POSSIBLE DOWNSIDES TO DATA SHARING IN THE RESEARCH COMMONS: ASSETS AND LIABILITIES, OPPORTUNITIES AND RISKS

Peter Schröder*

**Data Archiving and Networked Services (DANS), Anna van Saksenlaan 51, 2593 HW Den Haag, The Netherlands*
Email: Peter.Schroeder@dans.knaw.nl

ABSTRACT

Large-scale data policies easily may have unplanned effects of ''homogeneity" on the available data supply. A narrowing of the scope of the data supply tailored to established research paradigms could limit the opportunities for unconventional, but also adventurous, new research directions with the risk of slowing down scientific progress. A systematic assessment of data portfolios not only on the aspects of quality, accessibility and sustainability, but diversity as well, could help to diminish this risk.

Keywords: Data policy, Research policy, Science policy, Data access, Diversity in data sources

1 INTRODUCTION

The idea of globally networked Commons for scientific data and information is a very realistic one. After all, scientists in many ways make up coherent communities already and chat with each other at their global village greens anyway. They also do not have to be afraid of overgrazing the Commons pastures because they live on air; not thin air, but air full of wear-proof ideas. When ideas are your core business, you can afford to act idealistically.

Newton advised us of the debt scientists owe to the virtual community of their predecessor giants worldwide. Robert K. Merton explained in detail how our scientific heritage has flowered as the result of cumulative, communal processes. Through these processes the sharing of data and information has been more or less standard practice in many research disciplines and contexts. So, in their own self-sufficient way, digital scientific Commons should be able exist in the midst of even the big bad world outside the intellectual community, uncorrupted by "greedy capitalist exploitation."

Data sharing may have an anarcho-idealistic ring to it, but it nonetheless will be in most cases the most practical and efficient arrangement for the exploitation of data resources. Communal open access to research data in principle does not have much to fear from the customs and habits of a hostile, profit- seeking business world – as long as taxpayers are prepared to foot the bill for the efficient management of research resources and then reap the rewards.

Of course, I am speaking of the main concept, the general principle. Scientists should know the vast differences between theoretical and empirical evidence, between the ideal world and everyday practice. Other authors in this journal issue explain the serious practical obstacles to efficient data sharing. Still others highlight the threats to the creation of knowledge as a public good posed by unbridled commercial interests in the public sphere.

2 THE ENEMY WITHIN

I will not speak about the outside threats to the benefits of data sharing, but of the risks that result from sensible communal data policies and data management from within—those originating in the community and the scientific process itself. At this moment, these risks are of a rather academic, theoretical order and are dwarfed by the current practical problems. Still it is worth the trouble to look in a common sense way into these inherent risks.

2.1 Openness and monopolies

Let us start with a truism: the community that inhabits the global scientific Commons is made up of individual scientists that have ideas and interests that may conflict with those of the community. Scientists are part of a community system, but at the same time are ambitious individuals as well. They can act as a common flock, but as individuals they will behave strategically, searching for recognition of their important contributions to the progress of science and perhaps the Nobel Prize. As a result, the scientific process typically consists of alternating phases of communalism and competition, of open sharing and of monopolies of resources of data, information and knowledge. It is good to realize that not only businessmen, but also scientists, can be keen practitioners of information monopolies. This goes for individuals, as well as research groups and organizations.

2.2 Homogenization of thought

Aspects of the inherent tension between communal effort and individual credits are exemplified in the concept of the *Matthew Effect*, a concept we can safely attribute to the abovementioned Robert K. Merton (for SCI purposes not to be confused with the other Robert Merton). We could look at the Matthew Effect in terms of personal glory, but should also remember that personal names are very useful as frames of reference that help provide order in complex information systems such as science. Not only a Google search works very well on personal names.

The interesting thing about the Matthew Effect is not the possible unfair credit given to only one individual, but its application to the large company of nameless others who practice their research under the same *scientific paradigm*, to use the concept of Thomas S. Kuhn. If we look at the big names of the scientific giants as labels for scientific schools, the Matthew Effect refers to processes of homogenization in scientific thinking that we rather tend to overlook.

2.3 Paradigmatic uniformity

A powerful body of assumptions shared by a majority of the community will have more impact on the direction of scientific developments than the inspiration and creativity of individual geniuses. Again, this is the balancing act between individual and social influence. The reigning scientific paradigms pervade the whole science system. They are the air the community breathes. But there is also the socio-organizational side. Researchers, reviewers, research managers, science policy makers, publishers--they all tend to dance to the same favorite tunes of the day. And in the end there is the economic side: it will not always be easy to get the right stamps of approval for grant applications that drift too far from the successful main roads.

2.4 Unplanned organizational processes

Decision making in research policies and management favor established scientific customs and traditions. Not only Adam Smith's *invisible hand*, but also Henry Ford's car paint color options are at work. An organized science system tends to promote a homogeneous output.

