Data citation and digital identifiers for time series data / environmental research infrastructures

Huber, R; Asmi, A.; Buck, J.; de Luca, J.M.; Diepenbroek, D.; Michelini, A. and participants of the joint COOPEUS/ENVRI/EUDAT PID workshop

Part I: Use cases and scenarios for data citation and digital identifiers for time series data

Abstract:

In the age of data driven science the re-use of data and the compilation of existing data from monitoring infrastructures has become an integral part of research. For the sake of transparency and reproducibility of research it is crucial to be able to unambiguously identify data that were used as the basis of a publication. Globally unique and resolvable, persistent digital identifiers (PID) for digital data sets are an important tool to achieve this goal enabling unambiguous links between published research results and their underlying data. In addition, this unambiguous identification allows citation of data. Proven and community based examples are the usage of GenBank identifiers in the biological literature or the data citation method by using DOIs (digital object identifiers) already used widely in the scholarly literature.

Identification of discrete digital objects is simple and citation can be formatted in analogy to citing literature. The identification of still ongoing, open time series does not seem to fit this pattern. A major prerequisite for the proper use of PIDs within data citations is the persistence of both, identifiers as well as the integrity of the associated data set. This poses questions when PIDs are to be used for unfinished data sets or open time series data. Such data is typically generated within research infrastructures during long lasting experiments such as satellite missions, environmental monitoring campaigns, or in permanent installations such as natural hazard detection and early warning systems (e.g., seismic traces acquired by field stations).

Open time series data are often used in research during ongoing experiments and potentially published earlier than the underlying data set has been closed and is publicly released. It is therefore important to enable the scientific community to properly cite these data in their publications. Yet what is the meaning of “persistence” of data in ongoing time series? How does it relate to versioning? What is the granularity of a time series? In this publication we discuss and compare solutions currently used in some major European research infrastructures and propose transparent solutions which allow the citation of time series data using PIDs.

Introduction:

Permanent installations of large scale scientific facilities such as environmental observation systems have gained an increasingly important role in the modern research landscape and in the advancement of knowledge and technologies in general. Such environmental research infrastructures (RI) provide a variety of physical resources and services which are used by the scientific community to perform high-level interdisciplinary research, ranging from astronomy, atmospheric research to ocean and earth observation.

Such mostly cost intensive infrastructures provide highly complex technologies and instrumentation which are essential to address the major societal challenges related to the Earth's threatened ecosystems, climate change and geohazards. Due to their frontier science position they bear an enormous innovative and integrative potential for researchers, funding agencies, politicians and industry and are therefore essential to handle the scale and complexity of their related scientific questions.

These challenges require international collaboration especially in the field of environmental monitoring or early warning systems. Therefore, several of these research infrastructures are joint, international efforts represented by supranational organisation such as EISCAT or ARGO or represent large integrative networks of shared national infrastructures such as EMSO or EPOS. In Europe developing world-class research infrastructures is one of the top priorities of the European funding and regulatory bodies. The integration of these infrastructures is supported by the European Strategy Forum on Research Infrastructures (ESFRI) which aims to "*support a coherent and strategy-led approach to policy-making on research infrastructures in Europe, and to facilitate multilateral initiatives leading to the better use and development of research infrastructures, at EU and international level*" (Rizzuto, 2010).

Several infrastructures now are on the ESFRI roadmap towards implementation, including those mentioned above. During this process considerations regarding the share, re-use and exchange of RIs research data are of key importance. Especially the interdisciplinary character of the addressed scientific questions requires a high level of interoperability of the involved data centers regardless the discipline of the contributing scientific community. The European Commission has therefore launched several projects such as COOPEUS, ENVRI or EUDAT which aim to coordinate the harmonisation of technologies, standards, workflows as well as policies in order to facilitate the find-ability and exchange of data.

One particular problem for the management of data originating from several research infrastructures is their dynamic nature in terms of growth and quality. Research infrastructures continuously collect data during long lasting experiments such as satellite missions, environmental monitoring campaigns or observatories. Such 'open time series' data is continuously added to repository systems, databases or files and may at the same time receive continuous calibrations or quality checks. Here, the identification of discrete digital objects, a closed and publishable representation of the data, is difficult. Therefore, such entities are frequently defined not before the end of a measurement campaign or in discrete time intervals.

The problem of identity of digital objects which are by nature highly mutable can partially be solved by providing peristent digital identifiers which allow at least to unambigously identify a data source. They further allow to simply find and access the data regardless the stewardship or physical location of this data (Dürr et al. , 2011)(Parsons et al., 2010). Peristent digital identifiers are therefore the basic prerequisite for the citation of data e.g. within traditional publications but as mentioned above, mutating or growing data sets require additional strategies to balance between the need of fast publication and reproducibility of results. Dürr et al (2011) distinguished four basic use cases for identifiers which range from 'Unique Identifier', 'Unique Locator' to 'Citable Locator' and 'Scientifically Unique Identifier'. This categorisation reflects the different purposes PIDs for a wide range of applications: to allowing the internal identification of a digital object or record; to provide the technological basis to unambiguously identify a digital object in an interconnected information infrastructure or to provide a method to reference or record the usage of a digital object e.g. within scientific publications.

As mentioned above, open time series data are often used in research during ongoing experiments and potentially published earlier than the underlying data set has been closed and is publicly released. It is therefore important to enable the scientific community to unambiguously identify the data which is the basis of such a publication to ensure transparency and reproducibility of results.

