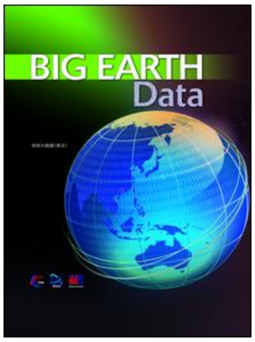


Big Earth Data

ISSN: 2096-4471 (Print) 2574-5417 (Online) Journal homepage: http://www.tandfonline.com/loi/tbed20

Building an Earth Observations Data Cube: lessons learned from the Swiss Data Cube (SDC) on generating Analysis Ready Data (ARD)

Gregory Giuliani, Bruno Chatenoux, Andrea De Bono, Denisa Rodila, Jean-Philippe Richard, Karin Allenbach, Hy Dao & Pascal Peduzzi

To cite this article: Gregory Giuliani, Bruno Chatenoux, Andrea De Bono, Denisa Rodila, Jean-Philippe Richard, Karin Allenbach, Hy Dao & Pascal Peduzzi (2017): Building an Earth Observations Data Cube: lessons learned from the Swiss Data Cube (SDC) on generating Analysis Ready Data (ARD), Big Earth Data, DOI: 10.1080/20964471.2017.1398903

To link to this article: https://doi.org/10.1080/20964471.2017.1398903

Published online: 30 Nov 2017.

BIG EARTH DATA, 2017
https://doi.org/10.1080/20964471.2017.1398903

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Building an Earth Observations Data Cube: lessons learned from the Swiss Data Cube (SDC) on generating Analysis Ready Data (ARD)

Gregory Giuliani[a,b], Bruno Chatenoux[a], Andrea De Bono[a], Denisa Rodila[a,b], Jean-Philippe Richard[a], Karin Allenbach[a], Hy Dao[a,c] and Pascal Peduzzi[a,c,d]

[a]Institute for Environmental Sciences/GRID-Geneva, University of Geneva, Geneva, Switzerland; [b]Institute for Environmental Sciences/EnviroSPACE, University of Geneva, Geneva, Switzerland; [c]Institute for Environmental Sciences/Environmental Governance and Territorial Development, University of Geneva, Geneva, Switzerland; [d]Science Division, United Nations Environment Programme, Geneva, Switzerland

ABSTRACT

Pressures on natural resources are increasing and a number of challenges need to be overcome to meet the needs of a growing population in a period of environmental variability. Some of these environmental issues can be monitored using remotely sensed Earth Observations (EO) data that are increasingly available from a number of freely and openly accessible repositories. However, the full information potential of EO data has not been yet realized. They remain still underutilized mainly because of their complexity, increasing volume, and the lack of efficient processing capabilities. EO Data Cubes (DC) are a new paradigm aiming to realize the full potential of EO data by lowering the barriers caused by these Big data challenges and providing access to large spatio-temporal data in an analysis ready form. Systematic and regular provision of Analysis Ready Data (ARD) will significantly reduce the burden on EO data users. Nevertheless, ARD are not commonly produced by data providers and therefore getting uniform and consistent ARD remains a challenging task. This paper presents an approach to enable rapid data access and pre-processing to generate ARD using interoperable services chains. The approach has been tested and validated generating Landsat ARD while building the Swiss Data Cube.

ARTICLE HISTORY
Received 5 September 2017
Accepted 18 October 2017

KEYWORDS
Data Cube; Earth Observations; Landsat; automatic processing; Analysis Ready Data

1. Introduction

Due to pressures from climate change, demographic, and economic growth, the land cover is changing (Rockstrom et al., 2009; Wulder, Masek, Cohen, Loveland, & Woodcock, 2012). To better preserve the quality of the environment and improve the management of natural resources and land planning, it is useful to monitor these changes through time (Wulder et al., 2008). One of the main advantages of remote sensing is to provide a synoptic view of

CONTACT Gregory Giuliani ✉ gregory.giuliani@unige.ch

a given spatial extent. With the archives from Landsat satellite sensors, the evolution of this coverage can be monitored all the way back to 1972 and with updates every 15 days at 30 m spatial resolution (Woodcock et al., 2008). Now with the introduction of new satellite sensors (e.g. Sentinel 2) both the spatial and temporal resolutions have increased (Gómez, White, & Wulder, 2016).

Remotely sensed Earth Observations (EO) data are increasingly available from a number of freely and openly accessible repositories. These data are highly valuable because of their unique and globally consistent information that they include (Lewis et al., 2016). Indeed, global observations together with scientific expertise and appropriate tools provide substantial benefit supporting economic development, decision-making, and policy implementation for all countries (Douglas McCuistion & Birk, 2005; Lehmann et al., 2017). However, the full information potential of EO data has not been yet realized. They remain still underutilized and stored in electronic silos of data (Gore, 1998; Lewis et al., 2016). This is due to several reasons: (1) increasing volumes of data generated by EO satellites; (2) lack of expertise, infrastructure, or internet bandwidth to efficiently and effectively access, process, and utilize EO data; (3) the particular type of highly structured data that EO data represent introducing challenges when trying to integrate or analyze them; (4) and the substantial effort and cost required to store and process data limits the efficient use of these data (CEOS, 2017; Lewis et al., 2016; Purss et al., 2015). Therefore, EO data can be considered as Big Data, data that are too large, fast-lived, heterogeneous, or complex to get understood and exploited (Baumann, Rossi, et al., 2016). Consequently, we need new approaches to fully benefit from EO data and (1) unlock the information power of EO data; (2) broaden the use of EO data to a wider range of communities; and (3) support decisions-makers with the knowledge they require by systematically analyzing all available observations and convert them into meaningful geophysical variables.

To address these Big Data challenges, it is necessary to move away from traditional local processing (e.g. desktop computer) and data distribution methods (e.g. scene-based file download) and lower the barriers caused by data size and related complexities in data preparation, handling, storage and analysis (CEOS, 2017). This paradigm shift is currently represented by EO Data Cubes (Baumann, Mazzetti, et al., 2016; Purss et al., 2015), an approach that is receiving increasing attention as a new solution to store, organize, manage, and analyze EO data in a way that was not possible before. Data Cubes (DC) are aiming to realize the full potential of EO data repositories by addressing Volume, Velocity, and Variety challenges, providing access to large spatio-temporal data in an analysis ready form (Baumann, 2017; Lewis et al., 2017). Currently, there are various operational DC like the Australian Geoscience Data Cube (AGDC – http://www.datacube.org.au), the Earth Observation Data Cube (EODC – http://eodatacube.eu), the Earth System Data Cube (ESDC – http://earthsystemdatacube.net), Earth on Amazon Web Services (EAWS – https://aws.amazon.com/earth/), and Google Earth Engine (GEE - https://earthengine.google.com). These different initiatives are covering different spatial scales (e.g. national for the AGDC, continental for the EODC, global for EAWS and GEE); storing different data (e.g. only Landsat 8 for the EODC while the AGDC stores Landsat 5, 7, 8, MODIS, and Sentinel 2 data; only processed products for the ESDC); using different infrastructure (e.g. high performance computer for the AGDC, cloud for EAWS and GEE); using different software implementations (e.g. Open Data Cube for the AGDC; RasDaMan[1] for the EODC; THREDDS for the ESDC); and using different interfaces to interact with a DC (e.g. Python API for AGDC, GEE, EAWS, ESDC; OGC WCS and WCPS for

EODC) (Baumann, Furtado, Ritsch, & Widmann, 1997; Baumann, Mazzetti, et al., 2016; Flach et al., 2016; Gorelick et al., 2017). This diversity of approaches asks also for a clear definition of an EO Data Cube. A tentative one has been recently made with the publication of "The Datacube Manifesto" that defines a Data Cube as

> a massive multi-dimensional array, also called "raster data" or "gridded data"; "massive" entails that we talk about sizes significantly beyond the main memory resources of the server hardware. Data values, all of the same data type, sit at grid points as defined by the d axes of the d-dimensional datacube. Coordinates along these axes allow addressing data values unambiguously. A d-dimensional grid is characterized by the fact that each inner grid point has exactly two neighbours along each direction; border grid points have just one. (Baumann, 2017)

The author defines also six principles of Data Cube services to ensure that services are significantly more user-friendly, efficient, and scalable than other data paradigms.

