**Proceedings for the 2015 PV Conference,
3-5 November 2015, Darmstadt, Germany**

BEST PRACTICES FOR PERSISTENT IDENTIFIERS IN EARTH OBSERVATION ARCHIVES

Tyler Christensen[1], Mirko Albani[2], Andrew Mitchell[3], Satoko Miura[4], Yoshiyuki Kudo[4], Iolanda Maggio[5], Razvan Cosac[5]

[1]DLR German Remote Sensing Data Center, Oberpfaffenhofen, D-82234 Weßling, Germany
[2]ESA ESRIN, Via Galileo Galilei, 00044 Frascati RM, Italy
[3]NASA Goddard Spaceflight Center, 8800 Greenbelt Road, Greenbelt MD 20771, USA
[4]JAXA Mission Operations System Office, 2-1-1 Sengen, Tsukuba-shi, Ibaraki 305-8505, Japan
[5]RHEA Systems, La Piramide, Via di Grotte Portella 6/8, 00044 Frascati, Italy

Abstract

Sharing and citing scientific data sources is becoming the global standard. Persistent identifiers (PIDs) allow data providers to permanently and uniquely identify a resource, track its use by the scientific community, and increase its visibility. To encourage and coordinate the implementation of PIDs in Earth Observation, the CEOS Working Group on Information Systems and Services (WGISS) recently developed a best practices document. The Best Practices include advice on choosing a PID system, assigning the IDs themselves, ensuring permanence, resolving the ID to the data resource, choosing an appropriate granularity, and documenting the PIDs properly. The recommendations are being applied in pilot implementations in DLR, ESA ESRIN, and NASA. Establishing common protocols and practices will help ensure interoperability among the global community of Earth Observation data providers. This paper will present the best practices, a few use cases, and the results of the pilot studies.

INTRODUCTION TO PERSISTENT IDENTIFIERS

Internet resources as well as their Uniform Resource Locator (URL) have a short life and do not ensure sustainable access to data resources. URLs, therefore, are not considered suitable as permanent reference to a resource. To ensure sustained and reliable resource discovery and facilitate citation and re-use, the implementation of a system for persistent identification of digital objects is a crucial prerequisite.

A Persistent Identifier (PID) is a unique alphanumeric code that is permanently assigned to an object. However, unlike a simple URL, persistent identifiers continue to provide access to the resource, even if the resource name changes, it is moved to other servers, or it is transferred to other organizations. Persistent identifiers are based on established and managed PID systems. Examples include the Digital Object Identifier (DOI) [3] and the Archival Resource Key (ARK) [1].

For publications, PIDs have been in use for many years. PIDs are now also becoming a global standard for referencing data resources, and establish a data provider's credibility and standards-compliance. Scientific journals and scientific project funding organizations increasingly require that the data related to a published study are properly archived and referenced by a PID. Persistent identifiers also lead to increased citation of data resources in published studies, so data providers can track the impact of their resources and receive credit when the data are used in future scientific studies.

BENEFITS OF PERSISTENT IDENTIFICATION OF DATA RESOURCES

Both data providers and users benefit from having persistent identifiers assigned to data resources.

Data providers clearly benefit from using PIDs. By tracking who uses a data set in publications and how it is used, data producers and repositories can receive credit for their work. This provides a parallel to the way that researchers are credited when journal articles are cited. The visibility and credibility of the data resources are also enhanced. It is expected that persistent identification – along with easy discoverability and accessibility - will also promote more frequent re-use of the resource and thus justify the data repositories' efforts in providing sustainable resource preservation and access.

PIDs can also improve data users' experience. When data sources are clearly cited with a PID in a journal article, a scientist who reads the study can seamlessly access the underlying data. In addition, when a study's authors make their data transparent and accessible, it improves the credibility of the research. Both of these benefits will strengthen science in the long run.

COMPONENTS OF A PID SYSTEM

Several interconnected components have to be established and maintained in order to set up and keep a PID system operational. The first component is a data resource that is expected to persist over time, as part of a long-term data preservation strategy. A data provider must then construct a globally unique identifier that complies with the chosen PID schema. A landing page must also be constructed. This is a web page with information about the data and a download link, hosted by the data provider's web server. The final requirement is a simple XML metadata file.

To register the PID, the XML metadata file, the landing page URL, and the identifier itself are sent to a PID resolver system. Depending on the PID system, this resolver could be one central database maintained by the system provider, or a local resolver on the data host's web server. When a data user clicks on a PID citation link, the resolver then redirects the user to the landing page. The published XML metadata can then be used for data discovery via online search, metadata harvesting services, data portals, and data repository catalogs. This provides visibility of the dataset beyond community-specific search and discovery tools and portals.

