SEABIRD DATABASES AND THE NEW PARADIGM FOR SCIENTIFIC PUBLICATION AND ATTRIBUTION

SCOTT A. HATCH

U.S. Geological Survey, Alaska Science Center, 4210 University Drive, Anchorage, Alaska, 99508, USA
(shatch@usgs.gov)

Received 20 February 2010, accepted 31 March 2010

SUMMARY

HATCH, S.A. 2010. Seabird databases and the new paradigm for scientific publication and attribution. *Marine Ornithology* 38: 1–6.

For more than 300 years, the peer-reviewed journal article has been the principal medium for packaging and delivering scientific data. With new tools for managing digital data, a new paradigm is emerging—one that demands open and direct access to data and that enables and rewards a broad-based approach to scientific questions. Ground-breaking papers in the future will increasingly be those that creatively mine and synthesize vast stores of data available on the Internet. This is especially true for conservation science, in which essential data can be readily captured in standard record formats. For seabird professionals, a number of globally shared databases are in the offing, or should be. These databases will capture the salient results of inventories and monitoring, pelagic surveys, diet studies, and telemetry. A number of real or perceived barriers to data sharing exist, but none is insurmountable. Our discipline should take an important stride now by adopting a specially designed markup language for annotating and sharing seabird data.

Key words: data sharing, bibliometrics, inventories and monitoring, pelagic surveys, seabird diets, telemetry, computer networks, professional ethics, conservation, Seabird Research Markup Language, database design, World Seabird Conference

INTRODUCTION

When the Royal Society of London launched its *Philosophical Transactions* in 1665, the stage was set for the journal article to become the primary means by which scientists share their research findings with peers. The move to journal publication was gradual at first (Kronick 1976), and contemporary journals continue to be supplemented by books, monographs, conference proceedings, and student theses. While this model of scientific communication has served well for more than 300 years, all print media share a significant drawback—they are good for conveying words and ideas, but poorly suited for sharing data.

It is estimated there are currently more than 23,000 scientific journals in production worldwide, and more than 1.3 million peer-reviewed articles appeared in those journals in 2006 alone (Björk *et al.* 2009). For any given practitioner, the stream of new information to assimilate is vastly smaller, but the ever-increasing volume of literature in even a narrowly specialized field like seabird biology is daunting. An inherent feature of traditional modes of publication that contributes to the information bottleneck can be summarized in a word: *overpackaging*. Scientists are under pressure to produce as many publications as possible, and some minimum requirements apply for introducing any given parcel of work in a journal article—establish the context, describe the materials and methods used, report the outcome (usually a distilled rendering of the data), and discuss the significance of the findings. However, with the advent of information processing tools scarcely imagined 50 years ago, there is reason to question whether the traditional model of scientific reporting is entirely appropriate. My objectives here are: (1) to consider the implications of a paradigm shift now underway in the dissemination of scientific information, especially for applied disciplines like conservation science, (2) to review the tasks facing seabird biologists in shaping and adjusting to the new paradigm, and (3) to address some frequently heard objections and offer suggestions for overcoming barriers to data sharing. This article is supported by supplementary materials accessible at the *Marine Ornithology* website. It is hoped the supplementary materials can serve as a starter kit for those who would embrace the concept and promote the development of global seabird databases.

OLD AND NEW PARADIGMS

Scientific progress depends on the initiative and creativity of individual scientists and also on information sharing and synthesis. With most of the data generated by scientists embedded in traditional journal articles, only limited means are available for collating related findings and doing synthetic analyses. One option is to prepare a review paper, perhaps employing the statistical techniques of meta-analysis to somehow quantify the abstracted results of published studies and devise combined tests of research hypotheses (Hedges & Olkin 1985). Combining data sets directly would generally be preferred over meta-analysis, but often it is impossible to extract from journal pages the raw measurements of interest—a tedious and error-prone process in the best cases. Consequently it is often necessary to contact authors about access to their original data, perhaps with an offer to collaborate on a product that achieves at least an incremental synthesis of the topic at hand. None of these approaches is efficient, and they rarely approach the ideal of bringing all potentially useful data to bear on a given scientific question.