Suitable and proven technologies and peristent identifier systems are available, ranging from simple unique identifiers, resolvable identifiers and persistent identifiers such as the handle based systems such as DOI or EPIC. However the individual application or usage of these PIDs e.g within publications or reports still is unclear with respect to the metadata and citation synthax needed to clarify the dynamic nature and status of the used data as well to enable the scientific community to identify the exactly the range or version of data which has been used.

Because of the significance of this problem for many European research infrastructures, the above mentioned projects COOPEUS, ENVRI and EUDAT have initiated a joint effort in cooperation with the german KOMFOR project to find a common, pragmatic solution for the application of PIDs for dynamic or time series data. Some of Europe's major research infrastructures in the field of environmental research including ARGO, EMSO, EPOS, LIFEWATCH as well as the infrastructures of the WDC-RSAT, germany and TTA, finland

Use cases: Persistent identifier usage strategies of European research infrastructures

Argo programme:

The Argo programme is an international collaboration responsible for the deployment and data management of several thousand autonomously operating autonomous drifting ocean profilers which continuously measure chemical and physical properties of sea water globally. In November 2012 the Argo programme passed the millionth ocean profile milestone.

Argo data are received and automatically processed almost continuously in real time by 10 national level data centres. Once processed data are forward onto the global telecommunications system for using in operational ocean forecasting and two mirrored Global Data Assembly Centres (GDAC) for public distribution. Once per year the data from each profiler are reviewed and checked against climatological data and nearby Argo data from different profiler. If necessary data are flagged and/or calibrated to produce a delayed mode version of data suitable for climate and oceanic heat content research.

Several needs for citing Argo data have been identified:

- A way for the Argo programme to track Argo data usage to help make the case for continued funding from national budgets.
- A way for the data used in high precision climate and heat content calculations to be cited to aid reproducible research.

The complication in Argo is constant mutation of the data on GDACs. This is both through the temporal extension of the data when new profiles are collected and updates to existing data when delayed mode quality control is done.

There are several obvious natural levels for assignment of persistent identifiers that do not meet the needs of data users or are impractical:

- When changes are made on GDAC data which is not practical or worthwhile with the number of changes that happen at GDAC level each day.
- When data for a profiler are processed in delayed mode a persistent identifier could be assigned for that time series. However, Argo data are used as a collection of profilers.

The current proposed solution to citing Argo data is as follows:

- Assigning a DOI to documents such as manuals and QC documents
- Assigning DOIs to the zipped releases of data which are then archived. Such an assignment means any scientific publication can cite the Argo data at a fixed point in time which meets the reproducible research need.

Such DOIs when minted can have a prescribed form too e.g. doi:[minting authority]/argo[rest of doi]. This would mean we can easily search the literature for usage of the data by looking for the “[minting authority]/argo” string.

The hard part is how to meet the needs of users of the real time mutating data stream such as ocean forecast groups. In this case timeliness of data balanced against quality and the zipped releases of data are not appropriate for these users. The holy grail is a single reference or identifier that can be cited. However, the concepts of data that behave in this way and operational usage of data go totally against historic data citation analogies of a data-set being akin to an unchanging book on a shelf. As a compromise the current proposal is single accession containing weekly snapshots of the Argo data-set, this can then be cited as a single DOI with an extension to the identifier that will resolve the state of Argo data to the nearest week, This proposal is being prototyped at the US National Oceanographic Data Centre.

WDC-RSAT:

Jens Klump wrote in a pers. comm via Email: As I mentioned in my previous e-mail, we do have the case of open time series in our geodetic satellite data. The procedure is as follows:
Each data product type (a class of data coming all from the same processing pipeline) is an open time series that is continuously appended until the end of the mission (or when the product is discontinued). Each product type is assigned a (parent) DOI.
In our case, each time series consists of dozens to millions of discrete data files. For historical reasons, we do not assign DOIs to each individual data file, although we could. This is done only in the case of highly processed data, e.g. global monthly gravity field models.
With assigning a DOI to a data product it becomes part of the record of science and should not be changed. It may be appended without issuing a new DOI.
It happens sometimes that data products are reprocessed, e.g. when a new version of the processing algorithm was developed. In this case we assign new DOI to all reprocessed objects. This is done to be able to distinguish different versions of the data.
An alternative could be to assign a DOI to the time series, which leads a user always to the latest version, and assign DOIs to each version of the time series (current and older versions). In addition, you can label data as "reference quality" or not, i.e. only label stable and quality controlled version of data as "reference quality".

EMSO (PANGAEA):

The European Multidisciplinary Seafloor Observatory (EMSO) is a network of seafloor observatories and platforms measuring environmental parameters related to the interaction between the geosphere, biosphere, and hydrosphere, including natural hazards. The EMSO infrastructure is geographically distributed in key sites of European waters, spanning from the Arctic, through the Atlantic and Mediterranean Sea to the Black Sea. The data infrastructure for EMSO is being designed as a distributed system. Presently, EMSO data collected during experiments at each EMSO site are locally stored and organized in catalogues or relational databases run by the responsible regional EMSO nodes. PANGAEA (Data Publisher for Earth & Environmental Science, www.pangaea.de) is one of these EMSO archive nodes and is responsible for the data management, long-term data archiving, data publication, and dissemination of the EMSO test site Koljöfjord as well as the EMSO arctic node, also known as the HAUSGARTEN site.

