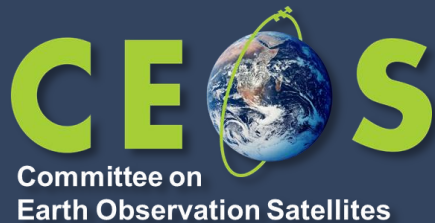


Data Citation White Paper discussion



I.Maggio, Starion for ESA
Agenda Item 8.15
WGISS-61
16-20 March 2026
IIRS/ISRO, India

WHAT IS THE DATA CITATION?



Acknowledgement
for the researchers
who produce the
object



Visibility for the
repositories that
host research
objects



Transparency &
reproducibility for
those using the
object



Evaluation of open outputs

Citing data properly is essential for ensuring transparency, reproducibility, and giving appropriate credit to data creators.

CORE COMPONENTS OF A DATA CITATION:

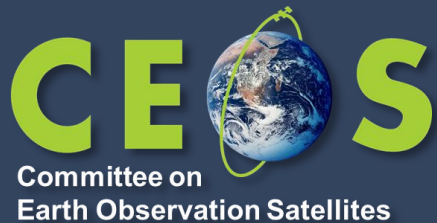
- ✓ Author: Name of each individual or organisational entity responsible for the creation of the data. This could also be the rights holder.
- ✓ Title of the data: Complete title of the data.
- ✓ Year of publication: Year the data was published or disseminated. This could also be the version of the data used.
- ✓ **Version**: Version or edition number associated with the data.
- ✓ Data publisher: Organisational entity (repository or archive) that makes the dataset available by preserving and/or disseminating the data.
- ✓ Persistent identifier: A unique identifier used to locate and access the data (such as a DOI). If available, the persistent identifier should be included as a link (e.g., <https://doi.org/xxxxxx>) and used instead of the URL or other Internet link.

CORE COMPONENTS SHOULD BE MANDATORY

ADDITIONAL COMPONENTS OF A DATA CITATION:

- ✓ Depending on the citation format requested by the guideline or style, additional components, such as the data number, the resource type, the place of publication, or the date of access, might be requested.

Data Citation Survey



First survey exercise – Metadata Required



Metadata Field	NASA (USA)	ESA (Europe)	JAXA (Japan)	ISRO (India)	CSA (Canada)	CNES (France)	DLR (Germany)	Common Standard (e.g. DataCite) 	Description
Dataset Identifier (DOI)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Required	Persistent identifier for the dataset
Dataset Title	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Required	Name of the dataset
Creator / Author	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Required	Person or organization responsible for dataset creation
Publisher / Repository	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Required	Organization distributing the dataset
Publication / Release Date	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Required	Date dataset was made publicly available

First survey exercise – Metadata Recommended



Metadata Field	NASA (USA)	ESA (Europe)	JAXA (Japan)	ISRO (India)	CSA (Canada)	CNES (France)	DLR (Germany)	Common Standard (e.g. DataCite) 	Description
Contributor	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Recommended	Additional contributors such as project members
Version	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Recommended	Dataset version number
Resource Type	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Recommended	Type of dataset such as imagery, model output, or product
Description / Abstract	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Recommended	Summary of dataset contents
Keywords / Subjects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Recommended	Thematic keywords for discovery
Spatial Coverage	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Recommended	Geographic region covered by dataset
Spatial Resolution	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Recommended	Spatial detail or grid resolution
Temporal Coverage	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Recommended	Time period represented in dataset
Temporal Resolution	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Recommended	Time interval between observations
Processing Level	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Recommended	Processing stage (e.g., Level-1, Level-2)
Data Format	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Recommended	File format such as NetCDF, HDF, GeoTIFF
Access URL / Landing Page	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Recommended	Webpage where dataset is accessed
Download Link	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Recommended	Direct link for dataset download
Access Date	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Recommended	Date dataset was accessed by user
License	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Recommended	Usage license for the dataset
Contact Information	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Recommended	Contact person or helpdesk
Metadata Creation Date	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Recommended	Date metadata record was created
Metadata Standard	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Recommended	Standard used (ISO 19115, DataCite, etc.)