Currently, Geoscience Australia, the National Aeronautics and Space Administration (NASA), the Commonwealth Scientific and Industrial Research Organisation (CSIRO), and the United States Geological Survey (USGS) under the umbrella of the Committee on Earth Observation Satellites (CEOS) have joined their expertise and are leading the development of the Open Data Cube (https://www.ceosdatacube.org). The main objective of this initiative is to provide a data architecture solution to lower the technical barriers for users to exploit EO data to its full potential and consequently solving the problem of accessibility and use while increasing the impact of EO data (CEOS, 2017). The primary problems for users are data access, data preparation, and efficient analyses to support user applications. The two first issues are essential challenges to tackle while building a DC. Indeed, these steps concern the generation of Analysis Ready Data (ARD). CEOS defines ARD as "satellite data that have been processed to a minimum set of requirements and organized into a form that allows immediate analysis without additional user effort" (Killough, 2016). It is envisioned that systematic and regular provision of ARD will significantly reduce the burden on EO data users. To be considered as ARD, data should satisfy the following requirements: (1) metadata description; (2) radiometric calibration; (3) geometric calibration; (4a) solar and atmospheric calibrations (for optical sensors) or (4b) speckle filtering (for radar sensors). ARD corresponds for Landsat 5/7/8 and Sentinel 2 to a surface reflectance (e.g. Level 2) or to gamma naught backscatter for Sentinel 1. Landsat ARD data can be ordered and accessed at the USGS Earth Resources Observation and Science (EROS) Center Science Processing Architecture (ESPA): http://espa.cr.usgs.gov. However, getting uniform and consistent ARD remains a challenging task. Indeed, ESPA has not yet all the entire Landsat data archived available; data ordering and delivery can be long (e.g. several hours or days); and the full process from ordering to getting the data has not been automated yet. Similarly Sentinel 1 and 2 data are not routinely generated and the Sentinels Data Hub (https://scihub.copernicus.eu/) suffers from the same drawbacks identified for ESPA. This clearly limits the accessibility and ingestion processes while building and updating a DC and consequently ask to find alternative ways to generate ARD products.

Recognizing these issues, the aim of this paper is to present an approach to enable rapid data access and pre-processing to generate Analysis Ready Data. The approach has been tested and validated by significantly facilitating the generation of ARD using Landsat medium-resolution imagery allowing to build the first version of the Swiss Data Cube (SDC – http://www.swissdatacube.org).

2. Background: setting the scene for the Swiss Data Cube

Traditional environmental data (e.g. field data collection) suffers from data inconsistencies caused by changes in reporting methodologies and from gaps (e.g. missing measurements/observations). EO provides a unique opportunity to build consistent time series (i.e. same measurements conducted at regular intervals), to compare different periods of time, and to derive trends. In this regard, Landsat data records are highly valuable for Global/Regional/National/Sub-national land changes studies. Landsat observations have the longest time-series, wide extent, medium spatial, and moderate temporal resolutions (Santos & Gonçalves, 2014; Wulder et al., 2008, 2012). Moreover, in 2008 all new and archived Landsat data held by the United States Geological Survey (USGS) were made freely available over the Internet followed by a tremendous increase in scientific investigations and applications (Woodcock et al., 2008; Wulder et al., 2012). Currently, most records of the Landsat archive are publicly available on platforms like USGS EarthExplorer,[2] LandsatLook Viewer,[3] GloVis,[4] the Global Earth Observation System of Systems (GEOSS) portal,[5] or on Google[6] and Amazon[7] Cloud-based Web Services. Before the free and open access policy, a daily average of 52 scenes of Landsat data were distributed, but since 2008 this number has dramatically increased to reach a value of 5,700 scenes (Ryan, 2016). The free and open access policy of Landsat data is a brilliant example on how to maximize the return on the large investments in satellite missions (Wulder et al., 2012). The benefits of similar Open Data initiatives were emphasized by the Group of Earth Observation (GEO) when pointing out the economic benefits brought back by these initiatives: *"The economic value of geospatial data lies in its utility"* (Ryan, 2016). According to GEO, so far more than 12 million Landsat images have been delivered across 186 countries enabling users to access multiple-year scenes for the same locations (Gray Davidson, 2014).

Harnessing the full potential of these EO data and getting meaningful information requires not only massive computing resources, but also specialized algorithms and dedicated tools, which have to be brought to data instead of moving the data to processing centers (Evangelidis, Ntouros, Makridis, & Papatheodorou, 2014; Karmas, Tzotsos, & Karantzalos, 2015). Data Cubes are aiming to address these Big Data challenges by providing an architecture allowing a time-series multi-dimensional (e.g. space, time, data type) stack of spatially aligned pixels ready for analysis. The concept has been proven and validated by Geoscience Australia together with CSIRO and the National Computing Infrastructure of Australia (NCI) who implemented the Australian Geoscience Data Cube, a national/continental scale DC of thousands of terabytes of EO data (Landsat, MODIS, Sentinel-2) making it quicker and easier to provide information on environmental issues that can affect all Australians (Evans et al., 2015; Lewis et al., 2016; Purss et al., 2015). It has allowed mapping the extent of surface water across the entire Australian continent using 27 years of Landsat imagery (Mueller et al., 2016), gaining knowledge on flood dynamics over Australia (Tulbure, Broich, Stehman, & Kommareddy, 2016), or extracting the intertidal extent and topography of the Australian coastline (Sagar, Roberts, Bala, & Lymburner, 2017).

Following the work done by Australia and CEOS, Switzerland has decided to follow their work in order to benefit from the long-term Landsat data archive and monitor environmental changes through space and time. Indeed, the pressure on Switzerland's natural resources is increasing and a number of challenges (e.g. pressure on surface water, land management, biodiversity loss) need to be overcome in order to meet the needs of a growing population in a period of environmental variability (Environment Switzerland 2015, 2015). Some of these important environmental issues can be monitored using remotely sensed Earth Observations and benefit from different data archives (e.g. Landsat, Sentinel). The Swiss Data Cube (SDC) is supported by the Federal Office for the Environment (FOEN). GRID-Geneva is a partnership between the United Nations Environment Program, FOEN, and the University of Geneva (UNIGE). The SDC has been developed, implemented and operated by the GRID-Geneva. The main objectives of the SDC are to support the Swiss government for environmental monitoring and reporting and enable Swiss scientific institutions (e.g. Universities) to facilitate new insights and research using the SDC and to improve the knowledge on the Swiss environment using EO data.