The arctic long term deep-sea observatory HAUSGARTEN is located at the eastern Fram strait and consists of a network of 17 permanent sampling sites. Since 1999 repeated sampling, deployment of moorings, lander systems and other long-term in situ instruments has taken place there. Due to its extreme geographic position, the HAUSGARTEN is not yet cabled or otherwise connected. Thus, the site has to be regularly visited during the ice free summer seasons. During these visits instruments are recovered, additional measurements are taken and water and sea-floor samples are collected. After the visiting cruises the recovered and collected data is processed and quality checked by the responsible principal investigators and undergoes a curatorial treatment such as formatting and metadata completion. Due to this reoccurring procedure, the resulting data sets already have a defined granularity and are archived accordingly at PANGAEA.

The Kojöfjord observatory is a permanent underwater installation located in the swedish Fjord Koljöfjord, approximately 100 km north of Gothenburg. Several instruments measuring sea water properties are arranged along a sensor string fixed at a mooring deployed at 45m water depth. The observatory is connected via a underwater cable with a shore based control cabinet from where the data can be transmitted via UMTS. In order to access the data in a standardized way, an OGC Sensor Observation Service (SOS) has been implemented and installed which delivers data according to OGC's Observations & Measurements (O&M) standard. A PANGAEA data harvesting service uses this SOS interface to regularly download data and transforms the O&M format into a proprietary PANGAEA import format. Metadata is added using the SOS OGC SensorML descriptions as well as the O&M results. The data is automatically uploaded to the PANGAEA import queue and subsequently archived as raw data in a monthly interval.

PANGAEA has pioneered in the field of data citation and is using DOIs as persistent identifiers since several years. DataCite is used as registry for PANGAEAs published data sets. In particular, PANGAEAs solution for using identifiers for observatory time series is as follows:

Currently, HAUSGARTEN data sets do not represent typical open-time series data sets as they are already reach the PANGAEA archive in a defined granularity and have frequently been scientifically exploited. Such data sets may have already been used as supplementary material for journal publications or have received a data peer-review and therefore reached a high level of quality and persistence. In this case, a DOI is assigned to the data set and registered at DataCite, the data set is citable without restrictions.

In addition to these archive formats, PANGAEA applies different PID strategies for e.g. raw or non quality checked data sets. For those data sets, DOI are internally assigned, but not registered. Further, the status of each data set is indicated in the accompanying metadata. PANGAEA hereby follows the proven concepts of the publication industry and assigns comparable publication status labels which range from ‘unpublished data set’ (=in prep.) or ‘DOI registration in process’ (= in press) to finally ‘published’. Consequently, the Koljöfjord raw data sets are labelled as ‘unpublished data sets’ unless they receive a quality check procedure.

However, there is a growing demand to accelerate both, citability and availability of open time series data. Therefore PANGAEA is also investigating additional strategies to handle such data. For example, citability of open time series could be increased and enabled earlier if a hierarchical model would be applied to the organisation of resulting data sets. A citable parent data set for which a DOI is assigned could here serve as a container for smaller child data sets which are continuously added to this parent. Another option would be to allow to simply assign a DOI to a initial data set to which data is continuously appended during a measurement campaign which would significantly speed up public availability of this data but would result in a certain loss of persistence. Both, metadata as well as the citation of such open time series data sets need to indicate its preliminary character.

EPOS:

The European Plate Observing System (EPOS) is an integrative research infrastructure for solid Earth Science in Europe. It is is composed by several communities in the solid Earth sciences and typology of data is very much diversified (see http://www.epos-eu.org/ride/). In general, the data span typologies ranging from time series acquired by continuously recording seismic stations (e.g., 100 Hz sampling rate, three component instruments) to GPS data and laboratory experiment data bases. In the preparatory phase of EPOS, much attention has been given to the seismic data since considered perhaps the most evolved one because it has been acquiring and storing digital data since 25-30 years. The community is evolving to the use of persistent identifiers and metadata.

One of the problems encountered by the community when seeking to uniquely identify the data digital objects is the incompleteness of the data acquired. This problem follows from data transmission from the remote station to the central data center and it consists of the presence of data gaps. These gaps are then filled in as the bandwidth of the transmission widens (i.e., the data compression varies depending on the amplitude of the signal recorded). Thus the problem of uniquely referring to data which may have undergone changes (addition of data) is very much felt especially because some users start working with these data before they are complete, i.e. they need references requiring the associations with separate PIDs. With the help of clever PID mechanisms referring specifically to the different versions of the “same” file one could help applications always to get the last version. It is in the responsibility of the data centres to maintain the stability of the references.

Windows: EPOS wants to let users define windows (see above) that include relevant phenomena, do this for many different types of streams that all cover the same time window, and add semantics to it (label, characteristics, etc) (which is the typical metadata). This could be done in two ways: a) the data is extracted from the files and copied to new data objects or b) the metadata simply refers to the PIDs (including fragment identifiers) to create a virtual aggregation. Given the likely scenario that many users will want to identify different phenomena, the first option is not recommended since a lot of data copying will occur. When using registered and thus stable PIDs for all files, no problems can occur when using the virtual method. The metadata object simply refers to many PIDs which, if done correctly, point to the different physical copies that are stored in some centre.

EPOS is currently experimenting assignment of PIDs to its data using EPIC (http://www.pidconsortium.eu). The real-time case described above and though important has not been addressed and seismic data are assigned PID on a 1-day data file object basis. This is currently carried out automatically by CINECA which is the main partner of EPOS within the EUDAT project (http://www.eudat.eu) and which is also replicating the seismological data on their storage. A solution for adopting PIDs to register data files automatically at a very early stage while being capable to keep track of all the different versions is under study. As in other domains, replicas would be registered in the PID record as well so that data management becomes simpler.

