Implementing Data Citation on the Earth Observation Data Centre for Water Resource Monitoring (EODC)

Adopting the RDA Data Citation Recommendations on an openEO[1]

The Earth Observation Data Centre for Water Resource Monitoring (EODC) located in Austria works with high-performance computing services and large amounts of data on a daily basis. Since 2017, it has been part of the consortium implementing the open Earth Observation (openEO) standard, which aims to standardize communication between EO scientists and data and service providers. It allows scientists to write code one time only and use it on different backend providers, but has the downfall of being not transparent. Adding data citation to the standardized process enables insights into what specific data was used in EO workflows.

Bernhard Gößwein, TUWien
Thomas Mistelbauer, EODS

The challenge

The technical challenges we had to address are related to having a minimal additional performance and storage impact on the backend, but at the same time having big evolving input data. This was also crucial to convince other backends of the openEO consortium like EODC to implement the data citation recommendations. The real challenge we faced was the complexity of the EODC infrastructure and standards used to embed the data citation implementation naturally into the existing software..

The RDA outputs adopted

We adopted the 14 Recommendations of RDA Working Group on Data Citation. These gave us step by step instructions on what needed to be done and what we had to consider. Therefore, little background knowledge on data citation was required to adopt them. Furthermore, our implementation was inspired by the successful adoption of the same RDA recommendations at the Climate Change Centre Austria (CCCA), whose code is open source and provided a starting point.

Implementation

To implement the RDA recommendations, we investigated the current state of the recommendations at the EODC backend. We discovered that time-stamping and versioning of the data records are already in place.

The backend uses the unique file paths as versions and creates time-stamps of the files in a meta-database. Additionally, we understood the query technology (Open Geospatial Consortium Catalogue Service for the Web (OGC CSW)) and that a PostgreSQL database stores the metadata. In contrast, the actual data is stored in files. So the result of the subset query is a list of file paths that the backend needs to fetch for processing. The filter arguments of a query are the identifier of the satellite, the spatial (geographical) extent, and the temporal extent (period) and the spectral bands used.

[1] https://openeo.org compliant earth observation backend.

We created a normalized query by sorting the filter arguments alphabetically since the order of appearance is not relevant. We identified the result by the hash value of the resulting stable sorted file list. We then introduced a Query Store in the database of the EODC openEO driver, which also happens to be a PostgreSQL database.

Therefore, we created a new query table, whereas every record represents a unique query and result tuple identifiable by a persistent identifier (PID). Giving the users of openEO the possibility to use this PID, we extended the openEO API to read metadata of a past dataset and reuse it in new processing chains. The implementation means that each request through the openEO driver of EODC stores the input data in the database and makes it accessible in the future, without having the users to change the way they work.

Lessons learned

When we started to think about implementing data citation at the EODC backend, we figured it would be difficult to achieve and maintain it. The successful implementation of the RDA recommendations on data citation gave us a structured plan to make it. The lessons learned here are that, by using the RDA recommendations, the implementation was easier and faster than estimated in the beginning. It made us wonder why we didn't think about implementing them earlier.

Related publications and presentations

Publication:

Bernhard Gößwein, Tomasz Miksa, Andreas Rauber, Wolfgang Wagner. Data Identification and Process Monitoring for Reproducible Earth Observation Research. IEEE eScience 2019, San Diego, USA. DOI: 10.1109/eScience.2019.00011

Presentation:

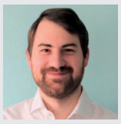
- eScience Conference 2019, San Diego – Oral Presentation: https://sched.co/UuUr
- RDA WGDC Webinar: Recordings and slides are available on the RDA website:
- https://www.rd-alliance.org/rda-wgdc-webinar-openeo-eodc-adoption

About the Earth Observation Data Centre for Water Resource Monitoring (EODC)

The EODC is located in Vienna and is an open international cooperation to foster the use of Earth Observation data. It connects to multi-petabyte storage infrastructures and highly efficient computer clusters, the biggest one being the Vienna Scientific Cluster (VSC). The VSC is a pool of high- performance computing resources by five Universities in Vienna and is ranked #85 in November 2014 in the TOP 500 Supercomputing Sites. EODC uses this computing power for EO data provision, distribution, procurement, as well as management of processing.

Contacts

Bernhard Gößwein: bernhard.goesswein@tuwien.ac.at



EODC:

Official Contact: office@eodc.eu

Thomas Mistelbauer: thomas.mistelbauer@eodc.eu

Find out more at:

www.rd-alliance.org/recommendations-outputs

Visit rd-alliance.org

or write us at enquiries@rd-alliance.org