At the same time, we know that in the end scientific progress is impossible without periodic revolutionary paradigm shifts brought about by unconventional ideas of outsiders or upstarts. Science cannot progress without the challenges of controversial ideas and opinions. The way the science system operates most of the time, however, tends to limit the indispensable diversity of the research field. Large-scale sharing of research data can thus easily enhance this homogeneity.

2.5 Data as an autonomous force

The most obvious advantages of improved data access and sharing are an increase in the quality and productivity of research. But decision making on major investments in data collection and management can also put further limits on the diversity of research resources. The larger the pool of available data resources, the more pronounced the uniformity of the data.

Research data have been the undistinguished servants of the research pathway for most of the history of science. In the pre-industrial environment, data were considered to constitute an inextricable part of (specific) research trajectories, practically useless in other contexts. Even subordinate to theory: in the breakthrough of "data driven" research that produced the laws of Gregor Mendel, the data used contradicted the theory.

The research process has developed further (division of labor, increasing scales and volumes, digitization) into organizational structures along more industrial lines. Larger-scale collection of digital data have made it possible to use datasets autonomously and outside the context for which they were originally intended, for more diverse purposes of an increased number of researchers, at different times and places. In this new research paradigm data are emerging as a separate and distinct category of research resources, increasingly accommodated in separate and specialized (divisions of) organizations.

2.6 Safe consensus

Generally speaking, data collection, processing, management, dissemination and archiving are becoming a routine business of specialized entities that support research. Effective data policies and efficient data management have become more important for obtaining successful scientific results, while at the same time the production of databases has become less eligible for scientific or proprietary exclusivity.

Limits on research budgets, combined with the broad consensus required by more complex processes of decision making from large constituencies in research programs, generates pressures that can easily result in predictably safe outcomes concerning the data inputs, particularly in observational research. Chances are that this process will favor a predominantly “common denominator” data supply, also leading to a loss of diversity in the available data sources. When a researcher’s data requirements differ from the consensus of the prevailing orthodoxy, it can be difficult to get the right data for that research in some cases.

If not counteracted by mechanisms that increase the opportunities for data diversity, data collections in major research programs can promote large-scale “content paradigm” monopolies in the data supply. Without procedural interventions, such an approach to management will reinforce the position of the “normal”, middle-of-the-road type of research. This could diminish the opportunities for unconventional or deviant research pathways that still could be the start of important new research directions and scientific progress in the long run.

2.7 Public-interest criteria

The Ministry of Education, Culture and Science that I worked for in The Netherlands carries the responsibilities of the government for what is called the public domain (“Het Publieke Domein”) in these areas. The actual governance of educational, cultural and scientific affairs, however, is the responsibility of autonomous parties. The government guarantees a minimum supply of Education, Culture and Science for its citizens and supports this with a generous amount of public funding. The minimum level funded by the government is measured against the criteria of *Quality* (assessed by an appropriate professional forum), *Accessibility* (within reach of every citizen), *Sustainability* (a dependable, durable supply) and, last but not least, *Diversity* (offering citizens alternative choices).

These four criteria support the public domain and the public interest in general, and especially in science. Consequently. they characterize important qualities of the scientific Commons and the emerging research data Commons. When looking for ways to counteract the possible downsides of data sharing, these criteria should be kept in mind.

3 MORE THAN QUALITY

In the current priority setting, data policies are predominantly focused on promoting the criterion of scientific quality of the output. Most researchers will agree on the importance of the quality of data that constitute the input for research, and consensus on the assessment of quality is not that hard to reach. Assessing data policies against the criterion of accessibility is not that controversial either. The assessment of data policies against the requirements of sustainability will not always be easy, however, because appreciating the value of this requires a long-term view that may conflict with direct, near-term benefits.

Specialized data organizations may be in the best position to turn the liability of increasing homogeneity into the asset of systematic policies to guarantee a decent level of diversity in the data supply. Although these organizations may be preoccupied with other urgent problems, it is important for them to consider the future diversity of the data supply.

Research assessments that systematically consider the criterion of diversity could pose paradoxical demands on reviewers, but reality probably will turn out simpler. Evaluating proposals for data collection for purposes one would in everyday research practice consider meaningless, or worse, heresy or quackery, is not everybody's job, although even the most unorthodox cases could make intriguing intellectual exercises.

The main point, however, is for the research review boards not to judge diversity on the basis of separate individual applications, but on the basis of the overall portfolio of grant applications in a certain period or a certain research sector. It is the larger data ''package" that should be assessed on the aspect of diversity. Coherent data collection and management programs for longer periods of time could make it possible to give the demands of outlying research due consideration in a balanced data supply, based on an adequate review of current research and expectations for the future. Supported by sensible, concise protocols, the review on the criterion of diversity could be an important contribution to improving overall access and sharing of research data.

National research policy and funding agencies could make a start with the development a research data agenda: an agenda that addresses strategically the demand and supply of research data, and that incorporates not only the criteria of *Quality*, *Accessibility* and *Sustainability*, but also the indispensable exigencies of *Diversity*.