Dynamic Data Citation for frequently modifying High Resolution Climate Data

RESEARCH DATA ALLIANCE

Climate Change Centre Austria (CCCA) Data Centre adopts Research Data Alliance (RDA) Recommendation on Data Citation of Evolving Data

An RDA adoption story written by **Chris Schubert**, Geologist and Geoinformatics,
Head of CCCA – Data Centre, Coordinator for Austria of the Group on Earth Observation (GEO)

Reading time: 6 minutes

The Climate Change Centre Austria (CCCA) Data Centre expected a comprehensive project outcome of completely new simulated High Resolution Climate Scenarios for Austria in the time range from 1965 till 2100 on a daily basis. For consumption, 13 model runs, 5 meteorological parameters like temperature, 3 emission scenarios, over 1600 NetCDF files with an average size of 13 GB were calculated. How could we implement proper data management processes on such data packages? We were looking for best practices on persistent identifiers and sub-setting tools for such big data containers. By chance, I met members of the RDA Data Citation Working Group. The idea of using the RDA recommendation on dynamic data citation as a pilot "NetCDF Pilot Implementation of Climate Scenarios" was born.

> ***"High Resolution Climate Data modify frequently, due to their complex dependencies and statistical methods for downscaling. In order to re-use these data and services in a reproducible manner, to share and cite, data analysts and researchers need a possibility to identify the exact version used."***
>
> **Chris Schubert,** Head of CCCA-Data Centre

The challenge addressed

The technical challenges we had to overcome in our project revolved around performance issues for the sub-set services of the big NetCDF files. The real challenge we faced, however, was a social one: we had to make sure our user community accepted and trusted reference data for Climate Services we offered. This certainly was the case for our "small" user group of Climate Services in Austria for the re-use of climate scenarios. Beyond trustful services, the readiness for good scientific practice on sub-set tools with dynamic data citation present barriers.

The RDA outputs adopted

The CCCA Data Centre adopted the Recommendations of the RDA Working Group on Data Citation: Data Citation of Evolving Data (Rauber et al. 2015[1]). Identification of Reproducible Subsets for Data Citation (Rauber et.al. 2016)[2], was used for further enhancement. Both documents, with their 14 recommendations, act as a kind of instruction manual and reflect almost a real implementation specification. **The RDA Working Group thus relieved us of the intellectual and conceptual work required. A further, and much appreciated advantage, is that these recommendations are based on the considerations of international experts.**

Find out more at:

www.rd-alliance.org/recommendations-outputs

Visit rd-alliance.org or write us at enquiries@rd-alliance.org

RDA added value for CCCA

I have been an RDA member since 2014 and have used RDA for my own investigation on recent activities and best practices. RDA offers channels like the Working Groups, webinars, common discussions and possibilities to present our own approaches and technical solutions. Consequently, we found the technical state of the art solution for data citation needed in our CCCA High Resolution Climate Scenarios for Austria pilot project in the RDA Data Citation recommendation. The RDA WG provided us with technical assistance and best practices, furthermore RDA raises awareness on the importance of data management principles and aspects to a data provenance techniques. Recommendations are real-life demands, but of course from a data management perspective. The demand is rapidly increasing and becomes "real" for scientist because the legal directives of research founder is currently changing and data citation with the alignment of persistent resources becomes more and more valuable.