First survey exercise – Metadata

Optional/Miscellaneous



Metadata Field	NASA (USA)	ESA (Europe)	JAXA (Japan)	ISRO (India)	CSA (Canada)	CNES (France)	DLR (Germany)	Common Standard (e.g. DataCite)	Description
Alternative Title	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Optional	Alternative dataset name
Language	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Optional	Language used in dataset documentation
Geographic Coordinates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Optional	Latitude/longitude bounding box
File Size	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Optional	Size of dataset files
Usage Rights	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Optional	Conditions for data reuse
Funding Information	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Optional	Funding organization or grant
Project Name	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Optional	Project under which dataset was created
Related Publications	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Optional	Research papers related to the dataset
Related Datasets	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Optional	Links to associated datasets

First survey exercise – Metadata Domain-Specific



Metadata Field	NASA (USA)	ESA (Europe)	JAXA (Japan)	ISRO (India)	CSA (Canada)	CNES (France)	DLR (Germany)	Common Standard (e.g. DataCite) 	Description
Instrument	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Domain-specific	Sensor or instrument used for data acquisition
Satellite / Mission	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Domain-specific	Space mission responsible for the data

Proposed Way Forward

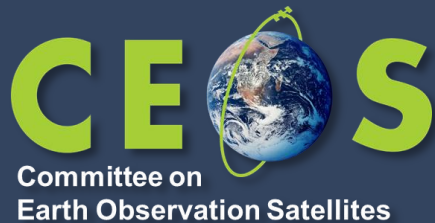


Number	Title	Completion Date	Description	Link to GEO Work Programme (if relevant)	External Reference (if relevant)	Responsible User(s) - Name-Surname	Responsible User(s) - Email address	Responsible CEOS Entity(ies)	Contributing Agency(ies)
DATA-25-04	EO Data Citation Guidelines	2026-12-31	These guidelines will provide information on EO data citation approaches and procedures			Mirko Albani	mirko.albani@esa.int	WGISS	ESA

- ❖ Additional Agencies to be added in the survey
- ❖ Propose (by June 2026) and Confirm the CORE COMPONENTS by October 2026 (WGISS-62 meeting)
- ❖ Propose (by June 2026) and Confirm ADDITIONAL COMPONENTS OF A DATA CITATION by October 2026 (WGISS-62 meeting)
- ❖ Draft and Circulate the EO Collection Data Citation Guideline by November 2026
- ❖ Formal issue of the EO Collection Data Citation Guideline December 2026

DISC#8	Resource metadata should contain the persistent identifier (e.g., DOI) of the corresponding resource.
--------	---

Persistent Identifier Use cases



I.Maggio, Starion for ESA
Agenda Item 8.15
WGISS-61
16-20 March 2026
IIRS/ISRO, India

Persistent Identifiers Best Practice

CEOS
Data Stewardship Interest Group

Doc. Ref.: CEOS/WGISS/DSIG/PIDBP
Date: March 2023
Issue: Version 1.4

Issue	Date	Comments	Editor
0.7	December 2014	First draft issue for CEOS-Data Stewardship interest group review	Tyler Christensen
0.8	March 2015	Second draft, incorporating some comments received during the CEOS-DSIG review	Tyler Christensen
1.0	March 2015	First version for public release	Tyler Christensen
1.1	January 2016	Added recommendation and use case scenarios for near real time products.	Tyler Christensen
1.2	January 2017	Added recommendations for the landing page. Updated applicable and reference documents.	Razvan Cosac
1.3	December 2019	Added new recommendations and use cases in line with received RIDs.	Razvan Cosac, Ville Saaristo
1.4	March 2023	Added new recommendations, use cases. Content refreshment and verification of links.	Sergio Folco, Iolanda Maggio, Bryan Keary

<https://ceos.org/ourwork/workinggroups/wgiss/documents/>

❖ B.1 Data User

- Citing Collection
- Accessing cited Collection

❖ B.2 Data Archive

- Adding data to a Collection
- Deleting a Collection
- Moving a Collection
- Distributed Collections
- Identifying additional information (PDSC – documentation, software)
- PID Granularity

❖ B.3 Data Producer

- Non-permanent products

- a. A user downloads a data set, uses it as part of their research, publishes the result, and cites the data source using a PID.

Recommendation: The PID should resolve to a landing page that provides access to the exact same data. The PID may link to the data set series from which the data set was taken. Citation guidelines should provide information on how to cite the individual data set used.

- b. A scientist uses and wants to cite two years' worth of data from a 20-year time series.

Question: Should “chunks” of a longer time series get their own PID?

