INTERNATIONAL COOPERATION FOR DATA ACQUISITION AND USE

Clark, M.J.
Department of Geography, University of Southampton, UK

Keywords: Data, error, data quality

Contents

Summary

Data can be regarded as both a scientific asset and a tradable commodity, and from both viewpoints there is a strong argument in favour of international cooperation to assemble, archive and disseminate such data. Routine deposition of data in an archive following completion of its operational use or the end of the project that generated it is widely regarded as highly beneficial. Access to such composite data sources can subsequently provide a basis for comparing sites or time periods, or for detecting change between them. This, in turn, helps to identify the underlying process controls, as well as providing a basis for hypothesis testing. Through these routes, it becomes possible to support more robust regional or global models, and to underpin the detection of impact or the design of an appropriate and evidence-based response. To yield these benefits, it is necessary to address the methodological and technical challenges of standardizing sampling, data model and data structure, without which integration of data is impossible. International data cooperation frequently starts as bottom-up initiatives

driven on a voluntary or pilot basis by individual scientists. Later, as the scale and investment increase, the professional data management infrastructures emerge and dominate. This can be demonstrated by the evolution of data cooperation within the discipline of geocryology, but applies in similar ways elsewhere. The successes of such initiatives are manifest, as are the widespread benefits, but it is instructive to also address ethical and professional cautions. Although the infrastructures may become independent on individual participants, data cooperation remains successful only if the professions sign up to their precepts.

1. A background to data cooperation

Most scientists and engineers prefer to consider the attributes of their data in the context of the specific problems that they are studying. At times, however, it is helpful to take a more general and abstract approach to the challenge of data and information management. This is particularly important when the aim is to design an information-handling strategy at national or international scale to serve the needs of a wide variety of present and future scientific or engineering projects. It is, however, difficult to design such a strategy without first considering the aims and functions of the data-handling process. These have been so long- and well-rehearsed in the literature since the mid-1980s, when the information technology and the commitment of scientists to data sharing began to converge substantially, that it is difficult to bring a fresh perspective. Yet the paradox remains that many scientists, projects and even complete programs continue to operate with little if any prior consideration to the ultimate destination of their data. Some reflections on the principles involved, and on a little of the practice in one specific area of the natural sciences (cold regions science), may thus serve to keep the debate alive and indicate the enormous progress that has been made (see *Global Data Networks in the Environmental and Life Sciences*).

1.1 Data as a scientific asset

If analogy is accepted as playing a role in scientific reasoning, then we could argue that data should be viewed as the fuel of the scientific engine. No matter how good the design of the engine itself, and no matter how skilled the engine driver, without data no work can be performed. No hypotheses, no tests, no operational models, no laws or theories. Of course, the analogy cannot be pressed too far, since it is possible both to create and conceptualize without data, but neither process could be envisaged as "scientific" in the absence of data, any more than the engine could be regarded as working in the absence of fuel. It merely has the potential to work, and herein lies much of the asset value of scientific data.

Every branch of environmental science has confronted the challenge of widening access to research data, but the argument is developed below in terms of cold regions science, which admirably demonstrates the best and the worst of the relationship between scientists (individually and as a professional community) and this most precious asset: in terms of data, as in so many contexts, we are a waste-producing society, and repeatedly ignore (or worse, reject) opportunities to preserve, recycle and enhance the "energy" which drives our work. There are many missed opportunities in the annals of the far North and South, but none so great as this.

To take this argument about the role of data much beyond the level of casual analogy would require an operational definition of science itself. For current purposes perhaps the simplest of outlines would suffice – with science being seen as common sense. *Common* in that it is based on replication and transparency such that results can be checked independently: *sense* in that it is based on measures or observations that can be sensed. Even at this level, it is clear again that data occupy a pivotal position. The observations (sensing) are captured directly by the data, and the common standard is achieved in part through data specification, standardization and quality control. From whichever direction we approach science, it thus appears that data lie close to its heart. Indeed, so self-evident is this statement that we could dismiss it as a trite truism were it not for the fact that data rarely head the list of hot science topics, and frequently are relegated to the level of taboo. We discuss our hypotheses, and may debate our theories, but to many scientists a critique of their data is tantamount to a personal attack. The data created by scientists are highly valued but frequently the focus of competition rather than cooperation.

