OPEN DATA FOR GLOBAL SCIENCE

Paul F. Uhlir[1] and Peter Schröder[2]*

**[1] National Research Council, 2101 Constitution Avenue NW, Washington, DC 20418, USA.* The views expressed in this paper are those of the authors and not necessarily those of their institutions of employment.

Email: puhlir@nas.edu

[2] Data Archiving and Networked Services (DANS), Anna van Saksenlaan 51, 2593 HW Den Haag, The Netherlands

Email: peter.schroeder@dans.knaw.nl

ABSTRACT

The digital revolution has transformed the accumulation of properly curated public research data into an essential upstream resource whose value increases with use.[1] The potential contributions of such data to the creation of new knowledge and downstream economic and social goods can in many cases be multiplied exponentially when the data are made openly available on digital networks. Most developed countries spend large amounts of public resources on research and related scientific facilities and instruments that generate massive amounts of data. Yet precious little of that investment is devoted to promoting the value of the resulting data by preserving and making them broadly available. The largely ad hoc approach to managing such data, however, is now beginning to be understood as inadequate to meet the exigencies of the national and international research enterprise. The time has thus come for the research community to establish explicit responsibilities for these digital resources. This article reviews the opportunities and challenges to the global science system associated with establishing an open data policy.

Keywords: Scientific data, Science policy, Information policy, Open access, Data management, Data licensing, International scientific cooperation, Cyberinfrastructure, e-Science, Internet

1 INTRODUCTION

The global science system stands at a critical juncture. On the one hand, it is overwhelmed by a hidden avalanche of ephemeral bits that are central components of modern research and of the emerging "cyberinfrastructure"[2] for *e-*

[1] See generally, National Research Council (1997), *Bits of Power: Issues in Global Access to Scientific Data*, National Academy Press, Washington, DC. "Data" may be defined as "facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors", National Research Council (1999), *A Question of Balance: Private Rights and the Public Interest in Scientific Databases*, National Academy Press, Washington, DC, p. 15. We define "public research data" as data that are generated through research within government organizations, or by academic or other not-for-profit entities, as well as public data used for research purposes, but not necessarily produced primarily for research (e.g., geographic or meteorological data, or socioeconomic statistics produced by or for government organizations).

[2] The U.S. Blue Ribbon Advisory Panel on Cyberinfrastructure anticipated an information and communication technology (ICT) infrastructure of "…digital environments that become interactive and functionally complete for research communities in terms of people, data, information, tools and instruments and that operate at unprecedented levels of computational, storage and data transfer capacity…" in (2003) *Revolutionizing Science and Engineering Trough Cyberinfrastructure: Report of the National Science Foundation Blue Ribbon Advisory Panel on Cyberinfrastructure*, National Science Foundation, available at: http://www.communitytechnology.org/nsf_ci_report/. We use the terms cyberinfrastructure and ICT infrastructure interchangeably in this paper.

science[3]. The rational management and exploitation of this cascade of digital assets offers boundless opportunities for research and applications. On the other hand, the ability to access and use this rising flood of data seems to lag behind, despite the rapidly growing capabilities of information and communication technologies (ICTs) to make much more effective use of those data. As long as the attention for data policies and data management by researchers, their organisations and their funders does not catch up with the rapidly changing research environment, the research policy and funding entities in many cases will perpetuate the systemic inefficiencies, and the resulting loss or underutilization of valuable data resources derived from public investments. There is thus an urgent need for rationalized national strategies and more coherent international arrangements for sustainable access to public research data, both to data produced directly by government entities and to data generated in academic and not-for-profit institutions with public funding.

In this paper, we examine some of the implications of the "data driven" research and possible ways to overcome existing barriers to accessibility of public research data. Our perspective is framed in the context of the predominantly publicly funded global science system. We begin by reviewing the growing role of digital data in research and outlining the roles of stakeholders in the research community in developing data access regimes. We then discuss the hidden costs of closed data systems, the benefits and limitations of openness as the default principle for data access, and the emerging open access models that are beginning to form digitally networked commons. We conclude by examining the rationale and requirements for developing overarching international principles from the top down, as well as flexible, common-use contractual templates from the bottom up, to establish data access regimes founded on a presumption of openness, with the goal of better capturing the benefits from the existing and future scientific data assets. The "Principles and Guidelines for Access to Research Data from Public Funding" from the Organisation for Economic Cooperation and Development (OECD), reported on in another article by Pilat and Fukasaku in this special issue of the CODATA *Data Science Journal*, are the most important recent example of the high-level (inter)governmental approach. The common-use licenses promoted by the Science Commons are a leading example of flexible arrangements originating within the community. Finally, we should emphasize that we focus almost exclusively on the policy—the institutional, socioeconomic, and legal aspects of data access—rather than on the technical and management practicalities that are also important, but beyond the scope of this article.

2 THE GROWING ROLE OF DIGITAL DATA IN THE RESEARCH PROCESS

The evolution of scientific research may be characterized by an accelerating growth in scale, scope, and complexity. These developments in scientific research have been accompanied by a substantial rise in costs. Overall expenditures on research and development (R&D) in the OECD countries increased from $163.2 billion in 1981 to $679.8 in 2003 (in constant prices, 2000 dollars: from $276.6 billion in 1981 to $638 in 2003)[4].

Not surprisingly, these trends also have elicited growing governmental policy involvement in scientific research at both the national and international levels. The research policy establishment has promoted greater cooperation between public researchers and the private sector, as well as greater international cooperation in public research[5]. The phenomenal growth of the cyberinfrastructure, particularly in OECD countries, has been both a facilitator and accelerator of these trends. It has further magnified the scale, scope, and complexity of scientific research by enabling the integration of research participants and information resources from multiple disciplines, sectors, and countries.

[3] "*e*-science" refers to "the large-scale science that will increasingly be carried out through distributed global collaborations enabled by the Internet. Typically, a feature of such collaborative scientific enterprises is that they will require access to very large data collections, very large scale computing resources and high performance visualisation back to the individual user scientist.....Besides information stored in Webpages, scientists will need easy access to remote facilities, to computer – either as dedicated Teraflop computers or cheap collections of PCs – and to information stored in dedicated databases." John Taylor, Director General of UK Research Councils. See: www.research-councils.ac.uk/escience/.

[4] Organisation for Economic Co-operation and Development (2005), "OECD Main Science and Technology Indicators", Paris.

[5] See, e.g., The Knowledge-based Economy (1996), OECD, Paris.

Continuously growing quantities of data about the universe around us are produced by government agencies, research institutions, and industry as a fundamental component of scientific research worldwide. Practically anything used for research purposes can be described and stored in a digital database. A genomic sequence, the speed of subatomic particles, a response in a social survey, the frequency of nouns in a text corpus, and satellite images of other planets all are used as research data. As described in the National Research Council symposium on *The Role of Scientific and Technical Data and Information in the Public Domain* in 2002:

> The rapid advances in digital technologies and networks over the past two decades have radically altered and improved the ways that data can be produced, disseminated, managed, and used, both in science and in all other spheres of human endeavour. New sensors and experimental instruments produce exponentially increasing amounts and types of raw data. This has created unprecedented opportunities for accelerating research and creating wealth based on the exploitation of data as such…. There are whole areas of science, such as bioinformatics in molecular biology and the observational environmental sciences, that are now primarily data driven. New software tools help to interpret and transform the raw data into unlimited configurations of information and knowledge. And the most important and pervasive research tool of all, the Internet, has collapsed the space and time in which data and information can be shared and made available, leading to entirely new and promising modes of research collaboration and production[6].