Recommendation: The whole data set series (collection) gets one PID. The data provider should give guidelines on how to cite the data (e.g. use the PID but specify the time and area so that others can retrieve the same subset).

B.1 Data User – Accessing cited Datasets



- a. A scientist reads a published article, sees a data citation, and wants to access the same data set used in the original research.
Recommendation: The PID should resolve to a landing page that provides access to the exact same data.
- b. A scientist reads a published article, sees a data citation, and wants to access the same data set used in the original research. The data set is not available anymore.
Recommendation: The PID should resolve to a “tombstone” landing page that explains why the data are gone, provides a contact for more information, and – if possible – refers to an alternative / similar data set which is accessible.
- c. A scientist reads an article from 15 years ago with a data citation. The cited data set is still archived and accessible, but it has been replaced with a new one reprocessed and/or computed using an improved algorithm.
Recommendation: The landing page should mention the changes and provide links between the different PIDs of the data sets.
- d. A scientist reads an article from 15 years ago with a data citation. The original data set is no longer archived, because it has been replaced with a new product that uses an improved algorithm.
Recommendation: The landing page should explain why the data are gone and link to the improved product.

B.2 Data Archive – Adding data to a Collection



- a. A data archive wants to assign a PID, either to a new data set or to an existing archived data set.

Recommendation: Develop a consistent workflow for assigning PIDs, and for evaluating whether a data set should receive a PID. Follow best practices and commit to maintaining the data for the long term.

- b. The data set series in the archive is dynamic. New data sets or products are being added daily as the mission or project continues.

Recommendation: The entire data set series should receive one PID. Citation guidelines should be provided to help the data users define which time span of data they used, so that future studies could repeat the research exactly. A further approach is to close the data set periodically (for example each year) and to assign a new PID to the next data set.

- c. The data set series in the archive is dynamic. New data sets or products are being added daily as the mission or project continues. The archived data set items are reproduced with minor changes or are removed.

Recommendation: The data set series keeps the same PID if the change is minor one. Details of the dataset change is described to help users.

- d. The static or dynamic data set series in the archive is found to be incomplete. The missing data sets are being recovered from external sources and added to the data set.

Question: Does the PID remain the same?

Recommendation: The data set series should keep the same PID, as long as the data source and processing algorithms are the same. A version history in the metadata should be used to clear up any confusion.

- e. A data set series includes data from the same sensor, carried on many different satellites (example: AVHRR on the NOAA POES satellites). The current satellite is replaced by a new operational satellite to continue the measurement.

Question: Does the time series keep the same PID?

Recommendation: Yes, the data set series keeps the same PID if the content of the existing data series has not changed. Details on inter-calibration and processing coefficients should be in the metadata.

B.2 Data Archive – Deleting a Collection



- a. A data set series is deleted from the archive.

Recommendation: The PID should resolve to a “tombstone” landing page that explains why the data are gone.

- b. A level-1b product is not permanently archived. Instead, the product is generated on-demand when a user requests the data. It is then deleted. As a result of the processor, operating system, and hardware used, it cannot be guaranteed that a reprocessing will produce the exact same product down to the individual precise pixel value.

Question: Should this kind of on-demand product get a PID?

Recommendation: No, do not give an on-demand product a PID. Collections or products for which no permanent archiving is planned do not receive a PID. Consider assigning a PID to the lower level data from which the product is generated. Citation guidelines should specify how to cite the product. The relevant information (e.g. processor version) should be available in the (temporary) product's metadata.

- c. Some level-2 products within a data set series are found to be faulty. They are replaced with corrected versions, generated with the same processors as the rest of the data series. The faulty products are deleted.

Question: Should the collection receive a PID, even though its content will change? Should it receive a new PID when the faulty products have been deleted?

Recommendation: The data set series keeps the same PID. The PID should resolve to a landing page that explains which data sets out of the series were replaced and why.

B.2 Data Archive – Deleting a Collection



- d. Several duplicate products, with slightly different calibration values, are found within a data set series. After careful testing of both versions, the faulty ones are deleted from the data set series.

Question: Does the PID remain the same? Should the data set series receive a new PID when the faulty products have been deleted? What if the faulty products have been used and cited?

Recommendation: The data set series should keep the same PID, as long as the data source and processing algorithms are the same. A version history in the data set series metadata and an explanation on the landing page should be used to clear up any confusion.

