Towards Sentinel-2 Analysis Ready Data: a Swiss Data Cube Perspective

Giuliani, Gregory; Chatenoux, Bruno; Honeck, Erica Cristine; Richard, Jean-Philippe

TOWARDS SENTINEL-2 ANALYSIS READY DATA: A SWISS DATA CUBE PERSPECTIVE

Gregory Giuliani[1], Bruno Chatenoux[1], Erica Honeck[1], Jean-Philippe Richard[1]

[1]University of Geneva, Institute for Environmental Sciences, GRID-Geneva

ABSTRACT

Earth Observations Data Cubes (EODC) are a new paradigm revolutionizing the way users can interact with Earth Observations (EO) data. They can provide access to large spatio-temporal data in analysis ready format. Systematic and regular provision of Analysis Ready Data (ARD) can significantly reduce the burden on EO data users by minimizing the time and scientific knowledge required to access and prepare remotely-sensed data having consistent and spatially aligned calibrated surface reflectance observations. Currently, Sentinel-2 ARD are not commonly generated by the Copernicus program and consequently getting uniform and consistent Sentinel-2 ARD remains a challenging task. This paper presents an approach to generate Sentinel-2 ARD using an automated processing services chain. The approach has been tested and validated to complete the Swiss Data Cube with Sentinel-2 data.

Index Terms— Sentinel-2, Open Data Cube, Analysis Ready Data, Earth Observations

1. INTRODUCTION

Remotely sensed Earth Observations (EO) data have already exceeded the petabyte-scale and are increasingly freely and openly available from various repositories (e.g., USGS Earth Explorer, Copernicus Open Access Hub, Global Earth Observation System of Systems, Google and Amazon Clouds). Landsat archive contains more than 2 Petabytes of data, and the European Space Agency (ESA) Sentinels program has already surpassed this data volume. Traditional approaches to the acquisition, management, distribution and analysis of EO data have limitations (e.g., data size, heterogeneity and complexity) that impede their true information potential to be realized [1]. To enable data analysts or scientists to transform these large amounts of data into useful information and decision-ready products, this requires to rapidly analyze data in a transparent and repeatable manner [2]. The fact that the full information potential of EO data has not yet been realized and therefore remains still underutilized is explained by various reasons: (1) increasing volumes of data generated by satellites; (2) lack of expertise and computing resources to efficiently access, process, and utilize EO data; (3) the particular structure of EO data and (4) the significant effort and cost required to store and process data limit its effective use [3]. Addressing Big Data challenges such as Volume, Velocity and Variety, requires a change of paradigm and moving away from traditional local processing and data distribution methods to lower the barriers caused by data size and related complications in data management. In particular, data volume and velocity will continue to grow as the demands increase for decision-support information derived from these data [2].

To tackle these issues and bridge the gap between users' expectations and current Big Data analytical capabilities, EO Data Cubes (EODC) are a promising solution to store, organize, manage and analyze EO data. The main objective of EODC is to facilitate EO data usage by addressing Volume, Velocity, Variety challenges and providing access to large spatio-temporal data in an analysis ready format [4]–[6]. Different EODC are currently operational such as Digital Earth Australia [7], the EarthServer [8], or the Google Earth Engine [9]. These initiatives are paving the way to broaden the use of EO data to larger communities of users; support decision-makers with timely and actionable information converted in meaningful geophysical variables; and ultimately are unlocking the information power of EO data.

Most of these EODC implementations are using Landsat data. This is explained by the fact that: (1) the Landsat program is the longest EO program initiated in 1972 and providing continuous observations for more than 45 years [10]; and (2) since 2008 the entire data archive has been freely and openly accessible [11], [12]. This created unprecedented opportunities to use Landsat data in different contexts such as land cover changes or ecosystem mapping [10], [13]. However, another valuable EO data source is represented by the Sentinels program. In particular, Sentinel-2 is a polar-orbiting, multispectral, high-resolution satellite for land monitoring. This mission, composed of two satellites, is developed by ESA and is part of a family of missions for the operational need of the European EO program called Copernicus.

