MOOC Cubes and Clouds - Cloud Native Open Data Sciences for Earth Observation

Peter James Zellner[1], Michele Claus[1], Tyna Dolezalova[2], Rufai Omowunmi Balogun[1], Jonas Eberle[3], Henryk Hodam[4], Robert Eckardt[4], Stephan Meißl[2], Alexander Jacob[1], Anca Anghelea[5]

[1]Eurac Research, Bolzano, Italy – (peterjames.zellner, michele.claus, rufaiomowunmi.balogun, alexander.jacob)@eurac.edu
[2]EOX IT Services GmbH, Vienna, Austria – (tyna.dolezalova, stephan.meissl)@eox.at
[3]Eberle Web- and Software-Development, Reutlingen, Germany - jonas.eberle@eberle-mail.de
[4]ignite education GmbH, Jena, Germany – (henryk, Robert)@ignite-education.io
[5]European Space Agency, Frascati, Italy - anca.anghelea@esa.int

Abstract

The "Cubes and Clouds" Massive Open Online Course (MOOC) addresses the growing challenges and opportunities in Earth Observation (EO) data analysis posed by the exponential growth of satellite missions and data volumes. This course introduces cloud-native EO concepts and open science principles, facilitating collaboration and data sharing within the EO community.

Key topics covered in the MOOC include data discovery, processing workflows, and data sharing using real-world examples like snow cover mapping in the Alps. Participants engage with interactive lessons, videos, and hands-on coding exercises, leveraging freely available geospatial data and emphasizing open data principles and interoperability.

The course infrastructure seamlessly integrates with cloud platforms like the Copernicus Data Space Ecosystem, enabling learners to apply concepts in a practical, cloud-based environment. Initial user statistics indicate strong interest, particularly among early-career professionals and researchers, with participant surveys suggesting increased confidence in using EO cloud platforms and embracing open science practices upon course completion.

Overall, "Cubes and Clouds" serves as a valuable resource for the EO community, fostering transparency, collaboration, and innovation in geospatial science. Future developments aim to expand course content and attract diverse user groups, ensuring broader accessibility and long-term impact in advancing cloud-native EO and open science initiatives.

1. Introduction

In the last years the number of satellite missions and the data volume they are producing has been growing exponentially. To exploit the full wealth of available satellite data, which means combining different data sources, creating long time series and mapping large extents of the world, the traditional approach of downloading data to a private PC is no longer viable. The necessity to overcome this technological burden has been addressed by the rapid development of cloud technology in the geospatial and EO field. Cloud platforms have been created, such as Google Earth Engine (Gorelick et al., 2017) or the Copernicus Dataspace Ecosystem (CDSE Development Team, 2023) to enable analysis of full data archives on remote servers. Using these services demands adaption in the working style in the EO field. Early adopters in research and industry have started the transition, but a large majority is still most comfortable with the traditional approach of downloading data to a local PC and setting up data infrastructures and codes themselves (Crowley et al., 2023).

In addition to the technological shift there is also the need to adhere to open science practices. The challenges of the 21st century cannot be addressed by isolated research activities. Collaboration within and across domains is key to address topics such as the impacts of climate change or food security. Open Science practices, such as the FAIR (Findable, Accessible, Interoperable, Reusable) movement for example (Wilkinson et al., 2016), are a set of methods and working style that facilitates and motivates collaboration by openly sharing every step of the research, allowing others to reuse and build upon relevant findings.

The Massive Open Online Course (MOOC) "Cubes and Clouds" (Zellner et al., 2024) teaches the concepts of data cubes, cloud platforms, and open science in the context of Earth Observation (EO). The course is designed to bridge the gap between relevant technological advancements and best practices and existing educational material in the cloud native geospatial and EO domain. Successful participants will have acquired the necessary skills to work and engage themselves in a community adhering to the latest developments in the geospatial and EO world. The course is available for free on EO College, an E learning platform dedicated to EO topics. The course material is furthermore available on GitHub and zenodo under the license CC-BY-4.0.

2. Data and Materials

2.1 Target group

The target group are earth science students, researchers, and data scientists who want to dive into the newest standards in EO cloud computing and open science. The course is designed as a MOOC that explains the concepts of cloud native EO and open science by applying them to a typical EO workflow from data discovery, data processing up to sharing the results in an open and FAIR way.

