if the aim is studying a large geographic region for solving a particular problem, the cloud processing environments can be a more viable solution.

A proposal for an ARD federated architecture is the focus of the next chapter. A requirement is to avoid "lock" into a particular cloud infrastructure.

Chapter 10. Federated architecture for ARD

In this chapter an architecture for exploiting ARD based on the following characteristics is proposed:

- Federated
- Heterogeneity
- Move Analytics to the Data
- Protect the "Crown Jewels"
- Event Driven

10.1. Federated

The concept of federation in information technologies is broad. The OGC Testbed 15 Federated Clouds Security Engineering Report (OGC 19-024r1) [19] suggests that the concept of federation can be interpreted as four different facets:

- *Identity Federation*: Allows sharing (in a secured manner) **identity credentials** with Service Providers (or other Identity Providers) that can either be internal or external to a specific Administrative Domain. A typical example is the use of Google or GitHub account for accessing other services. This approach is used to avoid users having to remember one username and password for each system. From the server point of view, the federated system no longer has to store or manage personal information describing the users and can delegate this responsibility to the federation. Managing personal information is becoming more difficult due to the emergency of laws protecting personal information imposing very strict rules on data e.g. the European General Data Protection Regulation (GDPR). Delegating personal information management to another actor in the federation is better.
- *Authorization Federation*: Allows common access control. Even if an Identity Federation can be used that prevents the user from managing personal information, the federation does not necessarily know about the federated resources that are made available. An authorization federation is able to determine for some given user the rights for access, processing, and so forth the user has for the resources available in the federation.
- *Resource Access Federation*: Provides discovery, registration and access mechanisms for resources. Access to a particular resource may require users to address the proper source hosting it or, alternatively, accessing any host in the federation may take care of providing the resource on request (location transparent federation). Resources can vary in granularity. A resource could be data behind an OGC Web Service, an OGC Web Service operation, or an OGC Web Service as a whole.
- *Service Federation*: Provides access (as a special, trivial case), filtering, processing, or general analytics on resources in a location transparent manner. The user does not have to know about the position of the resources or the processing capabilities of each node to get the job done. In a federation like this, distributed data fusion is accomplished transparently: Users send a request which may involve data from different hosts, and the federation orchestrates dynamically how to get the data and where the data are going to be processed. The user only gets back the final

result.

In the OGC 19-024r1 ER, the concept of federation is interpreted as an *Identity* and *Authorization* Federation where authentication and authorization are managed in a coordinated way across a group of computing or network providers with and agreed standards of operation (such as OpenID connect).

In OGC Testbed 16, in the context of ARD, federation is interpreted as a *Service Federation*: A group of organizations providing their own infrastructure, analytics, and data. In this federation, elements are part of a loosely coupled solution for performing automatic analysis (ARD) across these multiple clouds. There are nodes that provide access to ARD (that can be preprocessed ARD coverages or on-demand ARD processing results) and there are nodes that provide processing capabilities that are combined together in a single infrastructure.

Currently a federation that integrates different cloud providers faces the difficulty that each provider has its own architecture and their own way to facilitate access to data and execution of analytical algorithms. There are some practical solutions to this issue available:

- The Terradue solution solves the problem by providing a neutral place to deploy and test analytical algorithms (the sandbox) that can later be deployed in one of a selection of cloud providers. The solution eliminates vendor lock-in but still does not provide a federated solution because the solution is deployed in only one cloud at a time.
- The OpenEO project [https://openeo.org/] is developing an open API to connect R, Python, JavaScript and other client languages to big Earth observation cloud back-ends in a simple and unified way. The project proposes a lingua franca for cloud processing. When completed, OpenEO will provide a way to operate with any cloud in the same way. OpenEO requires that accessible cloud providers are willing to adopt this approach. This does not necessarily provide a federated approach but having a common language is the first step. If all providers in the network adopt OpenEO, distributing processing in a federated infrastructure could be easier.
- The Rasdaman product is a truly federated solution for coverages. Deployed in several nodes, is able to interpret a query language similar to SQL but with some extensions and then distribute operations automatically in an efficient way.

With some exceptions, such as Google Earth Engine, it seems that heterogeneity of alternative solutions for the cloud is converging into a common solution similar to the Terradue proposal: A Docker package based on Linux virtual machines. A virtual machine of this kind contains all the necessary elements that allow executing code written in a common languages. R and Python are interpreted languages that do not depend on specific hardware, so they are preferable in heterogeneous hardware environments. Java and C++ are also alternatives but they require some level of compilation in any new hardware environment adopted by the federation. In principle a Docker package could access data directly from files but organizing big data repositories as files is not scalable.

An alternative is to access data through an API. This is the configuration that four of the Copernicus Data and information Access Systems (Copernicus DIAS) are using: CREODIAS, MUNDI, SOBLOO and ONDA. They use a form of storage called *object storage*. This is the preferred storage for immutable "Big Data" (up to Petabytes). The storage is designed for write once, read often as indexed blobs. This approach is much simpler to manage and extend than file or *block storage* and

much cheaper. An *object storage* cannot be requested as a normal file and requires an API to transfer the data that will then later be handled as "normal" files. CREODIAS, MUNDI and SOBLOO all use S3 (AWS, GCS standard) *object* storage. ONDA uses ENS (OpenStack Swift). An example is how the Wekeo API [https://www.wekeo.eu/hda-api] enables retrieving the data from the object storage into a "file" that becomes available.

In this situation, the starting point for a federated architecture will assume that for each node of the federation:

- Access to data will be provided by an API.
- Code will be defined in a high-level language (e.g. R, Python) and interpreted and executed by an engine in a virtual machine.
- Authentication is provided by a single sign on mechanism. A username created in one member of the federation will be valid in the entire federation.
- Applications can be registered as trusted applications by users.
- A common authorization mechanism enables granting access to the right computer resources and data.

10.2. Heterogeneity

A Federated ARD architecture should support a broad variety of data types and analytics. The data should not be limited to remote sensing imagery and should include other forms of information such as a long time series of in-situ measurements. Handling the variety of formats and data types cannot be left to the final user. Instead, a common data model should be used. The *data cube* approach provides a layer of abstraction where data are sequences of values associated to a position and time (and eventually other dimensions). Products are then mapped into this data structure that defines the physical dimension of the space and the way the space is mapped into a discrete multidimensional grid. Each measurement arriving (each new remote sensing scene or each in-situ measurement data point) is mapped into the grid definition and the values are stored in cells of the grid. The data cube can be subsetted and resampled in each dimension to retrieve the information that is relevant for a study. The data is recovered in a data structure that contains the definitions of each dimension and the values available on the grid as well as the meaning of the values attached to the cells of the grid. This common data structure is very similar to the memory representation of a NetCDF/HDF [https://www.unidata.ucar.edu/software/netcdf/] format. The same data structure is directly mappable into the OGC Coverage Implementation Schema (CIS) used by the OGC Web Coverage Service (WCS) Standard. The CIS also defines a *DomainSet* that is equivalent to the definition of the dimensions and a *RangeType* that provides information about the values type and meaning.

