Towards Sustainable Stewardship of Digital Collections of Scientific Data

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

[1]Columbia University, Center for International Earth Science Information Network (CIESIN), rdowns@ciesin.columbia.edu
[2]Columbia University, Center for International Earth Science Information Network (CIESIN), bchen@ciesin.columbia.edu

Abstract

The digital revolution has vastly increased the ability of the scientific community to collect and store a tremendous variety and quantity of data in digital form, representing a potentially irreplaceable legacy that can support scientific discovery and scholarship in both the present and the future. However, it is not yet clear what organizations or institutions can and should maintain and store such data, ensuring their long-term integrity and usability, nor how such long-term stewardship should be funded and supported. Many traditional information preservation and access institutions such as libraries and museums are struggling to develop the skills, resources, and infrastructure needed for large-scale, long-term digital data stewardship. Government agencies often have strong technical capabilities, but are subject to political and budgetary pressures and competing priorities. Private organizations and companies can bring to bear innovations not only in technology but also in economic approaches that could provide financial sustainability. Developing long-term collaborative partnerships between different types of organizations may be one approach to developing sustainable models for long-term data stewardship. The development of objective criteria and open standards for trusted digital data repositories is another important step towards sustainable data stewardship. A critical challenge is the development of viable economic models for ensuring that the resources needed for long-term stewardship are put in place, while at the same time addressing the needs of the scientific community and society more generally for open access to scientific data and information resources. The development of a robust spatial data infrastructure can not only help reduce both the short- and long-term costs of data stewardship, but also provide a framework for the establishment and evolution of trustworthy data repositories that will be available for future generations of users to discover, access, and use the scientific heritage that is being created today.

Keywords: scientific data archives, data management, digital preservation, data stewardship, organizational strategy

1 INTRODUCTION

Scientific data stewardship has enabled scientists and other users of scientific data to access and analyze data that were collected in the past. Considering the value of these capabilities for continuing scientific progress, scientific data stewardship services, such as those enabled by spatial data infrastructure (SDI), are critical to the future of science. Scientific data stewardship encompasses the preservation and curation services to save the records of non-replicable observations, which otherwise could be lost. Establishing scientific data stewardship programs that preserve and curate scientific data and research-related information for future use can contribute to the scientific legacy that fosters the continuation of the scientific process. Scientific data stewardship programs can expand opportunities for data-driven science and foster longitudinal studies of observations captured in digital and digitized data sets.

By continuing scientific data stewardship programs, such activities can enable future communities of users to engage in data discovery and use of previously collected data and of data that will be collected in the future. Sustainable scientific data stewardship programs are needed to ensure that the legacy of scientific data can be accessed and used in the future (National Research Council, 2009). Preserving the scientific data that is created today contributes to the scientific legacy by enabling continuation of their analysis in the future (National Science and Technology Council, 2009). Without sustainable programs that can continue providing scientific data stewardship services, the scientific digital heritage will be at risk. The legacy data that underpin the cumulative scientific knowledge base could become undiscoverable and unusable. In addition, without access to these data, future scientists could lose the ability to replicate scientific analyses that have been conducted recently or in the past.

Research projects, which are temporary by definition, often provide the necessary support for studies that develop collections of scientific data and scholarly works in the form of digital objects. Similarly, projects can provide support to develop infrastructure for accessing and preserving digital objects. However, when funding for project activities ends, the resulting digital resources need to be supported by sustainable organizational, technological, and financial structures that can continue to provide the stewardship that is necessary to support their long-term preservation and dissemination. Challenges for providing sustainable stewardship for scientific data apply across disciplines, types of digital data, and infrastructure deployed, including SDI. The results of a recent analysis of physical science research practices, including Earth science disciplines, recommended removing barriers to research by "identifying sources of sustainable funding for information resources developed as part of funded projects." (Meyer et al., 2011, p. 92).

While cost savings and extensible architecture contribute to sustainable SDI (Czerwinski et al., 2007), institutionalization also is required (Yawson et al., 2010). Without continuing leadership and capabilities for data stewardship, there will be limited resources and incentives to support such activities. In addition to recognizing the "cultural divide" between rewards for short-term research projects and rewards for establishing infrastructure to support long-term research efforts (Appelbe and Bannon, 2007), recognition and rewards also need to support efforts to preserve the data that are collected (Lesk, 2008).

