LEONSEGS

LARGE EARTH OBSERVATION
NEW SPACE ECOSYSTEM
GROUND SEGMENT

D5.9 SCIENTIFIC AND TECHNICAL ABSTRACTS COMPILATION V2

30/11/2025

Dissemination Level: **PUBLIC**

Grant Agreement No.: 101082493
Call: HORIZON-CL4-2022-SPACE-01
Topic: HORIZON-CL4-2022-SPACE-01-13
Type of action: HORIZON-IA

D5.9 Scientific and Technical Abstracts Compilation v2

VERSION 2

Work package	WP 5
Task	5.1
Due date	30/11/2025
Submission date	28/11/2025
Deliverable lead	PLUS
Version	2.0
Authors	Dirk Tiede (PLUS)
Reviewers	Martin Sudmanns (PLUS) Carolina Pascaru (F6S) María Jesús Gutiérrez (GMV) Kathrin Lunzner (PLUS) Luis Saturnino (AISTECH)
Consortium Internal Code	D5.9_Scientific and Technical Abstracts Compilation v2
Abstract	This version updates and extends the first report on the abstracts disseminated by project partners in the second project period. It

Dissemination Level: **PUBLIC**

	briefly summarises the project's publications and presentations. demonstrating the growing scientific impact and visibility of the project.
Keywords	EO, New Space, flexible multi-mission EO ground segment, scientific abstracts, scientific publications

Document Revision History

Version	Date	Description of change	List of contributor(s)
1.0	27/01/2025	Initial version (Del. 5.3 v1)	PLUS, GMV, AISTECH, F6S
2.0	19/11/2025	Updated and extended compilation (this Del 5.9 v2)	PLUS, GMV, AISTECH, F6S

DISCLAIMER

Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.

COPYRIGHT NOTICE

The LEONSEGS Consortium is the following:

Participant number	Participant organisation name	Short name	Country
1	GMV AEROSPACE AND DEFENCE SA	GMV	Spain
1.A	GMV GmbH	GMV	Germany
1.B	GMV Romania	GMV	Romania
2	PARIS-LODRON-UNIVERSITAT SALZBURG	PLUS	Austria
~~3~~	~~SATELLOGIC SOLUTIONS SL~~	~~SATL~~	~~Spain~~
4	F6S NETWORK LIMITED	F6S	Ireland
5	AISTECH SPACE SL	AISTECH	Spain

EXECUTIVE SUMMARY

This deliverable presents the collection of scientific and technical abstracts published or presented by the consortium in relation to the LEONSEGS project developments. It documents the scientific output and technical dissemination activities carried out by partners, covering publications, conference contributions, posters, and other relevant presentations that communicate LEONSEGS results to the research community, industry stakeholders, and the wider public.

It will be periodically updated throughout the project's duration. The first version was released as Deliverable 5.3 (Version 1), this is the second version released as Deliverable 5.9 (Version 2) and the final version will be Deliverable 5.10 (Version 3).

TABLE OF CONTENTS

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

BiDS	ESA Big Data from Space Conference
CCM	Copernicus Contribution Missions
EARSeL	European Association of Remote Sensing Laboratories
EO	Earth Observation
ESA	European Space Agency
LPS	ESA Living Planet Symposium
MIGARS	Conference on Machine Intelligence for GeoAnalytics and Remote Sensing
IGARSS	International Geoscience and Remote Sensing Symposium
ISDE	International Symposium on Digital Earth

1. OVERVIEW SCIENTIFIC AND TECHNICAL ABSTRACTS

This section provides an overview of the scientific and technical abstracts produced and disseminated by the LEONSEGS consortium during the first and second project periods.

Scientific abstracts from the project's first period were published and presented at the ESA BIG Data from Space (BiDS) conference in Vienna (November 6-9, 2023) and at the EARSeL symposium in Manchester UK (June 17-20, 2024). Scientific abstracts and publications from the second project period were presented and/or published at the ESA Living Planet Symposium (LPS) (06/2025), MIGARS (09/2025) and IGARSS (08/2025).

LEONSEGS has been also promoted in additional big EO events like ESA Big Data from Space 2025 conference (10/2025), AGIT 2025 (07/2025), ISDE 2025 (04/2025), the GEO Global Forum (05/2025) and the 3rd Annual Copernicus Contributing Missions (CCM) Workshop (10/2025). A high-level journal publication has been published in July 2025. Publications from the project's second period are highlighted in the table below in orange color.

The following scientific abstracts have been published:

No	Title	Authors	Conference/Paper
I	An Advanced Framework for Semantic Querying of The Dynamic World Dataset	Martin Sudmanns, Lisah Ligono, Hannah Augustin, Lucas van der Meer, Dirk Tiede	Proceedings of the 2023 conference on Big Data from Space, Soille, P., Lumnitz, S. and Albani, S. editor(s), Publications Office of the European Union, Luxembourg, 2023, doi:10.2760/46796, JRC135493, pp 357-360
II	Semantic World – A Novel Benchmark Dataset for Semi-Supervised Semantic Segmentation	Felix Kröber, Dirk Tiede, Andrea Baraldi, Sébastien Lefèvre.	43rd EARSeL Symposium, Manchester, UK, June 17th to 20th, 2024; conference proceedings (book of abstracts)

III	An Approach for the Semantic Enrichment of Sentinel-1 Imagery Suitable for Large-scale Analysis	Luke McQuade, Martin Sudmanns, Dirk Tiede	43rd EARSeL Symposium, Manchester, UK, June 17th to 20th, 2024; conference proceedings (book of abstracts)
IV	One-layer RGB representation of big EO data analyses for supporting the visual communication of multi-temporal change detection	Dirk Tiede, Hannah Augustin, Thomas Strasser, Steffen Reichel, Markus Kerschbaumer, Kristýna Měchurová, Martin Sudmanns	43rd EARSeL Symposium, Manchester, UK, June 17th to 20th, 2024; conference proceedings (book of abstracts)
V	On-demand, semantic EO data cubes – knowledge-based, semantic querying of multimodal data for mesoscale analyses anywhere on Earth	Felix Kröber, Martin Sudmanns, Lorena Abad, Dirk Tiede	ISPRS Journal of Photogrammetry and Remote Sensing, Volume 228, 2025, Pages 552-565, https://doi.org/10.1016/j.isprsjprs.2025.07.015.
VI	On-demand data cubes – knowledge-based, semantic querying of multimodal Earth observation data for mesoscale analyses anywhere on Earth	Felix Kröber, Martin Sudmanns, Dirk Tiede	Poster and abstract presented at ESA LPS 2025 - based on the publication "V"
VII	Efficient aggregate land cover queries	Luke McQuade, Martin	Paper presented at the Machine Intelligence for GeoAnalytics and Remote Sensing (MIGARS)

	with cloud-optimized raster formats	Sudmanns and Dirk Tiede	conference at POLITEHNICA Bucharest, Bucharest, Romania, from 2 to 4 September 2025. https://doi.org/10.1109/MIGARS67156.2025.11231788
VIII	Semantic content-based image retrieval in semantic EO data cubes	Martin Sudmanns, Dirk Tiede and Andrea Baraldi	Scientific abstract and presentation presented at IGARSS 2025, Brisbane, Australia, 3 - 8 August 2025

Table 1: OVERVIEW SCIENTIFIC ABSTRACTS PUBLISHED IN THE FIRST AND SECOND PROJECT PERIOD

The following posters or presentations have been presented without specific scientific abstracts (usually on invitation):

No	Event	Title	Presenter
I	3rd Annual Copernicus Contributing Missions (CCM) Workshop (October 20 to 22, 2025, ESRIN, Frascati, Italy)	Poster: LEONSEGS - Large Earth Observation New Space Ecosystem Ground Segment	Luis Saturnino (AISTECH) https://leonsegs.eu/leonsegs-featured-at-copernicus-contributing-missions-workshop/
II	ESA Living Planet Symposium 2025 (25-27 June, 2025, Vienna, Austria)	Poster: LEONSEGS - Large Earth Observation New Space Ecosystem Ground Segment	Joint presentation at the European Union booth of Leonsegs representatives at LPS from GMV, PLUS, AISTECH and F6S
III	GEO Global Forum 2025 (05-09 May 2025, Rome, Italy)	Poster: LEONSEGS - Use cases for end	Presenter: Kathrin Lunzner (PLUS) Authors: Dirk Tiede, Martin Sudmanns, Kathrin Lunzner, PLUS | Gianni Marasca, Álvaro Núñez

		users and new space players	Martínez, GMV | Cátia Figueiredo, Carolina Pascaru, Alexandre Relvão, F6S | Sonia Navarro, Luis Saturnino, AISTECH
IV	14th International Symposium on Digital Earth (ISDE)	Presentation: An overview over open-source Earth observation data cube infrastructures: Perspectives, challenges, opportunities	Presenter: Martin Sudmanns (PLUS) https://leonsegs.eu/leonsegs-showcased-at-international-digital-earth-symposium/

Table 2: OVERVIEW SCIENTIFIC POSTERS AND PRESENTATIONS

1.1. FIRST PROJECT PERIOD

The presentations and scientific abstracts presented at the 43rd EARSeL symposium were part of a scientific workshop organisation related to LEONSEGS research activities from PLUS. The prestigious conference, focused on Earth Observation (EO) research, attracted around 200 researchers, students, and professionals from Earth and environmental sciences who work with remotely sensed data.

PLUS organized a dedicated workshop on "Semantic & Explainable Analysis of Big Data," directly related to research conducted within LEONSEGS. The workshop was planned by Martin Sudmanns, Dirk Tiede, and Hannah Augustin from PLUS, and included external expert Gregory Giuliani from the University of Geneva.

At the event, organized by the European Association of Remote Sensing Laboratories (EARSeL), PLUS contributed with three presentations, highlighting the project's commitment to advancing Earth observation analysis through innovative research and collaboration.

Workshop:

https://manchester2024.earsel.org/?page_id=184

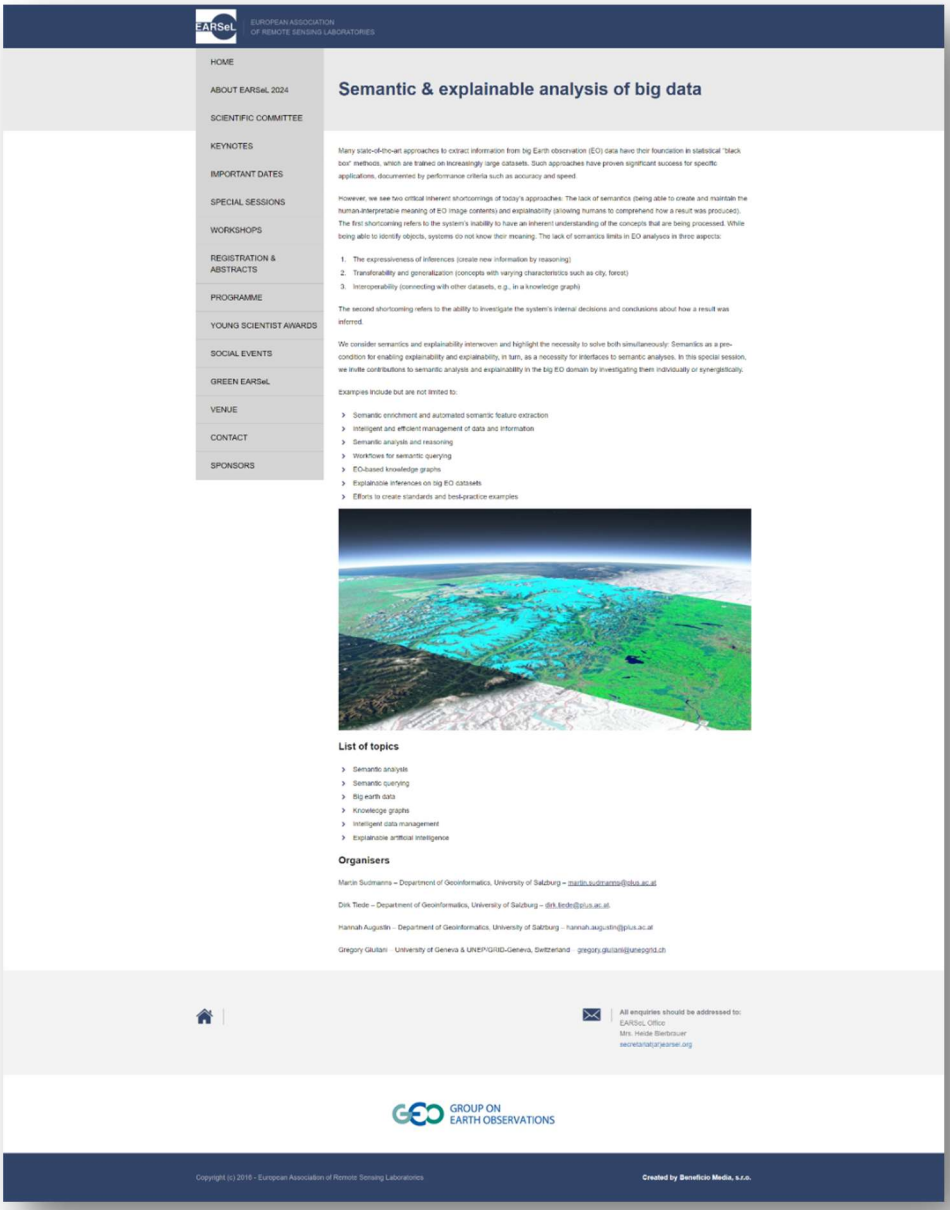


FIGURE 1-1: SCIENTIFIC WORKSHOP ORGANISATION RELATED TO LEONSEGS RESEARCH ACTIVITIES FROM PLUS AT THE EARSEL CONFERENCE 2024

The scientific abstract, published in the conference proceedings of the ESA Big Data from Space conference in Vienna, included a poster presentation that was awarded with the best poster award:

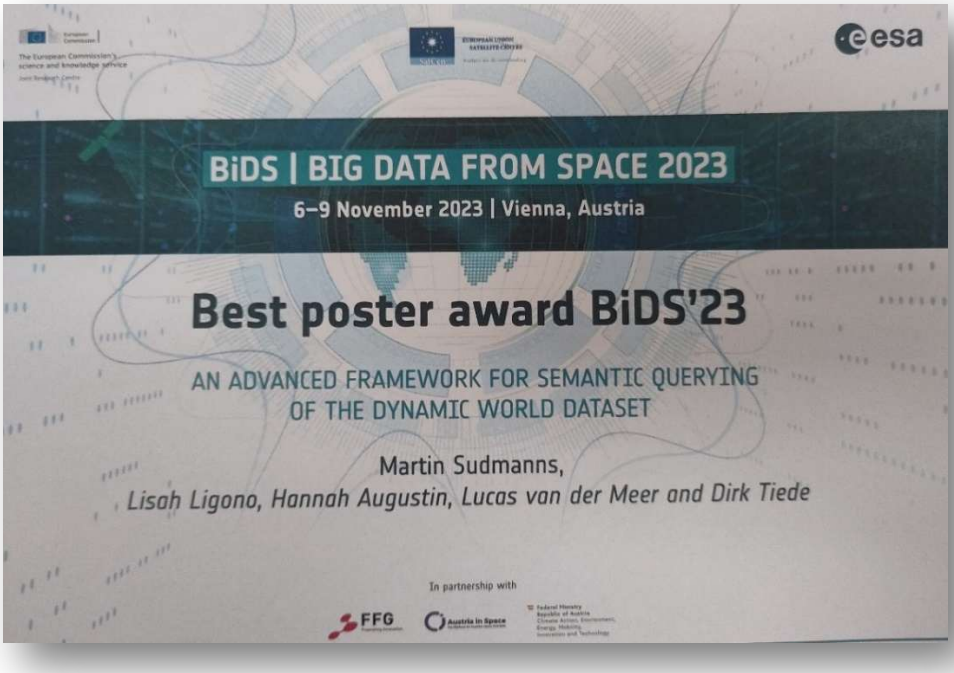


FIGURE 1-2: BEST POSTER AWARD FOR THE SCIENTIFIC POSTER SUBMITTED BY PLUS ("AN ADVANCED FRAMEWORK FOR SEMANTIC QUERYING OF THE DYNAMIC WORLD DATASET") FOR THE ESA BIG DATA FROM SPACE CONFERENCE IN VIENNA, 2023

1.2. SECOND PROJECT PERIOD

Within the second project period two high-level publications about scientific work co-funded through Leonsegs could be already published: the first one about on-demand semantic EO data cubes has been published in one of the highest ranked remote sensing journals (ISPRS Journal of Photogrammetry and Remote Sensing, impact factor 12.2), the second one published in IEEE Xplore and is based on a MIGARS conference contribution focusing on scientific and technical developments in semantic querying and how to optimize them efficiently.

Based on Leonsegs developments in the first period, the second project period has been used to promote the project in different, highly relevant conferences and workshops.

2. COLLECTION OF ABSTRACTS

In the following, the presented and published abstracts are reproduced as originally printed.

2.1. SCIENTIFIC ABSTRACT I

AN ADVANCED FRAMEWORK FOR SEMANTIC QUERYING OF THE DYNAMIC WORLD DATASET

Martin Sudmanns, Lisah Ligono, Hannah Augustin, Lucas van der Meer, Dirk Tiede

[1]Department of Geoinformatics, Paris Lodron University Salzburg, 5020 Salzburg, Austria

ABSTRACT

Within this contribution we show how the Dynamic World data by Google and the World Resources Institute can be semantically queried, using a methodology originally developed for use within semantic Earth observation (EO) data cubes. We demonstrate in a minimal working example how the Dynamic World dataset can be analyzed through space and time based on the given categories/classes but also on aggregated classes. This is beyond selecting the most occurring class (i.e., using the mode operator) and opens new possibilities for using this dataset and a new direction to unfold its potential.

Index Terms— Semantic Querying, Data Cubes, Dynamic World, Categorical Time Series Analysis, querying aggregated classes.

1. INTRODUCTION

Transforming reflectance values, which do not have inherent semantics, into categorical information, which can be understood by users, is an ongoing endeavor (e.g., for creating land cover maps, time series analysis). An approach that we developed is the semantic Earth observation (EO) data cube, which provides categories that users can combine on a case-by-case scenario in custom spatial and temporal ranges. A semantic EO data cube (i.e., semantics-enabled EO data cube) is defined as "a data cube, where for each observation at least one nominal (i.e., categorical) interpretation is available and can be queried in the same instance" [1]. Based on this definition, we developed a scalable architecture covering Austria, called Sen2Cube.at [2], utilizing a generic semantic enrichment of each available Sentinel-2 image. To facilitate semantic querying of the semantic EO data cube and analyzing categorical variables, we developed an inference engine, *semantique* (https://github.com/ZGIS/semantique) [3].

Google and the World Resources Institute pushed the boundaries for what is possible in the big (Earth) data era when they released the Dynamic World dataset because they do not provide a single, global product with a fixed legend. Instead, they have classified (and are continuously classifying) all Sentinel-2 images having less than 35% cloud cover according to their image-wide metadata [4]. This is a different approach compared to most land cover products that are generated and released every few years. Since this dataset is provided to Google Earth Engine (GEE) users, the users can create their own (land cover) maps by specifying temporal rules that are applied to the dataset, so-called reducers. While in their publication by Brown et. al 2022 they use the mode to reduce the temporal dimension of the dataset to select the most often occurring class, they argue that with "a more advanced decision framework […] it is possible to customize a discrete classification as is appropriate for a user's unique definitions or downstream task." [4]. We argue that this statement is in-line with the objectives of a semantic EO data cube architecture and to the best of our knowledge, such a decision framework has not been developed or proposed.

In this contribution, we show that our *semantique* inference engine can be such a decision framework for working with Dynamic World land cover classes. We use the classes as input for *semantique* and thus allow users to perform more complex operations than using only the mode. We demonstrate this approach by transferring our inference engine to Dynamic World classes using a minimal working example and argue for considering approaches that are inherently transferable to a variety of datasets and systems.

2. METHOD, DATA, AND IMPLEMENTATION

Our implementation uses the *semantique* Python package for semantic querying of the Dynamic World dataset. The Dynamic World dataset, available via GEE , is a near real-time global Sentinel-2 land use / land cover (LULC) mapping generated by a Fully Convolutional Neural Network (FCNN) [4]. The dataset provides nine classes, which can be accessed either per-pixel by probability or by directly obtaining the class with the highest probability (Figure 1). The classes were derived from level-1C data available since 2015 because Sentinel-2 level-2A data are systematically produced globally only after 2017.

semantique is an open-source Python package for semantic querying of EO data [3]. While it was originally developed for querying the spectral categories produced by the SIAM Software [5], it is designed to be generic and can be applied to any kind of image categorization. The abstraction levels of semantic querying using *semantique* consist of three elements referred to as the layout, mapping, and query recipe.

Fig. 1. Sample of the Dynamic World dataset of Lake Baringo and its environments from September 03, 2020. Note that this is a mono-temporal classification, while the entire Dynamic World dataset covers temporal stack of images that are typically reduced to a single map by using the mode operator (most often occurring class).

The **layout** is a human- and machine-readable description of the semantic EO data cube's content [6]. It defines the individual categories or classes of the semantic layers as well as the storage structure. For example, it specifies that categories 'A', 'B', and 'C' are stored in file 'X' and categories 'D', 'E', and 'F' in file 'Y'. Therefore, it allows connecting to different storage systems using the same content name and optimizes the storage system without affecting analysis.

The **mapping** is a human- and machine-readable connection between the available categories in semantic EO data cubes and the user's target classes. In the context of our original Sen2Cube.at implementation, it is used to bridge the high-level semantic domain (with vocabularies containing terms like "Forest") and the image domain (with vocabularies containing terms like "greenness above an index threshold"). This approach goes back to existing ideas of knowledge-based image interpretation [7]. The mapping can be either defined in a stable way or on a case-by-case, user-defined scenario. In the context of the Dynamic World dataset, the classes are already on a high semantic granularity; however, users may be still requiring referring to the (thematically) aggregated class "Vegetation" as a combination of "Trees", "Grass", "Crops", and "Shrub and scrub" or to the class "Surface Water" as a combination of "Water" and "Flooded Vegetation". References and – in this case – grouping of classes like these are defined in the mapping.

The **query recipe** is a set of high-level instructions where operations are defined and applied to the entities (i.e., classes in the case of the Dynamic World) specified in the mapping. Common examples are operations such as reduce, group, or extract.

While the layout defines what is available and storage parameters and the mapping connects categories to target semantic entities, the query recipe defines spatio-temporal operations for these entities when applied to a spatio-temporal subset of a data cube.

In our proof-of-concept implementation, we created a layout and a mapping using the Dynamic World classes and ad-hoc querying using the *semantique* API. In the example, we use the classes "Water" and "Flooded vegetation", respectively, with the aim to identify basic temporal dynamics of surface water. This type of analysis is similar to calculations using the Water Observations from Space (WOFS) algorithm [8] or the JRC Surface Water product [9].

3. EXAMPLE

To showcase our approach using semantic querying, we selected an area in Kenya covering several Dynamic World classes but focused on water dynamics of Lake Baringo (Figure 2). In this example, we want to demonstrate the added value of a more advanced querying framework instead of using only the mode as temporal reducer. Lake Baringo is a suitable test site for applying this approach because it has recently experienced an unusual increase in surface area extent due to rising water levels.

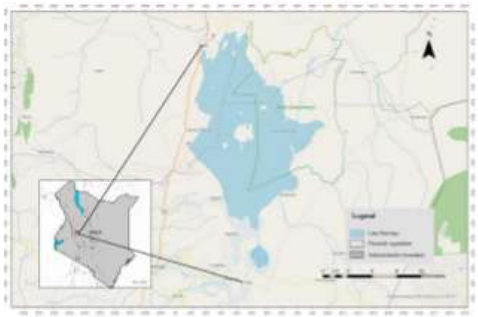


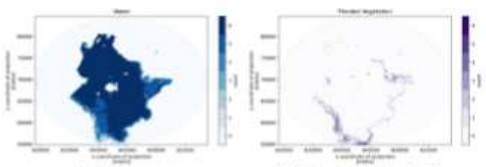
Fig. 2. Visualization of the test site, Lake Baringo which is a Great Rift Valley Lake. ©OpenStreetMap and contributors, CC-BY-SA.