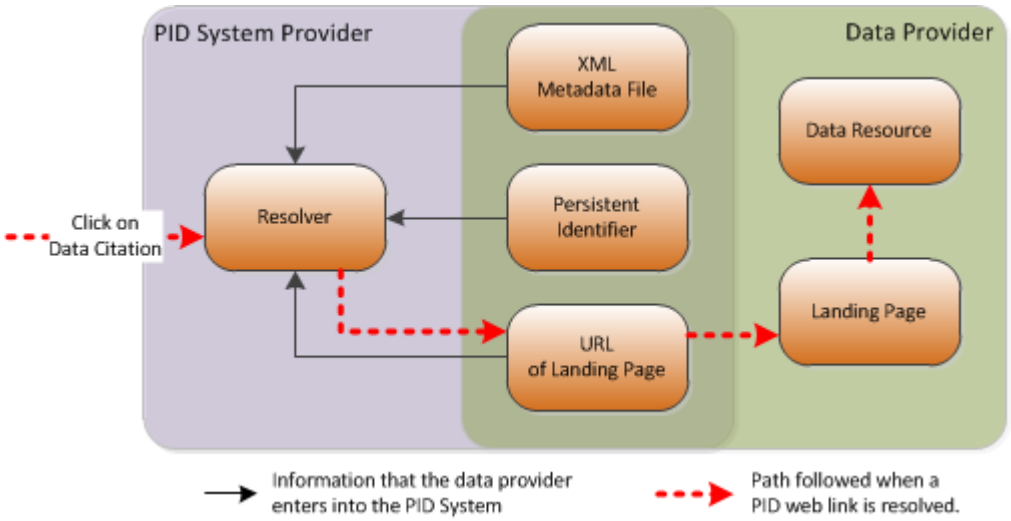

Figure 1: **Information flows in a PID system.**

The PID 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 metadata by sending an updated XML file to the resolver
- send a new URL to the resolver if the landing page location changes

THE MAIN PID SYSTEMS IN USE

There are many different PID systems available, and it can be a challenge to choose the right system. Some examples are: URN, URI, PURL, Handle, DOI, LSID, ARK, etc. Of these, DOI and ARK appear to be the most common in the Earth Observation (EO) community.

The Archival Resource Key (ARK) system has been developed at the California Digital Library [1]. The ARK system requires a data provider to maintain the data resource, the ARK ID itself, a data description, and a link to download the data. The ARK system is free and open with no cost to assign identifiers, and is most widespread in the library and museum communities.

The Digital Object Identifier (DOI®) system is maintained by the non-profit DOI Foundation [3]. It is widely used in the publishing industry, but it is also widely used for data citation. Unlike ARK, there is normally a cost associated with registering DOIs. Registration of DOIs is implemented by registration agencies, and the most common global agency for data citation is DataCite [2]. The main requirement is a commitment to provide and maintain a persistent data set, XML metadata, and a landing page.

PERSISTENT IDENTIFIERS FOR EARTH OBSERVATION DATA

Persistent identification systems were initially developed for citing publications, so using these systems for referencing data resources can be pose some challenges. Even more challenging is adapting persistent identifiers to the unique requirements of Earth Observation data management. The examples provided below highlight some of the issues specific to Earth Observation.

Example 1 – In theory, persistent identifiers should only be assigned to a static resource. That is, the content of the archived data resource should never change. However, many EO missions are ongoing, collecting data at regular intervals and populating the resource collection. The data set is therefore growing and extending every day. Assigning a PID to this growing collection would violate the fundamental requirement of a static resource. Does that mean that these active time series data are not eligible for a PID?

Example 2 – The EO community is constantly searching for better processing algorithms. It is common practice to reprocess a data set with the best and most modern techniques. The parameter is still the same, e.g. Sea Surface Temperature, but it has been calculated by a different method. Should the PID remain the same?

To account for these community-specific challenges and establish a unified approach, PID best practices for EO data were established within CEOS WGISS [4]. The hope is that these recommendations will help data repositories to implement PIDs and provide a level of harmonization. It is very important for data providers to be consistent with global best practices, because a chaotic PID implementation will prevent data users' acceptance of data citation.

BEST PRACTICES FOR IMPLEMENTING PERSISTENT IDENTIFIERS

The Persistent Identifiers Best Practices guidelines were developed by the CEOS/WGISS Data Stewardship Interest Group [4]. The document provides recommendations and best practices on the use of Persistent Identifiers for Earth Observation mission data, allowing globally unique, unambiguous, and permanent identification of a digital object. Meeting the harmonized CEOS guidelines for PID implementation improves interoperability with other EO data providers. The full document is available on the CEOS website: http://ceos.org/ourwork/workinggroups/wgiss/current-activities/data-stewardship/

The document gives 29 recommendations for various aspects of implementing a PID system:

- Choosing a PID system
- PID Numbering
- Permanence
- Resolving
- Granularity
- Documentation
- Interoperability

Some of the key recommendations are:

- 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.

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

- The identifier should never change, neither the identifier itself nor the resource it refers to.

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

- 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.

- 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.

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

- 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.

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

- 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.

