JESLIB 2012; 1(1): 57-59
doi:10.7191/jeslib.2012.1007

Journal of eScience Librarianship
putting the pieces together: theory and practice

A Review of ‘DataSpace: A Funding and Operational Model for Long-Term Preservation and Sharing of Research Data’

Raquel Abad

University of Massachusetts Medical School, Worcester, MA, USA

“DataSpace: A Funding and Operational Model for Long-Term Preservation and Sharing of Research Data,” is a 2010 paper authored by Serge Goldstein, Associate Chief Information Officer and Director of Princeton University’s Academic Services, Office of Information Technology, and Mark Ratliff, Digital Repository Architect of Princeton’s DataSpace repository (Princeton University 2011). Goldstein and Ratliff’s paper provides an overview and explanation of the funding and operational models applied to the DataSpace repository at Princeton University.

With recent data management requirements and policies enacted by federal funding agencies such as the National Science Foundation (NSF) and National Institutes of Health (NIH), higher education has quickly responded to these changes. Many academic departments and libraries have worked diligently to provide a wide variety of support services— from smaller scale, often individualized services such as assistance with creating a data management plan, to the construction of large scale, multi-departmental repositories designed specifically for the needs of data storage and sharing. The DataSpace repository, created through a partnership between the Office of Information Technology (OIT) and the Library, is Princeton University’s attempt to address its data storage needs.

Clearly, with the construction of a repository comes financial demands, e.g. the personnel necessary to build and maintain the repository and its services, and the cost of the storage for the data. There are currently many data repositories in operation today, for instance: Dryad, D2C2, arXiv.org, and Merritt (National Evolutionary Synthesis Center and University of North Carolina Metadata Research Center 2011, Purdue University Libraries 2011, Cornell University Library 2011, University of California Curation Center 2011), each reflecting different operational and funding models. The construction and maintenance of many repositories have been made possible by grant money; another common funding model is based on annual payments by the researcher depositing their data. However, as each of these models are reliant on grant-based funding, the primary challenge for these regards sustainability: when the grant funding ends, how will the repositories and the data be funded?

The DataSpace repository was constructed with this question of sustainability in the forefront of the project developers’ minds. Rather than enact an annual payment-based funding model, the DataSpace model operates by charging a one-time payment, based on the amount of data stored, which is due upon the time of initial data storage. This model, “Pay Once, Store Forever,” (POSF) (Goldstein and Ratliff 2010) fundamentally functions by way of the proposition that long-

Correspondence to Raquel Abad: raquel.abad@umassmed.edu
Keywords: data repositories, funding models, operational models

term data storage is “funded by one-time payments that cover the current costs of storage, and leave enough excess funds to cover on-going replacement and management of that storage.” In their paper, Goldstein and Ratliff outline the formula that is used as the foundation for POSF; this formula takes into account the rapidly depreciating value of data storage over time. For instance, the cost of 1 terabyte of storage ten years ago was close to $15,000; today, in 2011, an external hard drive with 1 terabyte of storage costs about $100.00.

DataSpace’s operational model was created with the main principles of POSF in mind: that it “makes sense only if storage costs decline steadily over time,” and “if management costs are kept to a minimum”. The operational model, “Write Once, Read Forever” (WORF) is intended to minimize ancillary costs that are associated with storing and disseminating data, while also ensuring that the data in DataSpace is publicly accessible. The principles of WORF include the following: the storage may not be re-used; the original data may not be changed; all the data is publicly accessible; the repository only provides storage for “the bits associated with the data, and a variable set of meta-data;” “once paid for, the repository assumes all responsibility for the storage and management of the data;” the storage operates on a pay-once basis; and the repository may offer, and charge for, ancillary services such as data conversion or specialized data delivery.

The funding and operational models of DataSpace have no doubt been met with many questions, many of which have been offered and explored in the DataSpace paper. Some of these questions include skepticism that a researcher only does need to pay the cost of storage once, skepticism that the payment will cover all costs associated with the data’s storage, management, and sharing, criticism that the cost is too steep ($6,000/TB), and perhaps the most important consideration— that in order for this funding model to work, the assumption that the cost of data storage will continue to depreciate is of critical importance.

Regardless of these considerations— all of which are entirely valid— both the operational and funding models of this repository carry significant promise. At the very least, as they are alternatives to currently existing funding/operational models, this will provide another option for researchers to explore in storing their data. Additionally, DataSpace was a joint partnership between the OIT and Library, and while the extent of the library’s participation is unknown, that the library does have a collaborative presence in an institutional project as large as this is important. By participating in this project, the library is exposed to a new source of publicity, perhaps extending its presence to a new group of researchers that they may not have otherwise been reaching. Moreover, while the OIT manages the storage and maintenance of the repository’s content, the library may offer services such as assigning metadata standards to repository records, thus ensuring the additional advantage of harvesting these records into the library’s catalog for searching and discoverability (Giesecke 2011, 541).

Regarding the view that the cost of storage is steep, as the payment is on a one-time basis, it may not be unreasonable for researchers to write a storage cost item into their grant proposals. Finally, although the funding model is essentially dependent on the depreciation of data, given the trend of data depreciation over the past few decades, it does seem safe to assume that this is a stable assumption. There has been a substantial amount of research on this topic: an initial search of the literature yielded numerous relevant papers, most specifically a white paper stating that the cost of magnetic disk storage has decreased annually by about 45% since 1989 (Gilheany 2004, 1). Princeton’s DataSpace repository, while it may not certainly be suitable for every researcher, does have its merit and seems promising for the management, storing, and sharing of data.

References

Cornell University Library. "Arxiv." Accessed December 12, 2011. http://arxiv.org/.

Giesecke, Joan. "Institutional Repositories: Keys to Success." *Journal of Library Administration* 51, no. 5/6 (2011): 529-42.

Gilheany, Steve. “The Decline of Magnetic Disk Storage Cost Over the Next 25 Years.” edited by Berghell Associates, 2004; http://www.archivebuilders.com/whitepapers/22004p.pdf.

Goldstein, Serge, and Mark Ratliff. “Dataspace: A Funding and Operational Model for Long-Term Preservation and Sharing of Research Data.” Published electronically August 27, 2010; http://arks.princeton.edu/ark:/88435/dsp01w6634361k.

National Evolutionary Synthesis Center and University of North Carolina Metadata Research Center. “Dryad.” Last Modified October 28, 2011. http://datadryad.org/.

Princeton University. “Dataspace.” Accessed December 12, 2011. http://dataspace.princeton.edu/jspui/.

Purdue University Libraries. “Distributed Data Curation Center.” Accessed December 12, 2011. http://d2c2.purdue.edu/index.php.

University of California Curation Center. “Merritt.” Accessed December 12, 2011. http://merritt.cdlib.org/.

Disclosure: The author reports no conflicts of interest.

ISSN 2161-3974