An essential pre-condition to support user applications and generating usable information products is to facilitate data access, preparation and analyses. The systematic and regular provision of Analysis Ready Data (ARD) can significantly reduce the burden of EO data usage. To be considered as ARD, data should be processed to a minimum set of requirements (e.g., radiometric and geometric calibration; atmospheric correction; metadata description) and organized in a way that allows immediate analysis without additional effort [14]. ARD correspond in optical imagery to surface reflectance products.

While Landsat ARD products are commonly used and generated either through USGS Landsat Collection 1 repository [15] or automated custom pre-processing workflows [3], Sentinel-2 ARD product generation is still an issue that has not been addressed yet as this is not commonly provided by the Copernicus Open Access Hub (http://scihub.copernicus.eu). This clearly limits the usage of Sentinel-2 data and methodologies are required to fully benefit from this European EO program.

Recognizing these issues, the aim of this paper is to present an approach to enable efficient data access and pre-processing to generate Sentinel-2 ARD. This approach has been tested and validated within the Swiss Data Cube (http://www.swissdatacube.org) initiative and has allowed to ingest more than 2TB of data corresponding to the entire Sentinel-2 data archive covering Switzerland.

2. SWISS DATA CUBE APPROACH FOR GENERATING ANALYSIS READY DATA (ARD)

Like many other countries in the world, Switzerland is facing various challenges (e.g., land management, environmental degradation) caused by increasing pressures on its natural resources. These challenges need to be overcome to meet the needs of a growing population. Some of these issues can be monitored using remotely-sensed EO data and may benefit from freely and openly available data archives such as Landsat and Sentinels programs. The Swiss Data Cube (SDC) is an initiative supported by the Federal Office for the Environment (FOEN) and developed, implemented and operated by the United Environment Program (UNEP)/GRID-Geneva in partnership with the University of Geneva. The objective of the SDC is to support the Swiss government for environmental monitoring and reporting, and enable Swiss scientific institutions to fully benefit of EO data for research and innovation. The SDC is based on the Open Data Cube (http://www.opendatacube.org) software stack which is an open source project initiated and led by Geoscience Australia, the Commonwealth Scientific and Industrial Research Organization (CSIRO), the United States Geological Survey (USGS), the National Aeronautics and Space Administration (NASA) and the Committee on Earth Observations Satellites (CEOS).Currently, the SDC contains 33 years of Landsat 5,7,8 ARD (1984-2017) corresponding to approximately 4000 scenes for a total volume of 1TB and more than 35 billion observations over the entire country.

Considering the need for routinely generating ARD products and the fact that currently ARD are not commonly generated by data providers like USGS Earth Resources Observation and Science (EROS) Center Science Processing Architecture (ESPA) (for Landsat data) and Copernicus Open Access Hub (for Sentinel-2 data), a central aspect is to have a method for generating ARD products ensuring that all observations ingested and stored in an EODC are consistent and comparable. Such a procedure should be automated as much as possible to discover, access and pre-process data from various repositories (e.g., Copernicus Open Access Hub, Earth Explorer, Google Earth Engine), handling different sensors (e.g., Landsat TM, ETM, OLI, Sentinel-2 MSI) and make it interoperable to be reusable. To meet these requirements, the Live Monitoring of Earth Surface (LiMES) framework has been used to generate Landsat 5,7,8 ARD products that are stored in the SDC [3]. LiMES is based on a composable chain of interoperable services for automating data discovery, downloading and processing, to transform EO data into information products [16]. This framework is supported by large storage capacities, high performance distributed computers, and interoperable standards for implementing a scalable, flexible and efficient analysis environment of decades of EO data for monitoring purposes in various domains.

3. SENTINEL-2 ARD FOR SWITZERLAND

Similar to the approach used to generate Landsat ARD products, the LiMES framework has been used to automatically produce Sentinel-2 ARD products. Hereafter, we present some of the challenges and issues identified during the ARD process generation (Figure 1).

