


Article

Dynamic Data Citation Service—Subset Tool for Operational Data Management

Chris Schubert ^{1,*} , Georg Seyerl ¹ and Katharina Sack ²¹ Data Centre—Climate Change Centre Austria, 1190 Vienna, Austria² Institute for Economic Policy and Industrial Economics, WU—Vienna University of Economics and Business, 1020 Vienna, Austria* Correspondence: chris.schubert@ccca.ac.at

Received: 31 May 2019; Accepted: 30 July 2019; Published: 1 August 2019



Abstract: In earth observation and climatological sciences, data and their data services grow on a daily basis in a large spatial extent due to the high coverage rate of satellite sensors, model calculations, but also by continuous meteorological in situ observations. In order to reuse such data, especially data fragments as well as their data services in a collaborative and reproducible manner by citing the origin source, data analysts, e.g., researchers or impact modelers, need a possibility to identify the exact version, precise time information, parameter, and names of the dataset used. A manual process would make the citation of data fragments as a subset of an entire dataset rather complex and imprecise to obtain. Data in climate research are in most cases multidimensional, structured grid data that can change partially over time. The citation of such evolving content requires the approach of “dynamic data citation”. The applied approach is based on associating queries with persistent identifiers. These queries contain the subsetting parameters, e.g., the spatial coordinates of the desired study area or the time frame with a start and end date, which are automatically included in the metadata of the newly generated subset and thus represent the information about the data history, the data provenance, which has to be established in data repository ecosystems. The Research Data Alliance Data Citation Working Group (RDA Data Citation WG) summarized the scientific status quo as well as the state of the art from existing citation and data management concepts and developed the scalable dynamic data citation methodology of evolving data. The Data Centre at the Climate Change Centre Austria (CCCA) has implemented the given recommendations and offers since 2017 an operational service on dynamic data citation on climate scenario data. With the consciousness that the objective of this topic brings a lot of dependencies on bibliographic citation research which is still under discussion, the CCCA service on Dynamic Data Citation focused on the climate domain specific issues, like characteristics of data, formats, software environment, and usage behavior. The current effort beyond spreading made experiences will be the scalability of the implementation, e.g., towards the potential of an Open Data Cube solution.

Keywords: dynamic data citation; subset; data curation; persistent identifier; data provenance; metadata; versioning; query store; data sharing; FAIR principles

1. Summary

Data with a spatial reference, so-called geospatial data, e.g., on land use, demographic statistics, geology, or air quality, are made accessible by interoperable and standardized web services. This means that data, whether stored in a database or file-based systems, are transformed into Open Geospatial Consortium (OGC) [1] conformal data services. These include catalog services for searching and identifying data via their metadata, view services for visualizing information in the internet browser, and more comprehensive services such as the Web Feature and Web Coverage Service (WCS) [2], which

allow access to more complex data structures, such as multidimensional parameters. Data and data services are becoming more sophisticated, more dynamic, and more complex due to their fine-grained information and consume more and more storage space. The high dynamics of the content offered can be explained by data updates, which take place at less and less frequent intervals, and the increasing number of new available sensors.

According to the objective and strategy of GEO—the Group on Earth Observation [3]—even more people, not only scientific domain experts, get access to climate, earth observation and in situ measures to extract information on their own.

Due to increasing interoperable and technologically “simplified” data access, the citation of newly created data derivatives and their data sources becomes essential for data analyses, such as the intersection of different data sources. The description of entire process chains with regard to information extraction, including the methods and algorithms applied, will become essential in the practice of data reproducibility [4]. In order to obtain this information in a structured system, the concept of data provenance [5,6] was defined, which describes the sequence of how data were generated.

It is common practice that behavior related to data usage goes away from downloading and using desktop tools to web-based analysis. The Open Data Cube (ODC) [7,8] as an open source framework for geospatial data management and effective web based data analysis for earth observation data. There is a growing number of implementations of ODC on national and regional level. Therefore, precise citation processes [9] should be considered in available data infrastructures.

For proper data management, data citation and evidence as robust information of data provenance in relation to the core principles on data curation [10–12] will be relevant. Each data object should be citable, referenceable, and verifiable regarding its creators, the exact file name, from which repositories it originates from, as well as the last access time.

The requirements [13] for citation of data should take into account: (i) the precise identification and time stamp of access to data, (ii) the persistence of data, and (iii) the provision of persistent identifiers and interoperable metadata schemes that reflect the completeness of the source information. These are the basic pillars of data citation, reflected in the Joint Declaration of Data Citation Principles [10] and the FAIR (Findable, Accessible, Interoperable, Reusable on Data Sharing) Principles [13].

These were considered in the Research Data Alliance Data Citation Working Group (RDA Data Citation WG) [14,15] and summarized as 14 recommendations of the document “Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation” (WGDC) [9,10]. This outcome forms the basis for the concept of dynamic data citation. Nevertheless, there are still barriers within the sophisticated offer on huge widespread characteristics on syntactical data formats and scientific domain issues. The Earth Observation domain is handling data curation in different principles than the climate model domain. Stockhause et al. [16] give a detailed overview of the evolving data characteristics and compare the different approaches.