- e. A project invents an improved algorithm for an atmospheric trace gas product, reprocesses the entire data set series, and releases the result for web download. This happens every few years. The data centre is not planning on keeping more than the current and two previous versions of the data set. The older versions are deleted from the archive.

Question: Does each new processing get a new PID? How should PIDs for deleted versions be dealt with?

Recommendation: Each reprocessed data set series receives a new PID. Ensure that the older versions are kept as long as possible, but the use of a “tombstone” landing page which refers to the updated versions for any old versions that have been deleted or are not available to users is necessary.

B.2 Data Archive – Moving a Collection



- a. A data set series, which has a PID assigned to it, cannot be archived any longer by the data centre. A purge alert is issued. Another data centre offers to take over the data set, permanently archive it, and make it accessible to users.

Question: Does the data set receive a new PID?

Recommendation: No, PIDs should be data-centre agnostic, i.e. they should not reflect the data centre in their name or number. The relocated data set series keeps the same PID but the PID metadata are updated to reflect the new location. A new landing page should be created and maintained by the new host. If the new host does not use PIDs and cannot maintain the metadata and landing page, a “tombstone” landing page must be created by one of the institutions, with information on the new data host. All PID management details should be specified in the purge alert agreement between the two institutions, e.g. who should update the PID metadata, commitment of persistence by the new host, is a tombstone page required, etc.

B.2 Data Archive – Moving a Collection



- a. A data set series, which has a PID assigned to it, cannot be archived any longer by the data centre. A purge alert is issued. Another data centre offers to take over the data set, permanently archive it, and make it accessible to users.

Question: Does the data set receive a new PID?

Recommendation: No, PIDs should be data-centre agnostic, i.e. they should not reflect the data centre in their name or number. The relocated data set series keeps the same PID but the PID metadata are updated to reflect the new location. A new landing page should be created and maintained by the new host. If the new host does not use PIDs and cannot maintain the metadata and landing page, a “tombstone” landing page must be created by one of the institutions, with information on the new data host. All PID management details should be specified in the purge alert agreement between the two institutions, e.g. who should update the PID metadata, commitment of persistence by the new host, is a tombstone page required, etc.

B.2 Data Archive – Distributed Collection



- a. Several copies of a data set series are archived in different data centres and delivered to users separately. For ESA, this could even be different national institutes that use different PID systems.

Question: Does each copy have its own PID?

Recommendation: If the data sets are completely identical, they should have the same PID. Collaborating data centres should coordinate their PID systems as much as possible. It is also possible to use a central ID resolving catalogue.

- b. The European NOAA AVHRR 1km data set series consists of partially overlapping data sets hosted by multiple data centres across Europe.

Question: Should these data set series each have their own PID?

Recommendation: The data sets are not identical. They were downlinked separately and processed on local systems, possibly using different algorithms and calibrations. Therefore, these data set series should each have their own unique PID.

B.2 Data Archive – Identifying additional information



- a. A data set has documentation associated with it. Some documentation is held by another organisation.

Question: Do these get PIDs also? Is there a way to link PIDs with each other, or are these “subsets” of the data set PID?

Recommendation: The documents should not be subsets of the data set series or of its PID. If the documentation is permanently archived and disseminated, it may qualify for its own PID. A link to the documentation associated with a data set series, or its PID, can be established via the landing page no matter where the documentation is being held.

- b. The archive contains auxiliary data that are used in calculating higher-level products. Some of these came from third parties and are probably archived there as well.

Question: Should these auxiliary data get a PID?

Recommendation: No, auxiliary data sets that are not meant to be released to the public do not receive PIDs. An internal ID is sufficient.

- a. A data centre in general holds individual data sets in collections, i.e. within data set series. PIDs usually are assigned at the level of the data set series. Additionally, individual data sets or products, such as global mosaics, are added to the archive. These products will be used in publications and scientists will want to cite them.

Question: Does the data centre deviate from its policy of assigning PIDs only to data set series and assign PIDs also to individual products for specific cases such as this?

Recommendation: For specific cases, and given sufficient importance is assigned to the product, the data centre should deviate from its general policy of assigning PIDs only to data set series and assign PIDs also to individual products if required.

- b. Within a satellite series, there may be variations in the sensors (e.g. AVHRR/2 had 5 spectral bands (NOAA7-14), whereas AVHRR/3 has 6 (NOAA15-19). Moreover there are also slight differences in the range of the spectral bands between NOAA 6, 8, 10 and NOAA 7, 9, 11, 12, and 14.

