Controlled Publication of Digital Scientific Data

How to balance free and open access to scientific data with privileged access to new results by authors while protecting them from being scooped by competing interpretations of their own data.

John J. Helly, T. Todd Elvins, Don Sutton, David Martinez, Scott E. Miller, Steward Pickett, and Aaron M. Ellison

Although the principle of equal access to data is a key aspect of U.S. government-funded science policy [10], there are strong, institutionalized, though sometimes contradictory, incentives for investigators to maintain proprietary control over that data; there is also increasing commercial and in some cases federal pressure to treat data as a commodity [4]. Efforts by the scientific community to prevent potentially deleterious international commercialization of scientific data through the World Intellectual Property Organization (WIPO) have had some success, thanks to support from the U.S. State Department. A recent example is the Anti-Piracy Bill (H.R. 2652) passed by the U.S. House of Representatives in 1998 but never approved by the Senate; it was similar in some ways to the WIPO proposal. Related pressures continue to build, including from within the U.S. private sector. It seems the commercialization of scientific data and treating it as a commodity represent an increasingly important aspect of how scientific data is published today; further complicating this scenario is the growth of the Internet-based business sector and the increasing commercial value of the data itself, especially biomedically significant data. These changes have influenced many aspects of scientific research, including the published content of professional journals, both online and on paper. The special role of research data in the advancement of science and its distinctly non-commodity character were identified as threatened by efforts to put a price on data [5].

Efforts by some scientists and policymakers to prevent the commercialization of scientific data reflect a certain irony and tension attending the purpose and politics of such data. On one hand is vigorous support for free and open access to the data consistent with the scientific

Function	Purpose
User Registration	A user ID and password are assigned to a given user while acquiring the user's email address and related contact information.The ID is used to audit data access and communication with users.
Data Acquisition	Data is acquired through contribution and submissions, along with at least a minimal set of metadata.This initiates the automatic creation of a unique name for the ADO and a transportable metadata file bundled within the ADO.
Search and Retrieval	A search system provides for spatial, temporal, and thematic (such as keyword) queries based on metadata content.
Deletion Control	The ability to delete an ADO is tightly controlled to prevent the arbitrary deletion of data copied by users. In a manner analogous to journal articles, no one should be able to unpublish data. Errata can be accommodated by publishing a revision of the data. An important special case to consider is the editorial peer-review process requiring confidentiality and the ability to remove an ADO if not accepted for peer-reviewed publication. A looser deletion policy might allow deletion of data if it had never been copied.
Assignment of Persistent Names	The persistent name, or accession number of an ADO, as in Figure 1, is used in the data repository to access the ADO, monitor updates of previously published ADOs, identify the retrieval of ADOs by users, notify users of anomalies or issues related to an ADO, establish precedence by publication date, and enable citation in other publications.
Quality Control and Quality Assurance Policy and Methods	This function can exist (or not exist) to varying degrees, exemplified by peer review and non-peer review, as well as by anomaly detection and reporting, though it must be stated explicitly. Some investigation is beginning on how to semiautomate QA/QC for specific types of data.
Access Control	Access control enables data contributors to specify a password only they know and that may be provided to other users to access the contributed ADO.This approach enables data submitters to independently control access to their own published data.Any user attempting to retrieve a password-protected ADO from the system needs to obtain that password from the data's contributor.
Traceability of Data Heritage	A mechanism for establishing the heritage of data contained within an ADO informs users of the data's measured, derived, or computed nature.This approach is also essential to preserving intellectual property rights analogous to claims of copyright or trademark.

Basic functions for the controlled publication of scientific data.

method and its emphasis on the reproducibility of results. On the other is a vigorous defense of the need for privileged access to new results by the data collector, as warranted by the need for scientific review to ensure that misleading or poor-quality data is not released [1]. Another factor is the competing incentive of protecting data submitters from being scooped by a quicker or alternative interpretation of their own data and consequent publication. The resulting ad hoc practice of delayed release of data has been tolerated within the scientific community and funding agencies, driven, in part, by the negative incentives inhibiting early publication.

Delayed release of data might also result from the lack of countervailing positive incentives for individual investigators to publish high-quality data as quickly as possible. The net result is that expensive, hard-won scientific data might go unnoticed by other researchers whose work could benefit from it. Major research funding agencies, including the U.S. National Science Foundation, expect that data resulting from their grants will be shared "... within a reasonable time"; the grant policy document of the U.S. National Institutes of Health states the grantee is the owner of the data, but investigators are "... expected to make the results and accomplishments of their activities available to the research community and to the public at large" [11, 12]. The oft-cited NIH data publication policy regarding its genome database comes to mind in this context, though it apparently contradicts overall NIH data policy. Despite such policies, the lack of positive incentives to publish data often results in the withholding of data as if it were owned exclusively by an individual researcher. Questions about the appropriateness of these delays are now being raised in the scientific community, especially as conventional journals move toward the integration of hyperlinks from their own online, electronic versions.