3. Building the Swiss Data Cube: challenges and lessons learned

A fundamental aspect while building a DC is having ARD products ingested, stored in the database, and readily available. Considering that ARD products are not commonly generated by data providers and the fact that current delivery mechanisms are not efficient, this requires finding a procedure to routinely generate ARD ensuring that all observations stored in a Data Cube are consistent and comparable (Figure 1). Ideally this procedure must be automated as much as possible (e.g. discover, download, and pre-processing), should be able to discover and access data from different repositories (e.g. ESPA, Sentinels Data Hub), should handle different sensors (e.g. Landsat MSS, TM, ETM, OLI; Sentinel 1 SAR; Sentinel 2 MSI), and should be interoperable (e.g. to enhance reusability).

To satisfy these requirements, the Live Monitoring of Earth Surface (LiMES) framework has been used. LiMES is a framework that helps automating EO data discovery and (pre-) processing using interoperable service chains for transforming observations into information products suitable for monitoring environmental changes (Giuliani et al., 2017). This framework is designed using a combination of large storage capacities, high performance distributed computers, and interoperable standards to develop a scalable, consistent, flexible, and

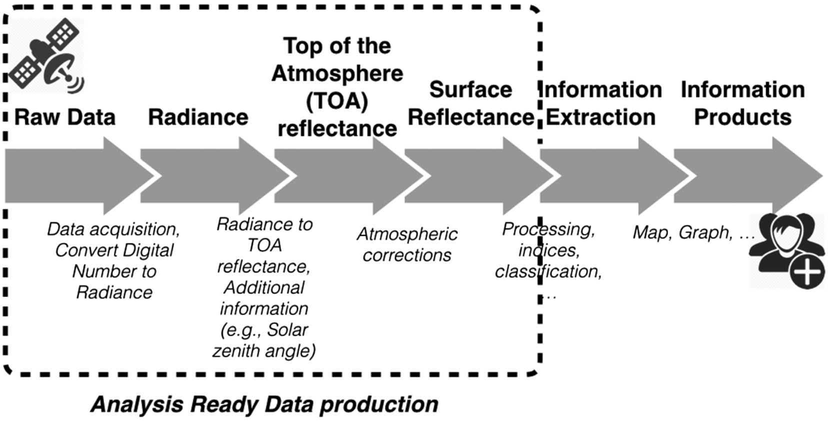

Raw Data
Radiance
Top of the Atmosphere (TOA) reflectance
Surface Reflectance
Information Extraction
Information Products
Data acquisition, Convert Digital Number to Radiance
Radiance to TOA reflectance, Additional information (e.g., Solar zenith angle)
Atmospheric corrections
Processing, indices, classification, …
Map, Graph, …
Analysis Ready Data production

Figure 1. General workflow for generating information products from observations.
Notes: Analysis Ready Data (ARD) are concerned by the four first steps (data acquisition, conversion to radiance, TOA reflectance and Surface reflectance) allowing then to analyze data and generate time-series.

efficient analysis system that can be used on various domains through decades of data for monitoring purposes (Figure 2).

While building the SDC, the LiMES framework has helped to automatically generate Landsat ARD products by overcoming the obstacles presented hereafter.

3.1. Landsat scenes discovery and availability

To get the full coverage of Switzerland, eight Landsat scenes are necessary. They are covered by the following WRS-2[8] Paths and Rows: 193/027, 194/027, 195/027, 196/027, 193/028, 194/028, 195/028, 196/028 (Figure 3).

An important issue identified is that any data repository used (e.g. USGS ESPA, GEE, AWS) does not have all the scenes available. Therefore, to have the most complete archive possible it is necessary to query all these repositories. A Python script was written to automatically generate the list of scenes available in these three repositories for a given coverage. The scene IDs will be further used in the next step of the workflow (i.e. data access and processing). To cover Switzerland, currently 3386 scenes for a total size of 867.5 GB over the period between 1984 to 2017 are available. This corresponds to 1155 Landsat 5, 1520 Landsat 7, and 711 Landsat 8 scenes. However, when looking at the availability of scenes across time, an important gap has been identified in the period 1991-1998 (Figure 4).

After discussing with CEOS, it seems that there are two explanations for this gap:

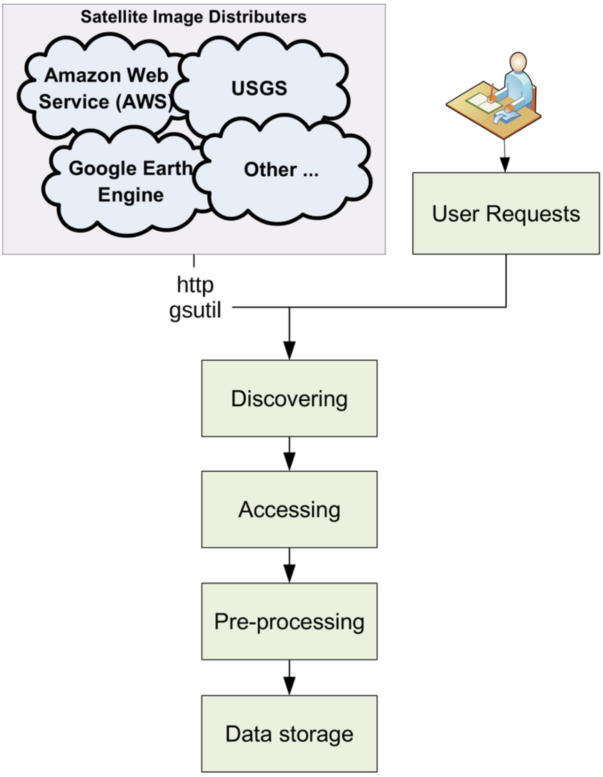


Figure 2. Processing workflow.

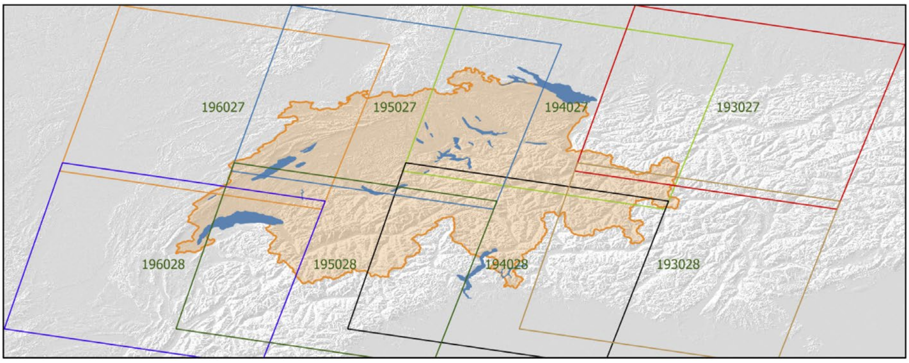


Figure 3. Landsat scenes coverage over Switzerland.