2.2 Geospatial Data

The geospatial data which is used throughout the course is openly available. The introduction to data cube processes relies on Sentinel 2 data fetched from public Spatio Temporal Asset Catalogues (STAC) (STAC Contributors, 2021). The final exercise, a full EO workflow, is using Sentinel 2 data from the Copernicus Data Space Ecosystem and snow depth measurements collected and shared openly by the ClirSnow

project (Matiu et al., 2021), as well as runoff data from the Alpine Drought Observatory project, which is also openly available (CC-BY-4.0).

2.3 Educational Data

The educational material produced in the MOOC relies to a large extent on existing sources that have been adapted and framed to fit the topic of cloud native geospatial. The Open Science block for example has reused many of the materials created by FAIR Data Austria which are intended for general use in research (FAIR Data Austria, 2021). This information has been adapted to fit the EO and cloud native geospatial domain. Likewise existing educational material, for example on data cubes (openeo.org, 2022; Söchting, 2022), has been integrated into the course and combined into a lecture. The complete references and suggested further reading can be found at the end of every lesson.

3. Methodology

3.1 Content

This MOOC is an open learning experience relying on a mixture of animated lecture content and hands-on coding exercises created together with community renowned experts. The course is structured into three main blocks *"Concepts"*, *"Discovery"* and *"Process and Share"*. The degree of interaction (e.g. hands-on coding exercises) is gradually increasing throughout the course. The theoretical basics are taught in the first block *"Concepts"*, comprising cloud platforms, data cubes and open science practices. In the second block *"Discovery"* the focus is on discovery of data and processes and the role of metadata in EO. In the final block *"Process and Share"* the participants carry out complete processing workflows on cloud infrastructure and apply open science practices to the produced results. Every lesson is concluded with a quiz, ensuring that the content has been understood.

The course contains 10 written chapters that convey the basic knowledge and theoretical concepts, 13 videos which have been created with a professional communication team and in collaboration with a leading expert on the topic and shines a light on a real world example (e.g. The role of GDAL in the geospatial and EO), 16 pieces of animated interactive content which engage the participants to actively interact with the content (e.g. Sentinel 2 Data Volume Calculator) and 11 hands-on coding exercises in the form of curated jupyter notebooks that access European EO cloud platforms (e.g. CDSE) and carry out analysis there using standardized API's like openEO (Schramm et al., 2021) (e.g. full EO workflow for snow cover mapping).

3.2 Course Outline

In detail the course comprises the following building blocks (Table 1).

Block	Chapter	Lesson
Concepts	What is a platform?	Why do we need EO cloud platforms?
		Types of platforms
		Components of platforms
	What is a data cube?	Data cubes in general
		Data cubes in EO
		Dimensions of data cubes
		Data cubes in EO cloud platforms
	Open Science, Open Data and the FAIR principles	Why do we need Open Science?
		What is Open Science?
		What is FAIR?
		What is Open Data?
		What is Open Source?
		The Open Science Journey in geospatial, EO and EO cloud platforms
Discovery	Data discovery	Geospatial data
		EO catalogue protocols
		The STAC data catalogue
		Data sources
	Data properties	What are data properties and metadata?
		Properties and metadata used for filtering
	Data access	Raster data loading
		Filter, Apply, Reduce, Resample, Aggregate
		Recap processes
		OpenEO – A standardized API for EO cloud processing
Process and Share	Data Processing	Case study: Snow cover in the Alps
	Data Validation	Validation of EO data
	Data Sharing	Publishing results using the STAC ecosystem

Table 1. Course outline of the MOOC Cubes and Clouds.

3.2.1 Concepts

The block *"Concepts"* teaches why using a cloud platform is useful, how to differentiate platform offerings, and explains the components and building blocks of a cloud platform.

The chapter *"What is a platform?"* teaches the essentials about cloud platforms in EO. Comprising the motivation why cloud platforms are relevant in EO for facilitating the analysis of EO data which typically involves several steps, including data discovery, data download, data pre-processing, and data analysis. This is especially relevant in the light of ever growing data volumes that cannot be handled on a single PC anymore and the homogenization of data access.

The chapter *"What is a data cube?"* teaches the concept of data cubes, which are multi-dimensional data structures suitable for describing EO data. The general concept of multi-dimensional data structures is taught and then applied to EO data, which usually comes in multiple dimensions, such as latitude, longitude, time, bands and more. It is shown how to interact with the dimensions and their specifics. Furthermore, the role of data cubes in cloud platforms is explained in order to overcome the dilemma of addressing single files and how data cubes allow unified access to different data sets on a platform.