The production of a data set thus constitutes the first stage of improving the knowledge of some part of nature and society for further research and innovation. Rather than a linear process, however, the use of digital data is better conceptualized as a series of dynamic "chain link" feedbacks, broadening the usability of separate and related chains (see Box 1). The increasing supply of data frequently may be useful for purposes beyond those contemplated in the original collection. Many publicly funded data can be of great value for reuse by a broad range of public and private researchers, other types of socioeconomic applications, and the general public.

Box 1:

Research data: their place in the research process

For most of the history of science, scientific data were usually inextricably embedded in an all-embracing research process. Researchers mostly collected and used their own data in their own research projects and had access to few external data sources. However, with the advent of digital technologies and networks, together with the growing scale and scope of research activities worldwide, the various parts of the research trajectory have been loosened into separate specialised activities (as, for example, data collection or technical support) that may be executed by different entities, in-house or outside the research institute. In large-scale research, specialised data service institutes may operate independently from the research projects they serve. Different parties will have differing responsibilities and may have differing claims on 'their' parts of the trajectories. The various phases of the research process, including the upstream data management process, may be subject to different policies, regulations, and legislation. This diagram shows the main elements of the research and data trajectories.

1. The Research Trajectory

The general outline:

Data ⇨ Information ⇨ Knowledge

Detailed stages in the process with feedback (↶ . ↷) chains:

hypothesis ⇉ (methodology/design) ⇉ data collection/measuring ⇉ analysis/synthesis ⇉ results/publications/patents

[6] Uhlir, Paul F. (2003), "Discussion Framework," in *The Role of Scientific and Technical Data and Information in the Public Domain*, Julie M. Esanu and Paul F. Uhlir, eds., National Academies Press, Washington, DC, at p.3.

2. The Data Trajectory

Possibilities for data sharing once primary data have been collected:

data collection ⇨ primary data ⇨ processing ⇨ documenting ⇨ final data ⇨ dissemination & archiving
↳ **data sharing options**

The changes in the research process have not only been quantitative, but qualitative as well, leading to discoveries never before possible. For example, hitherto unconnected data elements can be assembled into unexpected new results. The research strategy developed by Rita Colwell, former Director of the U.S. National Research Foundation, in her studies on cholera is a case in point[7]. By combining large sets of data on sea life, earth observation, historical epidemiology, DNA analyses, and social anthropology she was able to demonstrate disease patterns that, without the use of ICT tools and access to all the diverse data, would have remained invisible. What is clear is that digital data play a central part in the emerging global science system and in the promise of *e*-science. And while most of the palpable progress to date has occurred in the more economically developed countries, the biggest payoffs from this new research paradigm could take place in the developing world.

These major changes in the structure and conduct of data-driven research using the cyberinfrastructure result in an increasing need for rational organisation and planning, however. A more transparent and predictable environment for access to and use of data resources would help to optimize the national and international research system.

3 THE EMERGING ROLES OF STAKEHOLDERS IN THE GLOBAL SCIENCE SYSTEM IN DEVELOPING DATA ACCESS REGIMES

Changes in the scientific research process are coupled with changing roles of the interdependent parties responsible for science policy and research management. Here we briefly examine the roles of these different stakeholders with regard to public science data policy and management in the context of the cyberinfrastructure. There are formal organisations, associations, and individuals involved at different (inter)national levels in the digital data activities. They represent specific economic, social, national, personal, and scientific interests, and play roles as experts and managers of research. These stakeholder groups all affect the development (or not) of data access regimes, both directly through governmental and institutional data management and policy implementation, and indirectly through normative and behavioural influences.

Governments are responsible for the legal and regulatory framework in which the research system operates, as well as for funding it with the taxpayers' money. Governments have core responsibilities for general public information rights, including overall policy over national science systems. More specifically, governments claim responsibility for overall policy over national science and innovation systems as a public good (e.g., research for public health, national security, general advancement of knowledge, and socioeconomic development). As funders of research, they have an interest in promoting accountability for the cost effectiveness and management of their public investments in research. Governmental policies are thus crucial for establishing a rational framework for managing and implementing the national science system and international scientific cooperation, most of which is now entirely dependent on digital networks. To the extent that public scientific data (and other types of information) are fundamental components of the modern research enterprise, governments have a responsibility to establish the policy framework in which the research organizations function and enable the rational development and exploitation of those information resources. This involves a balance between protecting and stimulating competitive and cooperative values at different levels of the research system.

[7] Colwell, Rita (2002), "A Global Thirst for Safe Water: The Case of Cholera", Abel Wolman Lecture at the National Academy of Sciences, available at: http://www7.nationalacademies.org/wstb/2002_Wolman_Lecture.pdf .

Research funding agencies are responsible for the actual allocation of taxpayer funds to the various research activities. They are accountable for the support and performance of the national science system. They comprise the experts who must develop and implement national research strategies and funding priorities in consultation with key representatives of the scientific community. Research funding agencies are also responsible for the more detailed allocation of public research funds, the support of specific elements of the research infrastructure (the people, facilities, and equipment), and the formation of policies specific to their constituencies. Digital science increasingly requires such specific policy and infrastructure support for networks, computing facilities, and institutional mechanisms for storing and making available the digital inputs and outputs of public research. This responsibility includes the possible establishment of specialized data centers, both within the funding agencies themselves and with their support at other research institutions. As the research funding agencies decide on the funding priorities, they are in a powerful position to influence the overall data policy and management regimes for the research institutions that they create or support.

Universities and not-for profit research institutes manage their employees' implementation of publicly-funded research programs and projects, subject to academic norms and the guidance of the sources of their funding (both public and private, and internal and external). These functions include support and management of ICT facilities and the resulting data collections and repositories for publications. Many academic research institutions now manage a large number of individual databases—as well as specialized data centers and more comprehensive institutional repositories and libraries—that are funded in whole or in part with public money. Whether or not they do have a data center, they have a responsibility for establishing policies for the access to and use of their expanding amounts and types or research data and information. These policies must be consistent with the requirements and interests of their funding sources, researchers, and other institutional stakeholders, and with the broader research community in which these institutions operate. Widespread uncertainty about possible conflicting interests and tasks of multiple stakeholders make the establishment of data access policies at research institutions crucial, though difficult. They require consistency at the higher policy level, as well as flexibility at the implementation level.

Learned and professional societies represent the formal side of the otherwise more loosely defined research communities. They provide a focal point for interaction and communication by their particular discipline communities, especially at the national level. They are major players in developing scientific norms, values, and standards such as academic freedom, scientific responsibilities, and increasingly regarding access to data produced by members of their research communities. They provide concentrated expert resources that combine the perspectives of the larger-scale changes in the operation of the science system with the first-hand experience from the specific changes in the day-to-day research practice in their disciple areas. The societies promote their views within their own communities by establishing formal and informal policies and codes of conduct for their members, through major conferences and their journal publications, and externally through interactions with policy makers and research managers.