The emergence in the late 1900s of microprocessors, devices for mass data storage, and computer networks is overtaking even the printing press (circa 1450) as a watershed in the history of knowledge. The transformation encompasses both the tools we use to gather data (Burger & Shafer 2008) and our ability to share information rapidly and comprehensively (Whitson & Davis 2001, Alberts & Papp 2002). For one thing, computer technology facilitates the production of print media. The huge growth in paper journals would not have been feasible without it, and the impact of photocopiers and laser printers on the proliferation of hard copy has been equally spectacular. Now, driven by cost factors and demand, the transition to open access and fee-based electronic journals is proceeding rapidly, and will likely be complete within a decade. Such applications have already had a profound impact, even as they preserve the journal article as the basic currency of scientific exchange. But computers excel at more than just word processing, and the real paradigm shift will occur when we resolve to exploit maximally the potential of computer networks for sharing original data. In principle, it is possible to reduce greatly, and in many cases bypass altogether, the traditional packaging of data—packaging that is inefficient at best and dysfunctional at worst (i.e., when no delivery of reusable data actually occurs).

In the new paradigm, a scientist's first responsibility is to contribute his or her data to a publicly accessible database designed to capture the essential measurements and observations entailed in the kind of work they do. Each discipline will settle upon appropriate guidelines regarding the need for additional packaging. To a much greater degree than at present, scientific creativity, and credit for such creativity, will reside not in analyzing and making sense of *your* data, but in analyzing and making sense of *the* data relevant to a given question: that is, all pertinent observations and measurements that have been duly recorded in established databases. To be sure, the prospect of one or a small number of minds (aided by computers) accomplishing such a task depends on the scope of the question that is posed. The rhetorical point is that the new paradigm emphasizes synthesis over atomization in the scientific process.

What will be the outcomes of this paradigm shift? I suggest the following will be found among the major benefits: (1) Bigger questions, more robust answers. With vast amounts of data available at minimal cost, the individual scientist is empowered to tackle problems with far-reaching implications, including those most sensitive to spatial and temporal scaling. This applies especially to conservation and environmental sciences, where the problem domain extends to large-scale changes in the biosphere occurring in time frames of decades or longer. Models and conclusions that incorporate more of the relevant information are bound to hold up better than those incorporating less. (2) Better use of the scientific method, hypothesis testing in particular. Much research published in conservation science is not hypothetico-deductive (H-D) in form, even when it purports to be. Descriptive studies are routinely couched in H-D terminology to improve their chances of being published. This situation is better than the alternative, because hard-won data are nearly always of value in the scientific process, and should be available. However, as the contribution of small amounts of new data ceases to be the main focus of the standard journal article, authors will focus their energies on devising and testing explanatory hypotheses and selecting best-fitting models, as opposed to keying on scientifically trivial null hypotheses (Burnham & Anderson 2002). (3) Fewer authors per paper. This may or may not be viewed as a "beneficial" outcome, but authorship inflation is a widely noted and often lamented phenomenon occurring in most scientific disciplines (Weltzen *et al.* 2006). As a corollary to points (1) and (2), we might expect to see more bylines reflecting the intellectual contributions of one or two, possibly three workers, as opposed to an exhaustive list of those who had some role in generating a particular installment of data. (4) Little or no preventable data loss. It is an apt cliché that information gathered to meet specific management needs often dies with the workers who collected it, rarely to be exhumed from office files or gray literature reports. Besides making the data referenced in journal articles more accessible, the new paradigm provides options for data sharing by those not otherwise mandated to publish their measurements and observations.

There is clear evidence that our scientific culture is embracing data sharing as standard professional practice. Many funding agencies now require, as a condition of financial support, the timely archiving of project data in open access databases (e.g., NIH 2003, NPRB 2005). Likewise, most journal policies strongly encourage, and some require, that data referenced in published articles be placed in a suitable repository for direct access by the scientific community at large. The Ecological Society of America (ESA) launched its *Ecological Archives* in 1997 to promote and support the practice. *Ecological Archives* accommodates appendices and supplements to papers published in ESA journals, as well as "Data Papers" that stand alone as peer reviewed compilations of data and metadata (Bain 2005). Even before the proliferation of institutional and personal websites, some authors routinely offered "data available on request," and the new tools greatly facilitate that method of data sharing.

Clearly, reasonable options exist for online sharing of the data generated in any research project, no matter how specialized its design and resulting data records may be. But the data-sharing paradigm has greatest impact when the repository is a communally shared database, in which the structure and meaning of data records are consistent throughout. In the new paradigm, it is the responsibility of each discipline to identify aspects of its research agenda that can be readily captured in standardized record formats, and to create and use shared databases accordingly. One of the earliest and best examples of this is GenBank (Benson *et al.* 2000). GenBank was a natural response to the needs of molecular geneticists, because the basic units of data, nucleotide sequences, are universal and easy to capture in a shared database. For most disciplines, the task will be more complicated, and striking an appropriate level of detail is an important issue. Fortunately for those who work in conservation science, much of the information that is highly sought by the public, decision-makers, and scientists alike is obtained from observational studies—surveys, inventories, and monitoring—that are relatively amenable to data sharing using standard record formats. Understandably, the administration of shared systems can be a sensitive topic. I visit that issue after characterizing seabird databases and their contents.

