Cooperative Research, Development and Operation of Earth Observation Services in the Framework of Copernicus

Werner Mücke[1], Wolfgang Wagner[1,2], Christian Briese[1], Michael Aspetsberger[3]

[1]EODC, Wien · werner.muecke@eodc.eu
[2]Department für Geodäsie und Geoinformation, TU Wien
[3]Catalysts GmbH, Linz

Abstract: *The Sentinel satellites of the Copernicus space programme generate huge amounts of data at high spatio-temporal resolutions. Higher resolutions lead to massive or big data volumes and higher demands on processing, storage and distribution infrastructure. This contribution provides an insight into the challenges for research in Earth Sciences, as well as the processing, archiving and distribution of big EO data acquired through the Sentinel missions. An overview on ESA distribution and exploitation strategies is given, followed by an introduction on the technological and organisational concept being currently realised in the Austrian data centre initiative EODC. Next to the aspect of community building and collaboration, the focus is laid on information exploitation and distribution of results to the users.*

Keywords: *Sentinel, Earth observation data, long term archive, information extraction, data distribution*

Zusammenfassung: Die Sentinel-Satellitenmissionen des Copernicus-Programms produzieren eine Datenmenge in beispielloser räumlicher und zeitlicher Auflösung. Diese Auflösungen gehen mit hohen Datenraten einher, welche die Erdbeobachtung und deren Anwendungen in die Problematik von big data führen. Die Anforderungen an Prozessierungs- und Speicherkapazität, sowie Verteilungsmechanismen steigen parallel dazu. Dieser Beitrag liefert Einblicke in die Herausforderungen, welche für die Geowissenschaften durch die Verwendung von Sentineldaten entstehen. Ein Überblick über die Datenverteilungs- und Nutzungsstrategie wird gegeben. Das Konzept des Österreichischen Datenzentrums EODC wird vorgestellt. Neben den Aspekten der Gemeinschaftsbildung und Zusammenarbeit liegt der Fokus auf Datennutzung und Verteilung von Ergebnissen durch das EODC-Rahmenwerk.

Schlüsselwörter: Sentinel, Erdbeobachtung, Langzeitarchiv, Datennutzung, Datenverteilung

1 Copernicus and the Sentinel Missions

1.1 The Era of Big Earth Observation Data and its Challenges

Copernicus is a cooperative initiative of the European Space Agency (ESA) and the European Commission (EC), introducing a generation of novel satellite missions. The so-called Sentinel satellites are providing Earth Observation (EO) data from a range of different sensors at unprecedented combinations of temporal and spatial resolutions. To guarantee a maximum of coverage and data availability, several Sentinel missions are going to operate two satellites in orbit simultaneously. Current estimates have shown that the synthetic aperture radar mission Sentinel-1 (S-1) will acquire around 25 Terabytes (TB) per year over land (considering only Interferometric Wide Swath mode, ground range detected, medium resolution) (ESA 2016), which is already more than ENVISAT's Advanced Synthetic Aperture Radar (ASAR) acquired over its complete 10 years lifetime of the ENVISAT satellite (WAGNER 2015). With

AGIT – Journal für Angewandte Geoinformatik, 2-2016, S. 38-45. © Wichmann Verlag, VDE VERLAG GMBH · Berlin · Offenbach. ISBN 978-3-87907-622-2, ISSN 2364-9283, eISSN 2509-713X, doi:10.14627/537622006.

two satellites in orbit, and considering all data products and formats, S-1 will provide raw data in the amount of 1,8 TB per day and roughly 1 Petabyte (PB) over the 7-year life span of the mission. Similar amounts accumulate for the Sentinel-2 (S-2) mission (1.6 TB daily) (DRUSCH et al. 2012) with its high-resolution optical imager that is planned to serve in combination and continuation of the Landsat and SPOT missions.

With the increased spatio-temporal resolution of the sensors, also the complexity of the observed phenomena is increasing, which leads to a need for more comprehensive analysis and information retrieval algorithms (WAGNER et al. 2014). But for the EO community this also opens up new opportunities for data intensive research questions, such as global or long time-series analysis. For the involved IT infrastructure this poses huge challenges for storing, transferring, processing and delivery of these big EO data.

