OGC® DOCUMENT: 23-043
External identifier of this OGC® document: http://www.opengis.net/doc/PER/T19-D051

Open Geospatial Consortium

OGC TESTBED 19 ANALYSIS READY DATA ENGINEERING REPORT

ENGINEERING REPORT
Implementation

PUBLISHED

Submission Date: 2024-02-20
Approval Date: 2024-03-28
Publication Date: 2024-07-05
Editor: Liping Di, David J. Meyer, Eugene Yu

Note

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.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

I EXECUTIVE SUMMARY

Implementations of the Analysis Ready Data (ARD) concept are consistent with the FAIR principles of finding, accessing, interoperating, and reusing physical, social, and applied science data with ease. The goal of this Testbed 19 OGC Engineering Report (ER) is to advance the provision of geospatial information by creating, developing, identifying, and implementing ARD definitions and capabilities. Specifically, this ER aims to increase the ease of use of ARD through improved backend standardization and varied application scenarios. Additionally, this work seeks to inform ARD implementers and users about standards and workflows to enhance the capabilities and operations of ARD. Ultimately, the goal of the work described in this ER is to maximize ARD capabilities and operations and contribute to the enhancement of geospatial information provision.

Four distinct scenarios – gentrification, synthetic data, coverage analysis, and coastal studies – are explored to reveal both the strengths and limitations of the current ARD framework. The gentrification scenario, which utilizes existing Committee on Earth Observation Satellites (CEOS) ARD data, highlights the need to expand ARD's scope beyond Earth Observation (EO) data. The integration of diverse data types, such as building footprints and socio-economic statistics, is crucial for comprehensive analysis. The synthetic data scenario explores the potential of simulated EO imagery to enhance data availability and diversity for machine learning applications. However, challenges in standardization and quality assessment require further investigation. The analysis of coverages for ARD reveals the importance of clear pixel interpretation ("pixel-is-point" vs. "pixel-is-area") and standardized units of measure for seamless integration and analysis. Additionally, enriching the metadata structure with defined extensions is crucial for efficient data discovery and understanding. The coastal study scenario, where in-situ data needs to be elevated to ARD, emphasizes the need for flexible levels of readiness. Different analytical tasks may require distinct data properties, necessitating adaptable standards that cater to temporal emphasis, spatial alignment, and non-GIS applications like machine learning.

This work identified several key areas for improvement:

- encompassing non-EO data such as building footprints, socio-economic statistics, synthetic data, and in-situ measurements;
- establishing guidelines and quality controls for incorporating diverse data types;
- tailoring data specifications to accommodate different analytical needs, including temporal emphasis and non-GIS applications; and
- implementing structured metadata with defined extensions for enhanced data discovery, understanding, and provenance tracking.

In addition to the above recommendations, the interoperability and support of ARD in wider communities warrants further exploration and implementation. Additionally, areas such as uniform evaluation and compliance certification could be further investigated to ensure consistency in data readiness across various hierarchies and application domains.

II KEYWORDS

The following are keywords to be used by search engines and document catalogues.

testbed, web service, analysis ready data, remote sensing, earth observation

III CONTRIBUTORS

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

Table — List of Contributors

NAME	ORGANIZATION	ROLE
Liping Di	George Mason University	Editor
David Meyer	NASA Goddard Earth Sciences Data and Information Services Center (GES DISC)	Editor
Eugene Yu	George Mason University	Editor
Josh Lieberman	OGC	Task Architect
Peter Baumann	Rasdaman	Contributor
Dimitar Mishev	Rasdaman	Contributor
Chris Andrews	Rendered.AI	Contributor
Matt Robinson	Rendered.AI	Contributor
Daniel Hedges	Rendered.AI	Contributor
Li Lin	George Mason University	Contributor
Glenn Laughlin	Pelagis Data Solutions	Contributor
Carl Reed	Carl Reed and Associates	Content Reviewer
Jim Antonisse	WiSC Enterprises	Contributor

1

INTRODUCTION

1 INTRODUCTION

In this Engineering Report (ER), Analysis Ready Data (ARD) refers to time-series stacks of overhead imagery that are prepared for a user to analyze without having to pre-process the imagery themselves [71][88][103][110]. The idea behind ARD is that providers of satellite imagery are in a better position to undertake these routine steps than the average user [71]. Analysis-ready datasets have been responsibly collected and reviewed so that analysis of the data yields clear, consistent, and error-free results to the greatest extent possible [71].

ARD is important because it saves time and resources by providing users with data that has already been preprocessed and rigorously validated and is ready for analysis. ARD also ensures that users have access to high-quality data that has been reviewed for accuracy and consistency [71]. ARD can be used in various applications such as land cover mapping, change detection, and environmental monitoring. The concept and implementation of analysis readiness can significantly address both climate and disaster resilience needs for information agility by improving access to interdisciplinary sciences such as natural, social, and applied sciences as well as engineering (civil, mechanical, etc.) public health, public administration, and other domains of analysis and application.

The CEOS Analysis Ready Data (ARD) strategy aims to simplify data handling by removing many of the fundamental data correction and processing tasks from users so that more users and more uses of the data are possible [113]. CEOS ARD involves 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 [1].

The Testbed 19 ARD ER reviews all existing standards and previous ARD work, including CEOS ARD efforts [116][7][10][12][14][16][18][20][22][29][103][110][4] and previous OGC efforts [24][42][56]. The ER will define foundational elements that allow for the mixing and matching of different standards and target its mission of implementing Findable, Accessible, Interoperable, and Reusable (FAIR) principles for scalable and repeatable use of data [31]. ARD is a key example of the capability to enable the FAIR principles of finding, accessing, interoperating, and reusing physical, social, and applied science data easily.

The Testbed 19 activities included:

- defining ARD scenarios;
- refining and implementing the ARD scenarios;
- coordinating with the ARD SWG on ARD standard development;
- demonstrating the ARD scenarios;
- describing the ARD scenarios in detail (including all aspects that are relevant for current and future standardization); and
- documenting the open delivery of the demonstration components in a software container for future use.

The ARD ER clearly describes and reports scope, objectives, methodology, and expected outcomes of the Testbed 19 ARD work. The ER describes ARD requirements, identifies initial use case objective(s), and the components and elements needed to achieve these objectives. Further, the ER describes how the Testbed participants achieve the objectives, and, if applicable, identifies technology gaps or elements for future work. Finally, this ER summarizes how this work is scalable within other domains or to be applied more broadly within the same domain.

2

ANALYSIS READY DATA

2 ANALYSIS READY DATA

This section provides the basic concepts of and requirements for analysis ready data (ARD).

2.1. Definition

CEOS defines ARD as “satellite data that have been processed to a minimum set of requirements and organized into a form that supports immediate analysis with a minimum of additional user effort and interoperability both through time and with other datasets.” This definition needs to be expanded to cover non Earth Observation data, such as model outputs, in-situ measurements, demographic data, and economic data which may need geocoding to register them geospatially. Broadening the definition to encompass all geospatial data, ARD is then defined as geospatial data that have been processed to a minimum set of requirements and organized into a form that enables immediate analysis with a minimum of additional user effort and interoperability both through time and with other datasets. The ultimate direction is working towards the FAIR principles[31] of finding, accessing, interoperating, and reusing physical, social, and applied science data easily.

2.2. Fundamental Requirements

The fundamental requirements for making a dataset analysis ready are as follows.

1. Ensure Data Quality: The quality of data is critical in preparing the data for analysis. Data need to be accurate, consistent, complete, and free of errors. Thus, ensure that all the datasets being used are of high quality.
2. Data Cleaning: The next step is to clean the data set. Data cleaning involves removing duplicates, filling in missing values, and removing any irrelevant variables from the dataset.
3. Standardize Data Formats: Data come in different formats and types. Differences in coding or labeling of datasets may become a major problem during analysis. Thus, standardize the format of the data to make the data analysis-ready.
4. Data Integration: Often, data for analysis come from multiple sources. In such a situation, different datasets might have varied column names, restrictions, clarifications, or even misalignments. So, it becomes essential to integrate the data sets into a single conflated and comprehensible dataset.

5. Variable Identification: Knowing what each variable of a dataset represents is important. Proper documentation of each variable makes it easier to understand the dataset and improves the quality of the analysis.

6. Data Segmentation: In addition to integrating data sets, segmenting or partitioning the results as per a certain criteria or logic may be required if time-series analyses or testing hypotheses are carried out.

7. Ensure Data Security and Privacy: Protecting trustworthy data ensures the access and integrity of valuable datasets for analysis. While the specific requirements may vary depending on the data characteristics and intended use, upholding a high degree of data security and privacy remains paramount.

8. Data Storage: The final step is to store the data in a secure, but accessible manner. Best practices recommend storing data at a secure location, where the data are accessible to authorized users, with proper backup and disaster recovery provisions in place.

2.3. Product Families

2.3.1. Earth Observations (Satellite remote sensing)

The major components of satellite remote sensing ARD typically include the following properties[1][110]:

- radiometric and geometric correction;
- mosaicing and tiling;
- cloud and shadow masking;
- atmospheric correction; and
- metadata and catalog.

In Testbed 19, several ARD products have been found and originally used for different analysis. These products were combined into a comprehensive dataset for focused analysis. These include Landsat data products for gentrification scenario and essential earth observations for marine and coastal study scenario.

2.3.2. Model Outputs

Model outputs carry different properties and characteristics for preparing and publication as ARDs. One example is synthetic data which have a clear underlying physical model but simulate

sensor observation under different conditions. The serving and preparing of synthetic results as ARDs is one of the scenarios studied in Testbed 19.

2.3.3. Other Geospatial Data

There still exist data that do not quite fall into the existing CEOS ARD families. For example, in-situ observations may need to be pre-processed through a series of algorithms to provide the observations as ARD ready for integration and interoperation with other data sources and analytical systems. The in-situ data were studied in the coastal stud scenario.

Demographic data and building information are other examples of non-EO data that need to be prepared and made analysis ready. Processing may involve geocoding and other specific pre-processing.

Besides “strict” geospatial data, there may be data that have certain geospatial properties, but the focus is rather aligning the data ready for analysis through AI/ML (artificial intelligence and machine learning). For example, training datasets may include broad ranges of labeled data for machine learning. The inclusion of such datasets in ARDs is also discussed in this ER.

3

SCENARIO ON GENTRIFICATION STUDY

3 SCENARIO ON GENTRIFICATION STUDY

NOTE: This scenario was led and implemented by the Center for Spatial Information Science and Systems (CSISS), an interdisciplinary research center chartered by the provost and affiliated with the College of Science at George Mason University, Fairfax VA, 22030, U.S.A.

3.1. Introduction

Due to advancements in technology and acquisition strategies, institutions, such as the USGS and European Commission, can release vast amounts of remote sensing images under free licenses. Simultaneously, storing and computing power has developed to accommodate this increase in data. Yet, the volume of data still poses a great challenge to data analysts, scientists, and non-experts alike. Data analysts report spending 80% of their time cleaning data to ensure interoperability for time series[25]. Furthermore, much of the storage and computing power required to store and process these data are inaccessible to non-experts. Many scientists believe that the solution to both problems is a concept termed “analysis ready data” (ARD). The committee of Earth Observation satellites defines ARD as satellite data that have been processed to a minimum set of requirements and organized into a form that supports immediate analysis with a minimum of additional user effort and interoperability both through time and with other datasets [43]. In other words, an ARD product undergoes common data processing before distribution. These preparations are time consuming, computationally taxing, and require expertise to perform. Currently, the closest thing to a standard for ARD is CEOS Analysis Ready Data for Land (CARD4L), which outlines the threshold and target quality of data for it to be considered ARD [57].

The first set of data to meet the CARD4L requirements was the USGS’s Landsat collection 2 surface reflectance and surface temperature products [43]. Collection 2 has a great depth of remote sensing imaging as the USGS has also reprocessed its images from collection 1 from Landsat 1-8 [72] which enables users to use ARD for a time series that spans back to the first Landsat mission. Collection 2 is organized into 3 levels [72]. Level 1 is geometrically and radiometrically corrected data [72]. Within level 1 a tier system is used to distinguish between the quality of data based on the radical root mean square error (RMSE) [72]. Tier 1 is the best contains data with a RMSE lower than 12, all data with an RMSE greater than 12 are grouped into tier 2 [72]. Near real time data are data that have yet to be categorized into either tier 1 or 2 and are available for rapid download [72]. Moving onto Level 2 data which are certified by CEOS as ARD, Level 2 products are only derived from level 1 data and use top of atmosphere corrections to provide surface reflectance and surface temperature [72]. Furthermore, all ARD products include a quality assurance band [72]. As of January 28th, 2022 in addition to Landsat, Sentinel 2’s Level-2A product has been certified for meeting the threshold of ARD by CEOS [89]. The Sentinel 2 product is broken into several levels. Level-0 is the raw compressed image from the satellite that is downlinked [104]. Level-1A involves creating a geometric model to locate pixels in the image and uncompressing the images [104]. Level-1B involves performing radiometric calibrations and geometric refining to the geometric model produced in Level-1A [104]. Then Level-1C involves Ortho-image generation, top of atmosphere (TOA) is computed,

and clouds are calculated [104]. Finally, Level-2A provides surface reflectance products based on the TOA product as well as scene classification for clouds [104].

In short, the goal of ARD is to standardize and centralize data to make it more accessible in a way that removes friction for users working with remote sensing data 3. CEOS created a guideline for what constitutes ARD data in the CEOS ARD for land (CARD4L) product, providing fundamental support to ARD standards [111]. Likewise, The Open Geospatial Consortium (OGC) Testbed-16 worked to solidify what ARD is by creating a list of characteristics for data to be considered ARD, including, but not limited to, homogeneous organization, georeferencing, units, and metadata detailing changes made [111]. This scenario is implemented to show how ARD helps improve the ease of use and accessibility of data through a case study. The scenario aims to provide insights and recommendations to the OGC Standards Working Group (SWG) responsible for moving the ARD standards set forth by the International Organization for Standardization and Open Geospatial Consortium.

3.2. Scenario Methodology

3.2.1. Datasets

The Earth Explorer data portal provides access to the USGS's Landsat collection 2 data. Collection 2 consists of three levels. Level 1 consists of data that have been geometrically and radiometrically corrected. Level 1 uses an internal tier system to organize data based on pixel quality and processing level [114]. Real time data is where imagery goes before being moved into either tier 1 or 2, which is determined by the image radical root mean square error (RMSE) [72]. A score of 12 or better is categorized as tier 1. Level 2 data are ARD certified by CEOS and are derived from tier 1 data only [117], [2]. Furthermore, there are level 2 science products that have a long enough time series and consistency that the products can be used to track climate change [5]. Lastly, level 3 data are data derived from level 2 science products. Using Earth Explorer, ARD surface reflectance and associated quality assurance products were downloaded of the tile horizontal 27 vertical 9 spanning from 2013 to 2019, which encompassed the city of interest, Washington DC.

While gentrification bears resemblances to community redevelopment, it usually progresses more rapidly and is frequently propelled by significant financial investments. Moreover, gentrification often brings about notable shifts in micro-level socioeconomic dynamics. To track gentrification, the Testbed participants focused on the construction of new buildings in DC. While ARD is available up until early 2023, the most recent building footprint of DC was from 2019. Like the DC building footprint, the building permit data were accessed from Open Data DC. The building permit data were organized by year, which was then filtered by the type 'construction' and subtype 'new building'. Since construction takes several years, analysis was not only for the year that the permits were approved but also the years before and after. Working within these constraints, the earliest building permit used was 2012, and the latest was in 2018.

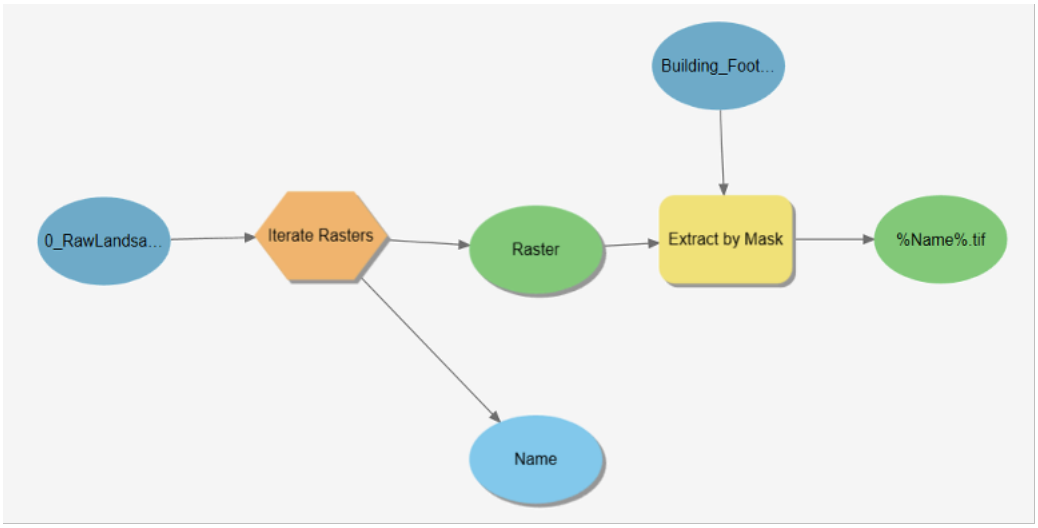

Figure 1 — Example of model builders used to create training data

Because the Landsat imagery is ARD, there was no need to perform any top of atmosphere corrections, so the process of turning the Landsat imagery into training datasets could begin immediately. First, all data were added to an ArcGIS map. As this scenario is exploring the capabilities of ARD, using a common and easily accessible tool like ArcGIS meant that the methodology could be easily replicated, and conclusions of ARD's limitations and abilities would be applicable to the largest number of people. Furthermore, as seen in Figure Figure 1, ArcGIS's built-in model builder expedited the process of creating training data by taking full advantage of the standardization of the ARD. After adding all the downloaded information, the building footprint was dissolved by shape area to remove shared boundaries. The dissolved building footprint was then used as the feature mask in the extract by mask geoprocessing tool to extract only pixels of DC from the Landsat imagery.

Then, an ArcGIS geoprocessing tool was used to remove all pixels that were not marked as clear pixels. All the data were reorganized into different folders by year, which were then grouped together by intervals of three. As a result, there would be a folder 2013-2015, with subfolders 2013, 2014, 2015 and so on for folders 2014-2016, 2015-2017, 2016-2018, and 2017-2019. Next, true values were created, by separating pixels known to be new buildings which was done by extracting by mask, with the building permit for the respective year acting as the mask. As mentioned before, for each permit the participants wanted to look at the year prior and after as well. For example the 2013-2015 images used the 2014 building permit data to extract the pixels that had new buildings. Pixels that were designated as not having new buildings were stored separately in another folder. In short, this resulted in three groups of data: Images with just the new building pixels, images without the new building pixels, and images of both types of pixels. The last set would be used as the unsupervised training data. Each group and set of years would be organized into separate mosaic datasets. In other words, there would be three 2013-2015 mosaic datasets, one for each group, and so on.

It was difficult to turn the mosaic dataset into a multidimensional dataset because ArcGIS was unable to read the ARD's metadata. So, information such as the product name and acquisition date were missing in the ARD. However, the ARD file names are formatted to be human-readable [104], and include the acquisition date as a part of the file name. Therefore, creating a new field in the mosaic dataset and splicing the file name, with the calculate field geoprocessing

tool, created the necessary temporal component to create a multidimensional raster, with the build multidimensional info geoprocessing tool. Finally, the Space time Cubes were created using the Create Space Time Cube from the Multidimensional Raster spatial analysis tool. One advantage of using this tool was the fill empty bins parameter, which used an interpolated univariate spline algorithm to create a temporal trend to fill the empty bins.

Using the datasets created from the ARD, other built-in machine learning tools within ArcGIS were used and tested. With the release of ArcGIS 3.1, 'Train Using AutoML' and 'Predict Using AutoML' became available. Both tools enable the user to streamline the machine learning process by determining the best models, hyperparameters, and creating the optimal ensemble of models for the validation set[Esri_automl]. So, the 'Train Using AutoML' tool was used for building a machine learning model optimized for the dataset given, and the 'Predict Using AutoML' was used to predict where gentrified pixels would be in the future.

3.2.2. Supervised Machine Learning

A feature class was needed to use the "Train using AutoML" tool. Therefore, the rasters had to be turned into polygons. Since the goal is for the model to predict future gentrified pixels, the dataset of rasters that contained only gentrified pixels was used. This dataset had previously been used as the input for the multidimensional rasters. Each raster was converted to an individual polygon feature using the 'raster to polygon' tool. The polygon feature stored each gentrified pixel as its row, so each pixel was assigned a date and satellite based on the raster it came from. Then, each polygon feature was added to a larger feature class grouped into the years 2013-2015, 2014-2016, 2015-2017, 2016-2018, and 2017-2019 like how the space time cubes were organized. All these steps were automated using the model builder feature of ArcGIS, which was made possible due to the ARD's standardization of file names and pre-processing. Finally, the feature classes could then be used as input into the 'Train Using AutoML' tool, with the grid code being the dependent value the model was trying to predict, and the date and satellites being the independent variables being used to explain that change over time. The models created could then be used by 'Predict Using AutoML'.

3.3. Results and Discussions

In this study, the strengths and limitations were evaluated for using Landsat ARD on the workflow of creating training data and used the training data to evaluate the AutoML tools and time series clustering tool built into ArcGIS. Overall, the evaluation of the various pre-defined machine learning tools in ArcGIS shows that standard machine learning algorithms are ideal in all use cases, such as monitoring gentrification. Moving on to look at how ARD affects workflow, the Landsat ARD was easily accessible due to the USGS Earth Explorer website and ARD tile tiling. Earth Explorer maintains accessibility by having free, indexable, and easily downloadable data. Furthermore, the radiometric and geometric correction standardization allows for immediate interoperability and comparison in time series. The benefits of this are threefold: firstly, increased accessibility to remote sensing data because the user doesn't need to be familiar with applying algorithms that perform radiometric and geometric corrections to use the images. Secondly, this reduces both the labor hours and computing power that would've

been needed to apply said corrections. Lastly, ARD maximizes accessibility by being processed enough that it can be used immediately upon download, but not so specifically that it can only be used in a narrow domain. This is exemplified by the creation of the ESRI's Space Time Cubes, which are similar to Data Cubes. Immediately after adding the data to ArcGIS, the images could be altered to fit the needs of the use case. Furthermore, the built-in quality assurance (QA) band that comes with each raster removed poor quality pixels from the data set.

However, there were some minor limitations of working with ARD. Firstly, the QA bands for the different satellites used different encodings to represent different pixel quality, meaning that two model builders had to be used to process the data. Another area for improvement was metadata incompatibility with ArcGIS, which supports the specific Landsat satellites but not the ARD. It is hoped that ESRI will address this in the near future. Lastly, it is recommended that users should be able to access ARD through a Data Cube-like format, especially in the described scenario, which prepares ARD to machine learning-ready training datasets. Currently, thanks to the standard human-readable file and raster names, acquisition date, band information, and satellite used could easily be added to datasets that had trouble reading the ARD metadata, but injecting ARD into Data Cube will provide a much smoother user experience. This recommendation was also mentioned in the Climate Resilience Pilot Engineer Report. Analysis Ready Data Cubes (ARDC) can be important in processing big data which will make ARD more accessible and easier to use [8]. Analysis Ready Data Cubes (ARDC) play a crucial role in efficiently processing large datasets, making ARD more accessible and user-friendly. The development of ARDC will not only enhance the practical application of ARD for Earth Observation (EO) data but will also facilitate the integration of non-EO datasets, fostering the development of applications that address real-world problems by seamlessly combining EO and non-EO data while adhering to ARD standards.