Lake Baringo is a desert lake [10] located in the Great Rift Valley in Kenya (0.6667° N and 36.0667° E) with an elevation of about 970 meters above sea level. It covers an

area of approximately 180 square kilometers, and as of 2020 is the second largest freshwater lake in Kenya. Its maximum depth reaches around ten meters, although these figures are seen to vary due to recent swelling of the Rift Valley lakes. The lake receives water primarily from two main rivers, the Molo and the Ol Arabel, as well as numerous small streams. It lacks any significant outlet, resulting in a relatively high mineral concentration due to evaporation. Nonetheless, it is still a freshwater lake [11].

In this minimal working example / proof of concept, we created a connection to GEE using the Python package wxee (https://wxee.readthedocs.io/en/latest/) to obtain the data directly as xarray. In this case, we used six different time steps between 2020 and 2022. After creating the layout for connecting the *semantique* package to the available classes of the Dynamic World dataset, we created the following mapping of the classes: a 1:1 mapping of the Dynamic World classes, extended with a mapping of the two water-related classes ("Water" and "Flooded Vegetation") into one semantically aggregated target class ("Surface Water"). As a result, we can directly conduct queries against these classes using the *semantique* API; in this case, we apply reducing operations over time and space.

The result of reducing over time by counting class occurrences is illustrated in two maps of Figure 3. In both maps, higher values indicate a higher occurrence of the class over time, from no observations up to six. In contrast to the mono-temporal selection or a mode operator, such a reducer is able to show temporal dynamics, which, in this example, are distinctly visible in the eastern, southern, and western part of the lake (Figure 3 – (a)). The map on the right side shows complementarily the flooded vegetation (Figure 3 – (b)). Both maps can already be used to visually identify areas with permanent surface water and areas that are occasionally dry as well as flooded.



(a) Water (b) Flooded Vegetation

Fig. 3. Visualized output of the query execution for 'Water' (a) and 'Flooded Vegetation' (b) on a six-timestamp date range between 2020-04-01 and 2022-08-31.

If users are not interested in the separated classes of "Water" and "Flooded Vegetation" but want to consider them both at the same time, it is possible to query them using the combined class specified in the mapping. Figure 4 shows the result of the same reducing operation of the combined class.

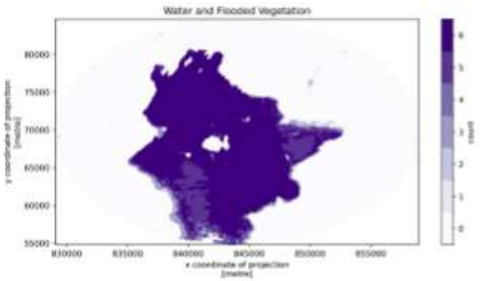


Fig. 4. Visualization of the analysis of the aggregated class "Surface Water", as a combined mapping of "Water" and "Flooded Vegetation".

Further, instead of reducing over time, it is also possible to reduce over space. The result of a reducing operation over space is not a map but a graph showing the aggregated values for each timestamp. In this case, it is the water count indicating the size of the lake and surrounding water areas (Figure 5).

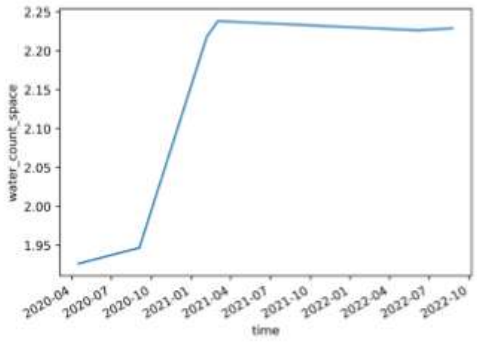


Fig. 5. The graph shows a steep rise in the observations of the class "Water" in October 2020 based on counted water pixels.

These examples use count as the aggregation function in both reducing operations. However, other functions are also possible and include aggregating or reducing based on the average, maximum, minimum, standard deviation, last, or first over time or space in the user-defined area of interest. It is the user's or developer's decision to choose the function and/or combine it with additional processing steps depending on their desired result, which is potentially much more sophisticated compared to the one in our example.

4. DISCUSSION

We demonstrated how our semantic querying and analysis approach can be extended and transferred to the Dynamic

World dataset. The semantic querying approach can be transferred to the Dynamic World dataset and individual query recipes can be generated and re-used due to the abstraction that semantic querying provides. The execution of the querying recipe is agnostic to the underlying dataset. For example, if other semantic enrichment approaches like the mentioned SIAM color names are mapped to the class surface water, the same querying can be applied.

While the algorithms and processing steps of the *semantique* package can be validated independently, the quality and validation of the overall result is dependent on the quality of the initial classification.

Instead of using the *semantique* Python package, it would be possible to develop such a querying and analysis framework directly in GEE's programming environment. A direct integration removes the necessity to download the data and process them locally. Instead, the processing would be done on GEE. There are other advantages of this approach as well such as good integration with existing scripts and workflows. However, has the main disadvantage of lacking transferability to other platforms. While the Dynamic World dataset is available in GEE and currently unique worldwide, we can expect that similar approaches will be developed and published soon.

5. CONCLUSIONS AND OUTLOOK

With continuous advancements in artificial intelligence and deep learning, land cover datasets that provide an interpretation of every available image, such as the Dynamic World, are possible. In contrast to other "fixed" processed land cover datasets, it does not provide a single layer of pre-calculated classes based on one or multiple years. Instead, it provides classes on a per-pixel level for all Sentinel-2 images and allows custom selection of spatio-temporal extents and classes. Hence, based on our work towards querying time series of categories to enable semantic queries, we argue that the *semantique* package can be one of the decision frameworks that the authors of Dynamic World dataset asked for in their publication.

To demonstrate the technical feasibility of connecting *semantique* to the Dynamic World dataset and the benefit of such a querying framework, we created an example for Lake Baringo in Kenya, which has had significant water dynamics in recent years.

Based on the promising results of the prototypical implementation and the experimental use cases, we conclude and argue that the full potential of datasets such as the Dynamic World unfolds when a semantic querying interface is available to users. Instead of reducing the dataset using a mode operator only, the spatio-temporal distribution of the classes and their occurrence can be analyzed in more detail. Therefore, a semantic querying interface could increase the dataset's uptake.

The example and technical implementation shown here can be considered the beginning, and there are several options for future developments. They include but are not limited to closer integration of such a querying framework into the GEE to be directly connected to the entire Dynamic World dataset, handling probabilities, and testing the dataset and approach in various applications.

REFERENCES

[1] H. Augustin, M. Sudmanns, D. Tiede, S. Lang, and A. Baraldi, "Semantic Earth observation data cubes," *Data*, vol. 4, no. 3, p. 102, 2019, doi: 10.3390/data4030102

[2] M. Sudmanns, H. Augustin, L. van der Meer, A. Baraldi, and D. Tiede, "The Austrian Semantic EO Data Cube Infrastructure," *Remote Sensing*, vol. 13, no. 23, p. 4807, 2021, doi: 10.3390/rs13234807.

[3] L. Van Der Meer, M. Sudmanns, H. Augustin, A. Baraldi, and D. Tiede, "Semantic querying in Earth observation data cubes," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XLVIII-4/W1-2022, pp. 503–510, 2022, doi: 10.5194/isprs-archives-XLVIII-4-W1-2022-503-2022.

[4] C. F. Brown *et al.*, "Dynamic World, Near real-time global 10 m land use land cover mapping," *Sci Data*, vol. 9, no. 1, p. 251, 2022, doi: 10.1038/s41597-022-01307-4.

[5] A. Baraldi, L. Durieux, D. Simonetti, G. Conchedda, F. Holecz, and P. Blonda, "Automatic Spectral-Rule-Based Preliminary Classification of Radiometrically Calibrated SPOT-4/-5/IRS, AVHRR/MSG, AATSR, IKONOS/QuickBird/OrbView/GeoEye, and DMC/SPOT-1/-2 Imagery—Part I: System Design and Implementation," *IEEE Trans. Geosci. Remote Sensing*, vol. 48, no. 3, pp. 1299–1325, 2010, doi: 10.1109/TGRS.2009.2032457.

[6] M. Sudmanns, H. Augustin, L. van der Meer, C. Werner, A. Baraldi, and D. Tiede, "One GUI to Rule Them All: Accessing Multiple Semantic EO Data Cubes in One Graphical User Interface," *giforum*, vol. 1, pp. 53–59, 2021, doi: 10.1553/giscience2021_01_s53.

[7] S. Growe, "Knowledge based interpretation of multisensor and multitemporal remote sensing images," *Int. Arch. Photogramm. Remote Sens*, vol. 32, no. pt 7, pp. 4–3, 1999.

[8] N. Mueller *et al.*, "Water observations from space: Mapping surface water from 25 years of Landsat imagery across Australia," *Remote Sensing of Environment*, vol. 174, pp. 341–352, 2016, doi: 10.1016/j.rse.2015.11.003.

[9] J.-F. Pekel, A. Cottam, N. Gorelick, and A. S. Belward, "High-resolution mapping of global surface water and its long-term changes," *Nature*, vol. 540, no. 7633, pp. 418–422, 2016, doi: 10.1038/nature20584.

[10] C. O. Olilo, C. O. Odoli, M. O. Obiero, J. O. Malala, and C. H. Bironga, "Chapter 8 - Microbial ecology of desert Lakes Baringo and Turkana, Kenya, East Africa," in *Lakes of Africa*, M. El-Sheekh and H. E. Elsaied, Eds., Elsevier, 2023, pp. 247–268. doi: 10.1016/B978-0-323-95527-0.00016-6.

[11] E. O. Okech, N. Kitaka, S. Omondi, and D. Verschuren, "Water level fluctuations in Lake Baringo, Kenya, during the 19th and 20th centuries: Evidence from lake sediments," *African Journal of Aquatic Science*, vol. 44, no. 1, pp. 25–33, 2019, doi: 10.2989/16085914.2019.1583087.

2.2. SCIENTIFIC ABSTRACT II

EARSeL Manchester 2024
Abstract
Corresponding Author: felix.kroeber@plus.ac.at

Semantic World – A Novel Benchmark Dataset for Semi-Supervised Semantic Segmentation

Felix Kröber[1], Dirk Tiede[1], Andrea Baraldi[2], Sébastien Lefèvre[3]

[1] University of Salzburg, Department of Geoinformatics – Z_GIS, Austria
[2] Spatial Services GmbH, Austria
[3] Université Bretagne Sud, IRISA UMR 6074, France

Keywords: Deep Learning, Supervised learning, Semi-supervised learning, Land cover, Sentinel-2

Challenge

Despite the increasing interest in deep learning models for remote sensing applications, large-scale data sets foundational to these models are still scarce. Data sets covering a range of biomes worldwide are of particular interest to train models generalisable across space. In this regard, the Dynamic World data set [1] represents an important milestone giving rise to the eponymous model [2] as well as the ESRI land cover product [3]. With these models global, continuously updatable land cover classifications have been produced for the first time. However, one factor that limits the accuracy and generalisability of the models is the availability of training data, based on a labour-intensive manual labelling process. Upscaling the training of such models requires an extension of the narrow view of supervised learning to semi-supervised learning, which enables the leveraging the much larger archive of unlabelled satellite scenes.

Methodology

This work utilises semantic enrichments of satellite data to develop the Dynamic World dataset into a more comprehensive dataset for semi-supervised semantic segmentation tasks. The use of the Dynamic World dataset as the basis for the Semantic World presented here enables corresponding benchmarking with the aforementioned models [2], [3].

The semantic enrichment is carried out using the Satellite Image Automatic Mapper (SIAM) [4]. SIAM is a physical model-based expert system capable of mapping, automatically and in near real-time, multi-spectral satellite imagery into a discrete and finite vocabulary of semi-symbolic spectral categories. Encoded as a decision tree, SIAM performs a deterministic, hyperparameter-free enrichment process translating reflectance values into a pre-classification. As an input, SIAM can employ any radiometrically calibrated multispectral image data, including Sentinel-2 L1C or L2A data. This allows to enhance the existing labelled Dynamic World patches by adding the pre-classifications as bands that can be used as input instead of or in combination with the original non-semantical reflectance values. Furthermore, adding pre-classifications for Sentinel-2 patches for which no land cover annotations are available enables to extend the dataset for purposes of semi-supervised learning. The data set designed in this way makes it possible to investigate the added value of knowledge-based, semantic enrichments in the context of various deep learning architectures.

Expected results

The data set produced comprises around 57.4K semantically enriched Sentinel-2 patches of 510 x 510 pixels – around 21.4K with and 36.0K without land cover annotations. The split between training and test data of 57.0K to 0.4K is based on the original split according to the Dynamic World dataset. The

reflectance values of all 10m and 20m Sentinel-2 bands for both processing levels (i.e. L1C & L2A) are provided as input information. In addition to any land cover annotations, the SIAM categorisations in four different granularities (i.e. 18, 33, 48 & 96 categories) are always available. The automated scene classification (i.e. SCL layer of the S-2 L2A products) is also stored in the label stack of an Sentinel-2 patch for comparison purposes of the added value of the SIAM categorisation. The overall organisation of the data set is summarised in Figure 1.

The diversity and scope of the dataset is evident from Figure 2. Based on the Dynamic World dataset, the sample patches are distributed across a total of around 9K Sentinel-2 scenes from the years 2017 to 2019, which in their entirety cover all main biomes globally. This spatio-temporal diversity is accompanied by a corresponding diversity of the spectral profiles of the Sentinel-2 input data and a broad coverage of 9 different land covers classes.

Outlook for the future

Building on the performed technical validation of the created data set, basic model tests will be carried out in the next step. These are intended to demonstrate the use of the dataset under fully- and semi-supervised paradigms. Well-established backbones for semantic segmentation networks (such as U-Net) will be used to perform these initial assessments. In a next step model architectures will be tailored to the unique structure and information content of the Semantic World dataset. The scalability of the semantically enriched part of the dataset offers the potential to establish a novel foundation model specific to remote sensing data. Conditioning such a model on the knowledge-based spectral categorisations can provide stronger guiding for learning physically reasonable feature representations within such foundation models. The Semantic World dataset along with basic model tests will be released publicly enabling users to train and develop their own architectures.

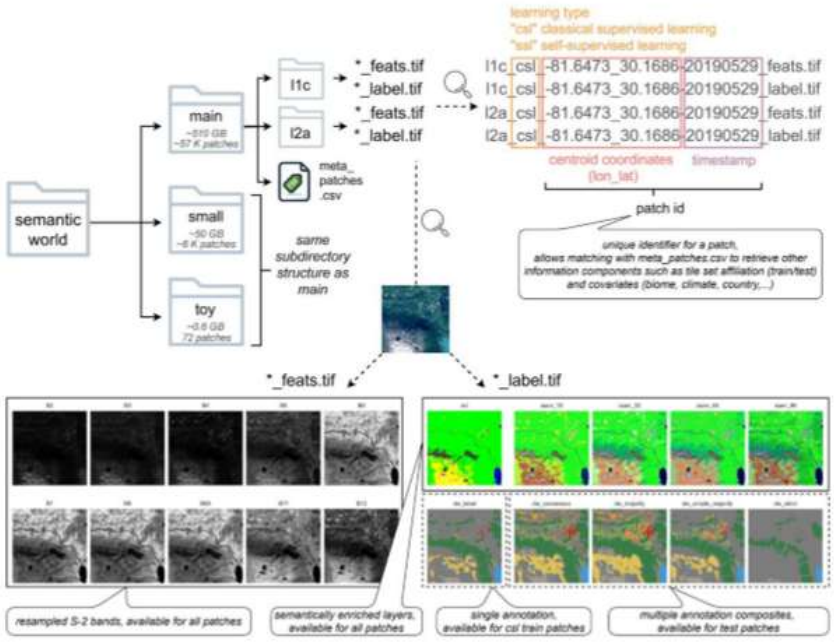


Figure 1 Structure of the Semantic World

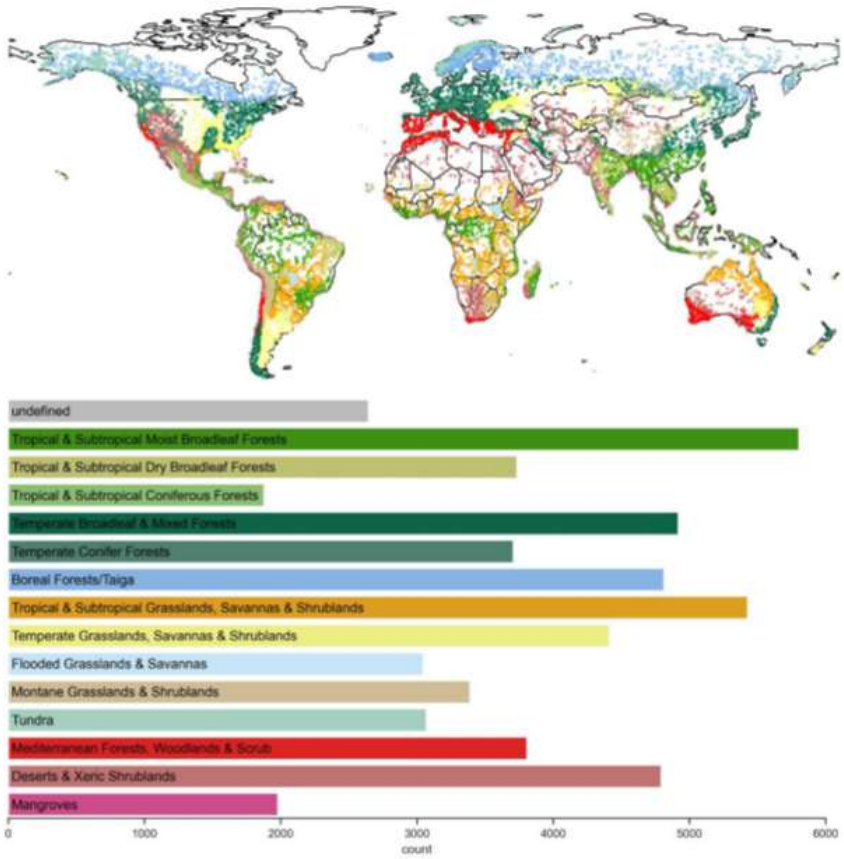


Figure 2 Spatial distribution of Semantic World patches with their relationship to global biomes

References

[1] A. M. Tait, S. P. Brumby, S. B. Hyde, J. Mazzariello, and M. Corcoran, 'Dynamic World training dataset for global land use and land cover categorization of satellite imagery'. PANGAEA, 2021. doi: 10.1594/PANGAEA.933475.

[2] C. F. Brown *et al.*, 'Dynamic World, Near real-time global 10 m land use land cover mapping', *Sci Data*, vol. 9, no. 1, p. 251, Jun. 2022, doi: 10.1038/s41597-022-01307-4.

[3] K. Karra, C. Kontgis, Z. Statman-Weil, J. C. Mazzariello, M. Mathis, and S. P. Brumby, 'Global land use / land cover with Sentinel 2 and deep learning', in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, Brussels, Belgium: IEEE, Jul. 2021, pp. 4704–4707. doi: 10.1109/IGARSS47720.2021.9553499.

[4] A. Baraldi, M. L. Humber, D. Tiede, and S. Lang, 'GEO-CEOS stage 4 validation of the Satellite Image Automatic Mapper lightweight computer program for ESA Earth observation level 2 product generation - Part 1: Theory', *Cogent Geoscience*, vol. 4, no. 1, p. 1467357, Jan. 2018, doi: 10.1080/23312041.2018.1467357.

2.3. SCIENTIFIC ABSTRACT III

L. McQuade et al

EARSeL Manchester 2024
Abstract
Corresponding Author: luke.mcquade@plus.ac.at

An Approach for the Semantic Enrichment of Sentinel-1 Imagery Suitable for Large-scale Analysis

Luke McQuade[1], Martin Sudmanns[1], Dirk Tiede[1]

[1] UNIVERSITY OF SALZBURG, DEPARTMENT OF GEOINFORMATICS - Z_GIS, AUSTRIA

Keywords (5): Earth Observation, Synthetic Aperture Radar (SAR), Big Earth Data, Land Use, Copernicus

Challenge

Synthetic Aperture Radar (SAR) Earth observation (EO) satellites have several advantages over their optical counterparts, such as being able to observe the Earth's surface at night, and through a wide variety of weather conditions. However, due to the nature of their sensors and mechanisms of capture, the resultant imagery is often difficult to interpret and use in downstream analyses. Several approaches exist for the semantic enrichment of optical data, such as the Satellite Image Automatic Mapper (SIAM*), which, coupled with their use in EO data cubes, can greatly improve accessibility and use of the original data. A system offering similar benefits for SAR EO data could be highly beneficial, especially considering the potential to complement optical data. Designing such a system to permit analyses across differing geographic areas globally presents an additional challenge which we have also attempted to address in this work.

Methodology

We devised an approach for the semantic enrichment of dual-polar Sentinel-1 radiometric-terrain-corrected (RTC) backscatter imagery (VV and VH polarizations). We refer to this as *polarimetric categorization*. It consists of binning the parameter space of VV and VH backscatter according to the scattering properties of known surface types; the result is that each pixel is assigned a category according to the scattering type(s) exhibited, e.g., surface scattering, volume scattering, double-bounce. The categorization/binning is performed with a decision tree algorithm, with set (constant) thresholds. These thresholds were determined in a part knowledge-based, part data-driven process. Figure 1 shows our preliminary categorization scheme.

Our categorization processor was implemented as a Python package, with working title, *dpolcat* (dual-polarimetric categorizer). It depends on standard EO packages such as xarray† and a STAC client. Just-in-time (JIT) compiler Numba‡ was also utilized to improve computational performance.

Several trials were conducted in using categorized scenes generated by our algorithm in downstream analyses, such as flood mapping, burned-area delineation, and vegetation change mapping.

Expected results

---

* Baraldi, A., Humber, M. L., Tiede, D., & Lang, S. (2018). GEO-CEOS stage 4 validation of the Satellite Image Automatic Mapper lightweight computer program for ESA Earth observation Level 2 product generation – Part 1: Theory. Cogent Geoscience, 4, doi: 10.1080/23312041.2018.1467357

† https://xarray.dev/

‡ https://numba.pydata.org/

Our preliminary version of *dpolcat* has facilitated several downstream trial analyses using models constructed in Python scripts. The first was an automatic flood mapping task. Observing the grid cells newly changed into strong surface scatterers (category 1 as in Figure 1), a model was built. Some simple spatial filtering and thresholding were also applied. The model was applied to map the flooding event of Duisburg, Germany, July 2021. Validated against the Copernicus Emergency Mapping Service reference, an F1 score of 0.80 was achieved. The source, intermediate, and resultant imagery are shown in Figure 2.

Other tasks include burned-area delineation and vegetation change mapping. Quantitative validation of these is ongoing. But, in all cases, the models using polarimetric categories were quick and simple to design and implement, compared to those in similar analyses using non-enriched images – models can be implemented in minutes rather than hours.

In terms of computational performance, it takes approximately 4 minutes to process an entire Sentinel-1 scene with *dpolcat*, using an Intel Xeon Platinum 8272CL CPU @ 2.60GHz.

Outlook for the future

As the category thresholds are globally constant, the algorithm can be applied to any Sentinel-1 scene, or across an area spanning multiple scenes, without a learning step - this is in contrast with similar techniques such as polarimetric decomposition and clustering. Also, compared to learning-based techniques, the algorithm is simple enough such that the computational cost for processing a scene is relatively low. This lends itself well to its use in semantic EO data cubes[§], where there is often a requirement to semantically enrich large numbers of scenes as a pre-requisite for further analyses.