We therefore propose a mechanism to encourage and enable early publication of scientific data in a manner that produces beneficial side effects, including: incentives for high-quality data publication; establishment of individual researcher precedence;

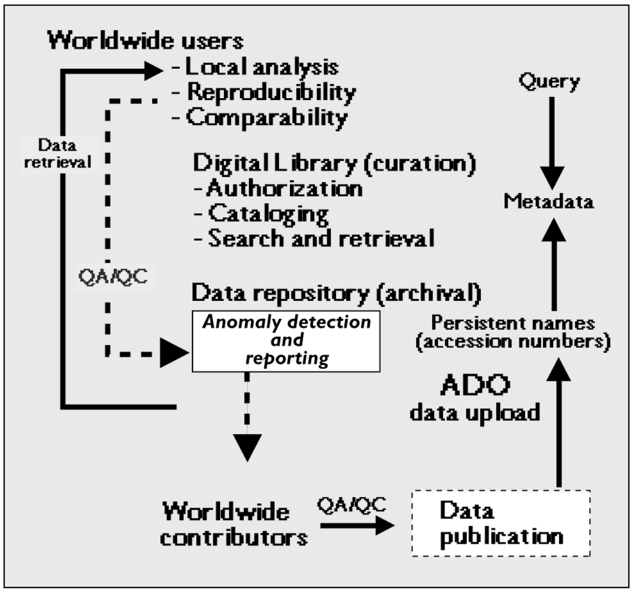

Figure 1. ADOs are produced when data is uploaded to the data repository where each ADO is assigned a persistent and unique name. This name and other metadata are passed to the digital library function where it is searched through a catalogue database.

off-site backup of data; a convenient mechanism for data sharing; a greater probability that a published data object is not overlooked; and a scheme for establishing a citable, persistent name for the published entity. Merely putting data on the Web is seriously deficient in this regard, as pointed out in [6], due in part to a lack of persistence and dubious quality. Indeed, one response to the study in [6] is the issue of citability of Web objects, as in [8].

Work on Web objects was directly influenced by a 1995 study by the Ecological Society of America's Future of Long-term Ecological Data Committee [1]. Although the Committee's initial motivation was to prevent the loss of at-risk ecological data, its members quickly realized it is really a special case of the larger issue of data sharing. For example, the seemingly simple task of identifying the existence of a particular type of data, locating the owner, and obtaining and comprehending the data can consume a great deal of time and effort. This investment often increases more quickly than the number of data sets involved; moreover, once appropriate data is found, its use is limited by related documentation, or metadata [7]. Questions about intellectual property also have to be resolved, and methods for the publication of data to ensure proper attribution and authorization for secondary use, or intellectual property rights, have to be developed and institutionalized. The Committee recognized that although the technology for establishing and maintaining such data collections was being developed throughout the Web, significant methodological and cultural hurdles to realizing their benefits were still not addressed.

Investigating solutions to these problems 1995–1999, we developed two experimental Web sites for the acquisition and dissemination of data: one for disseminating environmental monitoring data and policy information (see the San Diego Bay Project, sdbay.sdsc.edu), the other for publishing non-peer-reviewed ecological data (see the Caveat Emptor Ecological Data Repository, ceed.sdsc.edu). A third site, developed by Robert Peet, then editor of the journal *Ecology*, and collaboratively supported at the San Diego Supercomputer Center, is editorially controlled through peer review and serves as a prototype for the exploration and development of peer-review policies for publishing appendices and supplements associated with articles in *Ecology* (see esa.sdsc.edu/Archive). Based on what we learned developing these sites, we now propose a new method for the controlled publication of scientific data applicable to both peer-review and non-peer-review methods.

Digital Libraries, Data Repositories, Arbitrary Digital Objects

In designing an ecology data publishing system, we differentiated the function of a digital library from that of a data repository to clearly separate curation (maintaining content to support future use with domain-specific expertise) from archival (requiring computer resources and system administration expertise). This distinction is important because digital libraries tend to emphasize metadata content, while data repositories tend to emphasize data content (see the table here). One description [9] of the function of digital libraries emphasizes this distinction: "The primary purpose of digital libraries is to enable the searching of electronic collections distributed across networks, rather than merely creating electronic repositories from digitized physical materials."