In conclusion, notwithstanding ARD's minor limitations, ARD significantly optimized the workflow process of turning the downloaded data into organized machine learning training datasets due to its accessibility and immediate interoperability.

4

SCENARIO ON ISO/OGC COVERAGE AND DATACUBE STANDARDS

4 SCENARIO ON ISO/OGC COVERAGE AND DATACUBE STANDARDS

NOTE: The scenario on the topic of reviewing ISO/OGC coverage and datacube standards for Analysis Ready Data was led and implemented by rasdaman.

4.1. Introduction

This analysis investigates how analysis-ready the OGC Coverage Implementation Schema (CIS) is[26][44][58]. Coverages are the accepted paradigm for modeling fields (in the sense of physics) across standard bodies with a geospatial focus [107]. Technically speaking, coverages encompass regular and irregular grids, point clouds, and general meshes. The gridded data, specifically, resemble datacubes, which is the particular focus of this Engineering Report.

One use case to investigate (with the help of GeoDataCube(GDC) that was set up in parallel to this scenario activity) is how far the ISO/OGC coverage standards carry in supporting the analysis readiness of geospatial data, in particular: CIS 1.1[75] and the OGC Web Coverage Processing Service (WCPS) 1.1[92] Standards.

4.2. Coverages — A Data Structure for ARD in Earth Observation

This section summarizes the key findings and recommendations regarding the use of coverages as a data structure for Analysis Ready Data (ARD) in Earth Observation (EO). For detailed information, please refer to Annex B.

4.2.1. Standards and Structure

- **Standards Alignment**: Coverages adhere to OGC Standards such as CIS 1.1 and WCPS, aligning with ISO 19123-2 and 19123-3 for consistency and interoperability.
- **General Grid Coverage**: This core structure defines the spatial and data aspects of ARD, consisting of the following.
  - **Domain set**: Geospatial reference system
  - **Range set**: Data values and their types

- **Range type**: Data format (e.g., numerical, categorical)
- **Metadata**: Additional information about the data

4.2.2. WCPS — A Datacube Language for ARD Processing

- **Datacube Model**: WCPS provides a framework for organizing and manipulating large EO datasets.
- **Common Operations**: WCPS offers a set of functions for processing and analyzing coverages.
- **User-Friendly Syntax**: Similar to FLOWR, WCPS enables users to express processing tasks intuitively.

4.2.3. Obstacles and Recommendations for Coverages as ARD

This study analyzed challenges in using coverages for ARD and proposes solutions.

1 Data Modeling

- **Pixel-in-X Misconception**: Clarify that pixels are associated with specific coordinates, not cells. Invest in educational resources to address this confusion.
- **Pixel Interpretation**: Standardize on "pixel-is-area" for consistency and avoid half-pixel shifts.
- **Units of Measure**: Adopt QUDT for its machine-readable format and conversion capabilities.
- **Tiling Transparency**: Make tiling an internal detail of ARD, transparent to users.
- **Structured Metadata**: Organize and structure metadata for improved access and comprehension. Consider a registry of defined extensions for efficient information extraction.

2 Data Processing

- **Context-Aware Interpolation**: Utilize appropriate interpolation methods such as kriging based on data type and context.
- **Compatible Image Pyramids**: Allow only compatible interpolation methods during retrieval and processing to avoid inconsistencies.
- **Data Summarization**: Clearly document appropriate aggregation methods for different data types (e.g., counts vs. averages) to prevent misinterpretations.

- **Dimension Hierarchies**: Capture and document hierarchical structures (e.g., time series) for efficient analysis and exploration.
- **Validity and Reliability Masks**: Implement masks to identify and filter out areas of uncertainty, improving data reliability.
- **Product Provisioning Coherence**: Track data processing history and ensure consistency across versions and providers to maintain data quality.
- **Numerical Effects Awareness**: Understand the inherent inaccuracies of floating-point numbers to avoid calculation errors.

4.2.4. Practical Examples in Context

This study examined how coverages can be applied to real-world scenarios.

- **Service Quality Parameters**: Define and communicate key parameters such as accuracy, resolution, and uncertainty to users for informed decision-making.
- **Coverage Fusion**: Leverage advanced techniques and cloud computing to combine data from diverse sources despite format variations, quality differences, and spatial/temporal overlaps.
- **Machine Learning Integration**: Utilize ML models for automated ARD processing and analysis, ensuring data quality and model suitability for specific tasks.

4.3. Recommendations for Standard Development

Based on the insights gained, the following are recommendations for improving ARD standards.

- **Refine Existing Standards**: Update OGC standards to address identified challenges and reflect current needs in EO.
- **Enhance Metadata Structures**: Design and implement standardized metadata structures to accommodate diverse use cases and scenarios.
- **Universal Units of Measure**: Promote QUDT adoption for consistent and interoperable data exchange.
- **User-Friendly APIs**: Focus on clear data access and processing functionalities in APIs, hiding technical details.
- **Interval Arithmetic Adoption**: Utilize interval arithmetic to quantify uncertainties and provide more reliable calculation results.

- **Fitness Negotiation and SLAs**: Develop mechanisms for users to specify quality requirements and services for guaranteed data suitability.
- **Model Applicability Parameters**: Define clear parameters for machine learning models to ensure appropriate use and reliable results.

4.4. Conclusion

By addressing the challenges and implementing the proposed recommendations, coverages can become a powerful and versatile data structure for ARD in Earth observation which will enable efficient and accurate analysis for diverse applications, advancing scientific research and decision-making in EO.

5

SYNTHETIC DATA SCENARIO

5 SYNTHETIC DATA SCENARIO

NOTE: This scenario was led and implemented by Rendered.AI.

5.1. Introduction

One of the fundamental reasons to define Analysis Ready Data standards is that it is common for real world datasets to have insufficient metadata or structure for common tasks and analyses. Synthetic data generation, the process of creating datasets that simulate real data according to predefined specifications, offers an opportunity to advance the concept and application of ARD by the following.

1. Providing benchmark or referenceable examples of how an ideal dataset would be composed including content, metadata, and structure.
2. Supporting specific use cases or examples of commonly used datasets, such as Earth observations satellite content, that can be used to test data processing tools and pipelines.
3. Providing experimental input for both training and validating algorithms used to process real sensor data.

This Testbed 19 ARD project provided a demonstration of a synthetic data generation pipeline that produces diverse datasets in an ARD-compliant format, specifically the CEOS ARD for Land – Surface Reflectance (CARD4L-SR) Standard. The goal of this process is to better understand how synthetic data can provide value to the creators and users of ARD, and how the framework for ARD can likewise benefit the creators and users of synthetic data. In the process, the implications of synthetic data generation within this ARD framework are explored as well as what elements of an ARD specification might be beneficial for supporting synthetic data and its uses.

5.2. Methodology

The synthetic data application produced for this project supports the simulation of Analysis Ready Data for electro optical remote sensing imagery. The data generation pipeline utilizes industry leading physics-based image simulation technology that enables for the creation of sensor model approximations of existing Earth imaging platforms currently in orbit. This application also leverages remote-sensing derived content that has been assembled to produce “digital twins” of locations on Earth at various scales. This content simulation capability was then configured into a synthetic data channel on the Rendered.ai platform that outputs all necessary

dataset and pixel-level truth information to meet the threshold requirements of the CARD4L-SR specification.

5.2.1. Sensor Simulation

The simulation capability demonstrated in this effort uses the Rochester Institute of Technology's Digital Imaging and Remote Sensing (DIRS) Laboratory simulation technology, DIRSIG. DIRSIG enables the simulation of physically accurate electro-optical data with accurate spectral properties and radiometric responses calculated at sub-pixel resolutions which is done using detailed models of sensor and platform properties that drive a path-traced radiometry estimation performed against provided spectrally defined 3D content.

The image capture platforms chosen to be modeled for this application were:

Table 1

PLATFORM TYPE	SENSOR APPROXIMATION	GSD	SPECTRAL BANDS	ARRAY SIZE
Medium resolution EO	Maxar WorldView-3	~1.24 m	9-channel VIS+NIR	640 x 480
High resolution EO	Planet SkySat 16-21	~0.75 m	5-channel PAN+VIS+NIR	1024 x 768

These platforms represent common data sources for users in the remote sensing and ARD community.

5.2.2. Scenes

Simulation scenes were selected from a set of available radiometrically annotated scenes that provide a variety of geospatial content which can be used to provide product family information. Scenes in DIRSIG are defined using the following.

1. 3D content, including terrain surface model and specific models of above-ground assets
2. Material maps that define material type for all surfaces in the scene
3. Material emissivity curves for all material types referenced
4. Texture maps that associate varied material curves within each material type

The scenes selected for this application are as follows.

- Suburban scene: This scene represents an 8 km2 area modeled after the Rochester suburb of Irondequoit, NY.

- Industrial scene: This scene represents a 10 km2 area modeled after a chemical plant in the desert town of Trona, CA.

Figure 2 — RAI Scene Side By Side (Left: Overhead image of the Suburban scene; Right: Overhead image of the Industrial scene.)

These scenes were constructed by researchers and engineers at RIT's DIRS Laboratory and represent high-fidelity 3D geometry and spectra purpose-built for simulation within DIRSIG.

5.2.3. Atmospherics

The capability to modify atmospheric conditions and visibility is included with the application. This is achieved using atmosphere models generated using MODTRAN spectral modeling software. Separate atmospheric models were generated for urban and rural aerosol levels, as well as summer and winter conditions found at mid-latitude. Within these four categories of atmosphere, five different visibility levels were also modeled, including 5, 10, 15, 30, and 50 km visibility. These variations put the total number of atmospheric combinations at twenty, allowing for a flexible determination of atmospheric properties.

Importantly, atmospheres can also be removed from the simulation to approximate the desired output of imagery post-processed for atmosphere removal. By default, when an atmosphere is selected in the simulation configuration, this channel outputs two images per run of the simulation: one with the atmosphere included and one without. This design was chosen to support the use case of atmospheric removal process development, testing, and validation.

Clouds and cloud shadows are also modeled within this application. This approach uses a voxel-based approach where individual voxels contain information about water vapor concentration, which is used to approximate absorption and scattering of various wavelengths of light.

5.2.4. Synthetic Data Application

The synthetic data application is developed using the "Channel" implementation based upon Rendered.ai's open source Ana framework. The channel is then deployed to the Rendered.ai

platform, allowing users to generate synthetic datasets on-demand using the web-based dataset configuration interface. This graph interface supports the explicit definition of which parameters to control and which to randomize to produce the desired diversity in the output dataset.

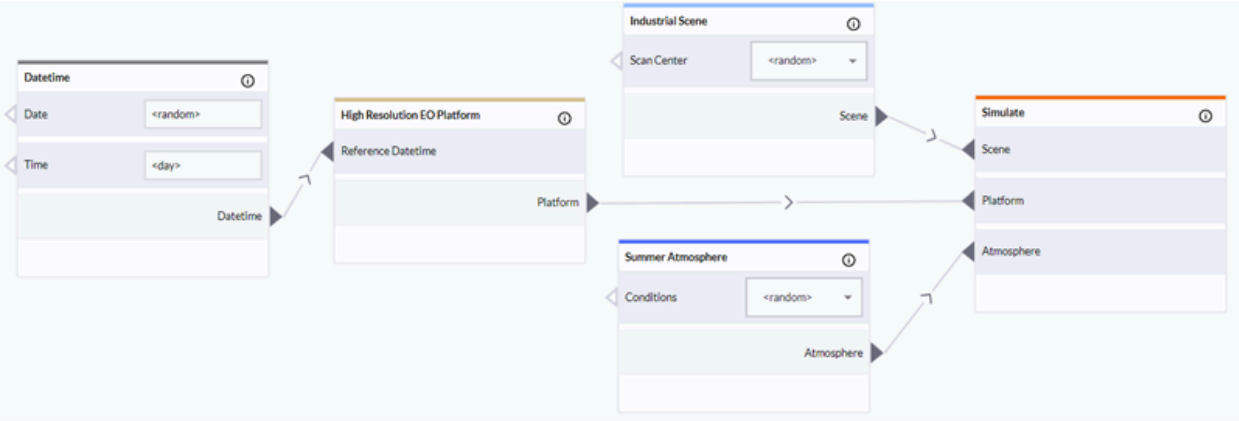

Figure 3 — RAI ARD Graph (Node and edge-based graph configures simulation inputs on the Rendered.ai platform.)

This channel is configured to output imagery with all information required to meet the threshold ARD requirements including dataset and pixel-level metadata. The output of running a simulation using this channel is a zipped folder containing output data cubes for each run of the simulation, JSON files containing dataset-level metadata, and pixel mask images that show pixel-level metadata designated in the ARD specification.

The synthetic data application developed for this project was deployed to the Rendered.ai platform, and can be utilized by the general public by using a content code within the web platform. The content code specific to this application is “ARD.” Within the workspace included in this content code, users will find pre-configured graphs for specific image scenario simulations, as well as pre-generated datasets that include all required metadata and annotations specific to the CARD4L-SR standard.

5.3. Discussion

5.3.1. Applications of Synthetic Data for the ARD Community

Synthetic data has many potential applications for ARD practitioners. The synthetic data channel developed for this project enabled creators of ARD data products to produce baseline datasets to develop and test algorithms for automated generation of ARD calibrations and metadata. This includes atmospheric calibration, water and ice pixel masking, cloud and cloud shadow detection, and terrain occlusion and shadowing. As these processes often require significant amounts of ground truth data to develop, the use of synthetic data can alleviate the need for expensive data collection and labeling campaigns.

Synthetic data also has uses for consumers of ARD datasets, supporting the development and testing of custom processing techniques with known land cover, sensor, and atmospheric inputs. The configurability of the data output enables fine-grained experimentation to understand the impacts that varied conditions and collection parameters have on algorithm performance. Also, because the scene and collection parameters can be fully customized, synthetic data allows users to approximate data collected from imaging platforms that do not yet exist, or of objects on Earth that have never been captured in imagery, thereby reducing barriers to innovation.

5.3.2. Synthetic Data and the ARD Standard

The CEOS ARD standard has been developed to enable users of real Earth Observation data with all information needed to perform common and complex analytics with those data. It is also a helpful guide for developers and users of synthetic EO data, as the standard lays out guidelines for associated descriptive data. Often, annotations and metadata constructed for synthetic data are customized to include only the information relevant to the task it was engineered for, but as synthetic data become more widely adopted, dataset structure metadata will need to become more standardized to be more widely useable. This effort serves as a step in that direction for synthetic EO imagery.

While the CEOS ARD standard was designed to apply to real EO datasets, much of the relevant dataset-level metadata of synthetic datasets generated in this exercise can be incorporated into the existing standard. For instance, the requirement to specify Auxiliary Data (section 1.14) can apply to all input content used in simulation, for example the 3D content and atmospheric databases used. Similarly, the requirement for all algorithms used in dataset generation to be listed (section 1.13) can apply to simulation algorithms used in image generation, for instance the use of DIRSIG for this simulation application. Due to the volume and granularity of this information, a distributed metadata approach, such as separate per-image metadata JSON files like those created by the Rendered.ai platform, is a preferred format for the exchange and transfer of this information.

The diversity needed for effective synthetic data requires the stochastic variation of input parameters to ensure sufficient domain coverage. To ensure data provenance, these stochastic inputs need to be traceable in image-level metadata. This is one area where synthetic data require metadata beyond what is defined in the existing CEOS ARD standard. By default, the Rendered.ai platform creates this level of metadata per run of a simulation. If synthetic data were to be considered as a novel family within the ARD standard, this would be an important element to include.

5.3.3. Lessons Learned and Next Steps

The work done for this project showed that a synthetic EO data pipeline can be developed that can auto-generate data that meet the requirements of a CEOS Analysis Ready Data standard. This standard is already well-defined for accommodating synthetic data, though there are areas where additional requirements may be specified to ensure the most useful output synthetic data products. With synthetic data growing in importance in the realm of AI and data analytics, it is important for this form of data to be considered in any standards development effort.

With this example established, further work could be done to develop Analysis Ready synthetic data with a specific use case in mind, or to supplement an existing real Analysis Ready dataset to address issues of bias or data scarcity in the real data. Beyond this, further work could be done to develop a formal substandard within ARD to specifically support synthetic data needs, including requirements discussed in this report surrounding stochastic simulation input information to ensure fully traceable data products and outcomes.

5.4. Conclusion

This effort serves to introduce the concept of synthetic data within the ARD community. To that end, the application developed as part of this effort will be included as a “Content Code” in the Rendered.ai platform, allowing new users to utilize the content and capabilities described in this report using a complementary thirty-day trial of the Rendered.ai platform allowing for simulation configuration and unlimited dataset generation within that period. From there, experiments can be run using the functionality and the meta described, including atmosphere removal processing, cloud detection, and detection of various land cover elements in varied scenarios.

Hopefully this effort will introduce users in the ARD community to the concept of synthetic data for EO applications. As the potential value of synthetic data is realized within this community, this will necessitate thoughtful planning around the incorporation of synthetic data techniques into demonstration and validation of Analysis Ready Data standards.

6

STUDY OF COASTAL ENVIRONMENTS IN THE ARCTIC

6 STUDY OF COASTAL ENVIRONMENTS IN THE ARCTIC

NOTE: This scenario was led and implemented by Pelagis Data Solutions. Pelagis is an ocean-tech venture located in Nova Scotia, Canada focused on the application of open geospatial technology and standards designed to promote the sustainable use of our ocean resources.

6.1. Introduction

Remote sensing of marine and coastal environments plays an increasingly important role towards monitoring the sustainable use of ocean resources. As the effects of climate change are especially impactful to coastal ecosystems, Earth Observation (EO) derived analysis ready datasets corrected for environmental bias and spatial and temporal resolution provide valuable insights into coastal areas that otherwise are very difficult, if not impossible, to monitor, such as mapping habitat extent and change, understanding biogeochemical processes, and monitoring human impacts and conservations.

This Testbed 19 project was designed to enhance previous work positioning the OGC suite of standards and best practices at the core of a federated marine spatial data infrastructure (MSDI). In particular, analysis ready datasets for marine and marine-terrestrial realms, as defined by the International Union for Conservation of Nature(IUCN) Global Ecosystem Typology, are reviewed and purposed towards the development of essential climate and biodiversity variables for coastal marine environments in the Canadian Arctic.

6.2. Challenge

The concept of analysis ready data has historically been targeted towards satellite-derived datasets *processed to a minimum set of requirements and organized into a form that enables immediate analysis with a minimum of additional user effort* and *interoperable through space and time*. Although the term “analysis readiness” appears relatively generic, in practice analysis ready datasets must adhere to a minimum threshold of requirements. These requirements include defining the characteristics of the dataset, the per-pixel properties and capabilities, and the metadata describing atmospheric and geometric corrections applied to dataset observations.

A key benefit of analysis readiness is that it hides the complexities of data collection and processing of raw satellite imagery and provides *application* ready datasets targeting specific scenarios. In terms of interoperability of these datasets, the key characteristics when applied to geospatial applications are the spatial and temporal properties of the dataset. Information

on these characteristics permits client applications (and users) to determine suitability of such datasets applied to specific domain problems over a specific temporal range and spatial extent.

Similarly, OGC provides Standards and specifications that address collections of observations provisioned through in-situ platforms and sampling programs. The OGC Abstract Specification Topic 20: Observations, measurements, and samples version 3.0 (OGC OMSv3)[60] models collections of observations associated with the properties of a feature of interest. Observations are modeled as collections over an observed property and allow for subsequent processing to derive 'analysis readiness'. In this context, it is important to understand the overlap of term definitions to further ensure the interoperability of analysis ready datasets independent of the platform from which the datasets were derived. This separation of concerns is well addressed by the OMSv3 specification.

The following scenario revisits the role of analysis ready datasets within a regionally applied climate monitoring system. The scenario was designed to leverage analysis ready datasets combined with in-situ observations to draw direct relationships between a changing environment and dependent human activities. The core of this exercise focuses on the application of OGC Standards and specifications as adapters to provision analysis ready datasets relative to key ocean and coastal climate indicators. The usability of satellite-derived observations is dependent on key processing algorithms that transform the raw observation collections into key environmental indicators that either directly measure essential variables associated with a region of interest or indirectly contribute to further processing towards similar goals.

6.2.1. Analysis Ready Datasets for a Digital Arctic

The Global Climate Observing System (GCOS) defines a set of Essential Climate Variables (ECVs) representing key variables that contribute to the characterization of Earth's climate. In particular, sea ice is a key indicator of climate variability in the polar regions. Three key components representative of sea ice variability are sea ice concentration, sea ice thickness, and surface albedo. Sea ice is defined as frozen sea water which floats on the surface of the ocean, excluding ice shelves which are anchored on land but protrude out over the surface of the ocean. Long-term monitoring of sea ice is important for understanding climate change and the related impact on regional biodiversity and ecosystem services. The sea ice - surface albedo relationship is a key component of climate monitoring. A decrease in sea ice coverage directly affects surface albedo with a corresponding increase in solar heating of ocean waters.

To support the integrity of climate observations, GCOS identifies the measurable parameters to be used to characterize each ECV. For example, sea ice concentration (coverage), sea ice surface albedo, and sea ice thickness. The requirements for each measured parameter are similar to the CEOS defined Product Family Specification (PFS) schema in that GCOS defines five criteria to be used to assess the quality of measurement – spatial resolution (horizontal, vertical), temporal resolution, measurement uncertainty, stability (i.e., effects of bias over time), and timeliness (how often is the phenomenon measured and made available, e.g., daily, monthly).

For each criterion, there is a set of guidelines that must be met to support the application of any measurement to the ECV. A *goal* (G) is the ideal requirement to be met by an observation collection, a *threshold* (T) representing the minimum acceptable value and a *breakthrough* (B) value representing an intermediate level between the goal and threshold identifying limits of

applying such measurements to specific use cases. For example, to analyze sea ice concentration for near-coast applications, the goal is set to a horizontal resolution of 1km whereas regional applications are limited to 5km resolution.

This work item complements the deliverables associated with the Digital Arctic theme of the OGC FMSDI 2023 project. In particular, the goal is to leverage satellite derived datasets to identify the changes in sea ice coverage, thickness, and surface albedo related to features of interest within the circumpolar Arctic.

There are several data sources available that provide measurements of sea ice coverage and surface albedo for the arctic region. Current work has focused on integration with data products provisioned through the Copernicus Climate Data Store (CDS) and NASA's National Snow and Ice Data Center Distributed Active Archive Center (NSIDC DAAC).

Processed observations of sea ice coverage and surface albedo provided through the Copernicus Data Store are made available through the ERA5 reanalysis product gridded to a regular latitude longitude of 0.25 degrees. In addition, there is the Copernicus Arctic Regional Reanalysis (CARRA) data product gridded to regional resolution of 2.5km. For the Testbed 19 exercise, the CARRA-WEST dataset was used as the baseline for monitoring the regional climate indicators for sea ice coverage and surface albedo.