1.2 Data as a commodity

While the data used in scientific investigations have been valued above all as a route to understanding, explanation, and prediction, they also have commodity value. At the same time, while the scientists have conventionally been seen as individual researchers driven by the quest for knowledge, most practicing scientists are employees whose skills and creativity contribute to an agency or corporate enterprise. The consideration of international data cooperation is thus driven as much by intellectual property rights, copyrights and patents as by optimizing the dissemination of knowledge. Data are thus capable of generating value (profit) for their user, and are therefore to be regarded as a tradable asset for their owner. In this context, "cooperation" requires a robust operational definition that moves beyond mere good will.

Enterprises will trade their data if they feel that there is net benefit to be gained from such a trade. This may, of course, take the form of a price that secures either inclusive or exclusive ownership of the data or, more likely, some form of licence to use the data. But it may be a less direct trade. For example, data exchange is a popular concept whereby donors gain immediate or ultimate benefit from being able to draw on data from a communal archive to which they donate their own data – a form of constrained altruism. The initial assumption here is that access to the archive is restricted to those who support it through donation of data, but the principle can be taken further. The public domain represents an unbounded archive from which all can draw, whether or not they donate. It is predicated on a blend of true altruism (some would say a belief in the values of classic science) and a realist model whereby advance and activity within the overall community are viewed as benefiting both providers and users of data, whether directly or indirectly.

The choice of model for data cooperation (priced, exchanged or open access) is far from just a philosophical issue. In the first place, it relates at least in principle to the funding of the data acquisition. "Data ownership" is relatively easy to allocate if individuals acquire data at their own expense, in their own time, and with their own facilities. But if an employer or other research funder is involved, then the rights will inevitably be more

complex. Increasingly, though, public- sector funding (whether on an operational basis using formal monitoring procedures, or commissioned as a one-off piece of research) is being linked to a requirement for ultimate public deposition of the resulting data, subject to the normal exercise of ethical considerations, including anonymity and confidentiality. This is a major boost to the creation and maintenance of the public domain data sources that are discussed below. In principle, at least, this confirms access to the data – but it also involves costs of deposition (including formatting, transmission, archiving and metadata) and risks of misuse (usually if metadata are inadequate).

It is easy to assume that the public domain is inevitably the right place for research data, but the full implications soon emerge as being more complex than this. Many societies are increasingly viewing an open approach to information as net beneficial, but arguments continue to focus on the possibility that some information can be damaging once publicized. A classic case involving natural science data is planning blight, whereby the market value or social satisfaction inherent in an asset is reduced by public release of detrimental information about it. For example, the public release of data suggesting that a house is in the active flood zone may reduce its market value and destroy its occupiers' peace of mind. Even if the zonation of risk is correct (which cannot be assumed), the intervention is a sensitive one where the overall social benefit of producing a more aware community that is better able to manage its own approach to risk is traded against individual good. If the zonation is speculative, for example, the suggestion that a property *may* be located on contaminated land thereby acquiring both a liability for clean-up and a drop in market value, the balance of social and individual good shifts. If the zonation can reasonably be demonstrated to be erroneous or unfounded, then the release of the data may trigger litigation. Without getting lost in either professional or ethical detail, it is plain that open access to data carries penalties as well as benefits.