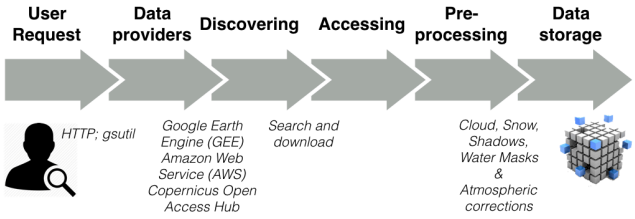

Figure 1: Workflow for generating Sentinel-2 ARD.

3.1 Sentinel-2 scenes discovery and access

The full coverage of Switzerland corresponds to twelve Sentinel-2 scenes (31TGM, 31TGN, 32TLR, 32TLS, 32TLT, 32TMR, 32TMS, 32TMT, 32TNS, 32TNT, 32TPS, 32TPT). To ensure having the most complete archive and to overcome the issue of slow performance of the Copernicus Open Access Hub (e.g., slow download) and the limitation of maximum 2 concurrent downloads while using the Application Programming Interface (API) Hub (https://scihub.copernicus.eu/userguide/5APIsAndBatchScripting); a python script has been developed to query the Google Earth Engine (GEE) data repository using the gsutil tool (https://cloud.google.com/storage/docs/gsutil). to download data in Cloud Optimized GeoTIFF format (http://www.cogeo.org). This appears the best solution in terms of download speed, stability, storage and data readiness.This allows having good performance for generating a list of available scenes for a given coverage and downloading them as fast as possible. Currently, this corresponds for Switzerland to more than 2400 scenes for a total size of 2TB over the period 2015-2017. Each image

requires around 10 minutes to be downloaded and processed. Since the launch of Sentinel-2B, a monthly mean of approximately 125 scenes can be downloaded. While automating the process to ensure that the archive is the most up-to-date, this seems a reasonable amount of data to ingest. A yearly update will correspond to approximately 1TB and will require 12 days of computation on the current server (Processors Intel Xeon E5-2660 v2 @ 2.2 GHz; 8 CPUs (6CPUs used for processing, 2CPUs for system and UI); 50Gb RAM; 6TB Hard Drive; Linux Ubuntu 16.04).

3.2 Sentinel-2 scenes pre-processing

After having developed an efficient strategy for discovering and accessing Sentinel-2 data, the pre-processing step is fundamental for generating ARD products. Pre-processing is done using the Sen2Cor algorithm (http://step.esa.int/main/third-party-plugins-2/sen2cor/). This is the algorithm proposed by ESA for Sentinel-2 Level 2A generation. It performs a scene classification to identify clouds, cloud shadows, snow and water and generates masks accordingly. In a second step, raw values are converted to Surface Reflectance (SR) by applying an atmospheric correction. At the end of this procedure, ARD products are ready to be ingested in a Data Cube.

However, two issues have been identified using Sen2Cor. First, an incomplete cloud shadow detection and second, it appears that data in mountainous regions have been overcorrected. Indeed, pixels in shadows became lighter than their surroundings if data are corrected using a Digital Elevation Model (DEM) (Figure 2). Consequently, it has been decided not to use any DEM in the ARD workflow generation in order to avoid generating those identified artifacts.

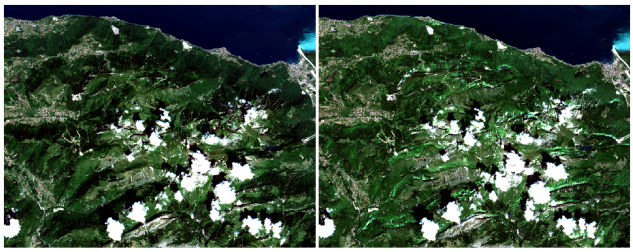

Figure 2: Correction artifacts with (right) or without (left) using a DEM.

An alternative to investigate would be to use the Atmospheric and Radiometric Correction of Satellite Imagery (ARCSI) software (http://rsgislib.org/arcsi/) that allows generating masks and applying atmospheric corrections on Sentinel-2 data (as it was done for Landsat pre-processing).