The NSIDC DAAC provides the ICESat-2 data collection derived from the Advanced Topographic Laser Altimeter System (ATLAS) instrument aboard the Ice, Cloud, and Land Elevation Satellite-2 (ICESat-2). The NSIDC DAAC distributes Level-1, Level-2, Level-3A, and Level-3B ICESat-2/ATLAS products, which range in temporal coverage from October 2018 to present. These datasets are based on the polar orbit of the ICESat-2 satellite separating each ground track revisit into separate data granules. There are 1387 reference ground tracks in the ICESat-2 repeat orbit. The reference ground track increments each time the spacecraft completes a full orbit of the Earth and resets to 1 each time the spacecraft completes a full cycle. Metadata specific to each polar orbit is maintained as a queryable and discoverable service identifying specific data granules based on spatial and temporal extents. A spatial query provides reference to the available data granules with reference ground tracks (RGTs) that intersect the extent of a feature of interest. Once established, temporal queries are available to isolate the set of data granules representing each ground track revisit. For the purpose of the Testbed 19 exercise, the focus was on the [74] data product providing along track heights for sea ice and open water leads.

6.3. Approach

This exercise leverages the emerging role of the OGC GeoDataCube initiative to transform the native formats provided through the CDS and NSIDC data stores into a Rasdaman vector database exposing the native data stores as "analysis ready." For context, the area of concern was established around the protected areas of the Canadian Arctic — specifically the Ninginganiq National Wildlife Area located on the east coast of Baffin Island, Nunavut.

Temporal analysis of Surface Albedo

The inherent value of the GeoDataCube (GDC) framework is its ability to scale both in terms of volume of data and processing capabilities. This use case focuses on delegating the analysis of surface albedo to the GDC service provider over a temporal extent to determine the magnitude of change for each gridded observation. In this case, the GDC provider translates the request to a native OGC Web Coverage Processing Service (WCPS) process graph for execution by the GDC service instance to determine the change in surface albedo for the region of interest over the temporal range of May 2019 — April 2023 (see Figure 4).

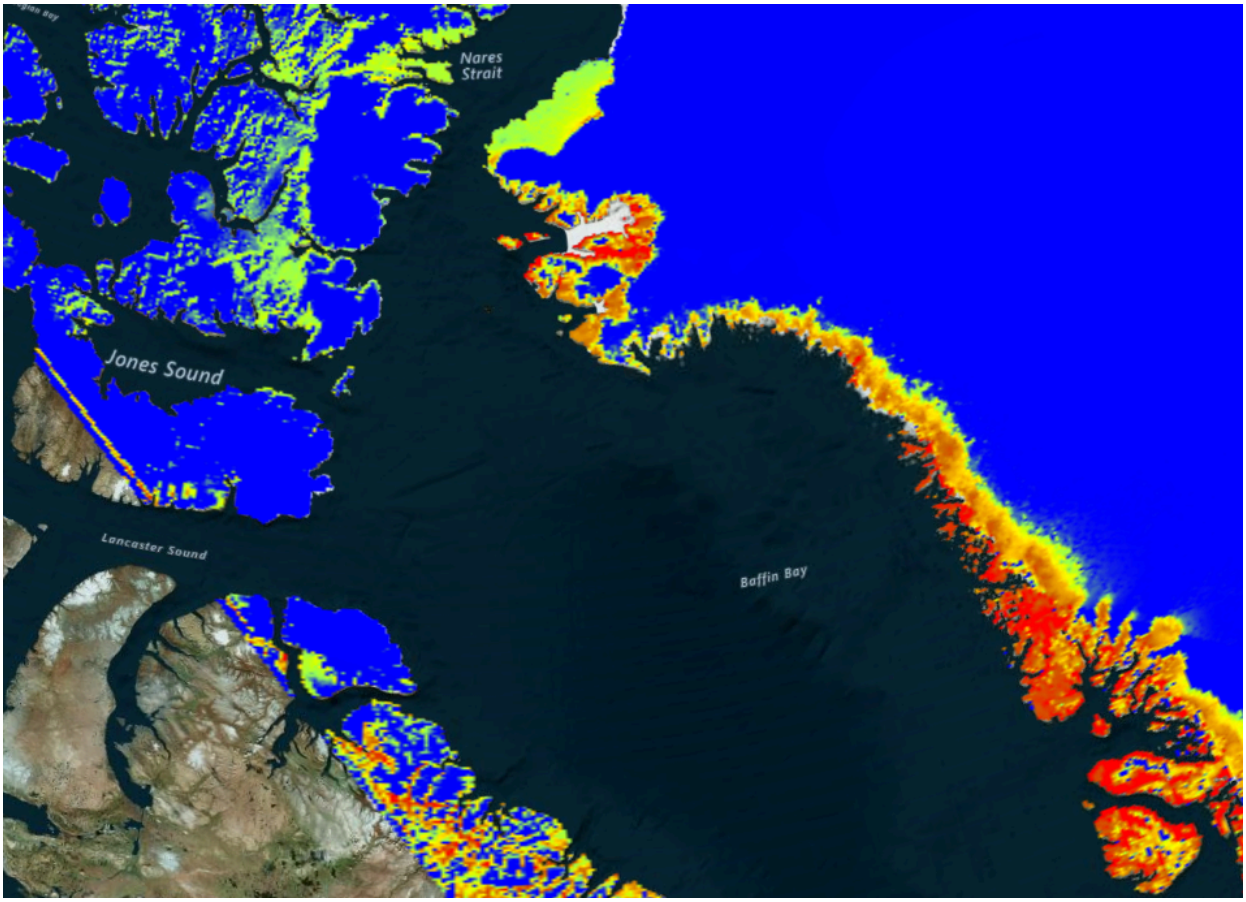

Figure 4 — Magnitude of change in surface albedo between May 2019 and April 2023

6.4. Standards and Interoperable Technologies

The following standards and interoperable technologies were evaluated as part of the Testbed-19 ARD initiative.

OGC Observation System Models

The coastal environments scenario extends the concept of Analysis Ready Data to include processed data pipelines sourced from in-situ observation collections and sampling programs. Raw data, such as NetCDF datasets provided through the NOAA Saildrone program for monitoring ocean conditions, are processed into an 'ARD' encoded using the OGC Moving Features Access Standard. Extending the concept of ARD to include datasets sourced from non-

satellite based observing platforms permits a consistent view of important datasets independent of the datasets' originating platforms and associated processes and procedures. This serves to validate the interoperability requirements of analysis ready datasets based on the minimum level of processing for spatial and temporal correctness.

This Testbed 19 work effort extended the OGC Observations & Measurements Abstract Specification (OMSv3) and the Connected Systems initiative to represent ARDs, with the appropriate metadata model, as a type of observation system providing essential variables and regional indicators for features and coverages of interest. To this effort, there is currently an initiative to align the OMSv3 observation systems metadata model with the GeoDCAT application profile of the ISO 19115-1:2014 Geographic information Metadata Standard. This effort may provide an opportunity to evaluate the requirements of the ARD metadata model in this same context.

OGC Moving Features

The OGC Moving Features Access [91] Standard defines a standard encoding for features in motion. Testbed-18 work extended this model to include observation collections made along a trajectory by an observer. Continuing to extend the principles of the Moving Features Standard, the Testbed 19 exercise uses terms defined in the OGC API — Connected Systems — Part 1: Feature Resources Standard[105] to model the ICESat-2 satellite as a *platform* hosting the ATLAS sensor. The *Reference Ground Track* represents the movement of a 'virtual' observer traveling along the earth surface providing height *observations* against a *Feature of Interest* with *observable property* 'sea ice height'.

Leveraging the Moving Features model to encode the sea ice height observations allows client applications to ingest the surface height along a trajectory and apply trajectory based analysis directly against the observation collection. For example, the ICESAT/2 revisit to the Ninginganiq National Wildlife Area may be down-sampled to intervals of 1/10th of a second (See Figure 5).

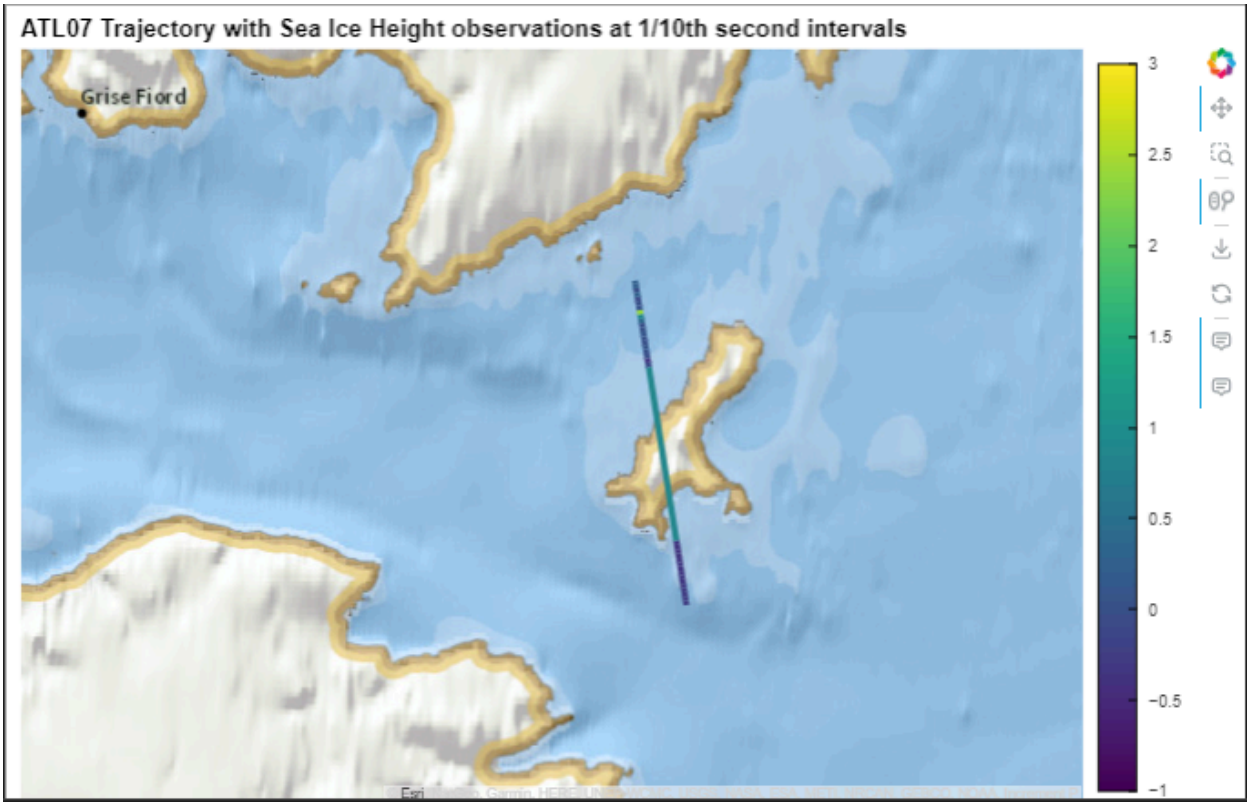

Figure 5 — Downsampled Sea Ice Height Observations over the Ninginganiq National Wildlife Area - May 2019

6.5. Future Work

Seasonal effects on Data Quality

Under investigation is an issue with measuring sea ice extent in the summer months in which melting ponds may affect the remotely sensed observations. The melting season sea ice radar freeboards require a correction for an electromagnetic range bias which is provided separately through the British Antarctic Survey (BAS), a component of the National Environment Research Council (NERC UK) [28]. It is unclear how a proposed Product Family Specification will address uncertainty issues for the overall data quality when such metrics are dependent on seasonal (temporal) conditions affecting observation measurements.

Spatial Resolution

Analysis of regional biodiversity indices and indicators requires a finer resolution of processed satellite-derived observation datasets. Level 3 processed earth observation data products tend to lose fidelity to offset exponential storage costs. As an example, the Copernicus CARRA-WEST reanalysis data product is provided as a 2.5km gridded coverage for specific regions of the Arctic whereas the global ERA5 dataset is provided at 25km gridded coverage. In line with the GCOS criteria for application of earth observations to coastal areas, the CARRA-WEST data product

meets the requirements of horizontal resolution for regional analysis whereas the ERA5 data product may only be used in a global context.

Future work efforts related to Testbed-19 are designed to evaluate the effectiveness of regionally specific ARDs that maintain the near-raw spatial resolution satisfying the criteria of GCOS while not incurring storage costs associated with global coverage models.

Temporal Resolutions

Satellite observations are periodic relative to a reference feature of interest. The ICESat-2 mission, for example, revisits the same reference ground track on a 91-day schedule resulting in sea ice observations that may not overlap with a specific event or period of interest. Inferring observations along a temporal coverage is required not only at time instant of an observation event but also for relevant periods of observations to infer causal analysis.

GeoDataCubes and Analysis Ready Datasets

The concept of multi-dimensional arrays and the arrays' use supporting earth observation missions are well established. However, as per OGC initiatives related to GeoDataCubes and Analysis Ready Datasets, further development to enhance the concept along a spatial and temporal coverage is required to support a scalable framework for climate observations and biodiversity.

7

RESULTS, DISCUSSIONS, AND CONCLUSIONS

7 RESULTS, DISCUSSIONS, AND CONCLUSIONS

NOTE: This section summarizes findings and conclusions.

7.1. Results and Discussions

The following three scenarios described in this Engineering Report demonstrated that ARD provides numerous benefits.

- Improved metadata: ARD data is well-documented with metadata and related catalog information making it easier for users to find and understand the data.
- Increased ease of integration: ARD data is easy to integrate with other data sources which makes it possible to create new and innovative applications that combine ARD with other types of data.
- Improved interoperability: ARD data is interoperable with other data systems making it possible to share and exchange ARD data between different organizations and communities.
- Uncertainty reduction: ARD data is processed to reduce uncertainty which makes it more reliable and accurate for use in decision-making.
- Provenance: ARD data has provenance metadata, which documents the source of the data and the steps that were taken to process the data. This helps users to understand the limitations of the data and to use it appropriately.
- Autonomous workflows: ARD data can be used to create autonomous workflows that can process and analyze data without human intervention which can save time and resources.
- Geolocation accuracy enhancement: ARD data is processed to improve geolocation accuracy making it more useful for applications such as mapping and navigation.
- Workflow characterization: ARD data can be used to characterize workflows, which can help to improve the efficiency and effectiveness of the workflows.
- Refined requirements on all aspects of ARD: The ARD community needs to be constantly working to refine the requirements for ARD which ensures that ARD data meets the needs of its users.

Overall, ARD makes data more accessible, reliable, and interoperable. With existing de facto standards or specifications for ARD (e.g., CARD4l), there are some challenges in preparing, publishing, and discovering ARD. The following are some of the challenges surfaced in Testbed 19.

- Incomplete Standards (metadata gaps): The testbed scenarios highlighted two critical areas where the existing CEOS ARD specifications need improvement in terms of metadata. First, an inconsistency in quality measures and encoding within CARD4L data was observed. Data from the same scenario may use different measures and encoding methods, which hinders overall consistency and interoperability. Second, the scenarios revealed compatibility issues with metadata for CEOS ARD-compliant data. Existing software tools may not support the metadata due to non-standard formats and a lack of widespread adoption, thereby creating challenges in accessibility and utilization.

- Limited Adoption and Support (recognition and participation gaps): Two key obstacles to user adoption and support were identified when implementing the Testbed scenarios. First, there is limited support and recognition for existing CEOS ARD standards. Commercial and open-source tools may not fully recognize or utilize data released based on current ARD standards, which hinders the accessibility and integration of the data. Second, there is a prevalent restriction on user adoption of existing ARD specifications. The lack of widespread recognition and support creates a barrier to broader user adoption and application of the valuable principles embodied in ARD.

- No Standard for Non-EO Data (data type Gaps): The current CEOS ARD specification does not include non-EO data. The CEOS ARD specification needs to be expanded to accommodate non-EO data, such as building data in a gentrification scenario, in-situ data in a coastal study scenario, and simulated electro-optical remote sensing imagery in a synthetic data scenario. Some data, such as training datasets that label scenes of images and/or objects for machine learning applications, may not need to be spatially aligned. These datasets may be considered ready for machine learning, but not necessarily GIS-ready. The inclusion of these datasets may necessitate further consideration of the expansion of the ARD scope and adjustments to its underlying common minimum requirements. In terms of the creation and publication of ARD, the extension of specifications is also necessary for non-EO data and data not intended for GIS applications.

- Levels of Readiness (readiness gaps): The testbed scenarios exposed a challenge regarding levels of data readiness. Different applications require varying degrees of preparedness, as evident in the gentrification scenario. Analyzing time series data within this scenario necessitates a strong emphasis on the temporal dimension, which might not be fully addressed by current standards. Readily utilizing such data demands a distinct level of preparedness compared to other analyses, potentially requiring further task-specific data preparation. To accommodate diverse analytical needs, flexible and adaptable standards encompassing various levels of readiness are necessary which will ensure data are appropriately equipped for different tasks and applications, maximizing the data's utility and impact across domains.

7.2. Recommendations to ARD standard

Recommendations:

- Leverage existing standards to facilitate the description and publication of ARD: Existing standards developed by organizations such as ISO, OGC, and QUDT for data description, publication, and interoperability are likely already supported by many existing tools and applications. Incorporating these established standards and specifications into the definition of ARD standards can potentially increase adoption and reduce development burden. For example, ISO/OGC Coverages (such as Coverage Implementation Schema CIS 1.1) can be reused to support ARD by making some adaptations for data quality and metadata requirements.

- Clearly Defining Scope and Levels of Readiness: Ensuring effective analysis ready data (ARD) requires clarity in both the scope and levels of readiness of the data which can be achieved through two key actions: Clarifying the scope and defining adaptable levels of readiness. The current definition of CEOS-ARD needs to be broadened to encompass non-EO data types which should include clear guidelines for incorporating diverse data such as building footprints, socio-economic statistics, and in-situ measurements within ARD standards. Secondly, the existing "minimum" and "goal" levels of conformance should be revised to better address diverse data types and user needs. Additionally, introducing new levels tailored to specific requirements, such as "GIS-ready" and "machine-learning-ready," would further accommodate temporal emphasis and other special needs. Furthermore, enabling flexible data provisioning empowers data providers to declare specific conformance levels based on the data's characteristics and intended applications enabling data to effectively cater to the diverse readiness needs of stakeholders such as data analysts, GIS analysts, and machine learning practitioners.

- Inclusion of Training Datasets in ARD: The inclusion of training datasets into the Analysis Ready Data (ARD) standard is recommended to facilitate the development and application of machine learning (ML) algorithms for geospatial data analysis. ML algorithms have gained significant prominence and are now employed in a wide range of geospatial applications, including land cover classification, change detection, and feature extraction. However, the development of ML algorithms necessitates substantial amounts of training data, which are often difficult to obtain. Incorporating training datasets into the ARD standard would provide a consistent and readily accessible source of training data for geospatial ML applications. The existing CEOS-ARD specification mandates the inclusion of metadata pertaining to the data source, the data collection process, and the data quality. This metadata model can be utilized to evaluate the quality and suitability of training datasets for specific ML applications. On the other hand, the OGC Training Data Markup Language for Artificial Intelligence (TrainingDML-AI) Part 1: Conceptual Model Standard defines a data model for datasets that encompasses information about the data, the labels, and the ML algorithm employed to create the labels. This unique emphasis on ensuring the compatibility of the training dataset with various ML frameworks and tools can be incorporated into the developing ARD standard by expanding the metadata model. The inclusion of training datasets in ARD standards would contribute to the achievement of increased availability of training data for geospatial domains, improved data quality and suitability for specific ML applications, and enhanced data interoperability with different ML frameworks and tools.

- Synthetic Data in ARD: The inclusion of synthetic data in ARD standards warrants further investigation due to the potential of synthetic data to enhance the availability, quality, and diversity of training data for geospatial machine learning (ML) applications. Synthetic data offers a unique advantage over real-world data in that it possesses

inherent tractability to underlying physical models and provides complete control over the simulated environment, enabling the generation of customized datasets tailored to specific ML tasks and geospatial applications.

Uniqueness of Synthetic Data:

(1) Tractability to Physical Models: Synthetic data is generated based on well-defined physical models, allowing for a deeper understanding of the relationships between input features and output labels, which can enhance the interpretability and generalization ability of ML models.

(2) Full Control of Simulated Environment: The controlled nature of synthetic data generation enables the creation of datasets with specific characteristics and scenarios that may be difficult or impossible to obtain from real-world data, providing valuable training material for a wide range of ML tasks.

(3) Customization and Augmentation: Synthetic data can be customized and augmented to address specific data biases or limitations in real-world datasets, ensuring that ML models are trained on a diverse and representative representation of geospatial phenomena.

Areas for Further Investigation:

(1) Standardization and Integration: Research on standardizing the format and metadata of synthetic data to facilitate integration into the ARD standard.

(2) Quality Assessment and Evaluation: Development of methods for assessing the quality and suitability of synthetic data for specific geospatial ML tasks.

(3) Generation of Realistic and Representative Synthetic Data: Exploration of techniques for generating synthetic data that accurately represents real-world geospatial phenomena and incorporates realistic levels of complexity and variability.

(4) Integration with Geospatial ML Frameworks and geophysical models: Investigation of methods for integrating synthetic data into existing geospatial ML frameworks and tools or geophysical models for earth science analysis.

- Data quality and representation: Several specific recommendations arose from the scenarios regarding data quality and representation. The first focuses on addressing pixel representation ambiguity. Establishing a clear consensus on interpreting pixel representation correctly (such as discussions on pixel as points or areas — ”pixel-is-point” vs. “pixel-is-area”) and ensuring consistency across datasets is crucial. Secondly, standardization of units of measure is imperative. Utilizing consistent and machine-readable units (e.g., QUDT) facilitates seamless data integration and analysis. Third, improved temporal dimension handling is necessary. Addressing the varying needs of time-series analyses by incorporating temporal considerations into data serving and standards will enhance flexibility. Finally, enriching metadata structure is key. Implementing structured metadata with defined extensions improves data discovery, understanding, and provenance tracking, empowering users to efficiently locate, comprehend, and utilize the data. Addressing these essential aspects can ensure high-quality ARD, maximizing its potential as a useful data ready for research and application.

7.3. Conclusions

Scenarios (gentrification, coverages, synthetic data, and coastal study) confirmed the benefits of ARD: improved data characterization, ease of integration for satellite and non-satellite data and services, interoperability (data, services, and tools), uncertainty, data provenance, autonomous workflows, geolocation accuracy, workflow characterizations, and ARD data organization. Several challenges have been identified: inconsistent metadata encoding, software support, high level of readiness, specialized data readiness, and non-geospatial domain priority access.

Recommendations for the development of an ARD standard include compatibility with existing standards, scope of geospatial analysis ready data, and levels of readiness or conformance.

Future directions for the development and testing of ARD standards may include expanding data products, applications, and readiness levels. Training datasets, GeoDataCubes, and link: CDB are other OGC standards activities related to the development of analysis ready data. The interoperation and support of ARD in these communities needs further study and testing. Uniform evaluation and conformance certification may be further enforced to ensure the consistency of data readiness in terms of hierarchies and application domains.