As a recently established research data infrastructure, the Data Centre at the Climate Change Centre Austria (CCCA) started with a dynamic data citation pilot concept focused on NetCDF Data for the RDA working group in 2016 and implemented completely the recommendation so that since 2017, operational service can be provided for regional and global atmospheric datasets.

The current development efforts are scaling up techniques with the aim to extend our coverage on existing services especially towards the objective to cover the requirements of scientific domain on Open Earth Observation and the Open Data Cube environment and to offer the technical approach as an extension for the domain.

The overall objective of this article was to demonstrate the technical implementation and to provide the future potential of benefits regarding the RDA recommendations, with operational service offered as evidence, such as sustainable storage consumption using the Query Store for the data subset, and automatic adaptation into interoperable metadata description to keep the data provenance information.

2. Introduction on Dynamic Data Citation

Citing datasets in an appropriate manner is agreed upon as good scientific praxis and well established. Data citation as a collection of text snippets provides information about the creator of the data, the title, the version, the repository, a time stamp, and a persistent identifier (PID) for persistent data access. These citation principles can easily be applied in a data repository to static data. If only a fragment of a dataset is requested, which is served by subset functionalities, a more or less dynamic citation is required [9]. The consideration is to identify exactly those parts, subsets, of the data that are actually needed for research studies or reports, even if the original data source evolves with new versions, e.g., by corrections or revisions.

With data-driven web services, the data used are not always static, especially in collaborative iteration and creation cycles [14]. This is particularly valid for climatological research, where different data sources and models serve as input for new data as derivatives, e.g., climate indices like calculation of the number of tropical nights, which is based on different climate model ensembles. From a data quality point of view, it is preferable that such derivatives also be affected and updated automatically by the performed correction chain. Such changes in consideration on dependencies in data creation should be communicated as automatically as possible. A research data infrastructure should be able to provide an environment for dynamic data. With the reproducibility of results in mind, it is essential to be able to accurately verify a particular dataset, its exact version, or the creation of data fragments. The reproducibility of the data fragments and their relationship to their originals is essential if data processing has to be repeated.

Creating subsets is a common procedure for setting up needed customized data extraction for experiments or studies. Either only specific areas of interest or only a certain time interval are needed, but also particular information layers, such as the distribution of the mean surface temperature, can be of interest for a further effective processing. However, it is also a known fact that the storage of subsets created cannot scale with increasing amounts of data [8]. This implies that subsets are always copies of the original, and redundant storage consumption is not an economical option for capacity reasons (storage costs). The objective on considerations of the RDA—WGDC is to store only the criteria that create the subsets as arguments in a query store. In general, these are only few kilobytes compared to mega- to gigabytes with a subset of, e.g., Austrian climate scenarios. Such a query can be executed again, and the subset will be created on demand for a needed use.

To ensure that the stored queries are available for long-term use, to be executed again, and created subsets are available to other users, they are assigned to unique persistence identifiers and verification techniques. These are the core concepts of the RDA recommendations on dynamic data citation.

With such implementation, an operator of data infrastructures or service provider has to allocate only temporary storage for access to a subset. For the aforementioned OGC-compliant web services, the storage plays a minor role too, as the mechanisms for the provision of data fragments are very similar to subset services, such as browser-controlled zooming in by controlling the bounding box parameters. However, for such web applications, the RDA recommendations provide the targeted added value that queries are provided by a persistent identifier and thus enable delivering information about the data origin, which is reflected in the inheritance and adaptation of metadata to newly generated data fragments.

The 14 RDA recommendations for the creation of reproducible subsets in a context of easy and precise identification for dynamic data is a very demanding but pragmatic guidance. The RDA—Recommendations for the Scalable Dynamic Data Citation Methodology serves as a guideline with technical requirements for implementation, which are underpinned with practical examples in an understandable manner.

The 14 RDA Recommendation on Dynamic Data Citation

The recommendations for creating reproducible subsets reflecting the results of expert discussion, which served as a guideline on how to identify dynamic subsets from existing data sources. Short core messages on each recommendation are given, based on Rauber et al. [14,15].

Four pillars for structuring the recommendations were identified, see Figure 1:

- Framework on preparing the data and a query store;
- Identifying specific data in a persistent manner;
- Resolving PID and retrieving data; and
- Guaranteeing modifications and adaptability for data infrastructures as well as changes in software environments.

R14 - Migration Verification									
R13 - Technology Migration									
R1 - Data Versioning	R2 - Event Time-stamping	Query			Result Set		Landing Page		
		R4 - Unique Queries	R7 - Query Time-stamping	R8 - Query PID	R5 - Stable Sorting	R6 - Result Verification	R10 - Citation Text	R11 - Human Readable	R12 - Machine Actionable
		R9 - Query Metadata							
Data Store		R3 - Query Store							

Figure 1. A structured order for the Research Data Alliance (RDA) recommendation on dynamic data citation.

The recommendations in detail are summarized and adapted according to the implementation at the CCCA Data Centre as listed below. More information equipped with practical examples can be found in Rauber et al. [10].

