Persistent Identifiers Best Practices

CEOS WGISS
Data Stewardship Interest Group

Doc. Ref.: CEOS/WGISS/DSIG/PIDBP
Date: March 2015
Issue: Version 1.0

Document Status Sheet

Issue	Date	Comments	Editor
0.7	December 2014	First draft for WGISS-DSIG review	Tyler Christensen
1.0	March 2015	First Issue, incorporating comments received during the WGISS-DSIG review	Tyler Christensen

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Purpose of the document

This document aims to provide recommendations and best practices on the use of Persistent Identifiers to Earth Observation mission data, allowing globally unique, unambiguous, and permanent identification of a digital object.

1.2 Document overview

This document is divided into:

Section 1: Introduction, including definitions, abbreviations, and related documents

Section 2: Background

Section 3: Objectives and Needs

Section 4: Components of a PID System

Section5: PID Best Practice Content

Annex A: More Information on Persistent Identifiers

Annex B: Use Case Scenarios, tailored for the Earth Observation community

1.3 Acronyms

Acronym	Description
ARK	Archival Resource Key
DOI	Digital Object Identifier
EO	Earth Observation
LTDP	Long Term Data Preservation
OGC	Open Geospatial Consortium
PID	Persistent Identifier
XML	eXtensible Markup Language

1.4 Definitions

Topic	Description

DataCite	International not-for-profit organization that aims to: • Establish easier access to research data on the Internet • Increase acceptance of research data as legitimate, citable contributions to the scholarly record • Support data archiving that will permit results to be verified and re-purposed for future study. The primary means of establishing easier access to research data is by DataCite members assigning persistent identifiers, such as digital object identifiers (DOIs), to data sets. Although currently leveraging the well-established DOI infrastructure, DataCite takes an open approach to identifiers, and considers other systems and services that help forward its objectives.
Landing Page	When a data user clicks on a persistent identifier link, the resolver will lead to a landing page. This webpage shows all of the relevant information about a data set, and (most importantly) will have a prominent link for downloading the data. Note that the PID resolver should not point directly to the data download. The landing page will display some metadata, and may also link to documentation, publications that have used the dataset, citation recommendations, data use policy, etc. The landing page should be actively updated and maintained. If the URL to the page changes, it must be updated in the resolver database. The following link could be useful to understand more about landing pages: http://vso1.nascom.nasa.gov/rdap/RDAP2012_landingpages_handout.pdf

1.5 Related Documents

1.5.1 Applicable Documents

Applicable Document ID	Document Title	Reference	Availability

1.5.2 Reference Documents

The following documents, though not formally part of this document, amplify or clarify its content.

Reference Document ID	Document Title	Reference	Availability

2 BACKGROUND

Internet resources tend to have a short life. Their identification and persistent location pose complex problems that affect many technological and organizational issues involving the citation, retrieval and preservation of cultural/scientific resources. This is by no means a technical problem alone: persistent digital object identification, including texts, music, video, still images, scientific documents and the like, is still a major issue that prevents the use of today's Internet as a trustworthy platform for the research and dissemination of scientific and cultural content.

The rapid increase of digital assets in recent years, especially in the context of e-science, has made this dependency even stronger, making it clear that digital identifiers are crucial in order to preserve, manage, access and re-use datasets over time. The implementation of a system for persistent identification of digital and non-digital objects is the first fundamental step to this purpose, becoming a crucial prerequisite for sustained and reliable resource discovery, citation and re-use.

The persistent identification of digital objects (e.g. articles, datasets, images, streams of data) as well as of non-digital objects (real-world entities, like e.g. authors, institutions, teams, geographic locations and so on) is a crucial issue for the whole information society. The capability to unambiguously locate and access digital resources, associate them with the related authors and other relevant entities (e.g. institutions, research groups, projects) is becoming essential to allow the citation, retrieval and preservation of cultural and intellectual resources.

An identifier is a unique identification code that is applied to "something", so that the "something" can be unambiguously referenced. For example, a catalogue number is an identifier for a particular specimen, and an ISBN code is an identifier for a particular book. Persistent identifiers (PIDs) are simply maintainable identifiers that allow us to permanently refer to a digital object. Identifiers are a way of giving digital resources, such as documents, images and data records, a unique reference number.

