A multi-talented datacube: integrating, processing and presenting big geodata for the agricultural end user

Christoph Friedrich[1], Johannes Löw[2], Insa Otte[1], Steven Hill[1], Sebastian Förtsch[1], Jakob Schwalb-Willmann[1], Ursula Gessner[3], Christoph Schierghofer[3], Sina Truckenbrodt[4,5], Eric Schonert[4], Thomas Piernicke[6], Denise Assmann[7], Christopher Conrad[2] and Michael Thiel[1]

Abstract: While scientific methods leveraging Earth Observation for agriculture are abundant, their actual application in Germany remains scarce. A key challenge in this context is to connect the end users to the data without the many technical obstacles. Therefore, we present a versatile platform that not only integrates and processes big geodata of highly diverse origin and type, but also provides access to these resources in ways that reflect the individual user's requirements and expertise. Based on free and open-source software building blocks, our datacube facilitates scientific computation through R and Python environments or direct API access, including emergent technologies such as openEO, STAC, and COG. At the same time, the results are delivered to easy-to-use applications that adequately present them to non-technical experts. We detail the architecture of the system and demonstrate a use case serving computed plant vitality information directly to farmers in the field.

Keywords: analysis-ready data, cloud processing, interoperability, data access, user interfaces

1 Introduction

Agricultural land management is a major application for Earth Observation (EO) research. It is capable of delivering a variety of information on crop land at high spatial and temporal resolution, ranging from plant vitality and phenology to yield and biomass estimations. This has been utilised for several use cases such as monitoring of subsidies and application maps based on vegetation indices [Eu22]. However, in Germany the agriculturalists

[1] University of Würzburg, Institute for Geography and Geology, Department of Remote Sensing, Earth Observation Research Cluster, John-Skilton-Straße 4, 97074 Würzburg, Germany; christoph.friedrich@uni-wuerzburg.de
[2] University of Halle-Wittenberg, Institute for Geosciences and Geography, Department of Geoecology, Von-Seckendorff-Platz 4, 06120 Halle (Saale), Germany
[3] German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), Münchener Straße 20, 82234 Weßling-Oberpfaffenhofen, Germany
[4] German Aerospace Center (DLR), Institute of Data Science, Department of Data Acquisition and Mobilisation, Mälzerstraße 3-5, 07745 Jena, Germany
[5] University of Jena, Institute of Geography, Department of Earth Observation, Leutragraben 1, 07743 Jena, Germany
[6] GFZ German Research Centre for Geosciences Potsdam, Department 1: Geodesy, Section 1.4: Remote Sensing and Geoinformatics, Telegrafenberg, 14473 Potsdam, Germany
[7] Deutscher Wetterdienst, Department of Agrometeorology, Kärnerstraße 68, 04288 Leipzig, Germany

working the fields – who are at the core of the system and should thus be the main beneficiaries of such research – are not yet applying EO-based information widely and to its full extent.

This is often attributed to the collected data being abundant in principle, but difficult to use in practice – especially when data from various sources or of different types needs to be combined. In addition to this *variety*, which gets increasingly common as increasingly complicated challenges are tackled, the sheer *volume* of today's geodata and the *velocity* of its creation add to the urgent need for Big Data management strategies. Therefore, our research project "AgriSens DEMMIN 4.0" not only addresses the creation of novel remote-sensing-based application techniques, but also puts an equally distinct emphasis on the development of an accompanying data integration and visualisation system. In the following, we describe how this essential tool closes the gap between data providers (e.g., satellite operators) and information consumers (e.g., farmers) by facilitating necessary analysis steps and combining these with adequate presentation for data-driven decision making in the agricultural reality.

2 Methods

In our IT architecture, we utilise one central *datacube* to conquer this problem, which acts as a cloud-based data holding and computation platform. It gathers a multitude of data (ranging from optical and radar raster imagery through climate data to in-situ field measurements) and pre-processes it into an interoperable, analysis-ready state. These resources can then be accessed through APIs for external usage; or computations can be carried out directly on the datacube and their results immediately visualised with tools hosted on the same server. Through this, we are aiming to make data usage more user-friendly. Figure 1 illustrates the architecture of our datacube, while the following text explains its capabilities in more detail.