Our preliminary results have identified some limitations of the approach. In time series analysis, especially with vegetated and mixed-use areas, the category assigned to a given location can vary unexpectedly, i.e., where there is seemingly no change to the underlying land-cover type. We refer to this as categorical instability. This could be due to, for instance, signal values close to category thresholds; and, with noise (which radar is prone to), a pixel can 'flip' between categories over time. Furthermore, there is scope to refine the implementation and integrate *dpolcat* processing into existing EO data cubes. We aim to address these in future work.

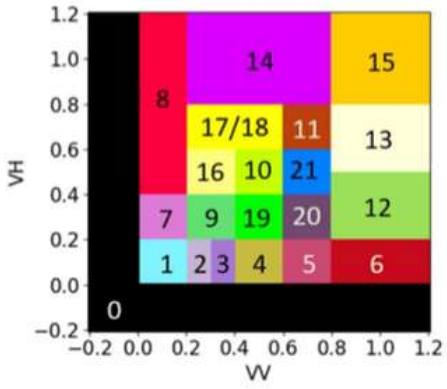


Figure 1 Division of the VV and VH parameter space into polarimetric categories: 1, 2, 3 and 4 represent mostly surface scattering; 9, 10, 16 and 19 represent mostly volume scattering; 8 is ill-defined (physically improbable); 0 is invalid or 'no data'; the remaining represent double-bounce and other phenomena.

[§] Augustin, H., Sudmanns, M., Tiede, D., Lang, S., Baraldi, A., 2019. Semantic Earth Observation Data Cubes. Data 4, 102. https://doi.org/10.3390/data4030102

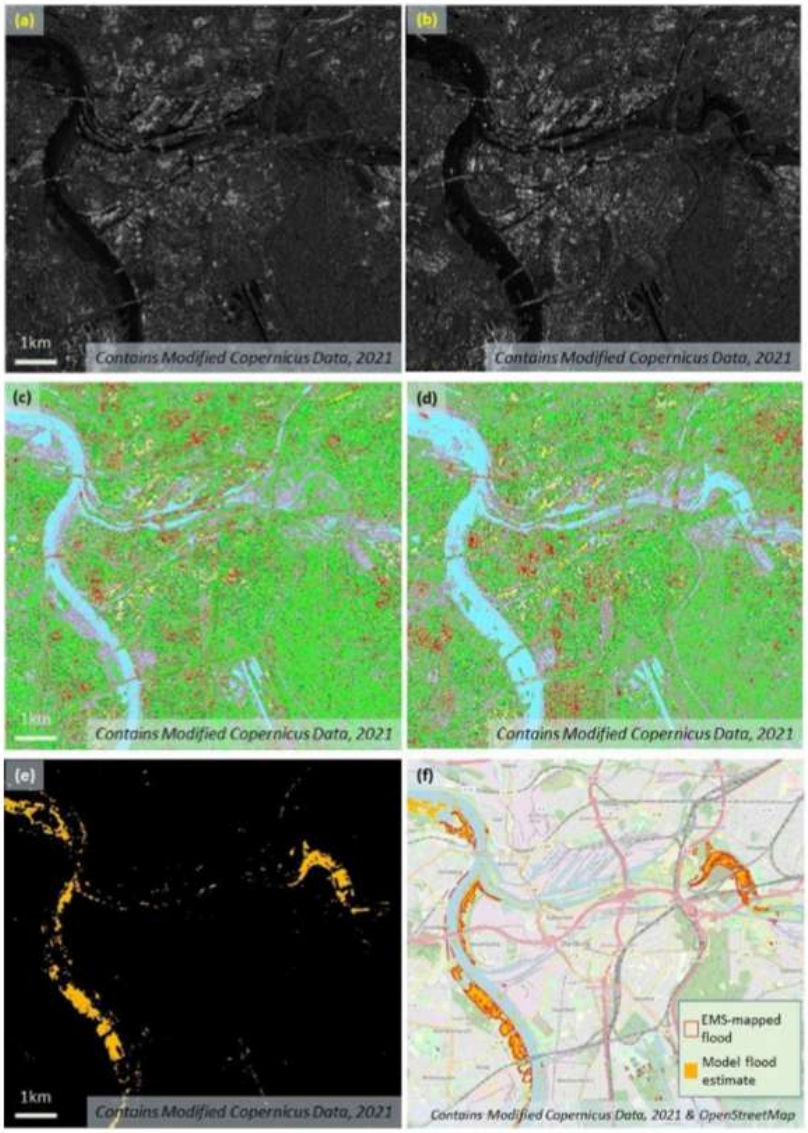


Figure 2 Example of using *dpolcat* in an automated flood mapping workflow – Duisburg, Germany, 13 July 2021 (pre-flood) to 16 July 2021 (in-flood). (a) Sentinel-1 VV backscatter, pre-flood. (b) Sentinel-1 VV backscatter, in-flood. (c) Pre-flood image categorized with *dpolcat*. (d) In-flood image categorized with *dpolcat*. (e) Flood map model output. (f) Flood map model output with Copernicus EMS reference**.

** Copernicus Emergency Mapping Service, product [EMSR517].

2.4. SCIENTIFIC ABSTRACT IV

EARSeL Manchester 2024
Abstract
Corresponding Author: dirk.tiede@plus.ac.at

One-layer RGB representation of big EO data analyses for supporting the visual communication of multi-temporal change detection

Dirk Tiede[1], Hannah Augustin[1], Thomas Strasser[1], Steffen Reichel[2], Markus Kerschbaumer[2], Kristýna Měchurová[2], Martin Sudmanns[1]

[1] University of Salzburg, Department of Geoinformatics – Z_GIS, Austria
[2] Spatial Services GmbH, Austria

Keywords (5): Earth Observation, Big EO Data, Change analysis, Sentinel-2, Geovisualization

Challenge

Big EO data, such as provided by the European Copernicus programme, are a great opportunity for continuous temporally high-frequent global monitoring of the environment. Challenges exist not only in the processing of the big multitemporal data* but also in communicating results in a meaningful and useful manner, especially for non-EO experts.

We present an approach for big EO data analyses in a semantic EO data cube and communicate results using a single-layer RGB (red, green, blue) representation, where each colour represents one of three different user-defined time periods. We focus on change analysis of observed vegetation, but the approach can be used in other applications. The resulting RGB layer serves as an interpretable base map that can be integrated in any GIS or browser interface. Multi-temporal information is encoded in different colour combinations. An adaptable colour cube legend aids interpretation (see Figure 1).

Methodology

The big EO data analyses behind the multi-temp thematic RGB layer are conducted in semantic EO data cubes†, where for each observation at least one nominal (i.e. categorical) interpretation is available and can be queried in the same instance. Our implementation - Sen2Cube.at‡ - is a semantic EO data cube available for all of Austria, where every Sentinel-2 satellite image taken since 2015 and their semantic enrichment can be analysed in the cloud. Data cubes have the advantage that the spatial and temporal extent to be analysed can be dynamically selected. Semantic data cubes extend this flexibility with a semantic query option that allows analyses directly in the selected area. No programming knowledge or additional software is required - everything can be done via the web browser and integrated with other data sets.

This approach uses semantic enrichment to count the percentage of vegetation / non-vegetation observed for all Sentinel-2 images in a user defined analysis period (e.g. years or seasons). Different to index-based approaches using only NDVI, no thresholds need to be defined since the semantic classes

---

* Sudmanns, M., Tiede, D., Lang, S., Bergstedt, H., Trost, G., Augustin, H., Baraldi, A., Blaschke, T., 2020. Big Earth data: disruptive changes in Earth observation data management and analysis? International Journal of Digital Earth 13, 832–850. https://doi.org/10.1080/17538947.2019.1585976

† Augustin, H., Sudmanns, M., Tiede, D., Lang, S., Baraldi, A., 2019. Semantic Earth Observation Data Cubes. Data 4, 102. https://doi.org/10.3390/data4030102

‡ Sudmanns, M., Augustin, H., van der Meer, L., Baraldi, A., Tiede, D., 2021. The Austrian Semantic EO Data Cube Infrastructure. Remote Sensing 13, 4807. https://doi.org/10.3390/rs13234807

(here: spectral categories) also reflect cloud-like / bare-soil-like / vegetation- and water-like categories. All available imagery can be used without additional pre-processing to filter cloud contaminated data. This has the advantage that smaller cloud-free regions are used even in very cloudy images, increasing the number of valid, clear observations and therefore the statistical soundness.

Expected results

The current implementation focuses on vegetation / non-vegetation changes based on Sentinel-2 big EO data, which means at least one image every 5 days, and in higher latitudes and overlapping orbits even more. The definition of the three time periods can be interactively conducted in the Sen2Cube.at interface and every Sentinel-2 image in the defined period will be used for the analysis. Based on the semantic enrichment, the spectral categories to be analysed in any given period are summed up and calculated as a percentage of the analysed images.

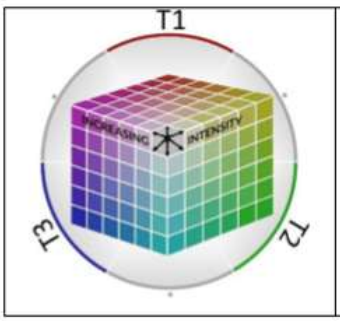
The RGB colour model is an additive colour model used to visualise the 3 different grayscale layers for each time period, each indicating the proportion of vegetation observed. The approach allows changes from 3 periods to be displayed on a map in one image using the different colour combinations. The interpretation of the colours can be drawn from the colour cube (Figure 1). The RGB colour palette and colour cube for the interpretation does not only communicate change, but also communicates changes in intensity and/or partly changed vegetation to non-vegetation and vice versa using main RGB colours and mixed colours plus their intensity (see Figure 2).

Outlook for the future

The single-layer representation is an approach to better communicate multi-temporal analyses to users (planning authorities, decision makers, non-EO scientists etc.). Our approach clearly indicates where changes happened and provides information on change intensity. This is different from base maps heavily used in GIS-based decision support systems, where often only mono-temporal information serve as background layers, such as static maps or aerial/satellite image mosaics with unclear observation dates. Application cases include supporting soil sealing monitoring, monitoring construction activity or natural disaster-based changes.

The layers can be accessed via WMS and soon also via STAC. Since the semantic EO data cube enables a spatio-temporal dynamic query, user-defined areas and time periods can be calculated on-demand at any time.

We will present different implementations for the Austrian federal states of Salzburg and Burgenland including different application scenarios.



- Black are areas: without vegetation for the whole-time span.
- White /bright coloured areas: highly vegetated in all time periods.
- Strong red means vegetation in T1, little to no vegetation in T2 and T3.
- Strong green means vegetation in 20T2, little to no vegetation in T1 and T3.
- Strong blue means vegetation in T3, little to no vegetation in T1 and T2.
- Mixed colours: Yellow shades are between T1 and T2 (a mixture of red and green) and mean vegetation in these years, but little to no vegetation in T3 (blue shade missing). Magenta: Colour mixture of red (T1) and blue (T3) - green shade missing, as little to no vegetation in T2 etc.

Figure 1 Example of the RGB legend and the colour coding in respect to the three time periods (T1-T3)

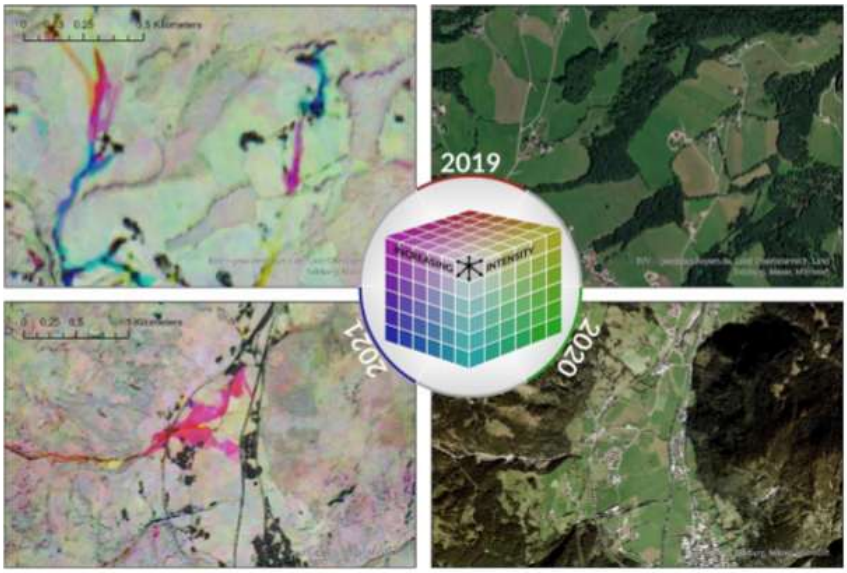


Figure 2 Example for a one-layer representation of the changes of observed vegetation counted from every Sentinel-2 image in the years 2019, 2020 and 2021 (can be adapted to any time period (e.g. different years or seasons)). Upper left: RGB layer representing changes in road construction based on vegetation change derived from all Sentinel-2 images, the colours represent the years when the changes occurred (removal of vegetation during construction, but also vegetation regrowth of parts of the area when the roads were finished). Upper right: VHR image of the same area taken after the changes happened (>2022). Lower left: RGB representation for a mudflow taken place in Bad Hofgastein, Austria, early July 2020. Since the vegetation was removed by the mudflow the colour changes to red (not vegetated parts of 2020 and 2021), for some parts to magenta, which indicates a regrowth of vegetation already in 2021. Lower right: VHR image of the same area taken after the changes happened (>2022)

2.5. SCIENTIFIC ABSTRACT V

ISPRS Journal of Photogrammetry and Remote Sensing 228 (2025) 552–565

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

On-demand, semantic EO data cubes – knowledge-based, semantic querying of multimodal data for mesoscale analyses anywhere on Earth

Felix Kröber [a,b,*], Martin Sudmanns [a], Lorena Abad [a,1], Dirk Tiede [a,*]

[a] Paris-Lodron University Salzburg, Department of Geoinformatics - Z_GIS, Schillerstraße 30, 5020 Salzburg, Austria
[b] Research Centre Jülich, Institute of Bio- and Geosciences, Leo-Brandt-Straße, 52425 Jülich, Germany

ARTICLE INFO

Dataset link: https://github.com/Sen2Cube-at/gsemantique, https://doi.org/10.5281/zenodo.15423258

Keywords:
Earth observation
Remote sensing
Big data analyses
Data cubes
Semantic querying

ABSTRACT

With the daily increasing amount of available Earth Observation (EO) data, the importance of processing frameworks that allow users to focus on the actual analysis of the data instead of the technical and conceptual complexity of data access and integration is growing. In this context, we present a Python-based implementation of ad-hoc data cubes to perform big EO data analysis in a few lines of code. In contrast to existing data cube frameworks, our semantic, knowledge-based approach enables data to be processed beyond its simple numerical representation, with structured integration and communication of expert knowledge from the relevant domains. The technical foundations for this are threefold: Firstly, on-demand fetching of data in cloud-optimized formats via SpatioTemporal Asset Catalog (STAC) standardized metadata to regularized three-dimensional data cubes. Secondly, provision of a semantic language along with an analysis structure that enables to address data and create knowledge-based models. And thirdly, chunking and parallelization mechanisms to execute the created models in a scalable and efficient manner. From the user's point of view, big EO data archives can be analyzed both on local, commercially available devices and on cloud-based processing infrastructures without being tied to a specific platform. Visualization options for models enable effective exchange with end users and domain experts regarding the design of analyses. The concrete benefits of the presented framework are demonstrated using two application examples relevant for environmental monitoring: querying cloud-free data and analyzing the extent of forest disturbance areas.

1. Introduction

In recent years, Earth Observation (EO) data access and processing has changed in a fundamental way. The opening of the Landsat archive in 2008 (Woodcock et al., 2008) enabled broad-scale analyses driven by a steadily growing user base (Wulder et al., 2012; Zhu et al., 2019). Free access to global satellite data at unprecedented spatial and temporal resolution has been further advanced by the Sentinel-2 (S-2) constellation (Drusch et al., 2012) launched in 2015. To facilitate the exploitation of the data, a new paradigm of big EO data processing has emerged (Guo et al., 2017; Sudmanns et al., 2020b). Among the underlying factors powering the evolution of this field, there are two major ones that keep posing challenges to users.

Firstly, accessible EO data is continuously increasing in terms of their volume. The Copernicus Open Access Hub, meanwhile replaced by the Copernicus Data Space Ecosystem, for example, provided a total of over 45 petabyte of data, which had been continuously built up over the previous years (Cipoletta and Sciarra, 2024). While this data availability enables broad-scale analyses in space and time, in practice there are usually limitations on users' ability to process large amounts of data. In addition to specific approaches such as optimizing the data selection (Kempeneers and Soille, 2017), these limitations gave rise to a fundamental change in data management. Instead of downloading data and processing it locally, throughout the last 10 years it became more common to analyze data in the cloud utilizing data models including data cubes (Sudmanns et al., 2020b). Still, a majority of users report limiting processing capabilities and growing data volumes as prevailing obstacles when working with big EO data (Wagemann et al., 2021). One reason for this can be seen in a reluctance to fully shift to cloud-based processing platforms due to their ongoing limitations. Many of these platforms are proprietary, closed-source (Gorelick et al., 2017; Microsoft Open Source et al., 2022), and their usage usually incurs costs. The lack of guarantees on the provision of the service can lead to

* Corresponding authors at: Paris-Lodron University Salzburg, Department of Geoinformatics - Z_GIS, Schillerstraße 30, 5020 Salzburg, Austria.
E-mail addresses: felix.kroeber@plus.ac.at (F. Kröber), martin.sudmanns@plus.ac.at (M. Sudmanns), lorena.abad@plus.ac.at (L. Abad), dirk.tiede@plus.ac.at (D. Tiede).
[1] Deceased author.

https://doi.org/10.1016/j.isprsjprs.2025.07.015
Received 13 February 2025; Received in revised form 20 May 2025; Accepted 8 July 2025
Available online 29 July 2025

unexpected shutdowns impeding analysis reproducibility, most recently experienced in June 2024 with Microsoft Planetary Computer Hub. Furthermore, such platforms exhibit restricted flexibility in terms of extensibility and customization. Meanwhile, open-source alternatives for large-scale data processing exist, but they are usually tedious to set up, e.g. due to data set indexing (Killough, 2018; Baumann et al., 2018). This effort currently limits the usability of open-source frameworks, especially for projects with a shorter runtime.

Secondly, supported by the launch of new satellites and sensors, the variety of data sets keeps growing. A rise in the availability of data sets with different spectral, spatial, temporal and radiometric resolutions poses challenges for multimodal data sources analyses. The SpatioTemporal Asset Catalog (STAC) fostering standardization in the structuring and publishing of geospatial metadata is an important means to facilitate data access. But data utilization and integration are not only about the technical means to query data. Data access is a natural prerequisite but by itself it does not support sophisticated image analytics. An example for optical images is the image understanding process of inferring information on 4-D physical world phenomena using a numerical model that runs on the 2D image domain. In order for EO data analysts to focus on image understanding they need to be provided with adequate means, allowing to query information and model knowledge in a consistent and transparent manner. Querying frameworks that provide a structured approach for EO information extraction are recently evolving (Van Der Meer et al., 2022) but not yet supported in most of the existing EO data cube systems. Many of them focus on technical solutions in terms of provision of data but they do lack the building blocks to aid knowledge-based image understanding.

In brief, we acknowledge the pressing need for open-source big EO data processing frameworks, which are easy-to-use, while having a sufficient conceptual basis to be able to deal with the semantics of EO data. To tackle this gap, we propose a data cube approach based on a semantic querying language that interprets knowledge embedded in semantic models to support big EO operational image understanding. Specifically, the contribution of this paper is a new Python package *gsemantique* that enables building ad hoc data cubes for semantic, knowledge-based EO analyses in on-premise or cloud-based infrastructures. We demonstrate the potential value of this implementation with two use cases.

This paper is structured as follows. In Section 2, we place our package in relation to existing data cube frameworks and recapitulate on the essence of semantic, knowledge-based querying. In Section 3, the technical implementation of the package and underlying design choices are presented. With Section 4 we showcase the general usage of the package. This is followed by the presentation of two specific use cases in Section 5, one focused on cloud-free imagery and the other one on forest disturbances. The paper concludes with a reflection on limitations and future works in Section 6 and a summary in Section 7.

2. Related works & conceptual foundations

2.1. EO data cubes

A data cube is a multi-dimensional array whose grid points are populated with data of the same data type (Baumann, 2017). The data values are indexed unequivocally by coordinates along the d axes of the d-dimensional data cube. In the EO domain, data cubes typically have at least two spatial and one temporal dimension, and the coordinates span the full spatiotemporal extent of a given set of observations (Lewis et al., 2016). The primary feature of EO data cubes is that the data is reorganized such that from a logical view data can be queried easily using spatiotemporal coordinates and abstraction of analytics from storage considerations. This data organization replaces the traditional file-based access for users, which is limited by files being organized in nested directory structures in a spatiotemporally inconsistent manner with custom naming patterns. EO data cubes therefore provide more convenient access to data, facilitating their analyses by reducing the pre-processing effort, which in turn is closely linked to the provision of analysis ready data (Giuliani et al., 2017). Beyond its array structure, conceptual disagreement about the essence of a data cube is still prevalent. Baumann et al. (2016) defined a set of technical requirements data cubes should adhere to. Strobl et al. (2017) extended this set of properties by providing a holistic view on six system-level aspects that need to be considered to realize the full potential of data cubes. Despite the valuable criteria provided, practical implementation considerations result in a broad variety of systems currently operating under the term 'EO data cube'. Therefore, we stick with the universal array definition of the EO data cube and subsequently highlight the specifics of our approach by comparing it with other EO data cube implementations. Note that proprietary geospatial web-based processing platforms including Google Earth Engine (Gorelick et al., 2017) and Microsoft Planetary Computer (Microsoft Open Source et al., 2022) are deliberately excluded from the comparison. A comprehensive overview on web-based processing frameworks can be found in Gomes et al. (2020). For the difference between such infrastructures and EO data cubes, the reader is referred to Giuliani et al. (2019).

One of the first operational national-wide EO data cube implementations was the Australian Geoscience Data Cube (AGDC) (Lewis et al., 2016). Whereas initially data was ingested, i.e. restructured via resampling and tiling, further developments shifted towards data indexing (Lewis et al., 2017), where the data is stored in its native format without being replicated. The evolution of the AGDC gave rise to the Open Data Cube (ODC) initiative (Killough, 2018) providing a set of open-source tools to create data cube infrastructures. The ODC approach gained attention rapidly (Dhu et al., 2019; Killough et al., 2020) with a variety of data cubes representing ODC instances including, for example, the Swiss Data Cube (Giuliani et al., 2017; Chatenoux et al., 2021), the Colombian Data Cube (Ariza-Porras et al., 2017), the Armenian Data Cube (Asmaryan et al., 2019), the Catalan Data Cube (Maso et al., 2019), the Vietnam Open Data Cube (Quang et al., 2019), the Austrian Semantic EO Data Cube (Sudmanns et al., 2021) and Digital Earth Africa (Yuan et al., 2021). An alternative array data base solution is provided by Rasdaman (Baumann et al., 2018) deployed in Baumann et al. (2016) and Storch et al. (2019), where the data is tiled into sub-arrays according to specific partitioning strategies and ingested into a data base to optimize the retrieval efficiency. All of the above-mentioned systems are united by their property of being extensive software infrastructures consisting of several components (e.g. modules for data pre-processing, data bases, APIs for data querying, monitoring tools).