However, institutional support for research could include support for preserving today's scientific data for use by tomorrow's researchers. Institutional responsibilities for long-term stewardship would involve the establishment and the continuing management of the infrastructure, including human resources, to preserve the scientific heritage of digital scientific data for future use. Enabling the scientific legacy to continue includes establishing sustainable programs, infrastructure, and incentives for preserving the scientific heritage that is being created today. If institutions that are participating in the creation of scientific data accept the responsibility for long-term stewardship of the digital resources and infrastructure that represent the legacy of their scientific research, these resources might be available for the future generations of researchers, educators, learners, and decision-makers to build on the scientific research that is being conducted today.

Sustainable stewardship, which is needed to ensure that scientific data collections do not languish when budgets and personnel change in the organizations that have accepted the responsibility for managing our scientific heritage, is reviewed in section 2. An experiment in collaborative sustainable stewardship for managing, preserving, and disseminating scientific data is described in section 3 and a model for collaborative sustainable stewardship of digital scientific data is presented in section 4.

2 SUSTAINABLE STEWARDSHIP OF SCIENTIFIC DATA

2.1 Need for Sustainable Stewardship of Scientific Data

The risks facing scientific and scholarly digital resources in terms of long-term preservation offer compelling reasons for institutions to explore alternative models for providing sustainable stewardship to manage digital collections of scientific data and scholarly works. Although developing infrastructure and services to preserve and provide access to digital collections is an important first step, institutions also need to ensure that long-term stewardship has been established for managing and operating such services in the future. Otherwise, the scientific data and scholarly works that are being ingested in digital archives

and repositories today could be at risk of abandonment in the future. Sustainable organizations are needed to provide long-term stewardship of the digital resources that are in their care (Ayris, 2009; Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010).

Sustainable stewardship is a necessary part of the SDI needed to ensure stakeholders that resources will be committed in the future so that digital information will continue to be preserved for the long-term. Providing sustainable stewardship is critical for an archive to be considered a viable entity for the long-term and trustworthy as a digital repository (OCLC and CRL, 2007). Emerging standards for trustworthy digital repositories recognize that conditions for sustainable stewardship of an archive include management responsibility, financial viability, and an organizational mission consistent with the preservation of the objects in its collections (Consultative Committee for Space Data Systems, 2011; Digital Curation Centre and DigitalPreservationEurope, 2007; Nestor Working Group – Trusted Repositories – Certification, 2006).

When weighing the costs and benefits of preserving research data, institutions also must recognize that such costs must be considered in terms of a long-term investment that is necessary (Ball, 2010). These, and other examples (Carlson et al., 2010), demonstrate the need for sustainable business models to ensure that data stewardship organizations are able to continue providing their services in the future (Lewis, 2010). Organizational partnerships also should be considered to support the infrastructure and services to provide access to digital content over time (Guthrie et al., 2008).

Government agencies also have recognized the need for sustainable organizations that can provide long-term stewardship of scientific data. The National Science Foundation has funded projects that have proposed the creation of sustainable organizations to preserve scientific data (Choudhury and Hanisch, 2009). The National Digital Information Infrastructure and Preservation Program (NDIIPP) of the Library of Congress has established the National Digital Stewardship Alliance (NDSA), an alliance of partners to form a national network to meet the need for a sustainable entity committed to the long-term stewardship of "historically, culturally or scientifically significant digital content" (Anderson et al., 2009). The final report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (2010) also recommended the establishment of public-private partnerships to attain financial sustainability for the preservation of digital resources, including research data.

2.2 Models for Sustainable Stewardship of Scientific Data

A critical challenge is the development of viable economic models for ensuring that the resources needed for long-term stewardship are put in place, while at the

same time addressing the needs of the scientific community and society more generally for open access to scientific data and information resources. Reviewing alternative models to public funding for supporting infrastructure for sustainable access to biological data, including commercial use fees, combined government and industry funding, stratified services fees, cost recovery fees, subscriptions, advertising, voluntary service contributions, and publishing industry partnerships, Bastow and Leonelli (2010) found that none of these alternatives would meet user expectations. Considering the rate of changes in technologies used to manage and enable the use of data, the diversity of data and the diversity of their potential uses, meeting user expectations may be quite challenging, especially when considering data that are used across disciplines.