A

ANNEX A (NORMATIVE) ABBREVIATIONS/ACRONYMS

A ANNEX A (NORMATIVE) ABBREVIATIONS/ACRONYMS

1D	One-Dimensional
3D	Three-Dimensional
4D	Four-Dimensional
API	Application Programming Interface
ARD	Analysis Ready Data
ARDC	Analysis Ready Data Cube
ASCII	American Standard Code for Information Interchange
ATLAS	Advanced Topographic Laser Altimeter System
BAS	British Antarctic Survey
BIPM	Bureau International des Poids et Mesures
CARD4L	CEOS Analysis Ready Data for Land
CARD4L-SR	CEOS ARD for Land – Surface Reflectance
CARRA	Copernicus Arctic Regional Reanalysis
CCD	Charge-Coupled Device
CDS	Climate Data Store
CEOS	Committee on Earth Observation Satellites
CEOS-ARD	Committee on Earth Observation Satellites — Analysis Ready Data
CIS	Coverage Implementation Schema
CODATA	Committee on Data of the International Science Council
CRS	Coordinate Reference System

CURIE	Compact URI
DAAC	Distributed Active Archive Center
DAG	Directed Acyclic Graph
DEG_C	Degree Celsius
DEM	Digital Elevation Model
DIRS	Digital Imaging and Remote Sensing
DIRSIG	Digital Imaging and Remote Sensing Image Generation
DRUM	Digital Representation of Units of Measurement
ECV	Essential Climate Variable
EDR	Environmental Data Retrieval
EEA	European Environmental Agency
EO	Earth Observation
EPSG	European Petroleum Survey Group
ER	Engineering Report
ERA5	ECMWF Reanalysis v5
EU	European Union
FAIR	Findable, Accessible, Interoperable, and Reusable
FAIRiCUBE	FAIR information cube
FLWOR	For, Let, Where, Order by, Return
FMSDI	Federated Marine Spatial Data Infrastructure
FOV	Field of View
GCOS	Global Climate Observing System
GDC	Geo Data Cube
GML	Geography Markup Language
GSD	Ground Sampling Distance
ICESat-2	Ice, Cloud, and land Elevation Satellite-2

INSPIRE	Infrastructure for Spatial Information in Europe
I/O	Input/Output
IRECI	Inverted Red-Edge Chlorophyll Index
ISO	International Organization for Standardization
IUCN	International Union for Conservation of Nature
JSON	JavaScript Object Notation
ML	Machine Learning
MODTRAN	MODerate Resolution Atmospheric TRANsmission
MSDI	Marine Spatial Data Infrastructure
NASA	National Aeronautics and Space Administration
NATO	North Atlantic Treaty Organization
NERC	National Environment Research Council
NetCDF	Network Common Data Form
NIR	Near Infrared
NOAA	National Oceanic and Atmospheric Administration
NSIDC	National Snow and Ice Data Center
OGC	Open Geospatial Consortium
OGC-NA	OGC Naming Authority
OLAP	Online Analytical Processing
PAN	Panchromatic
PFS	Product Family Specification
QUDT	Quantities, Units, Dimensions, and Types
RDF	Resource Description Framework
RIT	Rochester Institute of Technology
SI	International System of Units
SLA	Service Level Agreement

SKOS	Simple Knowledge Organization System
SOS	Sensor Observation Service
SPS	Science for Peace and Security
SQL	Structured Query Language
SWE	Sensor Web Enablement
TIFF	Tag Image File Format
TIRS	Thermal Infrared Sensor
UCUM	Unified Code for Units of Measure
UML	Unified Modeling Language
URI	Uniform Resource Identifier
URL	Universal Resource Locator
USGS	United States Geological Survey
UTC	Coordinated Universal Time
VIS	Visible
WCS	Web Coverage Service
WCS-T	WCS Transaction
WCPS	Web Coverage Processing Service
WGS84	World Geodetic System 1984
WMS	Web Map Service
WMTS	Web Map Tile Service
XML	Extensible Markup Language

B

ANNEX B (INFORMATIVE) RASDAMAN ARD ANALYSIS

B

ANNEX B (INFORMATIVE) RASDAMAN ARD ANALYSIS

NOTE: This contribution was prepared by rasdaman GmbH, a German academic spinoff whose product is the rasdaman ("raster data manager") federated datacube engine. The company has been and continues to be active in OGC, ISO, and EU INSPIRE standardization The focus is on the coverage standards that define the datacube data and service models.

B.1. Executive Summary

The rasdaman ARD analysis investigated how analysis-ready (better: consumption-ready) the OGC coverage standards are. Across all relevant standardization bodies, coverages are the accepted paradigm for spatiotemporally varying data ("fields" in the sense of physics). Technically speaking, coverages encompass regular and irregular grids, point clouds, and general meshes. Gridded data, specifically, resemble datacubes, is the particular focus of this report.

Data about the Earth, like in many other domains, are too difficult to access. In order to perform some insight-gaining task, a series of steps must be performed. These tasks often require a spectrum of detailed technology skills which are not related to the original Earth science task on hand. Special-purpose file formats with sometimes rather peculiar mechanics, juggling with horizontal, vertical, and time reference systems, and scaling up processing to large amounts of data are just a few of such common issues. One reason is that data often are provided in a more generator-centric (where generator can be a sensor or a program, such a weather forecast) rather than a user-centric manner, which might also be called "too upstream."

As is well-known, generator-centric (rather than user-centric) data hinder EO exploitation significantly, making such tasks impossible to conquer for non-experts and tedious for experts. For the desirable, user-friendly opposite approach, the term Analysis-Ready Data (ARD) was coined by the USGS Landsat team and has since gone viral.

However, despite significant work and visible progress, such as in CEOS, ultimately it is by no means clear what ARD exactly means and how it can be achieved.

In this Appendix, a fresh look is taken at the problem. The focus is on spatiotemporal raster data, i.e., datacubes, modeled as coverages according to the predominant OGC and ISO standards. The Holy Grail of this study is the ability to provide automatic data fusion of Earth data. The analysis provided in this Appendix is based on long-term practice (and suffering).

Current shortcomings and proposals for the way forward are listed, including research and standardization directions.

B.2. Section 1 — Introduction

This Appendix provides a concrete application and format independent discussion of the current status of ARD and provides recommendations. Further, the OGC Web Coverage Processing Standard (a potential datacube analytics language) provides a framework for the discussion of operations as a convenient way to write down atomic and composite operations on coverages. Results, however, are independent from any particular request expression style.

In the following discussion, the ultimate exercise of ARD and interoperability is considered to be automated data fusion. This is the process of combining two independently produced data sets in a mathematically rigorous way, automatically, and without human intervention. Obviously, this poses particularly high requirements on the "compatibility" of the data. Therefore, particular attention will be devoted to this use case.

Since at least 1997, the need for harmonizing remote sensing data was expressed [27]. The definition cited by Strobl, for example, calls EO data homogenized if "geophysical values [are] agnostic of the originally acquiring sensor and observation condition" and concludes that in this case the data are "directly comparable." Algorithmic comparison requires combination and, hence, appears to be a kind of data fusion. However, there is more homogenization needed before such a combination can be done. Therefore, such homogenization was chosen as one of the readiness test cases.

This research work was done in the context of OGC Testbed-19, EU FAIRiCUBE, and NATO SPS Cube4EnvSec.

The remainder of this Appendix is organized as follows. Section 2 is a brief primer on the OGC/ISO coverage data model. Section 3 is where existing and missing ARD qualities are inspected and proposals for ARD enhancement are made. Section 4 provides a synopsis of the recommendations. Finally, Section 5 concludes the analysis on achieving analysis ready data using the coverage standards.

B.3. Section 2 — Coverages

In this section, an overview of the OGC Coverage Implementation Schema (CIS) Standard is presented[45]. Version 1.1 introduces the *General Grid Coverage* structure. This structure is both a simplification and an extension of the previously considered coverage types, *Rectified Grid Coverage* and *Referenceable Grid Coverage* of CIS 1.0. The OGC CIS 1.0Standard is also ISO 19123-2 [44] which is currently equivalent to OGC CIS 1.0. A new work item is being processed to adopt OGC CIS 1.1 as an update to 19123-2. Bottom line, OGC CIS 1.1 is currently the most advanced — and handy — coverage data model standard.

Later discussions on operations references the OGC WCPS geo datacube analytics language standard [75], which is also ISO 19123-3 [58]. However, this is for discourse convenience only — operations can be expressed in any style, in both desktop and cloud environments, etc.

B.3.1. Overview

In this section provides an overview of the coverage model and structure as a basis for what follows. Readers familiar with it may safely skip this section. For more detail, there exist tutorials [59][73], an interactive sandbox [90], and a public OGC wiki on coverages and datacubes [106].

Coverages are standardized by both ISO and OGC in close collaboration. Concepts and Terminology are established in ISO 19123-1 [26][112] which has also been adopted by OGC as an update to OGC Abstract Topic 6 [115]. A concrete, interoperability-testable data structure based on these concepts is given by the OGC Coverage Implementation Schema 1.1 [45][118], specifically: its *General Grid Coverage* structure with its schemata for XML, JSON, and RDF encoding [45][3].

B.4. The Coverage Structure

Conceptually, a coverage is a function mapping location (i.e., coordinates) to data values. In plain words, a coverage offers some value (such as a color pixel) for each of the coordinates the coverage covers. These coordinates are called *direct positions* and only at these direct positions in a coverage is there a value. In discrete coverages, these are the only points from which data can be retrieved. In continuous coverages, values between the direct positions can be derived through interpolation.

Technically, the coverage data structure consists of four main components (plus some details which are ignored at this level of detail) as follows.

- *domain set*: Where can values be found?
- *range set*: The values.
- *range type*: What do the values mean?
- *metadata*: What else should be known about these data?

The UML diagram in Figure B.1 illustrates this structure. It shows the four components domain, range, range type, and metadata, of which the last one is optional. Each component is now described.

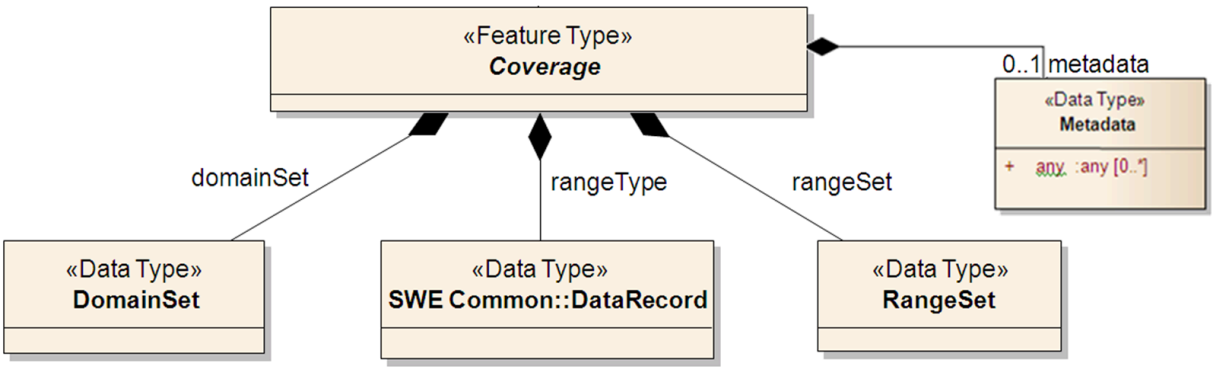

Figure B.1 — High-level coverage structure of OGC CIS 1.1 [118]

Domain set. The domain set consists of direct positions where values are located. As raster data/datacubes are focused only on grid coverages, the domain set forms a (regular or irregular) grid. Such a grid, as well as its grid coordinates, can be of any number of dimensions (better said: axes), made up from spatial, temporal, and other axes such as spectral frequencies. The underlying grid space of a domain set is defined by its corresponding Coordinate Reference System (CRS). More about CRS later in this Appendix.

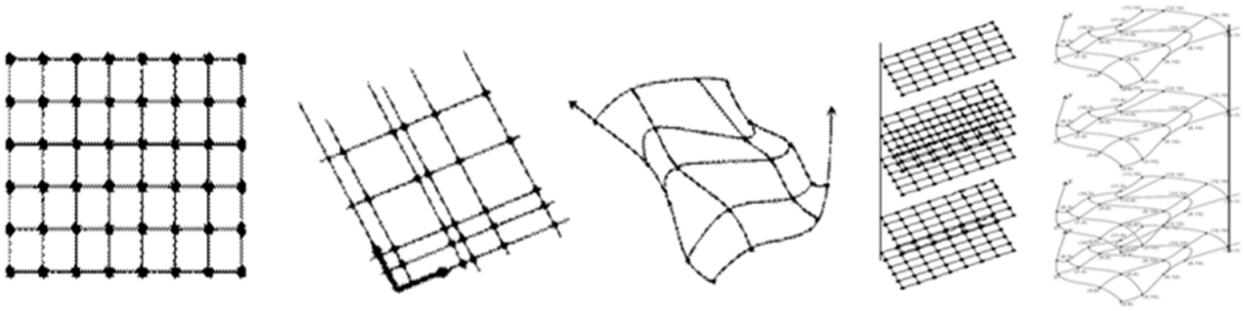

Figure B.2 — Examples of regular and irregular grids [26]

In gridded coverages (aka datacubes), coordinates are aligned on some grid. Still, there is a wealth of variety for possible grid types: Cartesian or geo-referenced, space or time, or something else, regular or irregular, etc. With growing complexity, the description of a grid, as part of the domain set, grows, and likewise so does the size of the corresponding domain set. The simplest case is a regular Cartesian or geo-referenced axis. In this case, what are simply needed to be stored are lower and upper bound as well as resolution. The case of an irregular axis is more involved: all the individual grid points on the axis between its lower and upper bound need to be stored explicitly. More complex grids require even more involved representations.

Below is an example for a CIS 1.1 *GeneralGridCoverage* grid (XML representation) that involves both regular axes (*Lat* and *Long*) and an irregular axis (time). Note the definition of *Lat* and *Long* as regular axis with a given resolution, as opposed to the explicit enumeration of the time steps in the *date* axis. The underlying (Cartesian) grid, modeling the array data structure, is given by the *GridLimits*. However, this is just a technical detail and often of no further concern. More important is the *srsName* (spatial reference system name — a GML legacy naming, as also time is covered) attribute which defines the coverage's coordinate reference system (CRS). In the

example the overall coverage is made up from the *Lat* and *Lon* axes which EPSG:4326 (aka WGS84) contributes, followed by a time axis provided via the OGC *AnsiDate* reference.

```
<GeneralGridCoverage xmlns="http://www.opengis.net/cis/1.1/gml" ...>
    <DomainSet>
        <GeneralGrid
            srsName="[ EPSG:4326,OGC:AnsiDate"]
            axisLabels="Lat Long date" uomLabels="deg deg d">
            <RegularAxis   axisLabel="Lat"  uomLabel="deg"
                lowerBound="40"  upperBound="60" resolution="10"/>
            <RegularAxis   axisLabel="Long" uomLabel="deg"
                lowerBound="-10" upperBound="10"  resolution="10"/>
            <IrregularAxis axisLabel="date" uomLabel="d">
                <C>2015-12-01</C>
                <C>2015-12-02</C>
                <C>2016-00-12</C>
                <C>2016-05-01</C>
            </IrregularAxis>
            <GridLimits srsName=" OGC:Index3D" axisLabels="i j k">
                <IndexAxis axisLabel="i"
                    lowerBound="0" upperBound="2"/>
                <IndexAxis axisLabel="j"
                    lowerBound="0" upperBound="2"/>
                <IndexAxis axisLabel="k"
                    lowerBound="0" upperBound="2"/>
            </GridLimits>
        </GeneralGrid>
    </DomainSet>
    <RangeType> ... </RangeType>
    <RangeSet> ... </RangeSet>
    <Metadata> … </Metadata>
</GeneralGridCoverage>
```

Notably, OGC GML does not support date strings for coordinates, only numbers. After multi-year discussions between the Web Coverage Service and the GML working groups the GML Standards Working Group (SWG) did not extend the coordinate format to include strings as requested by the WCS SWG. For this and other overly restrictive settings in GML, the WCS SWG departed from GML with CIS 1.1 and defined an XML schema more suitable for non-numerical coordinates. Meanwhile, in the latest version of ISO 19111:2019 Geographic information Referencing by coordinates [6], date/time strings finally are included, but the effort of changing the coverage XML schema to achieve the same effect via 19119 is currently not being proposed.

Range values. For storage, these values need to be linearized following one of many possible schemes, but this is an implementation detail of the particular representation chosen and does not affect the fact that coordinates are determined by the coverage axes. There is also the question about how the direct positions of the domain set are connected to their respective values. There are several ways of achieving this, as follows.

- The domain set, together with a sequentialization rule (which is ignored here), indicates a sequence of direct positions. The sequence in the range set follows this pattern.
- Domain and range sets are stored interleaved, as a sequence of coordinate/value pairs so that the correspondence is clear.
- Domain and range sets are tiled or partitioned into smaller parts. Inside each tile any of the above techniques can be used.

- Sometimes the domain set is not explicitly available, but just some information to derive the underlying grid. A typical case is a sensor model which stores Ground Control Points out of which the sensor model generates the grid coordinates for the range values.

Range type. A coverage's range type captures the semantics of the range set values. The definition of range is based on SWE Common so that sensor data can be transformed into coverages without information loss, thereby enabling seamless service chains from upstream data acquisition (e.g., through OGC Sensor Observation Service Standard) to downstream analysis-ready user services (such as OGC WMS, WCS, and WCPS). Notably, the range type can go far beyond just a datatype indicator (such as *integer* versus *float*). For example, unit of measure, accuracy, nil values, and the semantics (by way of a URL reference), and more information can be provided with a range type, thereby accurately describing the meaning of the values. The following is an example range type definition for panchromatic optical data, encoded in GML:

```
<GeneralGridCoverage xmlns="http://www.opengis.net/cis/1.1/gml" ...>
    <DomainSet> … </DomainSet>
    <RangeType>
        <swe:field name="panchromatic">
        <swe:Quantity definition=
                 "http://opengis.net/def/property/OGC/0/Radiance">
                <swe:description>panchromatic sensor</swe:description>
                <swe:NilValues>
                    <swe:nilValue reason=
"http://www.opengis.net/def/nil/OGC/0/AboveDetectionRange">
                        255
                     </swe:nilValue>
                </swe:NilValues>
                <swe:uom code="W.m-2.sr-1.nm-1"/>
            </swe:Quantity>
        </swe:field>
    </RangeType>
    <RangeSet> ... </RangeSet>
    <Metadata> … </Metadata>
</GeneralGridCoverage>
```

Metadata. This optional part is left unspecified in the CIS Standard and can contain any number of literally anything, (in XML *xs:any*). In addition to domain set and range type, the mandatory technical metadata of a coverage, these optional metadata are completely application dependent. Of course, the coverage user cannot understand the metadata, but the metadata will duly be transported so that the connection between data and metadata is preserved. One example of such metadata is given by the European INSPIRE legal framework for a common Spatial Data Infrastructure. INSPIRE prescribes canonical metadata for each object following a specific schema. This demonstration showcases use of INSPIRE metadata. Note the "any number:" different applications may add their own metadata, and each application in practice would only look at those metadata slots it recognizes, ignoring all others.

Returning to the discussion of CRSs and the *srsName* attribute, a coordinate is meaningless if there is no indication about the reference system in which it is expressed. A value of 42 — is that degrees (referring to what datum?), meters, years since epoch, or million years backwards? All that information is provided with the CRS.

As per an OGC Naming Authority (OGC-NA) decision, CRSs shall be expressed in URLs, and that is what is found in the *srsName* attribute. These URLs resolve through a special service operated

by OGC providing definitions for CRSs such as http://www.opengis.net/def/crs/EPSG/0/4326 and http://www.opengis.net/def/crs/OGC/0/AnsiDate.

A so-called CRS resolver service, running an open-source implementation by Constructor University, delivers the CRS details [9]. As this results in quite unwieldy URLs, OGC has resolved that CURIEs alternatively can be used everywhere in place of URLs. The CRSs above then can be written as [*EPSG:4326*] and [*OGC:AnsiDate*].

Such CRSs can be plugged together to define higher-dimensional spaces. The EPSG catalog is large, but preparing all possible axis combinations is not feasible. Therefore, and following ISO 19111-2, CRS and axis composition is provided where the base URL ends with *crs-compound*, followed by an ordered list of component CRSs and axis. This is the structure shown in the *srsName* attribute before, better digestible with a slight reformatting:

```
http://www.opengis.net/def/crs-compound
  ?  1=http://www.opengis.net/def/crs/EPSG/0/4326
  & 2=http://www.opengis.net/def/crs/OGC/0/AnsiDate
```

The alternative CURIE shorthand notation for this example is [*EPSG:4326,OGC:AnsiDate*].

Above is the information necessary to understand the concept "domain set." But there is also the following consideration: Services would be slowed down considerably if, for each coverage decoding, the application first needs to retrieve the CRS definition. Actually, not all the information is needed — the most important are the axis names and units of measure. The *axisLabels="Lat Long date"* and *uomLabels="deg deg d"* attributes provide this excerpt directly. Additionally, the axis labels define the sequence of axes. The axis sequence ambiguity actually is a problem recurring in GeoJSON and other OGC services where axis order is implicitly assumed. For example, the GeoJSON specification [11] states: "The coordinate reference system for all GeoJSON coordinates is [...] WGS 84." [11] also states that "the use of alternative coordinate reference systems [...] has been removed from this version of the specification." The RFC suggests that "where all involved parties have a prior arrangement, alternative coordinate reference systems can be used without risk of data being misinterpreted." This means that information about the CRS used must be transported outside the GeoJSON file which, consequently, is no longer self-contained. Hence, a conclusion is reached that GeoJSON may not be viable for use in ARD-ready applications and content.

B.5. Coverage Processing

For the sake of the discussion in the next section, a very brief introduction to ISO 19123-3 [58] and its geo datacube query language, also known as Web Coverage Processing Service (WCPS), is presented. Strictly speaking, WCPS defines a higher, abstract level language and not a concrete service as it is agnostic of the underlying protocol. Such protocols are provided in the OGC WCS Processing Standard [13]. This separation of concerns allows discussing concepts in an API-independent manner.

The current version of WCPS addresses the most widely used coverage subtype, regular and irregular grid coverages, also known as raster data or, more generally, datacubes. The concept of queryable datacube services was first introduced in [15] and refined in [17]. The concept follows the general Big Data definition of "data too big to download" therefore adopts an approach of

"shipping code to data." WCPS is embedded in a long tradition of special-purpose languages for data analysis, with SQL as its most prominent example. Having a special data language has several advantages: more concisely fitting the task, thus leading to more compact, high-level phrasing than with general programming languages; safe in evaluation, thus less prone to attacks than general programming languages; server-side optimization; and several more.

WCPS as a datacube language conveys several key properties as follows.

- The datacube model, OGC Coverage Implementation Schema [45], is embedded — no particular structure definitions have to be made. The language already syntactically enforces succinct and complete writeup (which means automatic query checking, syntax highlighting, and editing hints are possible, as provided with the rasdaman query editor).
- Common operations are readily available and easy to formulate, such as Tomlin's Map Algebra categories (local, focal, zonal, global) and many common Tensor Algebra operations reaching up to, for example, the Discrete Fourier Transform.
- There is a formal basis of static and dynamic semantics laid down in the WCPS Standard so that all implementations are guaranteed to return the exact same result. In contrast, OGC Environmental Data Retrieval (EDR) [21] which also knows a datacube subsetting request can return literally "anything" which makes client design and development potentially rather difficult.
- Some dangerous constructs — such as explicit loops with unverifiable termination — are unavailable, thereby disallowing a class of denial-of-service attacks. A WCPS query can get a "price tag" of how many data reads are required, how big the result will be, and how much processing is involved prior to executing the request.
- The syntax is close to the FLOWR expressions of XQuery [23] so that both can be integrated, forming a unifying query model for both data and metadata, supporting hierarchically structured XML or JSON.

