Comparing Approaches for the Sustainability of Scientific Data Repositories

Robert R. Downs[1] and Robert S. Chen[2]

NASA Socioeconomic Data and Applications Center (SEDAC),
Center for International Earth Science Information Network (CIESIN),
Columbia University

[1] rdowns@ciesin.columbia.edu

[2] bchen@ciesin.columbia.edu

2013 Columbia University Research Data Symposium
February 27, 2013, New York, NY

Sustainable data systems are critical components of the cyberinfrastructure needed to provide long-term stewardship of scientific data, including Earth science data, throughout their entire life cycle. A variety of approaches may help ensure the sustainability of such systems, but these approaches must be able to survive the demands of competing priorities and decreasing budgets over long time periods. Analyzing and comparing various approaches can identify viable aspects of each approach and inform decisions for developing, managing, and supporting the cyberinfrastructure needed to facilitate discovery, access, and analysis of data by future communities of users. A typology of sustainability approaches is proposed, and example use cases are offered for assessing the approaches over time. These examples help illustrate the potential strengths and weaknesses of each approach under various conditions and with regard to different objectives, e.g., open vs. limited access. By applying the results of these analyses to their particular circumstances, systems stakeholders can assess their options for a sustainable systems approach, which may incorporate multiple sustainability options, along with other metrics to ensure the sustainability of the scientific data and information for which they are responsible. In addition, clarifying and comparing sustainability approaches should inform the design of new systems and the improvement of existing systems to meet the needs for long-term stewardship of scientific data, and support education and workforce development efforts needed to ensure that the appropriate scientific and technical skills are available to operate and further develop sustainable cyberinfrastructure.

Typology of Sustainability Approaches for Scientific Data Stewardship

Discrete Revenue Stream Models

- Establishing and maintaining individual sources of revenue
- Multiple revenue streams may be needed to ensure continual funding

Fees
- Usage fees (commercial use fees vs. non-commercial use fees)
- Depositor fees

Subscriptions
- Annual or multi-year institutional subscribers (members)

Grants
- To acquire a specific collection
- To maintain a collection for a specified time period

Advertising or sponsorship
- Revenue dependent on site traffic by target users

Donations/Endowments
- Cultivating benefactors for collections or services

Subsidies
- Direct and in-kind support from activities that benefit from data, e.g., undergraduate and graduate education

See Baranski et al. (2010), Bastow and Leonelli (2010), Donker (2009), and Kintigh and Altschul (2010).

Cooperative Models

- Establishing mutually-beneficial relationships with partners
- Harnessing distributed stakeholder resources and in-kind support

Institutional commitments
- Cost sharing or resource sharing

Network development
- Development of bilateral and multilateral sharing, backup, and mutual assistance arrangements

Commitments from stakeholder communities
- Multiple stakeholders or stakeholder categories
- Funding or in-kind contributions (e.g., open source software development, crowd sourcing)

Incentives from funders
- Short-term funding or other resources in recognition of long-term commitments

See Beagrie et al., 2010; Finney, 2007; Halbert, 2009; Kwon et al., 2006; Lee, 2009; Uhlir et al., 2009; Walters et al., 2010.

Use Case: Subscription Services

A data center offers annual paid subscriptions to academic institutions and libraries for unlimited access to its data holdings. Alternatively, individuals or organizations can purchase a license to access a particular data product or data service.

Use Case: Institutional Commitments

A grant-funded data center based at a large research university establishes a long-term archive cooperatively with the University Library to manage, preserve, and disseminate selected scientific data sets housed in the data center and establishes an inter-departmental board to govern the archive.

Use Case: Deposit Fees

A data center requires a fee from individuals and organizations when they deposit data. The fee aims to cover both the short- and long-term costs of managing the data, preserving the data in the archive of the data center, and curating the data for dissemination to the targeted community of users.

Use Case: Cooperative Archives/Network

A group of universities establish a distributed network of archives to host the scientific data produced by each of the cooperating universities. The faculty, staff, and students of the cooperating universities have access to the data and services.

Conclusion

Given unavoidable uncertainties in future revenue streams, user interest, community capacity, institutional longevity, preservation costs, and other components of sustainability on decadal to century time scales, it seems clear that multiple and flexible strategies may be essential to long-term sustainability. Combinations of different types of institutions with long histories of persistence (e.g., universities, museums, libraries, and some government agencies) may also be needed as the foundation for long-term scientific data stewardship.