After assessing alternative models for financing the stewardship of digital archaeological data, such as access fees, grants, advertising, donations, endowment, and in-kind contributions, Kintigh and Altschul (2010) suggested that legal and ethical requirements necessitate the preservation of these digital resources and proposed that the contributor pays model, as a primary approach, offers the most potential as a sustainable model. In this model, the sponsors of the research would cover the deposit fees, which would be included in the awards to grantees performing the research. This contributor pays model also would rely on high-volume to minimize costs, as well as data preparation and description by data providers. Baranski et al. (2010) offered an architecture to support alternative business models for sustainable SDI, proposing revenue models that would utilize pay-per-use geoprocessing services made available to customers as cloud services. Such models would combine data products, infrastructure, and services from various providers, which would each be paid from the revenue obtained from customer fees.

Analyzing various business and financial models for their suitability to support web services for a sustainable SDI, Donker (2009) identified alternative revenue models, including a subscription model, a usage model, a royalty model, a free model, and hybrid models, which were categorized as a community model, an enticement model, and a street performer model. When evaluating these models for services that disseminate data for viewing by society, Donker (2009) recommended the Free Model, but also included the possibility of the royalty model, recognizing that, in effect, it also would be free and that government funding would be necessary to support such services. Adopting models for sustainable data stewardship that offer open access without requirements for subscriptions or fees would enable use by more stakeholders, such as those representing small science or those in developing countries, who do not have the ability to pay. However, providing free and open access to data collections may require a combination of cost reductions, such as those attained through economies of scale, cooperative resource sharing, and interoperable sharing of data and services, as well as support from committed stakeholders. Community

contributions to the development of open standards and open source software can reduce costs and improve the efficiency of SDIs (Finney, 2007).

Commitments from stakeholders are critical for the sustainability of geospatial data collections that are available for free or for modest fees that enable cost recovery (Uhlir et al., 2009). Obtaining sustainable funding may require the identification of multiple categories of stakeholders, such as scientists, publishers, and funding agencies, which receive value and are able to support the data services offered (Beagrie et al., 2010). Cooperatives that are supported by multiple institutional stakeholders offer another approach to support the sustainable preservation of digital research resources (Halbert, 2009; Walters and Skinner, 2010). The DataNet program of the National Science Foundation has encouraged institutional commitments for infrastructure sustainability from partnering organizations that are award recipients (Lee et al., 2009). For example, the University of Minnesota has committed organizational resources to sustain the infrastructure for the Terra Populus project to establish and maintain an SDI of environmental and population data as a partnership between the Minnesota Population Center, the Center for International Earth Science Information Network (CIESIN) of Columbia University, and the Inter-university Consortium for Political and Social Research (ICPSR), of the University of Michigan, which is described by Sobek et al. (2011).

2.3 Stewards of Scientific Data

While society has much to gain from the preservation of scientific data, the traditional stewards of scientific heritage, including universities, museums, libraries, and archives, are primary stakeholders that have an interest in ensuring the availability of legacy scientific data for future research. As long-term stakeholders in the creation and preservation of knowledge, these traditional stewards also should be considered when assessing alternatives for supporting scientific data stewardship.

Universities have an opportunity to experiment with new organizational forms and collaborations to improve capabilities for facilitating long-term stewardship of research resources (National Research Council, 2002). Similarly, recognizing the importance of sustainability for SDI, Olsson (2009) asked about the relationship between sustainability of an SDI-based network and cooperation. By collaborating within their own institution and cooperating with other institutions to preserve collections in digital repositories, universities can serve as laboratories where possible models for sustainable stewardship of such collections can be identified, explored, and tested. The benefits of such experiences can contribute to the trustworthiness of archives and help them to overcome the challenges of digital preservation that will undoubtedly be encountered over future years.

Universities have a longstanding history of support for the preservation of knowledge and offer an ideal environment to explore collaborative capabilities for providing scientific data stewardship (Association of Research Libraries, 2006). The faculty, researchers, and students who produce scientific data have an interest in ensuring the future use of their data. University libraries, laboratories, and scientific data centers share mutual interests in preserving scientific data and research-related information. Libraries also recognize that the increasing amount of digital research resources that are produced by faculty are at risk. Laboratories see the need for reusing and integrating data to conduct longitudinal investigations and to answer new research questions. Scientific data centers need to provide sustainable approaches for supporting the use of the research data that they offer.

Libraries and museums also have a long history of enabling access and use of scientific artifacts. These organizations also have been adopting the use of infrastructure to manage intellectual assets in digital form. In light of recent changes in the patterns of library use and developments in information and communications technologies (ICT) for the stewardship of digital resources, libraries have recognized opportunities for managing scientific data in addition to other scholarly resources in digital form.