These developments require significant IT infrastructure capacities to download, process, archive, and distribute the EO data and products based thereon. However, an average organization's own infrastructure investments are typically limited and not ad-hoc expandable. Therefore, in order to exploit the richness of the data, sharing infrastructure and transferring of processing towards a scalable multi-user cloud platform with on-demand hardware booking and further data availability for the interlinking of the data sets are becoming a more and more interesting option. As a result, a paradigm change in the entire EO (processing) landscape can be observed: instead of bringing the data to the software, the software needs to be brought to the data. From a practical point of view, two important questions are raised for the every-day users: (1) How can one access these data and information derived thereof, and (2) how can one efficiently and economically work with it?

1.2 Distribution and Exploitation Strategies

ESA and EC have started to organise the respective Sentinel EO ground segment and downstream activities. Due to the sheer amount and information richness of Sentinel data, combined with the inherent requirements for processing, storage and distribution of the data, it becomes a necessity to move away from the traditional centralized approaches of data access and delivery. The current strategy foresees that instead of a single data centre handling data production and delivery on its own, it rather promotes a federation and cooperation of platforms supporting and complementing each other.

ESA and EC have both made significant financial investments during the development, implementation and operation of the Sentinel missions' space segments, which refers to the satellite infrastructure itself. The realisation of the respective ground segments, where data reception, quality control, pre-processing and re-distribution of data take place, are a component in itself and are established in parallel. The ground segment infrastructure consists of the so-called core ground segment and the collaborative ground segments (ESA 2016a). Their main task being the delivery of the Sentinel core products, which comprise data sets in certain pre-processing stages for all Sentinel missions (ESA ESRIN 2016) (Fig: 1a). Sentinel core products are delivered free-of-charge to all users, regardless of their scientific, public or commercial background. This open data policy is meant to provide systematic and straightforward access to the observations, with the ultimate aim to foster innovation, business and employment.

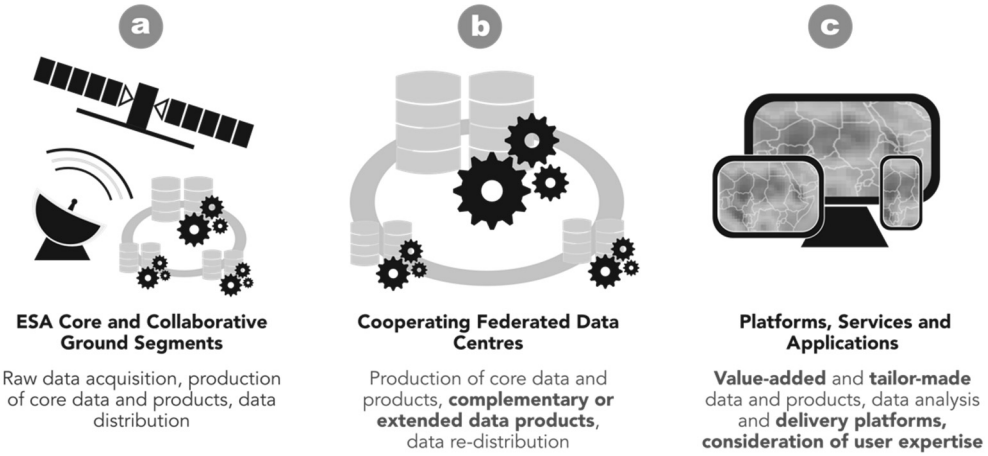

Fig. 1: (a) Data acquisition and processing in ESA ground segment, (b) production of core and complementary/extended products at cooperating/federated data centres, and (c) distribution of value-added, tailor-made products

For the implementation of the collaborative ground segments, ESA and EC rely heavily on the EU member states and funding of third parties in order to develop, implement and optimize distribution and exploitation schemes for the Sentinel EO Data (ESA 2016b). In principle, the data supply through the core and collaborative ground segments stops at the processing levels of the aforementioned core products. Currently, several private and public initiatives are being developed and setup (SINERGISE 2016, CLOUDEO 2016, SENSYF 2016) in order to support the handling of the enormous amount of data. These data centres and platforms usually cooperate with the respective ground segments, but are also aimed at creating value-added data and services. Thereby they are extending the Sentinel ground segment and core product set (Fig. 1b) by offering data and products tuned for thematic, regional, and technological requirements of specific applications (Fig. 1c).

