THE ROLE OF SCIENTIFIC AND TECHNICAL DATA AND INFORMATION IN THE PUBLIC DOMAIN

PROCEEDINGS OF A SYMPOSIUM

Julie M. Esanu and Paul F. Uhlir, Editors

Steering Committee on the Role of Scientific and Technical Data and Information in the Public Domain

Office of International Scientific and Technical Information Programs

Board on International Scientific Organizations

Policy and Global Affairs Division

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

17

Fundamental Research and Education

R. Stephen Berry

In this presentation I will emphasize fundamental research and focus less on education, but I will comment on the impacts on education. First, I am going to look at this issue from the viewpoint of a scientist.

Two fundamental characteristics govern the way scientists carry out their activities. First, they depend on open access to information, because that information continually is expected to be used and to be challenged. One of the most important ways in which that information is used is in sustaining the verifiability that makes science different from virtually any other subject. It is the verifiability, which is the second characteristic, that makes scientific knowledge a firmer kind of knowledge that anything else we have. This information includes not only data in databases, but also the information found in journals and textbooks, the interpretation of data, and the concepts that underlie these.

I want to address almost exclusively information that is generated by either governments or not-for-profit institutions; I will not address proprietary information. In Session 2, we heard that information generated by science supported this way constitutes a public good. The justification for the support of that research is the production of the public good that comes from the science. A public good is one that does not diminish with use and has virtually no marginal costs for all of the users after the first user. But there is a special characteristic to scientific public goods: not only does the value of the scientific information not diminish, but it increases with its use. To satisfy the intent of the supporter of the research, society has to use that information and maximize its use, if possible, to achieve the values of the public good.

Historically, the scientific community and the publishing community in the broad sense—that is, the private publishers, the professional societies, and government through its own publications—always had a symbiotic relationship, as long as paper publishing was the sole outlet for the distribution and archiving of this information. That all changed with the Internet, which provided a faster, cheaper, and more efficient way for the scientific community to distribute and share its information. I think probably the sharpest example of that is the online e-print archive that Paul Ginsparg started in the area of high-energy physics.[1]

When that technological development happened, the relationship changed. It was no longer that comfortable symbiotic relationship; many scientists wanted to make use of the new medium. Many publishers, including some professional societies, did not want to use the Internet as a principal mode of distribution. In fact, many publishers

[1]See the arXiv.org e-print archive Web site at http://arxiv.org for additional information.

saw the use of electronic distribution not as a new way to provide a different kind of added value, but as a threat to the way that they make their living.

We have to keep one point in mind, which is very difficult for people outside the scientific world to realize. This issue became apparent to me when we were carrying out the *Bits of Power* study.[2] The study committee consisted of scientists, technologists, economists, and lawyers. During the first two meetings, the scientists and technologists and the lawyers and the economists were making no contact. They were talking as if they were in different worlds, but there was a key step that was a breakthrough. That was the articulation of the realization that for scientists the motivation is not the same as it is for the author of a novel. It is not making money from publication. For the scientist, the primary motivation, the currency if you will, is the propagation of ideas. This is the reason why scientists want to publish the results of their work. The scientist's primary goal is to distribute ideas and influence the thinking of others. If you use that as the basis of a value system, then an economist can slightly recast traditional economics with this other currency.

With this realization, the scientists and the lawyers and the economists on the committee were able to talk to each other in a very productive way. We simply had to find a way to establish the bridge to allow the economists to use their tools with the analog of what they normally use as the basis of evaluation. The financial monetary basis and the idea distribution basis were compatible when there was only one way to distribute the information in a storable, preservable way. In addition, the existing social and legal structure made open access via copyright and its exceptions. Protective, or restrictive, approaches changed that, or at least raised the specter. Those restrictions basically created an incompatibility, or threatened to create an incompatibility, between the way that the scientists operate and the way that the publishers operate. That incompatibility has been very difficult to explain, because people outside the scientific world usually do not understand the motivations of scientists.