Data archives, including scientific data centers and repositories, have emerged during the last half of the twentieth century to address the data stewardship needs of scientific communities. Often managing and disseminating data for a particular scientific discipline or domain of inquiry, many scientific archives and data centers rely on government contracts and grants to support their operations (Ruusalepp, 2008).

As strong advocates of scientific data that recognize the future value of these data for scientific communities and society, universities, libraries, museums, and archives should collaborate with government agencies to ensure that the scientific heritage endures for use by future generations. Universities, libraries, museums, and government agencies have an opportunity to realize long-term benefits from their short-term projects by fulfilling the legacy entrusted to them by the great leaders who initially established their institutions. By organizing collaboratively, these institutions can leverage their capabilities to establish the organizational structures, technological infrastructure, and financial resources needed to provide sustainable stewardship of scientific data for current and future generations of scientific data users.

3 EXPERIMENT IN SUSTAINABLE SCIENTIFIC DATA STEWARDSHIP

An experiment in the provision of sustainable stewardship for the preservation of a digital collection of scientific data is being conducted by Columbia University

(Downs et al., 2007). The plan for the Long-Term Archive of the NASA Socioeconomic Data and Application Center (SEDAC) serves as a model for ensuring the long-term sustainability of digital repository collections and their objects. Columbia University has established an intra-institutional Board for continuing management of the SEDAC Long-Term Archive, with representation from the Columbia University Libraries, the Earth Institute of Columbia University, and the SEDAC, which is operated by the CIESIN, of Columbia University.

The criteria for the composition of the Long-Term Archive Board was intended to provide a diversity of organizational and disciplinary perspectives for managing the Long-Term Archive. Representatives from diverse units within an institution can garner support from multiple resources and potentially reduce the demand placed upon any one resource. Obtaining representation from various disciplines also can help to ensure that disciplinary bias does not determine the fate of particular resources or the general approach to be adopted for scientific data stewardship. Assembling representation from the physical and social sciences, as well as from the computer and information sciences, offers various disciplinary traditions on which to draw for expertise on preserving scientific data and research-related information. In practice, establishing an interdisciplinary Board enables a variety of opportunities and approaches to be considered from among the separate disciplines represented, increasing the possibilities for managing digital resources and contributing to the organizational knowledge on digital preservation and stewardship of scientific data.

Each of the organizational units represented by the members of the Board brings its own strengths and perspectives that contribute to the decisions and processes that facilitate the stewardship of the digital scientific data of the Long-Term Archive. The Columbia University Libraries have considerable experience in acquiring, managing, preserving, and sharing knowledge with the university community. SEDAC has experience and skills in managing specific types of scientific data using state-of-the-art tools, and information technology resources to develop, test, and implement archival systems for effective and efficient data curation. The Earth Institute of Columbia University represents various social science and physical science disciplines that focus on the creation, use, integration, and analysis of scientific data to address current problems that confront society.

The activities of the collaborative Board also improve the implementation and sustainable management of the Long-Term Archive. Recognizing that the costs of long-term preservation persist for the University, the Board carefully considers the costs of accessioning each data set in light of its potential value. Similarly, the appraisal criteria, preservation procedures, and the current and planned technological infrastructure are assessed periodically to identify improvements for

the preservation of the scientific data and to identify opportunities to reduce costs.

In addition, the diverse perspectives represented on the Board increases awareness of current and future challenges in data stewardship. While some collaborating units recognize a particular challenge that has not been considered by collaborating units, other units are able to offer suggestions to address the challenges recognized by others.

Contingency plans also have been established for the sustainable stewardship and management of the Long-Term Archive. If the current funding lapses, the composition of the Board and the management structure for the Long-Term Archive will change. The University Libraries would be represented on the Board by four members, which would include the member who serves as the Chair of the Board. The Earth Institute would be represented on the Board by two members and CIESIN would be represented on the Board by one member. In addition, personnel would be appointed by the University to manage and staff the Long-Term Archive. Establishing such contingency plans to specify changes in leadership and management of the Long-Term Archive in the event of a lapse of funding provides capabilities for the continuing stewardship of the archive and its resources over time.

Recognizing the need to improve and evaluate the long-term archive to ensure that it can serve as a trustworthy steward of the data entrusted to its care, a self-assessment was conducted using the Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC) document (OCLC and CRL, 2007). In addition to addressing technical issues, the criteria of the TRAC included requirements for ensuring the sustainability of the organization and the infrastructure to provide continuing long-term data management and stewardship. Results of completing the self-assessment included improvements to policies, plans, and procedures, as well as recommendations for further improvements and continuous reviews (Downs and Chen, 2010).