A Persistent Identifier is an identifier that is effectively permanently assigned to an object. The only useful persistent identifiers are also persistently actionable (that is, you can "click" them); however, unlike a simple hyperlink, persistent identifiers are supposed to continue to provide access to the resource, even when it moves to other servers or even to other organizations. A digital object may be moved, removed or renamed for many reasons. A solution is to give the object a persistent identifier (PID) that will remain the same regardless of where the resource is located and how it is stored.

Long term data preservation, dissemination and access of scientific digital objects are now among the core missions of international institutions. The use of URLs can't be considered a reliable approach for addressing these issues due to the structural instability of links (ex. domains no longer available) and related resources (relocation or updating). The current use of the URL approach increases the risk of losing documents or under-using available collections. In the Spatial & Scientific Heritage domain it is essential not only to identify a resource but also to guarantee continuous access to it.

PIDs are also an increasingly global standard. Not using a reliable PID system could harm a data provider's credibility and standards-compliance. PIDs also lead to increased citation of data resources used in published studies, so that data providers can better track the impact of their data resources.

3 OBJECTIVES AND NEEDS

The growth of scientific and non-scientific digital data is resulting in an increasing number of digital objects and resources that has to be managed. This data intensive environment offers many new opportunities: the possibility of accessing a massive amount of scientific and cultural data in digital format, the increasing linkage across authors and their publications, the development of new and much more powerful metrics for assessing the impact of scientific production, etc.

However, this scenario has led to the emergence of new challenges such as digital preservation, data integration, quality assessment and provenance. These challenges become magnified in global contexts where resources are distributed across systems and standards, and the movement of data across disciplines and organizations is very intensive.

The main purpose of a persistent identifier is to help data users cite and find specific data sets. In this context the Earth Sciences and Earth Observation mission data identified objectives and needs are listed below:

Objectives & Needs

- Globally unique, unambiguous and permanent identification of a digital object for locating and accessing over time.
- Improve discoverability and accessibility.
- Enable users to retrieve objects without knowing their location.
- Enable repositories to change the location of objects internally.
- Enable repositories to share objects with other services where appropriate.
- Enable researchers to cite objects consistently over time, which also benefits data holders.
- Increase data visibility and use.
- Increase credibility and value of data holdings.

4 COMPONENTS OF A PID SYSTEM

The first component is, obviously, the data resource that will receive a persistent identifier. This data resource itself is expected to persist over time, as part of a long-term data preservation strategy.

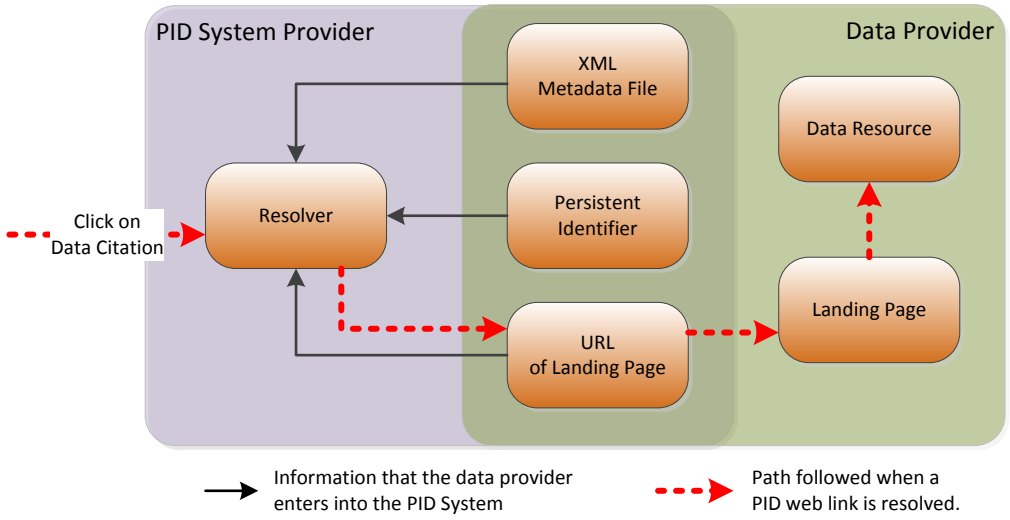
Note that in talking about the identifier, the example of a DOI that is registered through a DataCite member organization will be used in this document. Some subtle differences could be found if a different system is chosen, but the overall structure will be similar.