Some efforts have been made to lower the hurdles for setting up such complex software infrastructures. In line with the idea of self-hosted deployments of local, federated EO data cubes (Sudmanns et al., 2023), Giuliani et al. (2020) proposed a proof-of-concept for the automated generation of ODC instances. The user only needs to specify an area, time frame and sensor of interest to retrieve an ODC instance. As an all-in-one solution, Frantz (2019) proposed FORCE for the processing of large amounts of S-2 and Landsat data. Without any data base-driven indexing or ingestion, FORCE provides a suite of algorithms to create regularly tiled, analysis ready data on Level 2 or even higher levels, where all data for a given tile is referred to as a data cube. Despite these developments, from the user's point of view the time required for the initial creation of a populated data cube is quite high as, in the cases mentioned, data cubes are created with the original data being first downloaded and persisted on the disc. This results in static, infrastructure-oriented data cubes that are tailored to a few data products. In contrast, nowadays, many usable EO data products are already available on the web as analysis ready data, and a fast and flexible integration of different data products for on-the-fly analyses is desired.

The idea of on-demand or on-the-fly cubes summarizes approaches to create data cubes in an ad-hoc fashion for any specified spatiotemporal extent of interest. While lacking some of the functionalities and

performance benefits of more comprehensive data cube approaches, the on-demand approaches have the advantages of being lightweight and easy-to-use. xcube (Brockmann Consult GmbH, 2021) creates self-contained EO data cubes by relying on Python's big data ecosystem, specifically xarray (Hoyer and Hamman, 2017) for in-memory representations, dask (Dask Development Team, 2016) for memory management and Zarr as a format for cloud-native, chunked storage. Implemented in the Euro Data Cube (Euro Data Cube Consortium) and the multivariate Earth system data cubes as part of the Earth System Data Lab project (Mahecha et al., 2020), xcube is used in operational systems. Supplemented by the ml4xcube library (Peters et al., 2025), not only the creation but also the data-driven analysis of EO data cubes is facilitated. To automate the creation of mini data cubes as xarray objects from STAC catalogs, cubo has been proposed by Montero et al. (2024a). The open source C++ library gdalcubes (Appel and Pebesma, 2019) natively provides chunked management and parallel processing of EO data cubes. It can integrate with scripting languages such as R, Python or Julia or with software that can handle data cubes such as GRASS GIS (Neteler et al., 2012). Its R implementation depends on the stars package (Pebesma and Bivand, 2023), which enables the reading and processing of spatiotemporal arrays and allows proxy objects with lazy loading for larger rasters. Additionally, gdalcubes offers a set of predefined formats to load various EO products as image collections and convert them to regularized data cubes. Image collections can also be built from STAC catalogs assets accessed via the rstac package (Simoes et al., 2021b), developed as part of the Brazil Data Cube (BDC) project (Ferreira et al., 2020). A more comprehensive effort, also stemming from the BDC project, is the sits package (Simoes et al., 2021a), built on top of gdalcubes. sits has a tailored focus on satellite image time series analysis using data-driven techniques. It allows to build data cubes from various cloud-based providers of EO images and train machine and deep learning models on them with support for parallel processing and chunking along the spatial dimension. While most of the listed approaches offer a specific solution for data cube creation, only some offer end-to-end frameworks that allow to realize a full processing pipeline with the final aim of producing tailored information from EO data. The framework proposed by us supports end-to-end EO analysis with a specific focus on semantic querying and knowledge integration. The essence and relevance of semantic, knowledge-based querying for remote sensing-based image understanding is detailed further in Section 2.2.

2.2. Remote sensing based image understanding

The general vision process based on remote sensing data amounts to reconstructing a semantic 4D scene reality from sub-symbolic 2D image data (Matsuyama and Hwang, 1990). This makes it an inherently ill-posed problem. The following describes how semantic and knowledge-based systems, which are forming the basis for the analysis backbone of our on-demand EO cubes, deal with this complex task.

2.2.1. Semantic systems

Semantics, as the study of meaning, deals with the relation between physical world phenomena, mental concepts, and the expressions used to interconnect both. Enabling machines to be capable of handling semantics is fundamental for any advanced human-machine or machine-machine interaction, and not specific to the EO domain, as exemplified by the semantic web (Berners-Lee et al., 2001). A central feature of semantically-enabled systems is that beyond data itself, information as interpretations of data can be accessed and handled. In the EO context, semantic data cubes thus refer to systems that leverage interpretations of EO images, where for each spatiotemporal observation at least one interpretation is available (Augustin et al., 2019). Those interpretations are effectively mapping numerical sensory data to stable concepts. Semi-symbolic spectral categorizations as proposed by Baraldi (2011), for example, provide low level interpretations. They allow an initial characterization of the data, e.g. by splitting the continuous multispectral reflectance space into a discrete set of physically meaningful categories, but they do not represent physical world entities. In contrast, high level interpretations are given by concepts that adhere to existing ontologies describing physical world entities such as land cover classes according to the FAO LCCS (Di Gregorio et al., 2016). The common property of both types of interpretations is that they represent semantic enrichments transforming continuous, numeric data into interpretable categorical data, which essentially lifts elements from the lower data level to the next level of the data-information-knowledge-wisdom hierarchy (Rowley, 2007). Since the essence of image analysis and understanding is to transform data into information, most computer vision tasks including EO analyses are inherently dealing with questions of semantics.

2.2.2. Knowledge-based systems

Knowledge is a vague concept, but commonly referred to as structured, contextualized, and synthesized information (Rowley, 2007). Relevant for computational systems, knowledge enables to translate information into instructions, thereby allowing the guidance of systems (Ackoff, 1989). For the purposes of an image understanding system, various types of knowledge are required. Those include generic knowledge on problem solving and image understanding, remote sensing domain specific knowledge to achieve a meaningful mapping between numeric and symbolic representations, and knowledge on user interfaces to allow interaction with humans as knowledgeable, intelligent system users (Crevier and Lepage, 1997). There is a variety of knowledge representation techniques dealing with how knowledge is embedded and represented in systems. In the field of image understanding, those representation techniques are essentially aiming to transform data into information, i.e. to gain actionable insights from data. Baltsavias (2004) presents an overview of knowledge representations that are commonly used in the domain of remote sensing-based image understanding. They belong to the realm of more established, old-school artificial intelligence systems.

More recently, the field of image understanding has been supplemented and in large parts dominated by the suite of machine and deep learning techniques (Mountrakis et al., 2011; Belgiu and Drăguţ, 2016; Zhu et al., 2017; Hoeser and Kuenzer, 2020; Hoeser et al., 2020). While the design of these techniques and their selection for a specific task certainly involve knowledge, the actual reasoning process itself is carried out in an inductive, data-driven way. In the classical supervised paradigm, machine and deep learning techniques are essentially statistical models learning from examples. Baraldi and Boschetti (2012) and Baraldi et al. (2023) argue that these models are not only poorly based on correlation instead of causation, but that the image understanding process with these models remains ill-posed, because external input is required for the scene reconstruction. They emphasize the need for a-priori knowledge to make the vision problem better conditioned for solution. While data-driven models can be adapted by incorporating a-priori knowledge, one can argue that knowledge-based systems operating in a deductive manner remain a valuable complementary approach to tackle image understanding. Arvor et al. (2021) promote knowledge-based approaches focusing on their advantage to explicitly deal with symbolic information and its organization. Craglia and Nativi (2018) emphasize that specifically in the big data era with the prevalence of data-driven inferences, a focus on transparent modeling and exchange of domain knowledge is needed to increase the trust in inferences made on big data. Scheider et al. (2017) focus on the relevance of integrating existing knowledge into analyses for reproducible generation of information and further knowledge instead of pursuing knowledge acquisition in a purely data-driven manner. This coincides with a broader epistomological belief that despite the prevalence of data-driven approaches, the synergy of deductive and inductive approaches is required to drive information derivation and knowledge acquisition (Mazzocchi, 2015).

With our system design, we are referring to expert systems (Good-enough et al., 1987; Laurini and Thompson, 1992; Matsuyama, 1993), which are representing the knowledge of a human expert in a declarative manner. Declarative knowledge ("knowledge-that") refers to factual information regarding physical world entities and their properties, and is often contrasted with practical knowledge ("knowledge-how") and knowledge by acquaintance ("knowledge-of"). Declarative knowledge can be represented explicitly, e.g. as symbolic data embedded in logical production rules such as "if-then" constructs. Importantly, the knowledge is independent from the procedural process of reasoning and its usage is not tied to a single use case. In terms of system design, this explains the typical decomposition of an expert system into two parts: the knowledgebase as a collection of symbolic data and the reasoning engine as the task-solving program that infers information by leveraging the knowledgebase within a specific reasoning process. The separation of knowledge from reasoning fosters modularity and enables knowledge sharing.

2.2.3. Synthesis: Semantic, knowledge-based systems

Given the descriptions on semantically-enabled and knowledge-based systems, one may ask to what extent both approaches can be seen independent of each other. Are all semantically-enabled systems inevitably knowledge-based systems and vice versa? Semantic enrichments are not exclusively generated by knowledge-based methods but can also be produced by data-driven methods (Baraldi and Boschetti, 2012). Semantically-enabled systems can be used within knowledge-based expert systems but can also be described on their own (Augustin et al., 2019). Knowledge-based systems aiming at image understanding are always concerned with semantics but not necessarily in an explicit manner. The degree of formalization of shared conceptualizations known as ontologies can be low in common rule-based modeling approaches that leverage symbolic knowledge implicitly (Arvor et al., 2019). Furthermore, knowledge-based systems do not per se target the inclusion of semantically enriched information in their reasoning processes. Therefore, semantic and knowledge-based systems are neither synonymous nor dependent on each other, but can indeed be used in a complementary, synergistic manner.

As a foundational work for the design of a corresponding semantic, knowledge-based image understanding system, we refer to the Austrian semantic data cube (Sudmanns et al., 2021). Following the design of an expert-based system, semantic models from the knowledgebase are processed by an inference engine against the factbase containing image data. Following the idea of semantically-enabled systems, the factbase stores additional semantically enriched information layers, and the inference engine allows to create and execute semantic models that are targeted to processing such categorical data. The way knowledge can be represented and embedded in semantic models is formalized via a semantic querying language, *semantique*, as described by Van Der Meer et al. (2022). Similar to Sudmanns et al. (2021), with our proposed *gsemantique* package, we built on top of this querying language to extend it towards an operational image understanding system. In contrast to Sudmanns et al. (2021) and according to Section 2.1, our system focuses on the creation of on-demand cubes in an ad-hoc fashion. To further illustrate the relationship between the existing systems in the field of semantic, knowledge-based image understanding, the reader is referred to Fig. 1.

3. System requirements & architecture

3.1. Design goals

We define the following design goals for our data cube framework. The goals are framed as user's requirements reflecting the specific expectations of users interacting with an on-demand data cube system that aims to support ad-hoc EO analyses and image understanding.

(A) Data Access

- (A.1) Coverage: Users can query data world-wide for any spatial or temporal extent of interest having a range of predefined data sets at hand
- (A.2) Extensibility: Users can easily index new data sets to integrate them in their analyses
- (A.3) Persistence: Data for generating the data cube can be persisted locally or in the cloud to enable inspection of input data, ensure reproducibility of analyses, and speed up the calculations in case of repeated execution of similar analyses by moving the data location closer to the computing infrastructure

(B) Analysis

- (B.1) Basic support: Provision of a standard set of spatial and temporal data cube operations
- (B.2) Semantic support: Modeling image semantics including…
  - model formulation from a semantic point of view and automated translation into procedural data cube operations
  - possibilities for custom modeling of entities of interest
  - support for categorical data operations to interact with semantically enriched data
- (B.3) Expert knowledge support: Means to represent expert knowledge in a model
- (B.4) Customization: Possibility to define analysis workflows including user-defined functions
- (B.5) Visualization, Communication, Exchange: Automated export and visualization of models to exchange with other domain experts and communicate models

(C) Processing

- (C.1) Scalability: Support for analyses with spatiotemporal extents up to mesoscale dimensionality
- (C.2) Efficiency: Fast data access and processing, minimization of redundancies in data loading, and exploitations of available processing resources
- (C.3) Abstraction of complexity: Automated model execution requiring minimal user interaction

(D) General software requirements

- (D.1) Portability: Executability on local devices as well as cloud-based platforms
- (D.2) Usability: Simple client-side installations via package managers; usability via common Python programming language enabling big EO analysis in a few lines of code

3.2. Implementation

To implement these requirements, we extended the existing semantic querying language *semantique* (Van Der Meer et al., 2022) and built a new package, *gsemantique*, on top of it. In terms of functionality, *semantique* represents the general modeling framework and inference engine responsible for the core analysis support (requirements (B)). *gsemantique* represents a wrapper around *semantique* to ensure data access

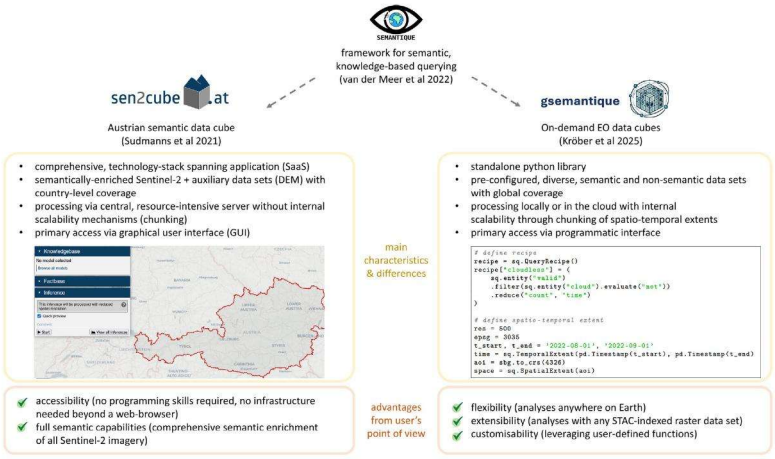


Fig. 1. Relationship of our work to other semantic, knowledge-based EO analysis systems.

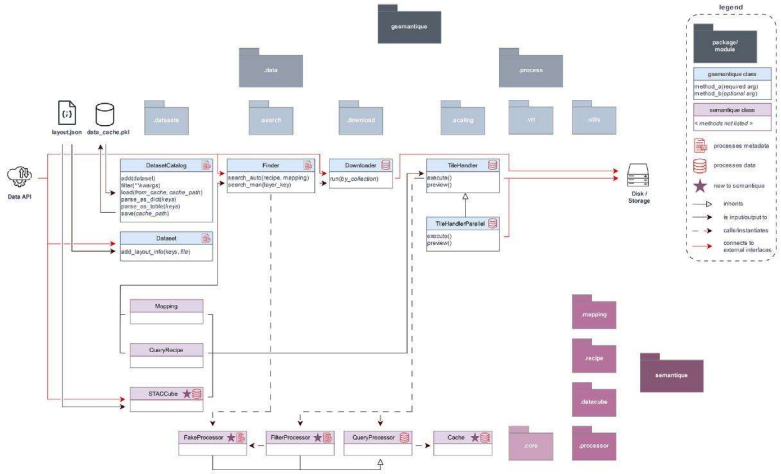


Fig. 2. Implementation of our on-demand EO cube architecture with package structure and main classes.

(requirements (A)) and efficient processing at scale (requirements (C)). Both packages are lightweight Python libraries fulfilling the general software properties as formulated in requirements (D). From a technical point of view, the core structure of both packages and their relationship is depicted in Fig. 2. Note that for *gsemantique* all classes are shown, whereas for *semantique* only a selection of relevant classes is shown to keep the figure clear and concise. The precise translation of the first three groups of requirements (A)-(C) into architectural design choices as shown in Fig. 2 is described subsequently.

In terms of data access (A.1–A.3), STAC is used as a widely accepted metadata standard to manage data description and retrieval in a consistent manner. The *STACCube* class implemented in *semantique* allows to create data cubes based on item collections as results of STAC metadata searches. The metadata searches themselves are encapsulated in the *Finder*. To search for STAC items, catalog endpoints and collection names need to be given. These are organized together with other metadata as *data set* objects added to a *DatasetCatalog*. A predefined *DatasetCatalog* is stored within *gsemantique* (A.1). Currently, it covers references to a total of 13 data sets (collections) with

Table 1
Data sets with global coverage ready-to-use in *gsemantique* via predefined data set objects and layout file. This list can be extended with any STAC-indexed data set.

Data set information			Data layer information			Temporal information	
STAC collection	**STAC catalog**	**Category**	**# Layers**	**Non-Semantic ones**	**Semantic ones**	**Extent**	**Frequency**
sentinel-1-rtc	[A]	SAR	4	Amplitude in four polarizations	–	2014 (Oct) – today	Sub-daily
sentinel-1-global-coherence	[B]	SAR	17	6-, 12-, 18-, 24-, 36-, 48-day coherence in VV and VH polarizations	–	2020	Quarterly
sentinel-2-l2a	[A]	Multispectral	13	Twelve reflectance bands	Scene Classification Map (SCM)	2015 (June) – today	Sub-daily
sentinel-2-l2a	[C]	Multispectral	13	Twelve reflectance bands	SCM	2015 (June) – today	Sub-daily
landsat-c2-l2	[A]	Multispectral	10	Nine reflectance bands	Quality assessment band	1982 (Aug) – today	Sub-daily
esa-worldcover	[A]	Landcover	1	–	10-class LULC layer	2020–2021	Yearly
io-lulc-annual-v02	[A]	Landcover	1	–	9-class LULC layer	2017–2023	Yearly
nasadem	[A]	DEM	1	Elevation layer	–	2000	Static
cop-dem-glo-30	[A]	DSM	1	Elevation layer	–	2010–2015	Static
modis-64A1-061	[A]	Fire Detection	3	Ordinal burn date, burn date uncertainty	Quality assessment band	2000 (Nov) – today	Monthly
modis-14A2-061	[A]	Fire Detection	2	–	Categorization of fire confidence, quality assessment band	2000 (Feb) – today	Monthly
jrc-gsw	[A]	Hydrogeography	4	Water frequency, frequency changes	Binary water existence, categorical changes in surface water status	1984–2020	Annual
glo-30-hand	[B]	Hydrogeography	1	Height above nearest drainage layer	–	2010–2015	Static

[A] https://planetarycomputer.microsoft.com/api/stac/v1, [B] https://stac.asf.alaska.edu, [C] https://earth-search.aws.element84.com/v1

more than 70 individual bands (assets) as shown in Table 1. Using predefined methods to add new data sets, the *DatasetCatalog* can be extended flexibly to include any data set for which a STAC catalog endpoint and collection name can be specified (A.2). This is not limited to dynamic STAC catalogs but also includes static ones. Users can therefore address a wide range of additional, publicly available data sets (https://stacindex.org) as well as local ones as long as STAC-conformant metadata is available. Importantly, *gsemantique* has another object containing data set information, which is the layout json file that is used to initialize the *STACCube*. This layout file contains metadata not on the data set level but the individual data layers (i.e. the assets). It is relevant to guide data fetching since it defines data types, value ranges and missing data values. Together, the predefined *DatasetCatalog* and layout file structure the variable parts of EO data and define which data is accessible for subsequent data cube construction. Finally, storage and persistence functionality for the input data is realized via the *Downloader* (A.3). It leverages a library for asynchronous, non-blocking I/O tasks to efficiently manage the high-concurrency operation of fetching data for the metadata search results as obtained by the *Finder*, and transfer it to a specified location. The data is automatically STAC-indexed by building a static STAC catalog and collection for seamless integration in further analyses including data cube construction.

Regarding analysis support (B.1–B.5), *semantique* (Van Der Meer et al., 2022) already provided a strong basis for standard data cube operations (e.g. aggregation via reduce-through-space/time) (B.1), modeling semantic concepts (B.2), and representing expert knowledge (B.3). These core components of the modeling are implemented using corresponding predefined structures (mapping and recipe) as described in Van Der Meer et al. (2022) and again briefly outlined in Section 4. The set of predefined modeling functions, which can be used within these structures, can be flexibly extended by user-defined functions. This is ensured by corresponding interface functions that enable the integration of any standard Python code (B.4). Extensions of this existing functionality mainly concern two points: The first one targets effective communication and exchange about analysis workflows by adding visualization options to represent semantic models graphically (B.5). Following Sudmanns et al. (2021), we rely on the JavaScript library Blockly (Google, 2024) with custom definitions of visual blocks for corresponding representations. The second extension of *semantique*, covers the increased complexity of translating semantic models into procedural code for on-demand cubes. The difficulty here is that the user defines a semantic model (e.g. using entities such as clouds), while the implementation requires to resolve the underlying data layer references (e.g. SCM of S-2) along with their queried spatiotemporal extent. Indeed, semantic querying on static, persistent data cubes such as ODC instances also requires this translation into numerical code. However, unlike their more comprehensive counterparts, on-demand cubes lack the underlying native capabilities for efficient retrieval of data, e.g. via data base indexing. This problem is exacerbated by the fact that the user can query data from the comprehensive totality of all EO data worldwide and the data are not necessarily stored on high-performance infrastructures with low retrieval latencies. A feasible implementation of semantic, on-demand data cubes therefore requires the metadata for constructing the data cubes to be narrowed down as far as possible in advance in order to speed up the subsequent actual data retrieval. This task is carried out by *FakeProcessor* and *FilterProcessor* objects. *FakeProcessor* instances resolve all semantic concepts in a model into data layer references prior the actual model evaluation. *FilterProcessor* instances evaluate the required temporal extents for which the referenced data must be loaded according to the model. If a model contains temporal filter operations for individual data layers, these are resolved using *FilterProcessor* before the actual data fetching. *FakeProcessor* and *FilterProcessor* together enable the user to perform modeling completely at the semantic level while ensuring an efficient translation into numerical code.

Finally, processing requirements concerning scalability and efficiency with abstracted complexity (C.1–C.3) are implemented in *gsemantique* via the *TileHandler* and *TileHandlerParallel*. Enabling scalability in a hardware-independent way, i.e. not relying on vertical scaling,

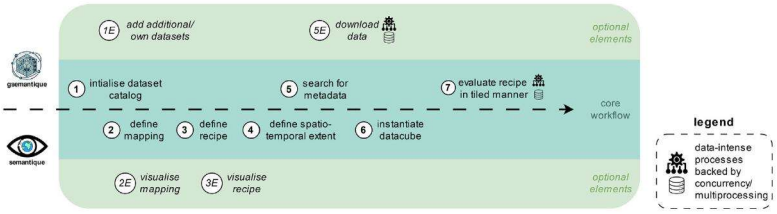


Fig. 3. Generalized end-to-end EO analysis workflow using *gsemantique*.

is realized by tiling the overall data cube into smaller chunks to run the code sequentially on them with a final merger of the chunked results (C.1). This is implemented by the *TileHandler* class. It analyses the recipe for operations executed on the spatial and/or temporal dimension and chooses the remaining one as a chunking dimension. A preview is calculated containing information on the number of chunks to be processed and the expected size of the output. Instead of sequential processing of the chunks, parallelization via multicore processing is possible using the *TileHandlerParallel* class (C.2). In both cases, the data for each chunk is cached using an instance of the *Cache*, ensuring that multiple requests of the same data within a semantic model are handled with an efficient one-time data fetching process (C.2). The chunked results are finally merged as single outputs or optionally virtual rasters in case of spatial outputs. Opting for virtual rasters enables storage-efficient processing of non-rectangular areas of interest. The overall processing complexity is abstracted from the user by full automation of the split-apply-merge workflow based on reasonable default values, e.g. for spatial and temporal chunk sizes (C.3).

4. System utilization

Using *gsemantique* to conduct EO analysis with on-demand data cubes is a matter of a few lines of Python code. The general workflow fundamental to any analysis is outlined in Fig. 3 and an example on how to translate this into code for a specific analysis is demonstrated in Fig. 4.