Subsequently, participating in the test audits that evaluated the draft standard for Audit and Certification of Trustworthy Repositories: Recommended Practice (Consultative Committee for Space Data Systems, 2011), which has since been adopted by the International Organization for Standardization (ISO) as ISO 16363 (Hughes et al., 2011), SEDAC was audited by a site visit team that examined the data center in terms of its compliance with the standard. While the auditors recognized the data center's potential for long-term stewardship, they also offered several suggestions for further improvement, including recommendations for additional efforts to ensure long-term sustainable management of its scientific data holdings. Based on the recommendations offered, a plan is being developed to improve capabilities for long-term

sustainable preservation and use of SEDAC data holdings and to apply for ISO 16363 certification when the certifying organizations have been established.

4 COLLABORATIVE SUSTAINABLE STEWARDSHIP

The stewardship of long-term archive collections of scientific data by collaborating university units and project partners serves as a model for sustainably managing the long-term stewardship of digital scientific data. With missions that include the preservation and dissemination of knowledge, universities are uniquely positioned to establish collaborative partnerships for scientific data preservation with data centers and funding agencies. In addition, with extensive experience in preserving knowledge and research resources, universities have the capabilities to accept the challenges of digital preservation.

Columbia University has over two-hundred and fifty years of experience in preserving knowledge to support learning, scholarship, and research. In comparison, many government agencies are quite young. Universities with such longevity in knowledge preservation have the experience and the responsibility to meet the challenges of digital preservation by engaging in collaborative partnerships with scientific projects, data centers, and the government agencies that fund the universities to engage in knowledge creation efforts. Similarly, government agencies, and other funders of scientific projects, need to recognize that it is not enough to simply support the creation of scientific knowledge. Support also is needed for managing and preserving new and existing scientific data and scholarly resources so that future generations of learners, researchers, and decision-makers can use these resources to learn and create new knowledge.

A collaborative stewardship model facilitates the natural evolution and potential transition of leadership among represented stakeholders. This model demonstrates how the establishment of contingency plans for cooperative stewardship may enable participating institutions, such as universities and government agencies, to provide long-term commitments for the management and care of collections that represent the irreplaceable scientific investments of an era. Establishing such institutional commitments for digital collections of scientific data can promote the continuous creation of new knowledge by enabling future scientists, scholars, students, and decision-makers to build on the foundations of previous science and scholarship (National Science Foundation Cyberinfrastructure Council, 2007).

Collaborative stewardship for preservation of digital data also offers potential benefits for collaborators, since it can facilitate knowledge sharing among collaborating organizational units to improve their understanding of digital preservation and stewardship issues. The shared perspectives on digital

preservation of each collaborating unit can contribute to the digital preservation knowledge of the other collaborating units. The experience gained from such collaborations can enhance the understanding of each contributing entity on issues and practices of digital preservation, enabling each collaborating unit to increase its potential for preserving collections of scholarly resources in digital form. Similarly, these units can share their digital preservation practices with other organizational units, either within the organization or in other organizations, to foster additional learning among collaborators.

5 DISCUSSION

Considering the large volume and diversity of scientific data and research-related information that can be created by the scientific and scholarly communities affiliated with educational and research institutions, organizations need to reduce the costs of preserving and providing access to archived resources on a continuing basis. This is an essential part of the commitment to the stewardship of their digital heritage. In light of the potential growth of digital collections that merit preservation, economies of scale will be necessary to ensure that all relevant scientific or scholarly work can be considered for long-term preservation and access. Collaborative efforts can foster economies of scale for digital preservation (Kwon et al., 2006)

In recognition of the potential costs of such institutional commitments, leadership also is needed to establish criteria for the appraisal and selection of scientific data considered for accession (Esanu et al., 2004). The continual costs for providing preservation services necessitate careful selection of data to ensure that the selected resources possess enduring value to warrant the costs of stewardship incurred over time (Downs and Chen, 2005; Gutmann et al., 2004;). Furthermore, recognizing that the perceived value of resources will evolve over time, plans must be established to reevaluate accessioned resources periodically to ensure that the scientific value of such resources continues to justify the continuing costs of their stewardship. Collecting metrics on the use and beneficial impacts of providing open access to data can help to justify continuing support for data collections and to encourage participation of stakeholders (Uhlir et al., 2009). On the other hand, plans for the reevaluation of resources also need to protect digital resources that happen to be reviewed when the represented topics or disciplines are unpopular or when budgets are being scrutinized injudiciously.