A data provider will get a DOI "prefix" from DataCite, simply a number that uniquely identifies that providers' subset of DOIs. The data provider then generates an internally unique "suffix" for the archived data resource. A landing page must also be constructed. This is a web page with information about the data and a download link (if available), hosted by the data provider's web server. Finally, an XML metadata file must be constructed according to the DataCite metadata standard.

To register the DOI, the XML metadata file, the URL for the landing page, and the identifier itself are sent to DataCite and from there into the DOI resolver system. When a data user clicks on a DOI citation, the resolver (http://dx.doi.org/) can then redirect the user to the landing page. The XML metadata will be used for data discovery via online search, metadata harvesting services, data portals, and data repository catalogs.

The DOI and the dataset itself should never change. However, the data provider must maintain these components as needed:

- update the landing page on its own web server
- update the DataCite metadata by sending an updated XML file
- send a new URL to DataCite if the landing page location changes



Information flows in a PID system. The data host provides the identifier itself, an XML metadata file, and the URL of the landing page to the PID system provider. These are then stored by the resolver system. When an interested user clicks on a PID web link, the resolver redirects them to the landing page with information on how to access the data resource itself.

5 PID BEST PRACTICE CONTENT

This section addresses the main "themes" that should be applied to guarantee a globally unique, unambiguous, and permanent identification of a digital object for locating and accessing it over time. Meeting these harmonized CEOS guidelines for PID implementation improves also interoperability with other EO data providers.

Some Use Cases are provided in Annex B in order to describe the Earth Sciences and Earth Observation mission context.

5.1 General Recommendations

5.1.1.1 Choosing a PID system

[REC_01]
In choosing a PID system, the following should be considered:

- Evaluate the technical reliability, authority, and credibility of the PID system.
- Ensure that the system has financial and/or institutional support, to ensure its long-term viability.
- Make sure that it is flexible enough to represent the granularity of the collections.
- Make sure that the service uses open standards for the implementation of PIDs.
- The IDs should be independently generated by the data provider, with no need for a centralized system.
- It is best to use an external resolver rather than an internal one that must be maintained by the data distributor.

5.1.1.2 PID numbering

[REC_02]
Numbering should be completely opaque. The identifier should not contain any information about the resource it identifies. Opaque IDs are easier to manage, less likely to become obsolete over time, and conform to global standards.

[REC_03]
PIDs must be universally unique, with a system to ensure each identifier is unique worldwide.

[REC_04]
No hierarchies/versioning should be used within the actual PID numbering, since that violates the opaqueness requirement. Use the landing page instead.

[REC_05]
Consider using a checksum, such as Mod97, to check for valid PIDs.

5.1.1.3 Permanence

[REC_06]
Data providers must commit themselves to the persistence of their PIDs, maintaining and updating metadata, URLs, and landing pages as needed.

[REC_07]
The identifier should never change, neither the identifier itself nor the resource it refers to.

[REC_08]

If data content changes (reprocessing, error correction, versioning), assign a new PID.

[REC_09]
If the file format of any data is changed, e.g. CDAT to GeoTIFF, assign a new PID.

[REC_10]
If the data are transferred to new physical storage media, it is not necessary to assign a new PID. It would be good practice to perform a validation/checksum to guarantee the integrity of the new copy and check for bitwise differences.

[REC_11]
Only assign PIDs to data sets in a long-term archive - not auxiliary data, experimental products, on-demand processing, or near-real-time products that are not archived.

[REC_12]
If a data set is transferred to a new institution, keep the same PID. In this case, a new landing page should be generated, and the URL and the resolver's metadata file should be updated. If the new host does not use PIDs, a tombstone landing page must still be maintained, either by the previous host or the new institution, giving information about the new data owner.

[REC_13]
The PID must remain resolvable. This means updating the landing page URL in the resolver if the location changes. If a data set must be deleted, a tombstone page must be maintained (e.g. explain why the data were removed, link to a new version, contact information for questions, etc.).