These design rationales differentiate WCPS from the family of raster processing languages which typically are designed in the spirit of general programming languages, such as the xarray python library [30], Matlab [32], or IDL [33]. In a nutshell, WCPS provides a crisp, concise geo datacube language for client-side ease-of-use and server-side evaluation and optimization.

The WCPS language works as follows. At the core, it consists of a *for* and a *return* statement. In the *for* part, variables are defined which iterate over coverage lists. For example,

```
for $c in ( A, B, C )
```

would assign coverages *A*, *B*, and *C* to variable *$c* in turn. This alone does not do anything because a way to generate output is lacking. Output generation is accomplished with the *return* clause. In the trivial case below it delivers a constant number:

```
return 42
```

If the goal is to download all three coverages from above this can be written as follows, selecting the preferred output format (assuming it technically can hold these coverages):

```
for $c in ( A, B, C )
return encode( $c, "png" )
```

Now pixelwise, processing can be started where all the operators are known: arithmetic, logical, exponential, logarithmic, trigonometric, case distinction, etc. While this works, it will download the complete coverage. It is better to add subsetting to extract a region of interest (note that the sequence of axes in the query does not matter):

```
for $c in ( A, B, C )
return
  encode( $c [ date( "2018-05-22" ),
      E( 332796 : 380817 ),
      N( 6029000 : 6055000 )
      ],
     "png"
  )
```

Notably, for each axis not mentioned, the full extent is used while subsetting is applied on the datacube axes in the query which allows writing queries that are to dimensions not of interest. For example, a timeseries analysis query would look perfectly the same for 1D sensor timeseries, 3D image timeseries, and 4D climate and weather timeseries.

So far, each coverage has been processed in isolation. Data fusion is possible through “nested loops”:

```
for $c in ( A ),
  $d in ( B )
return encode( $c + $d, "png" )
```

Aggregation plays an important role for reducing the volume of data transported to the client. With the common aggregation operators — in WCPS called “condensers” — queries like the following are possible (note that no format encoding is needed, numbers are returned in ASCII):

```
for $c in ( A )
return max( $c )
```

As a final example, the following WCPS query computes the Inverted Red-Edge Chlorophyll Index (IRECI) on a selected space/time region, performs contrast reduction for visualization, and delivers the result reprojected to EPSG:4326:

```
for $c in (S2_L2A_32633_B07_60m),
   $d in (S2_L2A_32633_B04_60m),
   $e in (S2_L2A_32633_B05_60m),
   $f in (S2_L2A_32633_B06_60m)
let $sub := [ date("2018-05-22"),
                    E(332796:380817),
                    N(6029000:6055000) ]
return
  encode(
   crsTransform(
    ( $c - $d ) / ( $e / $f ) [ $sub ],
    { E: " EPSG:4326", N: “EPSG:4326” }
   ) / 50,
   "png",
  )
```

For more detail, refer to the WCPS tutorial provided by EarthServer [59][73].

Such queries can be sent via the Web through the WCS *ProcessCoverages* request, with a structure as shown below where {*wcps-expression*} represents the WCPS query sent:

```
https://acme.com/rasdaman/ows
  ? SERVICE=WCS&VERSION=2.0
```

```
& REQUEST=ProcessCoverages
& QUERY={wcps-expression}
```

In the rasdaman server, for example, queries undergo highly effective optimization prior to execution, including parallelization among cores available and distributed processing (in case of federated queries).

B.6. Section 3 — ARD Obstacles in Coverages

In this section, shortfalls of ARD in coverages are inspected. The discussion is based on both conceptual considerations obtained from manifold standards development discussions as well as empirical experience from projects dealing with datacubes as a special, yet prevailing, category of coverages. The discussion below is loosely grouped into data and processing issues, acknowledging that both are tightly intertwined and cannot always be separated completely.

B.6.1. Data

This section is written with the coverage structure in mind, consisting of domain, range, range type, and metadata. Generally, it can be observed that scientists and developers so far have put more effort into the design of the domain (such as coordinate handling, at least horizontally) than in the measured/generated values of the range.

B.6.1.1. Pixel-in-X

A recurring discussion on geo raster data is whether the pixels are assumed to sit in the center or a corner of the cell. The root cause is a gross misinterpretation of diagrams.

In a 2D diagram the coordinate axes determine the position of points drawn in the diagram. Auxiliary lines go through the integer values or some other meaningful line spacing. Crossing auxiliary lines encloses areas, but in mathematics this has no meaning. If points are to be plotted at coordinates where auxiliary lines cross, then the points will invariably be plotted at the cross-section of two lines.

In programming, arrays (which are used to ultimately store the pixels) are diagrammatically usually shown as a lineup of boxes where each box symbolizes a storage area that holds some value. The borders in this diagram have no meaning.

Figure B.3 puts both types of diagrams next to each other. Note how in the left Cartesian diagram numbers (i.e., coordinates) are associated with the lines whereas in the right array diagram numbers (i.e., storage addresses) are associated with the boxes. Combining both concepts results in the notion of pixels that (i) have been acquired for some geographic point (left interpretation) and (ii) have been stored in some memory cell.

This has caused confusion among some geo scientists who started mixing both concepts and coming up with the idea that a pixel could sit in the “center” of the cell, leading (with regular grids) to a half-pixel offset of the pixel’s direct position. Consequently, the scientists differentiate

between “pixel-in-center” and “pixel-in-corner” often making some silent assumption that the corner under consideration is the upper-left one while for *d* dimensional data *2d* corners exist.

This interpretation is factually wrong. On the level of abstraction of gridded geographic data there is no concept of a “cell” in the array sense, just direct positions (in coverage terminology) sitting at the given real-world coordinates. Each sensor delivers data for the given coordinates, meaning at the crossing of the auxiliary lines for these coordinates. If the value was assumed to sit, say, “between x_i and x_{i+1}”, then it would have the direct position coordinates given by $(x_i + x_{i+1})/2$. On a side note, some formats offer slots to store a local reference frame which could capture such an offset.

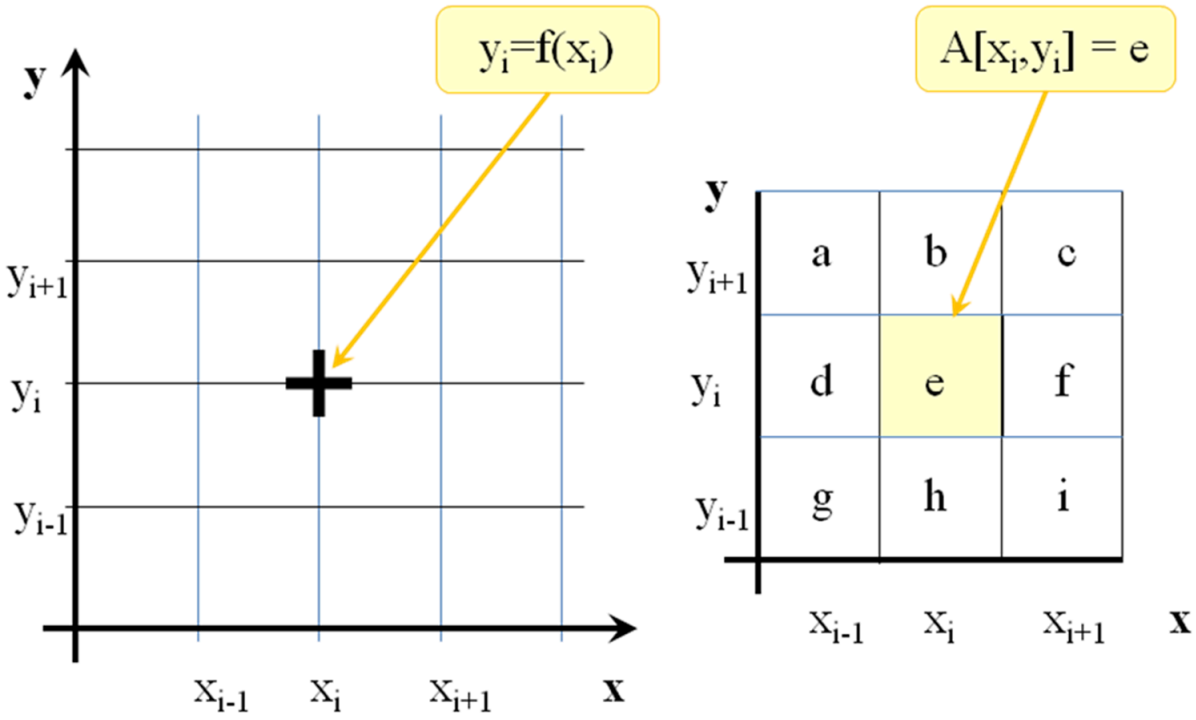

Figure B.3 — Cartesian diagrams (left) versus array symbolization (right)

The situation is aggravated by the lack of indicating the pixel-in-X assumption in some data formats. Combining datasets (including a background map) with different assumptions can lead to unwanted half-pixel shifts, potentially leading to dramatic deviations in an analysis. Discussions of how to handle this appear periodically on the Web, such as on gdal-dev where a contributor recently looked at the NetCDF format and Proba-V satellite data observing that wrong interpretation “could be dangerous or at least inconvenient. In my case I have a Python code that does zonal statistic and when a geometry is converted to a little raster binary mask, it does not match the same area in the ProbaV data and the results is a crash in Fiji because ReadRaster goes overboard or lost of NDVI data in other regions. The half pixel shift grows as you go east.”

This offset issue becomes even more spectacular when applied beyond the normally considered horizontal axes. In x/y/z coverages including a height or bathymetry axis there is no generally accepted convention that would determine a half-voxel offset upward or downward.

Similar questions arise on the temporal axis. Consider an example where some coverage with temporal resolution 1 year provides weather observations every January. Naturally, it would be expected that data "sit" in January. A pixel-in-center assumption would have a half-resolution offset, so 6 months. Hence, the weather observations would be attributed to June, rather than January.

Fortunately, the pixel-in-X discussion has not yet reached vertical and temporal axes and remains an item of discussion and confusion on the horizontal axes. Another difficulty is when non-regular grids come into play. Gridded coverages can be of the regular structure assumed so far (Figure B.4 left), or can be irregular of several types (Figure B.4 center and right).

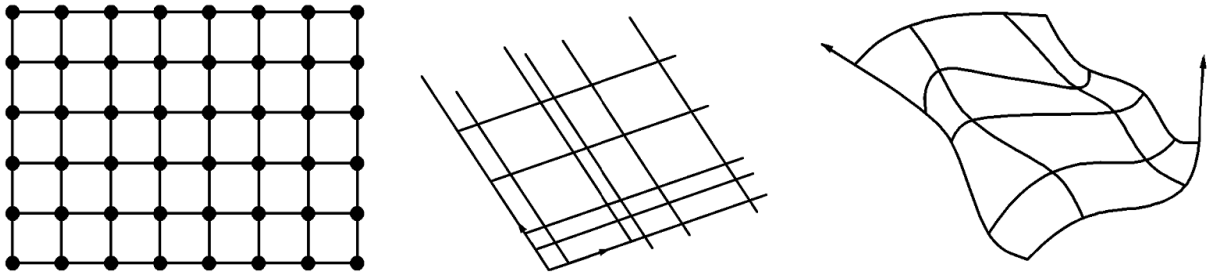

Figure B.4 — Sample regular and irregular grid coverage types [26]

The recommendation is to improve education on this misconception and always assume a pixel sitting at the coordinate indicated, with no offset to anywhere. Course material should be developed and distributed widely which discusses pixel-in-X in particular on 4D spatiotemporal coverages.

B.6.1.2. Pixel-is-X

Another issue centers around the question whether a pixel is considered a point (named "pixel-is-point") or an area "around" that point (named pixel-is-area"). Typically, in regular grids this area would stretch halfway to the neighbor positions (Figure B.5). Motivation is that optical sensors average light values over an area while elevation data represent a sample valid at exactly (and only) at the measured point.

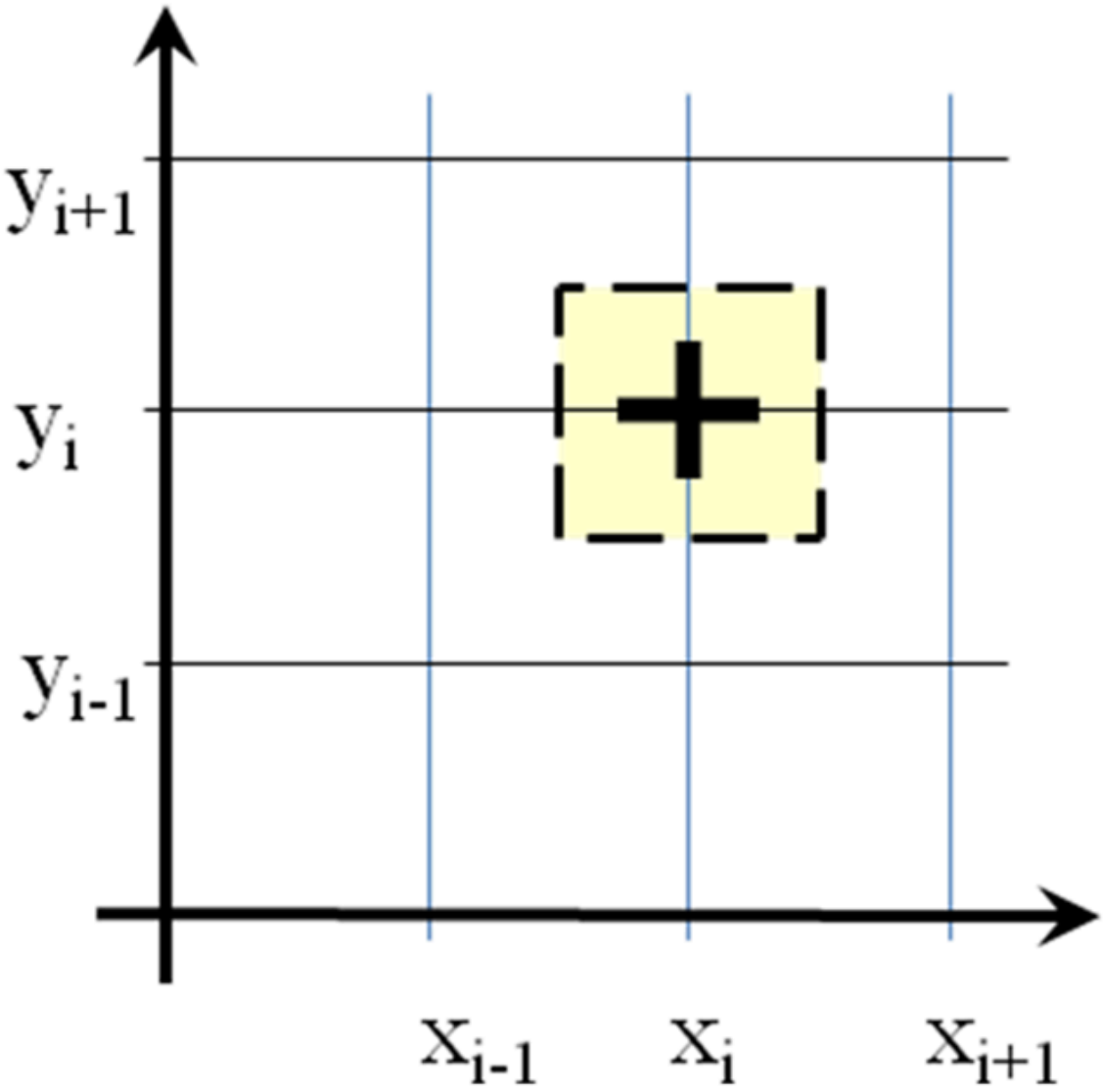

Figure B.5 — Pixel-is-area perception

There is no commonly defined handling. Users are expected to understand and interpret data appropriately. For example, the United States Geological Survey (USGS) points out [34] that Landsat uses a pixel-is-point interpretation and warns that some tools assume pixel-is-area designation. According to the above argument, however, Landsat should have a designation pixel-is-area.

To investigate this further, the characteristics of a sensor are considered. Details like viewing angles of a sensor are ignored, assuming an orthorectified, radiometrically corrected situation for simplicity. Likewise, the sensor instrument structure and its common variants such as pushbroom, whiskbroom, array, etc., are ignored.

Regardless of whether sensors are single or combined (e.g., CCD arrays), each single sensor has a so-called field of view (FOV) which is defined as the maximum angle of view that a sensor can

effectively collect signals (such as photons). The width on the ground corresponding to the FOV is called the *swath width*.

A sensor will collect signals over some time over its FOV, resulting in the final pixel value. This FOV, due to the physical nature of the sensor, will be a cone which is narrow close to the sensor and wider as distance increases. The area from which a sensor captures signals can be described through an integral over the FOV area, which is circular or more generally elliptic.

In any case, the FOV cone does not "know" about the neighbor sensors, and by no means can it be assumed that signal capture stops midway between the FOV centers (Figure B.6 left), or conversely that every point in the swath width contributes to some pixel in the sensor array (Figure B.6 right). Even in the case of no overlapping and maximum FOV relative to sensor distance, it noticed that there are areas that do not contribute to any pixel.

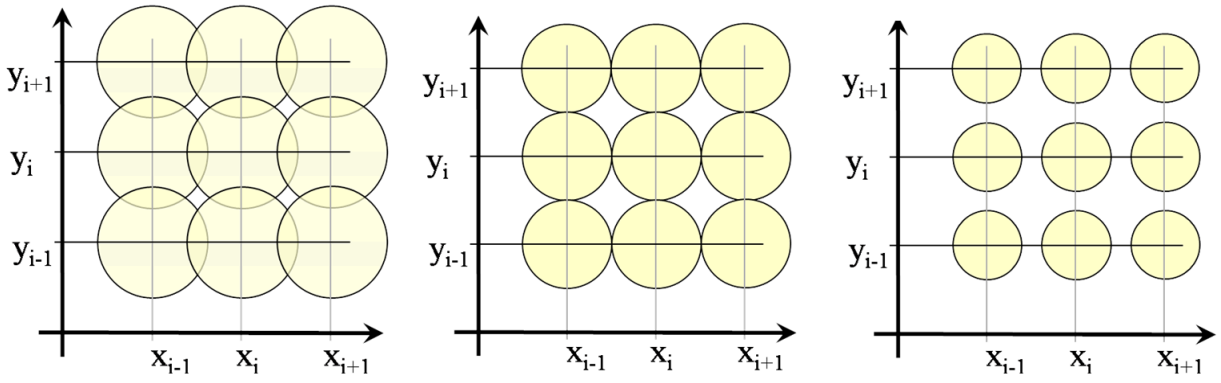

Figure B.6 — Various FOV situations

Therefore, the assumption is that a pixel can always represent an area of signal collection (pixel-is-area), even if the area shrinks to a singular point (pixel-is-point). The main point is that this differentiation should not lead to a substantially different treatment, such as introducing a map offset — at least not on the level of abstraction discussed and notwithstanding particular sensor characteristics to be captured.

The main difference remaining is interpolation: The DEM should not be interpolated while it is perfectly fine for the optical sensor. In this opinion, it is captured already in the interpolation information of the coverage where the DEM would have no interpolation associated whereas the optical sensor is fine with various sorts of interpolation.

At a next level of differentiation, interpolation might be restricted to the FOVs of the sensors, but that information is not captured in the current standards and would possibly lead to a massive increase in complexity.

Finally, an observation is that pixel-is-x exclusively focuses on horizontal space. No guidance was found on how to apply these concepts on vertical and temporal coordinates. Additionally, reservations from above concerning shifting by a half-resolution offset apply.

Notably, in the OGC there is some controversy on this topic[35].

Further investigation into the technical details of where in the coverage extent interpolation should be allowed is recommended which should not stop at area versus point discussions

but also drill more deeply into theory and practice. In any case, it is strongly recommended to abandon the idea of half-pixel shifts.

B.6.1.3. Units of Measure

A long-standing, yet not satisfactorily solved issue is the unit of measure (*uom*). CIS tentatively does not make any coverage-specific assumption about its representation. After all, *uom* is an overarching concept which should not be made specific to coverages. A string-valued attribute for identification, as adopted from OGC Sensor Web Enablement (SWE) Common where attributes *code* and *href*, is suggested. SWE requires either a code attribute using UCUM or a reference to an external unit definition.

However, diving into the corresponding XML schema, it appears that the underlying uom code type, *UnitReference*, is not defined. Instead, there is a *UnitReferencePropertyType* type whose attribute code is of type *UomSymbol*. Even in the examples provided with the SWE Common Standard, a wide range of usage is found:

- use of the code attribute for a symbol, name, abbreviation thereof, or a formula: *code="%"*, *code="mbar"*, *code="deg"*, *code="km/h"*; and
- use of the *xlink:href* attribute for a reference to a concept or an abbreviated unit name (for which, however, no registry exists).

```
xlink:href="http://www.opengis.net/def/uom/ISO-8601/0/Gregorian"
xlink:href="Cel"
```

In CIS, the range type — which is adopted from SWE Common — contains a corresponding element which is supposed to contain a UCUM unit or a URI, as per OGC convention. A URL would act as an identifier of some unit, and with a suitable mechanism behind this would allow even automatic transformation between units, say between feet and meters. Similarly, in the domain definition CIS remains agnostic and just requires units — such as geographic degrees or km, or temporal timestamps — which are understood in the context used.

Actually two issues can be identified. First, an incoherence in names can be observed. This happens when abbreviations are used. Even on very basic geographic units, uniformity is lacking as the following two random examples demonstrate.

- OGP at some time changed “long” for longitude to “lon” in the EPSG axis abbreviation to have a three character abbreviation like “lat”. This led to major issues in the handling of existing and new data as software usually did not automatically recognize that “long” and “lon” both denote the same axis. In the end, the solution in OGC was to allow in a coverage to freely name the axis labels, not tied to the name used in the EPSG CRS definition. These axes could be identified by position in the axis sequence — for example, CRS axes “Lat Lon h ansi” in a coverage can be named “lat long height date”. This decoupling allowed tools to keep legacy “long” and also to liberate applications from using cumbersome OGC names such as “ansi” for dates.
- Another issue of the same kind, which remains unsolved today, is the use of both “deg” and “degree” for the unit of measure along horizontal geographic axes.

Second, coverages are lacking a framework which would allow automatic conversion across the manifold units in practical use worldwide.

Obviously, there is a lack of standardization for such information, leading to nonhomogeneous or missing information in operational data. This has been observed already earlier [36]. Relevant candidates are inspected.

UCUM (Unified Code for Units of Measure) is a code system intended to include all units of measures being contemporarily used in international science, engineering, and business [37]. Units can be constructed following a mini language. For example, radiation per area might be expressed as "*W/cm2*" for Watts per square centimeter. If no unit is to be provided, by CIS convention, uom is "1" which in UCUM is expressed as "10^0". Notably, for complex physical phenomena UCUM can become quite involved, as the radiance definition "*W.m-2.sr-1.nm-1*" in one CIS schema example shows [45].

The European Environmental Agency (EEA) has published a dictionary with concepts relevant for environmental monitoring and land management [38]. This hierarchical classification system uses the Simple Knowledge Organization System (SKOS), a common data model for sharing and linking knowledge organization systems via the Web [39]. According to the website, this RDF-based data dictionary was last updated in 2015, so does not appear well maintained. Further, the power of ontologies seems not fully utilized as the main connectivity between terms is the hierarchy, without a fully-fledged explicit ontology network.