The physical system is entirely located at the Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities (LRZ). Apart from utilising the computing resources available there, this also opens up synergies with already-existing projects: we can directly access the enormous amount of EO data that is already available within the LRZ's "Data Science Storage" and the DLR's "terrabyte" platform. These storages are directly mounted into our server so that the datacube can access petabytes of imagery without having to duplicate it again, saving costs and emissions. This provides data from major EO satellite missions such as Landsat, MODIS, Sentinel-1, and Sentinel-2, which are essential to many EO workflows and thus a necessity for many EO users. As the supplied archives are not (yet) complete, we partly augment it by ingesting the needed missing parts from the "Open Data on AWS" programme and hosting them locally. All data is pre-processed (e.g. atmospheric correction) and provided in the Cloud-Optimised GeoTIFF (COG) format, resulting in a large collection of analysis-ready data (ARD).

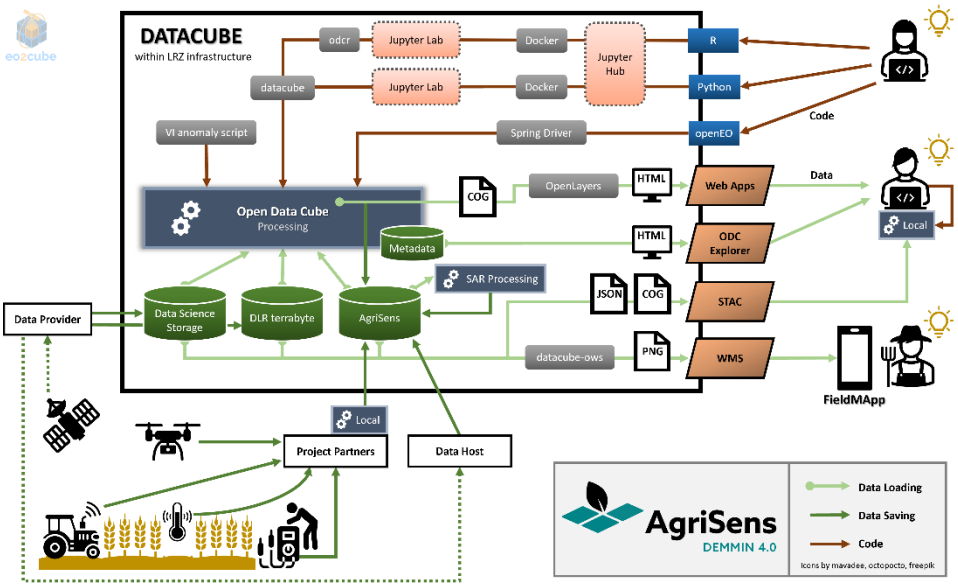

Fig. 1: The architecture of our datacube within its ecosystem. Data that has been acquired (bottom left) is processed by the datacube (big box) so that it can be utilised easily by end users (right-hand side). The latter, depicted by the pictograms, may be farmers, consultants, scientists, policy makers, or basically any agricultural stakeholder wishing to utilise EO data

Extending this rather common data offering, we also compute more special data that is currently intensively researched: Synthetic Aperture Radar (SAR) data has demonstrated its value for agricultural applications such as crop inventories and monitoring applications [Lo23; LUC21]. Its processing, however, is complex and computationally intensive. We set up a pipeline based on "ESA SNAP" (version 9) and "pyroSAR" [Tr19] that consumes Sentinel-1 data and generates products such as InSAR coherence as well as Alpha and Entropy, which are not easily and openly available elsewhere in an analysis-ready format.

Apart from this extensive satellite data repository, our catalogue also contains derived products such as Digital Elevation Models (DEMs), as well as project data like drone imagery or in-situ measurements from field campaigns. Another asset is (agro-) meteorological data provided by the German Weather Service (DWD). It is released every month through their FTP server, from where it is automatically downloaded into the datacube infrastructure, converted from the typical meteorological netCDF format into the more cloud-friendly GeoTIFF format, and then ingested into the datacube's database. It originates from a network of climate stations of the DEMMIN long-term test site and is delivered as hourly values interpolated over an area of 37 by 43 kilometres [HL22].