The chapter *"Open Science, Open Data and the FAIR principles"* introduces the participants to the concepts behind openness and FAIRness in research and EO. The different lectures explain the respective topics, for example the role of open source software in geospatial. The final chapter shows the whole open science journey a researcher in geospatial goes through by analyzing the project ClirSnow (Matiu et al., 2021), which adheres perfectly to open science practices and explains how cloud platforms facilitate open science practices.

3.2.2 Discovery

The block *"Discovery"* teaches everything related to finding the right data in the geospatial and EO world, which role metadata

and data properties play and how to search, filter and access data sets. The level of hands-on exercises increases in this block, as the participants start applying the learned concepts by accessing cloud platforms and programmatically processing data cubes.

The chapter *"Data discovery"* deals with data types, catalogue protocols and highlights the STAC data catalogue as an example of how to organize metadata in a standardized way. Then the participants search for data in different catalogues independently.

The chapter *"Data properties"* focusses on metadata and data properties, the difference between them and how to exploit them for filtering without opening the actual data.

The chapter *"Data access"* introduces the participants to interact with real data. First the loading of data cubes is explained and then carried out via the openEO client side processing library and publicly available STAC catalogues. Subsequently the standard processes on data cubes are examined in the same way. The participants learn what filter, apply, reduce, resample and aggregate processes do and apply them to the data cube they have loaded. Finally, the motivation and benefit behind openEO, a standardized API for EO cloud processing, is explained with a focus on reproducibility and portability between different platforms.

3.2.3 Process and Share

The block *"Process and Share"* further increases the level of hands-on coding exercises. It introduces an end-to-end EO workflow, from data selection up to sharing the results obtained. The chapter *"Data processing"* guides the participants through a real life EO use case: Snow cover mapping in the Alps. First, the research question is defined, which is to create a time series of the snow-covered area of the catchment of interest. Then the approach is described in theory. The choice of the data sources is explained and the steps which lead to the result. Firstly, the data cube is defined to the needs of the test catchment in the Alps (Meran, Italy) for the year of 2018. Then the normalized difference snow index (NDSI) is calculated and subsequently used to create a binary snow map. In the next step, the clouds are masked out. Finally, the catchment statistics are calculated by aggregating spatially (counting the number of cloudy pixels, number of snow covered pixels, number of snow free pixels in the catchment). The result is visualized as a time series of the snow covered area in the catchment (Figure 1).

The chapter *"Result Validation"* explains the importance of validation, by introducing concepts that deal with the validation of global EO products, such as the area of applicability (Meyer and Pebesma, 2022). The hands-on exercise builds on the results obtained in the previous chapter and the participants apply different validation approaches (e.g. comparing the snow cover to the snow station measurements in the analyzed catchment and comparing the snow cover to the runoff as a plausibility measure, see Figure 1).
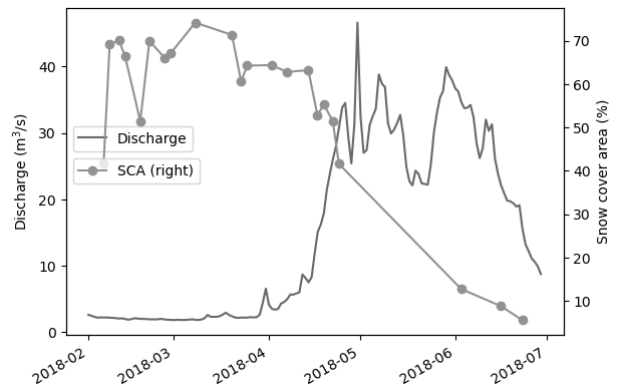

Figure 1. Time series of snow cover in the catchment Meran and the according discharge at the main outlet of the catchment.

The chapter *"Data Sharing"* teaches how to share data correctly and effectively. The participants reuse the workflow they have learned previously and a adapt it to a region of interest of their choice. Then the data properties and metadata are extracted from the result and a STAC metadata item is created for their result. Finally, the participants upload their metadata and result to a publicly available STAC browser "Cubes and Clouds – Snow Cover" (Cubes and Clouds Development Team, 2024). Every participant contributes in mapping a patch of snow cover over the alps. This last exercises closes the circle of theory and application, as the participants themselves contribute to the EO community by adhering to standards and sharing their results publicly.