Just as an overall infrastructure strategy is needed to keep up with the rapid evolution of technology, an institutional strategy and a commitment for sustainable stewardship are needed to address the evolution of budgets that could reflect short-term plans at the peril of the long-term preservation of scientific data. Contingency plans for sustainable stewardship can help to ensure that institutional commitments for management of digital collections will continue

when future levels of support for their stewardship cannot be predicted. Universities that establish multi-departmental and multi-institutional agreements to provide sustainable stewardship can begin addressing the financial threats as well as the technological threats to our scientific and scholarly heritage. In addition, research is needed on developing organizational capacity for sustainable digital preservation (Ross and Hedstrom, 2005).

Establishing infrastructure and partnerships to preserve future access to digital scientific and scholarly artifacts is even more critical as new scientific approaches emerge for mining legacy data to create new knowledge and understanding. In addition, as archives, repositories, and data centers increasingly provide services that integrate data from multiple sources, the need for sustainable capabilities will become even more pressing, as a failure by one provider could affect the services received from an interdependent partner. Just as institutions have collaborated on scientific programs to explore uncharted areas and produce new knowledge, institutions also need to collaborate to establish the sustainable capabilities needed to preserve the knowledge that they helped to produce. Collaborative approaches for continuing stewardship are vital to the sustainability of digital collections and add to the trustworthiness of such institutions as long-term stewards of science and scholarship. In addition, observations of such collaborations can contribute to the knowledge of sustainable organizational capabilities to support the digital preservation of scientific data and scholarly works.

6 CONCLUSIONS

A strategic approach is needed to address the long-term sustainability of organizations that provide stewardship for current and legacy scientific data in digital form. Organizations that operationally manage projects will need to consider new models for ensuring sustainable organizational stewardship to manage digital research resources over time. Strategic partnerships and alliances also may be necessary to ensure that resources can be operated on a continuing basis to preserve and provide access to these data for future users. Managing digital resources will require long-term commitments from organizations that have a strategic interest in the future accessibility of our digital scientific heritage.

Establishing collaborative partnerships can enable the stakeholders to contribute to the decision-making that will influence the future state of the archive and its contents. In addition to contributing to the sustainable stewardship of the data being managed and preserved, bringing together stakeholders with diverse interests and expertise to manage archives and repositories also offers benefits for the organizations and the organizational units represented by the collaborating stakeholders. Since all of the answers to digital preservation have

not been found, establishing a long-term archive partnership of concerned parties with mutual interests in digital preservation also creates a testbed for exploring ideas and sharing knowledge about digital preservation issues.

The individual organizational units represented on the Board each contribute to the overall management of the long-term archive. These units also gain knowledge and understanding from each other on the preservation of digital resources, which improves practices within the represented units as well as improving digital preservation practices within the university and other collaborating organizations. The collaborating units also share information and resources for improving the common capabilities and shared infrastructure for the preservation of digital resources, including the long-term archive collection that they collaboratively manage. Furthermore, current and future challenges for data stewardship are shared amongst the collaborating units, enabling awareness of such challenges and sharing of possible approaches for mitigating potential risks.

With over two and one-half centuries of experience in preserving knowledge for future generations, the Columbia University community is preparing to preserve scientific data and scholarship that is represented in digital form. The Columbia University Libraries, the Earth Institute, and the Center for International Earth Science Information Network recognize their long-term role in digital data stewardship. In addition to the collaborative long-term archive activities described above, these organizational units have been collaborating with the rest of the Columbia University community to establish a university-wide e-science task force to identify immediate and long-term infrastructure and e-research needs that must be addressed to ensure long-term use of the digital legacy that is being created by the university community and its collaborators.

ACKNOWLEDGEMENTS

This paper is based on earlier work that was presented by the authors to the Digital Archive Preservation and Sustainability (DAPS) 2008 Workshop of the 2008 IEEE Symposium and Workshops on Large Databases. The authors gratefully acknowledge the contributions of the SEDAC Long-Term Archive Board members and the support received from NASA for work conducted under contract NNG08-HZ11C.

REFERENCES

Anderson, M., Gallinger, M. and A. Potter (2009). *The National Digital Stewardship Alliance Charter: Enabling Collaboration to Achieve National Digital Preservation*, UC Office of the President: California Digital Library, at http://escholarship.org/uc/item/11n4t7rk, [Accessed 29 January 2012].