In Python, a good data structure for representing this multidimensional data cube is the XArray. XArray is based on NumPy [https://numpy.org/] that provides the fundamental data structure and API for working with raw multidimensional arrays. However, the NumPy data structure lacks the necessary metadata to assign array values to locations in space, time and to describe the meaning of the values in the array.

The following example describes a multidimensional array of a matrix of 9 points in space each one with an array of thirty measurements. Each measurement contains two variables: Liquid

precipitation and ice precipitation. The three coordinate values and the array of thirty values are listed.

XArray example

```
<xarray.Dataset>
Dimensions:               (latitude: 3, longitude: 3, time: 30)
Coordinates:
  * time                  (time) datetime64[ns] 2015-01-01T11:59:59.500000 ...
  * latitude              (latitude) float64 12.95 12.85 12.75
  * longitude             (longitude) float64 14.25 14.35 14.45
Data variables:
    liquid_precipitation  (time, latitude, longitude) int32 0 0 0 0 0 0 0 0 ...
    ice_precipitation     (time, latitude, longitude) int32 0 0 0 0 0 0 0 0 ...
Attributes:
    crs:      EPSG:4326
```

10.3. Move Analytics to the Data

In a federation, deploying the same application into different nodes is possible. Each node will get access to the object storage easily accessible by that node. Only the results of the process are transmitted back to the user. Ideally, results will be classifications or summaries of the data that will be smaller in size than the original data.

There are cases where this can be done easily. Let's imagine a use case that requires the most updated imagery available to look for anomalies. A USGS cloud facility provides access to Landsat imagery and a Copernicus DIAS facility will provide access to Sentinel-2 imagery. To find the most recent imagery a process in both facilities is run to get the date of the most current product. The user only gets back two dates and compares which one is newer and determine the facility that has that content. Then the user executes the process that extracts anomalies in that facility. Only an anomalies map is sent back to the user. In terms of data traffic, the solutions resulted in a very efficient flow.

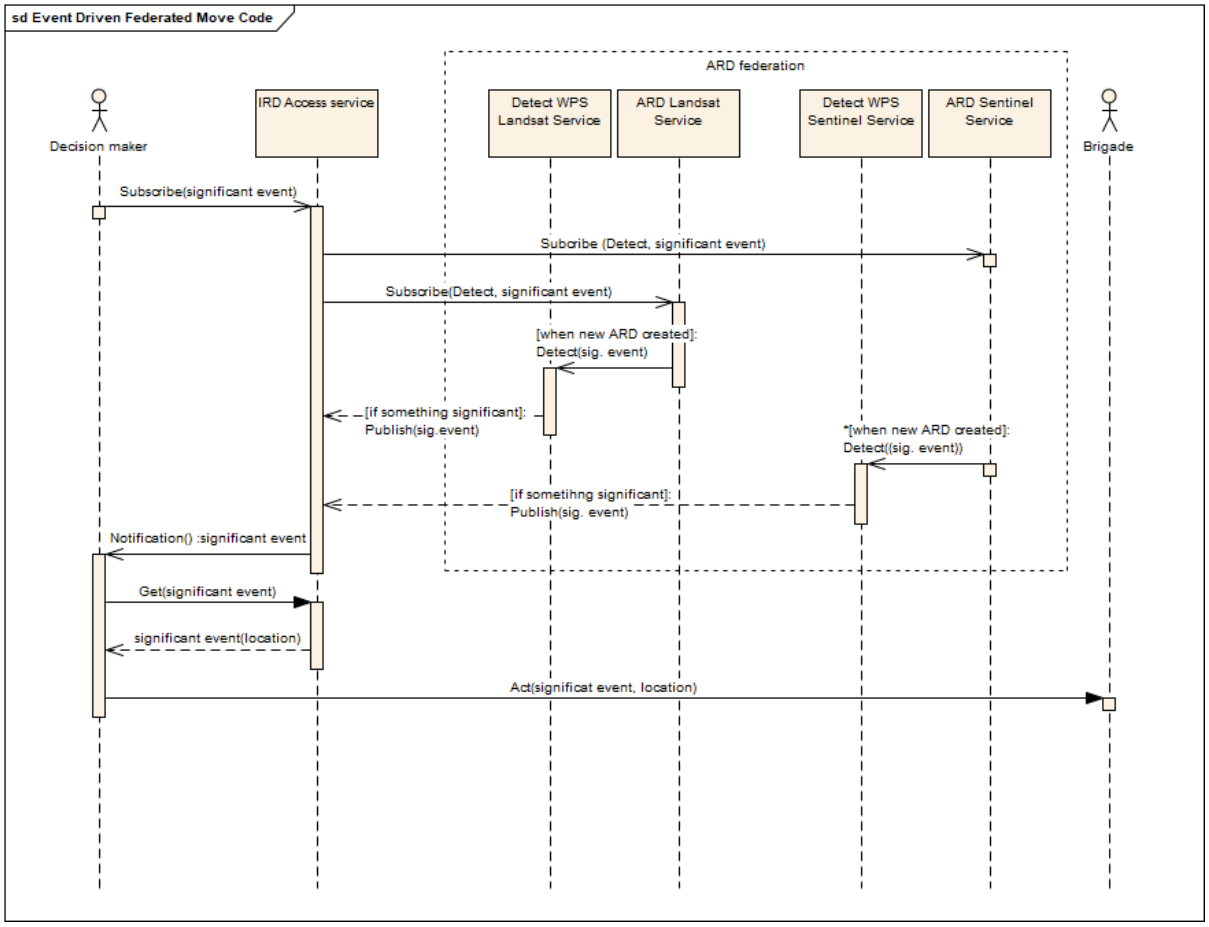

Figure 19. Distributed processing for an event driven use case

There are other use cases where distributing efforts in an optimal way in a federation is not so easy. Imagine a use case where a time series of optical data (as many points as possible) is needed for a phenological study. Phenology is the scientific study of periodic biological phenomena, such as flowering, breeding, and migration, in relation to climatic conditions. The arrival of the spring is perceived as a sudden appearance of leaf. Since most of leaves are green this change can be detected in a time series of optical remote sensing images covering the year. There is sudden increase in the green and the infrared component in one particular day of the year that will depend on weather conditions as well as in the species of plant present. A map that represents the day of the year with the sudden increase in the green and infrared components is called a *greening map*. Again, the USGS (with Landsat data) and Copernicus cloud facilities (with Sentinel-2 data) is available. If the user wants to analyze a single point in space, a simple request for a time series extraction of the point to both facilities is issued and the user gets back two time series: One for each sensor. The user could then combine both series and extract the *start of the greening period*.