Finally, a choice has to be made about which bands should be processed among the 13 bands in the visible, near-infrared, and short-wave infrared wavelengths. In the SDC, all bands are ingested except band 10 (Cirrus) that is processed by Sen2Cor. Especially bands 5,6,7,8 are bands at 20m useful for agriculture and vegetation [17].

3.3 Sentinel-2 data storage strategy

A final challenge identified concerns the data storage strategy. Due to the fact that Sentinel-2 data are stored in a rolling archive in the Copernicus Open Access Hub, we have decided to keep a copy of the original data downloaded from the different repositories. This duplicates data (1.5 Tb for the period 2015 - 2017, then 1 Tb per year) but we consider it to be important to have a full copy of the archive on our premises in case one has to reprocess data (e.g., new atmospheric correction algorithm).

Additionally, when building a multi-sensor data cube (e.g., Landsat and Sentinel-2) an important aspect to consider is ensuring that all observations (i.e., pixels) have the same spatial resolution and therefore can be spatially aligned. In the case of the SDC, all Landsat observations (at 30-meter resolution) are resampled at 10-meter to match the Sentinel-2 resolution. This is justified by the fact that this will allow to develop algorithms that can use both sensors for efficient time-series analysis. This is a promising solution to analyze multi-sensor data in a consistent way and to move towards the objective to be sensor agnostic while analyzing EO data.

4. DISCUSSION

The proposed approach has proven to be simple to implement using interoperable components and has facilitated the efficient and consistent production of Sentinel-2 Analysis Ready Data. It has allowed to pre-process and to incorporate the entire archive for 2.5 years covering Switzerland corresponding to more than 2400 scenes. The main benefit is that this solution fully automates discovery, access and pre-processing of Sentinel-2 data. The proposed workflow ensures replicability and homogeneity in results. Moreover, this approach is scalable and it is simple to add or complete with new components (e.g., replacing Sen2Cor by ARCSI).

In terms of limitations, addressing Big Data characteristics such as Volumes, Velocity and Variety requires computing performance. Within the current server environment, it took 16 days to execute the entire workflow for all Sentinel-2 scenes with six processes in parallel. To increase performance, cloud computing appears a good solution and distributed computing environments can be optimized for large processing tasks by optimizing the number of CPUs, memory, storage and best options for parallelizing computation [18]. The diversity of protocols, interfaces, and APIs required to handle the variety of data can be difficult. A good solution can be using an API that allows discovering and accessing heterogeneous EO data repositories in a seamless and homogenous way [6]. Finally, Sentinel-2 analysis tools currently implemented have not yet been extensively tested (compared to Landsat) in the current

version of the Open Data Cube software and therefore some bugs may appear while analyzing data stored in the SDC.

The next important challenge concerns the operationalization and usability of the Swiss Data Cube. After the first run of ingestion for storing the entire Sentinel-2 archive, an automated procedure to regularly add newly available Sentinel-2 scenes is necessary. This will ensure that the SDC is up-to-date and in turn can improve the quality of products generated by adding more observations. Additionally, to harness the full potential of EO data for supporting informed decision-making requires developing new algorithms and tailored applications for generating useful information for governmental offices. Furthermore, it is important to continue research (e.g., machine learning) and capacity building efforts associated with this new technology to ensure technology transfer and leverage the potential of EO data for various communities of users. Finally, a future challenge would be to routinely generate Sentinel 1 ARD products and expand the analytical capabilities of the SDC to radar imagery.

5. CONCLUSIONS

EO Data Cubes is a disruptive technology that has the potential to routinely transform EO data into useful and actionable information. An essential task is to reduce the data preparation burden on users by generating ARD products. However, Sentinel-2 ARD are not commonly generated by the Copernicus Open Access Hub impeding a wide usage of these data.

The proposed approach using an interoperable data processing chain enabled the generation of Sentinel-2 ARD products in a seamless and consistent way. It has allowed to efficiently discover, download and pre-process the entire Sentinel-2 archive for Switzerland. This solution has lowered the barrier to Sentinel-2 ARD products generation by automating pre-processing steps and increased the use of Sentinel-2 data at a country scale, allowing users to concentrate on data analysis. This is an essential condition for unlocking the information power of Big EO data and to become an essential asset for environmental monitoring.