Appelbe, B. and D. Bannon (2007). eResearch – Paradigm Shift or Propaganda?, *Journal of Research and Practice in Information Technology*, 39(2): 83-90, at http://www.jrpit.acs.org.au/jrpit/JRPITVolumes/JRPIT39/JRPIT39.2.83.pdf , [Accessed 29 January 2012].

Association of Research Libraries (2006). *To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering. A report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe*, September 26–27, 2006, Arlington, VA, at http://www.arl.org/bm~doc/digdatarpt.pdf, [Accessed 29 January 2012].

Ayris, P. (2009). LIBER's Involvement in Supporting Digital Preservation in Member Libraries. *Liber Quarterly*, 19(1).

Ball, A. (2010). *Review of the State of the Art of the Digital Curation of Research Data*, Project Report. Bath, UK: University of Bath, (ERIM Project Document erim1rep091103ab12).

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 [Accessed 22 January 2012].

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, [Accessed 22 January 2012].

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, [Accessed 28 January 2012].

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (2010) *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*, at http://brtf.sdsc.edu/publications.html, [Accessed 22 January 2012].

Carlson, J., Ramsey, A.E. and J.D. Kotterman (2010). "Using an institutional repository to address local-scale needs: a case study at Purdue University", *Library Hi Tech*, 28(1): 152-173, at http://dx.doi.org/10.1108/07378831011026751, [Accessed 29 January 2012].

Choudhury, S. and R. Hanisch (2009). The Data Conservancy: Building a Sustainable System for Interdisciplinary Scientific Data Curation and Preservation. *PV2009: Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data, 1-3 December 2009, Madrid, Spain at* http://www.sciops.esa.int/SYS/CONFERENCE/include/pv2009/papers/47_Choudhury_DataConservancy.pdf, [Accessed 29 January 2012].

Consultative Committee for Space Data Systems (2011). *Audit and Certification of Trustworthy Digital Repositories: Recommended Practice. Magenta Book, Issue 1*, at http://public.ccsds.org/publications/archive/652x0m1.pdf, [Accessed 22 January 2012].

Czerwinski, A., Sandmann, S., Stöcker-Meier, E. and Plümer, L. (2007). Sustainable SDI for EU Noise Mapping in NRW – Best Practice for INSPIRE, *International Journal of Spatial Data Infrastructures Research*, 2: 90-111, at http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/view/63/65, [Accessed 21 January 2012].

Digital Curation Centre and DigitalPreservationEurope (2007). *Digital Repository Audit Method Based on Risk Assessment, Draft for Public Testing and Comment Version 1.0*.

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 [Accessed 21 January 2012].

Downs, R.R. and R.S. Chen (2010). Self-Assessment of a Long-Term Archive for Interdisciplinary Scientific data as a trustworthy digital repository. *Journal of Digital Information* 11(1), at http://journals.tdl.org/jodi/article/view/753, [Accessed 25 January 2012].

Downs, R.R. and R.S. Chen (2005). Organizational needs for managing and preserving geospatial data and related electronic records. *Data Science Journal*, 4: 255-271, at http://dx.doi.org/10.2481/dsj.4.255, [Accessed 29 January 2012].

Downs, R.R., Chen, R.S., Lenhardt, W.C., Bourne, W. and D. Millman (2007). Cooperative management of a long-term archive of heterogeneous scientific data. *Proceedings of Ensuring the Long-Term Preservation and Value Adding to Scientific and Technical Data (PV 2007).*

Oberpfaffenhofen/Munich, Germany, at http://www.pv2007.dlr.de/Papers/Downs_CooperativeManagementOfALongTermArchive.pdf, [Accessed 29 January 2012].

Esanu, J., Davidson, J., Ross, S. and W. Anderson (2004). Selection, appraisal, and retention of digital scientific data: Highlights of an ERPANET/CODATA workshop. *Data Science Journal*, 3: 327-332, at http://dx.doi.org/10.2481/dsj.3.227, [Accessed 29 January 2012].

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, [Accessed 29 January 2012].

Guthrie, K., Griffiths, R. and N. Maron (2008). *Sustainability and Revenue Models for Online Academic Resources: An Ithaka Report*, at http://ithaka.org/strategic-services/sca_ithaka_sustainability_report-final.pdf, [Accessed 29 January 2012].

Gutmann, M., Schürer, K., Donakowski, D. and H. Beedham (2004). The Selection, Appraisal, and Retention of Digital Social Science Data. *Data Science Journal*, 3: 209-221. Retrieved May 8, 2011, from http://dx.doi.org/10.2481/dsj.3.209, [Accessed 29 January 2012].