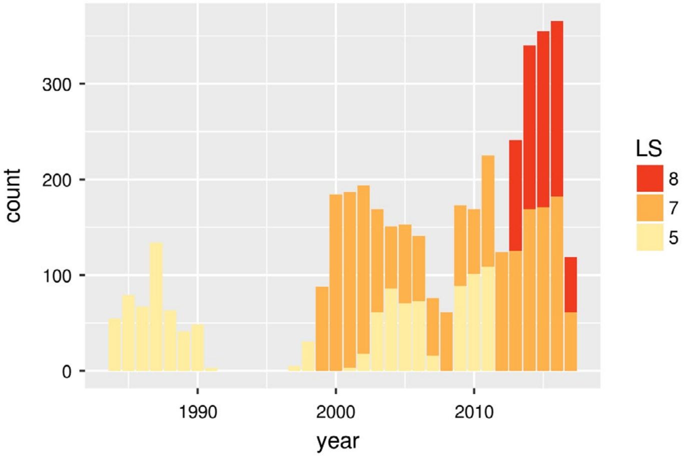


Figure 4. Data gap in Landsat scenes availability between 1991 and 1998.

(1) the European Space Agency (ESA) that is operating the Fucino (FUI) and Matera (MTI) ground stations in Italy have not yet provided the data. ESA is still loading those data to USGS and maybe reprocessing effort on their end is causing delays.

(2) there was a lack of data ingested over the period 1987-1999. Indeed, to ensure that data are of sufficient quality, scenes can be rejected for technical reasons and consequently not stored in the online archive. This might be caused for example by a paucity of Payload Correction Data (PCD). These ancillary data include information that accompanies the scenes (e.g. location of spacecraft during acquisition, pointing information) and are used for calibration purposes (Goward et al., 2006).

It seems that this gap does not only affect Switzerland. The same gap over Georgia, Moldova, Nauru, Vanuatu, and Solomon has been identified. This gap probably affects a large geographical coverage. For the SDC, the solution has been to get the data directly from ESA holdings, process them to Surface Reflectance (SR), and ingest them into the SDC.

3.2. Landsat scenes access

There are several options available to access Landsat imagery. It is possible to order images directly on web interfaces like EarthExplorer or Glovis. Another option is to use the USGS Application Programming Interface (API) to programmatically retrieve Landsat scenes. Alternatively, GEE and AWS are accessible either directly via HTTP protocol, or via gsutil (https://cloud.google.com/storage/docs/gsutil), which is a Python application that gives access to Google Cloud Storage from command lines.

To be able to perform an automatic data acquisition flow, it is necessary to test these different methodologies and find the optimal solution to efficiently download satellite imagery. Using the LiMES framework, a testing instance illustrated in Figure 2 has been implemented in Python. The testing instance was used to monitor and analyze different performance parameters of the download process from each data provider on each access method (Table 1) referred as protocol in the following lines.

The main objective of the following tests is to define and analyze which combination of provider and access method is faster and what are the possible network instabilities or the potential speed limitation after downloading a large number of scenes.

In order to avoid possible interaction between providers, each test session was performed randomly (i.e. protocols were run in random order). Three profiles of bands combinations were randomly used as follows:

- 3 bands (e.g. 2, 3, 4) to simulate the case of a simple composite process,
- 3 bands (e.g. 2, 3, 4) and panchromatic (e.g. 8) to simulate the case of a simple composite process with pan-sharpening,
- All bands (e.g. 1 to 11) to simulate the case of a more complex process.

To avoid possible download speed limitation in case of recurrent download of the same scene (noticed during the development phase), the sites (path and row) were selected randomly. The number of acquired scenes on a site randomly varied from 1 to 3.

The way providers distribute scenes varies as well. USGS and GEE are providing a single zipped file with all the bands, but they are using different compression formats (respectively tar.gz and tar.bz). AWS allows a direct access to each band in a geotiff format with the compress deflate option (meaning they can be used directly). Consequently, unzipping of data was added in the test process, as well as the process of cleaning (removing zipped file and unnecessary bands).

A comparison between the data access phases of two different providers (GEE and AWS), using gsutil, is shown in Figure 5. This figure shows the distribution of these phases (download only for AWS, shown in red; and download, unzip and clean for GEE, shown in blue) using as an example the access of all bands of the scenes with IDs: LC81980212015103LGN00, LC81980212015279LGN00, and LC81980212014260LGN00. It emphasizes the differences in the data acquisition process by providers.

Table 1. Landsat data providers and accessing methods.

Provider/access method	http	gsutil
USGS	usgs_http	N/A
GEE	gee_http	gee_gsutil
AWS	aws_ pds_http	aws_pds_gsutil

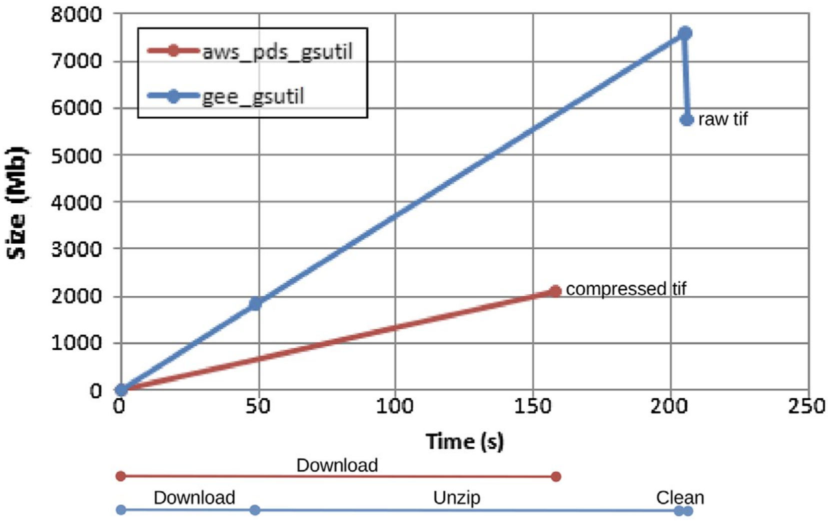


Figure 5. Acquisition phases comparison between GEE and AWS.

Figure 5 emphasizes that the size of acquired bands differs between providers. This means that for a given scene the size of downloaded data (zipped or not) and of the acquired band will not be comparable, as well as any download speed issued from these values. Consequently, the total time needed to get “ready to work” bands for a given set of scenes was used as performance parameter (158 and 206 seconds in Figure 5). To compare protocols independently of the size of the data-set, the total time of each protocol in each round of test was normalized by the minimum total time of the session (Total Time Ratio). In Figure 5, the minimum time was 158s, and the aws_pds_gsutil ratio was 1, while the gee_gsutil ratio was 1.3 (206/158). This identifies the faster combination (provider/access method) to access a Landsat scene.

The described benchmarking tests were performed on a server located in the University of Geneva Network (1 Gbps connection) on a Xen virtual machine (2.9 GHz, 8 Gb RAM, 4 CPUs). All protocols were tested in random order every three hours during one week. Each protocol was then tested 56 times with 140 scenes acquired for a total of 409 Gb of downloads.

After analyzing the benchmark tests results, we can conclude that AWS through gsutil is the best combination of provider/access method, for any bands combination, in terms of download speed, stability, storage, and data readiness (Figure 6), indicating that USGS protocol should be used as a last option when imagery is not available through other protocol.