Now imagine that the user wants to build a greening map at the same resolution as the original imagery. Each cell of the map will represent a date in the year where the greening starts. If the previous workflow is repeated for each pixel of the intended raster, this will result in millions of atomic requests - one for each point in space. This is going to be very inefficient. In this case, the reasonable thing to do is to request the imagery to both facilities and combine it locally. This minimizes the number of requests but requires transferring the data to the remote processing. Moving the analytics to the data seems impractical for this use case. However, the amount of data generated by the Sentinel-2 A and B tandem is about 3 times bigger than the data produced by Landsat 8. Considering that, a solution that will minimize the amount of data transmitted via the Internet would be to have the processing code in same facility as the Sentinel-2 repository and only

moving (duplicating) Landsat data into the Sentinel repository. This way an enriched Landsat-Sentinel time series can be created without the need to move and duplicate the Sentinel-2 data. In this configuration, the user will generate the greening map at the Sentinel-2 processing facility and send the map back to scientists to compare the current map to previous years and eventually detect and report about interannually changes.

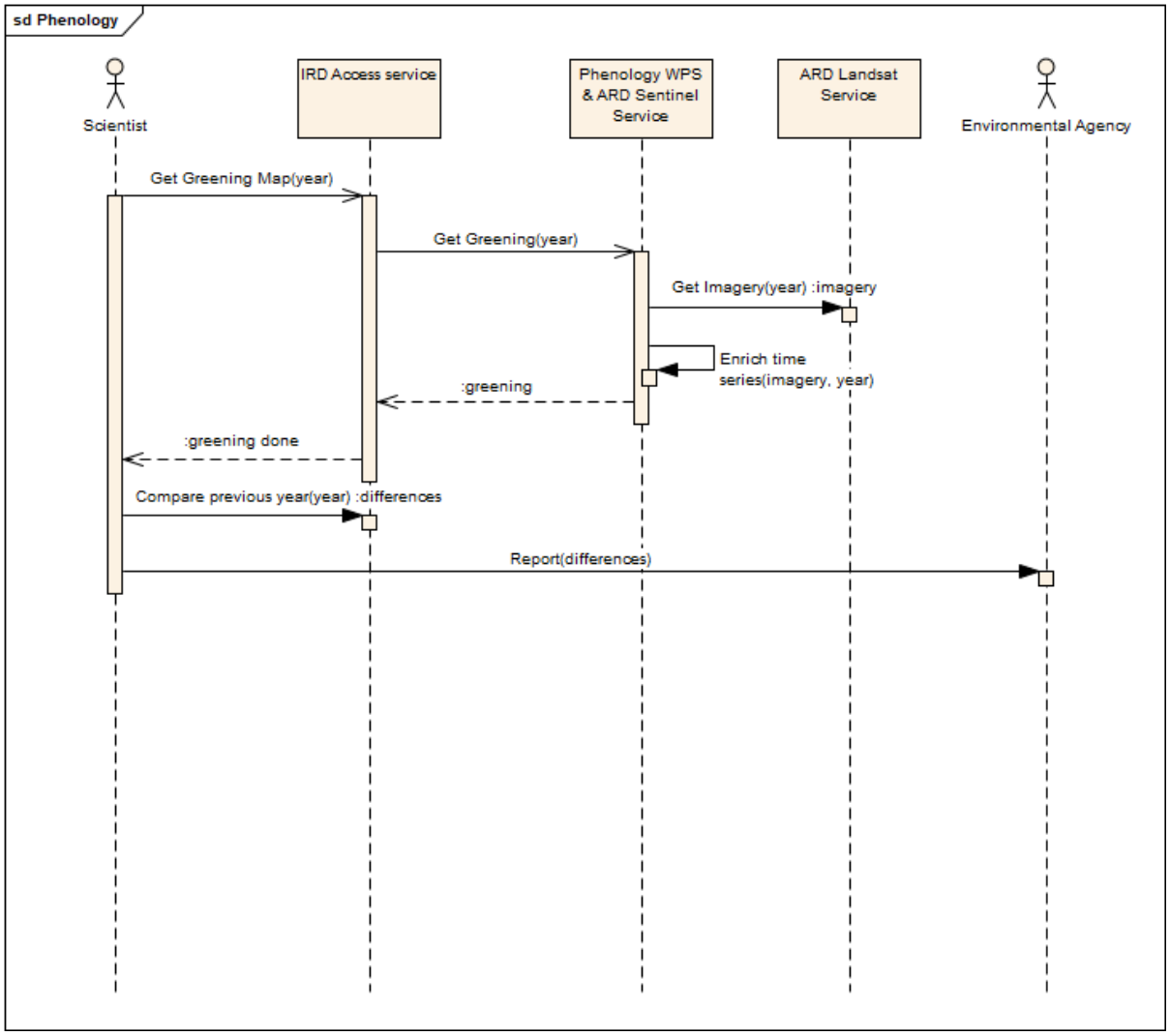

Figure 20. Distributed processing for a phenology use case

10.4. Protect the "Crown Jewels"

As a general principle, once the data was been downloaded, protecting that data by technical means is almost impossible. Only encrypted data that is shown on the screen (or reproduced as multimedia) can be protected. Data for analysis needs to be accessed and, at that instant, can be duplicated. One way to avoid the problem in the first place is to allow running processes at the premises of the provider. In this model, the remote processes access the data and produces the result. If the process serves as a proxy (shares identity and privileges) for the user, then the process will have access only to a subset of data originally agreed to. The process is remotely executed and only the result of the execution is sent back to the user.

Even if viewed as a good solution, in practice executing code in a remote facility generates all sorts of potential security problems for the provider. One can easily imagine a process that copies the data into an encrypted format that is still considered a valid result. Once the result is downloaded

and decrypted in the user's premises, the user gets the original data. To prevent this situation the amount of information extracted can be controlled. However, this limits the size and kind of results that can be extracted. For example, the process could allow for downloading a small file in a well-known vector format (e.g. a common result for an image classification process). The use case described above is common for companies trying to sustain a business model for Very High Resolution (VHR) data exploitation and, at the same time, trying to prevent costumers downloading the original raw data. This is in order to ensure that they are repeat customers.