International scientific organizations have a role similar to the learned societies, but at regional or global levels. The international non governmental scientific organizations (NGOs) must be distinguished from the intergovernmental organizations (IGOs). Among the IGOs relevant in this context are the Organisation for Economic Co-operation and Development (OECD), and some of the specialized agencies of the United Nations, such as the United Nations Educational, Scientific, and Cultural Organization (UNESCO). Relevant NGOs include the International Council for Science (ICSU), the interdisciplinary Committee on Data for Science and Technology (CODATA), the InterAcademy Panel on International Issues (IAP), and the Academy of Sciences for the Developing World (TWAS). These organizations have the subject matter interest and expertise to develop improved data policies and practices, as well as important contacts with the policy and research communities to promote them.

Industry research institutions generally benefit from greater access to scientific data produced by others. Traditionally, industrial laboratories and researchers tend to keep their own data outputs proprietary and inaccessible to other scientists and engineers. Keeping proprietary data inaccessible might entail lost opportunity costs for the owners as they will not be able to benefit from the results of additional research by other experts using those data. Industry research institutions increasingly outsource research to universities, however, partnering with university researchers often keeping the data on a proprietary basis. Industry-academic research partnerships are growing because of public policies favouring such arrangements and economic pressures on both academic and industrial research organizations. Public-private research partnerships may further complicate the management of the resulting data and the optimal allocation of rights to those data, as discussed further in Section 5.2.

Individual researchers generate increasing amounts and types of data, both as individuals and as participants in various kinds of formal and informal collaborations. Individual researchers sometimes show a different attitude to accessing data from colleagues for their own research than towards sharing 'their' data with colleagues. The informal culture at the working research level, with its strategic relations among researchers that are often invisible to outsiders, is dominated by traditions that in many cases have not yet caught up with recent developments in data policies and data management. However, much of the formal decision making on data access and sharing increasingly takes place at the institutional level. As the main producers and users of public scientific data, individual researchers ultimately have the greatest stake in the development of rational data access regimes and in the adequate funding and management of data collections and centers. Because researchers typically have been at the forefront of both developing and using the ICT infrastructure, they also have been some of the most influential players, together with their employing institutions, in creating new models of data access regimes from the bottom up. A great deal of data exchange and collaboration takes place informally on the internet between scientists as a result of their personal and professional relationships and in support of their respective research activities. Many researchers also have become part-time or specialised data managers.

The general public includes the taxpayers whose money is invested in public research and related data activities. Society in general has a strong interest in seeing that the fruits of those investments are effectively managed and used. The lay public generally is not concerned directly with the policy and management issues pertaining to national R&D, or to data from publicly funded research. Nevertheless, action groups of citizens may get involved in data access issues for various specific reasons and circumstances (e.g., local environment, health, or consumer safety). Increasingly, journalists do their own analyses of datasets used in the social sciences and the humanities. Moreover, with the broad public access to the internet in many countries, the potential user base for many kinds of public research data has expanded greatly, adding a further important dimension to the data policy debate, as discussed further in Section 5.

Each of these major stakeholder groups in the research enterprise has a major and growing interest in the development of more effective policies for access to and use of publicly funded research data. Although the sharing of data resources in networked cooperation has become standard practice in some fields, particularly in the more economically developed countries, in many cases researchers and their institutes experience too much uncertainty and barriers to make the most effective use of the new possibilities. This situation is exacerbated in less developed countries that also have less fully developed technical and human infrastructure for research, as well as institutional mechanisms and policy frameworks.

4 THE HIDDEN COSTS OF CLOSED DATA SYSTEMS

As described in Box 1, digital research data are emerging in the research system as autonomous resources, the uses of which are no longer tied solely to their original producers or purposes. There are, of course, data that have little value outside the narrow research project for which they were collected or that are not useful for lack of quality, insufficient documentation, or other deficiencies. Many types of data, however, can be used beyond the ambit of the original producers and users in diverse and unlimited ways, at different times and places, and potentially by anyone with access to the ICT infrastructure. The sharing of public research data opens up new opportunities to raise the quality and productivity of research, but the full realization of this potential requires additional attention to data policy and practice.

At the same time, there are competitive values and other legitimate reasons for restricting access to data from publicly funded research, as reviewed in the next Section. The different stakeholders involved may perceive conflicting interests when considering the benefits and drawbacks of open access to data. Many researchers tend to treat the data they produce through publicly-funded research as individual or institutional property, and this view frequently is reinforced by the lack of adequate policy guidance from their public funding sources.

There are, however, a number of negative implications[8] to the efficiency and effectiveness of the research system from unnecessarily balkanized and closed access regimes in light of the (quasi) public good[9] nature of such digital data resources.

Higher research costs. Most obviously, restricting access imposes structural inefficiencies and higher research costs. Many factual databases cannot or should not be independently recreated, either because they contain observations of unique phenomena, historical information, or cost a great deal to generate[10]. Moreover, databases with a monopoly status that are maintained on a closed proprietary basis will tend to result in higher, anti-competitive pricing[11]. Managing publicly funded databases on a restrictive, proprietary basis also adds substantial administrative overhead on both ends to make each transaction, further taxing the public research system. This is particularly exacerbated by public institutions that license data at high costs and restrictions to other public institutions.

Lost opportunity costs. Perhaps not as obvious, there is much less data-intensive research possible if the publicly-funded data are not shared or made easily available online. This results in significant lost opportunity costs that are certain to occur, but are difficult to measure[12]. A simple analogy might suffice to illustrate this effect. Just as it would hardly be cost-effective research management to limit the use of a telescope or an accelerator to the researchers and engineers who designed the instrument, it is a waste of effort and money to limit the use of data to the researchers responsible for their original collection and lose the potential benefits of greatly expanded applications for those data that may have some broader utility.

Barriers to innovation. The production downstream of copyrightable or patentable intellectual goods by both the public and private sectors depends to a large extent on access to the free flow of upstream public factual data and information. The overprotection or unavailability of public databases leads to deadweight social costs, taxing the innovation system in each country and slowing scientific progress[13].

Less effective cooperation, education, and training. A failure to make research data easily available, or erecting barriers that are too high, necessarily results in less effective interdisciplinary, inter-institutional, inter-sectoral, and international cooperation. In the same way, students may be less effectively educated and trained if they are unable to work with a broad cross-section of data. These barriers are reinforced in many cases by myopic policies that provide access and restricted use for a small number of pre-approved investigators formally associated with specific research projects and programs, even at an international level, while greatly constraining both access and use of those data by researchers and other potential users in "non-approved" disciplines, institutions, sectors, and nations.

---

[8] Reichman, JH, and Paul F. Uhlir (Spring 1999), "Database Protection at the Crossroads: Recent Developments and Their Impact on Science and Technology", *Berkeley Technology Law Journal*, Vol. 14, No. 2, at 819-821.

