Geophysical Research Abstracts
Vol. 18, EGU2016-14673, 2016
EGU General Assembly 2016

Data DOIs – virtues and weak points from the perspective of the data journal ESSD

Hans Pfeiffenberger (1) and David Carlson (2)

(1) Alfred Wegener Institute, Services/Computing and Data Centre, Bremerhaven, Germany (hans.pfeiffenberger@awi.de), (2) World Climate Research Programme, WMO, Geneva, Switzerland (dcarlson@wmo.int)

Fundamentally, data publication seeks to provide improved access, trust and utility for data users accompanied by improved recognition for data providers.

The features of persistent identifier schemes used in this context and the policies of providers of their resolver systems can greatly contribute towards these aims! Unfortunately, we also recognize situations where identifier systems undermine some of our lofty data publication aims.

From our 7 years of successful experience with the data journal ESSD, we will present our practical observations as well as future requirements on the DOI-system in particular, such as:

* Most scientists contributing to or drawing benefit from ESSD seem to understand the notion of published data in the sense of a static, citable entity to build on, relying on its long term availability and integrity (as is the case with journal articles), secured by DOIs.

* It is still true that many data providers – individuals or projects – can not find a repository for their (type of) data which would provide a DOI (or other persistent identifier) and that many data centers and databases holding valuable data have a hard time adding DOIs as a feature of their systems for conceptual, legacy or funding reasons.

* The typical evolution of many important data collections or data products – such as the Global Carbon Budget – can adequately be tracked with yearly (or less frequent) issues of dataset and data article, each having a new DOI. However, issuing a new DOI for a dataset extended in time clashes with many producers' wish to amass all references to the developing dataset under one citation (and one DOI).

* Current practises with DOIs and metadata associated with DOIs need to be clarified and amended so the identifier systems can and do express and track the relationships between versions of data, addressing extensions, corrections, modifications, revisions and retractions in an unambiguous (and machine readable) manner, while also removing extraneous information (dates, page numbers) from the identifier itself.

* Similarly, we recognize a clear need to hold one or more copies of important data in different locations (continents), under different governance and technologies but to still be able to identify these copies as intellectually identical. Today, however, we observe identical datasets held in separate repositories with different metadata and DOIs without a process to easily identify the duplicates.

* The dataset and data article – coupled with a research article citing both and drawing conclusions and perhaps with an additional article describing and publishing the software – offer a valid and vivid example of linking objects reliably across many players' (publishers') systems. Too often, however, establishing such linkages requires substantial human intervention to initiate and synchronize the necessary processes, in particular to ensure high-fidelity mutual citation among the individual items.

As editors of a data journal we firmly embrace the notion of persistence and persistent identifiers, in particular DOIs. We see the need and applaud all efforts to enhance its features to serve the important needs of published and linked data.