10.5. PubSub and event driven

One of the characteristics of ARD is the existence of a continuous flow of information about the same observed property in the same area. In these circumstances most of the information becomes redundant. Nevertheless, two interesting patterns could emerge: Anomalies which are significant events that are detected as uncommon values for a particular time, or gradual changes that progressively deviate from an initial measurement due to cumulative effects.

In this section the focus is on the detection of anomalies or significant events that can be extracted by a continuous analysis of new ARD detected by an ARD Access services. Such events would result in the activation of an alert that will result in the deployment of a team to fight the effects of the event in the ground.

A use case that follows the "event driven pattern" is the detection and evolution of forest fires. The US Forest Service could have a subscription to an ARD service providing imagery to detect the presence of fire and track its evolution in near real time. If the fire is big and appends in a remote area, the information extracted from remote sensing could be more comprehensive that the one obtained from a fire brigade deployed on the ground. In some cases, the addition of remote sensing data could be critical for faster, better informed and well-coordinated action.

The diagram describes a generic "emergency management" use case that can be applied also to forest fires. The use case **follows these steps**:

- Steps done **only once** at the beginning
  - An ARD access service is a *PubSub client* that is permanently subscribed to a raw data service.
    - The "raw data service" is an access service as well as a *PubSub server*.
  - A decision maker subscribes to a significant event by adding some widget in his dashboard (called IDR Access service in the diagram).
    - Internally a *PubSub client* integrated in a dashboard requests a subscription to a ARD access service that is a *PubSub server* too.
    - The subscription is done in an uncommon way: The notification should not be sent back to the "ARD access service" but to a "detect WPS service" that is able to process ARD and create new data ONLY if the significant event emerges.
- Things happening **regularly** as a result of the *subscription*
  - As soon as new raw data becomes available, the ARD access system receives a notification and starts the creation of a new time slice that becomes available except if the data is completely useless (e.g. full cloud overcast in the scene).

- As a result of the subscription, the ARD access service starts a "detect WPS process" by sending a "execute" command to the WPS.
- If the process detects a significant event the WPS publishes an IRD alert (the WPS is configured to save the data in the IRD access service database).
  - For example, the format of the saved data could be a feature collection with one feature with the position the significant event and the type and intensity of it.
- The IRD service detects that new data in its database. Since the decision maker is subscribed to the IRD data access service, the IRD access services sends a notification to the user.
  - For example, a red flag in the dashboard and/or and email.
- The decision maker evaluates the situation and, if the event seems worth to investigate further, asks for a briefing session, gives them the position where they should go and the type and intensity of the significant event and the actions to be done when the event is spotted.

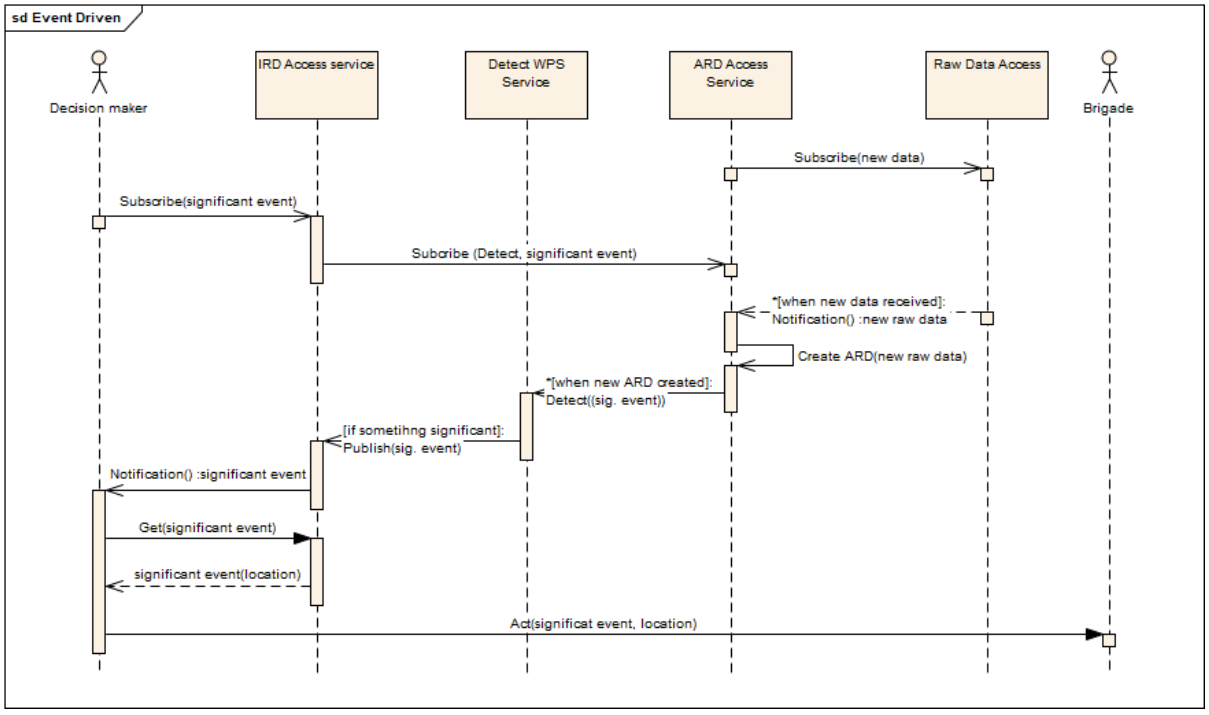

Figure 21. General PubSub and event driven sequence diagram

This general diagram can be applied to two more concrete scenarios that results in variations of the diagram.

Coming back to the use case example, a forest fire has been detected and a US Forest Service official subscribes to an ARD service providing imagery over the area where the fire was detected to evaluate the damage and the perimeters of the fire. Each time new raw data arrives, new ARD will be produced and a notification will be sent to a fire perimeter WPS that will estimate the perimeter of the fire.

If the image is cloud free and it is possible to extract a good perimeter, the new perimeter will be stored as IRD and the US official will get a notification. After analyzing the perimeter, the official can decide the proper course of action. If the fire progresses, the new perimeter is sent to the fire brigade. If the fire can be considered under control, a damage report is produced, and the subscription is cancelled.

