BIG DATA MANAGEMENT IN EARTH OBSERVATION THE GERMAN SATELLITE DATA ARCHIVE AT DLR

Stephan Kiemle, Katrin Molch, Stephan Schropp, Nicolas Weiland, Eberhard Mikusch

German Aerospace Center DLR, German Remote Sensing Data Center, Oberpfaffenhofen, Germany

ABSTRACT

The German Satellite Data Archive at the German Aerospace Center (DLR) has been managing large volume Earth observation data for more than two decades. Hardware, data management, processing, and user access, as well as long-term preservation are under one roof and interact closely in the payload ground segment (PGS). Several examples will demonstrate how Earth observation data life cycles benefit from close interaction between the PGS and the application scientists and from operational experience gathered over time.

***Index Terms*—** Data management, PGS, Earth observation, data life cycle

1. INTRODUCTION

The German Satellite Data Archive (D-SDA), operated at the German Remote Sensing Data Center of the German Aerospace Center DLR, is the central German infrastructure providing large volume Earth observation (EO) data archiving and access functionality. For many years it has been a key component in the payload ground segments (PGS) of numerous national and ESA Earth observation missions and in addition serves as a data center for scientific and civil service Earth observation applications.

With a 50 petabyte total capacity D-SDA currently holds about 3.8 petabyte of EO data and thematic information products in a hierarchical storage system. In response to current mission requirements for large volume processing, 170 terabyte of online cache allow for maximum throughput in the range of six terabytes per day. The Data and Information Management System (DIMS) – developed in collaboration with an industrial partner – ensures dedicated and swift data flows from archive to processing systems and users – assisted by a high capacity local area and wide area network infrastructure.

The D-SDA in its PGS function ensures flawless data reception and delivery during current Earth observation missions. However, its data management extends far beyond mission lifetime. Data are being curated and valorized throughout the entire data life cycle which contributes to ensuring long-term data accessibility and usability.

Collaborating closely with the user community and driven by their needs, a PGS is inherently concerned with the data content and the applications as it accompanies the data life cycle. The following three examples will present and discuss specific competences accumulated at the D-SDA over time. These are considered essential for providing comprehensive and efficient PGS data management services.

2. SENTINEL 5 PRECURSOR – BIG DATA PGS BASED ON EXTENSIVE EXPERIENCE

On behalf of ESA DLR develops and operates the Sentinel-5 Precursor PGS [1]. In the mission PGS, D-SDA provides archiving and data access for lower level and higher level processing. Long-standing experience with a variety of aspects and scenarios of EO data management facilitated PGS set-up and operations for this new mission.

Familiarity with atmospheric data workflows - and with user centered data modeling in support of these specific workflows - was available at DLR from contributing e.g. to the EUMETSAT Satellite Application Facility for Atmospheric Composition and UV Radiation (O3M SAF) [2]. This experience proved valuable in setting up the Sentinel-5 Precursor PGS. Similar to O3M SAF, Sentinel-5 Precursor processing workflows are based on numerous product inter-dependencies and intensive use of auxiliary data (Figure 1). Controlling such workflows requires more than a generic scheduling engine on a large, distributed computing facility. It also requires specific EO metadata for selecting input data and verifying the consistency of output data as well as knowledge on the behavior of the individual processor in order to optimize product quality and processing throughput.

One key component of the PGS of the German national TanDEM-X radar mission is large volume data management and systematic processing involving transferring large amounts of data to and from processing systems with data transfer rates reaching a maximum of 10.2 terabyte per day during peak processing phases [3]. This experience facilitated design and set-up of the Sentinel 5 Precursor PGS with daily throughput requirements totaling 250 gigabyte of level 0 data, 510 gigabyte of level 1b near-real time products, 905 gigabyte of level 1b offline products, and 465 gigabyte of auxiliary data.