2 A Platform for Science, Development and Operations

2.1 EODC – Connecting the Community

For organisations involved in the distribution and exploitation network for Sentinel data, the focus has to shift from solely being a data provider to additionally becoming an information service provider. Ideally, these services include (i) data procurement and provision, (ii) know-how and strategies for data analysis, as well as (iii) tools, hard- and software for data processing. One such organization is the EODC Earth Observation Data Centre for Water Resources Monitoring (EODC), which was founded as a public-private partnership in 2014 (https://www.eodc.eu/). In EODC's cooperation model, partners from the private and public sectors join together into communities sharing similar foci, research and development goals, e. g. the federation of infrastructure resources or joint software developments. The idea is for the individual organization to take advantage of the experience and skills of all involved

partners and to participate in collaborative development processes, while at the same time gaining increased external visibility via the larger group.

In order to support community building, EODC organises regular community events for information exchange. In an effort to pro-actively stimulate cooperation, several tools are implemented on the EODC IT platform, such as a shared code library (based on GitLab) or a knowledge base (similar to wiki) covering important aspects on the whole range of applications of interest to EODC partners.

2.2 The Basic Concept of EODC

EODC builds its IT capacities on three basic pillars (Fig 2): (1) NORA, which stands for "Near real-time operations and rolling archive", (2) SIDP, the "Science Integration and Development Platform", and (3) GTR, the "Global testing and reprocessing facility". NORA offers four central services: (1) data input from external (non-EODC) satellite data archives, (2) output of these data to the EODC data warehouse, (3) status monitoring of the system's overall performance and data transfer, as well as (4) near real-time (NRT) processing of selected products (e. g. a 500x500m^2 soil moisture data set from S-1). SIDP is EODC's fully equipped and flexible cloud infrastructure, where pre-configured virtual machines and other services (e. g. continuous integration and deployment) are being hosted, supporting the remote development and testing of methods and source code for novel EO services. SIDP is also intended to test new or revised algorithms on small scale use cases. On the otherhand, the GTR is a top500-ranking supercomputer (top500.org, 2016), namely the Vienna Scientific Cluster 3 (VSC 2016), and it is intended for large scale processing. These three components are complemented by a shared code library and a Petabyte (PB)-scale data archive (i. e. the EODC data warehouse), which is physically co-located with SIDP and GTR and connected via InfiniBand network technology to minimise transfer times and increase I/O. By 2018, the EODC archive will be capable of hosting up to 20 PB of EO data and derived products.