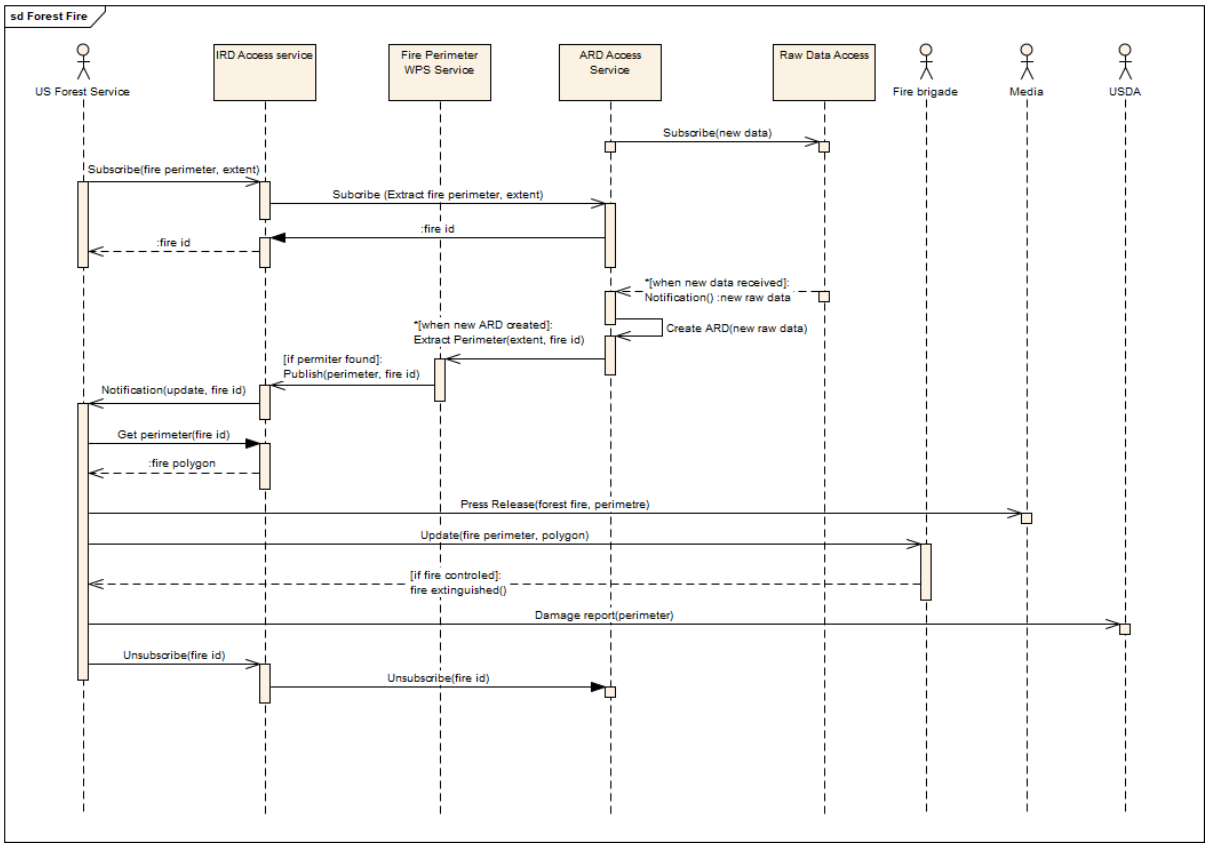

Figure 22. Forest fire use case sequence diagram

The second scenario is related to the detection of illegal logging in the Amazon forest. In this scenario an IBAMA official subscribes to a land cover WPS service. Each time new ARD in the area arrives, a new land cover map is automatically generated. If variations in the forest coverage are detected, a deforestation map is produced and notified sent to the IBAMA official. The official then retrieves the information and analyzes the data. If a deforestation patch is detected and associated with illegal activities, another service for road extraction is triggered using the same ARD or other related very high-resolution data. The extracted roads are conflated with the current road information and the new map sent to the military police with an indication of the position and extend of the illegal logging with the hope that they can avoid future illegal deforestation and arrest the people responsible.