R1—Data Versioning: Versioning ensures the former states of available datasets which can be retrieved. This information about this version is described within the metadata and the URI, which directs to the query store.

R2—Timestamping: Ensuring that all operations on data get timestamps is part of each data repository or database. The timestamp is provided in metadata.

R3—Query Store Facilities: Enabling a query store is an essential building block, with queries and associated metadata in order to enable re-execution in the future. The [UNI DATA] subset service (NCSS) provides a catalogue of subset arguments which are prepared in URIs.

R4—Query Uniqueness: Detecting identical queries and its arguments, e.g., by a normalized form and its comparison.

R5—Stable Sorting: Ensuring a stable sorting of the records in the dataset is unambiguous and reproducible. Executed queries are available in a query library, and if R4—Query Uniqueness response is a positive true result, the user has to apply still existing ones.

R6—Result Set Verification: Computing a kind of checksum generates a hash key as fixity information on the query result to ensure the verification of the correctness of re-execution. The check sum algorithm runs on each created subset and its execution.

R7—Query Time-stamping: A timestamp is assigned to the query, based on the last update to the entire database.

- R8—Query PID: Each new query with a purpose of republishing is assigned a new handle identifier as a PID.
- R9—Store the Query: Storing query and all related arguments, e.g., check-sum, timestamp, superset PID, and relation, based on R3—Query Store Facilities.
- R10—Automated Citation Texts: Generating citation texts based on snippets of authors, title, date, version and repository information. It lowers the barrier for citing and sharing the data.
- R11—Landing Page: PIDs resolve to a human readable landing page that provides the data and metadata, including the relation to the superset (PID of the data source) and citation text snippet. The metadata are held in DCAT-AP Schema, adapted by the European Commission [17]
- R12—Machine Actionability: Providing an API/machine actionable interface to access metadata and data via the provided ckan API. The query re-execution creates a new download link which is available for 72 h.
- R13—Technology Migration: When data are migrated to a new infrastructure environment (e.g., new database system), ensuring the migration of queries and associated fixity information.
- R14—Migration Verification: Verify successful data and query migration, ensuring that queries can be re-executed correctly.

3. Purpose of Implementation and Development Tasks

The CCCA—Data Centre operates a research data infrastructure for Austria with highly available server cluster, storage capacity, and linked to high-performance computing facilities of the Vienna Scientific Cluster and the Central Institute for Meteorology and Geodynamics (ZAMG), the national weather service. The main portfolio of CCCA Services is to enable a central access point of Austrian research institutions and the Greater Alpine Region for storing and distributing scientific data and information in an open and interoperable manner regarding FAIR principles.

The CCCA—Data Centre developed a web-based tool for dynamic data citation. The main motivation in 2015 was simply to have a technical solution to providing a persistent identifier and an automatically generated citation text. At this point, the issue of what happens with evolving data and its version concept arises. Consequently, this led to the incentive to provide proper components for an appropriate data lifecycle and assign a dynamically persistent identifier (PID) for all associated data derivatives. With the RDA recommendations, the approach of a query store was convincing, and an appropriate decision base to follow this concept on identifying uniquely queries which can be executed again when needed was created. With the CCCA—Data Centre’s task to provide large file sizes like climate scenarios, the argumentation to reduce redundancies for the storage consumption was the most convincing argument for the planned implementation at this time.

In cooperation with the Data Citation Working Group, a concept for a technical pilot implementation was accompanied.

This pilot implementation on Dynamic Data Citation at the CCCA Data Centre focused on CF standard [18] compliant NetCDF data to manage high-resolution climate scenarios for Austria in a time range from 1965 until 2100. NetCDF is an open standard and machine-independent data format for structured and multidimensional data. It includes attributes, dimensions, and variables. For example, for the Austrian Climate Scenarios, calculated temperature records on daily basis are available in 1×1 km gridded, geo-referenced data in multiple single files. The scenarios include different “representative concentration pathways” (RCPs) [19], ensembles of different GCM (general circulation models) and RCM (regional climate model) runs, for high-resolution conclusions, which are combined with statistical methods for the integration of in situ observations. The open accessible entire data package includes, for Austria, over 1200 files with a size up to 16 GB per file. Due to user requirements, in particular for the development of data-driven climate services and the characteristics of the climate scenarios provided, a subset service, Figure 2, was required.

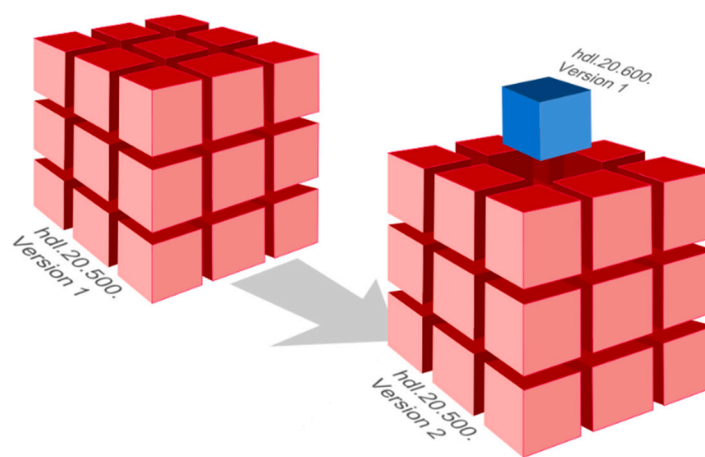


Figure 2. Schematic draft of subset needs, which includes the control on versioning and the alignment with the persistent identifier (PID), here handle.NET identifier—hdl. For the fragmented subset (blue cube), a new identifier is aligned, coupled with its own version number.