SEABIRD DATABASES – DELINEATION AND DESIGN ISSUES

In the early 1990s, the Pacific Seabird Group resolved to create a number of shared databases for seabirds of the North Pacific region (Hatch 1992). The suite included a comprehensive colony catalog, a seabird monitoring database, a pelagic surveys database, and an index to world seabird literature (Hatch 1993). With powerful bibliographic tools now widely available in libraries and on the Internet, the need for any special effort in the last category has largely vanished. It was considered that seabird diet information might be captured adequately in a monitoring database, but that proved to be difficult. Until recently, telemetry of oceanic birds was so limited as not to merit a data warehousing effort, but clearly that is no longer the case. Today, seabird professionals envision products of the following types and appreciate the value and feasibility of proceeding on a global basis.

World Seabird Colony Register

This most basic of seabird databases presents the state of knowledge of breeding seabird distribution and abundance—the complete list of all known colonies and best available information on species composition and population sizes. Such information is much in demand for habitat (land and sea) planning, for damage assessment in the event of oil spills, and for the general information of everyone interested in seabirds. Population estimates are of whole colony sizes and inevitably are crude in many instances. In principle, the register is the easiest database to realize on a global basis, because some version of a "seabird colony catalog" already exists for most regions of the world, and data formats are reasonably consistent.

World Pelagic Surveys Database

Counts of birds at sea, whether from ships, aircraft, points on land, or small boats working the shoreline, comprise another natural grouping of seabird information. This database complements the colony register, and serves a similar purpose by documenting the distribution and abundance of seabirds over water, in all seasons of the year. Standard techniques were developed and used for most ship-based surveys in the last 30 years or so, but the database must accommodate an assortment of abundance measures. Wherever possible, the database should render survey results in consistent units—e.g. birds per square kilometer—and preserve a small cell-size in the rasterized data. Users may want to compute "best blend" estimates of seabird densities from multiple data sources and will need flexibility to combine and analyze data on different spatial scales.

World Seabird Monitoring Database

The monitoring database works with observations on seabird population parameters (numbers, productivity, survival, breeding chronology, and others) that are replicated over time. Generally, only a few of the colonies in a given region will be represented, and data often refer to sample plots rather than whole colonies. The format is annual observation records, grouped as time series, that document the means and variability of measured parameters for a given species and location. Supporting records provide essential metadata—series units and sampling design, reference to documents, project sponsors, observers, and contributor contact information. The monitoring database fosters comparative analyses of variability and trends over multiple scales and supports the use of seabirds as indicators of change in marine ecosystems.

World Seabird Trophic Studies Database

Early efforts to lump seabird diet information with monitoring results proved unwieldy, given the special nature of data records resulting from food habits analysis. A seabird diet database (or more inclusively, a *trophic studies database*) is required to capture the results of diet sampling—examinations of stomach contents, recovery of whole prey specimens, and indirect methods such as stable isotopes or fatty acid analysis. This information places seabirds properly in the context of management schemes for marine resources and is essential for understanding changes in seabird distribution, abundance, reproduction, and survival. The trophic studies database includes proximate analyses of birds and prey and facilitates sharing of archived tissue samples.

World Seabird Tracking Database

The arrival since the early 1990s of miniature tracking devices: satellite transmitters, GPS receivers, solar geolocators, suitable for deployment on seabirds, has created a wealth of new information and the need for a clearinghouse to serve that rapidly expanding field. With the fusion of surface locating devices and time-depth recorders, the research agenda becomes nothing less than a complete description of seabird movements and habitat use in three dimensions, over all phases of the annual cycle. Telemetry of individual seabirds complements both pelagic surveys and diet studies, and integrates the summer and winter ecology of birds at colonies chosen for in-depth monitoring. Telemetry data, originating in digital form and stored in consistent record formats (as far as possible), present few technical barriers to realizing a globally shared database.