The seven stage workflow starts with the initialization of the *DatasetCatalog*. Usually, this amounts to reading the metadata for all predefined data sets as listed in Table 1. However, if a user wants to add custom data sets to be employed in the following analysis, the *DatasetCatalog* can be instantiated based on any custom layout file that extends the predefined data sets. The subsequent definition of the analysis model, is split into two parts. First, the definition of a *Mapping* as a collection of entities is carried out. The entities represent semantic concepts defined by a set of properties. In the case shown in Fig. 4, the entities are 'valid observations' and 'clouds', both being defined by their corresponding classification according to the S-2 Level 2 A SCM. More advanced cases, where the properties that define entities are not immediately drawn from already existing classifications, are shown in Section 5.2. Second, the definition of a *Recipe* as the application part is required. The entities of interest are transformed via spatiotemporal processing and other data cube operations to derive analysis results. Both parts, *Mapping* and *Recipe* represent Python dictionaries, but can be visualized in a graphical block structure (Fig. 7). The fourth stage of the workflow requires the user to define the area and time frame of interest to select the spatiotemporal extent for which the analysis should be carried out. Given this selection and the data references as encoded in the *Mapping* and *Recipe*, the user can query the metadata of all EO files necessary to calculate the results. Optionally, the user can query the raster data and download them, to persist the data on the disc. The sixth stage is the instantiation of the *STACCube*, which takes the previously acquired metadata with pointers to the data locations as an input. The creation of the data cube object itself does not load any data. The same lazy-loading strategy applies to the instantiation of the *TileHandler* in the final stage, where the general structure to calculate results is set up based on the previously defined *Mapping*, *Recipe*, *STACCube*, spatiotemporal extent. With a single call to execute the processing, the *TileHandler* takes care of calculating the results while abstracting the technical complexity of chunking the data cube with sizes that fit into the main memory during processing. If desired, the user can decide to tune the default chunk sizes along the spatial or temporal dimension by specifying them as arguments upon the instantiation of the *TileHandler* object.

5. Demonstration

The following application cases demonstrate the strength and flexibility of the on-demand data cube approach for semantic, knowledge-based querying of EO data. A summary of the use cases and their specific objectives is provided in Table 2.

5.1. Application I – cloud-free scenes

Frequently, clouds are obscuring the Earth's surface, complicating analyses based on optical remote sensing. Large parts of Europe, western North America, and areas within the equatorial low-pressure trough are characterized by cloud frequencies greater than 50% (Wilson and Jetz, 2016). Spatial information about the frequencies of cloud-free observations for a given satellite can be leveraged to anticipate possible complications in applying models in areas with few available cloud-free scenes, or to select areas rich of cloud-free observations as promising study areas. A corresponding evaluation on the availability of cloud-free scenes for S-2 is presented by Sudmanns et al. (2020a), for example. However, this analysis is based on scene-wide metadata on cloud coverage. With semantically enriched data, where label information on the existence of clouds is available at the pixel level, such analyses can be carried out with increased spatial precision. Beyond semantically enriched data availability, the prerequisites encompass a modeling framework that supports the construction of semantic queries. Furthermore, a processing engine allowing efficient and scalable computations is required since aggregating cloud-free observations over time on the pixel level is a resource- and data-intensive process (Table 2). Finally, the task is well-suited to be solved via on-demand data cubes as users may want to retrieve cloud-coverage data once to select their study area of interest without the need for extensive area-wide computations justifying the effort to setup a persistent, more comprehensive data cube framework. As shown in Fig. 4, conducting such analyses is a relatively simple task using *gsemantique*. Applying the script in a slightly modified version on a continental scale leads to the results as depicted in Fig. 5.

Conclusions that can be drawn based on Fig. 5 about the availability of cloud-free observations include spatially precise details, e.g. on the recording geometry of S-2 data in strips as well as topographical effects such as orographic cloud formation. It should be noted that, on a

```python
# general imports
import geopandas as gpd
import json
import os
import pandas as pd
import semantique as sq
import gsemantique as gsq

# step 1: load data catalog
ds_catalog = gsq.DatasetCatalog()
ds_catalog.load()

# step 2: define mapping
# i.e. relationship semantic concepts <-> numeric values
mapping = sq.mapping.Semantique()
mapping["entity"] = {}
mapping["entity"]["valid"] = {
    "class": (
        sq.layer("Planetary", "classification", "scl")
        .evaluate("not_equal", 0)
    )
}
mapping["entity"]["cloud"] = {
    "class": (
        sq.layer("Planetary", "classification", "scl")
        .evaluate("in", [8, 9, 10])
    )
}

# step 3: define recipe
# i.e. processing of semantic concepts
recipe = sq.QueryRecipe()
recipe["cloudless_count"] = (
    sq.entity("valid")
    .filter(sq.entity("cloud").evaluate("not"))
    .reduce("count", "time")
)
```

```python
# step 4: define spatio-temporal extent
epsg = 3035 # coordinate reference system (CRS)
res = 500 # resolution in CRS units
t_start, t_end = '2022-01-01', '2023-01-01'
time = sq.TemporalExtent(
    pd.Timestamp(t_start),
    pd.Timestamp(t_end)
)
aoi = gpd.read_file("polygon.geojson").to_crs(4326)
space = sq.SpatialExtent(aoi)

# step 5: search for metadata
fdr = gsq.Finder(
    ds_catalog,
    t_start,
    t_end,
    aoi
)
fdr.search_auto(recipe, mapping)

# step 6: instantiate datacube
with open(gsq.LAYOUT_PATH, "r") as file:
    dc = sq.datacube.STACCube(
        json.load(file),
        src = fdr.item_coll
    )

# step 7: evaluate recipe
# create TileHandler instance & execute processing
context = dict(
    recipe = recipe,
    datacube = dc,
    mapping = mapping,
    space = space,
    time = time,
    spatial_resolution = [-res, res],
    crs = epsg
)
th = gsq.TileHandlerParallel(n_procs = 12, **context)
th.execute()
```

Fig. 4. Code example for end-to-end EO analysis workflow using *gsemantique*.

Table 2
Summary of use cases realized via *gsemantique*.

Use case description		Analysis extent		Processed input data[a]		Design rationale & aim
Main focus	**Subpart**	**Space**	**Time**	**Number of scenes**	**Download volume**	
Cloud-free scenes	Evaluate number of cloud-free observations through time	Europe (excl. French overseas territories), 5.837.000 km^2	2021–2023	S-2: 872,541	S-2: 1128.75 GB	• Highlight the benefits of basic semantic querying capabilities in an EO data cube framework • Showcase scalability possibilities and limits
	Create monthly cloud-free composites	Lower Austria, 19.200 km^2	2022	S-2: 1475	S-2: 773.89 GB	
Forest disturbances	Analyze magnitude and persistence of forest disturbances	Lake Irrsee at the border between Salzburg and Lower Austria, 78.5 km^2	2020–2023	ESA Worldcover: 1 DEM: 1 S-1: 8 S-2: 564	ESA Worldcover: 84.8 MB DEM: 18.7 MB S-1: 14.7 MB S-2: 945 MB	• Demonstrate possibilities for semantic modeling beyond using predefined entities by defining own entities using multimodal queries • Showcase flexibility to define and compare different models based on different data sets & modeling assumptions

[a] *The number of scenes and the download volume both depend on the native format of the data. The number of scenes is calculated as the sum of the unique files (e.g. for S-2, the sum of the unique product URIs), while the download volume includes the size of all bands that are actually processed (e.g. for S-2, the SCM for the cloud statistics, and the R,G,B & NIR bands for the cloud composites).*

consumer-grade infrastructure, the calculations for Fig. 5 can only be created on a monthly basis and aggregated post-hoc annually due to the northern areas with multiple overlapping orbits causing a high dimensionality of the chunks with more than a thousand observations in the temporal domain. For local calculation on consumer-grade infrastructure, the analysis for a single month takes several hours, as large amounts of data (Table 2) have to be transferred and processed. Efficiency gains can be realized by deploying *gsemantique* on well-equipped cloud infrastructures close to the data servers. A detailed comparison of total model execution times under varying cloud server configurations can be found in the supplementary material of this article.

Beyond the calculation of cloud coverage statistics, the usability of cloud cover information also concerns filtering data with high cloud coverage as a pre-processing step in many EO processing workflows. A common technique is the creation of cloud-free composites as a higher level, analysis ready data product for subsequent processing. Creating such composites via semantic querying involves the pixel-based exclusion of all cloudy observations, conducting a temporal aggregation of reflectances for all observations that are flagged as non-cloudy. Using the SCM accompanying S-2 Level 2 A data for the definition of cloud entities, exemplary results for this simple semantic approach are shown in Fig. 6 (proposed approach). This can be contrasted with the non-semantic median composite as an alternative approach. This compositing technique is based on statistical assumptions on how clouds can be filtered indirectly, namely that the median reflectance of an entire time series is likely to represent cloud-free conditions (Fig. 6, baseline a). If scene-wide metadata on cloud-coverage is available, it can be used to further enhance the applicability and robustness of the median compositing by pre-filtering scenes with low cloud coverage, building the median only among those pre-filtered scenes (Fig. 6, baseline b).

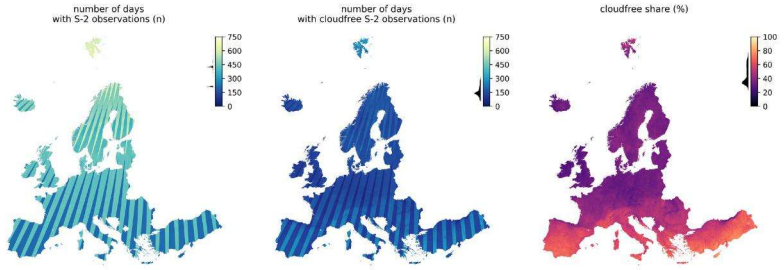


Fig. 5. Availability of S-2 observations over Europe for the years 2021–2023. The proportion of cloud-free observations shown on the right represents the ratio between the middle and left subfigure.

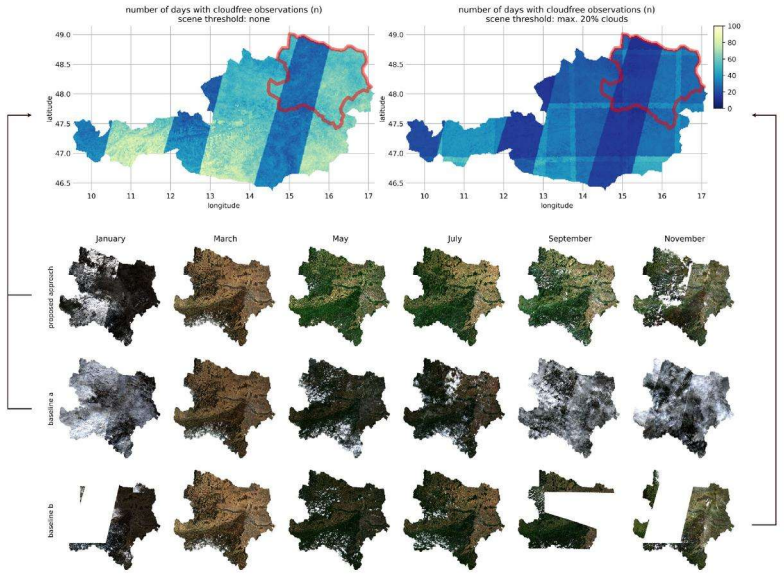


Fig. 6. Cloud-free composites using semantic and non-semantic median compositing. The upper subfigures indicate the number of observations available for composite formation. For a detailed description of the compositing techniques themselves (lower subfigures), refer to the explanations in Section 5.1.

As evident from the true color RGB visualizations for a selection of six out of twelve monthly composites, the semantic approach leads to superior results in terms of a consistent exclusion of clouds on a pixel-level. Under very cloudy conditions, the non-semantic median composite is natively incapable of selecting cloud-free observations. Pre-filtering the scenes reduces the number of available observations, possibly leading to a complete lack of observations for a given month. The observation is neither surprising nor is the applied semantic compositing technique novel. The point here is that the example clearly illustrates the effectiveness and relevance of a knowledge-based querying framework capable of exploiting the full amount of semantically enriched observations in a spatiotemporal data cube.

5.2. Application II – forest disturbances

Assessing the status of forests via remote sensing is a common research and application field with partially overlapping branches of investigating forest degradation (Hirschmugl et al., 2017; Gao et al., 2020), forest disturbance (Frolking et al., 2009; Hirschmugl et al., 2017; Paulino et al., 2024) and forest health (Lausch et al., 2016, 2017). With our use case, we follow Paulino et al. (2024) in a defining forest disturbances as natural or anthropogenic events resulting in forest changes that are identifiable by means of EO.

The task of forest disturbance mapping has three distinct aspects that makes it well-suited to be tackled with our framework: Firstly,

forest disturbances are inherently process-based, i.e. they are characterized by a temporal change in the forest's status. A data cube-based approach is therefore well positioned, as it allows to query every single observation through time. Secondly, the task is characterized by terminological complications with the existence of slightly different definitions of the phenomenon to be modeled. Remote sensing experts thus need to make their specific modeling assumptions explicit and transparent to enable meaningful exchange with others. The definition issues together with difficulties in acquiring ground truth data also cause a lacking availability of homogeneous label data sets impeding data-driven ways to solve the task. A semantic, knowledge-based modeling approach that allows to create a set of human-readable models based on different modeling assumptions is therefore a viable alternative. The third and final aspect of forest disturbance modeling is that neither the entity of interest, i.e. the forest, nor the process, i.e. the forest's change, are straightforward to be modeled. The entity and process of interest can be characterized by more than one property, which may require multimodal data usage to combine sensory information. Our framework provides explicit means to connect physical world and numerical views by mapping multiple properties of an entity in the semantic domain to features of the object in the image domain, leveraging a diverse set of predefined data sets for feature definition. For demonstration purposes, we generate a model suite with a total of four models based on two entity definitions of forest in an undisturbed state crossed with two process definitions that model potential disturbances (Fig. 7).

The entity definitions showcase two possibilities to calculate the extent of forests either in a data-driven or knowledge-driven manner. For the former, we simply equate forests with the tree cover class of ESA Worldcover (Zanaga et al., 2021), which has been generated via a gradient boosting decision tree applied to multimodal data. For the expert-knowledge-based definition, we create a set of forest properties with a corresponding mapping to numerical representations by relying on our own knowledge but also insights of former remote sensing studies on forests. We assume forests to be characterized by temporal stability translated to low radar coherence (Jacob et al., 2020; Nikaein et al., 2021; Borlaf-Mena et al., 2021), complex structure translated to a low-to-medium radar backscatter intensity (Dostálová et al., 2018; Nikaein et al., 2021; Borlaf-Mena et al., 2021) and an altitude below the tree line translated to an elevation below a thresholded elevation level (Hagedorn et al., 2006).

The two process definitions showcase how different modeling assumptions regarding the phenomenon of interest can be integrated. Both process definitions are based on the annual proportion of vegetation counts as given by the S-2 SCM, which is taken as a simplified proxy for forest vitality in a given year. Starting with the vegetation proportion statistics representative of the undisturbed forest state in the first year (Fig. 7, *status_original*), a comparison with the vegetation proportion statistics for the following years (Fig. 7, *status_post*) is carried out. Vegetation proportion decreases beyond a certain threshold are counted as relevant changes, which are then used to define disturbance magnitude and persistence as two exemplary custom properties of interest. The two process definitions differ in their threshold values as well as reducer functions used to calculate magnitude and persistence. The reducer function for persistence, for example, is either counting every year in which the threshold has been exceeded (Fig. 7, *sensitive model*), or only the amount of consecutive years with vegetation decreases larger than the threshold (Fig. 7, *robust model*).

The results of applying the models to a forested area close to Irrsee, Austria, analyzing forest disturbances in a four-year period are shown in Fig. 8. For the forest extent delineation, it is evident that both entity definitions correctly identify the central large forest areas as such. However, the ESA Worldcover definition additionally includes many smaller areas as fine-grained forest patches, which roots back to the fact that the class definition used actually identifies trees on a pixel-basis rather than larger spatially contiguous forest patches. The knowledge-based definition has a stronger tendency to spatial generalization given its foundation of Sentinel-1 coherence and backscatter data with coarser spatial resolution. The comparison of the model results thus reflects the different modeling assumptions made, which could now be used in an iterative process to refine the models. Given the semantic and visualizable design of the models, domain experts can be easily integrated in this discussion and adjustments of models. The ease to create and compare multiple models based on different data sources and methods offers further potential to estimate the degree of uncertainties in the final results, while leveraging the unified suite of models as an ensemble model with increased robustness. This not only applies to the entity definitions but equally to the process definitions with the results of forest disturbance magnitude and persistence.

It is worth emphasizing that the focus of this application example is not to create an optimal model achieving state-of-the art results in competition with other, more elaborated study designs. Intentionally, a conceptually rather simple model design was chosen to focus on highlighting the conceptual advantages brought by our data cube approach in facilitating semantic, knowledge-based analysis in line with the goals defined in Table 2. In terms of accuracy, the only aim is to showcase that our models lead to reasonably realistic results. This is given for both, forest extent delineation and disturbance assessments, as can be confirmed visually by comparing the modeling results in Fig. 8 to the true-color RGB timeseries visualizations. Moreover, the results are largely in line with data obtained from the European Forest Disturbance Atlas (Viana-Soto and Senf, 2025), providing further evidence for the basic reasonability of our model. The model can be easily refined and adapted to a specific target application for flexible use anywhere in the world.

6. Limitations & outlook

In terms of software implementation, further improvements concerning aspects of scalability are envisioned. Currently, features of parallelization and scheduling are performed on a chunk level with uniform chunk dimensionality as described in Section 3.2. This rigid tiling scheme with parallelization achieved on a high-level works but can be improved to achieve higher efficiency. Data of neighboring chunks are likely to be loaded multiple times as the tiling schema is not optimized with regard to the spatiotemporal extents of the native files. Caching, currently realized only within chunks but not across chunks could offer potential to mitigate redundancy in querying data. Still, sequentialization or parallelization of processing on a chunk level, where the execution of full models for a given chunk is considered the smallest unit, remains sub-optimal for at least two reasons. Firstly, this approach is inherently limited when a model involves both spatial and temporal operations that make model-wide chunking along the space or time axis impossible. Secondly, parallelization at the lower level of the individual functions within a model enables a better, more even utilization of the processing resources. To this end, we consider the integration of the Dask framework (Dask Development Team, 2016) as a promising way to improve the current implementation by relying on an established standard for the efficient parallelization of array-based processing. Prioritizing user-friendliness, we currently do not use Dask in our framework. The flexible creation of complex models requires a thorough understanding of task graph optimization on the user's side in order to exploit potential efficiency benefits of the full task scheduling provided by Dask.

A second point on extending software functionality refers to stronger support for multimodal data fusion. As of today, *gsemantique* offers predefined data set connections and means to integrate different data sets during the modeling process. The currently prevailing way of integrating multimodal data sets is to map different sensory information to properties of entities as demonstrated in Fig. 7. This data integration approach is primarily relevant for data sets derived from a heterogeneous set of sensors acquiring different information

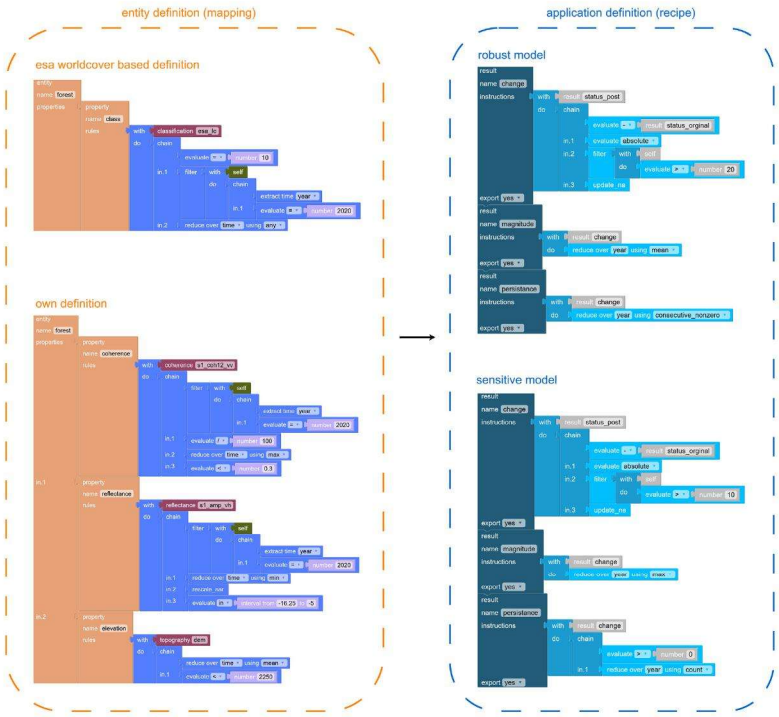


Fig. 7. Semantic models used for the assessment of forest disturbance.

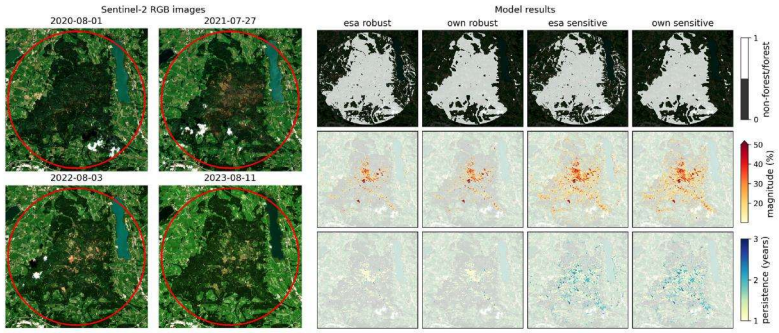


Fig. 8. Timeseries of true-color RGB visualizations of the area of interest (left) and forest disturbance maps (right). For the different underlying entity and process definitions, refer to Fig. 7.

about the entities of interest. For more homogeneous data sets, such as S-2 and Landsat with a large degree of overlapping bands, additional support for unifying data to model the same entity property would be beneficial. Corresponding automated routines for data harmonization are, for example, provided by Frantz (2019).

As a final point regarding software improvements, we would like to foster the better integration of data-driven and knowledge-based workflows. Demonstrated in Section 5.2, *gsemantique* allows to leverage data-driven modeling by integrating data layers as input, which in turn have been produced using machine learning. Furthermore, one can leverage user-defined functions to integrate machine learning or other data-driven elements during the model definition. Therefore, while our framework provides strong support for knowledge-based analysis, it does not exclude data-driven modeling. Still, the explicit integration of both domains could be improved in a way that machine and deep learning practitioners could seamlessly integrate data cubes prepared in a knowledge-based manner into their workflows and vice versa. Corresponding developments could be seen as complementary to the directions already indicated elsewhere with regard to improving data-driven modeling in EO cubes (Montero et al., 2024b).

Looking ahead on the more application-related future works, there are several interesting directions in the context of which our new framework could be explored. One of these concerns the question of analyzing synergies between multimodal data usage and semantic, knowledge-based modeling. With mono-modal data, only a narrow set of the properties of physical world entities can be modeled, which is different if multimodal data is available. Suggested by Bahmanyar et al. (2015), increasing the diversity of data sets in analyses could help to close the semantic gap as the difference between the users and the computers view on a given entity. Using *gsemantique*, one can model multiple different perspectives on the same entity, drawing on a variety of data sets and their possible combinations, in order to explore corresponding hypotheses in greater detail. Linked to this, ideas for further research designs include the analysis of uncertainty in modeling. In Section 5.2, we have indicated the potential for such analyses using multiple definitions of the forest entity and the phenomenon of forest disturbance. The possibility of integrating one's own data into the analysis in *gsemantique* allows for far-reaching comparisons to be made as to how far modeling results depend on the data basis and the combination of data in a model.