5.1.1.4 Resolving

[REC_14]
A PID must be actionable, meaning that the identifier will lead the user to information about the resource. This is also called resolution of an identifier.

[REC_15]
The ID should resolve to a landing page, not a direct link to data download. The page should be hosted by the data holder and updated as needed.

[REC_16]
The landing page should contain provenance, quality, and access constraints, but doesn't have to show the entire metadata record. The goal is a readable summary of the data set so potential users can see if it meets their needs. Information on how to access the data should also be included, with an online download link if available.

[REC_17]
Use the landing page to link data sets that are related (e.g. reprocessing, versioning, subsets, and supersets). These relationships can also be reflected in the resolver's metadata fields.

5.1.1.5 Granularity

[REC_18]
As a general rule, assign PIDs to data collections (e.g. a consistent time-series) rather than an individual scene or data product. This can be flexible, depending on how users will want to cite the data.

[REC_19]
Data from the same source but at a different processing level (e.g. L1B vs. L2) should receive separate PIDs.

[REC_20]
Assign a single PID for a whole time-series, even if new data are still being added. It is more convenient to cite a subset of a larger data source, rather than many PIDs to make up a single

time-series. Note that no retrospective changes will be made to historical data records that are already in the archive.

[REC_21]
Use one PID for a multi-satellite time-series, as long as the series is internally consistent.

5.1.1.6 Documentation

[REC_22]
An institution that uses PIDs should have an official, written PID policy.

[REC_23]
Make sure that the uses of PIDs are part of the written policy of the institution.

- Be clear, and make public, in which environments the PIDs are unique and how they resolve to an available resource.
- Clarify what is meant by 'persistent', whether there are any limitations, and how this will be implemented.

[REC_24]
Provide citation guidelines that use the PID, including how to cite a subset in space and time.

[REC_25]
Documents related to a data set should be linked on the landing page, not as subset PIDs.

[REC_26]
Add the PID to the metadata for individual products in the fields provided, if applicable.

[REC_27]
Comply with the latest metadata standards from your PID provider, and update as needed.

5.1.1.7 Interoperability

[REC_28]
The same data should have the same PID, even for duplicates in different archives. If your archive hosts a copy of a static data set that already has a PID, then keep the same one. A single landing page may have several different links for data access and download.

[REC_29]
In case of inter-agency data records, a PID (e.g. based on DOI) can be associated with a list of all contributing collections and their respective PIDs (e.g. DOIs, ARK). No translator is needed in case different PID systems are used by different providers.

5.1.1.8 A special note on DOI

Recommendations in this document are generic so that a data provider can be 100% compliant no matter which PID system is chosen. However, as a result of the analysis and investigations carried out in finalizing these recommendations, the utilization of DOIs (see Annex A) is considered the most suitable approach for persistent identifiers in the Earth Observation domain as:

- DOI is the most widely used system globally - for publications and increasingly for data – and is rapidly becoming the global standard for data citation. Several international organizations (also in the Earth Science domain) are using it and many publications, e.g. the PLOS journals, now encourage or require data citation via DOI.
- Many millions of data resources now have a DOI, making it very unlikely that the system will disappear.
- Most organizations pay a small fee to register DOIs, providing funding to maintain the DOI infrastructure.

5.1.1.9 A special note on registration agents

Regardless of the chosen PID system, it is also recommended to work with a regional or local PID "agent", if available. These nonprofit groups provide a simple way to register PIDs in the resolver system. Examples are DataCite, a global consortium that encourages data citation via DOIs, and EZID, a service of the California Digital Library that supports many different PID types. A registration agent can provide information, advice, and tools to help manage persistent identifiers and, most importantly, can connect providers with a data citation community of practice. For instance, DataCite works closely with journal publishers to promote data citation and help providers to track the impact of their valuable data resources. DOI metadata that are registered with DataCite are in addition actively pushed to publishing agencies, e.g. to Thomson-Reuters.

5.2 PID Policy - Example

This section can be customized by organizations to reflect their own PID policy decisions. The result could be used as an official persistent identifiers policy statement. An example of a policy statement for the use of persistent identifiers is given below. This will have to be adapted to suit the needs of the individual organization.