3.3 Infrastructure

The MOOC Cubes and Clouds has developed a fully integrated learning infrastructure (Figure 2). The EOCollege E-learning platform (EO College, 2021) hosts the lectures and the animated content (e.g. videos, animations, interactive elements) of the course. It is also the main entry point for users providing a single sign on mechanism connected to the JupyterHub environment for the coding exercises. The hands-on exercises are directly accessible from EOCollege via a dedicated JupyterHub environment, which accesses European EO cloud platforms, such as the Copernicus Data Space Ecosystem, using its open science tools like the Open Science Data Catalogue, openEO and STAC. This guarantees that the learned concepts are applied to real-world applications. In the final exercise the participants map the snow cover of an area of interest they choose and make their results openly available according to the FAIR principles on an web viewer (STAC browser). The integration between the STAC browser and the JupyterHub environment is seamlessly integrated. This community mapping project actively lives the idea of open science, collaboration and community building. The content is created in a dedicated GitHub repository (Zellner et al., 2024). It hosts the lectures and the exercises and is the forum for issue tracking and collaboration. Upon every GitHub release a new version of the content is automatically released on zenodo (Zellner et al., 2024). The zenodo community "Cubes and Clouds" holds all material created during the project and assigns DOIs to all of them. The written lectures are in the markdown format to allow for easy integration into EOCollege and other frameworks. The coding exercises can be pulled directly into the JupyterHub environment, so that changes can be integrated seamlessly. The videos are hosted on the EOCollege YouTube channel and embedded into the lectures. The animated content is created via H5P on the EOCollege Creation Hub (Content Management System) and is also embedded into the lectures.
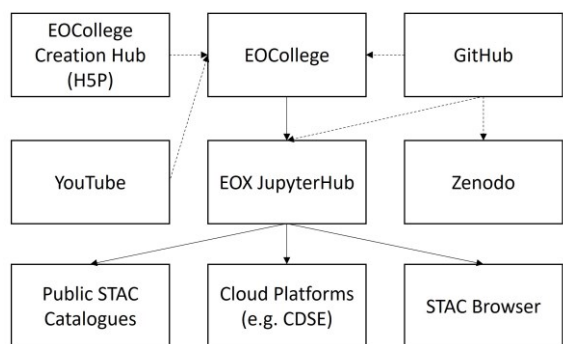
Figure 2. Infrastructure of the MOOC Cubes and Clouds. Arrows show the user flow. Full arrows show the user flow. Dashed arrows show the content flow.

3.4 Testing

As described in the previous chapter, the course contains multiple parts of infrastructure that need to function together flawlessly in order to guarantee a smooth learning experience (e.g. single sing on mechanism between EOCollege and the hosted JupyterHub environment, automatic update of bugfixes in the exercises from GitHub to JupyterHub environment, etc.). Before opening the course to the public two test rounds have been implemented. In a first test all internal project partners have completed the course and gathered feedback. In a second test round, 15 potential users have been contacted and asked to complete the course and give feedback so that the content and infrastructure could be tested in a real world scenario.

4. Results

4.1 Learning Achievements

After finishing the course, the participants will understand the concepts of cloud native EO, be capable of independently using cloud platforms to approach EO related research questions and be confident in how to share research by adhering to the concepts of open science. After the successful completion of the course the participants receive a certificate and diploma supplement and their personal map is persistently available in the web viewer (Cubes and Clouds Development Team, 2024) as a proof of work.

4.2 User Statistics

The course has been online for developers from mid-October 2023. It has been in development until the official opening. The course has opened for public on 2024-03-01. From this date on users could inscribe to the course on EO College and start learning. The promotional activities, including posts on social media (LinkedIn, Twitter/X), blog posts and news entries on the communities homepages (openEO, CDSE, etc.) have started on 2024-03-12. From this date on a rapid rise in user numbers has set on. Within the first month more than 200 people have started the course (Figure 3).
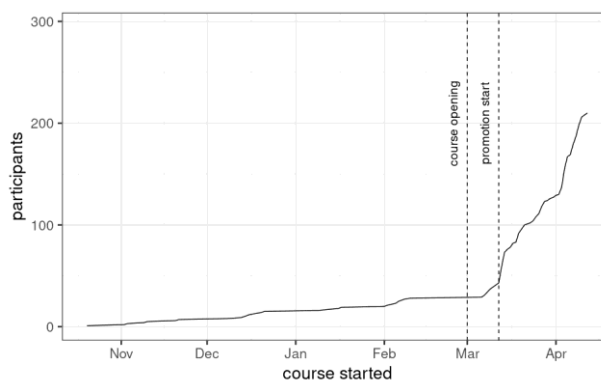