[9] Both the public nature of the research and the resulting data have public-good characteristics. A public good is both non-rival and non-excludable. The former means that it costs nothing to provide the good to another person once someone has produced it (i.e., it has a zero marginal cost of distribution). The latter refers to the characteristic that once such a good is produced, the producer cannot exclude others from benefiting from it. Inge Kaul, *et al*. (1999), "Defining Global Public Goods", in *Global Public Goods: International Cooperation in the 21st Century*, Kaul, *et al*., eds. Public research and publicly funded scientific data on digital networks may be considered as "quasi public goods" in that they are to a certain degree appropriable, although they nonetheless have public-interest characteristics that make them capable of production only if subsidized by public funding. See Callon, Michael (1994), "Is Science a Public Good?", in Science, Technology and Human Values, Vol. 19, p. 395.

[10] National Research Council (1999), *op. cit.* note 1, at p. 19-20.

[11] Weiss, Peter (2003), "Conflicting International Public Sector Information Policies and Their Effects on the Public Domain and the Economy", in National Research Council (2003), *op. cit*., note 6, p. 129-132; and Reichman & Uhlir, *op. cit*., note 8.

[12] It is difficult to determine what might have been possible if only the data were openly available. This was analyzed in at least one instance when the U.S. Landsat program was privatized in the mid-1980s. National Research Council (1997), *op. cit*., note 1, at p. 121-124.

[13] Reichman, JH, and Paul F. Uhlir (Winter/Spring 2003), *A Contractually Reconstructed Research Commons for Scientific Data in a Highly Protectionist Intellectual Property Environment*, in Law and Contemporary Problems, Vol. 66, Duke University School of Law, at p. 410-416.

Sub-optimal quality of data. Data organized in a closed environment frequently will be subject to a process of validation and verification from a substantially smaller and less diverse scientific community than data that are openly available. This will increase the risks of lower data quality and consequently of the quality of research outcomes. Less comprehensive opportunities for quality control will diminish the return on investments in data as well as research.

Widening gap between OECD nations and developing countries. Developing countries are particularly disadvantaged by a lack of availability or high barriers to access. Although not all databases produced in OECD countries are relevant in less developed ones, either because of their subject matter or geographic focus, those that do have broad applicability as a global public good will typically be unused in the developing world if there is a high price for access, and in many cases, any charge at all.

Unnecessary access barriers to publicly funded research data therefore result in diminished returns on the social and scientific capital investments in public research and in the inefficient distribution of benefits from those investments, even as the improving technological capabilities offer ever greater opportunities to increase that return.

5 THE SCIENTIFIC AND SOCIOECONOMIC BENEFITS OF GREATER OPENNESS

In view of the trends and the role of public data in science discussed above and the inefficiencies of the current ad hoc system, there are many compelling reasons for developing more comprehensive access regimes at the institutional, national, and international levels, with open access as the default rule. This is the case whether the data are produced within government or by entities funded by government sources, although some important distinctions apply, as outlined below.

Open access in the context of public research data may be defined as access on equal terms for the international research community, as well as industry, with the fewest restrictions on (re)use, and at the lowest possible cost[14]. This definition is also consistent with the "full and open" data policy used in various international environmental projects and in environmental (and other) research in the United States over the past two decades[15].

Because the value of scientific data lies in their use, open access to and sharing of data from publicly-funded research offers many advantages over a closed, proprietary system that places high barriers to both access and subsequent re-use. Open access to such data:

- reinforces open scientific inquiry,
- encourages diversity of analysis and opinion,
- promotes new research and new types of research,
- enables the application of automated knowledge discovery tools online,
- allows the verification of previous results,
- makes possible the testing of new or alternative hypotheses and methods of analysis,
- supports studies on data collection methods and measurement,
- facilitates the education of new researchers,
- enables the exploration of topics not envisioned by the initial investigators,
- permits the creation of new data sets, information, and knowledge when data from multiple sources are combined,
- helps transfer factual information to and promote capacity building in developing countries,
- promotes interdisciplinary, inter-sectoral, inter-institutional, and international research, and

[14] Preferably at no more than the marginal cost of dissemination (the cost of fulfilling a user request), which is essentially zero online.

[15] National Research Council (1997), *op. cit.*, note 1, at p. 1, 15-16.

- generally helps to maximize the research potential of new digital technologies and networks, thereby providing greater returns from the public investment in research[16].

Open access to factual data plays a vital enabling role in all these areas. Creating a level playing field for researchers and their institutes is impossible without broad and effective access to publicly funded research data. Nevertheless, there are essential distinctions to be made between data produced by government entities and by entities funded by government sources, as well as across disciplines and types of data. Moreover, there may be important and legitimate reasons for not making publicly funded research data openly accessible, but rather keeping them secret or proprietary, at least for limited times and in specific circumstances. These nuances and exceptions are complex, but important to understand in the development of access regimes. We touch on them only briefly below.

5.1 Policy considerations for data produced by government entities

The data and databases generated directly through government research have the following additional policy considerations favoring their open availability and unrestricted reuse[17]:

Legal considerations. Consistent with Article 19 from the Universal Declaration of Human Rights, national law on information rights should include public access to data and information produced by the government, and related freedom of expression by the public. Moreover, a government entity needs no legal incentives from exclusive property rights to create the data. Both the activities that the government undertakes and the information produced by it in the course of those activities are a public good, properly in the public domain. Data produced through public research frequently have global public-good characteristics[18].

Socio-economic considerations. Open access is the most efficient way to disseminate public data and information online in order to maximize the value and return on the public investment in its production[19]. There are numerous economic and non-economic positive externalities—especially through network effects—that can be realized on an exponential basis (though they may be difficult to quantify) through the open dissemination of public-domain data and information on the internet[20]. Conversely, the commercialization of public data on an exclusive basis produces *de facto* public monopolies that have inherent economic inefficiencies and tend to be contrary to the public interest.

Ethical considerations. The public has already paid for the production of the information. The burden of fees for access falls disproportionately on the poorest and most disadvantaged individuals (and researchers), including those in developing countries when the information is made available online. This is an important consideration for public, governmental scientific data that constitute a global public good.

Good governance considerations. Transparency of governance is undermined by restricting citizens from access to and use of public data and information created at their expense and on their behalf. Rights of freedom of expression are compromised by restrictions on re-use and re-dissemination of public information. It is no coincidence that the most repressive political systems make the least amount of government information, especially factual data, publicly available.

---

[16] See, e.g., Feinberg, S.E., Martin, M.E., and Straf, M.L., eds. (1985), *Sharing Research Data*, National Academy Press, Washington DC; National Research Council (1999), *op. cit.*, note 1; and Arzberger, *et al*. (2004), "Promoting Access to Public Research Data for Science, Economic, and Social Development", *Data Science Journal*, CODATA, p. 135-152.

[17] Uhlir, Paul F. (2004), *Policy Guidelines for the Development and Promotion of Governmental Public-Domain Information*, UNESCO, Paris, 49 p.

[18] See, e.g., Dalrymple, Dana (2003), "Scientific Knowledge as a Global Public Good: Contributions to Innovation and the Economy, National Research Council (2003), *op. cit.*, note 6, p. 35-51.