Statement of Persistence: [Organization] commits to assigning permanent identifiers to the data sets that are released to the public. The data, identifiers, metadata, and landing pages will be maintained indefinitely. Tombstone pages will be provided for any data that are no longer available for any reason.

PID system: [Organization] chooses to use the DOI system for identifiers, conforming to the [e.g. DataCite] standards for metadata and landing pages. Identifiers will be opaque strings of random letters and numbers, ending with a two-digit checksum calculated according to [e.g. a Mod97 algorithm]. These identifiers are globally unique within the DOI system. Metadata conforming to the [e.g. DataCite] standard will be made freely available for data discovery.

Resolving PIDs: Because [Organization] uses DOI, all identifiers are resolvable online through the DOI web interface (e.g. http://dx.doi.org/10.4225/25/5487CC0D4F40B). The URL will always redirect to a landing page that shows the current location of the original data set, even if it has moved. The landing page will also provide dataset details and information on any access restrictions.

Citing Data: Scientists who use [Organization] data in their research are asked to use the persistent identifier when citing the data source in their publications.
Example: Cooper, L.; Lamont-Doherty Earth Observing Laboratory (LDEO); (2009): HLY-08-01 POES Satellite Images (Version 1.0); UCAR/NCAR - Earth Observing Laboratory. http://dx.doi.org/10.5065/D6G73BQC

Granularity: [Organization] identifiers are assigned at the collection level, e.g. to the entire time-series of a data parameter. If a researcher uses a subset of the data in their research, the citation should include the subset boundaries so that others who wish to repeat the research can extract the same subset.
Example: Fiedler, E.K.; McLaren, A.; Merchant, C.J.; Donlon, C. (2014): ESA Sea Surface Temperature Climate Change Initiative (ESA SST CCI): GHRSST Multi-Product ensemble (GMPE). NERC Earth Observation Data Centre, 24th February 2015. Time subset: 1991-09-01 to 2004-01-01. http://dx.doi.org/10.5285/7BAF7407-2F15-406C-8F09-CB9DC10392AA.

Contact Information: If there are any questions about these identifiers or to report a broken link, please contact (email address and phone).

ANNEX A - MORE INFORMATION ON PERSISTENT IDENTIFIERS AND DIGITAL OBJECT IDENTIFIER (DOI)

Persistent identifiers (PIDs) are simply maintainable identifiers that allow to refer to a digital object. An identifier is a unique identification code that is applied to "something", so that the "something" can be unambiguously referenced. Identifiers are a way of giving digital resources, such as documents, images and data records, a unique reference number. A Persistent Identifier is an identifier that is effectively permanently assigned to an object.

The only useful persistent identifiers are also persistently actionable (that is, you can "click" them); however, unlike a simple hyperlink, persistent identifiers are supposed to continue to provide access to the resource, even when it moves to other servers or even to other organizations. A digital object may be moved, removed or renamed for many reasons.

A solution is to associate a persistent identifier (PID) with a digital resource that will remain the same regardless of where the resource is located.

Notions:

1. Persistent Identifiers must be globally unique
2. Persistent Identifiers must exist indefinitely

These are the main steps to be performed in order to implement a PID system:

1. Selection of resources that need a PID;
2. Resource name assignment and register creation;
3. Resolution of a PID with the associated URL;
4. Maintenance of the register that associates PID-URL and guarantee of continuous access to the resources.

Persistent Identifier system requirements

The PID system requirements are:

- Global uniqueness: an identifier is a label that is associated with an object in a certain context. "Context" is intended as both the kind of standard used for the name syntax and the identification of the authority (sub-namespace) that assigns this label.
- Persistence: refers to the permanent lifetime of an identifier. It is not possible to reassign the PID to other resources or to delete it. That is, the PID will be globally unique forever.
- Resolvability: refers to the possibility of retrieving a resource once it is published.
- Reliability: the PID infrastructure must always be active (service redundancy, back-up deposit services, etc.) and the register updated (through automatic systems).
- Authority: is who assigns, manages and resolves the identifiers.
- Flexibility: an identifier system will be more effective if it is able to accommodate the special requirements of different types of material or collections.
- Interoperability: this aspect is fundamental for guaranteeing the possibility of disseminating and accessing science digital objects.