The commodification of data thus presents the natural sciences with significant challenges. It is probable that the great majority of data in the natural and environmental sciences are generated in the public domain, but this does not guarantee either responsible archiving or ease of access to the archive (where ease is seen in terms of few restrictions as well as low costs). Even with government-generated data, it is often the case that the governments will see scope for recouping part of the acquisition costs by trading the data, with or without value-added services. Demographic (census) and topographic (map) data are classic cases where many governments place substantial costs on data products. The decision to take NASA Landsat data from the public domain into a commercial market also points up the dilemma. Commercial costs have undoubtedly reduced access to these data, but accelerated the spread and capability of competing systems to provide enhanced availability. And this seems to be the nub of the argument: availability (and quality) of data often has to be traded off against access: in order to pay for the acquisition (and thus availability) of high quality data, a charge is placed on user access to the data that is so high that it prohibits use.

The above paradox has polarized, over some 20 years, to the point where two distinguishable data strategies have emerged. The public domain (or low end-user cost) model argues that ease of access to data generates indirect social and economic benefit through the many activities that are fuelled by the data, and will also generate tangible, if

indirect, returns to government through an enhanced tax base and rateable value. The tradable asset (or high end-user cost) model is predicated on the notion that data acquisition, quality control, archiving, retrieval and dissemination are high-cost processes that should be funded by immediate users, with the cost being passed along the value-added chain to the end user. This implements a kind of value-added tax (VAT) or sales tax market for data. Since so many countries already operate conventional VAT and sales tax systems to net social benefit, it is unwise to dismiss this commodification model out of hand, though most scientists intuitively lean towards a low-cost or no-cost approach. Ultimately, of course, there is no "free lunch" in the data domain any more than elsewhere. The costs still have to be born, and the question is whether they are easier to manage and prioritize at the point of sale or use, or whether they should be handled more communally.

In attempting to resolve, or at least clarify, the above issues it is helpful to distinguish between operational data production or use and research data production or use. Production and operational environments, whether public sector or commercial, often generate large quantities of data both routinely (monitoring) and on a project basis. The research environment, however, may well suffice with pilot data and will often use existing data rather than generating new data to save cost and time. The two domains thus tend to approach data strategy with different value systems. To the researcher, data quality defined as a gold standard can be the key, and ease (time, cost, simplicity) of data acquisition is less important because quantities of data are small. To the operational user, data quantity is often the main challenge, quality is defined through fitness for purpose and acquisition ease becomes the key constraint. Such issues become mission critical in major information-based operational ventures, where data costs frequently dominate overall cost. In the analogy of data as fuel, it is tempting to see the physics and chemistry of the system dominating the research phase, whereas fuel economics will dominate strategies and options at the operational phase. The implementation choice is then between tuning the activity (the engine) to run on the existing data and refining the data (either for new acquisition or retrospectively) to support new performance capabilities. Research scientists may have the luxury of designing their own data model, data capture and data structure, but operational agencies and enterprises are often locked into existing legacy data sets (or legacy information systems) to the extent that they prefer to restrict performance rather than face the enormous cost and upheaval of re-architecting the system. Small local examples abound, but major cases are also easy to find, and include the national cost of moving from imperial to metric measures.

1.3 Data origination, archiving and rescue

A final stage in setting the background to the issue of international data cooperation is to acknowledge the different value systems that apply in the various phases of the data lifecycle: origination, archiving and rescue. Data origination can be regarded as the initial process of recording an observation. At this phase, data cooperation is important mainly in as far as it encourages the use of data and metadata standards and of the data quality control that will facilitate subsequent cooperation. Standards are further considered below (see section 3.1), and serve as a key attribute influencing the scope for cooperation. However, the imposition of standards is driven as much for temporal, spatial and inter-operator consistency within the originating organisation as for subsequent cooperation between organizations. Data archiving also in practice often

drives a data quality and standards exercise that builds on but is separate from that of the origination phase. Long-term archiving is a challenging process that requires both professional approach and technical resources, thereby generating substantial costs that become part of the business case behind data cooperation. Digital archiving media, now the *de facto* standard for data cooperation, are in constant flux and thus require repeated investment decisions – at each of which there is a temptation for those who do not understand, share or support the need for data cooperation to argue that retention of data is an expensive and unjustifiable luxury. It would be comforting to regard such short-sightedness as an aberration of bigoted individuals, but the evidence of the last few decades suggest that it can emerge even in the most prestigious and professional organizations. In the archiving phase, therefore, probably the main driver is an ability to establish a clear case for the value of long-term data archiving and dissemination (see section 2).