The core of our infrastructure is an instance of the "Open Data Cube" software package, which is open source. It provides the functionalities to handle user-specific data requests for computation and visualisation purposes and therefore acts as the link between

metadata, data, and our output channels. The metadata is ingested into its PostgreSQL database and can be retrieved via its built-in "Explorer", a web-based data discovery application. It is also exposed via an API endpoint of the emergent STAC standard, which in turn can be accessed via the "STAC Browser" or any other compatible software. As we use the COG format for raster data, access is efficient even from remote machines.

Our main interface for scientific computation is Jupyter Hub, enabling collaborative work across institutions. For each user a dedicated Jupyter Lab instance is spawned in its own Docker container that can access all the data of the previously mentioned storages and has a certain amount of computing resources allocated to it. Using containerisation allows to manage resources and minimises the interference for other users while maintaining collaborative pathways via shared folders. Users can write their code in Python or R, which offer the straightforward packages "datacube" or "odcR", respectively. Another way to work with the data is via openEO, a standardised way to interact with big EO data cloud processing backends via so-called process graphs. This integration is achieved via the "openEO Spring Driver".

However, to reach our target audience of people who are not EO data experts, these very technical interfaces are not suitable, but easy and straightforward access to data and functionality is required. Therefore, our priority is on developing easy-to-use graphical interfaces that visualise the data and scientific products in a manner that effectively helps the end users. One way to achieve this, which we are applying extensively, is to host not only the data, but also purpose-built web applications that present it via powerful JavaScript mapping libraries like OpenLayers. However, in the next section we explain in detail a case study that utilises a more special output channel.

3 Case Study: from Satellite to Field

One use case being realised in this context is a near-real-time plant vitality information system. Vegetation indices (VIs) are computed on incoming Sentinel-2 scenes as soon as they become available and compared to the average VI value that has been measured over the four years 2019 to 2022 for the same crop type and time of year. An interactive mobile application, called *FieldMApp*, then consumes this data, enabling farmers to directly assess the current status of their field on-site in comparison to the long-term average.

The baseline time series of multi-annual average VI values per crop type are calculated based on Sentinel-2 optical satellite data. In order to know which pixels to consider for which crop type's average, crop type maps that have been created within the AgriSens DEMMIN 4.0 project are utilised. They were produced using a method similar to the one described by Asam et al. [As22] and are available for the years 2019 to 2022, covering the area of the German state of Mecklenburg-Vorpommern. Accordingly, all available Sentinel-2 imagery for this spatio-temporal extent is utilised, atmospherically corrected and cloud masked using the PACO and Fmask algorithms, respectively. Data gaps are

filled by linear interpolation before the data is resampled into continuous 4-year time series of 5-day frequency (analogous to the 5-day revisit frequency of Sentinel-2A/B). Based on this reflectance data, the four VIs NDVI, SAVI, NDYI, and EVI are calculated, resulting in time series of VIs. These are subsequently stratified by the 16 crop types (e.g. winter wheat, maize, potatoes, rapeseed) using the aforementioned crop type maps. Then, for each 5-day time step within one year, the multi-annual (2019-2022) mean and standard deviation are calculated for each pair of VI and crop type. These values are finally exported in the CSV format for further usage.

The datacube is set up to automatically process new Sentinel-2 scenes as soon as they become available. Upon their ingestion, an automatic pipeline is triggered that computes the four aforementioned VIs on the new imagery and produces anomaly products for each of the 16 crop types. To do so, the applicable long-term average closest to the capture date is looked up and subtracted from the VI raster as a constant value. The resulting rasters are then exposed directly as COGs as well as through a Web Map Service (WMS) for increased compatibility with existing software. For the conversion, the "datacube-ows" package is employed.