7. Conclusion

This work was motivated by the identified need for new frameworks that mitigate the discrepancy between extensive data availability and restricted possibilities of analyzing them to achieve big EO image understanding. Reviewing the current state of the art, we noted that the variety of existing frameworks do not cover the requirements for ad-hoc, mesoscale analyses, which are commonly conducted in practice. Whereas such analyses exceed standard resources in terms of local hardware limitations for data processing, and require the shift to modern data cube- and cloud-based processing paradigms, they are not justifying the effort to set up complex persistent infrastructures. Additionally, approaches supporting the actual analysis of data, i.e. the modeling process that integrates different data sets to move from the non-semantic data level to condensed information, are integrated in existing data cube approaches insufficiently. The prevailing focus so far has been on technical means of data access, possibly extended by data-driven means of modeling. As a consequence of these deficiencies, we extended an existing semantic querying language towards an on-demand EO data cube system with end-to-end support for the whole EO analysis workflow. We outlined the implementation of the proposed system focusing on its main functionalities, which are pre-configured, extensible access to a variety of EO data sets with global coverage, strong analysis support through a framework for semantic, knowledge-based image understanding, and processing support to scale analyses in space and time. Those properties have been demonstrated by two application examples, both designed in a conceptually simple yet effective way to derive useful information from EO data in a few lines of code. Our work thus makes a valuable contribution to foster the structured transformation of data into condensed information with the aims of enabling analytical insights, supporting decision-making and generating further EO knowledge.

CRediT authorship contribution statement

Felix Kröber: Writing – original draft, Visualization, Software, Formal analysis, Conceptualization. **Martin Sudmanns:** Writing – review & editing, Project administration, Funding acquisition, Conceptualization. **Lorena Abad:** Writing – review & editing, Software. **Dirk Tiede:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition.

Funding

The research leading to these results has received funding from the European Union's Horizon Europe research and innovation program under the Grant Agreement No. 101082493 (Project: LEONSEGS).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Luuk van der Meer for his support in integrating the adaptations of the semantique framework. Furthermore, the authors gratefully acknowledge the reviewers for their comments which have contributed to the improvement of this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.isprsjprs.2025.07.015.

Data/code availability statement

- name of the software: gsemantique
- hard-/software requirements: any OS (Windows, Mac or Linux); Python installation
- source code availability: https://github.com/Sen2Cube-at/gsemantique
- analysis data availability: https://doi.org/10.5281/zenodo.15423258

References

Ackoff, R.L., 1989. From data to wisdom. J. Appl. Syst. Anal. 16 (1), 3–9.
Appel, M., Pebesma, E., 2019. On-Demand Processing of Data Cubes from Satellite Image Collections with the gdalcubes Library. Data 4 (3), 92. http://dx.doi.org/10.3390/data4030092.
Ariza-Porras, C., Bravo, G., Villamizar, M., Moreno, A., Castro, H., Galindo, G., Cabera, E., Valbuena, S., Lozano, P., 2017. CDCol: A geoscience data cube that meets colombian needs. In: Advances in Computing. CCC 2017. Communications in Computer and Information Science. 735, Springer, pp. 87–99. http://dx.doi.org/10.1007/978-3-319-66562-7_7.
Arvor, D., Belgiu, M., Falomir, Z., Mougenot, I., Durieux, L., 2019. Ontologies to interpret remote sensing images: why do we need them? GIScience Remote. Sens. 56 (6), 911–939. http://dx.doi.org/10.1080/15481603.2019.1587890.
Arvor, D., Betbeder, J., Daher, F.R., Blossier, T., Le Roux, R., Corgne, S., Corpetti, T., De Freitas Silgueiro, V., Silva Junior, C.A.D., 2021. Towards user-adaptive remote sensing: Knowledge-driven automatic classification of Sentinel-2 time series. Remote Sens. Environ. 264, 112615. http://dx.doi.org/10.1016/j.rse.2021.112615.

Asmaryan, S., Muradyan, V., Tepanosyan, G., Hovsepyan, A., Saghatelyan, A., Astsatryan, H., Grigoryan, H., Abrahamyan, R., Guigoz, Y., Giuliani, G., 2019. Paving the Way towards an Armenian Data Cube. Data 4 (3), 117. http://dx.doi.org/10.3390/data4030117.

Augustin, H., Sudmanns, M., Tiede, D., Lang, S., Baraldi, A., 2019. Semantic Earth Observation Data Cubes. Data 4 (3), 102. http://dx.doi.org/10.3390/data4030102.

Bahmanyar, R., Murillo Montes De Oca, A., Datcu, M., 2015. The Semantic Gap: An Exploration of User and Computer Perspectives in Earth Observation Images. IEEE Geosci. Remote. Sens. Lett. 12 (10), 2046–2050. http://dx.doi.org/10.1109/LGRS.2015.2444666.

Baltsavias, E., 2004. Object extraction and revision by image analysis using existing geodata and knowledge: current status and steps towards operational systems. ISPRS J. Photogramm. Remote Sens. 58 (3–4), 129–151. http://dx.doi.org/10.1016/j.isprsjprs.2003.09.002.

Baraldi, A., 2011. Satellite Image Automatic Mapper - a Turnkey Software Executable for Automatic Near Real-Time Multi-Sensor Multi-Resolution Spectral Rule-Based Preliminary Classification of Spaceborne Multi-Spectral Images. Recent. Patents Space Technol. 1 (2), 81–106. http://dx.doi.org/10.2174/1877611611101020081.

Baraldi, A., Boschetti, L., 2012. Operational Automatic Remote Sensing Image Understanding Systems: Beyond Geographic Object-Based and Object-Oriented Image Analysis (GEOBIA/GEOOIA). Part 1: Introduction. Remote. Sens. 4 (9), 2694–2735. http://dx.doi.org/10.3390/rs4092694.

Baraldi, A., Sapia, L.D., Tiede, D., Sudmanns, M., Augustin, H.L., Lang, S., 2023. Innovative Analysis Ready Data (ARD) product and process requirements, software system design, algorithms and implementation at the midstream as necessary-but-not-sufficient precondition of the downstream in a new notion of Space Economy 4.0 - Part 1: Problem background in Artificial General Intelligence (AGI). Big Earth Data 7 (3), 455–693. http://dx.doi.org/10.1080/20964471.2021.2017549.

Baumann, P., 2017. The Datacube Manifesto. URL: https://earthserver.eu/tech/datacube-manifesto/The-Datacube-Manifesto.pdf.

Baumann, P., Mazzetti, P., Ungar, J., Barbera, R., Barboni, D., Beccati, A., Bigagli, L., Boldrini, E., Bruno, R., Calanducci, A., Campalani, P., Clements, O., Dumitru, A., Grant, M., Herzig, P., Kakaletris, G., Laxton, J., Koltsida, P., Lipskoch, K., Mahdiraji, A.R., Mantovani, S., Merticariu, V., Messina, A., Misev, D., Natali, S., Nativi, S., Oosthoek, J., Pappalardo, M., Passmore, J., Rossi, A.P., Rundo, F., Sen, M., Sorbera, V., Sullivan, D., Torrisi, M., Trovato, L., Veratelli, M.G., Wagner, S., 2016. Big Data Analytics for Earth Sciences: the EarthServer approach. Int. J. Digit. Earth 9 (1), 3–29. http://dx.doi.org/10.1080/17538947.2014.1003106.

Baumann, P., Misev, D., Merticariu, V., Huu, B.P., Bell, B., 2018. Rasdaman: Spatio-temporal datacubes on steroids. In: Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, Seattle Washington, pp. 604–607. http://dx.doi.org/10.1145/3274895.3274988.

Belgiu, M., Drăguţ, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31. http://dx.doi.org/10.1016/j.isprsjprs.2016.01.011.

Berners-Lee, T., Hendler, J., Lassila, O., 2001. The Semantic Web. Sci. Am. 34–43.

Borlaf-Mena, I., Badea, O., Tanase, M.A., 2021. Assessing the Utility of Sentinel-1 Coherence Time Series for Temperate and Tropical Forest Mapping. Remote. Sens. 13 (23), 4814. http://dx.doi.org/10.3390/rs13234814.

Brockmann Consult GmbH, 2021. Xcube - An xarray-based EO data cube toolkit — xcube 1.8.0.dev0 documentation. URL: https://xcube.readthedocs.io/en/latest/.

Chatenoux, B., Richard, J.-P., Small, D., Roeoesli, C., Wingate, V., Poussin, C., Rodila, D., Peduzzi, P., Steinmeier, C., Ginzler, C., Psomas, A., Schaepman, M.E., Giuliani, G., 2021. The Swiss data cube, analysis ready data archive using earth observations of Switzerland. Sci. Data 8 (1), 295. http://dx.doi.org/10.1038/s41597-021-01076-6.

Cipoletta, S.R., Sciarra, R., 2024. copernicus Sentinel Data Access Annual Report 2023. URL: https://sentiwiki.copernicus.eu/_attachments/1673407/COPE-SRCO-RP-240021%20-%20Sentinel%20Data%20Access%20Annual%20Report%202023%20-%201.1.pdf?inst-v=86d5ab7c-f08e-4690-a070-6d2a33e3cade.

Craglia, M., Nativi, S., 2018. Mind the Gap: Big Data vs. interoperability and reproducibility of science. Earth Obs. Open Sci. Innov. 121–141.

Crevier, D., Lepage, R., 1997. Knowledge-Based Image Understanding Systems: A Survey. Comput. Vis. Image Underst. 67 (2), 161–185. http://dx.doi.org/10.1006/cviu.1996.0520.

Dask Development Team, 2016. Dask: Library for dynamic task scheduling. URL: http://dask.pydata.org.

Dhu, T., Giuliani, G., Juárez, J., Kavvada, A., Killough, B., Merodio, P., Minchin, S., Ramage, S., 2019. National Open Data Cubes and Their Contribution to Country-Level Development Policies and Practices. Data 4 (4), 144. http://dx.doi.org/10.3390/data4040144.

Di Gregorio, A., Henry, M., Donegan, E., Finegold, Y., Latham, J., Jonckheere, I., Cumani, R., 2016. Land Cover Classification System: Advanced Database Gateway. FAO, Rome, Italy.

Dostálová, A., Wagner, W., Milenković, M., Hollaus, M., 2018. Annual seasonality in Sentinel-1 signal for forest mapping and forest type classification. Int. J. Remote Sens. 39 (21), 7738–7760. http://dx.doi.org/10.1080/01431161.2018.1479788.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P., 2012. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. Remote Sens. Environ. 120, 25–36. http://dx.doi.org/10.1016/j.rse.2011.11.026.

Euro Data Cube Consortium, Euro Data Cube. URL: https://eurodatacube.com/.

Ferreira, K.R., Queiroz, G.R., Vinhas, L., Marujo, R.F.B., Simoes, R.E.O., Picoli, M.C.A., Camara, G., Cartaxo, R., Gomes, V.C.F., Santos, L.A., Sanchez, A.H., Arcanjo, J.S., Fronza, J.G., Noronha, C.A., Costa, R.W., Zaglia, M.C., Zioti, F., Korting, T.S., Soares, A.R., Chaves, M.E.D., Fonseca, L.M.G., 2020. Earth Observation Data Cubes for Brazil: Requirements, Methodology and Products. Remote. Sens. 12 (24), http://dx.doi.org/10.3390/rs12244033.

Frantz, D., 2019. FORCE—Landsat + Sentinel-2 Analysis Ready Data and Beyond. Remote. Sens. 11 (9), 1124. http://dx.doi.org/10.3390/rs11091124.

Frolking, S., Palace, M.W., Clark, D.B., Chambers, J.Q., Shugart, H.H., Hurtt, G.C., 2009. Forest disturbance and recovery: A general review in the context of spaceborne remote sensing of impacts on aboveground biomass and canopy structure. J. Geophys. Res.: Biogeosciences 114 (G2), http://dx.doi.org/10.1029/2008JG000911.

Gao, Y., Skutsch, M., Paneque-Gálvez, J., Ghilardi, A., 2020. Remote sensing of forest degradation: a review. Environ. Res. Lett. 15 (10), 103001. http://dx.doi.org/10.1088/1748-9326/abaad7.

Giuliani, G., Chatenoux, B., De Bono, A., Rodila, D., Richard, J.-P., Allenbach, K., Dao, H., Peduzzi, P., 2017. Building an Earth Observations Data Cube: lessons learned from the Swiss Data Cube (SDC) on generating Analysis Ready Data (ARD). Big Earth Data 1 (1–2), 100–117. http://dx.doi.org/10.1080/20964471.2017.1398903.

Giuliani, G., Chatenoux, B., Piller, T., Moser, F., Lacroix, P., 2020. Data Cube on Demand (DCoD): Generating an earth observation Data Cube anywhere in the world. Int. J. Appl. Earth Obs. Geoinf. 87, 102035. http://dx.doi.org/10.1016/j.jag.2019.102035.

Giuliani, G., Masó, J., Mazzetti, P., Nativi, S., Zabala, A., 2019. Paving the Way to Increased Interoperability of Earth Observations Data Cubes. Data 4 (3), 113. http://dx.doi.org/10.3390/data4030113.

Gomes, V., Queiroz, G., Ferreira, K., 2020. An Overview of Platforms for Big Earth Observation Data Management and Analysis. Remote. Sens. 12 (8), 1253. http://dx.doi.org/10.3390/rs12081253.

Goodenough, D., Goldberg, M., Plunkett, G., Zelek, J., 1987. An Expert System for Remote Sensing. IEEE Trans. Geosci. Remote Sens. GE-25 (3), 349–359. http://dx.doi.org/10.1109/TGRS.1987.289805.

Google, 2024. Google blockly - The web-based visual programming editor. Google, URL: https://github.com/google/blockly.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27. http://dx.doi.org/10.1016/j.rse.2017.06.031.

Guo, H., Liu, Z., Jiang, H., Wang, C., Liu, J., Liang, D., 2017. Big Earth Data: a new challenge and opportunity for Digital Earth's development. Int. J. Digit. Earth 10 (1), 1–12. http://dx.doi.org/10.1080/17538947.2016.1264490.

Hagedorn, F., Rigling, A., Bebi, P., 2006. Wo Bäume nicht mehr wachsen können: Die Waldgrenze. Die Alp. 9, 52–55.

Hirschmugl, M., Gallaun, H., Dees, M., Datta, P., Deutscher, J., Koutsias, N., Schardt, M., 2017. Methods for Mapping Forest Disturbance and Degradation from Optical Earth Observation Data: a Review. Curr. For. Rep. 3 (1), 32–45. http://dx.doi.org/10.1007/s40725-017-0047-2.

Hoeser, T., Bachofer, F., Kuenzer, C., 2020. Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review - Part II: Applications. Remote. Sens. 12 (18), 3053. http://dx.doi.org/10.3390/rs12183053.

Hoeser, T., Kuenzer, C., 2020. Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review - Part I: Evolution and Recent Trends. Remote. Sens. 12 (10), 1667. http://dx.doi.org/10.3390/rs12101667.

Hoyer, S., Hamman, J., 2017. Xarray: N-d labeled Arrays and Datasets in Python. J. Open Res. Softw. 5 (1), 10. http://dx.doi.org/10.5334/jors.148.

Jacob, A.W., Vicente-Guijalba, F., Lopez-Martinez, C., Lopez-Sanchez, J.M., Litzinger, M., Kristen, H., Mestre-Quereda, A., Ziolkowski, D., Lavalle, M., Notarnicola, C., Suresh, G., Antropov, O., Ge, S., Praks, J., Ban, Y., Pottier, E., Mallorqui Franquet, J.J., Duro, J., Engdahl, M.E., 2020. Sentinel-1 InSAR Coherence for Land Cover Mapping: A Comparison of Multiple Feature-Based Classifiers. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 13, 535–552. http://dx.doi.org/10.1109/JSTARS.2019.2958847.

Kempeneers, P., Soille, P., 2017. Optimizing Sentinel-2 image selection in a Big Data context. Big Earth Data 1 (1–2), 145–158. http://dx.doi.org/10.1080/20964471.2017.1407489.

Killough, B., 2018. Overview of the Open Data Cube Initiative. In: IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE, Valencia, pp. 8629–8632. http://dx.doi.org/10.1109/IGARSS.2018.8517694.

Killough, B., Siqueira, A., Dyke, G., 2020. Advancements in the Open Data Cube and Analysis Ready Data — Past, Present and Future. In: IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium. IEEE, Waikoloa, HI, USA, pp. 3373–3375. http://dx.doi.org/10.1109/IGARSS39084.2020.9324712.

Laurini, R., Thompson, D., 1992. Fundamentals of spatial information systems, vol. 37, Academic Press.

Lausch, A., Erasmi, S., King, D., Magdon, P., Heurich, M., 2016. Understanding Forest Health with Remote Sensing - Part I - a Review of Spectral Traits, Processes and Remote-Sensing Characteristics. Remote. Sens. 8 (12), 1029. http://dx.doi.org/10.3390/rs8121029.

Lausch, A., Erasmi, S., King, D., Magdon, P., Heurich, M., 2017. Understanding Forest Health with Remote Sensing - Part II - a Review of Approaches and Data Models. Remote. Sens. 9 (2), 129. http://dx.doi.org/10.3390/rs9020129.

Lewis, A., Lymburner, L., Purss, M.B.J., Brooke, B., Evans, B., Ip, A., Dekker, A.G., Irons, J.R., Minchin, S., Mueller, N., Oliver, S., Roberts, D., Ryan, B., Thankappan, M., Woodcock, R., Wyborn, L., 2016. Rapid, high-resolution detection of environmental change over continental scales from satellite data – the Earth Observation Data Cube. Int. J. Digit. Earth 9 (1), 106–111. http://dx.doi.org/10.1080/17538947.2015.1111952.

Lewis, A., Oliver, S., Lymburner, L., Evans, B., Wyborn, L., Mueller, N., Raevksi, G., Hooke, J., Woodcock, R., Sixsmith, J., Wu, W., Tan, P., Li, F., Killough, B., Minchin, S., Roberts, D., Ayers, D., Bala, B., Dwyer, J., Dekker, A., Dhu, T., Hicks, A., Ip, A., Purss, M., Richards, C., Sagar, S., Trenham, C., Wang, P., Wang, L.-W., 2017. The Australian Geoscience Data Cube — Foundations and lessons learned. Remote Sens. Environ. 202, 276–292. http://dx.doi.org/10.1016/j.rse.2017.03.015.

Mahecha, M.D., Gans, F., Brandt, G., Christiansen, R., Cornell, S.E., Fomferra, N., Kraemer, G., Peters, J., Bodesheim, P., Camps-Valls, G., Donges, J.F., Dorigo, W., Estupinan-Suarez, L.M., Gutierrez-Velez, V.H., Gutwin, M., Jung, M., Londoño, M.C., Miralles, D.G., Papastefanou, P., Reichstein, M., 2020. Earth system data cubes unravel global multivariate dynamics. Earth Syst. Dyn. 11 (1), 201–234. http://dx.doi.org/10.5194/esd-11-201-2020.

Maso, J., Zabala, A., Serral, I., Pons, X., 2019. A Portal Offering Standard Visualization and Analysis on top of an Open Data Cube for Sub-National Regions: The Catalan Data Cube Example. Data 4 (3), 96. http://dx.doi.org/10.3390/data4030096.

Matsuyama, T., 1993. Expert Systems for Image Processing, Analysis, and Recognition: Declarative Knowledge Representation for Computer Vision. In: Advances in Electronics and Electron Physics. 86, Elsevier, pp. 81–171. http://dx.doi.org/10.1016/S0065-2539(08)60154-7.

Matsuyama, T., Hwang, V.S.-S., 1990. SIGMA: A Knowledge-Based Aerial Image Understanding System. Plenum Press, New York, NY, USA; London, UK.

Mazzocchi, F., 2015. Could Big Data be the end of theory in science?: A few remarks on the epistemology of data-driven science. EMBO Rep. 16 (10), 1250–1255. http://dx.doi.org/10.15252/embr.201541001.

Microsoft Open Source, Emanuele, R., Morris, D., Augspurger, T., McFarland, Matt, 2022. Microsoft/PlanetaryComputer: October 2022. http://dx.doi.org/10.5281/zenodo.7261897.

Montero, D., Aybar, C., Ji, C., Kraemer, G., Söchting, M., Teber, K., Mahecha, M.D., 2024a. On-Demand Earth System Data Cubes. In: IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium. pp. 7529–7532. http://dx.doi.org/10.1109/IGARSS53475.2024.10640742.

Montero, D., Kraemer, G., Anghelea, A., Aybar, C., Brandt, G., Camps-Valls, G., Cremer, F., Flik, I., Gans, F., Habershon, S., Ji, C., Kattenborn, T., Martínez-Ferrer, L., Martinuzzi, F., Reinhardt, M., Söchting, M., Teber, K., Mahecha, M.D., 2024b. Earth System Data Cubes: Avenues for advancing Earth system research. http://dx.doi.org/10.48550/arXiv.2408.02348.

Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: A review. ISPRS J. Photogramm. Remote Sens. 66 (3), 247–259. http://dx.doi.org/10.1016/j.isprsjprs.2010.11.001.

Neteler, M., Bowman, M.H., Landa, M., Metz, M., 2012. GRASS GIS: A multi-purpose open source GIS. Environ. Model. Softw. 31, 124–130. http://dx.doi.org/10.1016/j.envsoft.2011.11.014.

Nikaein, T., Iannini, L., Molijn, R.A., Lopez-Dekker, P., 2021. On the Value of Sentinel-1 InSAR Coherence Time-Series for Vegetation Classification. Remote. Sens. 13 (16), 3300. http://dx.doi.org/10.3390/rs13163300.

Paulino, E.R., Schlerf, M., Röder, A., Stoffels, J., Udelhoven, T., 2024. Forest disturbance characterization in the era of earth observation big data: A mapping review. Int. J. Appl. Earth Obs. Geoinf. 128, 103755. http://dx.doi.org/10.1016/j.jag.2024.103755.

Pebesma, E., Bivand, R., 2023. Spatial Data Science: With applications in R. Chapman and Hall/CRC, London, http://dx.doi.org/10.1201/9780429459016.

Peters, J., Neumann, A., Jaeger, M., Gienapp, L., Umlauft, J., 2025. Ml4xcube: Machine learning toolkits for earth system data cubes. In: Proceedings of the AAAI conference on artificial intelligence. 39, pp. 28302–28311. http://dx.doi.org/10.1609/aaai.v39i27.35051.

Quang, N.H., Tuan, V.A., Hao, N.T.P., Hang, L.T.T., Hung, N.M., Anh, V.L., Phuong, L.T.M., Carrie, R., 2019. Synthetic aperture radar and optical remote sensing image fusion for flood monitoring in the Vietnam lower Mekong basin: a prototype application for the Vietnam Open Data Cube. Eur. J. Remote. Sens. 52 (1), 599–612. http://dx.doi.org/10.1080/22797254.2019.1698319.

Rowley, J., 2007. The wisdom hierarchy: representations of the DIKW hierarchy. J. Inf. Sci. 33 (2), 163–180. http://dx.doi.org/10.1177/0165551506070706.

Scheider, S., Ostermann, F.O., Adams, B., 2017. Why good data analysts need to be critical synthesists. Determining the role of semantics in data analysis. Future Gener. Comput. Syst. 72, 11–22. http://dx.doi.org/10.1016/j.future.2017.02.046.

Simoes, R., Camara, G., Queiroz, G., Souza, F., Andrade, P.R., Santos, L., Carvalho, A., Ferreira, K., 2021a. Satellite Image Time Series Analysis for Big Earth Observation Data. Remote. Sens. 13 (13), 2428. http://dx.doi.org/10.3390/rs13132428.

Simoes, R., Souza, F., Zaglia, M., Queiroz, G.R., Santos, R., Ferreira, K., 2021b. Rstac: An R Package to Access Spatiotemporal Asset Catalog Satellite Imagery. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. pp. 7674–7677. http://dx.doi.org/10.1109/IGARSS47720.2021.9553518.

Storch, T., Reck, C., Holzwarth, S., Wiegers, B., Mandery, N., Raape, U., Strobl, C., Volkmann, R., Böttcher, M., Hirner, A., Senft, J., Plesia, N., Kukuk, T., Meissl, S., Felske, J.-R., Heege, T., Keuck, V., Schmidt, M., Staudenrausch, H., 2019. Insights into CODE-DE – Germany's Copernicus data and exploitation platform. Big Earth Data 3 (4), 338–361. http://dx.doi.org/10.1080/20964471.2019.1692297.

