For Official Use **DSTI/STP(2005)10**

Organisation de Coopération et de Développement Economiques
Organisation for Economic Co-operation and Development **05-Oct-2005**

English - Or. English

DIRECTORATE FOR SCIENCE, TECHNOLOGY AND INDUSTRY
COMMITTEE FOR SCIENTIFIC AND TECHNOLOGICAL POLICY

DSTI/STP(2005)10
For Official Use

MAXIMIZING THE VALUE OF PUBLIC SCIENTIFIC DATA FOR GLOBAL SCIENCE: ISSUES AND PRACTICES

24-25 October 2005, OECD, Salle des Nations, La Défense

This document, authored by Paul F. Uhlir and Peter Schröder, is presented as a background paper on issues and practices for the policy discussion on Access to research data. The paper will be presented as a lead presentation by the United States under item 5a) of the draft agenda.

Contact persons: Paul F. Uhlir (puhlir@nas.edu) ; Peter Schröder (p.schroder@minocw.nl) ; Yukiko Fukasaku, Tel: (33 1) 45 24 18 69, E-mail: yukiko.fukasaku@oecd.org

JT00190933

English - Or. English

I. INTRODUCTION

1. The digital revolution has transformed the accumulation of properly curated public research data into an essential resource whose value increases with use.[1] Their potential contributions to the creation of new knowledge and downstream economic and social goods is 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 spent on maximizing 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.

2. We are thus at a critical juncture. On the one hand, we are overwhelmed by a hidden avalanche of ephemeral bits that are central components of modern research and of the emerging "cyber-infrastructure"[2] for *e*-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 general lack of attention by the research policy and funding entities in many cases is perpetuating the systemic inefficiencies, and the loss or underutilization of existing data resources derived from public investments. Despite the rapidly growing capabilities of information and communication technologies (ICTs) to make much more effective use of those data, the ability to access and use the data remains suboptimal. 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. International Guidelines for Access to Research Data from Public Funding, as proposed by the Organisation for Economic Co-operation and Development (OECD), would be one important step toward achieving this goal.

---

1 See National Research Council (1995), *Bits of Power: Issues in Global Access to Scientific Data*, National Academy Press, Washington, DC.

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.

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/.

Note: The views expressed in this paper are those of the authors and not necessarily those of their institutions of employment.

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 global science system. We focus on the rationale and requirements for developing overarching principles and for establishing 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.

II. THE GROWING ROLE OF DATA IN THE RESEARCH PROCESS

3. The evolution of scientific research over the years 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 has risen 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].

4. Not surprisingly, these trends also have elicited an increasing 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 cyber-infrastructure, 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 mutliple disciplines, sectors, and countries.

5. Continuously growing quantities of data[6] about the universe around us are produced by government agencies, research institutes, and industry as a fundamental component of scientific research worldwide. Practically anything measurable can be described and stored in a digital database. A genomic sequence, the speed of subatomic particles, the orbit of the earth, the temperature of a liquid, 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:

6. 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

4 OECD Main Science and Technology Indicators 2005/I, Paris.

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

6 *"Scientific data"* may be defined as *"the numerical quantities or other factual attributes generated by scientists and derived during the research process (through observations, experiments, calculations and analysis)"*. CODATA Working Group on Archiving Scientific Data,, available at www.nrf.ac.za/codata.

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[7].

7. 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 inextricably embedded in an all-embracing research process. For researchers the distinction between data input, analysis, and findings output only made sense at the theoretical level. With the advent of digital technologies and networks, however, the various parts of the research trajectory have been loosened into separate specialised activities (as for example data collection, technical, and methodological 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. Different activities may have differing consequences for the applicable regulations and legislation. This diagram shows the main ingredients of the research and data trajectories.

1. The Research Trajectory

Data ⇨ Information ⇨ Knowledge

hypothesis (technical support ---------- research proper) conclusion
(creative,inventive) (routine ---------- creative, inventive)

↶ (feedbacks) ↻ ↶ ↻ ↶ ↻
↳ (methodology/design) ⇉ data collection/measuring ⇉ analysis/synthesis ⇉ results/publications/patents

2. The Data Trajectory

data collection ⇨ primary data ⇨ processing ⇨ documenting ⇨ final data ⇨ archiving ⇨ dissemination

↳ ***Data sharing options***

8. 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 elements of the research process 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[8]. By

7 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.

8 Rita Colwell (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.

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 these developments could take place in the developing world.

9. These major changes in the structure and conduct of data-driven research using the cyber-infrastructure 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.

III. THE EMERGING ROLES OF STAKEHOLDERS IN DATA ACCESS REGIMES

10. 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 cyber-infrastructure. These stakeholder groups all affect the development of new institutional data management and policy models, as discussed in Section VI below.