TTA "National Research Data of Finland"

TTA ("Tutkimuksen tietojärjestelmät" in Finnish) is a national project for storing all of the research data produced by public funds in Finland. The PID chosen in general for this work is URN, a well suitable PID for many traditional data types (e.g. libraries, etc). As several of the environmental science datasets (atmosphic and biospheric datasets) are also included in TTA data storage system, the issues of especially granularity and versioning have proven to be difficult. The issue is not solved yet, but we are observing the current developments in the field.

For example, many of the atmospheric datasets are stored in very high time resolution, and depending of the use case, the data citation could be relevant to specific short time period (from minutes to hours) or to decades long timeseries. Thus choosing a single granularity for the datasets is challenging. Also the versioning is of critical importance in data citation, as many analysis products could be dependent on the version the data used in the analysis. The data versions are usually not inclusion of missing data in between (as in some of the other cases above), but more commonly new improved calibrations of the instruments, resulting in more physically representable end products. For reproducibility, the data user must be able to detect the version of the dataset included, but also to be able to find the latest ("best") dataset derived from the same source.

LifeWatch

LifeWatch represents Europes main infrastructure related to biodiversity research. It collects a significant amount of data continuously measured during environmental monitoring campaigns. Typical example are installations deployed by LifeWatch Spain such as autonomous multisensorial devices to monitor and forecast the dynamics of toxic cyanobacteriae in a remote water reservoir. These installations, typically running for ca. 1 month before revisions provide continous measurements of e.g. Fluorometers and these raw data (level 4) are used within early warning systems prior to QC routines. These early warning systems are designed to detect toxic phycocyanin concentrations in order to generate various warning levels. The underlying raw flouarometer measurents which represent open time series data are used as a pilot for the application of PIDs for dynamic, open data series within LifeWatch.

Currently, LifeWatch makes use of EPIC's PID service and PIDs are being tested within wireless sensor network (WSN) deployments to assign unique IDs to WSN resources (e.g. sensor nodes), provide standardized access to associate (meta-)data, e.g. via standardized attributes (URL, SERIAL, LOCATION, DATASHEET etc), to assign unique IDs to WSN-related concepts that feature in the meta-data e.g. to identify phenomenon that is being observed.

ICOS

ICOS represents a world-class research infrastructure to quantify and understand greenhouse gas fluxes in Europe and key regions of interest for Europe. The infrastructure has important implications for climate policy such as the detection of ‘hot-spots’ of the carbon cycle and

it defines backbone observations and metrology for treaty verification. ICOS provides long-term measurements at a network of sites including 50 ecosystem observation sites, 50 atmospheric concentration sites and 10 ocean ship-lines & stations. The performance of the ensemble is greater than the one of any combination of national networks and has a large potential for discoveries and re-analysis. It further provides the metrology for emission and sink monitoring and verification. All ICOS stations have deployed identical sensors. Data processing is centralized in thematic centers, in addition near real time data delivery and processing is implemented.

PID are intended to be used for near real time measurements - as well as for long time series of validated measurements. Initially ICOS intends to apply PIDs for static, archived annual releases of atmospheric data in accordance to the ICOS policy as defined in Bejing 2013: “A persistent identifier with the information of the data providers/authors will be accompanied with every ICOS data set, tracking the use of the ICOS data to individual research paper, with a strong traceability even to the individual sites/instruments”. Currently, granularity of data and the definitioj of data sets are investigated. The application of PIDs for near real time data is planned and currently ICOS discusses related problematc issues i.e. for data which can evolve from day to day, due to quality check back propagation in time or the potential application of PID for data dynamically delivered through database queries when datasets are generated on-the-fly, by user requests as extracted/subset from a relational database.

Use case analysis: A classification of ‘PIDs for open time series’ strategies

During the planning of PID usage for open time series data set several decisions have to be met regarding the granularity of data sets, timing of PID assignment, identifier target object definition and timing of registration of PIDs. For example the current practise to define a digital object for which a PID is assigned differs among research infrastructures and we can distinguish between several strategies to define ‘PID –able’ objects which can roughly be categorized as follows::

Placeholder strategies:

Abstract data set strategy: Creating an abstract or initial data set as an initial placeholder:

In some cases it might be desirable to have a PID available at a very early stage of a measurement campaign. A pragmatic approach would be to define an **abstract or initial data set** as a placeholder to represent the open time series data set. Such data sets not necessarily contain data but represent an abstract concept of e.g the expected data produced during a measurement campaign. Basically it consists of an initial metadata description and an empty data set or a minimal set of initial data. Data is either continuously added to the abstract or initial data set or bulk inserted after the end of the measurement campaign. Alternatively, an abstract data set can be populated by adding linked child data sets within its metadata each containing data from e.g different disciplines or instruments. Such a strategy was choosen e.g. by the WDC-RSAT for their data products. PANGAEA considers to create initial data sets for its marine observatory real time data streams delivered by the Koljoefjord observatory which would be continously updated during a defined period of time of the measurement campaign.

Such data sets require special considerations regarding citation, metadata as well as the registration of identifiers. In analogy to traditional citations which frequently state ‘in press’ or ‘in prep’ for preliminary research results, PANGEA indicates within the citation as well as the metadata of such data that the data set is ‘unpublished’ or ‘unfinished, the registration of the data set DOIs is hold back until its finalisation.