Strobl, P., Baumann, P., Lewis, A., Szantoi, Z., Killough, B., Purss, M., Craglia, M., Nativi, S., Held, A., Dhu, T., 2017. The six faces of the data cube. In: Big Data from Space. European Commission Joint Research Centre, Toulouse, France, pp. 32–35. http://dx.doi.org/10.2760/383579, ISSN: 1831-9424.

Sudmanns, M., Augustin, H., Killough, B., Giuliani, G., Tiede, D., Leith, A., Yuan, F., Lewis, A., 2023. Think global, cube local: an Earth Observation Data Cube's contribution to the Digital Earth vision. Big Earth Data 7 (3), 831–859. http://dx.doi.org/10.1080/20964471.2022.2099236.

Sudmanns, M., Augustin, H., Van Der Meer, L., Baraldi, A., Tiede, D., 2021. The Austrian Semantic EO Data Cube Infrastructure. Remote. Sens. 13 (23), 4807. http://dx.doi.org/10.3390/rs13234807.

Sudmanns, M., Tiede, D., Augustin, H., Lang, S., 2020a. Assessing global Sentinel-2 coverage dynamics and data availability for operational Earth observation (EO) applications using the EO-Compass. Int. J. Digit. Earth 13 (7), 768–784. http://dx.doi.org/10.1080/17538947.2019.1572799.

Sudmanns, M., Tiede, D., Lang, S., Bergstedt, H., Trost, G., Augustin, H., Baraldi, A., Blaschke, T., 2020b. Big Earth data: disruptive changes in Earth observation data management and analysis? Int. J. Digit. Earth 13 (7), 832–850. http://dx.doi.org/10.1080/17538947.2019.1585976.

Van Der Meer, L., Sudmanns, M., Augustin, H., Baraldi, A., Tiede, D., 2022. Semantic Querying in Earth Observation Data Cubes. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. XLVIII-4/W1-2022, 503–510. http://dx.doi.org/10.5194/isprs-archives-XLVIII-4-W1-2022-503-2022.

Viana-Soto, A., Senf, C., 2025. The European Forest Disturbance Atlas: a forest disturbance monitoring system using the Landsat archive. Earth Syst. Sci. Data 17 (6), 2373–2404. http://dx.doi.org/10.5194/essd-17-2373-2025.

Wagemann, J., Siemen, S., Seeger, B., Bendix, J., 2021. Users of open Big Earth data – An analysis of the current state. Comput. Geosci. 157, 104916. http://dx.doi.org/10.1016/j.cageo.2021.104916.

Wilson, A.M., Jetz, W., 2016. Remotely Sensed High-Resolution Global Cloud Dynamics for Predicting Ecosystem and Biodiversity Distributions. In: Loreau, M. (Ed.), PLOS Biology 14 (3), e1002415. http://dx.doi.org/10.1371/journal.pbio.1002415.

Woodcock, C.E., Allen, R., Anderson, M., Belward, A., Bindschadler, R., Cohen, W., Gao, F., Goward, S.N., Helder, D., Helmer, E., Nemani, R., Oreopoulos, L., Schott, J., Thenkabail, P.S., Vermote, E.F., Vogelmann, J., Wulder, M.A., Wynne, R., 2008. Free Access to Landsat Imagery. Science 320 (5879), http://dx.doi.org/10.1126/science.320.5879.1011a.

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 Sens. Environ. 122, 2–10. http://dx.doi.org/10.1016/j.rse.2012.01.010.

Yuan, F., Lewis, A., Leith, A., Dhar, T., Gavin, D., 2021. Analysis Ready Data for Africa. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. IEEE, Brussels, Belgium, pp. 1789–1791. http://dx.doi.org/10.1109/IGARSS47720.2021.9554019.

Zanaga, D., Van De Kerchove, R., De Keersmaecker, W., Souverijns, N., Brockmann, C., Quast, R., Wevers, J., Grosu, A., Paccini, A., Vergnaud, S., Cartus, O., Santoro, M., Fritz, S., Georgieva, I., Lesiv, M., Carter, S., Herold, M., Li, L., Tsendbazar, N.E., Ramoino, F., Arino, O., 2021. ESA WorldCover 10 m 2020 v100. http://dx.doi.org/10.5281/zenodo.5571936.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. IEEE Geosci. Remote. Sens. Mag. 5 (4), 8–36. http://dx.doi.org/10.1109/MGRS.2017.2762307.

Zhu, Z., Wulder, M.A., Roy, D.P., Woodcock, C.E., Hansen, M.C., Radeloff, V.C., Healey, S.P., Schaaf, C., Hostert, P., Strobl, P., Pekel, J.F., Lymburner, L., Pahlevan, N., Scambos, T.A., 2019. Benefits of the free and open Landsat data policy. Remote Sens. Environ. 224, 382–385. http://dx.doi.org/10.1016/j.rse.2019.02.016.

2.6. SCIENTIFIC ABSTRACT VI

On-demand data cubes – knowledge-based, semantic querying of multimodal Earth observation data for mesoscale analyses anywhere on Earth

Authors: F. Kröber, M. Sudmanns, D. Tiede
Session: D.3.2 Free Open Source Software for the Geospatial Domain: current status & evolution
Type: Poster

Introduction

In recent years, Earth observation (EO) data access and processing have undergone a transformative shift, driven by the advent of novel big EO data paradigms [1,2]. With the increasing volume and variety of EO data, the significance of cloud-enabled processing frameworks that allow users to focus on the actual analysis of data with an abstraction of technical complexities is growing. However, many cloud-based platforms are proprietary or closed-source [3,4], imposing costs and service uncertainties, as illustrated by the unexpected shutdown of the Microsoft's Planetary Computer Hub in June 2024. However, open-source, free alternatives like the Open Data Cube [5] may require significant setup effort, with dataset indexing being one reason. This effort is only justified for larger data cubes with long-term infrastructure goals, whereas for shorter-term projects the practicality is restricted. Moreover, while current systems offer technical solutions in terms of data access and scalability of analyses, many approaches are still lacking in image-understanding capabilities. A modern processing framework needs to provide adequate means to address the semantic complexity of EO data [6]. Analysts still grapple with raw data structures, rather than having frameworks at hand to focus on data meaning. In brief, there is a pressing need for open-source EO data processing frameworks that are both user-friendly and capable of representing the semantics of EO data. To this end, we introduce a novel Python package (*gsemantique*) for building ad hoc data cubes for semantic EO analyses. We demonstrate its utility for querying multi-modal data by focussing on the use case of forest disturbance modelling.

Design choices & Technical implementation

The technical foundations for the *gsemantique* package are threefold:

First, data in cloud-optimised formats is fetched on-demand to regularised three-dimensional data cubes. The SpatioTemporal Asset Catalog (STAC) [7], fostering standardisation in the structuring of geospatial metadata, is leveraged to facilitate data access. A pre-defined suite of STAC endpoints including several common EO datasets such as Landsat, Sentinel-1 and Sentinel-2 along with additional datasets such as a global DEM is part of the package. The way data access is modelled easily allows to extend the set of pre-defined datasets for custom ones.

Second, the creation of comprehensible, knowledge-based, transparent models is supported by providing a semantic querying language to address and model the data. Here, we build on the foundation laid by the semantique package [8], which introduced a structured approach to semantic querying of EO data. This can be used to supplement conventional, non-semantic approaches. To facilitate effective exchanges with end users and domain experts regarding the design of analyses, graphical visualisation options for models are integrated. Specifically, the model coded in a python structure can be represented using graphical blocks as defined by Google's Blockly library [9].

Third, the scalable execution of the models is enabled by an internal tiling of the queried spatio-temporal extent into smaller chunks. The complexity of the chunking mechanism with the decision on the dimension (i.e. chunk-by-space or chunk-by-time), the execution of the recipe and the merging of the individual chunks into a single result is abstracted from the user. Focussing on efficiency, the chunked execution of the model supports multiprocessing.

As data dependencies are not fixed or can be replaced and extended, the presented python package offers a very flexible and portable way of performing data analyses. Big EO data archives can be analysed both on local, consumer-grade devices, and on cloud-based, high-performance processing platforms without being tied to a specific platform.

Application case: Forest disturbance analyses

To prove the value of the proposed package, we focus on the use case of analysing forest disturbances via remote sensing data. The focus here is deliberately not on the optimisation of the model, i.e. to create the best performing forest disturbance model. Instead, we intentionally choose to address the example with a simple but still effective analysis model aiming at highlighting the conceptual advantages of our approach. Specifically, three beneficial properties of the processing framework are showcased.

First, the entity of interest (forest) is a 4D real-world phenomenon, that needs to be translated to features in the 2D image domain. This is indeed not unique to the forest entity but applies to all entities in the 4D world (including their relationships). However, in the image domain, entities such as water bodies are spectrally distinct relative to other objects. This makes the selection of useful image features a straightforward task, even without an explicit model that translates properties of the entity to features of the object. Forests, on the other hand, represent a type of vegetation, which is more challenging to distinguish in the image domain. Similar image features may be observed for other vegetated surfaces such as meadows or bogs. Here, the advantage of knowledge-based, semantic modelling with the possibility of an explicit definition of multiple relevant entity properties (and their translation into object features) becomes clear. We create such a model for the entity forest in an undisturbed state by defining the properties of temporal stability (translated to low radar coherence), vitality (translated to a positive NDVI) and altitude below the tree line (translated to an elevation below a thresholded DEM level). We compare this entity definition with a pre-defined one that was derived in a data-driven way. Both entity definitions can be generated without further effort leveraging the data connections pre-implemented in the package. The comparison of both definitions allows an estimation of the uncertainty when modelling the entity forest based on different data sets and approaches.

Second, the phenomenon of disturbance is an ambiguous concept. There is no unique, crisp definition of forest disturbances such that a remote sensing expert needs to make his/her specific assumptions in modelling the phenomenon explicit and transparent in order to discuss it further with other domain experts. Also, there is no simple data-driven way to solve the task of disturbance modelling since there is a lack of available label data. Hence, this example is well suited to be approached by a semantic, knowledge-based modelling approach that allows to visualise and communicate the resulting human-readable model to others.

Third, forest disturbances are inherently process-based, i.e. they are characterised by a temporal change in the forest's status. A datacube-based approach is therefore well positioned to approach this task, as it allows to query every single observation through time instead of relying on pre-processed, aggregated EO products. Using a multiannual use case design incorporating Sentinel-1, Sentinel-2 and DEM data for a spatial extent of more than 1000 km^2, we demonstrate the usability of our package for meso-scale analyses querying all available data references through time.

[1] H. Guo, Z. Liu, H. Jiang, C. Wang, J. Liu, and D. Liang, 'Big Earth Data: a new challenge and opportunity for Digital Earth's development', International Journal of Digital Earth, vol. 10, no. 1, pp. 1–12, Jan. 2017, doi: 10.1080/17538947.2016.1264490.

[2] M. Sudmanns et al., 'Big Earth data: disruptive changes in Earth observation data management and analysis?', International Journal of Digital Earth, vol. 13, no. 7, pp. 832–850, Jul. 2020, doi: 10.1080/17538947.2019.1585976.

[3] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, 'Google Earth Engine: Planetary-scale geospatial analysis for everyone', Remote Sensing of Environment, vol. 202, pp. 18–27, Dec. 2017, doi: 10.1016/j.rse.2017.06.031.

[4] Microsoft Open Source, R. Emanuele, D. Morris, T. Augspurger, and McFarland, Matt, microsoft/PlanetaryComputer: October 2022. (Oct. 2022). Zenodo. doi: 10.5281/zenodo.7261897.

[5] B. Killough, 'Overview of the Open Data Cube Initiative', in IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia: IEEE, Jul. 2018, pp. 8629–8632. doi: 10.1109/IGARSS.2018.8517694.

[6] H. Augustin, M. Sudmanns, D. Tiede, S. Lang, and A. Baraldi, 'Semantic Earth Observation Data Cubes', Data, vol. 4, no. 3, p. 102, Jul. 2019, doi: 10.3390/data4030102.

[7] 'STAC: SpatioTemporal Asset Catalogs'. Accessed: Nov. 24, 2024. [Online]. Available: https://stacspec.org/en/

[8] L. Van Der Meer, M. Sudmanns, H. Augustin, A. Baraldi, and D. Tiede, 'Semantic Querying in Earth Observation Data Cubes', Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., vol. XLVIII-4/W1-2022, pp. 503–510, Aug. 2022, doi: 10.5194/isprs-archives-XLVIII-4-W1-2022-503-2022.

[9] Google, Google blockly - The web-based visual programming editor. (Nov. 25, 2024). TypeScript. Google. Accessed: Nov. 25, 2024. [Online]. Available: https://github.com/google/blockly

2.7. SCIENTIFIC ABSTRACT VII

2025 International Conference on Machine Intelligence for GeoAnalytics and Remote Sensing (MIGARS) | 979-8-3315-7920-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/MIGARS67156.2025.11231788

EFFICIENT AGGREGATE LAND COVER QUERIES WITH CLOUD-OPTIMIZED RASTER FORMATS

Luke McQuade
Department of Geoinformatics – Z_GIS
Paris Lodron Universität Salzburg
Salzburg, Austria
luke.mcquade@plus.ac.at

Martin Sudmanns
Department of Geoinformatics – Z_GIS
Paris Lodron Universität Salzburg
Salzburg, Austria
martin.sudmanns@plus.ac.at

Dirk Tiede
Department of Geoinformatics – Z_GIS
Paris Lodron Universität Salzburg
Salzburg, Austria
dirk.tiede@plus.ac.at

***Abstract*—Semantic queries of Earth observation (EO) imagery, such as "Find the images with less than 5% cloud cover in this area between two dates", rely on aggregating some kind of scene classification data. Given the millions of pixels in a typical image, this can be resource-intensive. If knowledge is required of only the relative distribution of classes, though, do we need to process every pixel? The size and shape (morphology) of natural features, when observed by high-resolution optical sensors such as Sentinel-2, mean that simple downsampling can be used to drastically reduce the processing required for such queries, with only small losses in accuracy. Given an error tolerance of 1%, memory usage can be reduced by 625× and query runtime by 12×, for areas of interest of 60 km × 60 km. Using cloud native technologies such as Cloud Optimized GeoTIFF (COG), this can also lead to significant reductions in network and disk usage.**

***Index Terms*—scene classification, land cover, Sentinel-2, cloud-native geospatial**

I. Introduction

Some typical use cases of satellite Earth Observation (EO) datasets are queries such as "Find the images with less than 5% cloud cover between two dates", or "Find the first image of the year where the vegetation content of this area rises above 20%". These require some kind of scene classification information, which is then aggregated to the user's area of interest. Given the millions of pixels in a typical image, and dozens to hundreds of images to query against, this can be resource-intensive. The natural features that underlie classification maps often exhibit some degree of spatial autocorrelation [3], [17], so if we are querying against only the relative distribution of classes rather than their precise location, do we need to process every pixel? In other words, to what extent does downsampling (upscaling) affect query accuracy?

The multispectral instrument (MSI) of the Copernicus Sentinel-2 mission captures data from 13 optical bands at spatial resolutions of 10 m to 60 m (ground sampling distance). From these, Level-2A (L2A) analysis-ready products are created. One component of this is the Scene Classification Layer (SCL)—an auxiliary information layer generated with a rules-based classifier, comprising twelve classes covering various land cover types, cloud, and anomalous readings (see Table I) [5].

The research leading to these results has received funding from the European Union's Horizon Europe research and innovation program under the Grant Agreement No. 101082493 (Project: LEONSEGS)

Cloud Optimized GeoTIFF (COG) is a geospatial raster file format where data are stored as tiles, such that spatial subsets can be read without needing to load the entire file. This subsetting works over a network using, e.g., HTTP range requests. The underlying GeoTIFF offers pyramid/overview features that allow access to data at different scales. This can drastically lower the data transfer and processing costs of geospatial analysis [2]. But whether this can be leveraged depends on the use case, two factors of which are the size of the user's area of interest (AoI) and the scale (areal unit size).

TABLE I
SCL Classes. Source: [5].

Label	Class
0	No data
1	Saturated or defective
2	Dark area pixels
3	Cloud shadows
4	Vegetation
5	Bare soils
6	Water
7	Unclassified
8	Cloud - medium probability[a]
9	Cloud - high probability[a]
10	Thin cirrus[a]
11	Snow/ice

[a] Combined in our analysis.

The Microsoft Planetary Computer (MSPC) platform [11] hosts a Sentinel-2 L2A data archive, including the SCLs. The asset data is stored as COGs, which can be located via a Spatiotemporal Asset Catalog (STAC) metadata search and retrieved via HTTP(S). The SCL data is stored as a raster with 20 m spatial resolution [10]. A typical S2 image granule spans approximately 110 km × 110 km [6]. An entire SCL raster, therefore, has around 30 megapixels. The class distribution for the entire scene is actually

pre-computed and stored in the metadata, accessible by parameters such as `s2:vegetation_percentage` and `s2:snow_ice_percentage` [12]. This is useful if a user's AoI matches the granule's footprint, but this is often not the case, e.g., for analyses based on administrative boundaries.

The role of scale in remote sensing has been covered extensively in [19], but this was before the time of the Sentinel missions, COGs, and Big Earth Data. Later works have revisited the topic, exploring different resampling approaches and demonstrating the effects of downsampling in a range of other settings, such as classification and land cover accuracy on a pixel-by-pixel basis [13], [14], [20], [21].

II. Methods

A. Study areas and period

Two study areas were selected: an area of the Douro Valley, Spain (between 41.47°N to 42.57°N latitude and 3.57°W to 5.40°W longitude), and Stockerau, Lower Austria (between 47.78°N to 48.87°N latitude and 15.01°E to 16.65°E longitude), representing two different agricultural areas in Europe with different land use/land cover. The study period was 1st June to 1st July 2023, when the expected conditions for these areas of Europe would be a mix of cloudy and clear days, with significant vegetation growth.

B. Downsampling analysis

Based on the study areas, a series of synthetic AoIs were generated: square tiles of side lengths (sizes) varying from 1 km to 60 km, using the coordinate reference system (CRS) of the corresponding Universal Transverse Mercator (UTM) grid zone. For reference purposes, these were grouped by size. An overview is provided in Table II.

TABLE II
Summary of AoIs and Valid Image Samples

AoI tile size (m)	Size group	# AoIs	Valid samples[b]
1000	Small	123	2315
2000	Small	37	680
3000	Small	19	340
4000	Small	12	210
6000	Medium	21	345
8000	Medium, Large	43	694
10 000	Large	15	214
12 000	Medium, Large	22	348
20 000	Large, Extra-large	76	1021
40 000	Extra-large	8	82
60 000	Extra-large	2	20

[b] Matching, complete image chips.

For each AoI, a STAC search was performed. For each matching image, the SCL was windowed to the AoI, and the class distribution (as percentages) at native resolution (20 m) was calculated. The SCL was then queried at a series of lower resolutions using regular nearest-neighbor downsampling. The resolutions were selected using a loosely exponential pattern, representing a variety of analytical scales: 50 m, 100 m, 200 m, 500 m, and 1000 m.

The class distributions were calculated for each of the downsampled variants and compared against that of the native resolution. The maximum absolute difference in percentages, which we refer to as the maximum classification error, is our key metric; it directly informs the accuracy of our use case queries. An illustrative example is given in Table III.

TABLE III
Hypothetical Class Distribution Change

	Vegetation	Bare soil	Cloud
Native (20m)	10%	10%	80%
Downsampled (100m)	8%	11%	81%
Difference	−2%	+1%	+1%
Absolute difference	2%[c]	1%	1%

[c] Maximum classification error

Some preliminary results showed low spatial autocorrelation among the cloud classes (8, 9, 10) producing high eventual maximum classification errors. However, our use cases suggested that an aggressive approach to cloud filtering would be appropriate, and these classes were combined via summation of their percentages.

Sometimes an AoI would only partially overlap with the footprint of an image granule, yielding areas of 'no data', or the image data itself would contain 'no data' values. These were excluded from the class distribution comparisons. Valid samples, as in Table II, therefore refer to image chips that contain zero 'no data' values.

The maximum classification errors were compared across all AoIs for each tile size and downsampling resolution. From this, we could determine the distributions of classification errors due to downsampling, and where the maximum exceeded a user-defined threshold/tolerance (1% as an initial suggestion).

The analysis was performed with Python 3.13 in a `marimo` notebook [1], using packages `pystac_client` [7] and `stackstac` [9] for data loading, packages `xarray` [8], `pandas` [16], `geopandas` [4] and `scipy` [18] for data processing and analysis, and `holoviews` [15] for visualization. This code was running on an HP EliteDesk workstation with an Intel Core i5-8500 processor (3 GHz, 6 cores) and a 256 GB solid-state system drive, in a Windows Subsystem for Linux (WSL) Docker container with 24 GB of memory available.

C. Performance

For each AoI, we timed the duration of querying the matching SCL assets from MSPC and calculating the class distributions at a given resolution. This consisted of remote cloud computation, storage access and network transfer, as well as local computation. The measurements are, therefore, highly subject to external conditions such as network, storage and computation contention, and various types of caching. Samples containing 'no data' values were included in these measurements. There were around 10 to 20 images matching each AoI in the date range; the duration to load and process all of these was recorded and divided by the number of

images, giving a mean per-scene runtime. The number of pixels in each image was also recorded, as this, rather than physical size, determined the amount of data involved. These ranged from single-pixel images (as in a heavily downsampled small AoI) to 9 megapixels (as in a 60 km size tile at native resolution). The Python standard library's `time` module was used. Note that the duration of the initial STAC search to find the matching images for the AoI was not included.

III. Results

Fig. 1 shows the distributions of the maximum classification errors at a downsampling resolution of 50 m for AoIs in the Small subset. Here it can be seen that maximum classification errors approach 6% at tile sizes of 1 km but fall under 1% at tile sizes of 4 km. For readability, similarly detailed presentations of the results for the remaining resolutions and AoIs have been omitted. However, the key results are summarized in Table V, which shows the maximum downsampling resolution for each tile size before the maximum classification error exceeds a tolerance threshold of 1%. Note how the maximum downsampling resolution rises to 500 m at AoI tile sizes of 40 km.

A summary of the performance measurements by pixel count (grouped exponentially due to the wide range of magnitudes) is given in Table IV. Individual data are shown in Fig. 2. Together, these show that the average runtime is approximately constant for images up to around 1000 pixels, before steadily increasing with the number of pixels.

Two queries were excessively long-running compared with those of similar pixel counts and were deemed outliers. It was observed that memory usage was very high during this time, and paging to disk could be the cause.

A simple least-squares linear regression model fits the data (with outliers removed), with a slope of 2.07×10^{-8} s/pixel, y-intercept of 0.016 s, and an R^2 of 0.77. From this, the query speed-up from downsampling can be estimated. These are shown for each tile size in Table V. Downsampling offers a mild benefit to query runtimes for AoIs of 10 km or less, then reductions quickly increase with larger AoI sizes.

TABLE IV
Summary of Mean-per-scene Query Runtimes

Image size (pixels)	Query count	Min. (s)	Max. (s)	Median (s)
1 to 10	302	0.009	0.054	0.012
10 to 100	267	0.009	0.061	0.012
10^2 to 10^3	607	0.009	0.042	0.013
10^3 to 10^4	469	0.010	0.054	0.015
10^4 to 10^5	417	0.010	0.050	0.017
10^5 to 10^6	202	0.012	0.077	0.023
10^6 to 10^7	112	0.026	9.781	0.041

IV. Discussion and Conclusion

Our results show that downsampling can significantly reduce the data usage and runtime of aggregate, distribution-based land cover queries, if afforded a small tolerance for error: up to a 625× data reduction and estimated 12× decrease in query runtime for the largest AoIs in our experiment. For smaller AoIs, the benefits were smaller. In terms of class distribution error, smaller AoIs are more sensitive. In terms of performance, the smaller number of pixels meant they occupied the 'sill' where runtime was predominantly made up of network latencies rather than data processing.