3.3. Landsat scenes pre-processing

Pre-processing is the essential step for generating ARD products. In the case of the SDC, using the LiMES framework, this procedure is executed in two steps. The first step concerns the assessment of clouds, cloud shadows, snow, and water, generating masks accordingly. All these factors can have a significant influence on the behavior of the spectral bands of optical sensors and it is therefore fundamental to detect them before any type of analysis

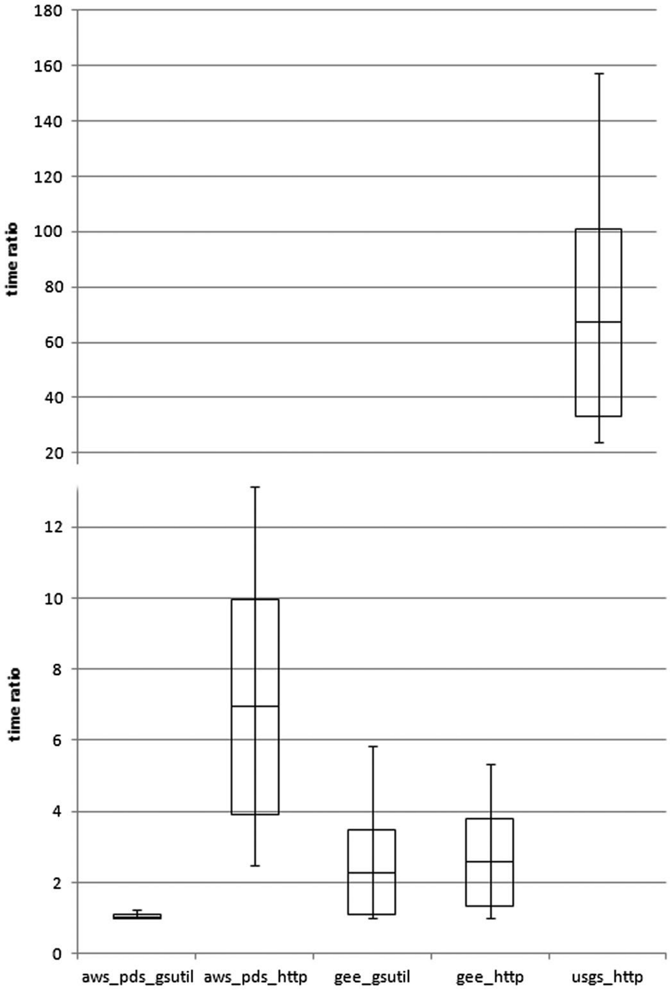


Figure 6. Total time ratio distribution per (provider, access method).

is performed. Several solutions have been tested (Fmask, ARCSI) and compared with ESPA products, the best option (the most conservative) is to use the Fmask algorithm (Zhu, Wang, & Woodcock, 2015). All these masks are stored in the DC as they can be useful when analyzing data. The second step concerns the conversion to SR by applying and correcting atmospheric effect (Song, Woodcock, Seto, Lenney, & Macomber, 2001). After testing different algorithms such as Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) (Schmidt, Jenkerson, Masek, Vermote, & Gao, 2013), R packages (RStoolbox and Landsat), Grass GIS algorithms (i.landsat.atcorr and i.atcorr), and the Simplifié Modèle d'Atmosphérique Correction (SMAC) (http://www.cesbio.ups-tlse.fr/multitemp/?page_id=2975), the final choice (based on efficiency, reliability, and easiness of integration in the workflow) was the Second Simulation of the Satellite Signal in the Solar Spectrum (6S) algorithm available in the Atmospheric and Radiometric Correction of Satellite Imagery (ARCSI[9]) software (Vermote, Tanre, Deuze, Herman, & Morcette, 1997).

Before ingesting any data in the Data Cube, data should be preprocessed. In particular, one should generate the proper metadata that will be used in the ingestion process for documentation. Useful information for systematic searching and archiving practices of data as well as information on data processing and values important for enhancing data (e.g. conversion to reflectance and radiance) are also included. However, an issue has emerged using multiple data providers. Indeed, USGS provides metadata in an XML-encoded file whereas GEE and AWS are providing metadata in text Metadata files (MTL.txt). Therefore, we modified the data preparation script in order to handle both XML and MTL files and generate the correct files used in the ingestion process.

3.4. *Data storage strategy*

A last import challenge to consider while building the SDC regards the data storage strategy and how to best resample data to ensure that all observations (i.e. pixels) have the same spatial resolution. In the case of the AGDC, they decided to resample Landsat data to a 25 meters' grid resolution (Lewis et al., 2016). This is justified by the fact that they also store MODIS data that have an original resolution of 250 meters. In the case of the SDC, all Landsat observations will be kept at a 30-meter resolution.

Another option can also be to keep the original resolutions for both sensors and store them in two different collections. However, this solution will impede the use of both sensors at the same time in a specific algorithm. In our view, this can be a limitation, especially if we consider the potential benefits of using multiple-sensors in virtual constellations to integrate data and derived information to contribute to (quantitative) analysis/measurement objectives (Wulder et al., 2015). Finally, when storing data, it is important to consider whether or not to keep the panchromatic band. Generally, this band is only used to pan-sharpen composite images (Irons, Dwyer, & Barsi, 2012). Depending on the envisioned usage, one can decide to ingest this band or not.

3.5. *Computing performances*

A major issue identified while building the SDC was the procedure to order scenes and download them. In particular, it takes a couple of hours from the manually requesting scenes before they are available for download. Within the LiMES framework, it is possible to automatically discover, access, unzip, (pre-)process, index, and ingest data in less than four minutes per scenes on average. The current infrastructure, used for the testing phase, to do all these tasks is the following: Processors Intel Xeon E5-2660 v2 @ 2.2 GHz; 8 CPUs (6CPUs used for processing, 2CPUs for system and UI); 50 Gb RAM; 2 TB Hard Drive; Linux Ubuntu 16.04. In terms of processing, six processes are executed in parallel. Indexation and ingestion cannot be performed as separate processes and therefore a bulk ingestion strategy has been implemented (e.g. by group of 50 scenes). After nine days, all Landsat data were processed, indexed, and ingested in the SDC. Finally, while moving to a production environment, an important element to consider is the optimization of computing performances. Even with 8 CPUs and 50 Gb of RAM, executing an algorithm at the sub-national scale can be long (e.g. 1 h for the NDVI algorithm). Being on the Cloud infrastructure of the University of Geneva, the SDC can be optimized for large collections by increasing the numbers of CPUs, Memory, HD, and best options to parallelize computing tasks will be explored.

