



Final LEARN Conference | London | 5th May 2017
How to Cost Data Curation: report from the Breakout groups

Breakout Tutorial Group 2 on Costing Data Curation

Lead: Paul Stokes (Jisc)

Rapporteur: Ignasi Labastida (University of Barcelona)

Paul Stokes started the session by defining data curation versus research data management remarking that the earlier focus in the long term archive and the later on the entire lifecycle of the data. With all his expertise on data curation, he has learnt that every higher education institution has different needs on preservation and also that it is a global community effort. It is not enough to say that we must preserve all the data until ten years after its last access without thinking of the long term cost implications of such diktats.

To start costing we need to know everything has a price, therefore we need to know that price. Someone has to pay this price, but who? In addition, it is important to know the benefits that offset the costs. We should need to track all the costs and try to figure out if it is possible to establish cost recovery mechanisms.

The main problem in costing data curation is that it is a difficult and complicated to achieve. Paul asked to the participants why they thought it is not easy to calculate the costs and he received the following answers:

- *There are many types of data*
- *Storage and curation processes depend on discipline*
- *In many cases the effort is unknown upfront*
- *The value of data is not known*
- *The cost of the people involved is not easy to calculate*
- *It is not easy to include specific services in a general cost*

To help in this costing evaluation, he suggested some best practices starting with the definition of what are you costing. Here he pointed out two methodologies: activity based costing and traditional costing. The first one is based on the determination of all the activities involved in a process establishing all the costs (direct and indirect), and then divide the total amount per cost unit. The traditional costing method is based on the calculation of direct costs and the division per cost unit.

Both systems are fine but we need to follow one of them along all the process of costing. Sometimes it is not easy to get the information we need to make all the calculations.

Paul suggested that every institution should establish its own cost tool, based (or not) on an existing one, because a single model does not fit everyone

He also asked the participants why they were concerned with costing data curation, He received the following answers:

- *There is a need to find sustainability at long term*
- *Researchers need to know it*
- *There is a need to defend costs*
- *To find business cases*
- *To compare the costs of doing it versus the costs of not doing it*

The session ended with some questions and answers. Among them we can remark some issues that arose:

- The need of define in the costing model for how long datasets must be curated.
- The importance of defining the current activities as the current costs and to foresee future costs also appeared.
- Sometimes in the costing evaluation, it has to be taken into account the payment of using certain data or the cost of reproducing data repeatedly.
- The institution can risk its own reputation by not doing a proper data curation. As an example, an institution was brought to court by an investigator caused by the loss of some data.
- Institutional “short termism” can be seen as a problem because curating data is a matter of long term.

As a conclusion, Paul encouraged all the participants to follow some of the models proposed and start to include some numbers in order to see the current spending in all those activities related with research data