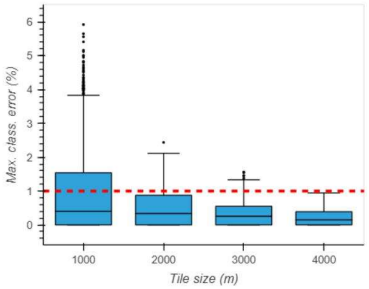


Fig. 1. Distributions of the maximum classification errors at a downsampling resolution of 50m, for the Small subset of AoIs. Quartiles are represented by the boxes, data within 1.5 × interquartile range by whiskers, and outliers as dots. The dashed red line marks a 1% error tolerance threshold.

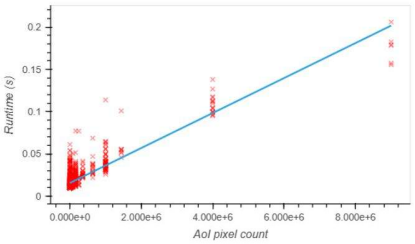


Fig. 2. Query runtime, mean per-scene. The blue line is a linear trendline, with $R^2 = 0.77$ and y-intercept= 0.016s. 2 outliers (per-scene runtime > 1.0s) have been omitted.

The average number of valid samples per AoI may seem high, given the revisit times of Sentinel-2 and the study period of only one month. This is partly explained by AoIs matching more than one granule footprint, but is mainly due to the STAC search returning several reprocessed versions of the original data. This behavior was uniform across all AoIs and therefore should not affect the conclusions drawn from the results.

There are likely more optimal resolutions to which AoIs can be downsampled, especially for the smallest AoIs and those larger than we tested. The experiment was kept constrained to avoid excessive use of cloud resources (a full run completes in approximately 1 h).

The linear regression model used to estimate the potential

TABLE V
MAXIMUM DOWNSAMPLING RESOLUTION BEFORE MAXIMUM CLASSIFICATION ERROR EXCEEDS 1% ("N/A", OTHERWISE), AND ASSOCIATED DATA AND ESTIMATED PER-SCENE QUERY RUNTIME REDUCTIONS

Tile size (m)	Max. downsampling resolution (m)	Pixels (native)	Pixels (downsampled)	Data reduction (multiple)	Est. runtime (native)	Est. runtime (downsampled)	Est. runtime reduction (multiple)
1000	N/A	2500	N/A	N/A	0.0157	N/A	N/A
2000	N/A	10 000	N/A	N/A	0.0159	N/A	N/A
3000	N/A	22 500	N/A	N/A	0.0161	N/A	N/A
4000	50	40 000	6400	6.25	0.0165	0.015 78	1.04
6000	50	90 000	14 400	6.25	0.0175	0.015 95	1.10
8000	50	160 000	25 600	6.25	0.0190	0.016 18	1.17
10 000	100	250 000	10 000	25	0.0208	0.015 86	1.31
12 000	100	360 000	14 400	25	0.0231	0.015 95	1.45
20 000	200	1 000 000	10 000	100	0.0363	0.015 86	2.29
40 000	500	4 000 000	6400	625	0.0983	0.015 78	6.23
60 000	500	9 000 000	14 400	625	0.2016	0.015 95	12.64

speed-ups from downsampling is also likely less than optimal, and this should be considered before extrapolating or expanding on the resulting numbers.

Only uniform, regular grid downsampling was used here. Other methods, such as the adaptive approach of Mîrt et al. [13], could perform better in terms of maintaining class distribution. But, this could lose the data handling performance benefits tied to the rectangular-tiled COGs.

Since the COG already stores pyramids with layers having different resolutions, there is no additional storage required on the server side. If the COG internal tile size and overview structure are known ahead of time, adjusting the downsampling resolution to match these could lead to additional performance gains. As a corollary, our results could inform the parameters for generating future COGs.

Repeating the approach on additional, geographically diverse study areas would be a useful extension of this work, as would applying it to land cover maps beyond the Sentinel-2 SCL, or even other binned/categorical data types.

ACKNOWLEDGMENT

Many thanks to our colleague, Felix Kröber, whose earlier explorations and Jupyter notebooks on Sentinel-2 STACs and the SCL were a great help; and to the Microsoft Planetary Computer team for making their service available to us.

REFERENCES

[1] A. Agrawal and M. Scolnick, "marimo," Apr. 2025, version: v0.12.9. [Online]. Available: https://github.com/marimo-team/marimo
[2] C. O. G. Community, "Cloud Optimized GeoTIFF," 2025. [Online]. Available: https://cogeo.org/
[3] P. J. Curran, "The semivariogram in remote sensing: An introduction," *Remote Sensing of Environment*, vol. 24, no. 3, pp. 493–507, Apr. 1988.
[4] J. V. den Bossche et al., "geopandas/geopandas: v1.0.1," Jul. 2024. [Online]. Available: https://zenodo.org/doi/10.5281/zenodo.12625316
[5] European Space Agency, "Sentinel-2 Level-2A Algorithm Theoretical Basis Document," ESA – European Space Agency, Frascati, Italy, Technical Report S2-PDGS-MPC-ATBD-L2A, Nov. 2021, version: 2.10.
[6] ——, "Sentinel-2 Products Specification Document," ESA – European Space Agency, Frascati, Italy, Tech. Rep. S2-PDGS-TAS-DI-PSD, Mar. 2021.
[7] M. Hanson et al., "stac-utils/pystac-client," Feb. 2025, version: v0.8.6. [Online]. Available: https://github.com/stac-utils/pystac-client
[8] S. Hoyer et al., "xarray," Mar. 2025, version: v2025.03.1. [Online]. Available: https://zenodo.org/doi/10.5281/zenodo.15110615
[9] G. Joseph, "gjoseph92/stackstac," Aug. 2024, version: v0.5.1. [Online]. Available: https://github.com/gjoseph92/stackstac
[10] Microsoft, "Sentinel-2 Level-2A | Microsoft Planetary Computer," accessed: May 13 2025. [Online]. Available: https://planetarycomputer.microsoft.com/dataset/sentinel-2-l2a
[11] Microsoft Open Source, M. McFarland, R. Emanuele, D. Morris, and T. Augspurger, "microsoft/PlanetaryComputer: October 2022," Oct. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7261897
[12] M. Mohr and P. Varner, "stac-extensions/sentinel-2," Nov. 2023, version: v1.0.0. [Online]. Available: https://github.com/stac-extensions/sentinel-2
[13] A. Mîrt, J. Reiche, J. Verbesselt, and M. Herold, "A Downsampling Method Addressing the Modifiable Areal Unit Problem in Remote Sensing," *Remote Sensing*, vol. 14, no. 21, p. 5538, Jan. 2022.
[14] R. Raj, N. A. S. Hamm, and Y. Kant, "Analysing the effect of different aggregation approaches on remotely sensed data," *International Journal of Remote Sensing*, Jul. 2013.
[15] P. Rudiger et al., "holoviz/holoviews: Version 1.20.2," Mar. 2025. [Online]. Available: https://zenodo.org/doi/10.5281/zenodo.15019128
[16] The pandas development team, "pandas-dev/pandas: Pandas," Sep. 2024. [Online]. Available: https://zenodo.org/doi/10.5281/zenodo.13819579
[17] W. R. Tobler, "A Computer Movie Simulating Urban Growth in the Detroit Region," *Economic Geography*, Jun. 1970, publisher: Routledge.
[18] P. Virtanen et al., "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020.
[19] S. Weigel, "Scale, Resolution and Resampling: Representation and Analysis of Remotely Sensed Landscapes Across Scale in Geographic Information Systems." *LSU Historical Dissertations and Theses*, Jan. 1996.
[20] H. Wu and Z.-L. Li, "Scale Issues in Remote Sensing: A Review on Analysis, Processing and Modeling," *Sensors*, vol. 9, no. 3, pp. 1768–1793, Mar. 2009. [Online]. Available: https://www.mdpi.com/1424-8220/9/3/1768
[21] K. Xu, Q. Tian, Y. Yang, J. Yue, and S. Tang, "How up-scaling of remote-sensing images affects land-cover classification by comparison with multiscale satellite images," *International Journal of Remote Sensing*, Apr. 2019.

2.8. SCIENTIFIC ABSTRACT VIII

Semantic content-based image retrieval in semantic EO data cubes

Martin Sudmanns[1], Dirk Tiede[1], Andrea Baraldi[2]

[1]*Paris Lodron University Salzburg, 5020, Salzburg, Austria*
[2]*Spatial Services GmbH, 5020, Salzburg, Austria*

EO (Earth observation) data cubes were developed and are used for time series analysis of (continuous) variables such as reflectance values or vegetation indices [1-3]. For such use cases, they have been proven powerful and support applications relying on aggregation of statistical values or trend analysis.

However, several use-cases require investigating images before and after an event. While an image time series is still useful to investigate and analyze the event, the tools available to users are not sufficiently applicable to support effective and efficient workflows. If only a date and a geographic area (area-of-interest, AOI) of an event are known, it is cumbersome to select the best images closest to the event, e.g., based on cloud cover or any other criteria. Workflows can use the global image statistics, which are in many cases not applicable to a geographic subset and generally require intensive human involvement. For example, an image-wide cloud cover statistic may report a value of 80%, suggesting to discarding the image, although the AOI may fit within the 20% cloud-free area. Such an image would be overlooked in automated processes based on this threshold. The alternative is a human operator selecting the images based on experience and visual inspection.

To fill this gap, we developed a semantic content-based image retrieval (SCBIR) approach that can automatically select best images using custom criteria. Leveraging semantic EO data cubes, where for each observation at least one categorical interpretation is available [4,5], users can select an AOI and a date of an event as well as their search criteria. An event can be defined as any user-defined date for which EO images should be analyzed (e.g. natural disaster, harvesting events, construction activity etc.). An automated inference evaluates the criteria for each available image tailored for the selected AOI. After successful evaluation, the best images before and after an event are automatically selected. The users can choose to obtain the images and forward them into their workflows or continue within the semantic EO data cube, e.g., conducting automatic change detection. While SCBIR using cloud cover is generic, some other operations are application-specific. However, once the workflow is defined, the semantic approach is generalized enough to be re-usable and transferable to any region worldwide.

Figure 1 illustrates an example application for SCBIR using the forest fire on Tenerife, Spain, in August 2023. From the stack of Sentinel-2 images, all acquisitions are considered without pre-filtering and semantically enriched within a semantic EO data cube using the SIAM-based [6,7] color33 semantic enrichment approach. This part is application-agnostic and does not have to be repeated for each application. In the next step, the target date, the approximate area of the forest fire, and a cloud filter are selected. Based on this, the cloud-free images closest in time before and after the target date are automatically selected. The selected images are forwarded to an automatic change detection process and the generation of a binary mask indicating the impact and extent of the forest fire. This application is defined and can be applied to other forest fires worldwide.

In this contribution, we present the approach as well as selected use-cases highlighting the importance of image organization in spatio-temporal EO data cubes, semantic analysis, and transferability and reusability of the method in a worldwide context.

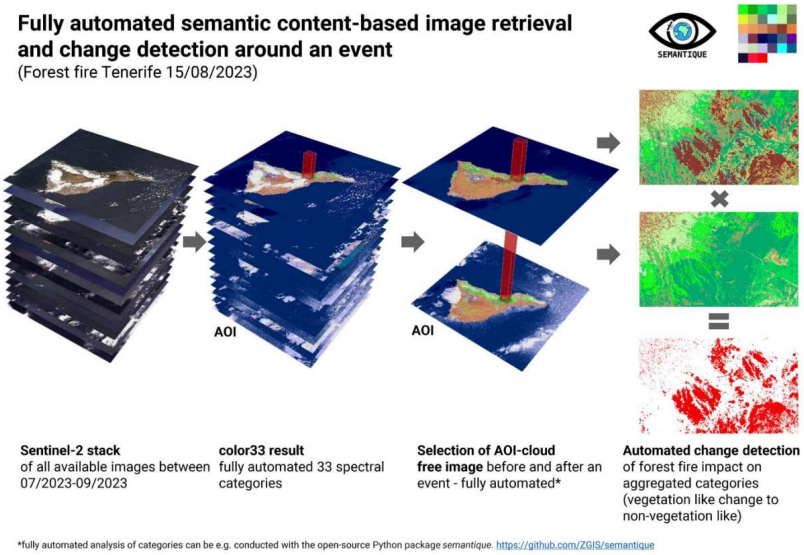


Figure 1: Example application of semantic content-based image retrieval for selecting images before and after an event using content-based criteria. The analysis is conducted in the image-subset (red cube) only, image wide cloud statistics are ignored to find any image which is cloud free at this specific location.

[1] Nativi, S.; Mazzetti, P.; Craglia, M. A View-Based Model of Data-Cube to Support Big Earth Data Systems Interoperability. *Big Earth Data* **2017**, *1*, 75–99 https://doi.org/10.1080/20964471.2017.1404232

[2] Lewis, A.; Lymburner, L.; Purss, M.B.J.; Brooke, B.; Evans, B.; Ip, A.; Dekker, A.G.; Irons, J.R.; Minchin, S.; Mueller, N.; et al. Rapid, High-Resolution Detection of Environmental Change over Continental Scales from Satellite Data—The Earth Observation Data Cube. *Int. J. Digit. Earth* **2016**, *9*, 106–111. https://doi.org/10.1080/17538947.2015.1111952

[3] Giuliani, G.; Chatenoux, B.; De Bono, A.; Rodila, D.; Richard, J.-P.; Allenbach, K.; Dao, H.; Peduzzi, P. Building an Earth Observations Data Cube: Lessons Learned from the Swiss Data Cube (SDC) on Generating Analysis Ready Data (ARD). *Big Earth Data* **2017**, *1*, 100–117. https://doi.org/10.1080/20964471.2017.1398903

[4] Augustin, H., Sudmanns, M., Tiede, D., Lang, S., & Baraldi, A. (2019). Semantic Earth observation data cubes. *Data*, *4*(3), 102. https://doi.org/10.3390/data4030102

[5] Sudmanns, M., Augustin, H., van der Meer, L., Baraldi, A., & Tiede, D. (2021). The Austrian semantic EO data cube infrastructure. *Remote Sensing*, *13*(23), 4807. https://doi.org/10.3390/rs13234807

[6] Baraldi, A.; Durieux, L.; Simonetti, D.; Conchedda, G.; Holecz, F.; Blonda, P. Automatic Spectral-Rule-Based Preliminary Classification of Radiometrically Calibrated SPOT-4/-5/IRS, AVHRR/MSG, AATSR, IKONOS/QuickBird/OrbView/GeoEye, and DMC/SPOT-1/-2 Imagery—Part I: System Design and Implementation. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 1299–1325. https://doi.org/10.1109/TGRS.2009.2032457

[7] Baraldi, A.; Durieux, L.; Simonetti, D.; Conchedda, G.; Holecz, F.; Blonda, P. Automatic Spectral Rule-Based Preliminary Classification of Radiometrically Calibrated SPOT-4/-5/IRS, AVHRR/MSG, AATSR, IKONOS/QuickBird/OrbView/GeoEye, and DMC/SPOT-1/-2 Imagery—Part II: Classification Accuracy Assessment. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 1326–1354 https://doi.org/10.1109/TGRS.2009.2032064

3. CONCLUSIONS

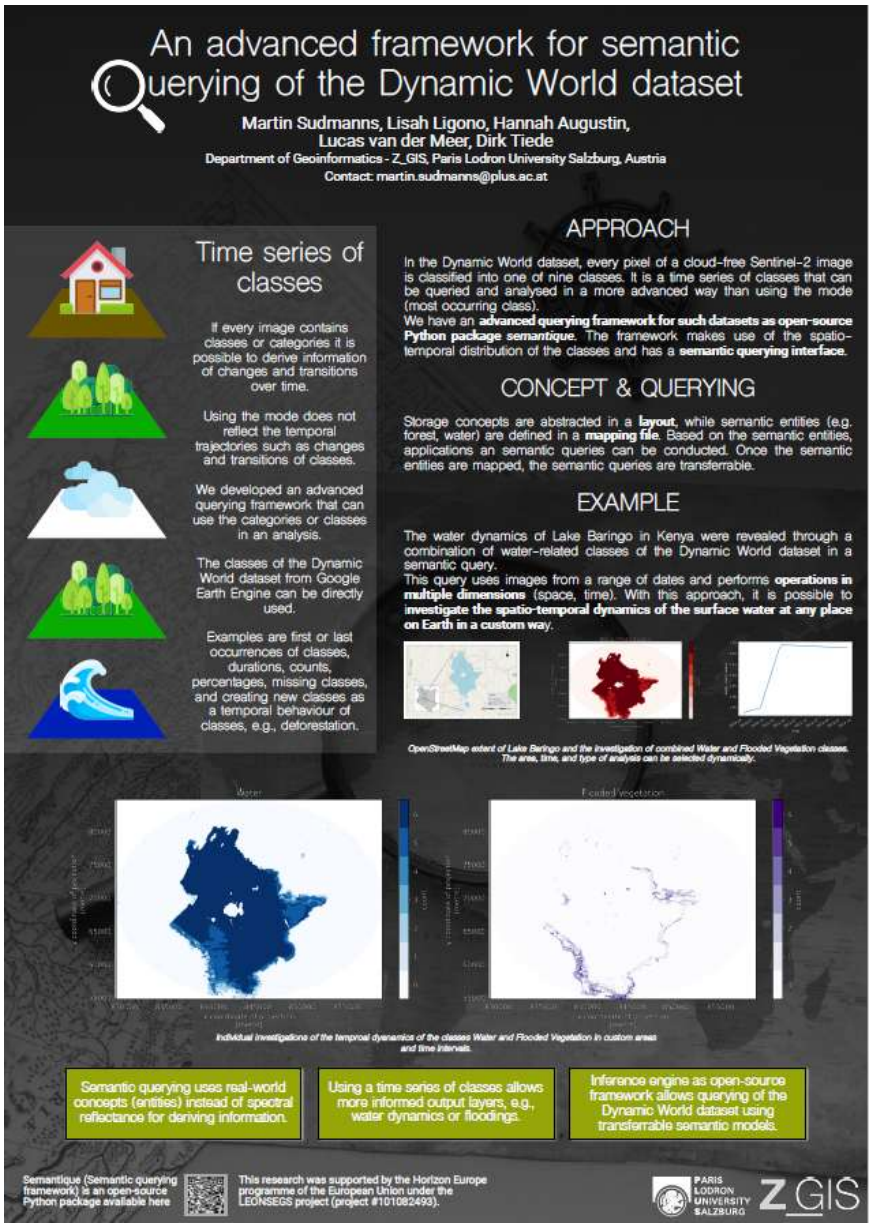
During the initial phase of the LEONSEGS project, research and development activities were extensively published and presented at various scientific conferences. A notable achievement was the organization of a dedicated workshop on semantic and explainable big EO data analysis at the EARSeL symposium. This workshop showcased advancements in EO data analysis, particularly in semantic and explainable methodologies, within an international scientific context, fostering discussions among diverse research groups. The receipt of the Best Scientific Poster Award at the ESA BiDS conference further emphasized the scientific community's interest in semantic information extraction.

As the project moved into the second phase, the primary objectives were (1) to enhance the publication of LEONSEGS scientific project outcomes, which could successfully achieved through high-level publications and presentations at some of the highest impact journals in the domain and (2) to promote the project and its outcomes to interested potential users / companies through participation in workshops and conferences with mixed audiences (companies, research organizations, authorities/stakeholders etc.) like ESA LPS, ESA Big Data from Space, AGIT 2025, ISDE 2025, the GEO Forum 25 and others.

Together, these activities have significantly increased both the scientific impact and the practical visibility of LEONSEGS, laying a solid foundation for the final project phase.

4. APPENDIX

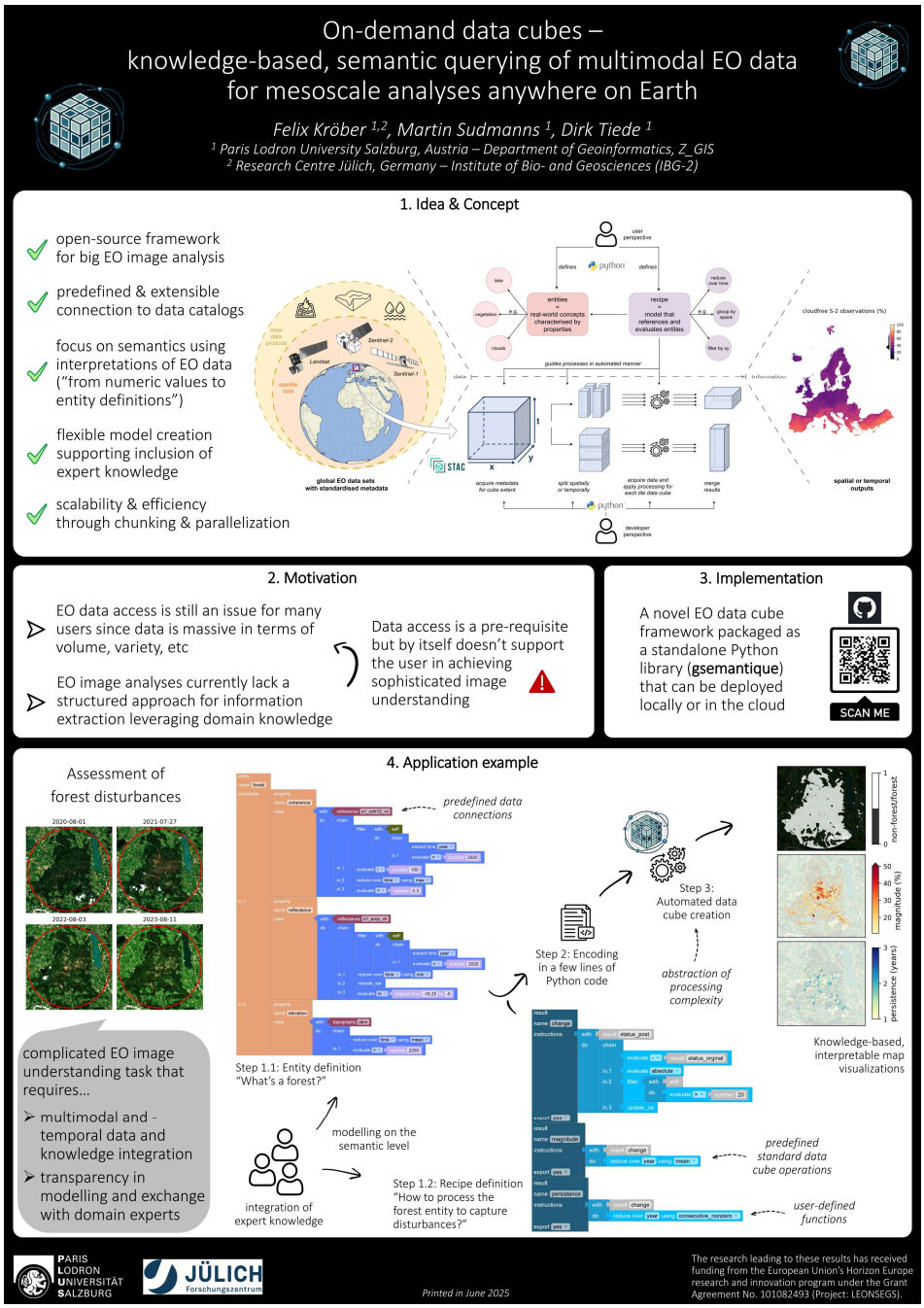
The following poster has been awarded with the best poster award at ESA BiDS 2023 (Scientific Abstract No I):



The following poster has been presented to promote LEONSEGS on different scientific events (e.g. LPS 2025, see Table 2)

The following poster has been presented at several events to promote LEONSEGS on different scientific events (e.g. GEO Global Forum, see Table 2)

The following poster has been presented at the LPS 2025 (see Table 1)



Dissemination Level: **PUBLIC**