Figure 3. Time series of the users of the MOOC Cubes and Clouds as of 2024-04-12. The first dashed line shows when the course opened officially. The second dashed line indicates when the promotion for the course has started.

Currently, many users are in the first lesson of the course (61 %). This is an effect of the course opening. Many users subscribe and decide not to continue in the beginning to get an overview of what the course is about. Nevertheless, a significant portion of the users (32 %) is actively following the content and are between the first and last chapter. 7 % have already completed the course (Figure 4). We are looking forward to review the usage statistics after a longer period of time. It is expected that the growth continues and more participants move into the active participation.
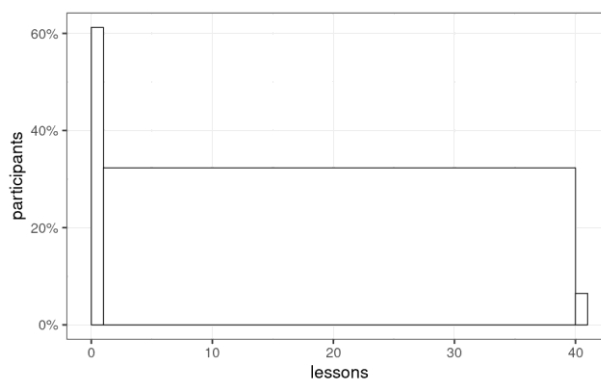

Figure 4. Histogram of the progress the users have made up to 2024-04-12. The first bin comprises users that are only subscribed. The second bin shows the active uses that are between lesson 1 and 40. The third bin shows the users who have completed the course.

4.3 User Survey

In addition to the standard user survey dedicated to the overall satisfaction with the course, an optional user survey is added to the course to trace the participants knowledge about cloud native EO and open science before and after the course. This serves to shed light on the interest in the topics and the feeling of competence the participants gain. The questions are rated with a five point system ranging from *"1 - strongly disagree"* to *"5 - strongly agree"*. So far there have been 86 respondents for the survey before the course and 10 for the survey after the course. Since the course has only opened previously most of the respondents are still actively following the course.

The first part of the survey depicts the general information about participants and the experience in the field the participants have before starting the course with an ample number of participants. The gender split is 30 % female and 70

% male participants with an even split between an engineering and a geoscience background. The envisioned target group is well represented. Geographically the participants come from 30 different countries split across all continents, whereas the majority comes from Europe (Figure 5). Still a very international group of participants is reached.
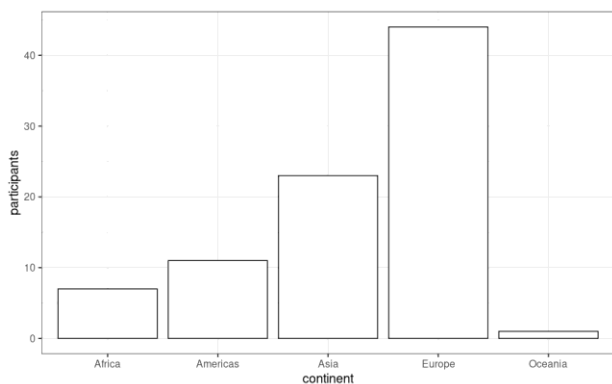

Figure 5. Geographical distribution of the participants of the MOOC Cubes and Clouds by continent.