ACKNOWLEDGMENTS

The authors would like to thank the Swiss Federal Office for the Environment (FOEN) for their financial support to the Swiss Data Cube. We would also like to acknowledge Dr. Brian Killough (NASA) and the ODC community for their valuable support in implementing and developing the Swiss Data Cube. The views expressed in the paper are those of the authors and do not necessarily reflect the views of the institutions they belong to.

REFERENCES

[1] M. B. J. Purss *et al.*, "Unlocking the Australian Landsat Archive – From dark data to High Performance Data infrastructures," *GeoResJ*, vol. 6, pp. 135–140, Jun. 2015.

[2] M. B. J. Purss, "The promise of Discrete Global Grid Systems," *Geoconnextion Int. Mag.*, pp. 16–17, 2014.

[3] G. Giuliani *et al.*, "Building an Earth Observations Data Cube: lessons learned from the Swiss Data Cube (SDC) on generating Analysis Ready Data (ARD)," *Big Earth Data*, vol. 1, no. 1, pp. 1–18, Nov. 2017.

[4] P. Strobl *et al.*, "The Six Faces of the Data Cube," in *Proceedings of the 2017 conference on Big Data from Space*, Toulouse, France, 2017, pp. 32–35.

[5] A. Lewis *et al.*, "The Australian Geoscience Data Cube — Foundations and lessons learned," *Remote Sens. Environ.*, 2017.

[6] S. Nativi, P. Mazzetti, and M. Craglia, "A view-based model of data-cube to support big earth data systems interoperability," *Big Earth Data*, vol. 0, no. 0, pp. 1–25, Dec. 2017.

[7] A. Lewis *et al.*, "Rapid, high-resolution detection of environmental change over continental scales from satellite data - the Earth Observation Data Cube," *Int. J. Digit. Earth*, vol. 9, no. 1, pp. 106–111, Jan. 2016.

[8] P. Baumann *et al.*, "Big Data Analytics for Earth Sciences: the EarthServer approach," *Int. J. Digit. Earth*, vol. 9, no. 1, pp. 3–29, Jan. 2016.

[9] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sens. Environ.*

[10] M. A. Wulder *et al.*, "Landsat continuity: Issues and opportunities for land cover monitoring," *Remote Sens. Environ.*, vol. 112, no. 3, pp. 955–969, Mar. 2008.

[11] M. A. Wulder, J. G. Masek, W. B. Cohen, T. R. Loveland, and C. E. Woodcock, "Opening the archive: How free data has enabled the science and monitoring promise of Landsat," *Remote Sens. Environ.*, vol. 122, pp. 2–10, 2012.

[12] B. Ryan, "The benefits from open data are immense," *Geospatial World*, pp. 72–73, 2016.

[13] V. J. Pasquarella, C. E. Holden, L. Kaufman, and C. E. Woodcock, "From imagery to ecology: leveraging time series of all available Landsat observations to map and monitor ecosystem state and dynamics," *Remote Sens. Ecol. Conserv.*, p. n/a-n/a, 2016.

[14] B. Killough, "CEOS Land Surface Imaging Analysis Ready Data (ARD) Description Document," 2016.

[15] M. A. Wulder *et al.*, "The global Landsat archive: Status, consolidation, and direction," *Remote Sens. Environ.*, vol. 185, pp. 271–283, 2016.

[16] G. Giuliani *et al.*, "Live Monitoring of Earth Surface (LiMES): A framework for monitoring environmental changes from Earth Observations," *Remote Sens. Environ.*, 2017.

[17] P. Addabbo, M. Focareta, S. Marcuccio, C. Votto, and S. L. Ullo, "Contribution of Sentinel-2 data for applications in vegetation monitoring," *ACTA IMEKO*, vol. 5, no. 2, pp. 44–54, Sep. 2016.

[18] Z. Q. Chen, N. C. Chen, C. Yang, and L. P. Di, "Cloud Computing Enabled Web Processing Service for Earth Observation Data Processing," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 5, no. 6, pp. 1637–1649, Dec. 2012.