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, [Accessed 28 January 2012].

Hughes, J.S., Giaretta, D., Ambacher, B., Ashley, K., Conrad, M., Downs, R.R., Garrett, J., Guercio, M., Lambert, S., Longstreth, T., Sawyer, D.M., Sierman, B., Tibbo, H. and M. Waltz (2011). Audit and Certification Process for Science Data Digital Repositories, IN53B-1629 Poster Presented at *2011 Fall Meeting, AGU*, San Francisco, Calif.

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, [Accessed 29 January 2012].

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://www-cs.ccny.cuny.edu/~jzhang/papers/AiChE09DataNet.pdf, [Accessed 29 January 2012].

Lesk, M. (2008). Recycling Information: Science Through Data Mining. The International *Journal of Digital Curation*, 3(1): 154-157, at http://www.ijdc.net/index.php/ijdc/article/view/71, [Accessed 29 January 2012].

Lewis, M. J. (2010). "Libraries and the management of research data", in: *Envisioning Future Academic Library Services.* London: Facet Publishing, pp. 145-168.

Meyer, E.T., Bulger, M., Kyriakidou-Zacharoudiou, A., Power, L., Williams, P., Venters, W., Terras, M., and S. Wyatt. (2011). *Collaborative yet independent: Information practices in the physical sciences*. Research Information Network (RIN) Report Series, IOP Publishing, at http://ssrn.com/abstract=1991753, [Accessed 27 January 2012].

National Research Council (2009). *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. Washington, DC: The National Academies Press.

National Research Council (2002). *Preparing for the Revolution: Information Technology and the Future of the Research University*. The National Academies Press.

National Science and Technology Council (2009). *Harnessing the Power of Digital Data for Science and Society. Report of the interagency Working Group on digital Data to the Committee on Science of the National Science and Technology Council*.

National Science Foundation Cyberinfrastructure Council (2007). *Cyberinfrastructure Vision For 21st Century Discovery*. Washington, DC: National Science Foundation, at http://www.nsf.gov/pubs/2007/nsf0728/index.jsp, [Accessed 29 January 2012].

Nestor Working Group – Trusted Repositories – Certification (2006). *Catalogue of Criteria for Trusted Digital Repositories, Version 1*, at http://edoc.hu-berlin.de/series/nestor-materialien/8en/PDF/8en.pdf, [Accessed 29 January 2012].

OCLC and CRL (2007). *Trustworthy Repositories Audit & Certification: Criteria and Checklist*, at http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf, [Accessed 29 January 2012].

Olsson, E.O. (2009). "Cooperation – a Key Factor for Sustainable Spatial Data Infrastructure", In van Loenen, B., Besemer,J.W.J. and J.A. Zevenbergen (Eds). *SDI Convergence: Research, Emerging Trends, and Critical Assessment*. Delft, The Netherlands: NCG, Nederlandse Commissie voor Geodesie, Netherlands Geodetic Commission, pp. 229-238, at http://www.ncg.knaw.nl/eng/publications/Green/48VanLoenen.html [Accessed 21 January 2012].

Ross, S. and M. Hedstrom (2005). Preservation Research and Sustainable Digital Libraries. *International Journal of Digital Libraries*, 5(4): 317-325, at http://eprints.erpanet.org/95/, [Accessed 29 January 2012].

Ruusalepp, R. (2008). *Infrastructure Planning and Data Curation: A Comparative Study of International Approaches to Enabling the Sharing of Research Data, Version 1.6*, at http://www.jisc.ac.uk/media/documents/programmes/preservation/national_data_sharing_report_final.pdf, [Accessed 29 January 2012].

Sobek, M., Cleveland, L., Flood, S., Hall, P.K., King, M.L., Ruggles, S. and M. Schroeder (2011). Big Data: Large-Scale Historical Infrastructure from the Minnesota Population Center, *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 44(2): 61-68, at http://dx.doi.org/10.1080/01615440.2011.564572, [Accessed 8 February 2012].

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, [Accessed 29 January 2012].

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, at [Accessed 28 January 2012].

Yawson, D.O., Armah, F.A., Dadzie, S.K.N.(2010). Ghana's Right To Information Bill: Opportunity For SDI As A Technical Infrastructure. *International Journal of Spatial Data Infrastructures Research*, 5: 326-346, at http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/view/172/261 [Accessed 21 January 2012].