The federal agencies that support the research have an interest in maintaining the distribution and archiving of the scientific information. And, of course, when a body, a law, or an activity acts to inhibit the distribution of that scientific information, then it is acting against the interests of the funding agency and acting against the interests of the national goals that justify the funding agency. That inhibition diminishes the public-good value of that information.

In extreme terms, which apply more to the case of the European Union Database Directive than to anything we have enforced in the United States right now, this thwarting of the distribution of information created in the national interest can be thought of as a theft of government property. The privatization, the inhibition of distribution, is in effect stealing from the government and putting into private hands the information that the government created for the purpose of public distribution.

People argue that scientists withhold information. However, the socially acceptable withholding of information in the scientific community is basically to allow scientists to (a) verify and establish the validity of what they are doing and (b) to be able to study, capture, and exploit their own research. So, for example, when crystallographers keep coordinates for one year, it is a way that the researcher with two graduate students in a small department can take the results of his own measurements and study them for that year. If the coordinates were published earlier, then a group of 30 could very easily do the studies much faster and publish in a few weeks something that would take the group of one faculty member and two graduate students several months to do. This is a kind of courtesy within the scientific community that is well accepted. It is a recognition, call it a soft spot or a weakness in the system, in which scientists compete with each other, and it is accepted.

There are some journals that require that data be deposited in publicly available databases. This is counter to that acceptance of the temporary withholding of information. By and large in the scientific community withholding data is really a bad thing socially. Scientists are very much looked down upon or scorned for withholding data. There is an ethic in this community that most of the time works pretty well.

Let us turn to the question of whether there could be a sound stable market for scientific information of the kind that would be captured by the legislation that has been proposed in the United States or by the European Union Database Directive. The value of small bits of scientific information is uncertain. The uncertainty of the

[2]See National Research Council. 1997. *Bits of Power: Issues in Global Access to Scientific Data*, National Academy Press, Washington, D.C.

value of scientific data diminishes as it is aggregated more and more. For example, all the data generated by the research supported by the mathematical and physical sciences division of the National Science Foundation have a high value. But what part of that body of data contributes the high value is very unpredictable. The value actually may not be achieved for a number of years.

One of the difficulties we have to contend with is that, although the aggregate data have a very high value, the way we decide what research to do or the way we fund the production of that information is at a much more desegregated basis. So the value of the data produced by the programs supported by one program officer at the National Science Foundation is very uncertain. The value of the data produced by one division of the National Science Foundation is somewhat less uncertain, because it is aggregated. But the funding is not done on the aggregated basis, the funding is done on a very disaggregated basis. Consequently, because of the high risk associated with each decision in the funding process, the result is that we cannot establish the value of the information produced by any one research project or even one program officer's set of projects. This is one of the ways that Congress justifies the relatively low support for the kinds of research that the United States supports, the National Institutes of Health currently notwithstanding. Real venture research is particularly unlikely to be adequately funded until it has proved itself.

What will privatizing do in this pinched market? It essentially will price the basic science community out of buying the information it needs. Or, alternatively, the scientific community may very well find its own ways to sustain itself, with its own new ways to distribute information outside the commercial market. The scientist does not have to publish in the existing journals, he does not have to deposit his data on a privatized basis. He can find his own pathways to do it.

To see how this is a plausible course, we can just look at the fact that scientists do have this other motivation—to maximize the distribution of information. The basic science community may find its own way to provide the information in a public domain, or some other open-access mechanism, outside the commercial publishing community. We have existing models for pathways that the scientific community can create for itself, such as the ArXiv.org and the Protein Data Base.

Professional societies represent a range of models that go all the way from astronomers, mathematicians, and physicists that have moved very much toward open access and public domain all the way to the other pole, to the American Chemical Society, which basically sees its publications as the principal source of its own support and therefore is very protective of its publications. One thing that we have not seen yet, and I think we will, is professional societies examining other models to support their publishing activities. The organizations that have considered it necessary to support themselves through publication have not yet started to look at other possible ways of doing this, but I think that we can expect to see that in the next four or five years.