RAINBOW is the OGC Definitions Server [40] which offers information about concepts relevant in the standards ecosystem, with units of measure, CRSs, etc. While the service offers a wide range (including, for example, planetary science CRSs) RAINBOW is relatively new. For example, for temperature there is Kelvin, but neither Celsius nor Fahrenheit are available, and the EPSG section is completely empty. UCUM is referenced, but the corresponding concept is not connected to anything. Most important for this discussion, no information is provided that could be used for some automatic conversion, such as from seconds to minutes or meters to feet. A discussion has been launched recently, but subsequently has been closed again [41].

An alternative is QUDT [46], a unified architecture for the conceptual representation of quantities, quantity kinds, units, dimensions, and data types. It appears that QUDT is substantially more comprehensive than the uom ontology started by OGC. For example, temperature is linked to Kelvin, Fahrenheit, and Celsius, among others. DEG_C (Degree Celsius) contains information on the derivation from the canonical (Kelvin) unit indicating a conversion multiplier of 1.0 and a conversion offset of 273.15, in a machine-readable format. This indeed allows for an automatic conversion.

A potential downside of QUDT is the lengthy URIs which are unwieldy, at least for human users (and even more so with composite CRSs, see discussion of multi-dimensional CRSs). One could argue, though, that only the developers see such URIs. In operational applications, there is only machine-to-machine communication where the length of the notation does not matter.

Notably, QUDT is not a standard, but a research outcome maintained by a small (but rather active) group. Custom datatypes for measurements as well as on-the-fly support of arbitrary custom datatypes are implemented in an Apache Jena fork.

Recently, another RDF ecosystem of practically relevant datatypes has been suggested [47], including a microformat for UCUM syntax next to energy, force, pressure, speed, temperature, time, etc., which combines the exact notation of UCUM with the reasoning capabilities as in

QUDT. An implementation based on the Apache Jena reasoning framework is available. In future this might open the door for an automatic conversion across atomic and composite units.

A further recent activity has been launched by of CODATA with its Digital Representation of Units of Measurement (DRUM) task group [48]. The Bureau International des Poids et Mesures (BIPM) [49], an international organization established by the Metre Convention and managing body of the International System of Units (SI) and the international reference time scale (UTC), has started, during Testbed 19, a Forum on Metrology and Digitalization with one aim being a SI Digital Framework. Both CODATA and BIPM goals appear similar to what is discussed in this ER Annex, but no results are available yet. At the University of North Florida, a prototypical Units of Measure Interoperability Service [50] has been established aiming at automatic units translation.

Comparing the approaches discussed so far reveals the following findings.

- Common sense names and abbreviations are not suitable for unit understanding and automatic conversion.
- UCUM is exact and can be parsed for automatic handling by introducing a microsyntax which is nicely compact but can become unwieldy for human users.
- OGC's ontology approach is going the right direction but is not (yet) usable in its current preliminary state.
- QUDT is amenable to reasoning as well as to conversion arithmetics.

Bottom line, it is recommended adopting QUDT for use in the range type — best as a normative requirement, but at least as a good practice. OGC might be the right body for that. Also, OGC could host the specification and service as part of OGC's "community standards" process which has become popular recently.

B.6.1.4. Tiling

Partitioning of large arrays into smaller sub-arrays, also known as tiling or chunking, is a common technique for storing large arrays. Queries requiring some region of a gridded coverage will need to retrieve the required tiles, cut out the intersection with the region, and reassemble the parts into the result.

A comprehensive analysis of tiling in general has been provided by Furtado [51]. Particular patterns were categorized and cast into several parametrized strategies for easier handling by the tiling responsible, normally the data manager (Figure B.7). In rasdaman, these strategies have been implemented in a storage layout clause extending the insert and update statements. The engine will use the tiling directives for all new incoming data arranging the data according to the pattern specified. Via an update statement, a repartitioning is possible, something useful if new insights about the access patterns have emerged.

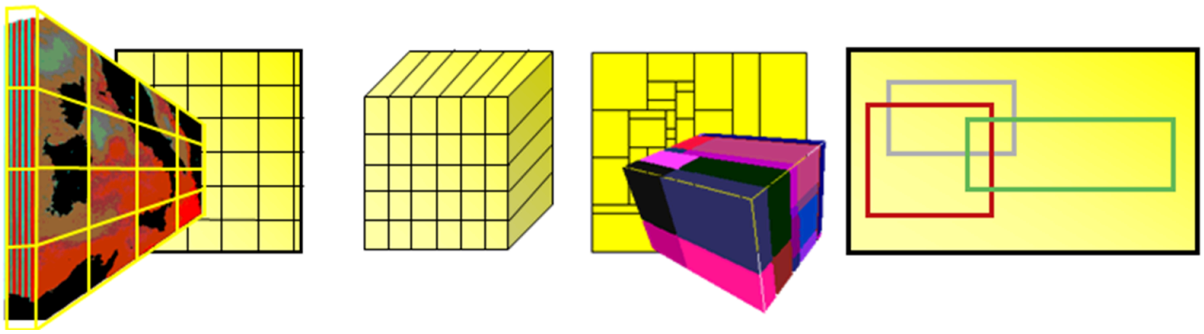

Figure B.7 — Sample tilings, after Furtado: from left, regular, aligned, non-aligned, area-of-interest strategy

The goal of tiling is always to minimize the number of disk accesses. Within limits it is of less importance how large the tiles are, but every I/O request costs. What is important is to minimize those I/O requests. If tiles are too small (like 4kB at some time used by Oracle and Esri, and about 100×100 pixels used in the 2D tiling of PostGIS Raster [52]) too many loads need to be performed, leading to significant performance losses as verified experimentally by Furtado. If tiles are too large, on the other hand, too much data are loaded and then dropped while cutting out the result. As such, tiling becomes a tuning parameter giving the administrator complete freedom to define any partitioning, or simply rely on some reasonable defaults.

Array databases hide the underlying storage structure of the datacubes, thus relieving users from a potentially challenging technicality which has proven useful in practice for simplifying access APIs and client codes. Some tools, though, require the user to reassemble the tiled and clipped data. In OGC standardization attempts exist, though, to make the tiling explicit.

OGC Web Map Tile Service (WMTS) "trades the flexibility of custom map rendering for the scalability possible by serving of static data (base maps) where the bounding box and scales have been constrained to discrete tiles"[19]. This reduced functionality is justified by easier implementation: "The fixed set of tiles allows for the implementation of a WMTS service using a web server that simply returns existing files." This tiling, though, is offered only for 2D x/y, any other potential dimensions a datacube can have are not considered. Elevation is considered in the 3D Tiles standard [53], however not time. On a side note, geographical parameters also are hardwired: Angles are fixed to radians, distances to meters.

OGC Abstract Topic 22 [54] specifies models for 2D tilings which, without further justification, are said to extend into higher dimensions. The 3D Tiles standard [53] is yet another specification addressing tiling. Further, in ISO a coverage tiling standard proposal has been submitted recently.

Notably, Abstract Topic 22 emphasized that a tile internally obeys a "homogeneity constraint" of following a single tessellation rule. A tile set making up one particular object is assumed to have a common CRS, uom, origin, and extent for all its tiles. In contrast, in the coverage world the coverage is the unit of homogeneity. All these tiling concepts and APIs have in common the introduction of a concept at the user level which is not required by the scientific task (such as EO analysis) and substantially complicates handling coverages.

One argument is that visual map clients request data in a tiled way and, therefore, providing tiles is natural. However, storage tiling and delivery tiling are two separate concerns. A client might get data in tiles, but a spatiotemporal access API should always allow user-selected arbitrary

bounding boxes and not constrain to some storage pattern the server finds useful. Finally, while visual clients like to access in tiles of their own chosen sizes, this certainly does not hold for analysis where, for example, the code for processing a 2D array should be a simple nested loop on a single, contiguous array. Everything else creates additional complexity, programming needs, and opportunities for introducing bugs — in short: it is not analysis-ready.

Notably, CIS supports tiled storage, but that is part of the internal data organization — neither WCS *GetCoverage* nor WCPS queries are aware of such internals. Both operate regardless of what the tiling of a coverage is like.

Therefore, it is ultimately claimed that for analysis-readiness tiling should be opaque and hidden from the user. If deemed necessary to standardize, then a separate tuning API might be defined which could fit well into the coverage administration API, WCS-T [55].

B.6.1.5. Coverage Metadata

The coverage metadata element contains any ancillary information beyond the canonical metadata fixed in the coverage structure. The corresponding XML schema part of CIS is as follows:

```
<complexType name="MetadataType">
    <sequence>
        <any namespace="##other" processContents="lax"
            minOccurs="0" maxOccurs="unbounded"/>
    </sequence>
</complexType>
```

Although by definition the CIS coverage metadata item can contain literally “anything” and, hence, is outside the scope of the Standard, other standards utilize the metadata in a normative manner. One example is the EU INSPIRE WCS standard which is based on CIS coverages with an INSPIRE-specific metadata contents.

Tools are expected to add further information. For example, rasdaman can add information about the contributing footprints of the original input files to allow tracing back the origin of pixel data if needed.

The INSPIRE Elevation Grid metadata are embedded as shown below, following the above metadata schema while honoring the INSPIRE schema.

```
<gmlcov:metadata>
    <el-covmd:ElevationGridCoverageMetadata xsi:schemaLocation="http:
//inspire.ec.europa.eu/schemas/el-covmd/4.0 http://schema.datacove.eu/
ElevationGridCoverageMetadata.xsd"
xmlns:el-covmd="http://inspire.ec.europa.eu/schemas/el-covmd/4.0" …>
        …
    </el-covmd:ElevationGridCoverageMetadata>
</gmlcov:metadata>
```

The following example extends this: three different, independent metadata records coexist in the coverage.

```
<GeneralGridCoverage xmlns="http://www.opengis.net/cis/1.1/gml" ...>
    <DomainSet> ...  </DomainSet>
    <RangeSet> ...  </RangeSet>
    <RangeType> ...  </RangeType>
```

```
    <Metadata>
        <el-covmd:ElevationGridCoverageMetadata
            xmlns:el-covmd=
            "http://inspire.ec.europa.eu/schemas/el-covmd/4.0" ...>
            ...
        </el-covmd:ElevationGridCoverageMetadata>
        <card4l:Card4lMetadata xmlns:card4l="..." ...>
            ...
        </card4l:Card4lMetadata>
        <special:MySpecialMetadata xmlns:special="..." ...>
            ...
        </special:MySpecialMetadata>
    </Metadata>
</GeneralGridCoverage>
```

How to combine various metadata items in this “bag” is not standardized, and hence tools will have varying issues in extracting specific information again. It would be helpful, at a minimum, to define a structure enabling the coexistence of metadata items without impact. This appears a relatively low-hanging fruit: by introducing a convention of parallel, independent sub-elements in the metadata slot, tools can provide and consume exactly those metadata the users are interested in, without getting confused by all the other slots that may be present. Simple path expressions would allow selecting such sections.

In rasdaman, for example, there are two such compartments already available. One is reserved for technical use by rasdaman, such as for metadata applying only to specific regions in a coverage. Another one is for metadata according to the EU INSPIRE standards.

A proposal is that OGC establishes a registry, maintained by the OGC Naming Authority (OGC-NA) and overseen by the Coverages SWG as is the case with other coverage-related registry information already, where extensions can be registered. Such an extension would consist of:

- a name not already registered;
- an XML namespace defining the structure; and
- any further pertinent information, such as description, examples, authoritative experts, etc.

Based on such registries, tools can extract the metadata of interest and be sure these are not “polluted” by other structures stored with the coverage, and as such ignore all other metadata items not of interest.

This has been discussed for XML, but it should be done in sync also for the JSON and RDF encodings which are also specified in the CIS standard.

B.6.2. Processing

After inspection of the coverage data structure, the focus now turns to coverage processing. As mentioned earlier, aspects sometimes are entangled so that processing is affected by data modeling, and vice versa.

B.6.2.1. Interpolation

In Earth sciences, coverage data in general describe spatially extended, time-varying physical phenomena, corresponding to the mathematical notion of a “field” in physics [61]. Sensors observing natural phenomena do not collect data only at the exact point of a direct position. Rather, date are integrated over some neighborhood of the direct position and present the result as the value at this position. Generally, these observations are aggregated over some region and time. Therefore, it may make sense to invert this quantization to discrete points and derive values for in-between positions from the direct positions available in the coverage.

Through interpolation, range values can be obtained for coordinates within the domain of a coverage which are not direct positions, usually by combining the values of several direct positions in the neighborhood of the coordinate location under inspection. Technically, interpolation is applied during resampling in reprojection and scaling operations where new direct positions are constructed which need to get values assigned.

As a first requirement, the domain must be in a CRS which allows expressing “in between” coordinates. Cartesian coordinates, for example, do not fulfill this while geographic and temporal coordinates do.

While computing such an interpolated value, an additional requirement is that the range type must support derivation of values. Only nearest-neighbor interpolation is always possible as it retains one of the original values. Further common interpolations, such as linear, quadratic, or cubic require additionally a continuous range data type supporting the interpolation math.

Obviously, the justification for interpolation cannot simply be deduced from the data type. Just because some range type is an integer does not preclude that new values can be derived through integer arithmetics. And just because the domain is continuous does not mean that interpolation can be applied freely. One example of a very careful handling of interpolation is kriging[62]. Kriging is a method similar to regression analysis developed originally for geostatistics (in particular, mining where underground data tend to be particularly sparse), but also applied in many other domains like hydrogeology, remote sensing, and astronomy. See [63] for an in-depth presentation of spatial data interpolation.

B.6.2.2. Image Pyramids

A special interoperability situation occurs in the context of image pyramids. Image pyramids are used by virtually every tool to speed up zoom access which is why pyramids are considered a form of processing but are actually an auxiliary data structure. When a user zooms out, then data get requested at a lower resolution to reduce client load and transfer times. For faster response the server caches lower-resolution versions of the data so that scaling loads and processes only from the next-closest higher resolution and not the high-volume base level. The problem is that while deriving the pyramid levels, interpolation may be applied. This raises the following concerns.

- With two scale steps being applied to the data, is the combined interpolation equivalent to a single scale operation, especially when considering numerical effects?

- During preparation of the pyramid when some interpolation gets applied, a choice must be made. If the user requests another interpolation method than the one used in pyramid building, then two different interpolation methods get applied to the data in turn, with unclear outcomes. Obviously, maintaining pyramid variants for different interpolations is unfeasible in face of the storage needs.

Obviously, this holds not only for a visual client's zoom-out, but also if the server uses pyramids for scaling operations during any other processing. Generalizing this idea, a coverage should only allow interpolation methods during retrieval and processing which are compatible with the interpolation methods applied during generation of this coverage.

In the case of Landsat 8, chosen as a random example, it is found in [64] that the TIRS bands 10-11 are collected at 100 meters but get resampled to 30 meters to match the OLI multispectral bands. Actually, handling different bands in satellite images with different methods is quite common, given that several different instruments contribute bands. For Landsat 8, resampling is reported to be Cubic Convolution.

B.6.2.3. Summarizability

It is well known in statistics that aggregation of values is meaningful only under particular conditions, and different types of aggregation are subject to different constraints.
A large body of research is available from OLAP and Statistical Databases, including [65][66][67][68][69][70][76][77][78]. These constraints are given by the type of data (the coverage range type), but also by the particular axis (the coverage domain), and aggregation constraints typically vary across the axes. Finally, the type of aggregation (count, min, max, sum, avg as well as modes, percentiles, etc.) is relevant. Examples for such impact factors include the following.

- On classification data, maximum and minimum are meaningful, but neither sum (as well as difference) nor average is. Counting of occurrences is allowed.
- Addition of timestamps is not meaningful — adding years 2023+2023 would result in the year 4046 which is likely not an intended result. Rather, it is necessary to differentiate between timestamps (where addition is not meaningful) and time periods (which can be added up and added to timestamps).
- To obtain the car throughput in an image timeseries it is incorrect to add the number of cars recognized in each time slice because cars driving through the area of interest likely will appear in several such slices.

The term *summarizability* has been introduced by Lenz and Shoshani [65] to denote whether some particular type of aggregation is meaningful. Categorial data, for example, are not amenable to summing up or averaging, but there are more cases. The paper lists a series of examples, such as in socio-economic data, "number of accident deaths by month" is summarizable whereas "number of children attending school by month" is not summarizable.

Mazon et al. [66] provide a review of summarizability issues with a focus on Online Analytical Processing (OLAP). In [67] non-, semi-, and fully-additive measures are discussed, and

beyond sums also averaging and rounding. Manifold approaches have been taken to ensure summarizability, including analytical [68], rule-based [69], and SQL integrity rules [70]. Inaccurate summary factors and the impact of inaccurate summary factors are treated in [76], providing practical rules on proper summation design. Altogether, there is a rich body of research and results available for business and statistical data.

Lenz and Shoshani [65] provide a categorization into flow (de-noting a cumulative effect over a period), stock (state at specific points in time), and value-per-unit, based on which they establish summation rules and compatibility matrices proving that these are necessary; no proof is provided for sufficiency. Kimball and Ross [77] adopt a slightly different perspective in distinguishing additive, semi-additive, and non-additive categories. On a side note, summarization does not preserve proportions [80], a fact which potentially can be relevant for model-based processing of EO data.

Niemi et al [78] combine these insights into the observation that a query can only produce meaningful results if (i) the aggregation operation is appropriate for the measure (in coverages: range values) and (ii) the measure is appropriate for the aggregation levels in the cube's dimensions (in coverages: axes).

Niemi et al [78] further establish a new categorization of dimension types and multi-dimensional structures, derive a measure categorization, give formal definitions for summarizability types based on the relational model of data, and then construct provably correct rules for summarization.

In the context of EO, direct aggregation is concerned as well as indirect aggregation performed in "zonal" operations like reprojection and scaling, plus "global" operations like radiometric corrections.

In the field of Earth data such considerations are at a very preliminary stage. OGC Sensor Web Enablement (SWE) Common [79] distinguishes an eclectic mix of *Count*, *CountRange*, *Category*, *CategoryRange*, *Quantity*, *QuantityRange*, *Time*, *TimeRange*, *Boolean*, and *Text* which does not appear helpful in providing summarizability guidance to tools.

More research is required to ensure that results can be transferred from the work mentioned or to perform any necessary adjustment. For example, if the data comprises wind directions taken at time intervals and there is a need to determine the prevailing wind during some period of time, then it is not appropriate to compute the sum or the arithmetic mean, but it is appropriate to compute the mode [78].

A goal should be to define a framework for categorizing data with respect to the data's summarizability which should allow finding provably correct constraints ideally derived automatically for some given coverage. In the SQL ecosystem, critical errors can be discovered automatically and, hence, avoided[81].

A first, admittedly simple and incomplete, step would be to reflect the well-known statistics categorization into categorical, ordinal, and numerical data by adding such an attribute into the coverage range type, plus corresponding protection of aggregation, interpolation, etc. With an increased understanding and better metadata such safeguarding might successively get increased.

B.6.2.4. Dimension Hierarchies

On dimensions, aggregation often is driven by the conceptual model of the axis. For example, over time it is common to aggregate up to days, up to months, or up to years. Aggregations might consider all data of some chosen interval or regular parts thereof, such as “every February.” A roll-up from days to weeks in OLAP sales data is mathematically very similar to a map zoom-out by a factor 7, maybe with the particularity that OLAP data regularly are null at weekends. It is typical to have several successive abstraction levels, commonly called dimension hierarchies or concept hierarchies.

Looking more closely, the time axis reveals a particularity: As weeks and months do not align in the calendar, OLAP time hierarchies are not strict (i.e., tree-like), but directed acyclic graphs (DAGs) where users can decide to roll up over weeks or over months. Figure B.8 schematically shows two dimensional hierarchies.

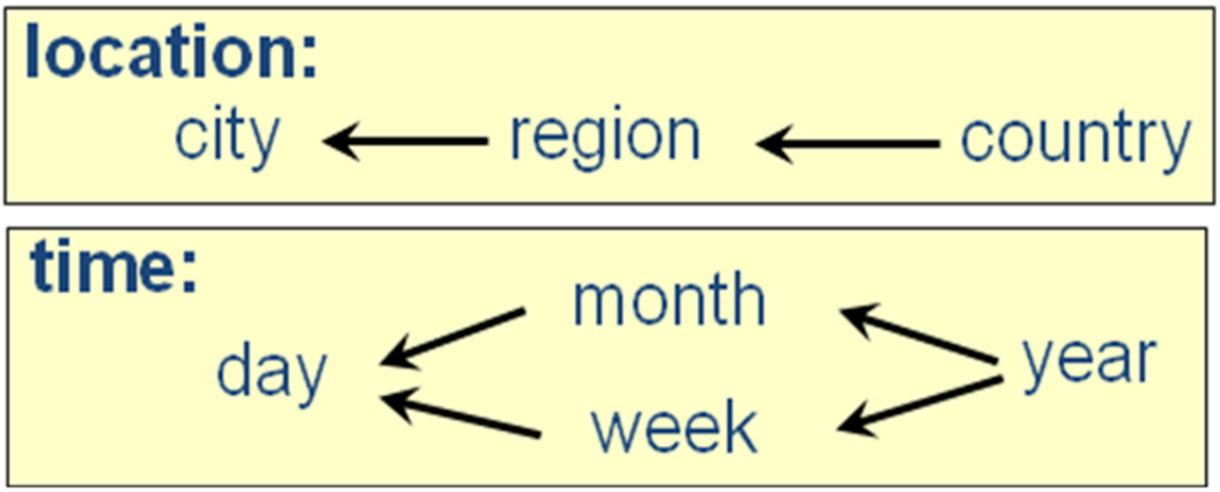

Figure B.8 — Sample dimension hierarchies for geographic names and time

For Earth data, use of the temporal axis is likewise found. However, spatial x/y/z axes differ in the zoom semantics. As opposed to many other axis types, these are continuous in aggregation, or, expressed in common terms, allow zoom by any real-valued factor. (This is not to be confused with image pyramids where the predefined pyramid members only act as an internal acceleration and optimization mechanism, but do not serve any conceptual purpose and are not visible to users.)

Altogether, dimension hierarchies are powerful concepts for adding more semantics to multi-dimensional data (including summarizability constraints, as discussed before). It is certainly worth looking into solutions in other domains, such as ISO SQL OLAP [82] and Microsoft MDX [83]. The recommendation, therefore, is to add dimension hierarchies to coverage axes and exploit their semantics in the GeoDataCube queries of ISO 19123-3 [58].

B.6.2.5. Validity and Reliability Masks

One way of characterizing validity of range values is to annotate every cell (pixel, voxel, etc.) with information about its validity which may consist of global metadata identical for all direct positions, or characterizations individual for direct positions.

For example, the SoilGrids service [84] offers an uncertainty layer which can be activated by the user for visual inspection (see Figure B.9).