Delegate data set strategy: Creating a delegate data set as a permanent placeholder:

If it is most unclear how a data set develops, e.g. when continuous reprocessing is expected it might be most convenient to assign a PID to a delegate data set which serves as a permanent placeholder for this data set. Such a delegate data set could be a metadata description or processing documentation file or any digital object which is suitable to document the evolution of the original dataset. Once the original data set reaches a stable condition, a new PID will be assigned. Such a strategy was considered by ARGO in order to avoid to assign PIDs to their highly dynamic data sets unless these data sets became persistent.

One of the advantages of using such a static, delegate data set is that it can immediately be cited and registered at e.g. a DOI registration agency. It is however critical to track provenance and evolution of such data to ensure its reproducibility. Further it is challenging to provide suitable links e.g within the metadata to the archival version of the original data set once it is finalised and receives its own PID.

Data product strategy: Providing data products prior to the publication of raw data

Data products such as maps, charts or images represent processing results, visualisations or modelling results derived from complex raw data which is continuously measured during an experiment or campaign. Examples for high quality data products are e.g. processed satellite images to provide vegetation maps, or atmospheric model results. Such data products also can represent different processing levels which are derived from the raw data, such level distinction is typically done during satellite imaginary processing e.g by the WDC-RSAT. The relationship or links to the underlying raw data is either within the metadata as practised by PANGAEA and WDC-RSAT, or -in some cases- the raw data is not preserved.

The resulting data product represents a discrete, static data set which can be archived and described by a suitable set of metadata and identified by a PID. As such the data product is available for citation and reuse.

Versioning strategy:

Versioning strategy: Creating new data set versions after each update or reprocessing step

If provenance information needs to be perfectly preserved, time series data needs to be stored as a new version of the original data set following each update of the data set. Reasons for creating new versions can be for example reprocessing of open time series data which may occur when sensor calibration adjustments or quality check routines reveal that the correction of previously measured is required. In this case a new reprocessing algorithm needs to be developed and applied on the raw data. The most obvious reason for versioning is of course simply an update of a data set following the addition of data.

Reprocessing versioning is e.g. applied to ARGO float data during their delayed mode quality control and to WDC-RSAT satellite data. This strategy is also considered by TTA following each instrument calibration. Such new versions are stored as discrete, static data sets, described by a suitable set of metadata which includes the provenance as well as the processing information. Each version is identified by an citable PID and is preserved in conjunction with their underlying raw data sets.

Fragmenting strategy:

Fragmenting strategy: Splitting time series data into discrete fragments or subsets of data

A very common approach to define the granularity of data sets for long term time series data is to store a subset of the incoming data in e.g. monthly or weekly intervals. The result of this very pragmatic approach are discrete data sets containing the data of this period of time only which can even be semi-automatically created and archived.

This data set definition strategy has been chosen by many EMSO observatories, e.g those maintained by the EUROSITES consortium and PANGAEA which both are stored after QC routines have been applied whereas only the latter currently assigns PIDs. In some cases, such subsets of data are stored as raw, unchecked data, for example the EMSO Koljoefjord data which is identified by unregistered DOIs and the preliminaty character of the data is identifies in the metadata.

Such subsets of the time series data are stored as discrete data sets which can unambiguously be identified by an set of metadata and a citable PID. In case these data sets are static and quality checked they are very well suited for archiving, data publication, citation and thus reuse.

Requirements and recommendations for PID usage within Research infrastructures:

A proposal for citation rules and synthax for dynamic data:

<author> . (**<release date range>)**: <*dataset title*>. [version: <version>|subset: <temporal range>]. <publisher>.[[**<resource type (growing dataset , evolving dataset , fragmented dataset)>]].** <PID>@<fragment identifier>. [accessed: <access date>]

Example:

Doe, J. **(2009-2011)**: *Dynamic Data Set Title*. version: 1.2. Responsible Data Archive. [**evolving dataset].** PID:123456789@version=1.2

Doe, J. **(2009-2011)**: *Dynamic Data Set Title*. subset: 2010-01-01 - 2010-12-13. Responsible Data Archive. **[growing dataset].** PID:123456789@range=2010-01-01-2010-12-13

Doe, J. **(2009-2011)**: *Dynamic Data Set Title*. version: 1.2. Responsible Data Archive. [**fragmented dataset].** PID:123456789. accessed: 2012-12-01@version=1.2

Literature

Bellini, E. et al. (2012), APARSEN Persistent Identifiers Interoperability Framework, Public Report, Fondazione Rinascimento Digitale, Florence, Italy. [online] Available from: http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=D22.1+Persistent+Identifiers+Interoperability+Framework

Bütikofer, N. (2009), Catalogue of criteria for assessing the trustworthiness of PI systems, nestor Materialien, Niedersächsische Staats und Universitätsbibliothek Göttingen, Göttingen, Germany. [online] Available from: http://nbn-resolving.de/urn:nbn:de:0008-20080710227

DIN (2012), Anforderungen an die langfristige Handhabung persistenter Identifikatoren (Persistent Identifier), Standard, Deutsches Institut für Normung, Berlin, Germany. [online] Available from: http://www.entwuerfe.din.de/cmd?level=tpl-art-detailansicht&committeeid=54738855&artid=150707873&bcrumblevel=2&languageid=de

Duerr, R., Downs, R., Tilmes, C., Barkstrom, B., Lenhardt, W. C., Glassy, J., Bermudez, L., & Slaughter, P. (2011). On the utility of identification schemes for digital earth science data: An assessment and recommendations. Earth Science Informatics, 4, 139-160.