The mean age of the participants is 30 and the years of experience in the EO field has a median of 3. The professional level is mostly young professionals in the beginning of their career (Figure 6). This is also well in line with the envisioned target group. Ideally more university students should be reached.
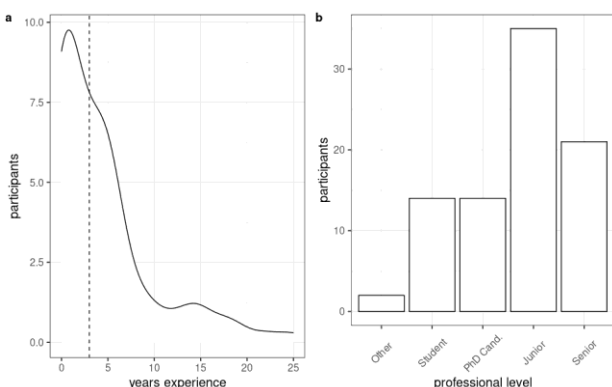

Figure 6. Years of experience (a) and professional level (b) of the participants of the MOOC Cubes and Clouds. The dashed line in (a) indicates the median professional level of 3 years.

The three questions to capture the status quo are "I am using EO cloud platforms", "I can carry out my research independently on EO cloud platforms" and "I can adhere to open science practices". The experience in the field of EO cloud platforms is low to medium, whereas the experience in the field of Open Science is rather high (Figure 7).
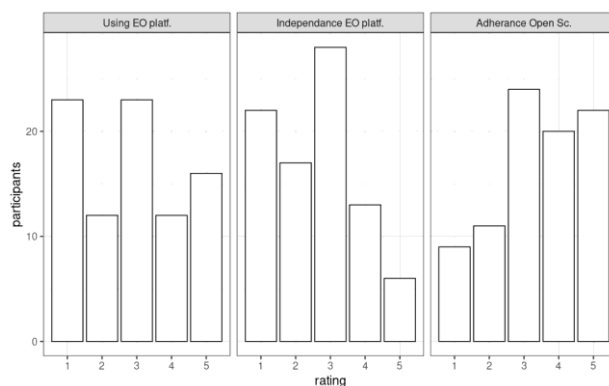

Figure 7. Experience with using EO cloud platforms, independence in using EO cloud platforms and the adherence to Open Science practices before the course started.

The analysis of the second survey, after the course, relies only on 9 participants who have completed the course so far. Nevertheless, some general insights about the feeling of competence and the satisfaction with the content can be derived cautiously. The same three questions that have been asked before the course are asked afterwards again: "I will use EO platforms from now on", "I can carry out my research independently on EO cloud platforms" and "I can adhere to open science practices from now on". In all three questions the ratings are higher than before the course. The median has risen from 3 to 4 in all three questions. Statistical significance is still to be tested as soon as more replies arrive. Furthermore, the difficulty level, the duration of the course, the proportion of theory and practical exercises, the usefulness of the exercises and the course structure were appreciated by the participants. The answers to these questions are all *"4 – I agree"* and *"5 – I strongly agree"*.

5. Discussion and Conclusion

The user statistics show good interest in the MOOC Cubes and Clouds. The participant numbers have grown steeply in the first month and have reached over 200. We expect that the portion of active participants will increase over time when the course establishes itself further. The user survey, which is split in one part before the course and one after already reveals some first insights. The participants of the course are international young professionals from the geosciences and engineering field. The gender split is tilted towards male participants which is common in technical fields. Nevertheless, the amount of female participants should be increased over time. Most participants are in their early stage of the career and the average age is 30. Overall the estimated target group is reached very well. The amount of university students should increase once the course finds its way into university domain.

The low experience with EO cloud platforms shows that the course is necessary and opens the opportunity to educate the new generation of EO researchers and professionals in this direction. Open science practices are more present in the portfolio of the participants. The preliminary analysis of the 9 graduates who have completed the survey after the course points into the direction that the confidence in using cloud platforms and open science rises after completing the course. The general quality of the course is well perceived. Over time more stable results will be obtained and will allow more robust evaluation.

The MOOC is valuable for the geospatial and EO community and open science as there is currently no learning resource available where the concepts of cloud native computing and

open science in EO are taught jointly to bridge the gap towards the recent cloud native advancements. The course is open to everybody, thus serving as teaching material for a wide range of purposes including universities and industry, maximizing the outreach to potential participants. In this sense also the raw material of the course is created following open science practices (e.g. GitHub repository, zenodo, STAC Browser for results) and can be reused, built upon and motivates contributions. The integrated infrastructure with one clear access point, the EO College E-learning platform, connects the users directly to the JupyterHub environment with one click and without further login and distraction by manually navigating between resources.