Collectively, the five databases outlined above encompass most of the information typically sought for purposes of environmental and species management. Some facets of seabird biology—anatomy, social behavior, physiology, and others—are generally of lower priority from a conservation standpoint, and their findings will be harder to codify in shared databases. However, any measure repeated across years in a longitudinal study is appropriate to enter as a time series in the monitoring database—corticosterone levels in the blood, food stealing behavior in a colony, ticks carried per adult or nestling, and many other examples come to mind.

Properly designed and implemented, the five seabird databases are complementary. Unfortunately, confusion arises when seabird specialists speak of a "seabird colony database," by which they usually mean a product that partially combines, and in so doing muddles, the functions of the World Seabird Colony Register (WSCR) and World Seabird Monitoring Database (WSMD). Population size is one parameter among many that can be monitored in a seabird colony. Because it is often the *only* parameter monitored, some workers envision a hybrid product that mixes inventory data (best current estimates of colony size) and monitoring data (historical estimates of colony size). A simple test of proper design is whether the WSCR contains a *single* (updateable) record of population size for each seabird species and location. Users interested in the distribution and abundance of breeding seabirds consult the WSCR; those interested in fluctuations or trends consult the WSMD. Where multiple estimates of population size exist, they should be included in the monitoring database, while inventory values should be *averages*, or *predicted values* from regression or other quantitative methods. When whole colonies are censused repeatedly, updates to the register can be triggered automatically by updates to the monitoring database. That solution is preferable to leaving the task open to individual interpretation, and having users arrive at varying answers to the question, "What is the status (distribution and abundance) of species x in region y?" It eliminates any possibility of duplicate records, and makes the disposition of data logical and unambiguous. Further details of design and connectivity among seabird databases are outlined in the supplementary materials.

OVERCOMING BARRIERS

Seabird databases—global in scope and readily accessible online—unquestionably hold great promise, but a number of practical challenges confront any such effort (Hatch *et al.* 1994, Nelson 2009). The main barriers are: (1) Intellectual property rights—how to define and instill ethical behavior among data users; (2) Scientific attribution—how to acknowledge in a fair and consistent manner the contributions of both individuals and institutions; and (3) Technical issues—how to move data between computer systems with a minimum of human intervention.

Intellectual property and professional ethics

For professional authors and publishing houses, uncontrolled access to literary materials is a matter of personal or corporate livelihood (NRC 2000). Scientists, by and large, do not create published materials for profit, but they may feel similarly threatened when asked to share hard-won data openly, especially in advance of publication. After all, without data to communicate, scientists do not publish, and without publications, they do not prosper. But those who work in conservation-oriented fields also recognize and welcome the fact that many potential uses of their data exist that do not involve formal publication. Seeking to accommodate both objectives—open access for some purposes and proprietary control where publication is concerned—designers of the Pacific Seabird Monitoring Database (PSMD) drafted ethical guidelines to govern the behavior of users. Contributors attach to each observation in the system a data release attribute—*restricted*, *unrestricted*, or *provisional*. Depending on their intentions, users are expected to contact contributors about using any data flagged as proprietary. Communication between users and contributors is expedited by email messages generated on-the-fly in the web application, but ultimately the solution is nothing *less* than an honor system. The premise is that a code of ethics, widely publicized and understood by authors, editors, and reviewers, is the best way to manage professional behavior. The PSMD statement of ethical guidelines is included in the supplementary materials.

Scientific attribution – individual

If curiosity is the prime mover in science, attribution runs a close second. It is important for individual satisfaction and potentially decisive in matters of career advancement. Asked to engage in wholesale data sharing, practitioners are justified in asking, "How shall I continue to receive full credit for my contributions when my unpackaged data are used and reused by others?" Part of the answer lies in maintaining a close connection between databases and conventional publication. Turning again to the PSMD as an example, contributors attach to each observation in a time series all documents (published or unpublished) that make any reference whatsoever to the data in question. User instructions include the ethical guideline that, wherever possible, citation of original sources is preferred over an anonymous reference to the Pacific Seabird Monitoring Database (Appendix 3). Documents attached to observation records are stored in the database, for viewing or downloading during a user session.

Bibliometrics is a notable trend in professional circles, and a variety of measures are in circulation that try to quantify the impact of a scientist's career in terms of number and quality of publications (Hirsch 2005). Such approaches would not be possible in the absence of bibliographic databases such as Thompson Reuters' Science Citation Index or Google Scholar. In the same vein, it is easy to imagine a system for tracking the use of unpackaged data. For example, as a user accesses data online, a record of data usage might be generated automatically. Papers that depend on shared data would record their sources concisely in acknowledgments, or better still, all such accounting could be updated and reported electronically. In the new paradigm, data contributors will continue to receive "royalties" for as long as their information has value, and in direct proportion to its use by others. Combined with traditional mechanisms of attribution, the resulting system will be more complete, quantitative, and fair than anything we have known in the past. Clearly, new conventions must be adopted by all of science to be viable, but digital accounting of who did what is a logical path for science management to take in the digital age.