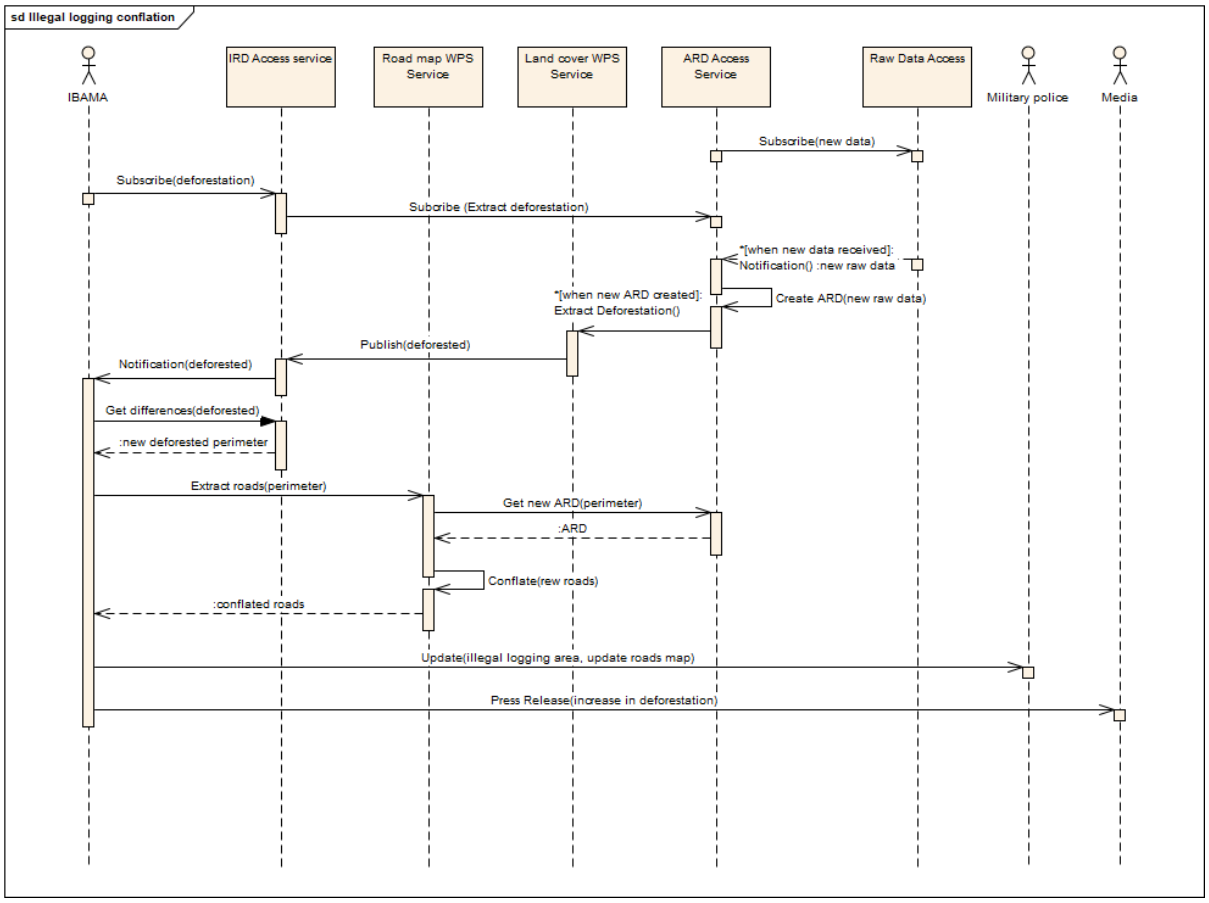

Figure 23. Illegal logging conflation use case sequence diagram

Other similar scenarios based on in-situ ARD (that replaces the satellite-based imagery used in the previous ones) could also be analyzed. One example was extracted from the OGC Critical infrastructure protection experiment (CIPI): "Pollution created by the explosion of a chemical lorry on a river span that contaminates both the water and the air". In this case, weather, air quality and river flow and velocity are the kind of ARD needed for immediate ingestion in a propagation model (WPS) that results in a prediction ready for interpretation and ulterior action.

10.6. The Proposed Solution

In the first subsection of this chapter, how cloud providers are providing access to data resources through an API was explained. Data access through APIs hides the physical storage structure and format allowing provision of resources that are in any place in the cloud or even accessed from a remote cloud. This approach provides a truly federated data access system. In contrast, the Docker approach also described in the first subsection is limited to a deployment in a single node and only provides a truly federated approach if deployed in several nodes and the code deployed has some coordination and parallelization mechanism in it. A truly federated processing capability can only be achieved by a high-level processing language that offers a set of high-level processing functions. A piece of software that knows about the existence of distributed data and processing resources in the federation will be able to interpret the high-level language. The software can then decompose that language into several sets of individual instruction sequences that use the resources in several configurations and select the optimal set of instructions to be executed. Criteria for selecting the "optimal" one could be to get the result faster, to optimize computational costs, to use the most power efficient and environmentally friendly workflow, or the one that minimizes data movement for security considerations.

Once the ideal set of instructions is known, the routine becomes available and published for subscription. The system will be able to invoke it every time new data is being made available as ARD.

Chapter 11. Applying ARD to Machine Learning

Machine Learning (ML) is a subset of Artificial Intelligence (AI). Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. ML is closely related to computational statistics, which focuses on making predictions using computers.

ML has been used in remote sensing for decades (and dates back to 1970s) in order to classify elements in imagery based on textures (object detection) or spectral behavior (Unsupervised image classification or Supervised image classification) [20]. In this case the result of the classification is automatically labelled data (features: geospatial objects with some properties (the annotations)) that is collocated with the original EO data. If the EO data is georeferenced, the labelled data is also georeferenced and becomes a geospatial feature collection.

ML algorithms are very processing intense algorithms that demand powerful processing facilities. In this decade, ML is becoming popular. This is in part due to

- The failure of the more deterministic algorithms to solve some problems (e.g. facial recognition),
- The abundance of processing power that can be bought in the form of cheap computer hardware or can be "rented" from a common cloud provider.

The latter includes to a specific research infrastructure (such as the European Open Science Cloud) or to specialized Earth Observation processing facilities (such as The ESA Thematic Exploitation Platforms (TEP) or the Copernicus Data and Information Access Services (DIAS)).

Many ML algorithms require training. A ML learning algorithm is trained when it is exposed to a training dataset. A training dataset is a dataset composed of a set of data similar to the expected input data and a dataset containing the expected result. For example, in a feature extraction algorithm, a training dataset would be a set of images and the features (geometries and attributes) that the ML algorithm is expected to extract from those images. Once trained the ML algorithm is ready to perform the task of feature extraction and annotation. ML algorithms are very dependent on the kind of inputs they receive and their performance rapidly degrades if the input dataset is different in nature (e.g. images with different contrast or different resolution). In practice, trained ML models are perfect users for ARD since ARD is subjected to a set of requirements specified in the PFS that make it homogeneous. As such, ML models become a new kind of user for ARD.

11.1. Annotating ARD for training

Most ML algorithms need training data to work. The process of exposing a ML algorithm to training data is called *train* or *retrain* ML and this creates a trained model. The model can be later used to annotate EO data automatically. Training datasets are pairs of examples of labelled data (independent variable) and the corresponding EO data (dependent variables). Together, these two variables are used to train an ML model that will be later used to make predictions on the target variable using previously "unseen" EO data. Test data is another set of paired data that is not used to train but to evaluate the performance of the model. In testing, the ML model is exposed to the EO

data and the response of the model is compared with the already known expected result (the labels). Then an overall performance metric is computed. In addition to the training and test data, a third set of observations, called a validation or hold-out set, is sometimes required. The validation set is used to tune variables called hyper parameters, which control how the model learns.

The set of training data, test data, and validation data are similar in structure and are considered to be all *labelled* or *annotated* data. In the geospatial world, where things are georeferenced in space, training data is named as *training areas*. Creating annotated data is a manual process that is time consuming and, paradoxically, precisely what we are trying to avoid by applying ML algorithms. Fortunately, training data is created from a small subset of the original data (as representative as possible of the whole set) and the user can select areas that are well-know or relatively easier to identify. Even for a small subset, the creation of training data prior to training the ML is costly. To avoid the work load of creating the training data , a citizen science project such as picture pile [https://blog.iiasa.ac.at/2016/05/17/picture-pile-gaming-for-science/] can be created. Alternatively reuse of existing annotated datasets may be possible. The Testbed 14 ML ER cites some annotated data sources that can be used to train ML models.

As a consequence, defining a new concept on top of ARD is possible: ML-ARD. ML-ARD is ARD that is paired to labels (feature data) for training purposes. Note that now dataset is composed of the annotation (a feature) and an image (a coverage) and both should be correlated or geolocated. The property type of the annotations in the training areas also determine the results of the classification process. For example, if the annotations contain "land cover types", the result of the ML model is a classification of land cover classes (a land cover map). This means that there are as many ML ARD as classification results. This also means that it should be better to separate ARD from "Training Ready Data" (TRD) that could be defined as training data targeting an ARD PFS and a result type. Since there is one TRD for each result type, the collection of these TRD, can be included in a TRD catalogue.

An example of a TRD small catalogue can be found in the Radiant MLHub [https://www.mlhub.earth/#datasets] with some training data ready for use in any project.

11.2. Analyzing ARD with trained ML models

Once the ML model is trained, the model is ready to be used. If the model was trained with TRD, the model will optimally run on the ARD products that are compatible with the same PFS as the TRD conforms to. Since training a ML model is a time-consuming process, the idea of a ML model catalogue [21] emerges. A ML model catalogue is a catalogue of ML trained models (eventually trained with TRD). The catalogue contains trained processes that are "compatible with" one type of ARD and have one result target.