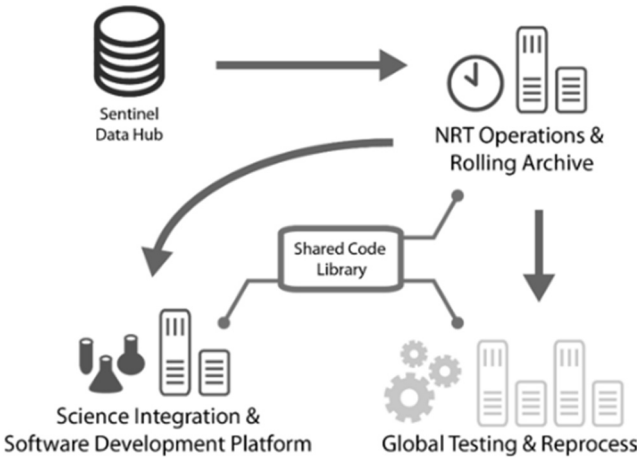

Fig 2: Basic concept and key parts of EODC's IT infrastructure

2.3 Accessing the Data Warehouse

EODC has been acquiring S-1 and S-2 data from day one and, in contrast to ESA's scientific data hub, is providing them through its long-term data archive. On average, the S-1 data are available in the archive approximately 2.5 hours after their processing time and 6.25 hours after acquisition time. To this day, a total of 117.625 S-1 and 13.649 S-2 acquisitions are hosted in the data pool (status: End January 2016). Once a user is logged on to the EODC network, he may access the data either through SIDP or GTR. Search and discovery are provided via a meta database that supports spatio-temporal search functions and scripting. Conventional access is established by means of file system access through the dedicated virtual machines of the users and http-export (e. g. webdav). Evaluations are currently ongoing to provide OGC WM(T)S and WCS interfaces for selected products, as such that users may not only download or view pre-created maps, but also directly access the underlying data using their preferred desktop applications.

2.4 Exploiting EO Information

The workflow from raw or pre-processed to refined and value-added data is as follows: Starting on SIDP, users or user-groups (i. e. communities) develop their methods, while testing it on small data sets (typically hundreds of Mega- and up to few Gigabytes). As soon as their code base has reached a certain state of maturity, they transfer to GTR, where the same codebase is available for testing on larger areas, usually regional or even global scale. If a NRT product is targeted, the developers move to NORA for the establishment of an operational NRT service. While working with SIDP and NORA is basically similar to working on personal computers, as both offer the possibility to host user-definable virtual machines, accessing the supercomputer GTR is a completely different environment. It involves a minimum level of understanding for high performance computing, parallelisation and certain programming experience. EODC is currently developing an integrated virtual workspace to create an environment that supports a variety of users with differing skill levels (Fig. 3). In this regard, a *basic* user would only be interested in accessing pre-configured processing chains or even pre-calculated products through web-based delivery services, while *advanced* users would like to adapt modules (i. e. building blocks) of processing chains according to their own wishes. The third level would be the *expert* users, being capable of designing new processing chains and modules. This workspace will offer highest flexibility and transparency to its users, giving them the freedom to adjust processing workflows to their special needs, and giving them enough insight in the system so that they can trust the data, the methods and the results.

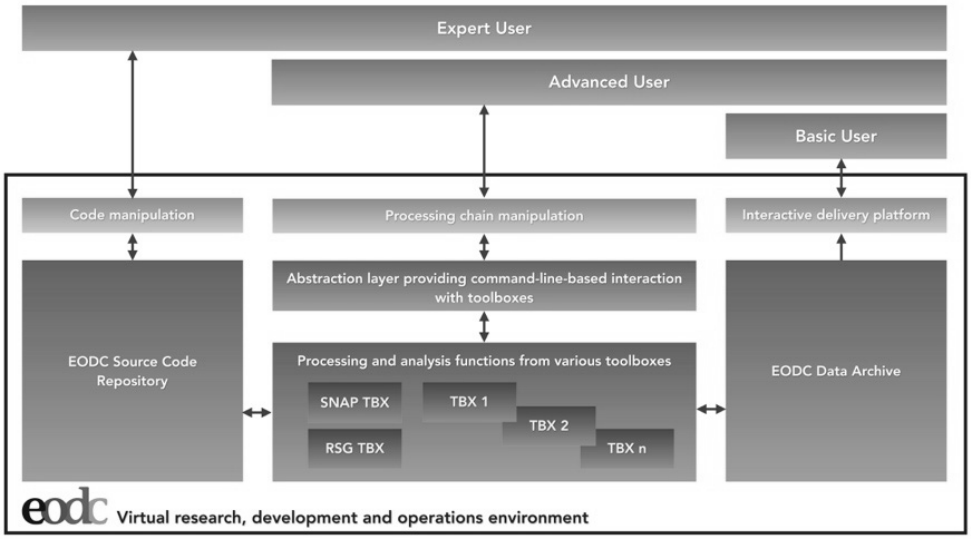

Fig. 3: Scheme of EODC's virtual workspace

To access the supercomputer VSC-3, a simplified processing submission and task scheduling interface (Fig. 4) is currently implemented. A browser-based graphical user interface allows users to select pre-defined processing chains from the EODC code library, use their own configuration files to define settings for included algorithms, and select the number of computing nodes they would like to employ.