In contrast, a data repository stores, maintains, and enables access to digital objects and manages hardware, software, protocols, interfaces, content synchronization, and related system-level infrastructure. In our approach, the basic objects to be published are computer files, either singly or as collections (see Figure 1). Each file is combined with its corresponding metadata in a public-domain archive TAR file format.[1] Other formats might be used, but an

[1] The acronym TAR stands for tape-archive and was originally implemented under the Unix operating system. The label gzip pertains to a common compressed archive format. We opted for this format in our implementation because other implementations of TAR and gzip are freely available through the Web for all major hardware platforms, though other formats might be used.

archive format enables multiple files and a directory structure to be stored together in a single file. This inherent convenience is especially important when a set of files, such as those containing, say, individual field surveys over a year logically comprise a single data object. The archive file is compressed using a public-domain method (such as gzip) to save space and time. We refer to the resulting file as an arbitrary digital object, or ADO, to emphasize the fact that it can contain anything that can be stored in a computer's file system, including measurements, images, sounds, and other digitally recorded data.

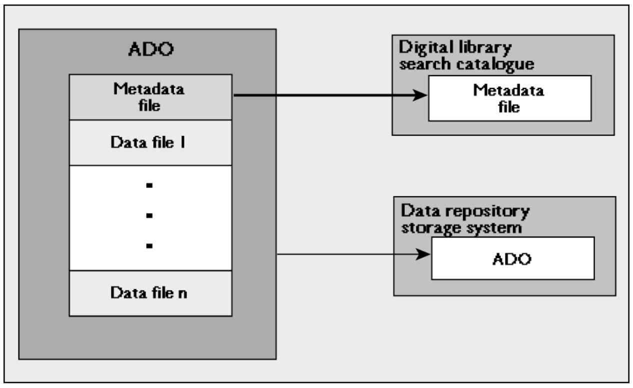
ADO
Metadata file
Data file 1
Data file n
Digital library search catalogue
Metadata file
Data repository storage system
ADO

Figure 2. ADOs consist of at least two files: one for data, one for metadata. The contents of the metadata file are copied and used as input to populate the search catalogue of the digital library function. The entire ADO is copied into the data repository's storage system.

Since ADOs are packaged as collections of data and metadata, their contents cannot be searched directly (see Figure 2). Searching is performed through a catalogue of metadata based on information provided by the data contributor during the data-publication process. The metadata, entered from a keyboard via a metadata editor application, is used to populate a database to enable the search and retrieval of the ADOs from a distributed data archive [2].

Data quality control and quality assurance (QA/QC) is an important part of the publication process we approach from two directions:

Peer review. Working in collaboration with the editors of the journals published by the Ecological Society of America, we developed a policy for the peer-review of digital appendices and supplements associated with articles published in the journals. This policy reflects the conventional notion of peer review of journal articles adapted to the unconventional notion of peer review of data. For example, the review process for the Society's Ecological Archives is organized around a human data editor responsible for soliciting reviews and deciding whether to accept or reject data papers. Either the data editor or the editor-in-chief of a particular journal makes an initial appraisal of a paper (data and metadata). If the topic and treatment seem potentially appropriate for the Ecological Archives, the paper is then reviewed by other experts in the field. It also undergoes technical review to ensure the data is organized logically and consistently, the metadata is comprehensive and adequate for secondary use, and the appropriate steps are taken to maintain data quality and integrity. Contributors can expect to hear whether their papers are accepted, rejected, or in need of revision within two or three months of submission.

Identification of ADOs. We developed a hardware and software infrastructure providing for the unique identification of ADOs, the acquisition of data and metadata, and the search and retrieval of contributed data (stored as ADOs) through a Web-based user interface; auditing and traceability of user retrieval of ADOs are provided via email. The ability to track user access to ADOs is important not only for determining who has obtained a particular author's data but for notifying these people when anomalies are found or a later version is available. The widely recognized problem of how to alert users to anomalies and revisions is emphasized by concerns regarding published gene sequence data, no matter who does the publishing.

We are also developing a formal method for data integration, or the combination of data, by merging or concatenating distinct computer files. It is predicated on a concept of levels of data, including explicit steps for QA/QC during the generation of a given level. Anomaly detection and reporting (ADR) is a key function of QA/QC processing, emphasizing the continuing interaction between a data submitter and the user community. Figure 1 outlines the ADR feedback loop, from data users to data submitters. Anomaly reports resulting from QA/QC processing are transmitted to the data submitters for resolution and subsequent notification of other users. Decisions about the correctness of any data must be made in the best possible way by the most qualified individual(s). Therefore, the most effective communication method is for the data submitter to also be the data originator. However, there is no way to enforce this correspondence, and in some cases (such as the death of an investigator) it would be impossible; we recommend that the data submitter be the authority on the data, playing an active role and performing maintenance and versioning as ADR proceeds.