This work is based on Downs and Chen (2012). The authors presented an earlier version of this poster at the 2012 AGU Fall Meeting in San Francisco, CA, on 6 December 2012.

References

Baranski, B., Deelmann, T. and Schäffer, B. (2010). "Pay-per-Use Revenue Models for Geoprocessing Services in the Cloud", in Brovelli, M.A., Dragicevic, S., Li, S. and Veenendaal, B. (Eds). ISPRS Archives, XXXVIII-4/W13, WebMGS 2010: 1st International Workshop on Pervasive Web Mapping, Geoprocessing and Services, August 26-27, 2010, Como, Italy,at http://www.isprs.org/proceedings/XXXVIII/4-W13/ID_20.pdf

Bastow, R. and S. Leonelli (2010). Sustainable Digital Infrastructure. EMBO Reports, 11 (October) 730-734, at http://dx.doi.org/10.1038/embor.2010.145

Beagrie, N., Eakin-Richards, L. and T. Vision (2010). Business Models and Cost Estimation: DRYAD Repository Case Study, at http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/beagrie-37.pdf

Donker, F.W. (2009). "Public Sector Geo Web Services: Which Business Model Will Pay for a Free Lunch?", In van Loenen, B., Besemer,J.W.J. and Zevenbergen, J.A. (Eds). SDI Convergence: Research, Emerging Trends, and Critical Assessment. Delft, The Netherlands: NCG, Nederlandse Commissie voor Geodesie, Netherlands Geodetic Commission, pp. 35-51, at http://www.ncg.knaw.nl/eng/publications/Green/48VanLoenen.html

Downs, R.R. and Chen, R.S. (2012). Towards Sustainable Stewardship of Digital Collections of Scientific Data. GSDI World Conference (GSDI 13) Conference Proceedings, Québec City, Canada, May 14-17, 2012. http://www.gsdi.org/gsdiconf/gsdi13/papers/130.pdf, http://www.gsdi.org/gsdiconf/gsdi13/slides/5.4c.pdf

Finney, K.T. (2007). A "bottom up" governance framework for developing Australia's marine Spatial Data Infrastructure (SDI). Data Science Journal, 6: 64-90, at http://dx.doi.org/10.2481/dsj.6.64

Halbert, M. (2009). Comparison of Strategies and Policies for Building Distributed Digital Preservation Infrastructure: Initial Findings from the MetaArchive Cooperative. International Journal of Digital Curation, 4(2): 43-59, at http://www.ijdc.net/index.php/ijdc/article/view/117/110

Kintigh, K.W. and J.H. Altschul (2010). Sustaining the Digital Archaeological Record. Heritage Management, 3(2): 264-274.

Kwon, H., Pardo, T.A., and G.B. Burke (2006). Building a State Government Digital Preservation Community: Lessons on Interorganizational Collaboration. Proceedings of the 2006 International Conference on Digital Government Research, 277-284, at http://doi.acm.org/10.1145/1146598.1146673

Lee, J.W., Zhang, J., Zimmerman, A.S. and A. Lucia (2009). DataNet: An Emerging Cyberinfrastructure for Sharing, Reusing and Preserving Digital Data for Scientific Discovery and Learning. AIChE Journal, 55(11): 2757-2764, at http://wwwcs.ccny.cuny.edu/~jzhang/papers/AiChE09DataNet.pdf

Uhlir, P.F., Chen, R.S., Gabrynowicz, J.I. and K. Janssen (2009). Toward Implementation of the Global Earth Observation System of Systems Data Sharing Principles, Data Science Journal, 8: GEO1-GEO91, at http://dx.doi.org/10.2481/dsj.35JSL201

Walters, T.O. and K. Skinner (2010). Economics, Sustainability, and the Cooperative Model in Digital Preservation. Library Hi Tech, 28(2): 259-272, at http://dx.doi.org/10.1108/07378831011047668

Center for International Earth Science Information Network
Earth Institute | Columbia University

The authors appreciate the support received from the National Aeronautics and Space Administration (NASA) under Contract NNG08HZ11C for the Continued Operation of the Socioeconomic Data and Applications Center (SEDAC).