[19] Stiglitz, Joseph, *et al*. (2000), *The Role of Government in a Digital Age*, CCIA, Washington, DC.

[20] Ibid. See also Weiss, *op. cit.*, note 11; European Union (1998), *Public sector information: A key resource for Europe,* COM; and PIRA International (2000) *Commercial Exploitation of Europe's Public Sector Information,* Final Report for the European Commission, Directorate General for the Information Society.

Although there are strong arguments in favour of a default rule of openness in support of publicly-funded research, at the same time there are various legitimate, countervailing polices that may limit the free and unrestricted access to and use of government information, including research data. For example, there are statutory exemptions to public access and use based on national security and law enforcement concerns, the need to protect personal privacy, and to respect confidential information (plus other exemptions to Freedom of Information laws, where applicable)[21]. Government agencies also should respect the proprietary rights in information originating from the private sector that are made available for government use, unless expressly exempted. Governments may adopt policies as well against competing directly with the private sector in providing certain information products and services. The next Section examines more explicitly some of the additional factors that need to be considered in limiting disclosure of data in research funded by the government.

5.2 Policy factors to consider in disseminating government-funded research data

The access policies for research data produced by non-governmental entities with government funds[22] have rationales similar to those outlined above for government-produced data. There are additional factors that may come into play, however.

In some areas of research or in certain research programs, the recipient of a government grant or contract may have a specifically established period of exclusive use of the research data or until publication of the research results. These policies vary across disciplines, institutions, and countries, and in many cases there are no expressly stated, formal rules, just community practice and norms. In some instances, it is appropriate for data to be withheld even after publication, either because of confidentiality or privacy requirements, or because the underlying data are part of a longitudinal study spanning many years. However, generally accepted scientific norms and the exigencies of the scientific process that require access to data underlying published results for the purpose of independent verification make disclosure of such data following publication an essential prerequisite for sound science, even if there is no formal rule in place[23].

Moreover, open access to research data will not in itself result in usability. Optimum accessibility and usability presuppose a trajectory of proper organization and curation of a database with "added" value, which also adds costs to its production. Investments in preparing factual data for broader use may easily qualify for intellectual property protection and require some source of funding for providing enhanced access to other users. In most cases, however, there is a compelling reason to develop legal and funding mechanisms that will actively promote public accessibility to those publicly funded data resources. Such complications strengthen the case for further cooperation among the different parties involved in developing the policies and institutional mechanisms for improved data management and access.

Some OECD countries or research funding agencies also have policies that favour the commercialization of government-funded research[24]. For research areas in which commercial applications are inherent or desirable, there will be additional motivations for the researcher to keep the data proprietary and under conditions of trade secrecy, at least until patent rights are secured. Furthermore, the non-governmental research may involve a mix of public and private funds or partners, or include parties from multiple countries, which can complicate the allocation of rights in

[21] For a compendium of freedom of information laws and their exceptions, see http://www.freedominfo.org.

[22] This is certainly the case in which public sources provide 100 percent of the funding. As the percentage of public funding in any given research project diminishes the corresponding rationale and arguments for full policy control become weaker as well.

[23] See, e.g., National Research Council (2002), *Community Standards for Sharing Publication-Related Data and Materials*, National Academies Press, Washington, DC..

[24] Perhaps the best known of these is the 1980 "Bayh-Dole Act" in the United States, which states in part: "It is the policy and objective of Congress to use the patent system to promote the utilization of inventions arising from federally supported research or development…[and] to promote the collaboration between commercial concerns and nonprofit organizations, including universities….", Public Law No. 96-517, Sec. 6(a), 94 Stat 3015 (1980), codified as amended at 35 United States Code, Sec. 200.

the research data. In such cases, the application of an open access data policy also may be inappropriate, unless expressly agreed to by all the participating parties.

The issues raised in public-private relationships take many forms and contain some inherent tensions, such as openness versus exclusivity, public goods versus private investments, public domain versus proprietary rights, and competition versus monopoly, among others. This mix of motivations, priorities, and requirements is context-dependent, typically unique to the parties involved, and frequently not amenable to inflexible statutory and regulatory frameworks. In such cases, the ordering of the respective rights and interests of the parties involved is most efficiently accomplished through contracts. Such private agreements provide maximum flexibility within the larger research policy context. What is especially important to emphasize here is that such agreements can in many cases provide for conditionally open access that advances the public interest goals associated with the public funding, while effectively protecting existing proprietary private interests[25].

This bifurcated ordering of interests can take many forms. At the most basic level, it is possible to provide free access for not-for-profit research and education (and other) users, while restricting commercial users and uses to a reimbursable, or even for-profit, basis. Various techniques of price discrimination and product differentiation may be similarly employed, based on factors such as time (e.g., real-time access for commercial users vs. delayed access for non-profits), scope of coverage (e.g., geographic or subject matter limitations), levels of customer support or service, and other possible distinctions[26]. Such strategies can help promote scientifically and socially beneficial access and use, not only in the complex public-private research relationships, but even in exclusively private-sector settings[27].

In addition to these complexities within the government-funded academic and not-for-profit research context, there are important distinctions that need to be made among different disciplines and types of research. A major difference is between those areas of science that are dominated by "big science" research projects and programs, and those that remain predominately "small science" research endeavours, performed by a single investigator (or small group)[28]. The former are typically cooperative, whereas the latter tend to be more competitive, or at least insular. Most big science programs have instituted a formal data access regime in established data centers, frequently on an open access basis (as discussed further in the next Section), whereas the latter generally have no formal access rules governing their research data.

Another key distinction across scientific disciplines is between the observational and experimental sciences, where the types of data that need to be preserved and made broadly available differ significantly[29]. Typically, for observational data sets, it is the raw or minimally processed data that have the greatest value for reuse in research, whereas in the experimental sciences, it is the highly evaluated and verified data that are preserved and made available for broad use.

---

[25] Reichman & Uhlir, *op. cit.* note 13.
[26] National Research Council (1997), *op. cit.* note 1, p. 124-126.
[27] See generally, Reichman & Uhlir, *op. cit.* note 13, Part IV.
[28] Traditionally, "small science" research was done primarily in experimental laboratory sciences, such as chemistry and biology; in fieldwork studies such as ecology, anthropology, and various areas of social science; and in studies of human subjects, such as the biomedical and behavioural sciences. The autonomous nature of the research, and in many cases the privacy concerns associated with human studies, have precluded the sharing of data or the pooling of small data sets in centralized repositories. Here the research has been more competitive than cooperative and any exchanges of data were typically done on an informal, collegial basis, rather than through some formally structured data access regime. With the advent of higher capacity computing and digital networks, however, some of these research areas have organized "big science" research programs (e.g., the human genome project) and become much more data-intensive. They have established their own specialized data centers (e.g., genomic and protein data in molecular biology) or formed distributed data networks with nodes (e.g., ecological or biodiversity data). Ibid., p. 343-344 and 426-427.
[29] National Research Council (1995), *Preserving Scientific Data on Our Physical Universe*, National Academies Press, Washington, DC, p. 34-36.

Finally, as already noted for government-produced data, an important distinction must be made between data collected on human subjects and data on other, impersonal, subjects[30]. Research data on human subjects are restricted in various ways on ethical and legal grounds to protect personal privacy.