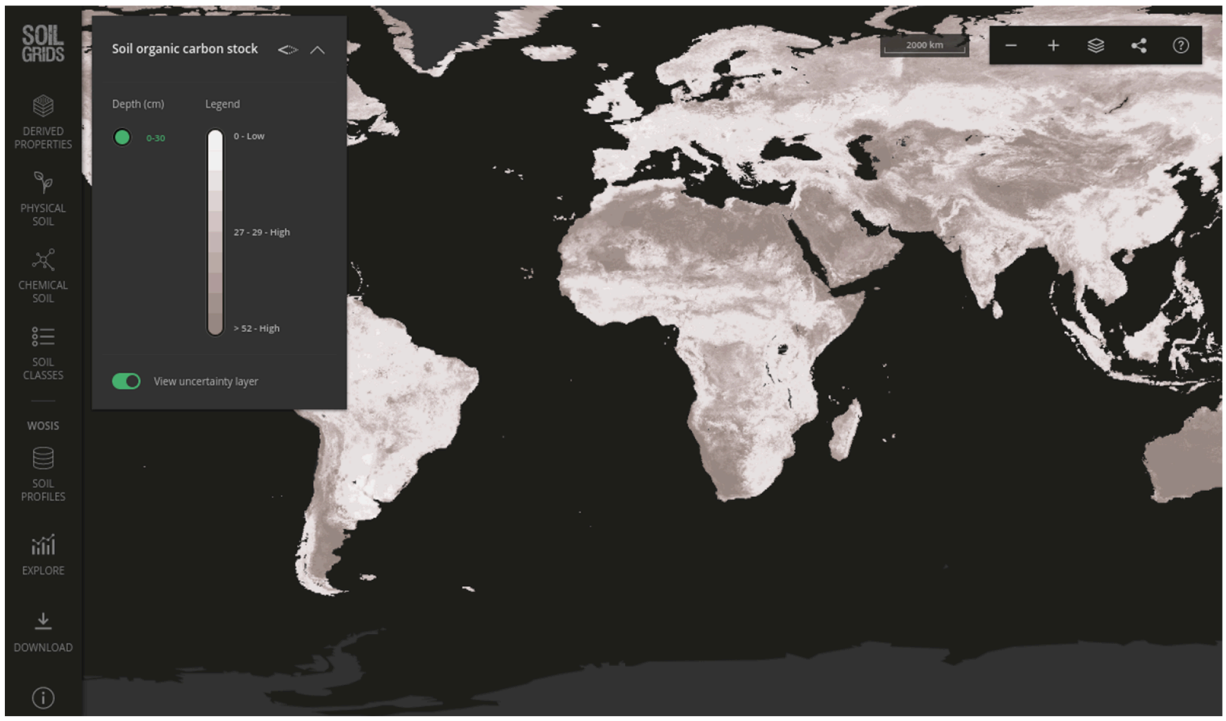

Figure B.9 — SoilGrid uncertainty layer visualization

The *l2gen* tool [85][86] which processes Level 1 data to Level 2 generates on the side a quality bit vector per pixel with a series of relevant information, including the following.

- Pixel is over land/over sea
- Sunglint
- Reflectance exceeds threshold
- Observed radiance very high or saturated
- Sensor view zenith angle exceeds threshold
- Solar zenith exceeds threshold
- Probable straylight contamination
- Probable cloud or ice contamination

- Atmospheric correction failure
- Very low water-leaving radiance
- Navigation failure
- Navigation quality is suspects

In WCPS, such masking can be used to disregard “bad” pixels, say, in an aggregation, expressed as

```
for $s in ( MyScene ), $m in ( MySceneMask )
return count( $s * ($m.sunglint = 0) )
```

In case of a statistical mask with, say, percent values between 0 and 100 a threshold might get applied in a query:

```
for $s in ( MyScene ), $m in ( MyScenePercentage )
return max( $s * ($m > 70) )
```

As can be seen, there is a much wider range of possible irregularities than just clouds and cloud shadows, which again requires some remote sensing experience to decide about relevance of individual flags for a given analysis task. As a Holy Grail, the system could determine which quality filters have to be applied for a meaningful result (and whether the remaining values still are sufficient).

Investigation is recommended on adding quality measures (as described or different) to coverages and investigate on automatically evaluating such quality measures in WCPS, maybe using ontologies. However, Machine Learning (ML) is not recommended.

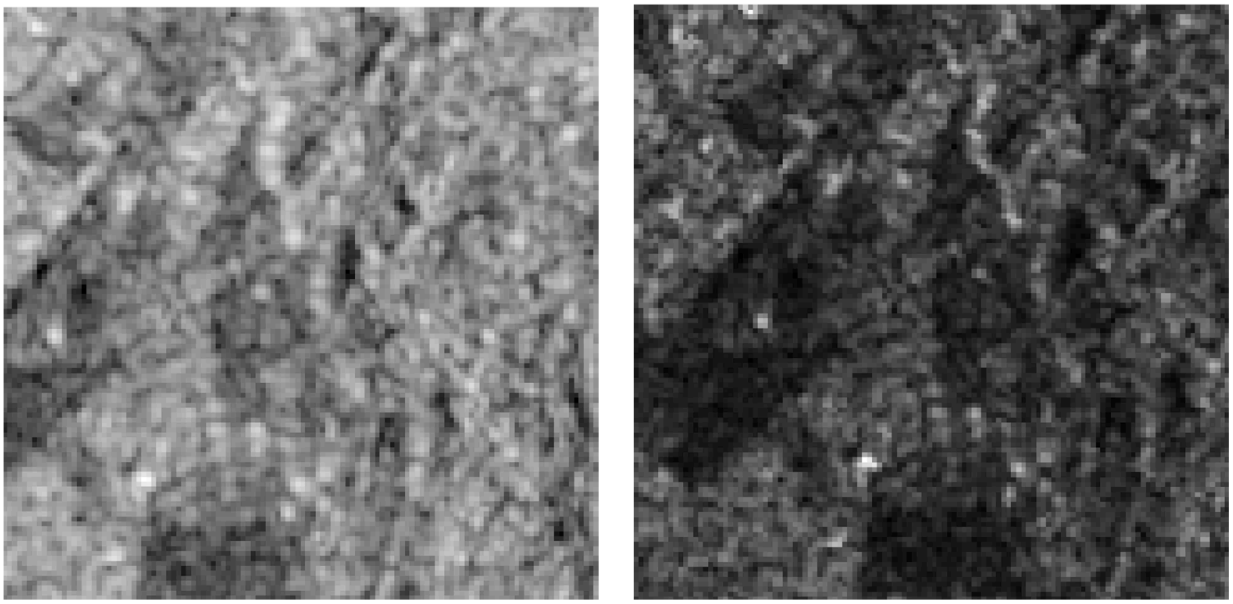

Figure B.10 — Sentinel-1 scenes delivered by ESA with different processing parameters applied

B.6.2.6. Product Provisioning Coherence

Data provenance is important information for assessing the usefulness of a particular data set for a given purpose. In particular, the processing applied can have an impact as studies [87] convincingly show. Even changes in software versions or processing parameters can have critical consequences. An example is given in Figure B.10 where both times Sentinel-1 radar satellite data are shown: left with an acquisition (and, hence, ESA processing) time between June 2017 and May 2018, right with a Sentinel-1 patch from 2023 extracted from the ESA Copernicus archive. Neural networks trained on the left-hand side image generation by experience fail miserably when applied to data of the generation shown on the right-hand side.

In the FAIRiCUBE project, several data sets were harvested from the European Environmental Agency (EEA). While building up datacubes from the the file sets, several inconsistencies were noted. In some cases, resolution got enhanced at some time. However, the dataset was continued to be advertised as the same and with the same name. In land use classification data, at some point in time new classes were introduced (fortunately none were deleted). In one dataset, the data semantics changes over time: The imperviousness dataset for the year 2018 has percentage values from 0 — 100% whereas for 2006 it has class numbers indicating 0-20%, 20-40 %, etc. Sometimes the legend was observed to change over time, and it remains to be investigated deeper whether this hints at deeper semantics changes in the data.

From the usability perspective and the coverage homogeneity idea, such changes establish new products — tools running automatically over the dataset time extent can generate disastrous errors without a chance of recognizing and reporting this.

Changes are not captured by the coverage constituents currently, and generally are hard to capture. Basically, a provenance record would need to capture not only the tools used, but also the software version and the concrete parameters used. However, this still leaves open the interpretation of how the changes affect the product properties.

Generally, the recommendation is to define processing levels rigorously enough so that measures for variation introduced by the processing get provided alongside with the result, or ideally variations are ruled out in the first place. Towards this Holy Grail, Strobl suggests a revision of the ubiquitously used processing levels [93]. The suggestion distinguishes processing along two “dimensions” (not related to coverage dimensions), measurand and spatiotemporal, which together span a 2D matrix of combinations where each field defines a possible processing status. This allows sorting the usual processing levels into the matrix, relating them, and spotting combinations not addressed by processing levels hitherto.

From a coverage standards perspective, Strobl’s measurand dimension relates to the coverage range set whereas the spatiotemporal dimension relates to the coverage domain set. However, while the disentangling of ARD impact factors is a good way forward there is no clear requirements analysis which could justify the distinction made between combination-ready, fusion-ready, analysis-ready, and inference-ready.

B.6.2.7. Numerical Effects

In the processing heavy EO world and the common use of floating-point numbers, a series of undesirable effects can occur which falsify results. Sometimes this influence is negligible, sometimes the result becomes completely wrong. It depends on the degree of deviation from the theoretically correct result, but also on the accuracy requirements of the problem to be solved.

One blatant example is defining null values as floating-point numbers. This is dangerous because (i) many numbers cannot be represented exactly (such as the beloved decimal-based fractions such as 0.1) and (ii) any operation applied can introduce inaccuracies. Stull [94] reports a simple example: add up 30 times 0.1 and you will get 2.9999993, rather than 3. This is with 32-bit floats; with 64-bit floats the result is 3.0000000000000013.

Now imagine a null value is defined as 3. Both computations would miss the null value. One may consider this example un-realistic, but it is felt that it is not: During import of data into a service it is common to preprocess from raw to say Level 2 (see discussion above), with all the numerical issues. Obviously, computations may accidentally distort the outcome and thereby generate a value that is treated as non-null, for example, in aggregations. Conversely, values which should count as non-null might accidentally get mapped to null.

This constitutes a huge and dangerous source of errors which often do not receive the attention required, among others, because data scientists rarely have an education in Numerical Mathematics as computer scientists usually get during undergraduate studies. In general, a rather naïve handling of floating-point numbers is often observed.

One of the insights Numerics teaches is interval arithmetics, used for floating-point numbers [95][96]. In this approach, numbers are considered inexact in the sense that the numbers do not have s single known value, but “smear” out over an interval of validity, in practice given by an interval. This can capture the behavior of floating-point numbers with finite accuracy. Error propagation rules can help estimating the ultimate accuracy of float computations [97].

While interval computations are significantly more involved (and require significant skills on the developer side) and drag down performance as compared to point Numerics, interval computations provide an automatic precision control which supports assessing processing chains, among others.

While tools generally do not implement this approach due to the above obstacles mentioned, these concepts might at least give hints on proper handling of finite-accuracy numbers. In rasdaman, for example, null intervals can be defined [98] which, when wisely used, can capture numerical deviations in the proximity of the conceptually foreseen null values. In the above example, a null interval around 3.0 might be defined as [2.9999993 : 3.0000007], based on the known processor accuracy. If a null value is defined as 3 then an equality comparison will fail in both cases.

The situation is further complicated through a parallel execution of code on heterogeneous hardware which can happen locally when CPU and GPU employ slightly divergent accuracies (and, hence, arithmetics), and even more so with distributed processing where, in the extreme case, supercomputer and edge devices jointly solve processing tasks.

Generally, for enhanced confidence, data tools should employ interval arithmetics so that the tools can provide error estimates. If provided by all tools along a processing chain, then the chain itself could be characterized in its overall error estimates. One starting point is shown with rasdaman where null intervals can be defined [98] and are strongly recommended on float and double range types.

B.6.3. Practical Examples

In this section, larger real-life examples are provided to put the ARD investigation in context.

B.6.3.1. Service Quality Parameters

There are manifold general service quality parameters such as bandwidth, availability, and response time from coverage-specific quality parameters. Only the latter are of concern in this context which supports concentrating on the coverage structure by inspecting every normative component for accuracy and fitness considerations. In the following incomplete walk-through, qualitative and quantitative criteria for fitness of purpose of a given coverage for some applications are addressed. Further research should conduct a systematic review with ideally formal arguments for correctness and completeness.

- Domain
  - Match with the area of interest: are the requested dimensions present, such as a temporal axis for timeseries analysis?
  - Is spatiotemporal resolution sufficient for the purpose?
  - Are data provided in a CRS which supports transformation into one of the CRSs understood by the client?
- Range type
  - Are data in some unit that can be converted to a unit understandable by the client? For example, feet can be converted to meters, but flight (pressure) levels cannot easily be converted into hekto-Pascal (hPa).
  - Accuracy: Are the range values exact enough for the analysis to be performed so that the targeted results can be achieved? In the simplest case, this would become a parameter in the range type. However, accuracy may vary across the coverage extent, so other means might be appropriate such as accuracy masks. Note that this is related to null values (a null value has an accuracy of 0), but not identical (non-null values can vary wildly in their accuracy).
  - Plausibility: With what probability can the data be trusted? Again, this can be represented by a single value or, in the other extreme, by a twin coverage containing a probability for every range value. Measures can be binary or percents, among others (such as the floating-point accuracy or smaller).

- Range:
  - Are there sufficient non-null values in the area of interest, in all dimensions?
- Processing:
  - Are all aggregation and other functions applicable on the type of data (categorial, ordinal, numerical)?
  - What interpolation must be applied, if any? Is a suitable interpolation method admitted by the coverage?

The key question is always: is the given coverage good enough for answering the question the client has — where "good" can mean a variety of things, as discussed?

Which, however, prompts for a counterpart on the client side: What is the required minimum accuracy for a given task? This question has not been addressed adequately to the best of knowledge. More generally, this addresses service quality parameters which actually may have strong mutual dependencies. For example, JPEG image quality may be reduced to obtain smaller files which transmit faster. This quality reduction (in this case: by filtering out higher spectral frequencies, but also introducing artifacts sometimes) may or may not be acceptable.

Therefore, the development of APIs is suggested which, in parallel to formulating extraction and processing requests, enable clients to express quality criteria. Corresponding frameworks exist already in the field of cloud computing [99][100] including, for example, agent-based fuzzy approaches [101]. For semantic handling in the coverage context these need to be augmented with coverage-specific parameters, such as the ones discussed above. Optimal flexibility is given by providing a language rather than just a list of parameters so that conditions of any complexity can be built easily. For the evaluation of such quality requests against the quality offered by the service — essentially, a multi-attribute prioritization decision making problem — methods such as dual hesitant fuzzy graphs [102] may turn out promising.

B.6.3.2. Coverage Fusion

As discussed in the Introduction of this Annex, data fusion is an excellent study field for ARD. Assume two georeferenced grid coverages, A and B. The goal is to obtain the mean square difference over all cells. In WCPS, this is expressed as:

```
for $a in (A), $b in (B)
return avg( pow( $a - $b, 2 ) )
```

If this is done as part of some large, unsupervised analysis process several facets must be investigated by the code (rather than a human), simply derived from the coverage definition. First, each coverage needs to undergo individual scrutiny as described in the previous subsection. Additionally, the following criteria are spotted; as before, rigorous research should establish a formally correct and complete criteria set.

- Domain

- Do A and B share the same dimension and axes? Otherwise, the coordinate tuple is ambiguous.
- Do both share the same CRS? If a reprojection is required, this could be performed automatically.
- Do further constituents fit in a compatible way, such as: coverage type; categorial vs ordinal vs numerical; etc. Again, in some cases an automatic adjustment may be possible.
- In case of resampling, are the interpolation methods of both coverages compatible?

- Range type
  - Through type casting, different numerical types (integer, float, complex) get adjusted automatically. This is part of the WCPS Standard already.
  - Determine proper treatment of null values (which may be different for A and B).
  - Verify the units of measure for compatibility between A and B. If there is an opportunity for automatic unit conversion, apply that. Further investigation should determine how much of this inter-coverage harmonization can be automated.

B.6.3.3. ML

From the perspective of this discussion, model-based prediction through Neural Networks constitutes just another "processing" request to a service. For example, in the UDF-enhanced WCPS of rasdaman a prediction can be invoked as a query function which internally the server, as part of the query processing sends over to pytorch. The following example takes some user-specified patch from a datacube built from Sentinel-2 optical satellite data, and applies a crop model to it.

```
for $c in (Sentinel_2a), $m in (CropModel)
return encode( nn.predict( $c[…], $m ),"tiff" )
```

Technically, function predict() from package nn (Neural Networks) receives the area of interest together with the model to be applied and passes both to pytorch. The result returned is processed further by rasdaman and in this case encoded as TIFF.

As it turns out in experiments as part of the AI-Cube [108] and FAIRiCUBE [109] projects, such pretrained models are extremely susceptible to even small deviations from the training data in practically all parameters: different resolution, different radiometry, different landscape, etc. Therefore, a current research topic is to investigate how to "fence" models in a way that the server can predict the result quality and possibly warn users or even refuse to apply a model in a given situation, such as "applying a fire forecast over the ocean does not make sense." Initial-stage considerations on criteria for such an assessment include the following.

- Landscape type: As a simple example, a seashore and water areas boundary polygon might automatically be applied on areas of interest when models classified as land-based are invoked.

- Temporal applicability: For example, a crop damage estimator should be applied at a time of year when normally crop grows in the area of interest.
- Data-inherent properties: This might address freeness of clouds and cloud shadows, data histogram, or other suitable statistics.

Obviously, to this end models also need to carry quality information, specifically: a description of applicability. This addresses on the one hand the input data for the model analysis, but also the quality parameters imposed by the client. It is understood that such research is in its infancy, although highly important when model-based prediction occurs somewhere in the middle of automated decision-making pipelines, without human plausibility checking.

B.7. Section 4 — Summary of Recommendations

An almost certainly incomplete set of issues with coverage handling related to ARD has been discussed. These issues are not always independent. Quite the opposite: many of them are disentangled and, hence, should be addressed in a synoptic manner. In this section, summary of the discussion is provided, providing the recommendations as defined above.

- OGC should update the GML Standard to (i) complete definitions and (ii) reflect needs that have emerged in other standards that use GML (and had to be implemented by these groups). One example is ISO 8601 / OGC WMS time strings as coordinates. On this occasion, care should be taken that any such changes — which recently tend to focus on JSON — are applied consistently over all encodings.
- OGC should establish conventions on the metadata structures so that independent metadata sections can coexist in the metadata slot of a coverage. For various situations such slots might be defined, such as for INSPIRE. Such conventions could become a Best Practice document, an annex to the existing Coverage Implementation Schema, or even a separate standard. The OGC registry should support this for an automatic validation.
- A common approach to the representation of units of measure (uom) should be adopted which ultimately allows for an automatic conversion of units. The best candidate so far appears to be QUDT. OGC should adopt QUDT as a “community project,” meaning at least some semi-normative “good practice” or alternatively some normative fixing. Additionally, development of tools and tutorials to boost community uptake should be pushed, for example in OGC Testbeds.
- Standards should strive to exclude technicalities from APIs as much as possible. A particular example is tiling of datacubes which is implementation specific, not adding any functionality, but complexity for users (including client developers).
- Interval arithmetics and systematic error propagation estimates should become common sense in and across scientific services.

- Concepts and APIs for fitness negotiation and service-level agreements should be established which enable clients to express quality requirements so that a service can decide whether it is able to honor those requirements or not.
- Applicability parameters should be developed for models so that the models can be matched against the invocation situation as a kind of “safety belt.”
- The product generation needs to become less hand-waving to more rigorously comply and keep product specifications and documentation coherent.
- ISO and OGC should extend the coverage data and processing model with the enhancements proposed in this paper. These enhancements affect the 19123-* series in ISO and the Abstract Topic 6, Coverage Implementation Schema, and WCPS Standards in OGC.
- Summarizability of Earth data should be investigated to find a framework for better safeguarding of semantically incorrect aggregation queries. As a start, OGC and ISO should enhance the coverage range type with a distinction of categorial, ordinal, and numerical range types.
- For coverage tiling, OGC should enhance the coverage maintenance standard WCS-T, rather than having completely independent, scope-limited standards not connected at all with coverages.
- An investigation is recommended on adding quality measures (as described or different) to coverages and investigate on automatically evaluating such quality measures in WCPS, maybe using ontologies. However, ML is not recommended for this task.
- OGC recently has adopted a strategy which maximizes the number of standards published, knowingly accepting that such standards may not only overlap, but effectively can be incompatible with existing standards (such as the recent CoverageJSON versus CIS). This damages the original goal of interoperability and, therefore, is being disputed controversially in the implementer, service operator, and data user communities.
- Abandon the misconception of pixel-in-X. In particular, do not perform half-resolution shifts on images. Obviously, this would need synchronous changes in many tools worldwide.

B.8. Section 5 — Conclusion

The issue of how to achieve analysis ready EO data has been inspected, starting from the coverage standards which provide already a canonical structure, but needs augmentation for more readiness. “Readiness for what?” is actually another central issue which is not always addressed sufficiently in research but is essential for having a clear goal. The following two aspects of readiness are found to be particularly important.

- Generally, clients and users should not need to be burdened with technical details not relevant for the action to be performed, such as data formats, protocol specifics, tiling, parallelization, etc.
- There is no absolute, free-standing analysis-readiness. Rather, readiness is specific to each task, and therefore applications requiring some analysis-readiness need a means to negotiate parameters. Research in this direction is encouraged.

Besides these general statements, several concrete suggestions have been contributed to relevant bodies like OGC for enhancing standards and registries for better ARD support.

It is hoped that these deliberations contribute to a better consumption readiness of Earth data and services.

B.9. Acknowledgements

This work has been supported through OGC Testbed-19 by the Testbed sponsors, by the European Commission through EU FAIRiCUBE, and by NATO through SPS Cube4EnvSec, which are gratefully acknowledged. Valuable discussions have occurred with the rasdaman team and OGC Testbed-19 participants.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] CEOS-ARD, 2023. *CEOS ANALYSIS READY DATA.* URL: https://ceos-dev.ceos.org/ard/

[2] Earth Resources Observation And Science (EROS) Center, "Collection-2 Landsat 8-9 OLI (Operational Land Imager) and TIRS (Thermal Infrared Sensor) Level-1 Data Products." U.S. Geological Survey, 2013. doi: 10.5066/P975CC9B.

[3] P. Baumann, E. Hirschorn, J. Maso, V. Merticariu, D. Misev: All in One: Encoding Spatio-Temporal Big Data in XML, JSON, and RDF without Information Loss. Proc. IEEE Intl. Workshop on Big Spatial Data (BSD), Boston, USA, 2017, https://ieeexplore.ieee.org/document/8258326,

[4] Siqueira A, Lewis A, Thankappan M, et al 2019. CEOS analysis ready data for land–an overview on the current and future work. In: *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium. IEEE*, pp 5536–5537

[5] "Landsat Science Level 3 Products | U.S. Geological Survey," Landsat Science Products. https://www.usgs.gov/landsat-missions/landsat-science-products (accessed Aug. 07, 2023).

[6] ISO 19111:2019 Geographic information Referencing by coordinates. 2019-01. https://www.iso.org/standard/74039.html

[7] CARD4L-SR, 2021. *CEOS Analysis Ready Data for Land, Product Family Specification, Surface Reflectance.* URL: https://ceos.org/ard/files/PFS/SR/v5.0/CARD4L_Product_Family_Specification_Surface_Reflectance-v5.0.pdf

[8] "OGC Climate Resilience Pilot ER," OGC Climate Resilience Pilot: Engineering Report.