Now, the question is how to know which process in the catalogue is compatible with one type of ARD and what is the expected result of the model. All processes are depending on the type of data applied in their inputs. Commonly, models are designed to work well at a particular resolution and with a particular type of values. ML models are even more dependent on data inputs but they have one advantage: They were trained with a type of ARD. This gives us an immediate reference to the type of ARD compatible with the ML model.

Assuming that each type of ARD (i.e. each Product Family Specification) has a URI that represents the ARD type, each model input should have the URI associated with it.

Chapter 12. Recommendations

This section includes a discussion on some recommendations extracted from the previous sections. Our intention is not to have a comprehensive list of recommendations but to select some of the most important ones.

12.1. Definitions

The ARD acronym has a clear meaning: *Data that can be immediately used*. In this study we have detected that there are three main ways of considering that the data is *ready*, depending on the kind of user or the expected usage. This has resulted in three main interpretations of the ARD concept:

1. CARD4L ARD (content ARD): Remote sensing product that had been geometrically and radiometrically corrected and has comprehensive metadata at the dataset level and at the pixel level (e.g. quality masks).
2. Homogeneous ARD (technical ARD): Product that follows CARD4L and is regularly distributed in time and space; commonly distributed in regular an adjacent tiles and time intervals that are not dependent on the irregularities of the acquisition process (e.g. satellite orbits).
3. ARD platform (technical ARD): Processing infrastructure that provides immediate access to one of the two previous kinds of ARD (commonly through an API) and that allows running processing code without the need to download the data.

Every actor in the RS market seems to favor the interpretation of the ARD that benefits her. Even if polysemic is a common feature in languages, this will result in confusion to the market. To avoid confusion, we recommend to use the ARD acronym accompanied by the "adjective" that helps pointing to the right meaning. For example: *CARD4L ARD*. This ER is not providing any new definitions to avoid creating more confusion.

12.2. Towards standardization

The CEOS Product Family Specifications are not formal Standards. The ARD Strategy item 4.4 foresees "discussions with standards organizations (e.g. OGC) to explore whether CEOS ARD Specifications might be used as the basis for broader, official community standards, and to ensure that COES work is recognized by others including the data research community". Standards are seen as key factor in CEOS engagement with the private sector.

Standardized ARD could help in avoiding the divergence that can be caused by various groups working towards different interpretations of the concept. Standardization increases the buy-in from the broader stakeholder community, reduces confusion, establishes consistency in terminology, concepts and procedures and formalizes compliance criteria. Standards require volunteer effort to develop and must follow procedures to ensure openness and fairness.

12.2.1. CEOS-OGC continued collaboration

In the opinion of the authors of this document, there is ground to advance in the CEOS and OGC

collaboration in the field of ARD. While the CEOS is capable of mobilizing the satellite ARD producers and the satellite industry and companies, OGC can bring his experience of government, industry and academy in developing standard services to make data discoverable, accessible, interoperable and reusable. CEOS is continuously pushing for ARD that is content ready, while the OGC is continuously working for standards and tools that contribute to make ARD technically ready.

The Earth Observation Exploitation Platform Domain Working Group [https://www.ogc.org/projects/groups/eoexplatform] is the Working Group in the OGC that can act as a point of contract between the content readiness and the technical readiness aspects of the ARD. The OGC can start working in the EO Exploitations Platform DWG by examining the ARD and formulate recommendations on how OGC can help the CEOS process. The group could canalize the technical readiness requirements of ARD to other DWG or SWGs in the OGC if necessary. This is particularly true for groups that can formalize data models (Coverage DWG, O&M SWG and SensorThings API SWG) and architectures (Architecture DWG) for creating and processing analysis ready data on the web (OGC API Processes) or to apply ARD to weather and climate (MetOcean DWG) or other GeoSciences (GeoSicences DWG).

In any case we have to take into account that the process in CEOS is already running. It is important to identify what parts of the CEOS process can have gaps that OGC can cover and be sure that by moving them to the OGC they will run faster or better. We should also consider what elements the OGC community of members can bring to the process. In other words, the most agile and capable party that can better deliver parts of the solution to the community should be identified.

12.2.2. Other Standards organizations

There are some other standard organizations that could collaborate in the adoption of the ARD concept. In the opinion of the authors of this document it really depends on the kind of standards we are trying to formalize.

- The IEEE standards tend to be more sensors and electronics focused. It could be the right place to formalize some of the Product Family Specifications, in particular the ones related with radar and SAR.
- The ISO standards tend to be more conceptual. ISO could help to consolidate the concept among nations. Unfortunately, ISO does not provide a good connection to the opinions from the private sector and rely on the participation of member state representatives so formalization on ISO needs to come after reaching the consensus in the public sector first and OGC and IEEE are better positioned to achieve that.

12.2.3. Other organizations

Due to the interest that the concept of ARD and datacubes have generated in the GEO community, we would like to propose to the GEO to consider the concept of ARD in the context of the informed decisions making process. We also see GEO as an organization that can help to promote the concept beyond satellite data and into in-situ observations.

12.3. An specific examples of collaboration between CEOS and OGC on ARD.

12.3.1. The definition service.

Assign a URI to each type of ARD (i.e. each Product Family Specification). This URI will act as a "variable name" in a dictionary of ARD variables. This concept is equivalent to the ObservedProperty in Sensor Things API (cite:[OGCSTA2015) or the RangeType in the Coverage Implementation Schema. So STA or CIS could use the ARD type URI to verbalize that the content of the service is an ARD type. There are other services that do not have this concept that should have it.

The PFS covers a set of requirements and recommendations that a product should follow to be considered ARD. This product represents a variable or a set of variables measured on the surface of the Earth (that is the current scope of CARD4L; other scopes are emerging). However, there is not a clear definition of the variable in the document apart from the name of the variable in the title of the document itself. In this sense we propose that CEOS includes a concrete definition of the variable or variables represented by a PFS as well as an agreed upon URI to designed it.

The OGC Definitions Server can be used as a repository of this ARD variable names and URIs.

12.3.2. Considering the ARD semantics in processing services.

All services in the OGC that provide data discovery, data access or data processing should be able to have URI to indicate that data provided or the data required is of ARD type. This engineering report mentioned before that STA and CIS already have this capability.