11. ***Governments*** have responsibilities 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). They have an interest in promoting accountability for the cost effectiveness and management of their public investments in research. Governmental policies are 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.

12. ***Research funding agencies*** are accountable for the support and performance of the national science system. They 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 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 establishment of specialized data centres 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 support.

13. ***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 one or more specialized data centres—and a large number of individual databases—which are funded in whole or in part with public money. Whether or not they do have a data centre, they have a responsibility for establishing policies for the access to and use of their expanding amounts and types or research data, consistent with the requirements and interests of their funding sources, researchers, and other institutional stakeholders, as well as the broader research community in which these institutions operate. The frequently conflicting interests of multiple stakeholders make the establishment of data access policies at the institutional level both crucial and difficult.

14. ***Learned societies*** 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. The societies promote their views within their own communities through major conferences and their journal publications, and externally through interactions with policy makers and research leaders.

15. ***International scientific organizations*** have a role similar to the learned societies, but at regional or global levels. By international scientific organizations we mean both intergovernmental organizations (IGOs) and international non-governmental organizations (NGOs). Among the IGOs relevant in this context are the Organisation for Economic Co-operation and Development (OECD), the Commission of the European Communities (CEC), and some of the specialized agencies of the United Nations, such as the United Nations Educational, Scientific, and Cultural Organization. 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.

16. ***Industry research institutions*** benefit from greater access to scientific data produced by others, although they tend to keep their own data outputs proprietary. They increasingly outsource research to universities, partnering with university researchers 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 further complicate the management of the resulting data and the optimal allocation of rights to those data, requiring express agreements in each instance.

17. ***Individual researchers*** generate increasing amounts and types of data both as individuals and as participants in various kinds of formal and informal collaborations. As the main producers and users of public scientific data, they have the greatest stake in the development of rational data access regimes and in the adequate funding and management of data collections and centres. 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 in creating new models of data access regimes from the bottom up. Many researchers also have become part-time or specialised data managers, either on an ad hoc, informal basis, or in a more formally structured context.

18. ***The general public*** typically does not become involved in the policy and management issues pertaining to national R&D, generally, or to data from publicly funded research, specifically. Nevertheless, as the source of taxpayer investments in public research and related data activities, there is a strong public interest in seeing that the fruits of those investments are effectively managed and used. 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 V.

19. Each of these major stakeholder groups in the research enterprise has a major and growing interest in the development of optimal policies for access to and use of publicly funded research data. Although 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.

IV. THE HIDDEN COSTS OF CLOSED DATA SYSTEMS

20. As described in Box 1, research data are emerging in the research system as autonomous resources, the uses of which are no longer inextricably linked to their original producers or purposes. Many types of research data can be used beyond the original producers and users in unlimited ways at different times and places by an unlimited number of researchers. 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 management policy and practice.

21. 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 is typically reinforced by their public funding sources.

22. There are, however, a number of negative implications[9] to the efficiency and effectiveness of the research system from unnecessarily balkanized and closed access regimes in light of the (quasi) public good[10] nature of such digital data resources. This is particularly true for data from fundamental public research, of course, rather than for research data with significant commercial potential.

23. ***Lost opportunity costs.*** Most obviously, there is much less data-intensive research possible if the publicly-funded data are not shared or made easily available. This results in significant lost opportunity costs that are certain to occur but are difficult to measure[11]. 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 that 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 (assuming those data have some broader utility).

24. ***Barriers to innovation.*** The production of copyrightable or patentable downstream 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

---

9 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.

10 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). 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 Michael Callon (1994), "Is Science a Public Good?", in Science, Technology and Human Values, Vol. 19, p. 395.

11 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. *Bits of Power*, op. cit., note 1, at p. 121-124.

leads to deadweight social costs, taxing the innovation system in each country and slowing scientific progress[12].

25. ***Less effective cooperation.*** 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. Moreover, 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[13].

26. ***Higher transaction costs.*** Databases with a monopoly status that are maintained on a closed proprietary basis will tend to result in higher, anti-competitive pricing[14]. Moreover, managing publicly funded databases on a restrictive, proprietary basis adds substantial administrative overhead on both ends to make each transaction, further taxing the public research system.

27. ***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.

28. Overall, unnecessary access barriers to publicly funded research data 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 technological capabilities offer ever greater opportunities to increase that return.

---

12 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.

13 National Research Council (1999), *A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases*, National Academy Press, Washington, DC, at p. 19-20.

14 Weiss, Peter (2003), "Conflicting International Public Sector Information Policies and Their Effects on the Public Domain and the Economy", in *The Role of S&T Data and Information in the Public Domain*, op. cit. note 7, p. 129-132, and Reichman & Uhlir, op. cit., note 10.