4. Discussion

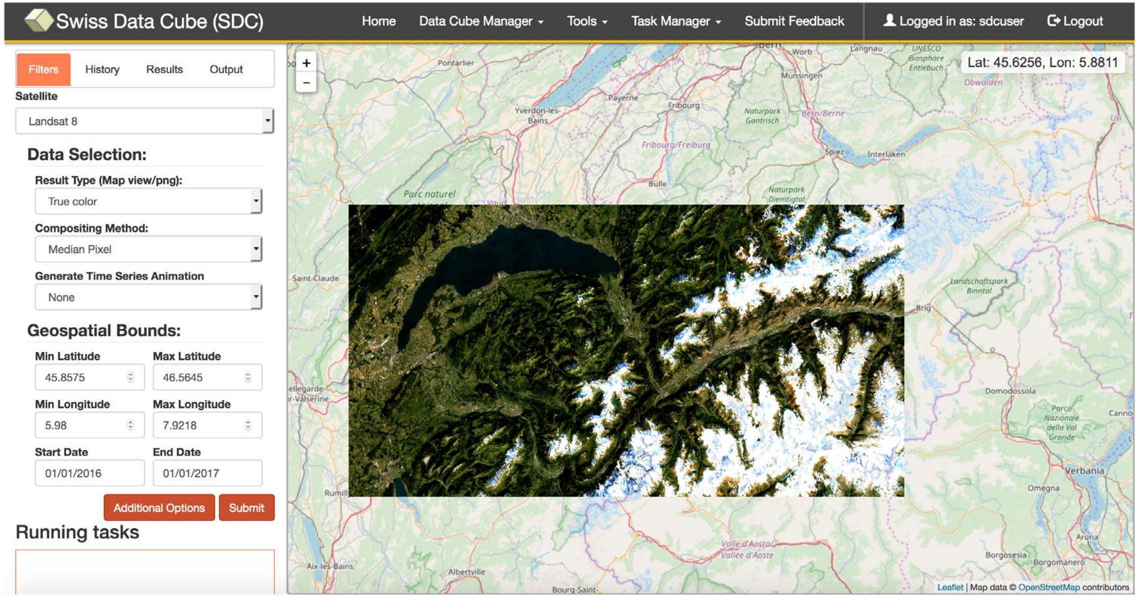
The proposed solution is simple to implement using interoperable components, and has provided the solution to some important challenges and facilitated the production of Analysis Ready Data. Currently, the SDC contains 33 years of Landsat 5,7,8 Analysis Ready Data (1984–2017) corresponding to more than 3,300 scenes. A prototype platform is running that allows the testing and visualizing several algorithms (Figure 7). Soon Sentinel 1 and 2 data will be added using the same methodology.

The major benefit of this approach is that it fully automates the discovery, access, and pre-processing of EO data. This enhances replicability and ensures homogeneity in results. Moreover, this solution is scalable, easily allowing the adding of new sensors. Finally, data access is not dependent on one repository but instead is helping to have the most complete archive of Landsat data using the best data access methodology. Recognizing the need for having efficient access to data, USGS has also developed the Landsat Global Archive Consolidation (LGAC[10]) process with the objective to have a single universally accessible and centralized global archive of analysis-ready products (Wulder et al., 2016). The proposed approach adds value because currently it is not possible to programmatically access this archive; is not yet complete; the order and download issues remain; and it allows only accessing Landsat data.

Among the identified limitations, it should be noted that especially in a multi-sensor and multi-repository context, this requires the handling of different protocols, interfaces, and APIs. One possible solution to get an homogenous access to EO data can be to go through the Global Earth Observation System of Systems (GEOSS) that provides a GEO Discovery and Access Broker (GEO DAB) API (Nativi et al., 2015). Through this API, users can discover and access various heterogeneous EO data in a seamless and homogenous way. This can be an interesting solution to investigate in order to be more flexible and access different types of EO resources. Another challenge relates to computing performances. Building a DC requires addressing the Big Data characteristics of Volume, Velocity, and Variety. To address these issues, Cloud computing appears as a viable solution for processing data and increasing the Value of these data by generating usable and useful products (Yang, Huang, Li, Liu, & Hu, 2017; Yang, Yu, Hu, Yongyao, & Yun, 2017).

The next challenge relates to the usability of the Swiss Data Cube. After the first run of ingestion, an automated update procedure is planned that will regularly add all newly available Landsat scenes. This will ensure that the SDC is always up-to-date and this will in turn improve the quality of the products (e.g. algorithms) that will be implemented. Developing specific algorithms to generate useful information products for supporting various governmental offices in decision-making is a fundamental task to leverage the potential of EO data. Furthermore, it is also important to continue the research effort associated with this new technology. For example, applying machine learning techniques can be valuable to enhance extraction of information and classification (Camps-Valls, 2009; Lary, Alavi, Gandomi, & Walker, 2016). Additionally, DC being a new promising technology, capacity building efforts should be considered to ensure both technology transfer and to enable users to become familiar with this technology and raise awareness on how this can be useful in leveraging the potential of EO data (Desconnets et al., 2017; Giuliani et al., 2016).

In a broader context, the DC technology can be extremely useful for environmental monitoring by providing insights into phenomena that are otherwise impossible to measure. For



Satellite: Landsat 8

Data Selection:

Result Type (Map view/png): True color

Compositing Method: Median Pixel

Generate Time Series Animation: None

Geospatial Bounds:

Min Latitude: 45.8575

Max Latitude: 46.5645

Min Longitude: 5.98

Max Longitude: 7.9218

Start Date: 01/01/2016

End Date: 01/01/2017

Running tasks

Figure 7. Processing results with the SDC. Landsat8 cloud-free true color composite for the year 2016 over the south-western part of Switzerland.

example, it can help in answering the needs of specific scientific communities like Biodiversity where spatial and temporal resolution, long-term data continuity, and data accessibility are major concerns (Kuenzer et al., 2014).

Moreover, with the concept of Essential Variables, new opportunities are emerging for monitoring the state of biodiversity and ecosystems in a more systematic way and this could be expended to national or international indicators (e.g. Aichi targets) (Vihervaara et al., 2017). Finally, Earth Observation can be a valuable resource for monitoring Sustainable Development Goals (SDG) (Anderson, Ryan, Sonntag, Kavvada, & Friedl, 2017). The Group on Earth Observations (GEO) together with CEOS have recently demonstrated that EO can be a promising complement to traditional national statistics (Group on Earth Observations, 2017). In particular, it can provide data at different scales; it can help track progress toward specific policy objectives and targets thanks to long time series and continuity; it provides consistent and effective means of comparison among different countries (can contribute to more detailed and more harmonized indicators, without requiring any additional reporting by countries); and offers a variety of measurements and therefore potential to generate useful information products.

The current trend in EO (e.g. open data policies, cloud computing, data cubes) is expanding the use of EO data beyond specialized scientific communities. These developments offer new opportunities for improving the scope and strengthen environmental data and indicators. In particular, efficient environmental policies require effective evidences that take into account both the spatial distribution of environmental issues and economic activity. EO data can provide an invaluable and timely source of information across various scales (e.g. local to global) helping to overcome current data gaps and incoherent time series needed for a strong evidence-based environmental policy process. It can support harmonizing reporting on environmental issues as well as being used with other geospatial data such as demographic, economic, or administrative data to make indicators and analysis more relevant and targeted (Lehmann et al., 2017; OECD, 2015).

5. Conclusions

Data Cubes are revolutionizing the way users can work with EO data. It is a disruptive technology that is significantly transforming the way that users interact with EO Data. It has the potential to routinely transform Earth Observations into useful and actionable information for users. To reduce the processing burden on users, generating Analysis Ready Data is a fundamental requirement. ARD products minimize the time and scientific knowledge required to access and prepare satellite data having consistent and spatially aligned calibrated surface reflectance observations. Current methods in ARD delivery (e.g. ESPA, SciDataHub) are not satisfactory principally because they are not commonly generated by data providers and because the process to order and download data are slow.

The proposed approach makes use of the LiMES framework to build interoperable data processing chains for generating ARD products. This methodology has been tested in building the Swiss Data Cube, a country scale DC for monitoring the environment in space and time, and has allowed to efficiently download, pre-process, and ingest thousands of Landsat scenes in a couple of days. In particular, it has significantly facilitated the generation of ARD products from various sensors (Landsat 5, 7, 8). Google- and Amazon Cloud-based Web Services appear to be very efficient solutions to gather Landsat data allowing the downloading of data faster than any other traditional data delivery mechanisms (e.g. EarthExplorer). The proposed solution lowers the barrier to ARD product generation by automating pre-processing steps and allowing users to concentrate on data analytics to support the utilization of the growing volume of EO data. This is a key requirement to enable unlocking the information power of Big EO data, expand the number of potential EO data users, and allowing EO data to become an essential asset for environmental monitoring.

