A Future Data Environment

reusability vs. citability and synchronisation vs. ingestion

Dirk Fleischer, Pina Springer, Andreas Czerniak
datamanagement@geomar.de

During the last decades data managers dedicated their work to the pursuit for importable data. In the recent years this chase seams 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.

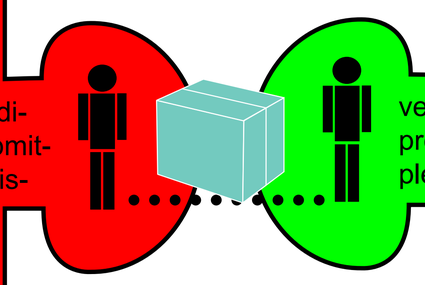
Data Environment Right Now

Providing Data

Scientists are struggeling with the additional work required to actually prepare a dataset for submission to an archive. There is not doubt about it that this work in necessary, but the down side of this workload is that such data submissions are regarded as one time effort and future improvemnts and additional data ware only rarely communicated to the archive or have to be resubmitted completely. Do we realy want data submissions to be more like a dead end road or something more up-to-date?

Reusability

Reusability of data is right now not really the main focuse for scientists while creating dataset for submission to an archive. Access to the data right now usually requires a lot of post processing like elimination of not needed parts of a dataset. Since reusability seams not to be in the focus of the users it also is not completly in the focuse of publishers, considering data access by author or some print publication. But are data always fixed to their initial purpose or is it necessary to safe data as a frozen snapshot to provide this to the scientific community? Convincing data providers to actually provide their data it was the main focus during the last decades to provide the citation of datasets just like print publications. But is a print publicaion or fix formated dataset really the right thing for reuse?



Ingestion

The ingestion process of data archives required right is dominated by the consistency checks and the communiation with the scientist due to quality issues before the data are actually imported into the data system. This is a very time consuming and uncomfortable procedure for data manager and data provider. Facing the future data avalanche do we really need a completly human based ingestion and can we afford the time and money that this process will cost?

Citability

For years it was necessary to seek a credit system assuring the credit to the data providers and creators. The wish to convince scientists was and still is very high. Accordingly the aspired system is similar to the citation and publication system driven by commercial publishers and already known to the scientific community.
The discussion of the last decade about the impact and citation of usual print publications seams to be neglected. But data and data publications will be quite different from the known publications as print media. Reconsideration for the term: ‘data publication’ seams more then overdue.
The generale citation procedure did not keep up with the technical possibilities. Why not using version controlled citation bookmarks instead of freezing the whole cited dataset? The chance to improve your data is not a bug its a feature of the digital age, why not using it?

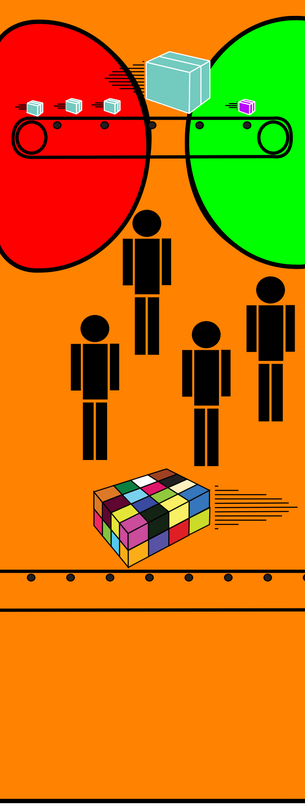
The Future Data Environment

Synchronisation

Data provider capture their data already during the creation process in a structured infrastructure. This master copy of data is the source for all further analysis procedures and the data quality is improved during this process already within the infrastructure.
Published and archived data can be improved on a technical basis by recording all improvemnts. This update service by the initial data provider is like a synchronisation with the publicly accessible data keeping the archive and the community always up-to-date. The dissemination as a key feature of archives would stay there.

Reusability

Due to the fact that a publication includes for each data archive the link to the applyed extraction query. Reuse of the presented and analysed dataset is very easy. No postprocessing is required to achieve the very same dataset used by the presenting author.
Reusability of data is the major point in the future data environment. This change in paradigmm requires the community to get involved by commenting and evaluating the used dataset. The obligation to evaluate data sets due to the particular use case is very valuable for reusability. Data providers sometimes do not see the hidden potentials within their data this potential needs to be revealed by third party evaluation.



Ingestion

In a future data environment the human interactions during the ingestion process of data into an archiving facility is reduced to a bare minimum. Data flow from a structured data capturing facility into the archive. The data were already used, analysed and improved based on this capturing infrastructure so inconsistency may not be an issue and the synchronisation will take care of the quality of the presented data. The combination of early data capturing and and archiving without leaving the connection to the initial capturing infrastructure enables faster and easier ingestion models.

Citability

Giving credit to the used datasets will become mainly a technical problem instead of a cultural problem as it was.
Data citations within a publication are hardly writen in human read able way and its not necessary. Machines follow up the links to the data archive and retrieve the actual data sets. Enabling the acquisition of citations in a recursive way will push everything towards the reusability of data for potential users.
The data can be tracked by the same way fllowing the same link as a hyperlink everyone could dig down to the bare numbers and the citation of the source publication of each number. Everything based on the data aggregation of a datawarehouse.