Especially for such large files, the first argument is decreasing the download rate, and the second is again storing the subsets on a desktop workstation. The continuous process chain on data fragments will be broken. Normally, GIS or data analytic tools are used to intersect the individual ‘area of interest’ or choosing a separate, distinguished layer or simply selecting a given time frame is still a common behavior. In a case of republishing, to offer a reuse or reproducibility study, all metadata and siblings’ relation to the origin and different version would be lost and have to be described again. To do this manually is time-consuming, while describing the processes with all arguments for the intersection procedure will be imprecise. The CCCA Data Centre wants to overcome these troublesome processes mostly related with complex data structures especially for the climate services.

The overall approach on the CCCA-DC software environment was to set up a system which follows open source licenses. All developments and modules are available on the CCCA GitHub [20]. The data in the storage system, which are embedded in a highly available Linux Server Cluster, are managed by the ckan [21] software packages as a Python application server. This collaborative development framework is specialized in data management and catalogue systems, which is used as a central system component. For ckan, many extensions especially for the geospatial scientific domain are available, which brings a lot of synergies and benefits in its own modular software developments. One essential component for a provided catalogue of services is the flexible metadata scheme functionality. The Data Catalog Vocabulary DCAT [22] as a ckan default metadata profile was extended by the DCAT-AP and GeoDCAT-AP [23], a development by the Joint Research Centre of European Commission, which meets interoperable requirements for data exchange between distributed data servers. With this solution, heterogeneous data formats can be described with a common core schema for metadata and enable a uniform transformation into other profiles, such as Dublin Core, INSPIRE, and ISO 19115 metadata for geographical information.

The graphical user interface of the CCCA data server is based on the ckan web server and includes all functionalities, such as catalog and search functions, view services for web-based visualization of data content, as well as the implemented subset service. A Python API interface is also provided via ckan, which enables machine-to-machine communication for automatically steered processes.

For the unique identification of a data object, persistent identifiers (PIDs) are used, see Figure 2, and its registry guarantees uniqueness according to the specifications of internet identifiers to other data objects. For the CCCA, the Handle.NET® Registry Server was used for PID assignment. The advantage of Handle is the unlimited and instant assignment of identifiers, the technical coherence on standards, and encoding, which is essential for each newly created query.

The primary component for processing and creating data fragments is the Unidata Thredds Data Server (TDS) [24]. This server is responsible for processing NetCDF data, such as visualizing the data. In addition to TDS, the NetCDF Subset Services (NCSS) was embedded. NCSS provides a catalog of subsetting parameters that allows creating data fragments while retaining the original resolution and characteristics of the original data. These parameters include geographic coordinates, date ranges, and multidimensional variables. NCSS uses “HTTP GET” [25] in the following structure:

`http://{host}/{context}/{service}/{dataset}/{dataset.xml | /dataset.html | {?query}}`

where elements proposed as:

{host}—server name

{context}—“thredds” (usually)

{service}—“ncss” (always)

{dataset}—logical path for the dataset, obtained from the catalog

dataset.xml—to get the dataset description in xml

dataset.html—to get the human-readable web form

datasetBoundaries.xml—to get a human-readable description of the bounding boxes

{?query}—to describe the subset that you want.

The subsetting parameters for the element {?query} allow a combination of different parameters, like the name of variables, the location points or bounding box, arguments which specify a time range, the vertical levels, and the returned format.

Figure 3 illustrates the implemented components and gives an overview about the relationships between requests (blue arrows) and responses (orange arrows) between the server. The application server takes the requests via the Web server and generates URL-based (HTTP GET) requests with the subsetting parameters (subset requests). These requests are stored in the query store and are assigned with the Handle identifier.