Notes

1. http://www.rasdaman.org.
2. https://earthexplorer.usgs.gov.
3. https://landsatlook.usgs.gov.
4. https://glovis.usgs.gov.
5. http://www.geoportal.org.
6. https://cloud.google.com/storage/docs/public-data-sets/landsat.
7. https://aws.amazon.com/public-data-sets/landsat.
8. https://landsat.gsfc.nasa.gov/the-worldwide-reference-system/.
9. http://rsgislib.org/arcsi/.
10. https://landsat.usgs.gov/landsat-global-archive-consolidation-lgac.

Data availability statement

The data that support the findings of this study are available from the corresponding author, [GG], upon reasonable request.

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 like also to acknowledge Dr Stuart Minchin (GeoScience Australia) and Dr Brian Killough (CEOS) for their valuable support in implementing and developing the Swiss Data Cube. We thank Martin Lacayo for the reviewing in an earlier draft. The views expressed in the paper are those of the authors and do not necessarily reflect the views of the institutions they belong to.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Anderson, K., Ryan, B., Sonntag, W., Kavvada, A., & Friedl, L. (2017). Earth observation in service of the 2030 agenda for sustainable development. *Geo-spatial Information Science, 20*, 77–96. doi:10.1080/10095020.2017.1333230

Baumann, P. (2017). *The Datacube Manifesto*. Retrieved from https://earthserver.eu/tech/datacube-manifesto

Baumann, P., Furtado P., Ritsch R., & Widmann, N. (1997). The RasDaMan approach to multidimensional database management. In *SAC'97 Proceedings of the 1997 ACM symposium on Applied computing* (pp. 166–173). 331732: ACM. doi:10.1145/331697.331732

Baumann, P., Mazzetti, P., Ungar, J., Barbera, R., Barboni, D., Beccati, A., … Wagner, S. (2016). Big data analytics for earth sciences: The earthserver approach. *International Journal of Digital Earth, 9*(1), 3–29. doi:10.1080/17538947.2014.1003106

Baumann, P., Rossi, A. P., Clements, O., Dumitru, A., Evans, B., Hogan, P., … Wagemann J. (2016). *Fostering cross-disciplinary earth science through datacube analytics* (p. 32). Retrieved from https://www.rd-alliance.org/sites/default/files/earthserver-chapter_ssr-issi.pdf

Camps-Valls, G. (2009). Machine learning in remote sensing data processing. In *2009 IEEE international workshop on machine learning for signal processing* (pp. 1–6). doi:10.1109/MLSP.2009.5306233

CEOS. (2017). *The CEOS open data cube initiative*. Retrieved from https://docs.wixstatic.com/ugd/f9d4ea_1aea90c5bb7149c8a730890c0f791496.pdf

Desconnets, J-C., Giuliani, G., Guigoz, Y., Lacroix, P., Mlisa, A., Noort, M., … Searby, N. D. (2017). GEOCAB portal: A gateway for discovering and accessing capacity building resources in Earth Observation. *International Journal of Applied Earth Observation and Geoinformation, 54*, 95–104. doi:10.1016/j.jag.2016.09.010

Douglas McCuistion, J., & Birk R. (2005). From observations to decision support: The new paradigm for satellite data. *Acta Astronautica, 4th IAA international symposium on small satellites for earth observation, 56* (1): 5–8. doi:10.1016/j.actaastro.2004.09.046

Environment Switzerland 2015. (2015). *er2015. State of the environment*. Bern: Swiss Federal Council. Retrieved from https://www.bafu.admin.ch/bafu/en/home/state/publications-on-the-state-of-the-environment/environment-switzerland-2015.html.

Evangelidis, K., Ntouros, K., Makridis, S., & Papatheodorou, C. (2014). Geospatial services in the cloud. *Computers & Geosciences, 63*, 116–122. doi:10.1016/j.cageo.2013.10.007

Evans, B., Wyborn, L., Pugh, T., Allen, C., Antony, J., Gohar, K., … Bell, G. (2015). The NCI high performance computing and high performance data platform to support the analysis of petascale environmental data collections. In Ralf Denzer, Robert M Argent, Gerald Schimak, & Jiří Hřebíček (Eds.), *Environmental software systems. Infrastructures, services and applications: 11th IFIP WG 5.11 International Symposium, ISESS 2015* (pp. 569–577), Melbourne, March 25–27, 2015. Cham: Springer. doi:10.1007/978-3-319-15994-2_58

Flach, M., Gans, F., Brenning, A., Denzler, J., Reichstein, M., Rodner, E., … Mahecha, M. D. (2016). Multivariate anomaly detection for earth observations: A comparison of algorithms and feature extraction techniques. *Earth System Dynamics Discussions, 2016*(November), 1–27. doi:10.5194/esd-2016-51

Giuliani, G., Dao, H., De Bono, A., Chatenoux, B., Allenbach, K., De Laborie, P., … Peduzzi, P. (2017). Live monitoring of earth surface (LiMES): A framework for monitoring environmental changes from earth observations. *Remote Sensing of Environment*. doi:10.1016/j.rse.2017.05.040

Giuliani, G., Lacroix, P., Guigoz, Y., Roncella, R., Bigagli, L., Santoro, M., … Lehmann, A. (2016). Bringing GEOSS services into practice: A capacity building resource on spatial data infrastructures (SDI). *Transactions in GIS, 21*(4), 811–824. doi:10.1111/tgis.12209

Gómez, C., White, J. C., & Wulder, M. A. (2016). Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing, 116*(June), 55–72. doi:10.1016/j.isprsjprs.2016.03.008

Gore, A. (1998). The digital earth: Understanding our planet in the 21st century. *Australian Surveyor, 43*(2), 89–91.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*. doi:10.1016/j.rse.2017.06.031

Goward, S., Arvidson, T., Williams, D., Faundeen, J., Irons, J., & Franks, S. (2006). Historical record of Landsat global coverage. *Photogrammetric Engineering & Remote Sensing, 72*(10), 1155–1169. doi:10.14358/PERS.72.10.1155

Gray Davidson, O. (2014 January 6). The group on earth observations looks toward a second decade of data sharing. *Earthzine*. Retrieved from https://earthzine.org/2014/01/06/the-group-on-earth-observations-looks-toward-a-second-decade-of-data-sharing/

Group on Earth Observations. (2017). *Earth observations in support of the 2030 agenda for sustainable development*.

Irons, J. R., Dwyer, J. L., & Barsi, J. A. (2012). The next Landsat satellite: The Landsat data continuity mission. *Remote Sensing of Environment, Landsat Legacy Special Issue, 122*(July), 11–21. doi:10.1016/j.rse.2011.08.026

Karmas, A., Tzotsos A., & Karantzalos, K. (2015). Scalable geospatial web services through efficient, online and near real-time processing of earth observation data. In *2015 IEEE First International Conference on Big Data Computing Service and Applications*, 194–201. doi:10.1109/BigDataService.2015.49

Killough, B. (2016). *CEOS land surface imaging analysis ready data (ARD) description document*.