The adoption process

In adopting the RDA recommendations on Data Citation, we extended our CKAN- based data management system by developing Open Source plugins for Data Versioning. This was to record an explicit history and relationships between the subsets created and the versions these were based on. The Query Store, which stores all arguments, was implemented. The subsetting uses coordinates for the geographical bounding box, time ranges and variables such as temperature, humidity, etc. that refer directly to the original NetCDF arrays. The data subsets are being preserved with the original resolution and accuracy. These subset requests are stored in a Query Store, in conjunction with their persistent identifiers (PIDs). Additional functionalities like the Query Library for re-using Subset arguments and Subset verification were developed and the checksum method was adapted to verify correctness. The CCCA Software ecosystem used for the Dynamic Data Citation Tool consists of the following main components:
Web server
Application server (CKAN) for access, data management and used as query store
Handle.NET® Registry Server for PID allocation
Unidata Thredds Data Server (TDS).
Only 0.5 FTE of our small CCCA Data Center Software Development Team was responsible for carrying out the operational services. We continuously work on minor improvements, visualizations, data collection, and predefined areas by political names. A big step forward for us was the incorporation of Global Atmospheric datasets, where more than one variable such as the 4th dimension was used for the subsetting process.

The impact of the adoption

With the operational application for Dynamic Data Citation the data becomes significantly more attractive for data analysts. The user gets a dynamic generated citation text, which contains the original author, label of the dataset, versions, selected and applied subsetting parameters as well the alignment to the persistent identifier. For a new created and published subset, all metadata are inherited from the original ones and supplemented by the defined arguments, like the adapted bounding box, observed parameter and the name of the subset creator.
If we had not adopted the Dynamic Data Citation Sub-Set Service, our users would be forced to download data themselves and thus create an unintended first disruptive point against data provenance information. Data would still, for example, be prepared by selecting the area of interest and time range on the user's desktop computer. Dynamic data citation clearly increases the handling of data quality through redraw-able corrections and improvements.

Lessons Learned

"With this application, the CCCA Data Centre will strengthen the potential and attractiveness for researchers by lowering the barriers for the re-use of data by citing the author of the original data and using our data hub according to good scientific practice. Adopting Dynamic Data Citation as one of our Services improved our collaboration and knowledge sharing."

Chris Schubert, Head of CCCA-Data Centre

Contacts here

The CCCA-Data Centre Software components are Open Source and published on GitHub[3].

Related publications and presentations

Publications:

- *Schubert C., Bamberger H. (2019) Handling Continuous Streams for Meteorological Mapping. In: Döllner J., Jobst M., Schmitz P. (eds) Service-Oriented Mapping. Lecture Notes in Geoinformation and Cartography. Springer, Cham; https://doi.org/10.1007/978-3-319-72434-8_13*

Presentations:

- *EuroGEOSS workshop 2018, Geneva – Poster Summaries; https://ec.europa.eu/easme/sites/easme-site/files/poster_summaries.pdf*
- *EGU2018, Vienna - Oral Presentation, https://meetingorganizer.copernicus.org/EGU2018/EGU2018-17117.pdf*
- *RDA 11th Plenary 2018 Berlin - Oral Presentation, RDA WG Dynamic Data Citation*
- *EODC (Earth Observation Data Centre for Water Resources Monitoring) Forum 2018, Vienna - Oral Presentation, Dynamic Data Citation on Climate Scenarios and Global Radio Occultation Data*
- *Data Stewardship Realized: From Planning to Action. Towards the Establishment of an Research Infrastructure, 2017 Vienna*
- *RDA WGDC Webinar: first release. Recordings and slides are available on the RDA website: https://www.rd-alliance.org/automatically-generating-citation-text-queries-recommendation-10-rda-data-citation-wg-webinar*

About the Climate Change Center Austria (CCCA)

The Climate Change Centre Austria (CCCA) is a research network supported by Austria's most important research institutions. It promotes climate and climate impact research and fosters collaboration in and among those domains. The CCCA Data Centre as an operational department is responsible for the research data infrastructure to promote applicable data management for Austrian researchers and the Greater Alpine Region. Our main objective is to provide a central climate data hub for models, climate scenarios, related research data and information services. The CCCA Data Centre fosters data sharing principles and standardized data services for download, view and data analyzing tools.

Endnotes

1 *https://rd-alliance.org/system/files/RDA-DC-Recommendations_151020.pdf*

2 *Rauber et.al. (2016) Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use, http://www.ieee-tcdl.org/Bulletin/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf*

3 *https://github.com/ccca-dc*