Persistent Identifier scheme

Main elements relevant to a dataset to be exposed using Persistent Identifiers:

1. Authority: this generally correlates to the institute or organization which has responsibility for the dataset (also known as the information resource). Important considerations when choosing an authority include:

   a. Stability – organizations that endure little change to their name and the names of their dependencies are a better choice for long term authority names.
   b. Longevity – organizations that have more community support are more likely to persist into the future.
2. Context: this correlates to the dataset (which might be a database or a clearly identified subset of a database). Considerations when choosing a context name include:
   a. Contexts must be unique within an authority.
   b. What subset of the information resource can be independently curated? If Persistent Identifiers are applied to a large, heterogeneous dataset (perhaps actively curated in some areas, and fairly static in others) or to a dataset which has been derived from multiple sources, consider splitting the dataset into subsets. Also consider what might happen to information resources if the authority were to wind down, split or merge with another organization.
3. Object: this correlates to the specific resource (possibly a database row). Object names (identifiers) must be unique within a context.

DOI: Digital Object Identifier

DOI identifiers are a managed identifier system, maintained and controlled by the DOI Foundation [http://www.doi.org/]. The foundation manages a commercial infrastructure for the assignment and use of DOI identifiers. DOI is the most widely used Persistent Identifier system globally for publications and increasingly for data. More than 3 million data resources now have a DOI, and more than 100 million DOIs have been assigned over all resource types. DOI metadata that are registered with regional or local PID "agents" can be in addition actively pushed to publishing agencies or journals to facilitate EO data visibility and discoverability beyond EO portals, e.g. in generic search engines. This wide visibility outside the EO community is a major advantage of using DOI.

Additional details on DOI are provided below or can be found at http://www.doi.org/:

- Registration, support, persistence control and policy making is provided by the DOI Foundation, ensuring a robust system for maintaining the identifiers.
- DOIs may be free for research institutions, but not for other organizations. When not free, DOI identifiers must be bought at a cost per identifier from the DOI Foundation.
- Funds raised by the DOI Foundation support the maintenance of the DOI resolver and associated infrastructure.
- The primary focus of the DOI system is on the management of entities of interest as intellectual property, but this does not preclude issuing a DOI name to any entity of interest to a user community.
- All DOI names must be registered in a DOI system directory. Registrants are responsible for the maintenance of current data relating to DOI names that they have registered.
- Resource may change, be updated, be renamed, be moved, or be removed (assignment of new DOI plus update of DOI metadata to reflect changes or point to updated resource).
- The DOI system will not accept duplicate DOI names on registration; no two DOI names from different registrants can ever share the same prefix; no two identical strings can be assigned within one prefix.
- DOI example: 10.1000/186, http://dx.doi.org/10.1000/186
- DOI are resolved through the online DOI resolver by appending the DOI to the URL http://dx.doi.org/
- Authority, context and object identifier components are obscured with the use of DOIs.
- DOI registration through local agents like DataCite, may be free or reduced cost for public research institutions.

ANNEX B – USE CASE SCENARIOS

The scenarios below address a number of situations that may occur in Earth Observation (EO) and which may affect the way persistent identifiers are applied to EO by data managers. The main purpose of a persistent identifier is to help data users cite and find specific data sets. The recommendations below are based on this main objective and provided as guidance to EO data managers in defining Persistent Identifiers for their data holdings.

B.1 Data User

1. **Citing data sets**
   a. A user downloads a data set, perform some scientific activity on it or with it, 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).

2. **Accessing cited data sets**
   a. A scientist reads a published article with a data citation, and wants to access the same dataset used in the original research.
      Recommendation: The PID should resolve to a landing page that provides access to the exact same data. If the content of a dataset has changed, a new PID should be assigned to the new resource. The old landing page should be maintained and refer to the updated data set series.
   b. A scientist reads a published article with a data citation, and wants to access the same dataset used in the original research. The data set is not available any more.
      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 dataset which is accessible.
   c. A scientist reads an article from 15 years ago with a data citation. The cited dataset is still archived and accessible, but it has been replaced with a new one, computed using an improved algorithm.
      Recommendation: The landing page for the original dataset should include a disclaimer, informing users that an improved version is available and giving a link to the new version.
   d. A scientist reads an article from 15 years ago with a data citation. The original dataset is no longer archived, because it has been replaced with a new product that uses an improved algorithm.
      Recommendation: The PID should resolve to a landing page that explains why the data are gone and links to the improved product.

