PRICING POLICY FOR DIGITAL RESEARCH DATA ARCHIVES

Torsten Rathmann[1*], Frank Toussaint[1], Michael Lautenschlager[1]
[1]Deutsches Klimarechenzentrum (DKRZ), Bundesstr. 45 a, 20146 Hamburg, Germany
[*]Email: rathmann@dkrz.de

ABSTRACT

Not every digital research data archive is fully funded. Charging fees for data services may help archives to survive. Archives which follow this funding stream have to decide who has to pay, data producers or users, and how much. For the calculation of data service prices a multi-linear price function is suggested. Variables are the number of datasets and the data volume. Different user requirements and data complexity are mirrored as service levels in the price function. Its utilization is shown in a case study at the World Data Centre for Climate (WDCC).

1 INTRODUCTION

Not every research data archive is fully publicly funded or may have such high internal budgets that charges can entirely be waived. Funding in the context of research infrastructure projects is usually limited and intended to cover the initial costs for setting up the infrastructure. Charges are a possibility to create the revenue which is necessary to cover the operational costs after the end of a project funding. Of course, prices are also the basis for the billing of data services to customers. But prices still serve as at least one other important function.

More and more, funding agencies are requiring researcher's information about which measures are taken to ensure that the research data is archived. Conversely, research funding sometimes includes money for long-term archiving of research data. An application is required for the approval of such funds. A realistic cost calculation is necessary for receiving approval for a research project. Thus, to enable applicants to apply for funding in the amount needed at a later time, prices must actually be known — not only when the data is created but even at the time of application for research funding. Laudable examples of open price information include ADS (2016), and Dryad (2016).

Who pays what now? Possibilities would be to charge data users a fee to access data or data producers to archive data. Some arguments for having the costs covered by the data producers:

- Data producers often have the means to get money for storing if this is applied, how it should be.
- Data producers are required by the funding agencies to store their data and are not in the position to omit archiving.
- Cost-oriented prices can be deduced from the costs of ingestion and storage. Therefore the prices are on a solid foundation.

On the other hand, an argument against having the costs covered by the data producers is:

- Producers already had and have a lot of work with their data and "pay" already in this way: the production of the data, a part of the quality control and the provision of meta-information about the data.

An argument for having users pay for access is that users have the benefit from the data. Arguments against this are:

- Publicly funded data should always be freely available.
- Free access stimulates further economic growth and, thus, leads, at least indirectly, to higher tax revenues, at least in some sectors such as public sector information (Houghton, 2011).
- Users (should) already appreciate the scientific, organizational and technical work of publicly providing research data in the form of a quote and already acknowledge the work in this nonmonetary way.
- To deduce a price based on access is difficult since the number of data users must be estimated. This number varies over time and can only be roughly estimated.

In any case, prices for data services are problematic, and adequate institutional funding of research data archives would be the better solution. However, if funding is insufficient, reasonable prices may help to ensure the existence of the archive.

2 BASIS FOR CALCULATING PRICES OF DATA SERVICES: THE PRICE FUNCTION

Prices for data archiving will mirror real costs of archiving services on total cost basis or on additional cost basis depending on the funding model of the responsible data archive. Archiving costs can be related to the data volume or the number of datasets (or more general: the number of individual data entities). Prices for extra services such as post-processing or the provision of persistent identifiers, e.g. Digital Object Identifiers (DOI), may be deducted from the consumption of computation time or the number of newly created identifiers.

Suppose a price is to be deduced from effective archiving costs, the price may be a function of the number n of datasets, the data volume V and the extra services (e.g. computation time, number of newly created persistent identifiers): $P(n, V, \text{Extra})$.

The number of datasets may be replaced by another measure that can quantitatively describe the logical structure and matches the granularity of the data. The price function should contain such a measure as the costs of the ingestion and other steps in the data life cycle depend considerably on the number of logical components (number of individual data entities). Each individual data entity usually has its own metadata and its own quality control steps.

An estimate of the storage costs at the German Climate Computing Centre (DKRZ) has shown that purely volume-dependent price calculation does not suite. This is discussed in more detail in the case study below. The observations at the DKRZ are consistent with those at the British Archaeology Data Service, where the twelve smallest data collections cost £ 88.06 / MB, but the largest twelve are only £ 1.54 / MB (each median) (Beagrie, Lavoie, & Woollard, 2010).