Scientific attribution – institutional

Some of the difficulty in creating shared databases results from turf battles—the reluctance of any work group to allow another's version of a database to emerge as the standard, thereby placing its originators in the coveted position of being "in charge" of all the available data. A possible solution for seabird biologists is to create an umbrella organization that provides neutral representation for everyone in the profession. Representatives from each of the world's seabird societies should convene as a World Seabird Consortium (WSC), a standing work group that, among other functions, manages the development and administration of global seabird databases. The physical system might consist of a server and Internet domain directly owned and managed by the WSC, or the consortium could link together any number of web resources distributed globally. The work of administering various components of the system might be delegated among constituent groups on a *revolving* basis, but ultimately the responsibility and attribution for seabird databases should accrue to the WSC.

It should also be emphasized that computer networks, seamlessly integrated, render concerns about institutional control essentially moot. In the parlance of object-oriented programming, a database is an object that furnishes outputs on request. Objects are cobbled together to create other, more powerful objects on a value-added basis. Innovation is an assured and desirable outcome when everyone has access to the same objects, because no endpoint exists in a continuing search for better tools to access, collate, and work with data. Institutional players should expect and welcome the development of systems that benefit from and outdistance their own efforts.

Technical issues

The last of the major hurdles in data sharing is a language barrier. Proprietary software, incompatible file formats, and network security issues all combine to frustrate the free exchange of information. In response, the World Wide Web Consortium created XML (Extensible Markup Language), a standard now universally embraced in the information technology industry (Møller & Swartzbach 2006).

XML itself is not the lingua franca one might imagine. Rather, it supports a growing *family* of languages used to serialize and annotate data in simple text files for transfer across the Internet. Within each discipline—whether it is chemical engineering, mathematics, graphics design, home gardening, seabird research, or any other—it is the practitioners' job to create and corporately adopt a dialect of XML to serve their needs. Just as web pages are composed in HTML (Hypertext Markup Language), chemists have CML (Chemical Markup Language), realtors have RELML (Real Estate Listing Markup Language), and space engineers have SML (Spacecraft Markup Language) to package their data using a standard vocabulary and syntax. To date, hundreds of interest groups have created their own XML applications to facilitate data sharing (see http://xml.coverpages.org/xmlApplications.html). Seabird biologists have only to follow suit.

An XML application, or dialect, is codified in an XML Schema—a formal definition of labels and data structures to which a document must conform to be a valid instance of its document type. Data in XML are structured hierarchically, in contrast to the relational model widely used to store data on disk. However, newer versions of relational database products such as SQL Server, Oracle, and Sybase are able to import XML data efficiently and can output query results as XML-formatted files. A relational system that captures all the data of interest is a good place to start in creating an application of XML.

A proposed schema for Seabird Research Markup Language (SRML) is included in the supplementary materials (Appendix 4). The specification includes language elements and data structures to support a world seabird monitoring database, and extensions for the world seabird inventory and trophic studies database can follow directly once those database schemas (Appendices 1 and 2) are adopted. Further extensions will be needed for pelagic surveys and seabird telemetry data. Certainly it is best if everyone agrees to speak the same dialect, but failing that, the situation is still recoverable. The flexibility of XML is such that one version of a markup language can generally be translated into others programmatically, with little or no data loss.

CONCLUSIONS

Constant pressure from reviewers and journal editors to condense papers as much as possible is a reminder that none of us can be afforded the luxury of overpackaging our information any longer—there is just too much of it. The rise of electronic data processing and a global computer network leads inexorably to open and timely data sharing as a modern standard of scientific practice. I take the optimistic view that no one loses and everyone gains under the new paradigm. But whatever else can be said of it, the new paradigm will clearly benefit (in our case) the birds, and for that reason alone no barrier should be permitted. For their part, journals can aid the transition by mandating submission of data to a shared database, if a suitable one exists, not later than the publication date of a paper, and also by encouraging and showcasing papers that achieve a substantial synthesis of data from one or more such databases.