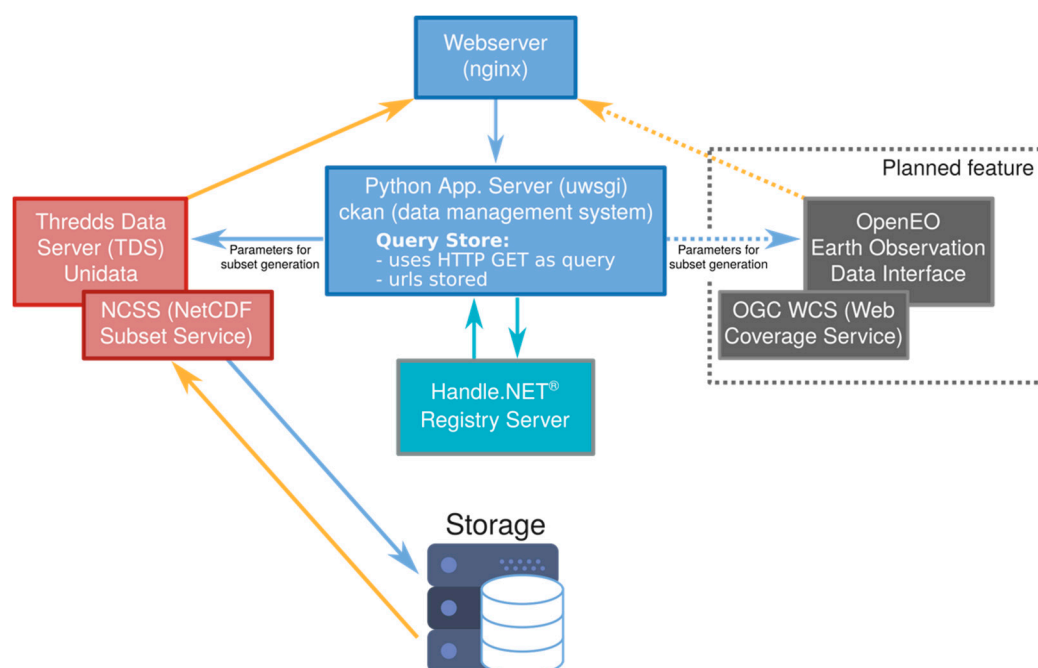


Figure 3. Simplified structure of server and hardware components for dynamic data citation within the CCCA Data Centre environment: (i) ckan web server, (ii) the application server for access, data management used as query store, (iii) Handle.NET® Registry Server for PID allocation, and (iv) the Unidata Thredds Data Server (TDS), NCSS Subset Service and planned features on Open EO support.

Within the ckan data management system, the required meta information for the subset dataset is compiled from the original meta data via adaptation and inheritance and tagged with the necessary description of the relationship as well as versions as supplementary meta data elements. The metadata of the newly created data subset also contain the original metadata elements, such as a short description, the data creator, licenses, etc. The supplementary elements are based on the query arguments and the meta information from the application server, which are automatically adapted. These are the title of the subsets, the selected parameters, the new spatial extent, and the changed time interval. In addition, there is the contact of the subset creator, the time of creation, the check-sum to verify if it is the same result if the request is repeated, the file size, and then the relationship to other records and their version.

The Thredds server retrieves the defined arguments from the query store via NCCS and thus creates the subset directly from the data store in which the original NetCDF data are contained. The data format is again NetCDF; other formats like comma-separated values (CSV) are also supported and return them to the web server. There, the subset is available as a resource for download, but also as a view service (OGC-WMS) for web-based visualization.

4. User Interface of the Application on Dynamic Citation Service

The Subset and Dynamic Data Citation Service at the CCCA Data Server is accessible for everyone. Due to performance reasons via Thredds, only registered users get access, Figure 4, for the comprehensible functionality on defining and republishing the subset at the data server.