The MOOC Cubes and Clouds equips participants with essential skills in cloud native EO and open science, enhancing their ability to contribute meaningfully to the open geospatial community. By promoting transparency, reproducibility, and collaboration in research, graduates of the course strengthen the foundations of open science within the community. Access to cloud computing resources and European EO platforms empowers participants to undertake innovative research projects and share their findings openly, enriching the collective knowledge base. Ultimately, the MOOC fosters a culture of openness and collaboration, driving positive change and advancing the field of geospatial science for the benefit of all.

It is planned to actively continue developing the MOOC Cubes and Clouds. Currently a add on project adds the subjects of cloud native data formats, concepts of scaling and the cost of cloud computations. Furthermore, the exercises are expanded to the PANGEO software stack in addition to openEO. This allows users to either choose which or complete both paths. It also opens the course to the PANGEO community (PANGEO Team, 2023) attracting another user group and highlights the diversity of approaches in the EO cloud computing field. In terms of promotion, universities will be targeted more to establish the course in their curricula which will guarantee stable user numbers.

Acknowledgements

References

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment.

CDSE Development Team, 2023. Copernicus Data Space Ecosystem. https://dataspace.copernicus.eu/ (27 April 2024)

Crowley, M., Stuhlmacher, M., Trochim, E., Van Den Hoek, J., Pasquarella, V., Szeto, S., Howarth, J., Platt, R., Roy, S., Tellman, B., Chakraborty, T., Ignatius, A., Cherrington, E., Markert, K., Wu, Q., Madhusudan, M., Mayer, T., Cardille, J., Erickson, T., & Saah, D., 2023. Pillars of Cloud-Based Earth Observation Science Education. AGU Advances, 4.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

Zellner, P. J., Dolezalova, T., Claus, M., Eberle, J., Balogun, R. O., Meißl, S., Eckardt, R., Hodam, H., Jacob, A., & Anghelea, A. (2024). Cubes & Clouds - Cloud Native Open Data Sciences for Earth Observation (v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.10869466

STAC Contributors, 2021. SpatioTemporal Asset Catalog (STAC) specification. https://stacspec.org (27 April 2024)

Matiu, M., Crespi, A., Bertoldi, G., Carmagnola, C., Marty, C., Morin, S., Schoner, W., Cat Berro, D., Chiogna, G., De Gregorio, L., Kotlarski, S., Majone, B., Resch, G., Terzago, S., Valt, M., Beozzo, W., Cianfarra, P., Gouttevin, I., Marcolini, G., Notarnicola, C., Petitta, M., Scherrer, S., Strasser, U., Winkler, M., Zebisch, M., Cicogna, A., Cremonini, R., Debernardi, A., Faletto, M., Gaddo, M., Giovannini, L., Mercalli, L., Soubeyroux, J.M., Sušnik, A., Trenti, A., Urbani, S., & Weilguni, V., 2021. Observed snow depth trends in the European Alps: 1971 to 2019. The Cryosphere, 15(3), 1343–1382.

FAIR Data Austria, 2021. Open Educational Resources Research Data Management. DOI: 10.5281/zenodo.6923397

openeo.org, 2022. What are Data Cubes. https://openeo.org/documentation/1.0/datacubes.html (27 April 2024)

Maximilian Söchting, 2022. LexCube - Leipzig Explorer of Earth Data Cubes. https://www.lexcube.org/ (27 April 2024)

Schramm, M., Pebesma, E., Milenković, M., Foresta, L., Dries, J., Jacob, A., Wagner, W., Mohr, M., Neteler, M., Kadunc, M., Miksa, T., Kempeneers, P., Verbesselt, J., Gößwein, B., Navacchi, C., Lippens, S., & Reiche, J., 2021. The openEO API–Harmonising the Use of Earth Observation Cloud Services Using Virtual Data Cube Functionalities. Remote Sensing, 13(6).

Meyer, H., Pebesma, E., 2022. Machine learning-based global maps of ecological variables and the challenge of assessing them. Nat Commun 13, 2208. https://doi.org/10.1038/s41467-022-29838-9

EO College, 2021. EO College. https://eo-college.org/ (27 April 2024)

Cubes and Clouds Development Team, 2024. Cubes and Clouds – Snow Cover. https://esa.pages.eox.at/cubes-and-clouds-catalog/browser/#/ (27 April 2024)

PANGEO Team, 2023. PANGEO - A community platform for Big Data geoscience. https://pangeo.io/ (27 April 2024)