The need for the above dedication to data availability is nowhere more obvious than in the case of data rescue. It has been suggested that many natural history archives and museum collections are close to crisis point, particularly in areas of instability or economic stress. Substantial proportions of the collections are not catalogued, poorly housed or actually at the point of destructive decay. Irreplaceable baseline and reference collections have deteriorated or disappeared. Creative scientific work with many of these resources is thus severely hampered. It is not unrealistic to suggest that the same is true of the scientific data sets in many fields and in many regions. The neglect that continues to be shown is worrying at best, and woeful at worst. The work of many of the subjects' pioneers has already been lost, or has been stored with such paucity of control and organisation that its value is decimated. Data search and rescue is the term used to project the urgency and commitment with which we must react to locate and save what is left. In countries such as Russia a decade ago, the enormity of the challenge was such that international mobilisation was required to mount data rescue, but during the 1990s it became apparent that countries such as the UK, USA and Canada also had no basis for complacency. Every data set is an asset that should be considered for passing on to future generations with the same dedication that we so theatrically proclaim in the context of the sustainable development of the environment. Destruction eliminates choice and like wetlands, historic data sets are near impossible to recreate.

Time and again, potentially valuable data sets deteriorate, decay and ultimately disappear, whether through benign neglect or deliberate act of termination. The reasons are manifest. Individual data held by their originator (an independent scientist) are lost when that individual ends a particular interest, moves employment, retires or dies. Many of the "classic" data sets on which great ideas have been built are lost in this way – and such loss is to be mourned even if the interest is "merely" historical. Data are abandoned or binned when organizations move their physical accommodation or change their departmental management structure. Whole areas of research, and thus data retention, are lost when financial or political priorities change. And political upheaval itself, in extreme cases associated with war, brings destruction to data as well as other parts of the social infrastructure. This catalogue of risk is not just a theoretical construct: each and every one of its components has applied to the history of data loss and rescue in the case of cold regions (permafrost) science that is briefly explored as an example of topic-specific data management in Section 5. Whether data rescue is driven by

professional priorities (as with the International Permafrost Association data rescue initiatives) or commercial potential (as with oil company data-set recovery forays into the states of the former Soviet Union), it implies a recognition of the long-term value of data archiving and access, and a determination to institute a reversal of the inexorable trend towards data loss.

2. Value from data integration: the case for data cooperation

Thus far we have assumed that data availability, access and international cooperation are self-evidently beneficial, but to take the proposition for granted carries all the dangers of complacency. International cooperation is a trade-off decision with penalties and costs as well as benefits, and if it is to be promulgated and sustained then the case must be both explicit and convincing. Cooperation requires investment and commitment, and both represent a competition for scarce resources and capabilities. It is also pertinent to reflect that the contention between cooperative and competitive behaviour characterizes all aspects of social life across the animal rather than just the human domain, and there are many who challenge altruism and suspect cooperation of being counter-evolutionary.