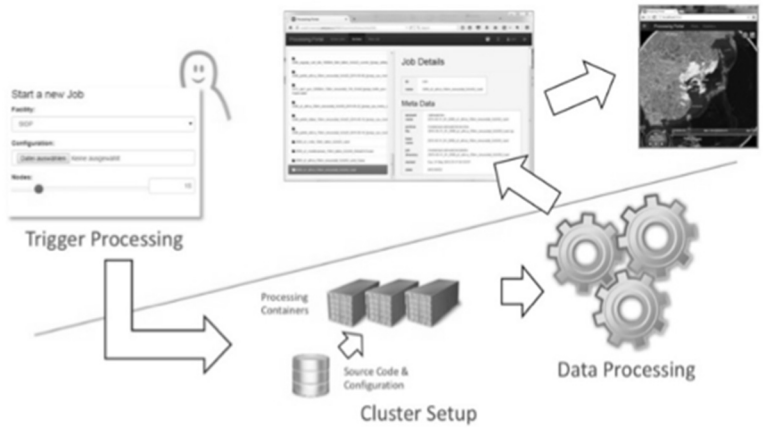

Fig. 4: Illustration of simplified process management through the EODC workspace on the VSC-3 supercomputer

After the processing is triggered by the user, everything else is done automatically: the respective codebase is pulled from the library and together with the configuration files processing containers are built, representing sand-boxed environments including all software packages necessary for successful computation. During data processing, the browser-based

job monitor gives access to task information, such as estimated run time. On job completion, the users are notified and may inspect their results using the available data viewer.

2.5 Distributing Results

Each EODC partner has the possibility to use a dedicated and private location on the EODC archive for storage of value-added data produced on the EODC platform. At the partner's discretion, these data may stay private for personal use, or may be shared with the community. Such data would first undergo a quality check carried out by EODC and would then be available through the general archive, instead of consuming partner's dedicated disk space.