B.2 Data Archive

1. **Adding data to a data set series (collection)**
   a. A data archive wants to assign a PID, either to a new dataset or to an existing archived dataset.
      Recommendation: Develop a consistent workflow for assigning PIDs, and for evaluating whether a dataset 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 dataset 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.

c. The static or dynamic dataset series in the archive is found to be incomplete. The missing datasets are being recovered from external sources and added to the dataset.
Question: Does the PID remain the same?
Recommendation: The dataset 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.

d. A data set series includes data from the same sensor, carried on many different satellites (example: AVHRR on the NOAA POES satellite series). A new operational satellite to continue the measurement replaces the current satellite.
Question: Does the time series keep the same PID?
Recommendation: Yes, the dataset 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.

2. Deleting a dataset series (collections)

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 have been deleted.

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 L2 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.

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 dataset series receive a new PID when the faulty products have been deleted? What if the faulty products have been used and cited?
Recommendation: The dataset series should keep the same PID, as long as the data source and processing algorithms are the same. A version history in the dataset series metadata and an explanation on the landing page should be used to clear up any confusion.

e. A project defines and implements an improved algorithm for an atmospheric trace gas product, reprocesses the entire dataset series, and releases the result for web download. This happens every few years. The data center is not planning on keeping more than the current and two previous versions of the dataset. The older versions are deleted from the archive.
Question: Does each new processing get a new PID? How to deal with PIDs for deleted

versions?
Recommendation: Each reprocessed dataset series receives a new PID. Ensure that the older versions are kept as long as possible, but definitely use a "tombstone" landing page which refers to the updated versions for any old versions that have been deleted.

3. Moving a data set series (collections)

a. A dataset series, which has a PID assigned to it, cannot be archived any longer by the data center. A purge alert is issued. Another data center offers to take over the dataset, permanently archive it, and make it accessible to users.
Question: Does the data set receive a new PID?
Recommendation: No, PIDs should be data-center agnostic, i.e. not reflect the data center in its name or number. The relocated dataset 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.

4. Distributed data set series (collections)

a. Several copies of a dataset series are archived in different data centers 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 datasets are completely identical, they should theoretically have the same PID. Since this, however, cannot be guaranteed - and since the dataset locations are different – each dataset series should receive its own PID. Nonetheless, collaborating data centers should coordinate their PID systems as much as possible. It is also possible to use a central ID resolving catalog, such as the OKKAM Entity Name System (http://www.api.okkam.org/).

b. The European NOAA AVHRR 1km dataset series consists of partially overlapping datasets hosted by multiple data centers across Europe.
Question: Should these dataset series each have their own PID?
Recommendation: The datasets are not identical. They were downlinked separately and processed on local systems, possibly using different algorithms and calibrations. Therefore, these dataset series should each have their own unique PID.

5. Identifying additional information (PDSC – documentation, software)

a. A dataset has documentation associated with it. Some documentation is held by another organization.
Question: Do these get PIDs also? Is there a way to link PIDs with each other, or are these "subsets" of the dataset PID?
Recommendation: The documents should not be subsets of the dataset series or of its PID. If the documentation is permanently archived, it may qualify for its own PID. A link to the documentation associated with a dataset 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 enough.

6. PID Granularity

a. A data center in general holds individual datasets in collections, i.e. within dataset series. PIDs usually are assigned at the level of the dataset series. Additionally, individual datasets 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 center deviate from its policy of assigning PIDs only to dataset 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 center should deviate from its general policy of assigning PIDs only to dataset 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 organized. 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.

B.3 Data Producer

1. 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 analyzes the experimental product, publishes his 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 dataset series is released in its final form.

b. For archiving efficiency, preliminary and final products are being held in one archive dataset 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 dataset series receives and keeps its PID. The PID should resolve to a landing page that explains which datasets out of the series were deleted and why (similar to scenario 2c and 2d).

c. A near-real-time 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. An archived dataset series 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: No PIDs are assigned to a dataset series that is not intended for long-term archiving.