



A future data environment - reusability vs. citability and synchronisation vs. ingestion

D. Fleischer

Helmholtz-Zentrum für Ozeanforschung Kiel GEOMAR, Germany (dfleischer@geomar.de)

During the last decades data managers dedicated their work to the pursuit for importable data. In the recent years this chase seems to come to an end while funding organisations assume that the approach of data publications with citable data sets will eliminate denial of scientists to commit their data. But is this true for all problems we are facing at the edge of a data avalanche and data intensive science?

The concept of citable data is a logical consequence from the connection of points. Potential data providers in the past complained usually about the missing of a credit assignment for data providers and they still do. The selected way of DOI captured data sets is perfectly fitting into the credit system of publisher driven publications with countable citations. This system is well known by scientists for approximately 400 years now. Unfortunately, there is a double bind situation between citeability and reusability. While cooperation of publishers and data archives are coming into existence, it is necessary to get one question clear: "Is it really worth while in the twenty-first century to force data into the publication process of the seventeenth century?"

Data publications enable easy citability, but do not support easy data reusability for future users. Additional problems occur in such an environment while taking into account the chances of collaborative data corrections in the institutional repository.

The future with huge amounts of data connected with publications makes reconsideration towards a more integrated approach reasonable. In the past data archives were the only infrastructures taking care of long-term data retrievability and availability. Nevertheless, they were never a part of the scientific process from data creation, analysis, interpretation and publication. Data archives were regarded as isolated islands in the sea of scientific data. Accordingly scientists considered data publications like a stumbling stone in their daily routines and still do. The creation of data set as additional publications is an additional workload a lot of scientists are not yet convinced about.

These times are coming to an end now because of the efforts of the funding organisations and the increased awareness of scientific institutions. Right now data archives have their expertise in retrievability and availability, but the new demand of data provenance is not yet included in their systems. So why not taking the chance of the scientific institutes sneaking in and split the workload of retrievability and provenance. Such an integrated data environment will be characterized by increased functionality, creditability and structured data from the creation and everything accompanied by data managers. The Kiel Data Management Infrastructure is creating such an institutional provenance system for the scientific site of Kiel.

Having data sets up to date by synchronisation with institutional provenance system capturing all changes and improvements right where they happen. A sophisticated and scalable landscape needs to combine advantages of the existing data centers such as the usability and retrievability functionality with the advantages of decentralised data capturing and provenance. This data environment with synchronisation features and creditability of scientific data to future users would be capable of the future tasks.