The bottom line in all of these categories of research and data types, however, is that open access to publicly funded research data should be the default rule and operating presumption, rather than the exception, and the exceptions to openness should be based on explicit, well-justified grounds.

6 EMERGING OPEN ACCESS MODELS

The presumption of openness and the implementation of an open access policy as the default rule in publicly funded research is certainly not a revolutionary concept. Not only are there solid justifications for such a policy as outlined above, but there are innumerable examples of successful implementations of this policy in practice in both government and government-funded institutions, in many fields of research, and in many countries. In this section we characterize these examples broadly and provide a number of specific references. Box 2 identifies a range of distributed, open, collaborative research and information production and dissemination activities using digital networks[31], while Box 3 provides details about one compelling example, identified in Box 2, of open access to academic materials at a world-class university.

Box 2:

There are many new kinds of distributed, open collaborative research and information production and dissemination on digital networks. Examples of open data and information production activities include:

- Open-source software movement (e.g., Linux and 10Ks of other programs worldwide, many of which originated in academia and are developed for research purposes);
- Distributed Grid computing or *e*-science (e.g., SETI@Home, LHC@home);
- Community-based open peer review (e.g., Journal of Atmospheric Chemistry and Physics); and
- Collaborative research Web sites and portals (e.g., NASA Clickworkers, Wikipedia, Curriki).

The following are examples of open data and information dissemination and permanent retention:

- Open data centers and archives (e.g., GenBank, the Protein Data Bank, The SNP Consortium, Digital Sky Survey);
- Federated open data networks (e.g., World Data Centers, Global Biodiversity Information Facility; NASA Distributed Active Archive Centers);
- Virtual observatories (e.g., the International Virtual Observatory for astronomy, Digital Earth);
- Open access journals (e.g., BioMed Central, Public Library of Science, + > 2500 scholarly journals);
- Open institutional repositories for that institution's scholarly works (e.g., the Indian Institute for Science, plus hundreds globally);
- Open institutional repositories for publications in a specific subject area (e.g., PubMedCentral, the physics arXiv);
- Free university curricula online (e.g., the MIT OpenCourseWare); and
- Emerging discipline-based commons (e.g., the Conservation Commons, the Geoscience Information Commons).

[30] Organisation for Economic Co-operation and Development (1980), "OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data," Paris.

[31] Uhlir, Paul F. (2006), "The emerging role of open repositories for the scientific literature as a fundamental component of the public research infrastructure", in *Open Access: Open Problems*, G. Sica, ed., Polimetrica, Monza, Italy.

Box 3:

The OpenCourseWare initiative at the Massachusetts Institute of Technology

The digital revolution is transforming information economics in a radical way. In the public science system one of the interesting trends is the development of additional user bases for 'secondary' use of data, information, and knowledge. When openly available, publicly funded digital resources can have many new useful 'lives' in addition to their primary uses. Use of the internet has minimised distribution costs. Open access is a way of cutting transaction costs. Low access barriers serve the original purposes of the public investment and increase the return on the investment: a broader scientific workforce can be put to work to get additional results without investments in additional resources.

Low access barriers make it possible to meet an important demand that cannot be served through traditional markets. For example, in 1999 the Massachusetts Institute of Technology (MIT) investigated a business model for selling its curriculum materials online. When it appeared that there would be an insufficient market for this service, MIT did not abandon the idea, but changed the original business model into one of open access: the "OpenCourseWare" initiative. The university now offers free access to well over one thousand courses and has gotten hundreds of million hits on its portal from educators, students, and self-learners from all over the world. Of course, the project initially was greeted with a great deal of apprehension among the MIT faculty, but eventually this bold vision was accepted. As expressed by President Emeritus of MIT Charles M. Vest: "*OpenCourseWare looks counterintuitive in a market-driven world. But it really is consistent with what I believe is the best about MIT. It is innovative. It expresses our belief in the way education can be advanced – by constantly widening access to information and by inspiring others to participate*."

Together, these various open access activities constitute an emerging globally networked "commons" for public science, representing a broad range of information types, institutional structures, disciplines, and countries. A common policy aspect of all these activities is their provision of free and open access online, with either reduced retention of intellectual property rights through permissive licensing mechanisms[32] or, much less frequently, a statutory public domain status[33].

In the area of data from publicly funded research, there already are many open access activities throughout the world, although no comprehensive compendium currently exists. As indicated in Box 2 there are at least two major types of institutional models specific to data: (1) open data centers or archives, and (2) federated[34] open data networks. The

[32] For a selection of such permissive licensing templates, which use statutory intellectual property protection, but with only "some rights reserved" instead of all the rights accorded under the statute, see the Creative Commons and its more recent Science Commons initiative at: http://www.creativecommons.org.

[33] The public domain status of factual data is a complex legal subject. Some countries expressly exclude government-generated information from copyright. Moreover, under traditional copyright law, factual compilations that lacked creativity or originality in their selection or arrangement, like many of the databases that are the subject of discussion in this paper, were not copyrightable and all the data in those compilations were in the public domain. However, some jurisdictions had so-called "sweat-of-the-brow" common-law protections (e.g., the United Kingdom and certain states in the United States), while others adopted more formal statutory protection of non-copyrightable compilations (e.g., the Scandinavian Catalogue Rule). More recently, the European Union enacted exclusive property protection of databases and compilations of information (Directive 96/9 of the European Parliament and the Council of 11 March 1996 on the legal protection of databases, 1996 O.J. (L77)), which has been implemented in all E.U. member States and Affiliated States, as well as in some other countries. This protection in most countries applies even to government and government-funded databases. In most countries there are very limited exceptions for public-interest uses of data (e.g., for public scientific research or education), and in some jurisdictions (e.g., France, Italy, Greece) there are no exceptions at all. For a comprehensive description and analysis of the E.U. Database Directive and its potential long-term effects of public research, see Reichman & Uhlir, *op. cit.*, notes 8 and 13.

[34] This type of management structure for distributed scientific data archives and data centers was first described in National Research Council (1995), *op. cit.*, note 30, p. 51-53. This model was based on a "flat" corporate management model described in Handy, Charles (1992), "Balancing Corporate Power: A New Federalist Paper,"

former is a centralized model whereas the latter has a connected set of distributed nodes. There are numerous examples of each type of open access data model operated either directly by government agencies or by government-funded entities (universities and not-for-profit research institutes).

Despite the successful adoption of open data access policies and practices in many areas of public research, the application of such regimes remains fragmented and inconsistent—a patchwork of uncoordinated and largely disparate activities, many of which are ad hoc, bottom-up endeavours. In too many cases, establishing satisfactory arrangements for data access seems to go beyond the means and imagination available at the working level. If finding adequate solutions without outside help is too much trouble, the researchers involved may easily succumb to passive risk avoidance. In view of the potential benefits that can be derived from increasing and improving access to such resources, establishing a more transparent and predictable environment that is coordinated at the national and international levels is desirable.