The best practices were designed so that a data repository can be fully compliant with the formal recommendations no matter which specific PID system is chosen. However, the document does express the opinion that DOI is the best choice for the Earth Observation domain. The system is widely used, well-funded, and robust. Most importantly, it is becoming the global standard for data citation among scientific journals, data providers, and scientists.

EXAMPLES OF PID IMPLEMENTATION IN THE EARTH OBSERVATION COMMUNITY

Many Earth observation data repositories have already started the processing of integrating PIDs into existing data curation workflows. To show the variety of implementations, we will provide a few examples. These are not meant to be a comprehensive inventory, but simply to give an idea of what it possible.

German Aerospace Center (DLR), Earth Observation Center (EOC)

For the EO data held in the German Satellite Data Archive (D-SDA), the EOC has chosen to implement DataCite DOIs, registered via the German National Library of Science and Technology (TIB). The DOIs are fully opaque, and are assigned at the collection level. Policy and governance documents are currently being finalized, and the process of integrating DOIs into the data ingestion workflows has begun. The DOIs are being incorporated as a new component in the database that underlies the automated Data and Information Management System (DIMS), which has been developed over many years to handle multi-mission ground segment data archiving and access. EOC is also developing scripts to generate landing pages and DataCite XML metadata directly from ISO/INSPIRE-compliant collection metadata. In the meantime, DOIs are being assigned manually to a few pilot data sets (http://dx.doi.org/10.15489/ak90g1wty909). The long-term objective is to assign DOIs to all publically available data collections in the D-SDA.

European Space Agency (ESA), European Space Research Institute (ESRIN)

In accordance with the PID Best Practices, ESA-ESRIN has chosen the DOI system for the implementation of PIDs in the Earth Observation domain, and DataCite as the registration agency. DOIs will be assigned at collection level, to long-term archived data. ESA is also considering assigning DOIs to near-real-time EO data that are archived. A pilot study has already been performed at ESRIN in order to test the various aspects regarding the implementation of DOIs, and showed significant results. A plan of activities for the full implementation has been drafted and is currently being discussed. Once this has been established, the process of incorporating DOIs can begin. As part of this process, it is envisaged to generate documentation regarding citation guidelines and PID Policy.

Japan Aerospace Exploration Agency (JAXA), Satellite Applications and Operations Center (SAOC)

The SAOC is in a planning phase. A study is being conducted to assess the impact of implementing PIDs in the JAXA data management systems, focused on DOIs provided through the Japan Link Center (JaLC). Data citation is gaining momentum in Japan, and JAXA is looking closely at related activities before finalizing its implementation plans.

National Aeronautics and Space Administration (NASA), Earth Science Data and Information System (ESDIS)

NASA's Earth Science Data and Information System (ESDIS) Project began investigating the use of PIDs in 2010 with the goal of assigning PIDs to all data products archived and distributed by the twelve NASA Distributed Active Archive Centers (DAACs) located across the United States. With input from its data provider community, ESDIS selected the DOI as the PID of choice because of its growing acceptance in the publishing community and its ability to function as both a unique identifier and a citable locator. ESDIS also encourages the use of the DOI to credit both the data producer and the data steward in science publications. ESDIS has developed an automated robust system that assigns and registers both structured and opaque DOIs at the collection level of the data products. All DOIs are registered using the California Digital Library (CDL) EZID registration service. One unique feature of the ESDIS DOI registration system is that it provides data producers and data distributors the flexibility of holding DOIs in "Reserve" while the data products are being developed or tested before distribution. This system also stores all DOI metadata for internal purposes before registering it with EZID. As of September 2015, ESDIS has assigned and registered over 3300 DOIs. Each data center serves a different Earth science discipline user community and, accordingly, has a unique approach and process for generating, archiving, and distributing a variety of data products. These varied approaches present a challenge for developing a uniform DOI solution. To address this challenge, ESDIS has developed processes, guidelines, and several models for creating and assigning DOIs.

CONCLUSION

Persistent identifiers can be a useful data curation tool for the Earth Observation community. The CEOS WGISS Best Practices will help foster a harmonized, responsible implementation of PIDs in global EO data repositories. A coordinated approach by data providers will enable data users to consistently cite their data sources, encouraging scientific transparency and allowing data repositories to measure their impact.

REFERENCES

[1] "ARK (Archival Resource Key) Identifiers." University of California Curation Center. Accessed 2 Oct. 2015. https://confluence.ucop.edu/display/Curation/ARK.

[2] “DataCite Homepage.” DataCite. Accessed 2 Oct. 2015. https://www.datacite.org/.

[3] "The DOI System." International DOI Foundation. Accessed 2 Oct. 2015. http://www.doi.org/.

[4] ”Persistent Identifiers Best Practices.” Committee on Earth Observation Satellites (CEOS), March 2015. http://ceos.org/ourwork/workinggroups/wgiss/current-activities/data-stewardship/