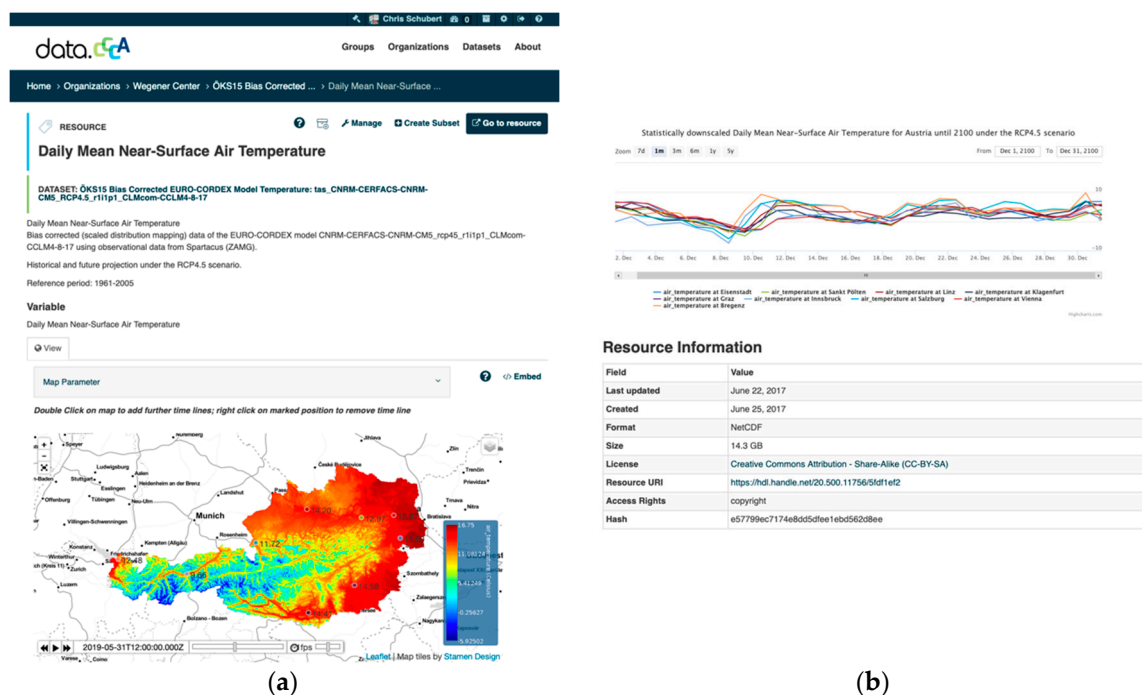


Figure 4. The general landing page of a data resource, after the personalized login: the general landing page of a dataset resource after login, where the subset can be created (on top): (a) The visualization is a view service (WMS), created by Thredds, and it allows the user by activating the time control to visualize each time step up to 2100; (b) additionally, it shows a timeline diagram after a point of interest on the map window is created.

After creating the subset, Figure 5, the user immediately receives a dynamically generated citation text containing the original author, the name of the subset, version, selected parameters, and the persistent identifier. This citation proposal can be used for correct reference in studies, publications, etc. and is clearly assignable to the entire research community. For a newly created and published subset, all metadata are inherited from the original data and supplemented by the defined arguments,

such as the customized bounding box and the name of the creator, as well as the relation as a first step for data provenance information.

Create Subset

Select Layers/Parameters
Daily Mean Near-Surface Air Temperature

Reuse Query
Select Query Template

Choose Geographical Extent

Choose Time Range
From: 2019/04/07 14:17 To: 2024/05/31 14:17

Would you like to create a new resource within the CCCA data portal?
No, just download the subset
Yes

Dataset Title
eg. A descriptive title
* URL: data.ccca.ac.at/dataset/<name> Edit

Organization
CCCA Data Centre

Resource Name
subset_Daily Mean Nea

Should the resource be quotable? (dataset becomes public and cannot be set private again)
Yes No, maybe at a later point

Submit

Figure 5. GUI of the subset creation function: (a) The upper part of web page for defining the parameter, or reuse of a still existing query, defining a bounding box either by polygon or predefined administrative units, (b) allows choosing a time range for other datasets like the globally available radio occultation data packages, a fourth dimension—e.g., the Potential High was introduced and can choose.

Versioning is used to ensure that previous states of records are maintained and made retrievable. Being able to refer to previous versions of datasets is important for reproducibility of simulation, calculations, and methods in general. The given Handle PID resolves into the landing page of the subset resource, where detailed metadata are provided. The web application generates automated citation texts. It includes predefined text snippets like the title, author, publishing date, version, and the data repository. For subsets, the aforementioned filter arguments based on queries were used and provided as text information, see Figure 6. The generated citation texts are in a form that lowers barriers for data sharing and reusability with proper credits.

Variable
Daily Minimum Near-Surface Air Temperature

Dataset Versions Citation

Dataset Versions:
This Version
Version 1 Release Date: 2017-06-22 10:24:54.504515
Latest Version
Version 1 Release Date: 2017-06-22 10:24:54.504515

Cite this dataset:
Using this data set or resource, you should cite this data set according to the given copyright conditions with following citation rules:
Lauprecht et al (2017). Test dynamic data citation TU Delft. Version 1. Vienna, Austria. CCCA Data Centre.
PID: <https://hdl.handle.net/20.500.11756/08399c90>. [May 31, 2019] Copy Text

Subset
This dataset is a subset of "OKS15 Bias Corrected EURO-CORDEX Model Temperature: tn_CNRM-CERFACS-CNRM-CMS_RCP4.5_r11ip1_CLMcom-CCLM4-8-17" Show relations

Original Version	Release Date	Subset Version
Version 2	2017-06-22 10:24:54.504515	Test dynamic data citation TU Delft (Version 1)
Version 1	2016-12-28 14:05:00	Create

Resources
subset_Daily Minimum Near-Surface Air Temperature Explore

Figure 6. The screenshot gives an impression of what versions, relations, and the suggested text for citation looks like. In addition, the user could create, with the same arguments, a subset based on oldest versions but normally on a new version published. If new versions are available, a notification will be sent to the subset creator, which is part of the metadata profile.

5. Discussion and Next Steps

The implementation of the CCCA Data Centre's Dynamic Subsetting of evolving data shows its feasibility on a Pilot for NetCDF software and data processing environment. Nevertheless, limitations exist and can be seen both in the particular scope of the data format and in the lack of hardware configurations that enable interfaces and connectivity to other data infrastructures. The given requirements for CCCA data only lie in the CF conformity. Thus, all described functionalities are automatically available to the data providers. Due to the performance of the NetCDF format, the system independence and the multidimensional structured description of geospatial content, this format is used as an ingest and transfer format for the Open Data Cube. Integrated Python libraries allow a seamless transformation of data formats that are commonly used in the Earth Observation sector, such as GeoTIFF. Open Data Cube is a Python-based software framework that allows analyzing and processing the entire data package as a Data Cube to generate new earth observation products and services. Further considerations for the described dynamic citation implementation consist of setting up the data management software components with regard to the linkage with PID and the automated extraction of metadata on local Open Data Cube implementation in order to apply exactly this gap of the dynamic data citation within the Data Cubes. A first showcase within the framework of the Austrian Data Cube in cooperation with the Vienna University of Technology and the EODC—Earth Observation Data Center in Austria is currently in the conception phase, see Figure 3.