There is a question now of who will pay for the distribution of scientific information. We have heard in this symposium that if a truly competitive market that would establish suitable pricing would do it, then that would be fine. If basic scientific information cannot be managed in a stable way by a competitive market, then society faces a choice: Which is more important to the society, the sustenance of scientific enterprise or the sustenance of the private information business? We have evidence from the fact that the federal government is a very important, key supporter of basic research and that we place a high value on the maintenance of science.

What is the responsibility of the private sector? If we look back at the publishing business during the 1960s and 1970s, when there was a lot of money for science and a lot of money for highly specialized journals, libraries were able to pay for the subscriptions on specialized journals. Publishing all kinds of scientific journals was a reasonable and profitable thing for publishers. However, a responsible publisher must monitor the profitability of every one of its ventures.

We have heard about the number of subscriptions that are being dropped by the university libraries that provide the principal sustenance for scientific publishers. As a result of the decreasing subscriptions, publishers must determine whether to discontinue these journals. Publishers have been very reluctant to take the responsibility to decide whether continuing to publish high-priced scientific journals is profitable. I challenge the publishers to examine one by one the specialized journals that they publish and decide whether they should continue to do so. I think it is a business decision that is hard to face up to, because it has been very profitable. But it is not clear that it is going to continue to be profitable. If the publishers decide it is not, then they should drop the journals. The

scientific community simply will have to find other alternatives to distribute its information, and we have some models for that to happen.

The public funder of research has some responsibilities. We talked about the costs of publication, but these costs along with those of collecting and distributing the information are far lower than the costs of the research. When we think of the observational sciences, I include the gathering of meteorological and astronomical data as part of the research, rather than as part of the publication process. This research done for public good is valueless unless the results are distributed. As such, the supporter of the research carries the responsibility to see that there is some mechanism to distribute the information. If the market mechanism does not do it, then the publisher of the information must be some institution or some mechanism supported by the supporter of the research. That small added cost for getting the information out has to be included.

Let us turn now to education. Education has thrived on access to scientific information through fair use for many years, and we will count on that in the future. But there is a problem that I will not discuss in much detail about the use of online and distance education, and the vehicles that are used for this. Are these going to become captured, privatized, and turned into the kinds of instruments that are not available for fair use? This is one of the new problems that education faces. As you know, there are open-source materials available, as well as commercially marketed counterparts.

One effect on education is already apparent, which is the impact of the nondisclosure constraints in some university-industry collaborations. Some of these collaborations have nondisclosure restraints that literally prohibit graduate students from one research group talking to the graduate students in another group about their work. This is an erosion of the environment in which we want our graduate students to be trained. This is a very serious surrender of principles of education to essentially gain a fast buck.

I think it is very disturbing that in much of our discussion even at this symposium, we have talked about universities as though their primary function is turning out commercially useful research. The primary purpose of a university, the primary product of a university, is educated students, and we must never lose sight of that. We must never surrender the mechanisms that produce truly educated students for secondary purposes such as commercially productive research. This is a very important perspective that we have to retain.

Let me go back now to the online education issue, which will lead me to a final perspective. In the case of online education where we have both models, open source and commercial, why not let them compete? Let us do the experiment and see whether the commercial products are the ones that people want to use, or the open-source ones, or both. We may very well have two kinds of users in the long run.

What are the next steps? In education, in scientific data, we are not at a stage where we have a clear-cut course ahead of us; we are going through a period of adaptation. We do not know what will be best. The only sensible thing for us to do is to try the different alternatives and see what works where. The worst thing would be to follow a restrictive course through legislation. The most productive course we could take is a permissive one to allow the different modes of activity to compete with each other. The Digital Millennium Copyright Act in this sense is going in exactly the wrong direction, because it is an inhibiting, rather than a permissive legislation. We need legislation that encourages the competition between different methods and allows us to try different options and see what works where.

I hope that we can recognize and adjust to that before we reach a crisis in which, for example, the scientific community strangles. My own personal hope, optimistic and naive as it may be, is that if we do face these restrictive forms of legislation, then the scientific community will be inventive enough to find its own way to solve its problems and sustain itself independent of those who insist on capturing the real estate and listing databases at the cost of whatever the scientific community might have to pay.