DATA PRACTICES, POLICY, AND REWARDS IN THE INFORMATION ERA DEMAND A NEW PARADIGM

C Barton[1], R Smith[2], and R Weaver[3]*

*[*1]Research School of Earth Sciences, Australian National University, Canberra, ACT 0200, Australia*
Email: cebarton@gmail.com
[2]Geophysical Institute, University of Alaska Fairbanks, PO Box 757320, Fairbanks, AK 99775-7320, USA
Email: roger.smith@gi.alaska.edu
[3]National Snow and Ice Center, Cooperative Research Institute for Research in Environmental Sciences,University of Colorado, Box 449, Boulder, CO 80309-2468, USA
Email: weaver@nsidc.org

ABSTRACT

As informatics becomes embedded in the scientific method, workload shifts from the user to the provider of data and information services and systems. Yet there is little incentive for research scientists to devote time to data management and system development. Our reward system can be adjusted to encourage responsible data management and open access practices, as well as motivate people to develop systems and services for the common good. At the same time, the status and professional infrastructure for those engaged in informatics needs to match traditional scientific and technical disciplines and create an attractive, competitive career path. Five readily achievable steps can be taken to redress these imbalances.

Keywords: Information era, Informatics, Data Science, Paradigm, Reward System, Data Publication, Data Accreditation, Data Attribution

1 INTRODUCTION

As we move further into the information era, informatics is becoming an increasingly important component of the scientific method (Baker et al., 2008; Baker & Barton, 2009, this issue). We use the term informatics[1] to refer to the science and engineering of the interface between users who need access to data, information, and services (e.g., research scientists, decision makers, teachers, the public) and computing and communications (cyber) infrastructures that provide access to data sources. Informatics encompasses the development of distributed data systems (virtual observatories), advanced computation and visualization technologies, interoperability protocols, semantics, development of data models and markup languages, data assimilation for numerical models, knowledge engineering for the discovery and integration of complex data, and innovative educational methods. We include activities that make data etc. available as well as the development of distributed data systems.

Through open access to data, information, and services, abundant new opportunities exist for delivering scientific results more efficiently, for broadening the scope, scale, and timeliness of research, and for communicating internally and externally. As a consequence, research scientists and others can achieve better results, faster, yet do less routine work. These benefits depend on the efforts of those who serve and preserve data and provide services.

The Electronic Geophysical Year, 2007-2008 (eGY, www.egy.org) embodied the underlying principles and practices for the establishment of the scientific information commons, the foundations of which were laid during the International Geophysical Year fifty years ago. The aim is to derive maximum benefit from the newly available information and communication technology opportunities (Baker et al., 2004; Baker, 2008). This

[1] Wikipedia (with minor editing) defines informatics as the science of information, the practice of information processing, and the engineering of information systems. Informatics covers the structure, algorithms, behaviour, and interactions of natural and artificial systems that store, process, access and communicate information. In addition to its own conceptual and theoretical foundations, informatics utilizes foundations developed in other fields and has computational, cognitive and social aspects, including studies of the social impact of information technologies.

requires that time and talent are devoted to informatics in all its aspects, and scientists and others need motivation to adopt responsible data management practices.

2 THE TRADITIONAL PARADIGM

The traditional paradigm for conducting scientific research places emphasis on publishing new scientific results, and it rewards researchers accordingly. In this context, data management in an “information common” sense (i.e., for the good of all and to facilitate re-use of data) is of secondary importance.

In the past, a typical scientific researcher undertook several time-consuming preparatory steps before being in a position to address a scientific issue. First, the researcher would acquire data – new data by direct acquisition and existing data by contacting colleagues, and then search through data centers, institutions, observatories, and so forth. This would commonly require physical visits to institutions in addition to postal and email exchanges. Second, the data accumulated would be re-formatted into a common style and probably ingested into a local database. Third, analytical code would be prepared – written anew if necessary, or customized from a proprietary code if available. Only then would the researcher be able to analyze the data, run models and simulations, prepare visualizations, and do the science. The end result of this process was typically a peer-reviewed journal article or monograph, with data reported in published tables.

3 A SHIFT IN THE WORKLOAD

The traditional paradigm has served us well and continues to do so in many areas of science. But today it is possible to streamline or bypass many of the preparatory steps through the use of distributed data and information systems and services. The reduction in workload for the user is impressive and unprecedented.