Another potential field of application is seen in the direction of OGC-compliant Web Services. The focus of these techniques is more on the interoperable web-based provision of data. The Web Coverage Service (WCS) describes the effective handling of subset generation and data fragments for effective further processing. The aspect to the requirements in the direction of dynamic data citation is taken into account but is not implemented so consistently in data infrastructures. This gap is not the aim of OGC standards themselves, but data infrastructure operators as well as their users should be guided towards these needs.

With this demonstrated implementation, an effort is undoubtedly made from a technical as well as a development as well as maintenance cost perspective. The big advantage, which can be shown here, is the avoidance of redundancies for storage consumption of generated subsets, whether locally or via cloud storage systems, and the exact citation to such individually created subsets so that they can be made accessible for other users.

The considered reflection and implementation regrettably go only in one direction, that is, from a dataset to its own data fragment. Inheriting the meta-information from an original to its subsets is not a dialectical challenge. What needs to come next in data curation and data management science is a method for how to deal with grouping of data ensembles and the merging of meta-information and contrary metadata elements.

6. Conclusions

The citation of data, which are mostly static, serves the description of the origin, the credits on authorship, and a link for accessing and downloading an entire dataset. In many research environments, data grow dynamically and through updating, which is a challenge for research data repositories. New versions can be created continuously through corrections; this can be done regularly, for example, on a monthly basis, but also quite agilely at irregular intervals and helps to improve data quality.

When data are used as the basis for a study or calculation, it can be ensured that the exact data version is available for verification in a study. This is especially the case for data derivatives where new algorithms are applied to the original data at a given point in time, e.g., the calculation of climate indices based on different climate models. The citation of the data should make it possible to identify the data fragment in a reliable and efficient process for all aspects of reproducibility of research and published studies.

The RDA recommendations of the Working Group on Data Citation (WGDC) enable researchers and data infrastructures to identify and cite data they are using. The recommendations support a

dynamic, query-centric view of the data and enable precise identification by associating the queries to the subsets that are generated.

The Subset and Dynamic Data Citation Service of the CCCA was one of the first operational adaptations of the RDA Citation Working Group recommendations. This implementation is also listed as an RDA Adoption Story [26] as a factsheet, which also contains some useful information about the development effort required for implementation and acceptance.

This ongoing operational service for subset creation and dynamic data citation is evidence of the applicable approach of the RDA Recommendation.

Nevertheless, the observation of user behavior shows that there are still obstacles to republishing the created subsets on the CCCA server. Reasons for this behavior could be the minor number of users in Austria, especially for the climate scenario scope. In order to expand the user community, the implemented subset service was applied to datasets with a global 5-dimensional atmospheric dataset. An extension was also made by providing climate scenarios for the Western Balkan region in Europe, where institutions, such as their national weather services, can create their scenarios covering the national territories as subsets.

The additional strategy for expanding the user community is to extend the service to the scientific field of satellite-based Earth observation, such as through the Open EO approach and the Open Data Cube environment. The RDA is supporting this planned activity at CCCA through the RDA Adoption Grant Program for the next 12 months.

With the present implementation of the dynamic data citation of evolving data, the feasibility is given on the one hand, while on the other hand, experiences as well as software developments can be passed on in order to obtain a more exact estimation of efforts for future implementation for other data infrastructures in order to realize mechanisms for proper data management.

Author Contributions: All authors were involved in the conceptualization of this paper. Software used for subsetting and dynamic data citation at the CCCA Data Centre was developed by K.S. and G.S., the Server Configuration was done by G.S.

Funding: This CCCA Data Centre Infrastructure has received funding from the Austrian Federal Ministry of Education, Science and Research (bmbwf) under the Research Investment Programme (HRSM) within the project CCCA and GEOCLIM as well as in-kind contribution by ZAMG. The planned activities for the Earth Observation Data extension as option of RDA outputs will be funded by the “RDA Europe 4.0” Project. Contracts and agreements are under preparation at the time of publishing this article.

Acknowledgments: In addition to the developer team at the CCCA Data Centre, I am very grateful to Andreas Rauber from the Technical University of Vienna and Chair the RDA Citation Working Group for his constructive discussion and his curiosity regarding the climate science as well as the geospatial domain, providing concrete consultancy. In addition, I would like to express my passionate thanks to the RDA, the fruitful discussion on plenaries and presentation, but in particular Marieke Willems for the support and marketing actions by RDA.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. OGC. The Open Geospatial Consortium. Available online: <https://www.opengeospatial.org> (accessed on 29 July 2019).
2. Baumann, P. OGC Web Coverage Service (WCS) 2.1 Interface Standard—Core. 2018. Available online: <http://docs.opengeospatial.org/is/17-089r1/17-089r1.html> (accessed on 29 July 2019).
3. GEO Group on Earth Observation. GEO Strategic Plan 2016–2025: Implementing GEOSS. Available online: https://www.earthobservations.org/documents/GEO_Strategic_Plan_2016_2025_Implementing_GEOSS.pdf (accessed on 25 May 2019).
4. Craglia, M.; Nativi, S. Mind the Gap: Big Data vs. Interoperability and Reproducibility of Science. In *Earth Observation Open Science and Innovation*; Pierre-Philippe, M., Christoph, A., Eds.; Springer: Cham, Switzerland, 2018; pp. 121–141. [CrossRef]
5. Lawrence, B.; Jones, C.; Matthews, B.; Pepler, S.; Callaghan, S. Citation and peer review of data: Moving towards formal data publication. *Int. J. Digit. Curation* **2011**, *6*, 4–37. [CrossRef]