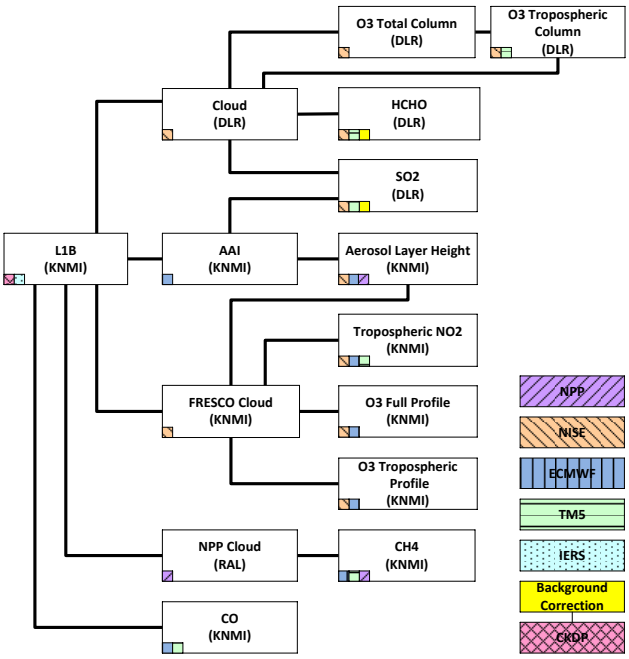

Figure 1. Expected product inter-dependencies and auxiliary data required for the production of level 2 products in the Sentinel-5 Precursor mission.

Knowledge about ESA-specific ground segment architectures and interfaces for discovery and access, which were essential for Sentinel-5 Precursor, was available from developments associated with the Copernicus Space Component Data Access and from developing the long-term archives for Sentinel-1 and Sentinel-3. Similar solutions, for example, were pursued in all three missions for integrating the DLR and Copernicus wide area networks.

The Sentinel missions PGS is closely monitored and thus a systematic reporting of system activities, such as the data processing completeness, auxiliary data usage, product quality, and service performance, is requested. D-SDA has established a reporting control service which systematically generates reports based on information collected from all PGS components.

3. TIMELINE – SERVICE-DRIVEN DATA MANAGEMENT

The DLR TIMELINE project will generate a consistent 30-year timeline of 18 level 2 and level 3 thematic information products, including sea surface temperature, snow cover, and cloud properties, in support of monitoring Global Change. Contributing to the TIMELINE project the D-SDA extends its PGS-focused archiving and access functionality towards servicing large volume data processing for scientific projects.

Hence, the PGS is moving closer to the scientific users, which requires a mutual adaptation of approaches, architectures, processes, and procedures. The project highlighted the advantages of designing and operating the complete service chain, from data acquisition to archiving, basic and thematic processing, as well as data and product access, under one roof and re-using existing PGS data management components.

During the initial project phases the close interaction between data managers and application scientists, processor developers and system operators helped recognize existing constraints on either side. A clear, mutually agreed view on roles and responsibilities was established which resulted in a constructive division of duties. The following roles have been defined:

- Scientist
- Tool developer
- Software engineer
- System engineer
- Data librarian
- Operational system engineer
- Operating / production
- Support levels 1-4

Striving for mutual optimization, the D-SDA data management has evolved towards a more flexible, service-driven archiving and access infrastructure. Improvements included creating a project-specific instance of the EOWEB Geoportal, the generic D-SDA data access portal, with project-specific discovery and download services, as well as developing new interfaces for bulk data handling. With a view to furthering interoperability, the project provided an opportunity to move towards providing standardized product metadata following the *Earth Observation Metadata profile of Observations & Measurements* defined by the Open Geospatial Consortium [4].

4. FROM SCIENTIFIC ALGORITHMS TO OPERATIONAL PGS-STYLE PROCESSING CHAINS - LESSONS LEARNT

One of the challenges faced during the TIMELINE project is converting scientific processing algorithms into large-volume PGS-style processing chains – while providing sufficient flexibility for ad-hoc modifications as required by the scientists. Specific issues arise from the different cultures of scientific and operational PGS-style data processing. Scientists are concerned with developing and fine tuning algorithms for accurate thematic product generation. Systematic, PGS-style processing requires

transferring these scientific algorithms into processing tools and ultimately into systematic, high-capacity, operational processing systems.

Experience indicates that this transfer is best done by an informatics specialist located within the scientific application environment. His understanding of the algorithm and the data, as well as close interaction with the scientists, will result in short evolution cycles and add to the flexibility of the resulting processing system.

5. USER COMMUNITY DRIVEN LONG-TERM EARTH OBSERVATION DATA PRESERVATION AND ACCESS

Earth observation data are unique snapshots of the condition of the Earth or atmosphere at a specific point in time. DLR and the D-SDA, therefore, put particular emphasis on long-term data preservation with the objective to keep the valuable data and products accessible and useable for future generations. Preservation is an inherent part of the data life cycle - its scope extending beyond keeping the instrument data safe from loss.