Publishing Non-Peer-Reviewed Data

The ADO approach, along with our experience in the San Diego Bay Project, has helped us implement a Web site for the publication of non-peer-reviewed data we call Caveat-Emptor Ecological Data, or CEED, mentioned earlier. This data is made available to the public by the data submitters in the interests of research and the advancement of collaborative ecological science. The Web site for the Ecological Society of America's electronic publications—including three of it's journals: *Ecology, Ecological Monographs, and Ecological Applications*—publishes data using a conventional HTML approach rather than the ADO approach. However, the emphasis in establishing the site is on the policy issues in peer-review for data. We are now planning to integrate these approaches, identifying eight functions that have to be supported (see the table here).

Publishing data this way makes it possible for anyone to use it while protecting the submitter's intellectual investment—analogous to the publication of journal articles. It may also provide protection against the risks presented by potential future laws, because it enables individual scientists to publish and claim copyright to a uniquely identifiable collection of data within an ADO. Publishing and copyrighting material alone may provide sufficient motivation to publish data this way. The culture of academic merit may also come to recognize the value of publishing potentially priceless research data and rewarding it accordingly [3].

Placing copies of data in the public domain establishes independently verifiable precedence but creates a problem managing the volatile nature of research data (including detection of data anomalies and the addition of observations) that must be addressed by a method of quality control and interaction among users and contributors. Adding a quality-control method is a costly enterprise for which long-term funding represents a significant obstacle. It may be impossible to charge data users enough to fund the cost of a data repository. However, it seems this problem also represents a significant new opportunity for discipline-specific professional societies. In the same way they charge modest fees for access to online versions of journals, they can also charge for access to authorized data collections.

Although the details of a self-sustaining economic model have yet to be worked out, the concept should be considered. For example, individual scientists and curators could be payed a small royalty from the fees obtained by professional societies to support data maintenance—analogous to the fees U.S. government agencies charge for reproducing, filing, and maintaining documents. This model, along with copyright protections from having data published, the accrual of academic merit from the quality of the data, and the security of having an off-site, backup copy of the data, may substantially increase the rate the data is released. One thing we can reasonably expect is a benefit to scientific progress.

References

1. Gross, K. et al. *Report of the Committee on the Future of Long-term Ecological Data.* Ecological Society of America, Washington, D.C., 1995.
2. Helly, J., et al. A method for interoperable digital libraries and data repositories. *Future Gen. Comput. Syst. 16,* 1 (Nov. 1999), 21–28.
3. Helly, J. New concepts of publication. *Nature, 393* (May 14, 1998), 107.
4. Kaiser, J. Database bill worries scientists. *Sci. 280,* 5369 (June 5, 1998), 1499.
5. Kanciruk, P. Pricing policy for federal research data. *Bullet. Ameri. Meteorolog. Soc. 78,* 4 (Apr. 1997), 691–692.
6. Lawrence, S. and Giles, C. Searching the World Wide Web. *Sci. 280,* 5360 (Apr. 3, 1998), 98–100.
7. Michener, W., et al. Nongeospatial metadata for the ecological sciences. *Ecolog. Appli. 7,* 1 (1997), 330–242.
8. Molloy, M. Searching the Web, continued. *Sci. 281,* 5374 (July 10, 1998), 176–177.
9. Schatz, B. Information retrieval in digital libraries: Bringing search to the Net. *Sci. 275,* 5298 (Jan. 17, 1997), 327–334.
10. Uhlir, P. *Bits of Power: Issues in Global Access to Scientific Data.* National Research Council, Washington, D.C., 1997.
11. U.S. Department of Health and Human Services, Public Health Service. *Grants Policy Statement,* Washington, D.C., 1999.
12. U.S. National Science Foundation. *Grant Proposal Guide NSF 99-2.* Washington, D.C., 1999.

John J. Helly (hellyj@ucsd.edu) is a scientist at the San Diego Supercomputer Center at the University of California, San Diego.
T. Todd Elvins (todd.elvins@oracle.com) is product director in Oracle Corp.'s Voice Laboratory, San Diego, CA.
Don Sutton (suttond@sdsc.edu) is a project scientist in the San Diego Supercomputer Center, La Jolla, CA.
David Martinez (damartin@sdsc.edu) is a scientist at the San Diego Supercomputer Center at the University of California, San Diego.
Scott E. Miller (miller.scott@nmnh.si.edu) is the chairman of the Department of Systematic Biology in the National Museum of Natural History, Smithsonian Institution, Washington, D.C.
Steward Pickett (picketts@ecostudies.org) is a senior scientist in the Institute of Ecosystem Studies, Millbrook, NY.
Aaron M. Ellison (aellison@MtHolyoke.edu) is the Marjorie Fisher Professor of Environmental Studies in Mount Holyoke College, South Hadley, MA.

This research was supported by National Science Foundation grant BIR/DBA/DBI-9631091 at the San Diego Supercomputer Center, University of California, San Diego.