Question: Should the time series from different versions of a sensor get different PIDs?

Recommendation: This is a question of how collections are organised. If data from the different sensor versions end up in the same data collection, they would get one PID. Each archive may have different ideas about grouping, and each archive must decide what grouping makes sense for their own data collections. Users should be carefully advised on how to cite the data, indicating the sensor(s) and/or satellites and/or processor that have been used. Larger and more heterogeneous collections have to be more carefully cited. It is good practice to keep the collections as homogeneous as possible, so the user can cite with just a PID (and maybe a time/area subset, if applicable). If the archive owner chooses something different, very detailed citation guidelines must be provided to the user.

- c. A dataset has been assigned a PID and it is also included in the metadata of the products of the dataset. Later on, a new collection is being created that also uses this dataset as a part of it, but in this first dataset there is already a PID assigned.

Question: How to handle PID insertion when the dataset already has a PID?

Recommendation: It is recommended to only keep the first PID inside the dataset as it is. The landing page of the new collection should clarify the existence of an earlier PID for this dataset and why the new collection PID cannot be found in the dataset.

- d. Some instruments generate hundreds of datasets, but it is also desired to have a single DOI that points to “all” of the data.

Recommendation: One PID is assigned to the dataset as it is. The landing page describes this dataset. Any minor change is reported in the landing page and, in case of major change, a new PID is assigned just to the specific dataset or to the full instrument dataset.

- e. Two different data sets are assigned two different DOIs and portions are extracted to create a new subset without changes.

Recommendation: A new DOI should be assigned to the sub data set.

B.3 Data Producer – Non-Permanent products



- a. The ground segment for a new satellite is planning the release of novel remote sensing data products. The products are currently in an experimental phase, released only to a few known project partners, but they will soon be released to the public. One of the project partners analyses the experimental product, publishes their results, and wants to cite the product used.

Question: Should experimental products with limited release get a PID?

Recommendation: No, preliminary or experimental products do not receive a PID. A PID is assigned when the data set series is released in its final form.

- b. For archiving efficiency, preliminary and final products are being held in one archive data set series. Within the collection they can be distinguished via a flag in the product metadata. The experimental products may at some point be deleted from the collection when they are no longer needed.

Question: Should the collection receive a PID, even though its content will change? Should it receive a new PID when the experimental products have been deleted? What if the experimental products have been used and cited in publications?

Recommendation: The data set series receives and keeps its PID. The PID should resolve to a landing page that explains which data sets out of the series were deleted and why (similar to scenario 2c and 2d).

B.3 Data Producer – Non-Permanent products



- c. A near-real-time (NRT) product is released to the public. These products are not archived, but only available for a couple of weeks – the emphasis is on releasing a product with as little turnaround time as possible. No archived version of the product is available.

Question: Should the NRT product receive a PID?

Recommendation: As a general rule, PIDs should not be assigned to a data set series that is not intended for long-term archiving. However, there may be cases where many scientific studies are based on the NRT product and the users are asking for a way to cite their data source. The data provider should then consider committing to archiving the NRT data files. If the data provider can safely archive the data for at least 10 years, following general scientific best practices, then a PID may be assigned to the NRT product. One PID should then be assigned to the entire data series, even if the processor version changes mid-stream. The landing page should clearly state the limitations of the NRT product and the archive commitment by the data provider.

- d. A near-real-time product is released to the public. An archived version of the same product is also available, processed after the fact with more quality control and perhaps slightly different data processing.

Question: Does the NRT product need a PID? Is that PID different from the archived time series product?

Recommendation: The recommendation is still to not assign a PID to a data set series that is not intended for long-term archiving. However, there may be a compelling reason to give a PID to the NRT product; for instance, it is being used in published studies and the authors want to cite their data source. In this case, the PID given to the NRT product should be different from the archived product's PID. The NRT product's landing page must clearly state the product's limitations and that it is not archived and provide a link to the landing page of the archived version.

Proposed Way Forward



- ❖ Ask the WGISS members for new Use Cases by end of April
- ❖ Draft (including any new use case proposed) and Circulate the EO Persistent Identifier BP by end of May 2026
- ❖ Formal issue of the updated EO Persistent Identifier BP by end of June 2026