One aspect for ensuring long-term data accessibility in the D-SDA is its stable and sustainable archive infrastructure. Hardware and software are maintained and upgraded on a regular basis following technology evolution cycles. On the data side, active data curation contributes to long-term usability. As new requirements emerge – from technology evolution, application scientists, or end users - the data are migrated, converted to new formats, re-processed using improved algorithms, transformed into user-friendly higher level products, and made accessible via the required interfaces.

For data sets to remain useable for future generations, however, preserving sensor data and metadata is not sufficient. Curation of the associated knowledge, such as mission and sensor documentation, data structure and format specifications, calibration and processing information, is as essential as maintaining the capability to visualize and re-process the data [5],[6]. For new missions appropriate data management principles including the generation of the associated information should be established already during the planning and preparation phases.

In addition to implementing sustainable data curation measures, the D-SDA contributes to developing ESA-wide harmonized and interoperable data management principles and procedures for Earth observation data. In collaboration with partner organizations operational know-how is being transferred into guidelines and best practices [5], [6]. Extending beyond Europe these procedures are now being introduced into the Group on Earth Observations (GEO) and the Committee on Earth Observation satellites (CEOS).

6. LOOKING BEYOND EO - FACILITATING INTEGRATED DATA EXPLORATION AND ANALYSIS

As an efficiently operated PGS, the D-SDA can well provide data discovery and access services to the emerging collaborative thematic exploitation platforms. This ability will maximize data use and facilitate cross-fertilization between different application communities. The use of standardized metadata formats and discovery and access protocols, such as those provided by OGC, HMA, or OpenSearch, become mandatory for ensuring smooth interoperability between the distributed data archives.

Data exploitation across archives requires seamless interoperability. While federated data discovery across distributed Earth science data archives is already common, interoperable data formats or data exchange formats, in particular at the lower PGS level, are still in their infancy. However, data preservation and curation will soon require operational solutions for scenarios such as consolidating time series across archives or for physically relocating complete historical archive holdings in case of a transfer of responsibilities [6].

7. SUMMARY AND CONCLUSIONS

"Big data" management in Earth observation is not just an IT issue - it is more than just handling large data volumes, providing large volume processing capacity and high network bandwidths for swift data access. The data sets as well as the valuable scientific results are best handled by a comprehensive data life cycle center in which data managers, IT engineers, and scientists collaborate closely. It is the knowledge of the data set content and the experience with data management systems serving specific Earth observation applications and workflows which make for optimum, integrated end-to-end service chains. Thus the maximum value is extracted out of the Earth observation missions – during the mission lifetime and beyond - for current and future generations.

8. REFERENCES

[1] S. Kiemle, R. Knispel, M. Schwinger, and N. Weiland, "Sentinel-5 Precursor Payload Data Ground Segment, proceedings of ESA Advances in Atmospheric Science and Applications," in Proc. *ATMOS 2012 - Advances in Atmospheric Science and Applications*, Bruges, Belgium, 18-22 June 2012, ESA Special Publication SP-708.

[2] S. Kiemle, P. Valks, M. Boettcher, D. Loyola, W. Zimmer, T. Ruppert, and T. Erbertseder, "DLR Data Services for GOME-2/MetOp Atmospheric Trace Gas Monitoring," in Proc. *Joint 20017 EUMETSAT*

Meteorological Satellite Conference, Amsterdam, 24-28 September 2007.

[3] S. Kroeger, M. Schwinger, M. Wegner, and M. Wolfmueller, "Data Handling and Preservation for the TanDEM-X Satellite Mission," in Proc. *PV 2009 - Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data*, 1-3 December 2009, Villafranca, Spain, pp. 1-7.

[4] J. Gasperi, F. Houbie, A. Woolf, and S. Smolders (eds.), "Earth Observation Metadata profile of Observations & Measurements," v. 1.0, OGC® Implementation Standard, reference number OGC 10-157r3, 12 June 2012.

[5] GSCB LTDP Working Group, "Earth Observation Preserved Data Set Content," accessed at http://earth.esa.int/gscb/ltdp/LTDP_PDSC_4.0.pdf on 12 October 2014.

[6] GSCB LTDP Working Group, "European LTDP Common Guidelines," accessed at http://earth.esa.int/gscb/ltdp/EuropeanLTDPCommonGuidelines_Issue2.0.pdf on 12 October 2014.