One of the services that does not provide this capability is the WPS (or the equivalent OGC API Process standard candidate). After realizing that, the testbed participants decided to fill a Change Request (CR) during the Public comment period for the OGC API - Processes - Part 1: Core standards candidate. The CR reads like this:

```
ObservedProperty in the description of inputs and outputs. It is becoming more and
more important that processes and data sources are used in a compatible way. There are
GIS processes that are very generic (such as a buffer) but there are others (some
times called models) that are specific and require a specific type of variable (e.g. a
temperature). The STA standard provides a way to describe the "observed property" that
allows you to describe the "variable" that you are exposing (e.g. temperature,
elevation, NO2 concentration etc); see (http://docs.opengeospatial.org/is/15-078r6/15-
078r6.html#figure_2). It has a name, a URI and a description. I would like to request
that a process input and a process output can include the same element (the most
important property is the URI) as an optional property. This way, coupling process
inputs and STA datastreams will be trivial (just look for the data in the web and
inputs that both have the same variable URI (URI of the observedProperty). OGC API
coverages uses CIS and they also have a RangeType that allows for specifying a URI of
the represented variable so it will also work. For the moment OGC API Features does
not define the properties of the features but that can be added later.
```

12.3.3. EOC Integration

Testbed 16 builds on previous OGC Earth Observation Cloud (EOC) efforts through the Earth Application Packages (EOAP) and Data Access and Processing API (DAPA) threads.

The DAPA thread seeks to "develop methods and apparatus that simplify access to, processing of, and exchange of environmental and earth observation data from an end-user perspective." That includes:

- "An abstract description of a resource model that binds a specific function to specific data and also provides a means of expressing valid combinations of data and processes."
- "A survey of which data formats work best in which situations of data retrieval informed by the different scenarios and user groups addressed in this ER. The result is a better understanding of how to encode data for transaction and exchange."

The applicability of Analysis Ready Data to this task is obvious. Further work should explore ARD as a component of the broader EOC architecture.

12.3.4. Analyzing the proposed federated architectures for taking advantage of ARD.

The previous chapter Federated architecture for ARD has proposed a federated ARD architecture. The chapter proposes several recommendations that will not be repeated here and that could be considered the based for future Testbeds and experimentation.

Appendix A: Revision History

Table 1. Revision History

Date	Editor	Release	Primary clauses modified	Descriptions
July 15, 2020	Joan Masó	0.1	all	Initial version
September 15, 2020	Joan Masó	0.9	all	All sections with good content
September 22, 2020	Joan Masó	1.0	all	Added a section with a list of ARD products and providers
October 21, 2020	Carl Reed	1.0	all	Formal phase 1 review
November 18, 2020	Gobe Hobona	1.0	all	Formal review
November 19, 2020	Joan Masó	1.0	all	Accepting reviews and final additions
December 22, 2020	Joan Masó	1.1	all	Preparation for publication

Appendix B: Bibliography

[1] Holmes, C.: Analysis Ready Data Defined, https://medium.com/planet-stories/analysis-ready-data-defined-5694f6f48815, (2018).

[2] Simonis, I.: Standardization Efforts Across Space Agencies: Applications and Analysis Ready Data Discovery in the Cloud. In: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium. pp. 4441–4443. IEEE (2019).

[3] Baumann, P.: From Sensor-Centric to User-Centric-when are data Analysis-Ready? In: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium. pp. 4779–4782. IEEE (2019).

[4] OGC Testbed-16: Call for Participation (CFP), https://portal.ogc.org/files/?artifact_id=91644#PartARD, (2020).

[5] CEOS Analysis Ready Data, http://ceos.org/ard/, (2019).

[6] Soenen, S.: Planet's Framework For Analysis Ready Data, https://www.planet.com/pulse/planets-framework-for-analysis-ready-data/, (2018).

[7] Interoperability, C.E.O.S.- W.G.I.S.S., Group, U.I.: CEOS Interoperability Terminology, http://ceos.org/document_management/Meetings/Plenary/34/Documents/CEOS_Interoperability_Terminology_Report.pdf, (2020).

[8] McCaie, T.: Analysis Ready Data, https://medium.com/informatics-lab/analysis-ready-data-47f7e80cba42, (2020).

[9] An Overview of Copernicus Sentinel-2 Surface Reflectance Products from an Analysis Ready Data Perspective. In: Living Planet Sympossium. Milan 2019. ESA (2019).

[10] Ochiai, O.: JAXA EO Data Strategy, http://ceos.org/document_management/Meetings/Future%20Data%20Architectures%20Big%20Data%20Workshop%20April%202018/FDA_WS_JAXA.PPTX, (2018).

[11] Rosenqvist, A., Tadono, T., Shimada, M., Itoh, T.: Jaxa Global Sar Mosaics–Assessing Compliance with CEOS Analysis Ready Data for Land (CARD4L) Specifications. In: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium. pp. 5545–5548. IEEE (2019).

[12] Generation of Analysis Ready Data for Indian Resourcesat Sensors and its Implementation in Cloud Platform. I.J. Image, Graphics and Signal Processing. 6, 9–17 (2019).

[13] Shang, R., Zhu, Z.: Harmonizing Landsat 8 and Sentinel-2: A time-series-based reflectance adjustment approach. Remote Sensing of Environment. 235, 111439 (2019).

[14] Sentinel-2 and Landsat-8 Analysis Ready Data: Towards a Service Prototype for On-Demand Processing Using the ESA Research and Service Support. In: Proc. of the 2019 conference on Big Data from Space (BiDS'19). pp. 189–192 (2019).

[15] PCI Geomatica Analysis Ready Data (ARD) Tools, https://www.pcigeomatics.com/pdf/geomatica/

techspecs/2018/ARD-Tools.pdf, (2019).

[16] Simonis, I.: OGC Earth Observation Applications Pilot: Summary Engineering Report (OGC 20-073), https://docs.ogc.org/per/20-073.html, (2020).

[17] Strobl, P., Baumann, P., Lewis, A., Szantoi, Z., Killough, B., Purss, M., Craglia, M., Nativi, S., Held, A., Dhu, T.: The six faces of the data cube. In: Proc. Conf. on Big Data from Space (BiDS'17). pp. 28–30 (2017).

[18] Frantz, D.: FORCE - Landsat+ Sentinel-2 analysis ready data and beyond. Remote Sensing. 11, 1124 (2019).

[19] Rodriguez, H.: OGC Testbed-15: Federated Clouds Security Engineering Report. Open Geospatial Consortium, http://docs.opengeospatial.org/per/OGC 19-024r1.html (2019).

[20] What is Image Classification in Remote Sensing, https://gisgeography.com/image-classification-techniques-remote-sensing/, (2020).

[21] Landry, T.: OGC Testbed-14: Machine Learning Engineering Report. Open Geospatial Consortium, http://docs.opengeospatial.org/per/18-038r2.html (2018).