In implementing its own version of the new paradigm, the seabird profession has some work to do. The World Seabird Conference scheduled for September 2010 will be the first of its kind and an excellent opportunity to begin creating the necessary infrastructure. A series of workshops on global seabird databases is planned. Seabird societies should take the essential step of instituting a World Seabird Consortium, increasing the likelihood of sustained action on seabird databases in the wake of the conference. An early item of business for the consortium should be the adoption of a standard markup language for annotating and sharing seabird data.

SUPPLEMENTARY MATERIALS

Readers are invited to access materials relevant to this article at the Marine Ornithology website (http://www.marineornithology.org/):

Appendix 1: World Seabird Databases – Schema Diagrams

Appendix 2: World Seabird Databases – Complete Documentation

Appendix 3: Guidelines for Ethical Data Sharing – PSMD Example

Appendix 4: Seabird Research Markup Language (SRML) – A Dialect of XML for Seabird Professionals

Appendix 5: Collegial Reaction to the Essay

ACKNOWLEDGEMENTS

I requested initially the comments of John Croxall, Ben Lascelles, and Bill Sydeman and am grateful for their thoughtful critiques of this essay. Somewhat emboldened, I approached a sizeable number of additional colleagues, whose kind indulgence exceeded any reasonable expectation (Appendix 5). I appreciate and have benefitted greatly from all the feedback provided. Mention of trade names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

REFERENCES

ALBERTS, D.S. & PAPP, D.S. (Eds). 2002. The information age: an anthology on its impacts and consequences. Washington, DC: NDU Press.

BAIN, J.L. 2005. An introduction to Ecological Archives. *Bulletin of the Ecological Society of America* 86: 86-91.

BENSON, D.A., KARSCH-MIZRACHI, I., LIPMAN, D.J., OSTELL, J., RAPP, B.A. & WHEELER, D.L. 2000. GenBank. *Nucleic Acids Research* 28: 15-18.

BJÖRK, B-C., ROOS, A. & LAURI, M. 2009. Scientific journal publishing: yearly volume and open access availability. *Information Research* 14: paper 391. [Available online at: http://InformationR.net/ir/14-1/paper391.html; accessed 19 October 2009]

BURGER, A.E. & SHAFFER, S.A. 2008. Application of tracking and data-logging technology in research and conservation of seabirds. *Auk* 125: 253-264.

BURNHAM, K.P. & ANDERSON, D.R. 2002. Model selection and multimodel inference. 2nd Edition. New York: Springer.

HATCH, S.A. 1992. Seabird monitoring in the Pacific. *Pacific Seabird Group Bulletin* 19: 3-4.

HATCH, S.A. 1993. A "PSG 2000" proposal for seabird databases. *Pacific Seabird Group Bulletin* 20: 13-14.

HATCH, S.A., KAISER, G.W., KONDRATYEV, A.Y. & BYRD, G.V. 1994. A seabird monitoring program for the North Pacific. *Transactions 59th North American. Wildlife & Natural Resources Conference (1994)*: 121-131.

HEDGES, L.V. & OLKIN, I. 1985. Statistical methods for meta-analysis. London: Academic Press. 369 pp.

HIRSCH, J.E. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences* 102: 16569-16572.

KRONICK, D.A. 1976. A history of scientific and technical periodicals. 2nd Edition. New York: Scarecrow Press.

MØLLER, A. & SWARTZBACH, M.I. 2006. An introduction to XML and Web technologies. Reading, MA: Addison-Wesley. 568 pp.

NELSON, B. 2009. Empty archives. *Nature* 461: 160-163.

NIH (National Institutes of Health). 2003. NIH data sharing policy and implementation guidance. [Available online at: http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm; accessed 22 October 2009]

NPRB (NORTH PACIFIC RESEARCH BOARD). 2005. North Pacific Research Board Science Plan. Anchorage, AK: North Pacific Research Board. 198 pp.

NRC (NATIONAL RESEARCH COUNCIL). 2000. The digital dilemma: intellectual property in the information age. Washington, DC: The National Academies Press. 340 pp. [Available online at: http://www.nap.edu (search ‘digital dilemma’); accessed 28 October 2009]

WELTZIN, J.F., BELOTE, R.T., WILLIAMS, L.T., KELLER, J.K. & ENGEL, E.C. 2006. Authorship in ecology: attribution, accountability, and responsibility. *Frontiers in Ecology and the Environment* 4: 435-441.

WHITSON, T.L. & DAVIS, L. 2001. Best practices in electronic government: comprehensive electronic information dissemination for science and technology. *Government Information Quarterly* 18: 79-91.