V. THE SCIENTIFIC AND SOCIAL BENEFITS OF GREATER OPENNESS

29. 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 true whether the data are produced within government or by entities funded by government sources, although some important distinctions apply, as outlined below.

30. 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[15]. This definition is also consistent with the "full and open" data policy used in various international environmental projects and in environmental research in the United States over the past two decades[16].

31. 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.
- 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 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.

---

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

16 *Bits of Power*, 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[17].

32. Open access to factual data plays a vital role in all these areas. 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 only touch on them briefly below.

A. Policy considerations for data produced by government entities

33. The data and databases generated directly through government research have the following additional policy considerations favouring their open availability and unrestricted reuse[18]:

34. ***Legal considerations.*** 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[19].

35. ***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, 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.

36. ***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.

37. ***Socio-economic considerations.*** Open access is the most appropriate way to disseminate public data and information online in order to maximize the value and return on the public investment in its production[20]. There are numerous economic and non-economic positive externalities—especially through network effects—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[21]. Conversely, the

---

17 Feinberg, S.E., Martin, M.E., and Straf, M.L., eds. (1985), *Sharing Research Data*, National Academy Press, Washington DC and *A Question of Balance*, op. cit. note 14, compiled by Arzberger, et al. (2004), "Promoting Access to Public Research Data for Science, Economic, and Social Development", *Data Science Journal*, CODATA, p. 135-152.

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

19 See, *e.g.*, Dalrymple, Dana (2003), "Scientific Knowledge as a Global Public Good: Contributions to Innovation and the Economy, in Julie M. Esanu and Paul F. Uhlir, eds., *The Role of Scientific and Technical Data and Information in the Public Domain*, National Academies Press, Washington, DC, p. 35-51.

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

21 Ibid. See also "Conflicting International Public Sector Information Policies", op. cit. note 15; European Union Green Paper (1998); and PIRA International (2000) [refs to be completed].

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 on other social, ethical, and good governance grounds.

38. 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. 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)[22]. 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.

B. Policy factors to consider in disseminating government-funded research data

39. Although access policies for research data produced by non-governmental entities with government funds have similar rationales as those outlined above for government-produced data, there typically are additional factors that need to be considered.

40. 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, data may be withheld even after publication. 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].

41. 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 maximize 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 data management and access.

42. 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, at least until a patent is filed. 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 the research data. In such cases, the application of an open access data policy will also likely be inappropriate.

---

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

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

24 [References to be added]

43. The issues raised in public-private relationships take many forms and contain some inherent tensions, such as openness vs. exclusivity, public goods vs. private investments, public domain vs. proprietary rights, and competition vs. monopoly, among others. This mix of motivations, priorities, and requirements is context-dependent, typically unique to the parties involved, and 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 maximizes the public interest goals associated with the public funding, while effectively protecting the proprietary private interests[25].

44. This bifurcated ordering of interests can take many forms. At the most basic level it is possible to provide for 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].

45. 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 centres, frequently on an open access basis (as discussed further in Section VI), 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].

---

25 *A Contractually Reconstructed Research Commons*, op. cit. note 13.

26 *Bits of Power, op. cit*. note 1, p. 124-126.

27 See *A Contractually Reconstructed Research Commons*, 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 centres (*e.g.*, genomic and protein data in molecular biology) or formed distributed data networks with nodes (*e.g.*, ecological or various biomedical sub-disciplines). 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.

46. Yet another 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.

47. 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.

30 [references to be added]

VI. STATE OF THE ART IN OPEN DATA ACCESS REGIMES

48. 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 provides one compelling example of open access to academic materials at a world-class university, while Box 3 identifies a range of distributed, open collaborative research and information production and dissemination activities using digital networks.

Box 2. The Open Course Ware initiative at the Massachussets 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 entry barriers will 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 entry 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 no market for this MIT did not abandon the idea, but changed the original business model into one of open access: the "OpenCourseWare" initiative. Today MIT offers free access to 1100 courses and as of April 2005 had gotten 556 million hits from educators, students and self-learners from all over the world. Of course, this project initially was greeted with a great deal of apprehension among the MIT faculty. Eventually, however, this bold MIT vision was accepted. As expressed by President Emeritus 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*."

49. 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).
- Distributed Grid computing (*e.g.*, SETI@Home, LHC@home).
- Community-based open peer review (*e.g.*, Journal of Atmospheric Chemistry and Physics).

- Collaborative research Web sites and portals (*e.g.*, NASA Clickworkers, Wikipedia, Project Gutenberg).