6. Pröll, S.; Rauber, A. Scalable data citation in dynamic, large databases: Model and reference implementation. In Proceedings of the 2013 IEEE International Conference on Big Data, Santa Clara, CA, USA, 6–9 October 2013; IEEE Press: Piscataway, NJ, USA, 2013; pp. 307–312.
7. Open Data Cube Initiative. Open Data Cube Whitepaper. Open Data Cube Partners. 2017. Available online: https://docs.wixstatic.com/ugd/f9d4ea_1aea90c5bb7149c8a730890c0f791496.pdf (accessed on 26 January 2018).
8. Sudmanns, M.; Tiede, D.; Lang, S.; Bergstedt, H.; Trost, G.; Augustin, H.; Baraldi, A.; Blaschke, T. Big Earth data: Disruptive changes in Earth observation data management and analysis. *Int. J. Digit. Earth* **2019**. [CrossRef]
9. Buneman, P.; Davidson, S.B.; Frew, J. Why data citation is a computational problem. *Commun. ACM* **2016**, *59*, 50–57. [CrossRef] [PubMed]
10. CODATA-ICSTI Task Group on Data Citation Standards and Practices. Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. *Data Sci. J.* **2013**, *12*, 1–67.
11. FORCE11 Data Citation Synthesis Group. *Joint Declaration of Data Citation Principles*; Martone, M., Ed.; FORCE11 Data Citation Synthesis Group: San Diego, CA, USA, 2014; Available online: <http://www.force11.org/datacitation> (accessed on 15 June 2019).
12. FORCE11 FAIR Data Publishing Group. FAIR Guiding Principles. 2017. Available online: <https://www.force11.org/fairprinciples> (accessed on 15 June 2019).
13. Silvello, G. Theory and Practice of Data Citation. *J. Assoc. Inf. Sci. Technol.* **2018**, *69*, 6–20. [CrossRef]
14. Rauber, A.; Asmi, A.; van Uytvanck, D.; Pröll, S. Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC). Result of the RDA Data Citation WG. 20 October 2015. Available online: http://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf (accessed on 25 May 2019).
15. Rauber, A.; Asmi, A.; van Uytvanck, D.; Pröll, S. Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use. *Bull. IEEE Tech. Committee Digit. Libr.* **2016**. Available online: http://www.ieee-tcdl.org/Bulletin/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf (accessed on 25 May 2019).
16. Stockhause, M.; Lautenschlager, M. CMIP6 Data Citation of Evolving Data. *Data Sci. J.* **2017**, *16*, 1–13. [CrossRef]
17. DCAT Application Profile for Data Portals in Europe (DCAT-AP). Available online: <https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe/releases;http://data.europa.eu/w21/c319d97f-19bf-4365-af0b-eaddcd3256293> (accessed on 29 July 2019).
18. CF Conventions and Metadata. Available online: <https://cfconventions.org/latest.html> (accessed on 29 July 2019).
19. Representative Concentration Pathways (RCP's). Available online: <https://www.aims.ucar.edu/docs/IPCC.meetingreport.final.pdf> (accessed on 25 May 2019).
20. CCCA Server Software. Available online: <https://github.com/ccca-dc> (accessed on 29 July 2019).
21. CKAN. Open Source Data Management System. Available online: <https://ckan.org> (accessed on 29 July 2019).
22. W3C. Data Catalog Vocabulary (DCAT). Available online: <http://www.w3.org/TR/2014/REC-vocab-dcat-20140116/> (accessed on 29 July 2019).
23. GeoDCAT-AP Is an Extension to the “DCAT Application Profile for European Data Portals” (DCAT-AP) for the Representation of Geographic Metadata. Available online: <https://inspire.ec.europa.eu/documents/geodcat-ap> (accessed on 15 June 2019).
24. Unidata Thredds Data Server. Available online: <https://www.unidata.ucar.edu/software/thredds/current/tds/> (accessed on 29 July 2019).
25. NetCDF Subset Service. Available online: <https://www.unidata.ucar.edu/software/tds/current/reference/NetcdfSubsetServiceReference.html> (accessed on 29 July 2019).
26. RDA Adoption Stories. Available online: <https://www.rd-alliance.org/dynamic-data-citation-frequently-modifying-high-resolution-climate-data> (accessed on 29 July 2019).