If the costs significantly depend on the complexity of data or on different user requirements, the price function should take these differences into account. Unfortunately, e.g. the type of data and Service Level Agreements (SLA) the customer can choose cannot easily be transferred into an analytical price function. Instead, service levels should be defined, which reflect different price levels.

If there are N service levels, data entities and volumes are distributed to these N levels. The number n of individual data entities and the data volume V are then vectors.

$$n = (n_1, \dots, n_N) \tag{1}$$
$$V = (V_1, \dots, V_N) \tag{2}$$

The price function P should be additively composed of component functions which depend only on one variable, to keep prices transparent for customers.

$$P(n_1, \dots, n_N, V_1, \dots, V_N, \text{Extra}) = B + \sum_{i=1}^{N} P_{\text{Entity},i}(n_i) + \sum_{i=1}^{N} P_{\text{Vol},i}(V_i) + P_{\text{Extra}}(\text{Extra}) \tag{3}$$

The base price B does not depend on any variable and is charged only once per order. It is advisable to have this constant in the price function because this definitively reduces the need to distribute variable-independent costs to variable-dependent price components. Via B all costs can be taken into account that are independent of the amount of data but often occur in connection with an order, e.g. creating a concept. Only if costs are taken into consideration, which are also independent of the number of orders, e.g. building occupancy costs, these costs cannot be assigned as a whole and must be distributed. Often such costs only occur in full-cost accounting.

The best choice for $P_{\text{Entity},i}$, which only depends on n_i, is a linear function unless the dependency is known in detail.

$$P_{\text{Entity},i}(n_i) = P_{\text{E},i} \cdot n_i \tag{4}$$

Such linear functions are the easiest for customers to understand. The coefficients $P_{\text{E},i}$ are constants, prices per individual data entity at service level i.

For the volume-dependent components $P_{\text{Vol},i}$ the best choice is again a linear function unless the dependency is known in detail.

$$P_{\text{Vol},i}(V_i) = P_{\text{V},i} \cdot V_i \tag{5}$$

The factors $P_{\text{V},i}$ are constants, the prices per unit volume at the service level i. The costs of bit-stream preservation should be included in the coefficients $P_{\text{V},i}$. Provided only small data volumes are delivered, for example, in the megabyte range, the volume-dependent component functions may be omitted, i.e. $P_{\text{V},i} = 0$. In this case, possible media costs may be included in the base price B.

In summary, the following multi-linear price function is recommended unless variable dependencies are known in more detail.

$$P(n_1, \dots, n_N, V_1, \dots, V_N, \text{Extra}) = B + \sum_{i=1}^{N} P_{\text{E},i} \cdot n_i + \sum_{i=1}^{N} P_{\text{V},i} \cdot V_i + P_{\text{Extra}}(\text{Extra}) \quad (6)$$

The price function P_{Extra} for extra services is not specified in any detail here but can be adapted for many types of extra services in a similar manner.

3 HOW TO FIND THE COEFFICIENTS OF THE PRICE FUNCTION

As research data archives are mostly non-profit organizations, prices are usually expected to be cost-covering, but not higher. Therefore, the most important cost components should be known, but there is no need to consider all costs. Most research data archives are not obliged to set their prices based on full cost accounting. Cost components covered from elsewhere, e.g. by the general budget of the institution, can simply be leaved out. Finally, the cost accounting also costs. A summary of all costs on a full cost accounting is more expensive. Of course, it is beneficial to know all costs, but for the purpose of pricing only, full cost accounting is not required.

Cost accounting can be top-down or bottom-up. The top-down approach determines the cost of a particular process on the basis of total expenditures, which are split up. The bottom-up approach starts with the individual steps. The individual steps are condensed into sub-processes and further into primary processes. The costs of the individual steps have to be known or estimated so that the costs of the sub- and primary processes may be calculated. Conversely, the percentages must be known in order to carry out the split for the top-down approach.