Keitel, C. (2012), DIN Standard 31644 and nestor certification, Fondazione Rinascimento Digitale, Florence, Italy. [online] Available from: http://nbn.depositolegale.it/urn:nbn:it:frd-9273 (Accessed 31 January 2013)

Parsons, MA, R Duerr, and JB Minster. 2010. Data citation and peer-review. Eos, Transactions of the American Geophysical Union. 91(34):297-98. doi:10.1029/2010EO340001

Simons, N. (2012), Implementing DOIs for Research Data, D-Lib, 18(5/6), doi:10.1045/may2012-simons. [online] Available from: http://dx.doi.org/10.1045/may2012-simons (Accessed 21 May 2012)

Rizzuto, C. "Research Infrastructures and the Europe 2020 Strategy." Science Business (2010).

Uhlir, P. F. (2012), For Attribution -- Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop, National Academies Press, Washington, DC. [online] Available from: https://download.nap.edu/openbook.php?record_id=13564&page=R1 (Accessed 21 November 2012)

Wren, J. D. (2008), URL decay in MEDLINE—a 4-year follow-up study, Bioinf., 24(11), 1381 –1385, doi:10.1093/bioinformatics/btn127. [online] Available from: http://bioinformatics.oxfordjournals.org/content/24/11/1381.abstract

Part II Minutes of the joint COOPEUS, ENVRI and EUDAT PID workshop

Joint COOPEUS, ENVRI and EUDAT workshop on persistent digital identifiers (PID) for open time series data

in cooperation with **KomFor** and RDA

Date: 25.-26. June, 2013
Location: room 2060, MARUM, Bremen, Germany
Hosts: COOPEUS, ENVRI, EUDAT

Motivation:
A major prerequisite for the proper use of persistent identifiers (**PID**) e.g. within data citations is the persistence of both, identifiers as well as the integrity of the associated data set. This poses questions when PIDs are to be used for unfinished data sets or **open time series data**. Such data is typically generated within research infrastructures (RI) during long lasting experiments such as satellite missions, environmental monitoring campaigns, or in permanent installations such as natural hazard detection and early warning systems. Open time series data are often used in research during ongoing experiments and potentially published earlier than the underlying data set has been closed and is publicly released. It is therefore important to enable the scientific community to properly cite these data in their publications and the proper use of PIDs is of key importance to reach this goal.

This workshop aims to continue ongoing joint efforts between the European projects COOPEUS, ENVRI and EUDAT in particular the common goals defined during the ***strategic workshop on future harmonization of data sharing among Research Infrastructures*** during the EGU 2013. We will discuss and compare solutions currently used in some major European research infrastructures with the overall goal to find a best practise solution for the usage of PIDs for open time series data.

Audience:

Representatives of national and international research infrastructures (e.g. EuroARGO, EISCAT, EMSO, EPOS, ICOS, LIFEWATCH); KOMFOR, ESFRI COORD project representatives (COOPEUS, EUDAT, ENVRI, iCORDI, ODIP); international initiatives: RDA, DataCite, WDS

Minutes:

After the short introductory and welcome talk by Robert Huber, Christoph Waldmann gave an overview on the activities of COOPEUS (the host of the meeting) and showed the cooperative context with the contributing European (ENVRI, EUDAT) as well as international initiatives and projects (RDA etc..)

The session then started with talks related to PID technologies, their providers and users.

- Ulrich Schwardmann gave an introduction to EPIC and its interesting features with regards to the support of fragment indication etc and
- Frauke Ziedorn
  gave an overview on DataCite and the usage of DOI for data citation.
- Michael Diepenbroek showed some examples how applied data citation works within PANGAEA and demonstrated the consequent linkages between publication and research data.
- Peter Wittenberg (EUDAT, RDA) gave together with Tobias Weigel (RDA, DKRZ) a presentation about the Data organisation in RDA perspective and the RDA working groups in general. Peter also showed a nice illustration of the problem of fragmented data, a specific problem within EPOS where data sets are frequently initially full of gaps which subsequently are filled.

The following discussion focussed on specific, basic requirements for PIDs intended to be applied for open time series data sets. The group concluded that existing technologies and initiatives are sufficient and can be used, given that some additional requirements are fulfilled. In particular the following list has been considered to be valuable:

- Fragmentation support
- Integrity (e.g hash tag, but community specific)
- Versioning support
- Aggregation / Relation support
- Notion of time as attribute