Nothing comes free. The new benefits to the user come at the cost of the effort that must be put into good data management practices and in building and maintaining data and information systems. Building systems and establishing underlying protocols is not easy. It requires skill, resources, time, and dedicated people exercising a high degree of cooperation worldwide.

We need to draw talented individuals into informatics as a career and to encourage research scientists to devote more time to informatics activities. Strengthening the professional community for informatics practitioners and elevating the status and visibility of informatics as a discipline are important factors in achieving these goals.

4 REWARD GOOD DATA MANAGEMENT

Despite the ongoing shift in data preparation from the user to the provider, our reward system for scientists continues to favor research output over data stewardship. Indeed, committees responsible for allocating research grants and for advising on career advancement are generally composed of successful research scientists who have a natural bias towards rewarding the practices that led to their own success. Little wonder then that many scientists continue to invest little time in data management and ensuring “their” data, which are often acquired using public funds, are available for re-use by others.

Optimizing output from the scientific community requires a balance between the contributions of data users and providers. An imbalance exists at present driven by insufficient motivation for scientists to devote enough time to good data stewardship and participation in developing systems and services for the common good. Complaining about any such imbalance will not produce change. Adjusting rewards will. Some readily achievable steps that will help are as follows. Progress in these directions is already well advanced in certain quarters, but much more can be done.

1. Establish procedures for publication of data (electronically) and citation thereof. This improves access to data, helps to preserve data, and gives due credit to both data providers and data centers for their work.
2. Require that all research proposals declare a data policy and have a data management plan. Policies need to ensure that data are available for re-use by others when legitimate intervals of restricted access have expired. This applies to both publicly- and commercially-funded research. Grant allocation bodies can ensure that declared data management plans are implemented by a combination of incremental funding (final

payment is made only when agreed data management obligations are acquitted) and denial of future funding to data management defaulters. For the latter, a research proposal must include a history of data management obligations and how they were met.

3. Make informatics (data) contributions a routine consideration in career advancement. Curricula vitae can all include a list of informatics contributions – systems that have been developed, data sets and databases that have been produced and made openly available, data services that have been provided, contributions to common standards and data models, data rescue and data preservation activities, and so forth. A data publication and citation capabilities, as described above, are obviously needed to make this possible.

5 PROFESSIONAL INFRASTRUCTURE

As the complexity, scale, and dependence on informatics grow, so does the need to migrate responsibilities to a professional class of experts. Like any other professional community of interest, informaticians have a need to share expertise and knowledge, cooperate, develop new opportunities, raise awareness, and engage with other professional groups, decision-makers, and the public. A manifestation of this is the emergence of special interest groups in existing professional bodies such as AGU, EGU, GSA, CEOS (WGISS), COSPAR, and CODATA. (Acronyms used in this paper are listed in Appendix 9.2.) Within ICSU, IUGS has a Commission for the Management and Application of Geoscience Information, and IUGG has just formed a Commission for Data and Information. The list continues to grow. ICSU regards data and information as one of the four priority development areas for science and has resolved to take a leading role by adopting the recommendations of the Strategic Committee for Data and Information (Appendix 9.1). The challenges of developing the Global Earth Observing System of Systems (GEOSS) illustrate the need for an established, interacting, inter-disciplinary community of informatics practitioners.

6 CONCLUSION

The way we conduct science is being transformed as we move further into the information era. The ready availability of data and information relieves research workers of much of the traditional burden of obtaining access to data and services and increases the efficiency and scope of their work. The cost is the additional burden carried by providers of data and services. To ensure that research workers adopt responsible data management practices and that there is an appropriate flow of time and talent into informatics, we must bring the reward system for informatics contributions into alignment with that of traditional scientific research. Five steps can be taken to establish informatics on a sound professional footing, with recognition, status, and career attractions commensurate with its present and expanding future role.