Some science policy leaders have begun to address these exigencies at the national level. For example, China established the Scientific Data Sharing Program in 2002[35]. Canada launched a National Consultation on Access to Scientific Research Data in 2004[36] and, that same year, the Research Council of Norway released a white paper documenting the important role of databases as a research infrastructure component[37]. In 2005, the U.S. National Science Board called for an initiative to develop a national policy framework for long-lived data collections[38], which was followed up by the establishment of an Interagency Working Group on Digital Data in the White House Office of Science and Technology Policy[39]. Most research funding agencies in the United States also have developed data policy guidelines for their grantees that encourage data sharing or deposits in established community data repositories, within specific discipline or research program contexts. However, the existing institutional policies still remain ad hoc and sub-optimally coordinated at the national level in the United States, as in most other countries.

At the international level, initiatives such as the Budapest Open Access Initiative, the Bethesda Declaration, and the Berlin Declaration [40], although focused more on open access to the scholarly journal literature than to the data, have helped to pave the way for further national policies. The new "Guidelines for Access to Research Data from Public Funding" from the OECD, endorsed by the governments of OECD countries (as discussed in the final Section of this paper and in another article in this volume), may be expected to play an important catalytic role.

While these incipient institutional models and policy approaches are commendable indicators that the scientific community is awakening to the opportunities and challenges of comprehensively rationalized data access regimes in

---

Harvard Business Review, Vol. 70, No. 6, p. 59-72. The key elements of a federated management model are: subsidiarity (the power is assumed to lie within the subordinate units of the organization), pluralism (interdependence of members), standardization of key elements to facilitate cooperation and interoperability, a separation of powers (responsibilities), and strong leadership from a small central directorate that is effective but not overbearing.

[35] CHENG Jinpei (2006), "Development of China's Scientific Data Sharing Policy," in Paul F. Uhlir and Julie M. Esanu, eds., *Strategies for Preservation of and Open Access to Scientific Data in China*, National Academies Press, Washington, DC. Also discussed in the article by XU Guan-hua in this special issue of the CODATA *Data Science Journal*.

[36] Strong, David F., and Peter B. Leach (January 31, 2005), *National Consultation on Access to Scientific Research Data*, National Research Council Canada, 82 p. Also discussed in the article by Sabourin and Dumouchel in this special issue of the CODATA *Data Science Journal*.

[37] The Research Council of Norway (2004), *The Need for Scientific Equipment, Databases, collections of Scientific Material, and Other Infrastructure*, report submitted as input to the 2005 White Paper on Research, Oslo (Abridged English version).

[38] National Science Board (2005), *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century,"* National Science Foundation, 64 p.

[39] Butler, Declan, "Agencies join forces to share data," *Nature* 446:354 (22 March 2007).

[40] The 2002 Budapest Open Access Initiative is available at: http://www.soros.org/openaccess/read.shtml/; the 2003 Bethesda Statement on Open Access Publishing is available at: http://www.earlham.edu/~peters/fos/bethesda.htm/; and the 2003 Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities is available at: http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html/.

public science, a great deal more can and should be done. And although the patchwork quilt of bottom-up data access regimes has served some research communities well in some cases, this loosely decentralized aggregation of approaches could achieve much greater results from a concerted national and international policy and funding focus.

7 TOWARD OPEN DATA REGIMES: GUIDING PRINCIPLES AND FLEXIBLE CONTRACTUAL TEMPLATES

The foregoing discussion has sought to develop a rationale for more formalized data access policies and procedures in public research, based on a core default principle of openness. The benign neglect of research data and databases thus far has not been regarded as a significant policy blunder. The most pressing database requirements seem to have been met through the ad hoc resourcefulness and volunteerism of dedicated individuals in public science[41]. But the brief history of the digital age already is replete with major losses of data and missed opportunities[42] that are certain to multiply in the absence of sustained focus and action.

As discussed in Section 5, it also is important to recognize that public policies in the developed and developing countries alike are shaped by legitimate considerations and interests that do not leave all scientific information and data in the public domain or under pure open access conditions. Instead, they impose limitations upon openness and cooperation in the conduct of public research and the utilization of its findings, in varying degrees and for a variety purposes Consequently, there is a need for public policies and institutional arrangements to seek a judicious balance between positive and negative effects upon the conduct of publicly funded research that are likely to ensue from the granting and enforcing of private ownership rights in scientific and technical data and information Yet, in recent decades the policy balance in this regard has been disrupted in ways that some science policy analysts perceive as threatening the long-term vitality of fundamental scientific research[43].

A successful data access regime must involve a comprehensive framework of policies and procedures that are based on a complete set of supporting principles and guidelines. Areas that require attention in developing principles and subsequent access regimes include organizational and management, financial and economic, legal, socio-cultural, and technical considerations[44]. The costs of inaction in the current state of affairs continue to accumulate, while the opportunities provided by the emerging cyberinfrastructure and new science initiatives will remain suboptimal.

Because of the diverse role of data in different fields of research, and the diverse and sometimes competing interests of the different stakeholders in the research enterprise, the formal data regimes need to be tailored to specific circumstances, but managed for the greatest return on the public investments. These conditions make it essential for most policy directives from the top at the national and international levels to be flexible and not rigidly prescriptive, while providing sufficiently strong and comprehensive guidance to the entities at the working level to implement effective regimes that are responsive to their particular interests.

In this final section we examine some mechanisms that can improve top-down guidance on the one hand, and bottom-up flexibility on the other. The former are the high-level international principles that can help guide the development of specific data access regimes at the (inter)national level. The latter involve the practical implementation through the development and voluntary adoption of new licensing templates that rights holders can select as standard options to provide access and use on less restrictive terms and conditions. We conclude with a brief overview of a major new initiative that seeks to integrate more effectively the top down and bottom up approaches.

7.1 Guiding principles

[41] Maurer, Stephen M., Richard B. Firestone and Charles R. Scriver, "Science's neglected legacy", Nature, Vol. 405, 11 May 2000.
[42] See, e.g., National Research Council (1997), *op. cit.* note 1, at p. 121-124.
[43] See, e.g., National Research Council (2003), *op. cit.*, note 6.
[44] Arzberger, *et al*., "Science and Government: An International Framework to Promote Access to Data, *Science* 303:1777-1778.

A good starting point for regulation at the more general level is the development of international principles, based on consensus by the national participants, which can help provide guidance to the governments, the public agencies, institutions, and individual researchers engaged in publicly funded research worldwide.[45] Coherent, consensus-based international principles, building on the experience of established successful models, should provide a number of benefits. They indicate the collective importance placed by science leaders in the national governments to the public research data issues. They can articulate a rationale and responsibility for improving the management and funding of the public data resources. They can provide guidance for the development of new access regimes based on a common set of values and objectives. And they can help establish an international level playing field for research and industry. The end result may be expected to lead to a higher return on public investments in research and substantial increases in productivity and cost-effectiveness.

The development of overarching international principles that cover publicly-funded research data in many countries can only be restricted to the essentials, of course. In the many different countries, disciplines, and institutes complete compliance with the principal rules will be difficult, and there will always be exceptions to the rules. Context-dependent solutions will have to be found, but all of these exceptions cannot and should not be part of the principles. The perspective can only be that of stating the *default* rules, including the core openness principle. Applying the principles and working out the specific details will be the responsibility of the stakeholders identified in Section 3 above—the national governments, public research funding agencies, and universities and public research institutes—in collaboration with the research community, as represented by the learned societies and the private sector. The principles therefore should offer the general international guidance for further regulation by the parties more directly involved.