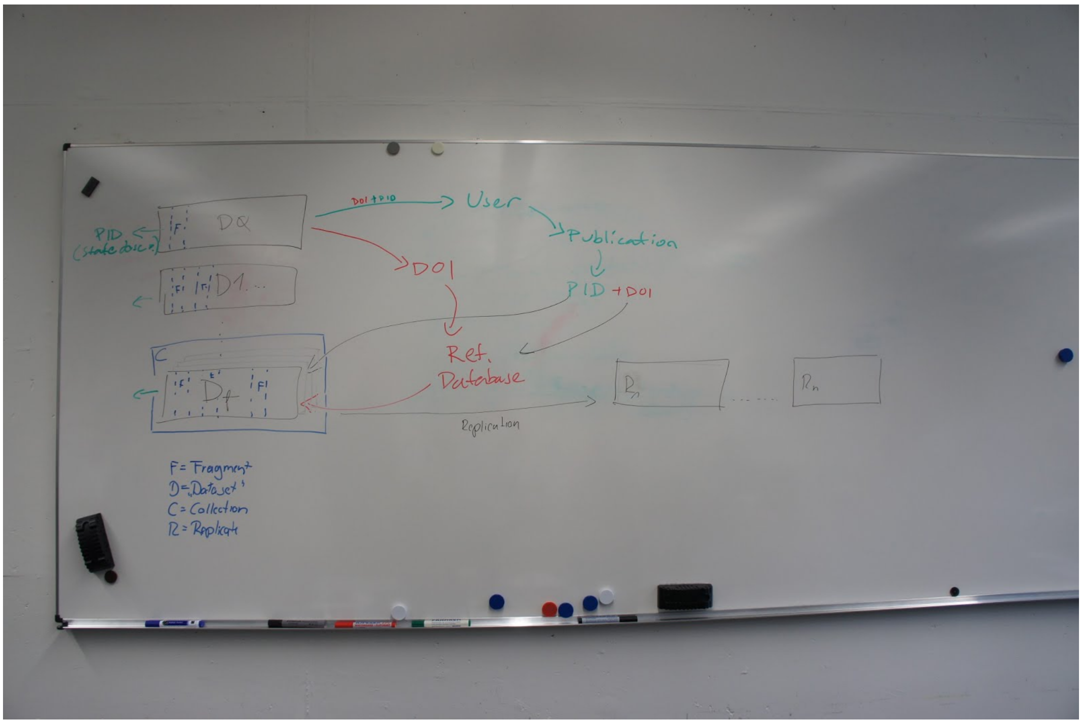
After this discussion, several case studies have been which introduced the current handling of PIDs for open times series within the RIs:

- Justin Buck presented impressive insights into the data management of ARGO in particular he showed the problematic of ARGO data which frequently mutates after QC procedures and their thoughts about using DOIs.
- Andree Behnken showed how the EMSO test site data from the Koljoefjord observatory is treated by PANGAEA and their usage of DOIs and data citation examples.
- Jesus Marco de Lucas gave an overview on the Spanish LIFEWATCH data architecture and their planned application of EPIC IDs in cooperation with EUDAT.
- Brian Wee introduced to the US NEON as well as the US Global Change Information System and their ideas about PID usage.

During the second day, the presentation of RI use cases has been continued.

- Jerome Tarniewicz gave an introduction to the distributed ICOS data architecture and their foreseen PID assignment for ICOS data set annual releases.
- Massimo Fares showed many interesting facts on the EPOS data landscape and their cooperation with EUDAT regarding PIDs.

The discussions during and after the presentations showed that each of these case studies demonstrated that both, requirements and strategies are heterogeneous but many commonalities exist. During the coffe break a first draft of a common, generic model has been developed at the flip chart:



Following the use cases first results of the Google docs use case analysis which have been presented wherein 3 general strategies to apply PIDs for open time series data have been identified,

1) Placeholder strategies:
2) Versioning strategies
3) Fragmenting strategies

The group agreed that most probably a combination of these strategies would allow proper usage of PIDs for OTD given that the used PIDs meet the requirements.

After the overview presentations and presentations of use cases, a fruitful discussion started about requirements for PIDs for open time series and basic requirements. The idea was to find the “10 golden rules for the selection and use of PIDs for open time series”.

Key point is to understand the conflicting interests of two uses of persistent identifiers:

1. For data citation and reference credit purposes (“Citation”)
2. For getting the data set actually used in a study (“Data Management”)
   One method suggested was to combine citation approach (registered DOI) with a data management addition (formatted PID which can be resolved by the data centre) by giving both when using datasets. Applicability of such model is still discussed.

The group has defined some basic requirements regarding metadata as well as citation of data which include: indication of access date, request, portion or fragment of original dataset, versioning, nature of dynamicity, creation date etc. In particular these requirements are:

The nine golden rules for selection and use of PIDs for open timeseries :

1. Persistence: Each datacenter must define a versioning and preservation strategy
2. PIDs must be persistent, even when datasets are deleted or changed. However, PID should give info on fate of data.
3. PIDs must be organized according to its use ... Publication vs. Data management recommendation: assign PIDs for „events of interest“ in datasets.
4. Time-fragmentation support (resolution). PIDs should as minimum specify „time of access“ and „time-frame“ (=temporal subsample) used. (Info assigned in attributes)
5. Transparency: level of dynamicity in the data-set must be defined in PID.
   a. a. dataset **growing** over time = data added at the end.
   b. b. Dataset is **evolving** (= changed back in time ex filling gaps, recalibrated)
   c. **c. fragmented**
6. Procedure for PID generation must be consistent, transparent , documented and financial affordable
7. PIDs should be assigned early as possible ...
8. Levels of granularities must be standardized within each scientific field
9. Data center must provide a citation template

Open timeseries specific metadata requirements / our recommendations

1) Level of dynamicity in the data-set.
2) Include timestamp to identify „*version*“ (identify time relative to changes to the dataset)
3) Fragment identification
   - o Define temporal subsample (ex from Jan 2001-dec2011)
   - o Specification of sub-sets (ex only Tuesdays bt 18:00-19:00)
4) Content of request selection used
5) Creation date of the whole timeseries

The recommendations for citation in relation to PIDs for open timeseries

Example: Smith, John, 1993-2000, „result from xxx expedition“, DATA-TWO, PID/DOI, Access-date,

1) Dataset must have DOI and/or PID with standard info, but also specifically for open timeseries information regarding the level of dynamicity.
1) (example: for MS, Nature assigns DOI to „web-only“ version. The DOI remains as it becomes a standard version and only then is it entered by the DBs like webofscience)
2) Access-date should be included
3) Needed indicator like „in press“ stating that this is a dynamic dataset

Brian Wee introduced the group to the US Interagency Data Stewardship Citations provider guidelines see: (http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines) for inspiration. The group agreed that these rules could be adapted or serve as a template for metadata element definition. It should be tested in use case.