The cooperative international data repositories (in practice, almost inevitably databases in the IT sense) are great facilitators of synthesis and integration. Integrated databases are categorically not the equivalent of 21st Century shoe boxes built simply to store facts until someone should chose to extract them. The modern data*base* is much more than a mere deposit box simply protecting whatever is stored within it, but has the potential to transform into a data *bank* and then a data *investment.* This distinction is substantive, and the financial analogy is helpful. A data bank puts its resources to work and thus adds value to them, while at the time honouring the integrity of the initial deposit so that the restoration of the original input is always guaranteed. The originator of the data set can remain at all times its owner and the controller of its employment. A data investment can be regarded as something more speculative, more entrepreneurial, and ultimately more profitable. It takes the deposited knowledge and actively puts it to work wherever the return on investment seems likely to be highest. The depositor (the data owner) may no longer be the active agent driving the use of the data, but stands to profit greatly through the productivity that is achieved by combining many deposits of information and by placing them in the hands of a skilled data investment manager. The payoff comes through co-operation and citation, thereby transcending altruism and creating sufficient benefit to encourage use of the cooperative approach.

But whether we opt for the notion of a database, a data bank or a data investment, we can view the agglomeration of data through a system of deposit and access (not, these days, necessarily a physical repository at some central location) as offering four interlinked roles:

- Organizational: Storage; archiving; quality assurance; added security.
- Facilitating: Accessing; selective access; visualization.
- Associative: Linking; sharing; analytical; spatial or temporal integration and modelling; linear, logical and deductive.
- Bisociative: Creative; chance; non-linear - the creation of new

knowledge.

Part of the data dilemma is that most scientists accept the benefit of data agglomeration at the level of principle, but many shy away at the level of practice if participation is optional. The challenge of the scientific community is thus to construct a system of data management that provides confidence in the ability of data sharing to lead to a strong net benefit to the individual. Mercifully, the case for international data cooperation does not rest on altruism alone (though this may motivate individuals), but on clearly defined mutual benefit from the availability of communal data resources (whether free or commercially-priced). Indeed, so broad is the range of benefits, that we can reasonably do no more than highlight some of the highlights, developing in particular aspects that have been prioritized in the Global Geocryological Database (GGD) of the International Permafrost Association (see Section 5).

-
-
-

Bibliography

The ideas developed in this chapter have drawn on a wide literature including, but not restricted to, the following:

Branson, J., Gregory, K. J., and Clark, M. J. (1996). Issues in scientific co-operation on information sharing: the case of palaeohydrology. *In* "Geological Society Special Publication." pp. 235-249.

Briggs, D. J. (1991). GIS development for broad-scale policy applications: the lessons from CORINE. *In* "Geographic Information 1991." (J. Cadoux-Hudson, and D. I. Heywood, Eds.), pp. 113-120. The Yearbook of the Association for Geographic Information. Taylor & Francis, London.

Clark, M. J., and Barry, R. G. (1998). Permafrost Data and Information: Advances since the Fifth International Conference on Permafrost. *In* "Proceedings of the Sixth International Conference on Permafrost, July 1998." pp. 181-188. International Permafrost Association, Yellowknife, Canada.

Kineman, J. J., Hastings, P. A., and Colby, J. D. (1986). Developments in global data bases for the environmental sciences. *In* "Proceedings 20th International symposium of Remote Sensing of the Environment." pp. 471-482. International research Institute of Michigan, USA, Nairobi, Kenya.

Mounsey, H., and Tomlinson, R. (1988). "Building databases for global science." Taylor and Francis, London.

Pickles, J. (1995). Ground Truth: the social implications of Geographic Information Systems, pp. 248. The Guildford Press, London and New York.

Rhind, D. (1992). War and peace: GIS data as a commodity. *GIS Europe* **1,** 24-26.

Webster, F. (1997). Threat to full and open access to data. *Science International, Newsletter***,** 11-12.

WMO. (1997). GCOS/GTOS Plan for Terrestrial Climate-related Observations. Version 2.0. World Meteorological Organization WMO/TD, Geneva.

Biographical Sketch

Mike Clark is Professor of Geography and Director of the GeoData Institute in the School of Geography, University of Southampton, UK. His research interests are information management, GIS, and environmental management. He is joint Chair of the Working Group on Permafrost Data and Information, of the International Permafrost Association.