50. 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, space science data centers).
- 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, + > 1500 scholarly journals).
- Open institutional repositories for that institution's scholarly works (*e.g.*, the Indian Institute for Science, + > 400 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).
- Emerging discipline-based commons (*e.g.*, the conservation commons or geoscience commons).

51. Together, these various open access activities constitute an emerging global "*e*-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[31] or, much less frequently, a statutory public domain status[32].

52. In the area of data from publicly funded research, there already are a great many open access models throughout the world, although no comprehensive compendium currently exists. As indicated in

---

31 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.

32 The public domain status of factual data is a complex legal subject. Some countries expressly exclude government-generated information from copyright and authors' moral rights (*e.g.*, the United States, 17 U.S.C. 105); others exclude all (*e.g.*, Finland, ref.) or some (*e.g.*, E.U. Directive on Environmental Information, ref.) government-generated information from copyright, but not from authors' moral rights. 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 (U.K. and certain states in the U.S., refs), while others adopted more formal statutory protection of non-copyrightable compilations (*e.g.*, Scandinavian Catalogue Rule, refs.). More recently, the E.U. enacted exclusive property protection of databases and compilations of information (Directive on the legal protection of databases, ref.), 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 op. cit., note 11 and note 24.

Box 3 there are at least two major types of institutional models specific to data: (1) open data centres or archives, and (2) federated[33] open data networks. The 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).

53. 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 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.

54. Some science policy leaders have begun to address these issues at the national level. For example, Canada launched a National Consultation on Access to Scientific Research Data in 2004[34] and China established the Scientific Data Sharing Program in 2003[35]. Most recently, the U.S. National Science Board called for an initiative to develop a national policy framework for long-lived data collections[36] and the Research Council of Norwegian released a white paper documenting the important role of databases as a research infrastructure component[37]. A number of 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[38]. However, the existing institutional policies still remain largely uncoordinated at the national level and the new national policy initiatives are not coordinated internationally (with the exception of a new policy initiative at OECD, as discussed in the final Section of this paper).

55. While these incipient top-down approaches are commendable indicators that the science policy community is awakening to the opportunities and challenges of comprehensively rationalized data access regimes in public science, clearly 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

---

33 This type of management structure for distributed scientific data archives and data centers was first described in *Preserving Scientific Data on Our Physical Universe*, op. cit., note 28, 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," *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.

34 Strong, David F., and Peter B. Leach (January 31, 2005), *National Consultation on Access to Scientific Research Data*, National Research Council Canada, 82 p.

35 CHENG Jinpei (publication pending), "Development of China's Scientific Data Sharing Policy," in Paul F. Uhlir, ed., *Strategies for Preservation of and Open Access to Scientific Data in China*, National Academies Press, Washington, DC.

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

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 [examples to be provided]

loosely decentralized aggregation of approaches could achieve much greater results from a concerted national and international policy and funding focus.

VII. TOWARD OPEN DATA ACCESS REGIMES BASED ON GUIDING PRINCIPLES

56. 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[39]. But the brief history of the digital age already is replete with major losses[40] and missed opportunities[41] that are certain to multiply in the absence of sustained focus and action.

57. 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 policy considerations[42]. The costs of inaction in the current state of affairs continue to accumulate, while the opportunities provided by the emerging cyber-infrastructure and new science initiatives will remain suboptimal.

58. 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 while optimized for the greatest return on the pubic 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. Here we look only at the high-level international principles that can help guide the development of data access regimes, rather than at specific national laws and policies, or specific data access regimes themselves (descriptions of those may be found in the examples referenced in the previous Section).

59. A set of internationally developed principles, based on consensus by the national participants, can help provide guidance to the public agencies, institutions, and individual researchers engaged in publicly funded research worldwide.[43] Coherent, consensus-based international principles, building on the

---

39 Maurer, Stephen M., Richard B. Firestone and Charles R. Scriver, "Science's neglected legacy", Nature, Vol. 405, 11 May 2000.

40 Many valuable data sets have been lost entirely or partially degraded as a result of improper management. [add refs.]

41 See, *e.g.*, *Bits of Power*, op. cit. note 1, at p. 121-124.

42 Arzberger, *et al.*, "Science and Government: An International Framework to Promote Access to Data, *Science* 303:1777-1778.

43 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. 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

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.

60. The development of international principles that cover all research data in many countries can only be restricted to the essentials, of course. In all the different countries, disciplines, and institutes complete compliance with the principal rules often will be difficult or even impossible. There will easily be more exceptions than there are rules. For all these complications, 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 III 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.

61. 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 from a broad consultation process.

62. 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 should empower those who have to find the practical solutions with the right guidance for implementation.

63. 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.

---

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 to be worked out in subsequent *Guidelines*.