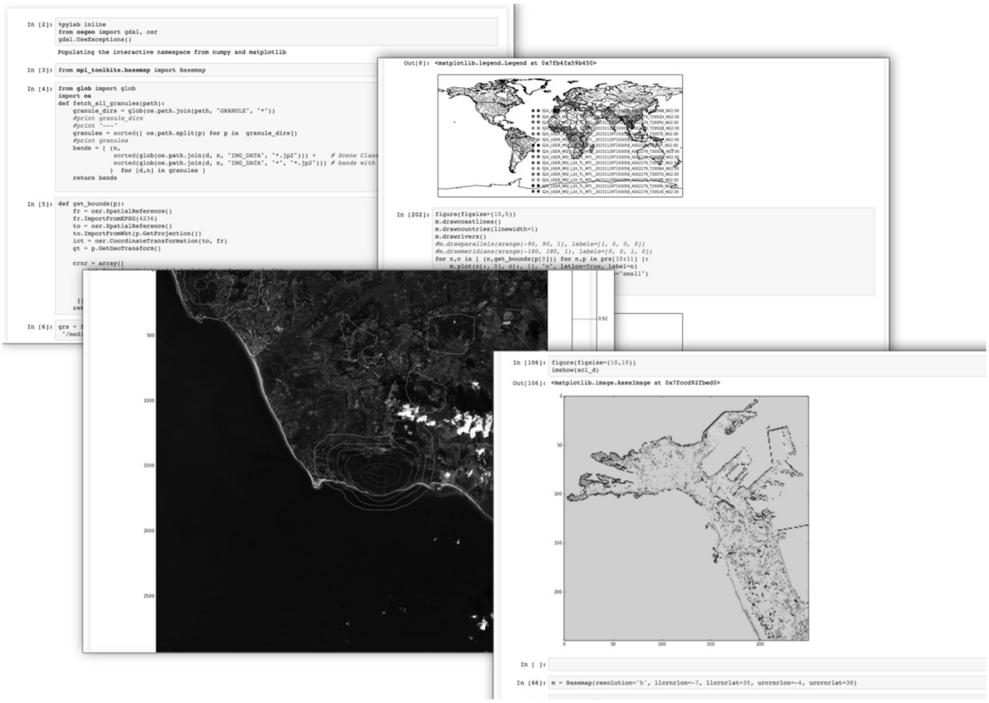

Fig. 5: Data analysis and visualisation in the EODC Interactive Delivery Platform

Data or product publishing or redistribution is then possible directly from the archive, depending on the users, creators or owners wishes, as all EODC services have access to respective archive locations. The EODC interactive delivery platform, which is based on common open source tools and techniques, such as jupyter (PROJECT JUPYTER 2016) or GeoServer (PROJECT GEOSERVER 2016) and is currently in development, will provide multi-channel online access and basic analysis features for various kinds of EO data and is intended to offer users tailor-made solutions to deliver their information products to a wider audience (Fig. 5).

3 Summary and Outlook

The ever-increasing amounts of EO data demand for a sophisticated expansion strategy for computational power, as well as for storage space. EODC is planning to receive and provide Sentinel 1, 2 and 3 data and therefore aiming to continuously extend its storage capacities. In parallel, the virtual research environment SIDP is currently being equipped to host 100+ users in the same time frame. Given the diversity of EO products in Europe, and the data centres producing them, EODC is strongly working towards a federation of data centres or other similar initiatives across Europe, in order to be able to exploit the information richness of up-to-date EO data to their full potential.

References

CLOUDEO (2016), CLOUDEO. http://www.cloudeo-ag.com.

DRUSCH, M., DEL BELLO, U., CARLIER, S, COLIN, O., FERNANDEZ, V., GASCON, F., HOERSCH, B., ISOLA, C., LABERINTI, P., MARTIMORT, P. et al. (2012), Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sensing of Environment, 120, 25-36.

ESA (2016), ESA Sentinel Online. https://sentinel.esa.int/web/sentinel/user-guides/sentinel-1-sar/resolutions/level-1-ground-range-detected.

ESA (2016a), Ground Segment overview. http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Ground_Segment_overview (accessed 28/01/16).

ESA (2016b), Sentinel Collaborative Ground Segment. http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Sentinel_Collaborative_Ground_Segment (accessed 28/01/16).

ESA ESRIN (2016), The Copernicus Space component: Sentinels Data products list. https://sentinels.copernicus.eu/documents/247904/685154/Sentinel+Products+List-Issue1-Rev1.pdf.

PROJECT GEOSERVER (2016), GeoServer. http://geoserver.org.

PROJECT JUPYTER (2016), Project Jupyter. http://www.jupyter.org (accessed 28/01/16).

SENSYF (2016), Sensyf. http://www.sensyf.eu.

SINERGISE (2016) Sinergise. http://www.sinergise.com.

TOP500.ORG (2016) top500.org. http://www.top500.org/list/2014/11/?page=1.

VSC (2016) VSC-3. http://www.vsc.ac.at.

WAGNER, W. (2015), Big Data Infrastructures for Processing of Sentinel Data. In: Photogrammetric Week 2015. Stuttgart, Germany.

WAGNER, W., FRÖHLICH, J., WOTAWA, G., STOWASSER, R., STAUDINGER, M., HOFFMANN, C., WALLI, A., FEDERSPIEL, C., ASPETSBERGER, M., ATZBERGER, C., BRIESE, C., NOTARNICOLA, C., ZEBISCH, M., BORESCH, A., ENENKEL, M., KIDD, R., VON BERINGE, A., HASENAUER, S., NAEIMI, V.& MÜCKE, W. (2014), Addressing Grand Challenges in Earth Observation Science: The Earth Observation Data Centre for Water Resources Monitoring. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, II-7, 81-88.