Overall, the price scheme should be simple and clear so that customers can understand it. The risks which are associated with the pricing of archive services should also be considered. Apart from the risk of prices being too high or too low in association with each price calculation, the main risk is, however, the risk of rapidly changing costs. Sudden price increases are harmful to research since researchers must apply for a sufficient amount for archiving, long before the data become available. The two largest cost blocks, the personnel and the hardware costs, are usually not associated with special risks that are higher than in other parts of the research infrastructure. Hardware costs usually decline if costs per MB are considered. Personnel costs rise slowly. However, risks lurk where requirements increase and affect prices. It is therefore important to identify costly work steps early, which will be of increasing importance in the following years — especially quality assurance and curation may be such steps.

4 CASE STUDY: WORLD DATA CENTER FOR CLIMATE

DKRZ runs the ICSU World Data Center for Climate (WDCC)[1], which is specialized in climate model data. Although most of the data is the result of numerical models, WDCC also stores processed observations to be used for validation or operating of climate models, e.g. measured values for precipitation or aerosol concentration.

The archiving costs must usually be born by those who have ordered the archiving. However, the download of data is free if the data is only used for non-commercial purposes and if no additional data processing is requested. So far, mainly climate data of institutions, that are also users of the DKRZ mainframe, has been archived. In this case, the costs are set off against the authorized quotas, and, thus, archiving is de facto free of costs. In the meantime, the WDCC archive services have become interesting for external customers, e.g. research institutions which want to avoid building their own long-term archive. Therefore, a pricing model was set up at the WDCC.

The basis for WDCC cost estimates have been tables using Eq. 6 (Luthardt, 2010). The working hours of the employees for the individual steps that are required for execution of an order, as well as for media and continuous operation costs are taken into account. All other costs, i.e. all fixed expenses as training and building occupancy costs, are so far not considered. Any distribution of fixed costs is therefore not necessary and cost accounting is comparatively easy. The bottom-up method was chosen for the working hours, estimated by the

[1] https://www.dkrz.de/daten-en/wdcc

employees who are involved in each step[2]. The individual steps have been condensed into the sub-processes of Table 1 and Table 2.

Some of the steps are taken only once per order, e.g. the creation of a concept. Other steps, such as quality control of the data and metadata are taken once for each experiment. In climate science an *experiment* is a simulation or an ensemble of related simulations with a well-defined set of physical parameters. Each experiment consists of datasets. At WDCC a dataset is typically a time series of one variable at one altitude level, e.g. the temperature at the 500-mbar pressure level. At each time step of the simulation, the time series is usually comprised of a two-dimensional array of values of the variable referring to a network of grid points. The grid is defined by the location coordinates, usually latitude and longitude.

Table 1. Ingestion, WDCC (working hours)

Sub-process	Per order	Per experiment	Per experiment with the same data structure
Information and consultation	4		
Project specification (data volume, formats, data organization, storage strategy, data path to WDCC, data policy, access limitations)	2		
Creating a concept (metadata, pre-processing, schedule) and cost evaluation	4		
Gathering, loading and quality control of metadata	10	5	3
Data transfer and insertion	7	1	1
Quality control of data including checking the consistency of metadata and data		10	4
Activation and final report	6		
Total	33	16	8

Table 2. 10 years of storage including curation, WDCC (working hours)

Sub-process	Per order	Per experiment	Per experiment with the same data structure
Metadata updates	10	10	8
Curation of datasets within the database		10	5
Adapting the access permissions	8	2	2
Continuous adaptation to the DKRZ infrastructure	10	5	3
Total	28	27	18

In a second step, the sub-processes have further been consolidated into the primary processes “Ingestion” (line “Total” in Table 1) and “10 years of storage including curation” (“Total” in Table 2), which have then been priced.

WDCC cost estimates refer to experiments having 500 datasets, which is a typical order of magnitude for the WDCC. The number of datasets does not currently influence the cost estimate. Instead, the decisive factor is the number of experiments. Two service levels are separately shown in Table 1 and Table 2. The column “Per experiment with the same data structures” (service level 2) is used if an order is given to archive multiple experiments, which were calculated with the same climate model, the same grid and the same climate variables (temperature, relative humidity, ...). Then there are synergies, especially concerning the metadata. For example,

[2] Pre-ingestion work, as the generation of metadata and the preparation of data for the WDCC ingestion interface, is not included in this calculation. Storage of metadata is needed for search and browse since the data itself are voluminous and mainly stored on tape.

experiments with different scenarios of anthropogenic CO_2 emissions, but otherwise the same specifications, have the same data structure.