One output channel is the *FieldMApp*. It is an application (app) for mobile devices that provides support to farmers in sustainable land management and crop production. The app aims at an optimized application of production resources by combining satellite-based crop monitoring data (e.g. from the datacube), freely available geodata, and digitised local knowledge of farmers. The latter can be acquired with the FieldMApp, based on (1) a tool for recording and annotating areal data during field management, (2) forms generated with the "Open Data Kit", and (3) a geographic information system (GIS). With the GIS, vector and raster data from different sources (local or from servers) can be visualised, and, moreover, vector data can be edited. The app is developed using the Flutter framework, employing an offline-first approach to meet the needs of farmers working under varying (i.e., often bad or inexistent) network coverage conditions. This framework was chosen for its ability to facilitate the writing of platform-agnostic code while maintaining native performance and accommodating the integration of platform-specific implementations that are crucial for incorporating hardware sensors.

The modular structure of the FieldMApp allows for easy extensibility and adaptability. As such, incorporating the plant vitality information explained above as an additional visualisation layer was uncomplicated. Since Flutter's geospatial visualisation library "Flutter_Map" does not yet support the rendering of GeoTIFFs, the datacube exposes the anomaly rasters via a WMS. Through this standardised data service, the FieldMApp can load the appropriate data by setting the "style" parameter of the request according to the needed crop type and desired vegetation index. While the latter is to be specified on presentation, information about the crop type is part of the field metadata that the app stores anyway, so is already known. They also contain the field's boundaries, which are utilised to mask out the areas of the delivered image that do not belong to the field.

4 Discussion and Outlook

While the individual building blocks of the solution presented here are not fundamentally new, we argue that our integrated end-to-end approach is quite distinguished and will be helpful in advancing the uptake of EO-based methods in the agricultural sector. As such, our datacube can be a facilitator towards increasing the technology readiness of EO products: the simplification of prohibitively complicated steps increases real-world usage. Our datacube offers both the familiar Jupyter interface as well as state-of-the-art APIs. This empowers scientists and programmers to utilise the wealth of available data with the flexibility of a coding environment and thus to supply the latest methodological developments to farmers. The developed infrastructure leverages software packages that are tried-and-tested yet incorporate recent developments, ensuring the system to be both stable and modern. The free and open-source concept of the datacube also ensures that data supply and interfaces remain customisable towards the specific needs of end users. It is also a strategic advantage over existing solutions such as Google Earth Engine, which is powerful, but eventually a black box dependent on the good will of one company.

Overall, our datacube is an effective tool for bringing together users of various backgrounds and expertise, due to its variety of access levels from simple app-based visualisation to coding environments. It is publicly available to some extent through the website https://eo2cube.org/. Field tests with farmers will bring insights into the real-world usage of the developed system. It will be expanded by including further project-driven use cases such as drone-based irrigation monitoring. Also, even more output channels are to be unlocked, e.g. through a dedicated plugin that connects to established GIS platforms.

Bibliography

[As22] Asam, S. et al.: Mapping Crop Types of Germany by Combining Temporal Statistical Metrics of Sentinel-1 and Sentinel-2 Time Series with LPIS Data. Remote Sensing 14/13, 2022.

[Eu22] European Commission; Joint Research Centre; Åstrand, P.; Devos, W.; Loudjani, P.: Controls with remote sensing in the CAP2020+. Publications Office of the European Union, 2022.

[HL22] Haßelbusch, K.; Lucas-Moffat, A.: Rasterdaten für die Agrarmeteorologie: Vergleich verschiedener Interpolationsverfahren am Beispiel AgriSens Demmin 4.0, 2022.

[Lo23] Lobert, F. et al.: A deep learning approach for deriving winter wheat phenology from optical and SAR time series at field level. Remote Sensing of Environment 298/, pp. 113800, 2023.

[LUC21] Löw, J.; Ullmann, T.; Conrad, C.: The Impact of Phenological Developments on Interferometric and Polarimetric Crop Signatures Derived from Sentinel-1: Examples from the DEMMIN Study Site (Germany). Remote Sensing 13/15, 2021.

[Tr19] Truckenbrodt, J. et al.: Towards Sentinel-1 SAR Analysis-Ready Data: A Best Practices Assessment on Preparing Backscatter Data for the Cube. Data 4/3, 2019