Kuenzer, C., Ottinger, M., Wegmann, M., Guo, H., Wang, C., Zhang, J., … Wikelski, M. (2014). Earth observation satellite sensors for biodiversity monitoring: Potentials and bottlenecks. *International Journal of Remote Sensing, 35*(18), 6599–6647. doi:10.1080/01431161.2014.964349

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers, Special Issue: Progress of Machine Learning in Geosciences, 7*(1), 3–10. doi:10.1016/j.gsf.2015.07.003

Lehmann, A., Chaplin-Kramer, R., Lacayo, M., Giuliani, G., Thau, D., Koy, K., … Sharp, Jr., R. (2017). Lifting the information barriers to address sustainability challenges with data from physical geography and earth observation. *Sustainability, 9*(6), 858. doi:10.3390/su9050858

Lewis, A., Lymburner, L., Purss, M. B. J., Brooke, B., Evans, B., Ip, A., Oliver, S. (2016). Rapid, high-resolution detection of environmental change over continental scales from satellite data – The Earth Observation Data Cube. *International Journal of Digital Earth, 9*(1), 106–111. doi:10.1080/17538947.2015.1111952

Lewis, A., Oliver, S., Lymburner, L., Evans, B., Wyborn, L., Mueller, N., … Wu, W. (2017). The Australian Geoscience Data Cube – Foundations and Lessons Learned. *Remote Sensing of Environment,*. doi:10.1016/j.rse.2017.03.015

Mueller, N., Lewis, A., Roberts, D., Ring, S., Melrose, R., Sixsmith, J., … Ip, A. (2016). Water observations from space: Mapping surface water from 25 years of Landsat imagery across Australia. *Remote Sensing of Environment, 174*(March), 341–352. doi:10.1016/j.rse.2015.11.003

Nativi, S., Mazzetti, P., Santoro, M., Papeschi, Fabrizio, Craglia, M., & Ochiai, O. (2015). Big data challenges in building the global earth observation system of systems. *Environmental Modelling & Software, 68*, 1–26. doi:10.1016/j.envsoft.2015.01.017

OECD. (2015). *Strengthening national statistical systems to monitor global goals* (p 8). 2017. Earth Observation for Decision-Making. Author.

Purss, M. B. J., Lewis, A., Oliver, S., Ip, A., Sixsmith, J., Evans, B., … Chan, T. (2015). Unlocking the Australian Landsat archive – From dark data to high performance data infrastructures. *GeoResJ, Rescuing Legacy Data for Future Science, 6*(June), 135–140. doi:10.1016/j.grj.2015.02.010

Rockstrom, J., Steffen, W., Noone, K., Persson, A., Chapin, F. S., Lambin, E., & Foley, J. (2009). Planetary boundaries: Exploring the safe operating space for humanity. *Ecology and Society, 14*(2). Retrieved from http://www.ecologyandsociety.org/vol14/iss2/art32/

Ryan, B. (2016). The benefits from open data are immense. *Geospatial World*, 72–73.

Sagar, S., Roberts, D., Bala, B., & Lymburner, L. (2017). Extracting the intertidal extent and topography of the Australian coastline from a 28 year time series of Landsat observations. *Remote Sensing of Environment, 195*(June), 153–169. doi:10.1016/j.rse.2017.04.009

Santos, N. D., & Gonçalves, G. (2014). Remote sensing applications based on satellite open data (Landsat8 and Sentinel-2). In Conferencia Nacional de Geodecisao, Barreiro, 15 y 16 de mayo de 2014.

Schmidt, G., Jenkerson, C., Masek, J., Vermote, E., & Gao, F. (2013). *Landsat ecosystem disturbance adaptive processing system (LEDAPS) algorithm description*. Reston, VA: USGS Publications Warehouse. Retrieved from http://pubs.er.usgs.gov/publication/ofr20131057

Song, C., Woodcock, C. E., Seto, K. C., Lenney, M. P., & Macomber, S. A. (2001). Classification and change detection using Landsat TM data. *Remote Sensing of Environment, 75*(2), 230–244. doi:10.1016/S0034-4257(00)00169-3

Tulbure, M. G., Broich, M., Stehman, S. V., & Kommareddy, A. (2016). Surface water extent dynamics from three decades of seasonally continuous Landsat time series at subcontinental scale in a semi-arid region. *Remote Sensing of Environment, 178*(June), 142–157. doi:10.1016/j.rse.2016.02.034

Vermote, E. F., Tanre, D., Deuze, J. L., Herman, M., & Morcette, J. J. (1997). Second simulation of the satellite signal in the solar spectrum, 6S: An overview. *IEEE Transactions on Geoscience and Remote Sensing, 35*(3), 675–686. doi:10.1109/36.581987

Vihervaara, P., Auvinen, A.-P., Mononen, L., Törmä, M., Ahlroth, P., Anttila, S., … Virkkala, R. (2017). How essential biodiversity variables and remote sensing can help national biodiversity monitoring. *Global Ecology and Conservation, 10*(April), 43–59. doi:10.1016/j.gecco.2017.01.007

Woodcock, C. E., Allen, R., Anderson, M., Belward, A., Bindschadler, R., Cohen, W., … Wynne, R. (2008). Free access to Landsat imagery. *Science, 320*(5879), 1011a. doi:10.1126/science.320.5879.1011a

Wulder, M. A., Hilker, T., White, J. C., Coops, N. C., Masek, J. G., Pflugmacher, D., & Crevier, Y. (2015). Virtual constellations for global terrestrial monitoring. *Remote Sensing of Environment, 170*(December), 62–76. doi:10.1016/j.rse.2015.09.001

Wulder, M. A., Masek, J. G., Cohen, W. B., Loveland, T. R., & Woodcock, C. E. (2012). Opening the archive: How free data has enabled the science and monitoring promise of Landsat. *Remote Sensing of Environment, 122*, 2–10. doi:10.1016/j.rse.2012.01.010

Wulder, M. A., White, J. C., Goward, S. N., Masek, J. G., Irons, J. R., Herold, M., … Woodcock, C. E. (2008). Landsat continuity: Issues and opportunities for land cover monitoring. *Remote Sensing of Environment, 112*(3), 955–969. doi:10.1016/j.rse.2007.07.004

Wulder, M. A., White, J. C., Loveland, T. R., Woodcock, C. E., Belward, A. S., Cohen, W. B., … Roy, D. P. (2016). The global Landsat archive: Status, consolidation, and direction. *Remote Sensing of Environment, 185*, 271–283. doi:10.1016/j.rse.2015.11.032

Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2017). Big data and cloud computing: Innovation opportunities and challenges. *International Journal of Digital Earth, 10*(1), 13–53. doi:10.1080/17538947.2016.1239771

Yang, C., Yu, M., Hu, F., Yongyao, J., & Yun, L. (2017). Utilizing cloud computing to address big geospatial data challenges. *Computers, Environment and Urban Systems, Geospatial Cloud Computing and Big Data, 61*(January), 120–128. doi:10.1016/j.compenvurbsys.2016.10.010

Zhu, Z., Wang, S., & Woodcock, C. E. (2015). Improvement and expansion of the Fmask algorithm: Cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images. *Remote Sensing of Environment, 159*(March), 269–277. doi:10.1016/j.rse.2014.12.014