The cost estimate for archiving of e.g. five experiments with the same data structure includes

1× the total no. of working hours by column "Per order",

1× the total no. of working hours by column "Per experiment" and

4× the total no. of working hours by column "Per experiment with the same data structures".

For the conversion of working hours into Euros a factor of € 31.25 / h has been used since 2010. Finally, the current costs for media and continuous operation are added, e.g. in the year 2015 at a rate of € 400 / TB for 10 years. Two media changes during the ten-year period are assumed.

Using Eq. 6, all these quantities result in the following approximate pricing formulas for the offered primary processes for ingestion

$$P_{\text{Ingest}} = € \, 1031 \; + € \, 500 \cdot n_1 + € \, 250 \cdot n_2 \tag{7}$$

and for storage for 10 years, including curation

$$P_{\text{Cur}} = € \, 875 + € \, 844 \cdot n_1 + € \, 563 \cdot n_2 + € \, 400/\text{TB} \cdot V \tag{8}$$

Here, P_{Ingest} and P_{Cur} are the prices, n_1 is the number of experiments at service level 1 (various data structures), n_2 the number of experiments at service level 2 (same data structures), and V is the volume of data. Eq. (7) and (8) are approximations; WDCC cost estimates include additional rounding steps.

All prices do not include VAT.

The coefficients $P_{\text{E},1}$ and $P_{\text{E},2}$ in front of n_1 and n_2 in Eq. (7) and (8) are so large that the price component dependent on the number of datasets is usually much greater than that dependent on volume. A pricing structure based only on data volume is, therefore, not an option for WDCC.

Pure single copy tape storage without ingestion into WDCC long-term archive and without any data curation has been offered for a fee at DKRZ since 2013. On this occasion, the media and continuous operation costs were recalculated. The recalculation resulted in € 165 / TB and copy for the ten-year storage. Since two copies are always available in the WDCC, the cost is € 330 / TB for the pure bit-stream preservation.

After years in use, the working hours listed in Table 1 and Table 2 have still to be verified and updated. Further software developments have led to savings through more extensive automation. On the other hand, the cost of quality assurance has increased due to additional requirements.

5 CONCLUSIONS

Well justified objections can be raised against making data users or data producers pay for archiving services. Comprehensive funding of digital research data archives would be the best for archives, users and data producers, but if funding is insufficient or ceases with projects' ending, fees may help to support archives to survive. Should an institution have decided to charge a fee for use of its archive, such information should be made known early on as researchers have to estimate their costs for data usage and archiving when they apply for funding. Further, in the case of publicly funded research, all fees should be based on costs. A simple cost calculation including the costs that are not covered from elsewhere is usually sufficient for deducing cost prices.

A multi-linear price function is recommended unless a precise dependence of the costs on the relevant parameters, e.g. number of individual data entities or data volume, is known. In a case study at DKRZ, the multi-linear price function has been used for the pricing of a) ingestion and b) 10 years of storage and curation of climate data. Two service levels are currently considered in long-term archiving. Customers have to pay for working hours, media and continuous operation only. Other expenses, such as building occupation costs, are not considered. With this choice, it has been possible to assign costs one-to-one to the parameters in the cost function.

6 REFERENCES

Archaeology Data Service: Charging Policy. (2016, November). Retrieved January 20, 2017, from Archaeology Data Service: http://archaeologydataservice.ac.uk/advice/chargingPolicy

Beagrie, N., Lavoie, B., & Woollard, M. (April 2010). *Keeping Research Data Safe 2.* Retrieved January 20, 2017 from JISC: http://www.webarchive.org.uk/wayback/archive/20101225023826/http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf

Houghton, J. (2011, September). *Costs and Benefits of Data Provision.* Retrieved January 20, 2017, from Australian National Data Service: http://www.ands.org.au/__data/assets/pdf-file/0004/394285/houghton-cost-benefit-study.pdf

Luthardt, H. (22. November 2010). *Datenspeicherung und Verteilung von Projektdaten am DKRZ (≥ 10 Jahre).* In German only. Retrieved Januar 23, 2017 from DKRZ: https://www.dkrz.de/daten/data-services/langzeitarchivierung/LZA_Kostenabschaetzung_generell_v05b.pdf/at_download/file

Pricing plans - Dryad. (2016, May). Retrieved January 20, 2017, from Dryad: http://www.datadryad.org/pages/pricing