The group again agreed that existing PID technologies are sufficient, however some specific requirements for PID providers have been discussed including, fragment and coverage support (in terms of parameterisation which already is supported by EPIC).

It was agreed by all participants to continue to work on the Google Docs which has been initiated several months ago by members of COOPEUS, ENVRI and EUDAT. Here, the participants will describe and discuss the RI uses cases, possible PID strategies as well as the ‘golden rules’ results of the workshop. In general, the outcome will be summed up soon as a report or a paper publication.

It was agreed that a more formalized way of drafting PID workflows etc would be highly desirable, ideally as ODP model as an extension of the ENVRI RM. Some flip chart sketches are available for initial considerations and will be sent to the ENVRI WP3 members.

The group agreed in trying to set up a new RDA working group. Ari Asmi has offered to draft a proposal for a RDA working group, aimed at the development and definition of persistent time series data indices. This in of interest for many ENVRI RI:s, as they could participate in development of such method to better facilitate the data usage and correct attribution. This work will start after the WS notes are distributed, and hopefully the Working Group can be accepted before next RDA plenary, although this is a very thigh schedule.

Agenda:

Day 1, 25.6.2013

1. **Welcome and introduction**

   12:00-12:30 Lunch and registration

   12:30-13:00 *Robert Huber, all participants*: welcome and introduction

   13:00-13:15 *Christoph Waldmann*: Introduction to COOPEUS, ENVRI, EUDAT

2. **Presentations: status quo and updates:**

   a. Overview presentations on PIDs, data citation and linked data

      13:15-13:30 *Ulrich Schwardmann*: Introduction to EPIC

      13:30-13:45 *Frauke Ziedorn*: DataCite and DOIs

      13:45-14:00 *Michael Diepenbroek (or other PANGAEA rep.)*: Data publication and data citation ‘in the wild’: some PANGAEA examples

      14:00-14:15 *Tobias Weigel*: PIDs and the Research Data Alliance

      14:15-14:30 *Peter Wittenburg*: Data organisation, the RDA perspective

      14:30-14:45 Coffee break

      14:45-15:30 Discussion: requirements for PIDs for open time series; intended result: short common catalogue of basic requirements, technical recommendations

   b. Status quo reports on current or planned usage of PIDs for open time series data at different ENV RIs

      15:30-15:45 *Justin Buck*: ARGO

      15:45-16:00 *Andree Behnken*: EMSO, the Koljöfjord test site

      16:15-16:30 *Jesus Marco de Lucas*:LIFEWATCH (Spain)

      16:30-16:45 *Brian Wee*: NEON ; US Global Change Information System

      ~ 17:00 End of day 1

Day 2, 26.6.2013

   c. Status quo reports continued

      09:00-09:15 *Jérôme Tarniewicz*: ICOS

      09:30-09:45 *Massimo Fares*: EPOS

      09:45-10:15 slot for spontaneous presentations and/or discussion of use cases

d. Presentation of results of the joint ENVR I/ COOPEUS / EUDAT case study analysis

10:15-10:30 *Robert Huber*: A classification of ‘PIDs for open time series’ strategies - results of the case study analysis

10:30-10:45 Coffee break

3. **Discussion: best practice for PIDs for open time series**
   a. Discussion on PID best practice for open time series

10:45:12:30 Determining best practises and workflows; intended result: definition of ‘*The 10 golden rules for the selection and use of PIDs for open time series*’

12:30-13:30 Lunch break

13:30:14:30 Determining best practises and workflows; continued

4. **Preparing further steps:**

14:30-15:00 Planning a joint publication on the workshop issue: writing team, journal selection etc..

15:00-15:15 Coffee break

15:15-16:00 Cooperation with RDA, support and input for RDA PID working groups

16:00 End of meeting

Participants

Aguilar, Fernando	CSIC	LIFEWATCH
Asmi,Ari	Uni Helsinki	EMVRI/COOPEUS
Azzarone, Adriano	INGV	EMSO
Behnken, Andree	UniHB	PANGAEA
Buck, Justin	BODC	ARGO
de Lucas, Jesus Marco	CSIC	LIFEWATCH
Diepenbroek, Michael	UniHB	PANGAEA, KOMFOR, WDS
Fares, Massimo	INGV	EPOS
Hausstein, Brigitte	GESIS	RDA
Hellström, Margareta	Lund University	ICOS
Huber, Robert	UniHB	EMSO, ENVRI
Koop-Jakobsen, Ketil	UniHB	COOPEUS

Marcucci, Nicola	INGV	EMSO
Pfeil, Benjamin	Uni Bergen	ICOS
Riedel, Morris	FZ Jülich	EUDAT
Schindler, Uwe	UniHB	KOMFOR
Schneider, Nadine	LSCE	ICOS
Schwardmann, Ulrich	GWDG	EPIC
Spinuso, Alessandro	KNMI	EPOS
Tarniewicz, Jérôme	LSCE	ICOS
Waldmann, Christoph	UniHB	COOPEUS
Wee, Brian	Smithsonian	NEON
Weigel,Tobias	DKRZ	RDA
Wittenburg, Peter	MPI Nijmegen	EUDAT
Ziedorn, Frauke	TIB Hannover	DataCite, KOMFOR
Sanchez Cano,Francisco Manuel	CSIC	LIFEWATCH