1. Elevate the status and visibility of informatics through demonstrations of achievement, presentations, publications, and engagement with all scientific communities, decision-makers, the public, and the media.
2. Establish a professional infrastructure for informatics scientists and engineers, with governance structures at national, professional society, and international levels.
3. Instigate data publication; provide and preserve attribution information for data, just as for published scientific results; and encourage citation of data (services etc.) so providers get credit.
4. Mandate a data policy and a data plan in research proposals; make research funding conditional on the adoption of good data practices and meeting a declared data management plan.
5. Make informatics (data) contributions a routine consideration in career advancement.

7 ACKNOWLEDGEMENTS

We acknowledge that others have already articulated the essential messages in this paper. We support their views! This publication was inspired by the Electronic Geophysical Year, 2007-2008 enterprise, which was supported mainly by IAGA, IUGG, LASP, NSF, and NASA.

8 REFERENCES

Baker, D.N., Barton, C.E., Rodger, A.S., Fraser, B., Thompson, B., & Papitashvili, V., (2004) Moving beyond the IGY: the Electronic Geophysical Year concept. EOS Trans. *Am. Geophys. Un.*, 85(11). pages 105 & 109.

Baker, D.N., (2008) A 21st-century vision for geophysical data management. *Physics Today*, pages 54-55.

Baker, D.N., C.E. Barton, W.K. Peterson, and P. Fox (2008). Informatics and the 2007-2008 Electronic Geophysical Year. *EOS Trans. Am. Geophys. Un*., 89(48), 485-486.

Baker, D.N. & Barton, C.E., (2009) eGY: Progress Toward a Global Earth and Space Science Information Commons. *Data Science Journal*, in press.

9 APPENDIX

9.1 SCID recommendations

The major recommendations of ICSU's Strategic Committee for Information and Data (SCID), adopted at the Maputo General Assembly of ICSU in October 2008, are that:

1) ICSU assert a much-needed strategic leadership role on behalf of the global scientific community in relation to the policies, management, and stewardship of scientific data and information;

In order to achieve this, ICSU must reform some of its current interdisciplinary bodies and establish a new committee that will provide overall strategic direction and advice.

2) a new World Data Services system be created (as an ICSU Interdisciplinary Body), incorporating the WDCs and FAGS as well as other 'state of the art' data centres and services;

This new structure or system must be designed clearly to support ICSU's mission and objectives, ensuring the long-term stewardship and provision of quality-assessed data and data services to the international science community and other stakeholders.

3) CODATA focus its activities on the three main initiatives identified in its draft strategy and extend its links to other organisations and networks to play a more prominent role within ICSU and within the wider scientific community;

This will require the close alignment of implementation mechanisms, e.g. working groups and task groups, with the three main initiatives identified in the draft CODATA strategic plan. The bi-annual CODATA conference should also be modified to provide closer links to ICSU priorities and the new World Data Services system.

4) a new ad hoc ICSU Strategic Coordinating Committee for Information and Data be established to provide broad expertise and advice to ICSU in this area;

This Strategic Coordinating Committee will act as an interface between scientists and data and information professionals that can advise on the data needs and possible solutions for existing and new ICSU programmes and other international initiatives. It will enable ICSU to establish visible and effective leadership and ensure proper coordination among ICSU activities.

5) ICSU National Members and Unions be strongly encouraged to establish committees or commissions, where these do not already exist, focussing on data and information issues;

Where national committees or liaison structures already exist for CODATA and/or the WDCs, consideration should be given to amalgamating and expanding these to integrate data policy, management and stewardship issues. Professional data services must be recognised and supported at the national level as part of the long-term infrastructure of science.

9.2 Acronyms

AGU – The American Geophysical Union
CEOS – The Committee on Earth Observation Satellites
CODATA – ICSU's Committee on Data for Science and Technology
COSPAR – ICSU's Committee on Space Research
EGU – The European Geosciences Union
GSA – The Geological Society of America
IAGA – IUGG's International Association of Geomagnetism and Aeronomy

ICSU – The International Council for Science
IUGG – The International Union of Geodesy and Geophysics
IUGS – The International Union of Geological Sciences
LASP – The Laboratory for Atmospheric and Space Physics, University of Colorado, USA
NASA – the US space agency
NSF – The National Science Foundation, USA
SCID – ICSU's Strategic Committee for Information and Data
WGISS – CEOS's Working Group on Information Systems and Services

(Article history: Received 31 January 2009, Accepted 17 February 2010, Available online 26 February 2010)