The principles should not conflict with national legislation, nor harm other national, institutional, or individual interests. Strong, simple principles should be distilled from a much more extensive body of input and from a broad consultative process.

At the level of international science policy, principles represent the broadest common denominator of existing policies and (best) practices. But from this common ground they should guide emerging processes of change. International principles ultimately may look like abstract noncommittal generalities, but they can empower those who have to find the practical solutions with the right guidance for implementation.

Finally, international principles should be part of a common policy strategy to seize the new opportunities to increase the return on public investment in research and enhance the productivity and quality of research. The high-level principles should have primacy—they are the *Why* in the process. The principles then need to be implemented in a sensible access regime by the research organisations – the *How* in the process.

7.2 Contractual templates for the flexible implementation of the openness principle

To implement the general guiding principles, one way to deal with the potential imbalance in the statutory intellectual property system is to seek to amend the aspects that affect public research most negatively. However, this is not easily done, especially in view of the fact that many of these laws are quite recent and largely have ignored such considerations as they were debated and enacted.

[45] One example of this type of consensus-building international process is the OECD Ministerial *Declaration on Access to Research Data from Public Funding* of 30th January 2004 and the 2007 OECD *Guidelines* that followed it, as described in the companion article in this special issue of the CODATA *Data Science Journal* by Pilat and Fukasaku. The Declaration was inspired by the successful examples of data sharing on the (inter)national and institutional levels. The science ministers agreed that OECD guidelines would contribute to reach common science policy goals by improving the quality and productivity of scientific research and increasing the cost effectiveness of public investment in scientific research. The essence of the Declaration lies in the Principles that systematically treat the main points of the data access issues that have been worked out in subsequent *Guidelines*.

There is, however, another and rather different approach whose practical aspects merit wide attention and support to its further development. The proposed approach consists of the voluntary use of the rights held by intellectual property owners, which allow them to construct by means of licensing contracts conditions of "common-use" that emulate the key features of the public domain that are most beneficial for collaborative research in all its forms. The intention is to promote the cooperative use of scientific data, information, materials and research tools that actually are not in the public domain, and whose licensed use is therefore legally protected by an intellectual property regime. Such an undertaking may be properly described as creating "global information commons for science", inasmuch as a "common" constitutes a collectively held and managed bundle of resources to which access by cooperating parties is rendered open (though perhaps limited in its extent or use) under minimal transactions cost conditions.

The economic logic and practical feasibility of the "contractually constructed commons" approach can be derived from non-market mechanisms constructed as systems of customary rights and restraints. Historically, it was deliberate acts of private enclosure rather than some imagined tragedy of over-grazing that often spelled the end of the agrarian commons. The legal system today makes it possible for the owners of a tangible resource held in common to protect their collective use-rights, and manage their contractually constructed common-pool so as to sustain and augment the benefits that it yields. Consequently, because information cannot be depleted by overuse, individuals having private ownership rights in intellectual property may voluntarily use contracts to construct a common use-rights area that is all inclusive, in granting access to those wishing to use the contents. Furthermore, and because the common in this case is owned and not part of the public domain, the benefits that all users can enjoy from such an arrangement may be preserved and enhanced. This can be accomplished by reserving the legal right to exclude certain usage practices that might otherwise undermine the willingness of others to similarly pool the information that they have created.

The respective rights of the participants in the public research system can be most effectively mediated through the use of contracts at the individual researcher and institutional levels. Common-use licensing approaches that promote broad access and reuse rather than restrict it, such as those being developed by the new Science Commons under the Creative Commons mentioned in Section 6, above, can preserve essential ownership rights while improving the social benefits and returns on the public investments in research[46]. They can help to achieve a productive balance between the domains of proprietary R&D and publicly- funded open science.

8 TOWARDS GLOBAL INFORMATION COMMONS FOR SCIENTIFIC DATA AND INFORMATION

The rationalization of policies and practices across nations, institutions, and disciplines may be expected to result in much greater social and economic impact from the investment in public research overall by enabling greater access to and use of scientific data and information resources, and by facilitating interdisciplinary and international cooperation in public science and education. Because of the international scope of digital networks and research collaborations, strategic international approaches for building information commons are both necessary and desirable. In short, the adoption in recent years of the many innovative and promising open initiatives and common-use licensing approaches from the bottom up, coupled with the introduction of some new top-down policy proposals at the international level (at the OECD) and at the national level in several countries, make this an appropriate time to integrate these efforts.

It is for all the reasons established in this article that several international science policy organizations—CODATA, ICSU, and Science Commons—are joining efforts to launch the Global Information Commons for Science Initiative. This Initiative[47] has the overall goal to accelerate the development and scaling up of open scientific data and

[46] See the companion article by Onsrud and Campbell in this special issue of the CODATA *Data Science Journal*.

[47] The original ideas for the Global Information Commons for Science Initiative were presented in a series of reports published at the U.S. National Academies, in a seminal article by Reichman & Uhlir, *op. cit*, note 13, and in David, P. A. and M. Spence (2003), "Toward Institutional Infrastructures for e-Science: The Scope of the Challenges," A Report to the Joint Information Systems Committee of the Research Councils of Great Britain, Oxford Internet Institute Report No. 2. (September) [Available at: http://www.oii.ox.ac.uk/resources/publications/OIIRR_E-Science_0903.pdf]. These ideas were more fully fleshed out following an international workshop at UNESCO Headquarters in Paris on 1-2 September 2005 on the theme "Creating the Information Commons for Science: Toward Institutional Policies and Guidelines for Action" [details

information resources on a global basis, with particular focus on "common use" licensing approaches. The specific objectives are to:

(1) *Improve understanding and increase awareness of the societal and economic benefits of easy access to and use of scientific data and information, especially focusing on those resulting from governmental or publicly funded research activities;*

(2) *Promote the broad adoption of successful institutional and legal models for providing open availability on a sustainable basis and facilitating reuse of data and information;*

(3) *Help coordinate the efforts of the many stakeholders in the world's diverse research community who are engaged in devising and implementing effective approaches to attaining these objectives, with particular attention to the circumstances of the developing as well as developed countries.*

(4) *Develop an online "open knowledge environment" to promote all of the objectives of the Initiative, including providing an online collaboratory for work with different research communities to define, test, analyze, and create new knowledge about the information commons paradigm.*

In our view, such an Initiative can help devise and promote new normative and legal structures for the exchange of data and information that are expected to be especially well-suited for the future conduct of collaborative research in many domains of science. By rationalizing the policy and management systems in publicly funded research, the value of global digital networks and related technological advances to the progress of science can be fully realized.

---

of the Workshop rationale and proceedings, are available at: http://www.codataweb.org/UNESCOmtg/index.html.]. That event was organized by CODATA with the joint sponsorship of ICSU, ICSTI, INASP, UNESCO, and TWAS, and with the collaboration of the OECD.