[9] D. Misev, M. Rusu, P. Baumann: A Semantic Resolver for Coordinate Reference Systems. Proc. 11th Intl. Symposium on Web and Wireless Geographical Information Systems (W2GIS), Naples, Italy, April 12-13, 2012, Springer LNCS 7236

[10] CARD4L-ST, 2021. *CEOS Analysis Ready Data for Land, Product Family Specification, Surface Temperature.* URL: https://ceos.org/ard/files/PFS/ST/v5.0/CARD4L_Product_Family_Specification_Surface_Temperature-v5.0.pdf

[11] H. Butler, M. Daly, A. Doyle, S. Gillies, S. Hagen, and T. Schaub (eds.): The GeoJSON Format. August 2016. https://datatracker.ietf.org/doc/html/rfc7946

[12] CARD4L-NRB, 2022. *CEOS Analysis Ready Data for Land, Product Family Specification, Normalised Radar Backscatter.* URL: https://ceos.org/ard/files/PFS/NRB/v5.5/CARD4L-PFS_NRB_v5.5.pdf

[13] P. Baumann and J. Yu (eds.): OGC Web Coverage Service WCS Interface Standard — Processing Extension. OGC Interface Standard, OGC 08-059r4, 2014. https://portal.ogc.org/files/08-059r4

[14] CARD4L-NRB-METADATA, 2022. *Metadata Specification for CEOS Analysis Ready Data for Land, Product Family Specification, Normalised Radar Backscatter.* URL: https://ceos.org/ard/files/PFS/NRB/v5.5/CARD4L_METADATA-spec_NRB-v5.5.xlsx

[15] P. Baumann: Language Support for Raster Image Manipulation in Databases. Intl. Workshop on Graphics Modeling, Visualization in Science & Technology, Darmstadt/Germany, 1992, pp. 236 — 245

[16] CARD4L-ORB, 2022. *CEOS Analysis Ready Data for Land, Product Family Specification, Ocean Radar Backscatter.* URL: https://ceos.org/ard/files/PFS/ORB/v1.0/CARD4L_Product_Family_Specification_Ocean_Radar_Backscatter-v1.0.pdf

[17] P. Baumann: On the Management of Multidimensional Discrete Data. VLDB Journal 4(3)1994, Special Issue on Spatial Database Systems, pp. 401 — 444

[18] CARD4L-ORB-METADATA, 2022. *Metadata Specification for CEOS Analysis Ready Data for Land, Product Family Specification, Ocean Radar Backscatter.* URL: https://ceos.org/ard/files/PFS/ORB/v1.0/CARD_METADATA-spec_ORB-v1.0.xlsx

[19] J. Masó, K. Pomakis and N. Julià(eds.): OpenGIS Web Map Tile Service Implementation Standard. OpenGIS Implementation Standard, OGC 07-057r7, 2010. https://portal.ogc.org/files/?artifact_id=35326

[20] CARD4L-PFS, 2022. *CEOS Analysis Ready Data for Land, Product Family Specification, Polarimetric Radar.* URL: https://ceos.org/ard/files/PFS/POL/v3.5/CARD4L-PFS_Polarimetric_Radar-v3.5.pdf

[21] M. Burgoyne, D. Blodgett, C. Heazel, and C. Little (eds.): OGC API — Environmental Data Retrieval Standard. OGC 19-086r6, 2023-07-27. http://www.opengis.net/doc/IS/ogcapi-edr-1/1.1

[22] CARD4L-PFS-METADATA, 2022. *Metadata Specification for Analysis Ready Data for Land, Product Family Specification, Polarimetric Radar.* URL: https://ceos.org/ard/files/PFS/POL/v3.5/CARD4L_METADATA-spec_POL-v3.5.xlsx

[23] J. Robie, M. Dyck, and J. Spiegel (eds): XQuery 3.1: An XML Query Language. W3C, 2017-03-21. https://www.w3.org/TR/xquery-31/

[24] Bermudez, L. (ed.), 2021. *OGC Testbed-16: Data Access and Processing API Engineering Report (OGC 20-025).* URL: http://docs.opengeospatial.org/per/20-025r1.html

[25] "OGC Testbed-19: Call for Participation (CFP)." https://portal.ogc.org/files/?artifact_id=104098#ARD (accessed Aug. 20, 2023).

[26] *ISO: Geographic information — Schema for coverage geometry and functions — Part 1: Fundamentals.* IS 19123-1, https://committee.iso.org/sites/tc211/home/projects/projects---complete-list/iso-19123-1.html

[27] G. Gutman, A. Ignatov: Towards a Common Language in Satellite Data Management: A New Processing Level Nom-enclature. IEEE Intl. Geoscience and Remote Sensing

Symposium (IGARSS), Singapore, 1997, pp. 1252-1254 vol.3, doi: 10.1109/IGARSS.1997.606413

[28] Landy, J., & Dawson, G. (2022). Year-round Arctic sea ice thickness from CryoSat-2 Baseline-D Level 1b observations 2010-2020 (Version 1.0) [Data set]. NERC EDS UK Polar Data Centre. https://doi.org/10.5285/d8c66670-57ad-44fc-8fef-942a46734ecb

[29] CARD4L-AR, 2022. *CEOS Analysis Ready Data for Land, Product Family Specification, Aquatic Reflectance.* URL: https://ceos.org/ard/files/PFS/AR/v1.0/CARD4L_Product_Family_Specification_Aquatic_Reflectance-v1.0.pdf

[30] Xarray N-D labeled arrays and datasets in Python. https://xarray.dev (last accessed: 2023-12-28)

[31] Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3:160018. https://doi.org/10.1038/sdata.2016.18

[32] Mathworks: Math. Graphics. Programming. https://www.mathworks.com/products/matlab.html (last accessed: 2023-12-28)

[33] NV5: IDL® Software. https://www.nv5geospatialsoftware.com/Products/IDL (last accessed: 2023-12-28)

[34] USGS: Differences between Pixel-is-Area and Pixel-is-Point Designations. https://www.usgs.gov/media/images/differences-between-pixel-area-and-pixel-point-designations (last accessed: 2023-12-28)

[35] OGC: Deal correctly with GeoTIFF Pixel-is-point vs Pixel-is-area. https://github.com/opengeospatial/ogcapi-coverages/issues/92 (last accessed: 2023-12-28)

[36] R. Hanisch, S. Chalk, R. Coulon, S. Cox, S. Emmerson, F. Javier, F. Sandoval, A. Forbes, J. Frey, B. Hall, R. Hartshorn, P. Heus, S. Hodson, K. Hosaka, D. Hutzschenreuter, C.-S. Kang, S. Picard and R. White: Stop Squandering Data: Make Units of Measurement Machine-Readable. Nature, 10 May 2022, https://www.nature.com/articles/d41586-022-01233-w

[37] UCUM. https://ucum.org (last accessed: 2023-12-28)

[38] EEA: EIONET Data Dictionary. https://dd.eionet.europa.eu (last accessed: 2023-12-28)

[39] W3C: Simple Knowledge Organization System (SKOS). https://www.w3.org/2004/02/skos (last accessed: 2023-12-28)

[40] OGC: OGC RAINBOW (OGC Definitions Server). https://www.ogc.org/resources/rainbow (last accessed: 2023-12-28)

[41] OGC: Units registered in the OGC Rainbow. https://github.com/opengeospatial/NamingAuthority/issues/263 (last accessed: 2023-12-28)

[42] Maso, J. (ed.), 2020. *OGC Testbed-16: Analysis Ready Data Engineering Report (OGC 20-041).* URL: http://docs.opengeospatial.org/per/20-041.html

[43] S. Hunt, “Review of Analysis Ready Data Specification.” European Space Agency, Oct. 31, 2021. Accessed: Aug. 20, 2023. [Online]. Available: https://earth.esa.int/eogateway/documents/20142/37627/Analysis+Ready+Data.pdf/f19baa0e-ef74-4037-c507-8b800ebb888d

[44] *ISO: Geographic information — Schema for coverage geometry and functions — Part 2: Coverage implementation schema.* IS 19123-2, https://committee.iso.org/sites/tc211/home/projects/projects---complete-list/iso-19123-2.html

[45] OGC: Coverage Implementation Schema. Document 09-146r8, http://schemas.opengis.net/cis/1.1 (last accessed: 2023-12-28)

[46] QUDT. FAIRsharing record for: Quantities, Units, Dimensions and Types. 2015. doi: 10.25504/FAIRsharing.d3pqw7, https://qudt.org (last accessed: 2023-12-28)

[47] M. Lefrançois, A. Zimmermann: Linked Datatypes — Light-weight Description for Key Information. https://ci.mines-stetienne.fr/lindt/ (last accessed: 2023-12-28)

[48] CODATA: Digital Representation of Units of Measurement (DRUM). https://codata.org/initiatives/task-groups/drum/ (last accessed: 2023-12-28)

[49] BIPM: Bureau International des Poids et Mesures. https://www.bipm.org (last accessed 2023-12-28)

[50] The Units of Measure Interoperability Service. https://umis.stuchalk.domains.unf.edu (last accessed 2023-12-28)

[51] P. Furtado, P. Baumann: Storage of Multidimensional Arrays based on Arbitrary Tiling. Proc. Intl. Conference on Data Engineering, Sydney, Australia, 1999

[52] PostgreSQL: Chapter 10. Raster Reference. https://postgis.net/docs/RT_reference.html (last accessed: 2023-12-28)

[53] P. Cozzi and S. Lilley (eds): OGC: 3D Tiles Specification. OGC 22-025r4, 2023-01-12. https://docs.ogc.org/cs/22-025r4/22-025r4.html

[54] C. Reed (ed.): OGC: OGC Abstract Specification Topic 22 — Core Tiling Conceptual and Logical Models for 2D Euclidean Space. OGC Absatract Specifcication, OGC 19-014r3, 2020-10-22. https://docs.ogc.org/as/19-014r3/19-014r3.html

[55] P. Baumann (ed.): OGC: Web Coverage Service Interface Standard — Transaction Extension. OGC Implementation Standard, OGC 13-057r1, 2016-11-17. https://docs.opengeospatial.org/is/13-057r1/13-057r1.html

[56] Vretanos, P. A. (ed.), 2020. *OGC Testbed-16: Data Access and Processing Engineering Report (OGC 20-016)*. URL: http://docs.opengeospatial.org/per/20-016.html

[57] W. Vlach, L. Lin, C. Zhang, L. Di, H. Zhao, and H. Li, “Enhancing Remote Sensing Based Machine Learning Applications through Analysis Ready Data: A Comprehensive Review,” in 2023 11th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Jul. 2023, pp. 1–5. doi: 10.1109/Agro-Geoinformatics59224.2023.10233548.

[58] *ISO: Geographic information — Schema for coverage geometry and functions — Part 3: Processing Fundamentals.* IS 19123-3, https://committee.iso.org/sites/tc211/home/projects/projects---complete-list/iso-19123-3.html

[59] Learn the Datacube Standards. https://earthserver.eu/wcs

[60] S. Schleidt and I. Rinne (eds.)(2023) OGC Abstract Specification Topic 20: Observations, measurements and samples. OGC 20-082r4, version 3.0.0. URL: https://docs.ogc.org/as/20-082r4/20-082r4.html

[61] E. McMullin: The Origins of the Field Concept in Physics. Physics in Perspective, 4(1)2002, pp. 13-39

[62] D.G. Krige: A statistical approach to some basic mine valuation problems on the Witwatersrand. Technometrics, 52:119-139, 1951

[63] M.L. Stein: Interpolation of Spatial Data. Springer Series in Statistics, New York, Springer, 1999

[64] Satellite Imaging Corp.: Landsat 8 Satellite Sensor. https://www.satimagingcorp.com/satellite-sensors/other-satellite-sensors/landsat-8 (last accessed: 2023-12-28)

[65] H.-J. Lenz, A. Shoshani: Summarizability in OLAP and Statistical Data Bases. Proc. Intl. Conf. on Scientific and Statistical Database Management, August 1997, Olympia, USA, pp. 132-143, doi: 10.1109/SSDM.1997.621175

[66] J.-N. Mazón, J. Lechtenbörger, J. Trujillo: A Survey on Summarizability Issues in Multidimensional Modeling. Data & Knowledge Engineering, 68(12)2009, pp. 1452-1469, doi: 10.1016/j.datak.2009.07.010

[67] J. Horner, I.-Y. Song, P.P. Chen: An Analysis of Additivity in OLAP Systems. Proc. 7th ACM Intl. Workshop on Data Warehousing and OLAP (DOLAP), 2004. pp. 83-91. doi: 10.1145/1031763.1031779

[68] F.M. Malvestuto, M. Mezzini, M. Moscarini: An Analytical Approach to the Inference of Summary Data of Additive Type. Theoretical Computer Science, 385(1-3)2007, pp. 264-285, doi: 10.1016/j.tcs.2007.07.004

[69] N. Prat, I. Comyn-Wattiau, J. Akoka: Combining Objects with Rules to Represent Aggregation Knowledge in Data Warehouse and OLAP Systems. Data & Knowledge Engin-eering, 70(8)2011, pp. 732-752, doi: 10.1016/j.datak.2011.03.004

[70] C.A. Hurtado, C. Gutierrez, A.O. Mendelzon: Capturing Summarizability with Integrity Constraints in OLAP. ACM Transactions on Database Systems, volume 30, 2005, pp. 854-886

[71] Havard, 2023. *Analysis Ready Datasets*. URL: https://datamanagement.hms.harvard.edu/collect-analyze/analysis-ready-datasets

[72] J. L. Dwyer, D. P. Roy, B. Sauer, C. B. Jenkerson, H. K. Zhang, and L. Lymburner, "Analysis Ready Data: Enabling Analysis of the Landsat Archive," Remote Sens., vol. 10, no. 9, Art. no. 9, Sep. 2018, doi: 10.3390/rs10091363.

[73] P. Baumann, D. Misev, V. Merticariu, B. Pham Huu: Datacubes: Towards Space/Time Analysis-Ready Data. In: J. Doellner, M. Jobst, P. Schmitz (eds.): Service Oriented Mapping — Changing Paradigm in Map Production and Geoinformation Management, Springer Lecture Notes in Geoinformation and Cartography, 2018

[74] Kwok, R., A. Petty, G. Cunningham, T. Markus, D. W. Hancock III, A. Ivanoff, J. Wimert, M. Bagnardi, N. Kurtz, and the ICESat-2 Science Team. 2023. ATLAS/ICESat-2 L3A Sea Ice Height, Version 6. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/ATLAS/ATL07.006.

[75] *OGC: Web Coverage Processing Service (WCPS) Standard.* OGC 08-068r3, https://docs.ogc.org/is/08-068r3/08-068r3.html

[76] J. Horner, I.Y. Song: A Taxonomy of Inaccurate Summaries and Their Management in OLAP Systems. In: L. Delcambre, C. Kop, H.C. Mayr, J. Mylopoulos, O. Pastor (eds): Conceptual Modeling — ER 2005. LNCS 3716, Springer, doi: 10.1007/11568322_28

[77] R. Kimball, M. Ross: The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (2nd. ed.). John Wiley & Sons, 2002

[78] T. Niemi, M. Niinimäki, P. Thanisch, J. Nummenmaa: Detecting Summarizability in OLAP. Data & Knowledge Engineering, Volume 89, 2014, pp. 1-20, doi: 10.1016/j.datak.2013.11.001

[79] A. Robin (ed.): OGC: SWE Common Data Model Encoding Standard 2.0. OGC Encoding Standard, OGC 08-094r1, 2011-01-04. https://portal.ogc.org/files/?artifact_id=41157

[80] A.S. Whittemore: Collapsibility of Multidimensional Contingency Tables. J.R. Statist. Soc. Series B, vol. 40, 1978, pp. 328-340

[81] P. Thanisch, T. Niemi, J. Nummenmaa, M. Niinimäki: De-tecting measurement issues in SQL arithmetic expressions and aggregations. Data & Knowledge Engineering, Volume 122, 2019, pp. 116-129, doi: 10.1016/j.datak.2019.06.001

[82] ISO: Information technology — Guidance for the use of database language SQL — Part 9: Online analytic process-ing (OLAP) capabilities (Guide/OLAP). ISO/IEC 19075-9:2022(en), https://www.iso.org/obp/ui#iso:std:iso-iec:19075:-9:ed-1:v1:en (last accessed: 2023-12-28)

[83] Microsoft: Querying Multidimensional Data with MDX. https://learn.microsoft.com/en-us/analysis-services/multidimensional-models/mdx/querying-multidimensional-data-with-mdx?view=asallproducts-allversion (last accessed: 2023-12-28)

[84] SOILGRIDS. https://soilgrids.org (last accessed: 2023-12-28)

[85] SeaDAS Tools. https://seadas.gsfc.nasa.gov/docs/SeaDAS_Tools.pdf (last accessed: 2023-12-28)

[86] SW. Bailey, P.J. Werdell: A multi-sensor approach for the on-orbit validation of ocean color satellite data products. Rem. Sens. Environ. 102, 12-23 (2006)

[87] N. Djamai, R. Fernandes: Comparison of SNAP-Derived Sentinel-2A L2A Product to ESA Product over Europe. Remote Sensing 2018, 10, 926, doi: 10.3390/rs10060926

[88] Di, Liping, 2021. ARD Services for ECV Data in CEOS WGISS Carbon Community Portal. *52th CEOS WGISS Plenary.* URL: https://ceos.org/document_management/Working_Groups/WGISS/Meetings/WGISS52/2.Wednesday/2021.10.20_10.05_ARDServiceForECVs.pdf Proposed Work Organized by Technical Activity Type.

[89] "New CEOS ARD Compliant Surface Reflectance Products: Sentinel-2 and EnMAP | CEOS | Committee on Earth Observation Satellites," Committee on Earth Observation Satellites, Jan. 28, 2022. https://ceos.org/news/ard-sentinel-2-enmap/ (accessed Aug. 20, 2023).

[90] Earth Datacube Playground. https://standards.rasdaman.com

[91] K.S. Kim and N. Ishimaru (eds.)(2020), OGC Moving Features Encoding Extension — JSON. OGC 19-045r3. URL: https://docs.ogc.org/is/19-045r3/19-045r3.html

[92] *OGC: Coverage Implementation Schema Standard.* OGC 09-146r8, http://docs.opengeospatial.org/is/09-146r8/09-146r8.html

[93] P. Strobl: A Revised Processing Level Scheme for Earth Observation Data. Big Data from Space (BiDS), Vienna, Austria, 2023, doi:10.2760/46796

[94] R. Stull: 20.4: Numerical Errors and Instability. https://geo.libretexts.org/Bookshelves/Meteorology_and_Climate_Science/Practical_Meteorology_(Stull)/20%3A_Numerical_Weather_Prediction_(NWP)/20.03%3A_Section_4 (last accessed: 2023-12-28)

[95] R. de la Llave: Computer Assisted Proofs of Stability of Matter. In: K.R. Meyer, D.S. Schmidt (eds): Computer Aided Proofs in Analysis. The IMA Volumes in Mathe-matics and Its Applications, vol 28, 1991. Springer, doi: 10.1007/978-1-4613-9092-3_11

[96] M.W. Gutowski: Power and Beauty of Interval Methods. 20 Feb 2003. http://arxiv.org/abs/physics/0302034 (last accessed: 2023-12-28)

[97] K. Atkinson: An Introduction to Numerical Analysis. John Wiley & Sons, 1991

[98] Rasdaman: Null Values. http://doc.rasdaman.org/04_ql-guide.html#sec-nullvalues (last accessed: 2023-12-28)

[99] R. El-Awadi, M. Abu-Rizka: A Framework for Negotiating Service Level Agreement of Cloud-based Services. Procedia Computer Science, Volume 65, 2015, pp. 940-949, doi: 10.1016/j.procs.2015.09.066

[100] KS. Sendhil Kumar, N. Jaisankar: An Automated Resource Management Framework for Minimizing SLA Violations and Negotiation in Collaborative Cloud. Intl. Journal of Cognit-ive Computing in Engineering, Volume 1, 2020, pp. 27-35, doi: 10.1016/j.ijcce.2020.09.001

[101] L. Li, L. Liu, S. Huang et al.: Agent-based multi-tier SLA negotiation for intercloud. J Cloud Comp 11(16)2022, doi: 10.1186/s13677-022-00286-6

[102] W. Bai, J. Ding, C. Zhang: Dual hesitant fuzzy graphs with applications to multi-attribute decision making. Intl. Journal of Cognitive Computing in Engineering, Volume 1, 2020, pp. 18-26, doi: 10.1016/j.ijcce.2020.09.002

[103] Dwyer J.L., Roy D.P., Sauer B., et al. 2018. Analysis ready data: enabling analysis of the Landsat archive. *Remote Sens* 10:1363.

[104] "Sentinel-2 MSI — Technical Guide — Sentinel Online," Sentinel Online. https://copernicus.eu/technical-guides/sentinel-2-msi (accessed Aug. 20, 2023).

[105] Draft OGC API — Connected Systems — Part 1: Feature Resources (0.0.1) URL: https://ogcapi.ogc.org/connectedsystems/

[106] OGC Spatio-Temporal Coverage / Datacube Standards. https://myogc.org/go/coveragesDWG

[107] OGC: Spatio-Temporal Coverage / Datacube Standards. https://myogc.org/go/coveragesDWG

[108] AI-Cube. https://ai-cu.be (last accessed: 2023-12-28)

[109] FAIRiCUBE. https://fairicube.eu (last accessed: 2023-12-28)

[110] Lewis A, Lacey J, Mecklenburg S, et al 2018. CEOS analysis ready data for Land (CARD4L) overview. In: *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE*, pp 7407–7410

[111] "OGC Testbed-16: Analysis Ready Data Engineering Report," OGC Testbed-16: Analysis Ready Data Engineering Report.

[112] P. Baumann: A General Conceptual Framework for Multi-Dimensional Spatio-Temporal Data Sets. Environmental Modelling & Software, 2021, doi: 10.1016/j.envsoft.2021.105096

[113] CEOS-ARD, 2019. *CEOS Analysis Ready Data Strategy.* URL: https://ceos.org/document_management/Meetings/SIT/SIT-35/Documents/3.1-CEOS-ARD-Strategy.pdf

[114] "What are Landsat Collection Tiers? | U.S. Geological Survey," What are Landsat Collection Tiers? https://www.usgs.gov/faqs/what-are-landsat-collection-tiers (accessed Aug. 07, 2023).

[115] OGC: Topic 6 — Schema for Coverage Geometry and Functions. OGC document 07-011, version 7.0.0, 2006-01-30. https://portal.ogc.org/files/?artifact_id=19820

[116] CEOS-ARD, 2021. *CEOS Analysis Ready Data Governance Framework.* URL: https://ceos.org/ard/files/CEOS_ARD_Governance_Framework_18-October-2021.pdf

[117] Earth Resources Observation And Science (EROS) Center, “Collection-2 Landsat 7 Enhanced Thematic Mapper Plus (ETM+) Level-1 Data Products.” U.S. Geological Survey, 1999. doi: 10.5066/P9TU80IG.

[118] P. Baumann, E. Hirschorn, J. Maso: Coverage Implement-ation Schema (CIS) with Corrigendum, version 1.1. OGC document 09-146r8, 2019. http://docs.opengeospatial.org/is/09-146r8/09-146r8.html