# TOWARDS INTERNATIONAL GUIDELINES FOR ACCESS TO RESEARCH DATA FROM PUBLIC FUNDING

# YOUR OPINION PLEASE !

The Committee for Scientific and Technological Policy (CSTP) of the Organisation for Economic Co-Operation and Development (OECD) is in the process of drafting International Guidelines for Access to Research Data from Public Funding. The point of departure is the OECD Ministerial Declaration from 30 January 2004. You will find the *Declaration* in this discussion paper accompanied by an *Explanatory Note* and a *Background document*. The paper should make clear the goal and the scope of the upcoming OECD guidelines: general rules applicable for all science all over the world, so necessarily not going into detail, stated in very general terms. A firm anchorage for research practice all the same. The value of the guidelines should lie in their usefulness as a basis for the more detailed, more concrete, practical guidelines that are required in the more specific disciplinary, institutional and national contexts.

The Declaration includes a set of *Principles* on which the upcoming *Guidelines* could be based. To assist CSTP in the drafting process, your expert opinion on these Principles would be most welcome. At this stage even your very general ("first sight"), short comments on the Principles can be valuable input for the drafting process. I hope you can find the time, during the conference or within the next month, to answer the following simple questions in a simple way:

1. Could these principles be the basis for international guidelines?
2. Will the implementation of the *Draft Principles* in your institution be feasible?
3. Do you have additional suggestions to be included in the Principles and/or the additional text?
4. Do you have any other general comments?

Could you please send your answers with your name, the name of your organisation and it's address by email to me?

Peter Schröder ; p.schroder@ minocw.nl

Thank you very much.!

# TOWARDS INTERNATIONAL GUIDELINES FOR ACCESS TO RESEARCH DATA FROM PUBLIC FUNDING:

*THE MODEL PROPOSED IN THE DECLARATION ENDORSED BY THE MINISTERS FROM OECD COUNTRIES RESPONSIBLE FOR SCIENTIFIC AND TECHNOLOGICAL POLICY.*

Contents:




Contact: Peter Schröder, Ministry of Education, Culture and Science, The Netherlands, p.schroder@minocw.nl

# MINISTERIAL DECLARATION ON ACCESS TO RESEARCH DATA FROM PUBLIC FUNDING

**adopted on 30 January 2004 in Paris**

**The governments[1] of Australia, Austria, Belgium, Canada, China, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Russian Federation, the Slovak Republic, the Republic of South Africa, Spain, Sweden, Switzerland, Turkey, the United Kingdom, and the United States**

Recognising that an optimum international exchange of data, information and knowledge contributes decisively to the advancement of scientific research and innovation;

Recognising that open access to, and unrestricted use of, data promotes scientific progress and facilitates the training of researchers;

Recognising that open access will maximise the value derived from public investments in data collection efforts;

Recognising that the substantial increase in computing capacity enables vast quantities of digital research data from public funding to be put to use for multiple research purposes by many research institutes of the global science system, thereby substantially increasing the scope and scale of research;

Recognising the substantial benefits that science, the economy and society at large could gain from the opportunities that expanded use of digital data resources have to offer, and recognising the risk that undue restrictions on access to and use of research data from public funding could diminish the quality and efficiency of scientific research and innovation;

Recognising that optimum availability of research data from public funding for developing countries will enhance their participation in the global science system, thereby contributing to their social and economic development;

Recognising that the disclosure of research data from public funding may be constrained by domestic law on national security, the protection of privacy of citizens and the protection of intellectual property rights and trade secrets that may require additional safeguards;

Recognising that on some of the aspects of the accessibility of research data from public funding, additional measures have been taken or will be introduced in OECD countries and that disparities in national regulations could hamper the optimum use of publicly funded data on the national and international scales;

[1] Including the European Community.


Contact: Peter Schröder, Ministry of Education, Culture and Science, The Netherlands, p.schroder@minocw.nl

Considering the beneficial impact of the establishment of OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (1980, 1985 and 1998) and the OECD Guidelines for the Security of Information Systems and Networks (1992, 1997 and 2002) on international policies for access to digital data;

***DECLARE THEIR COMMITMENT TO:***

Work towards the establishment of access regimes for digital research data from public funding in accordance with the following objectives and principles:

**Openness:** balancing the interests of open access to data to increase the quality and efficiency of research and innovation with the need for restriction of access in some instances to protect social, scientific and economic interests.

**Transparency:** making information on data-producing organisations, documentation on the data they produce and specifications of conditions attached to the use of these data, available and accessible internationally.

**Legal conformity:** paying due attention, in the design of access regimes for digital research data, to national legal requirements concerning national security, privacy and trade secrets.

**Formal responsibility:** promoting explicit, formal institutional rules on the responsibilities of the various parties involved in data-related activities pertaining to authorship, producer credits, ownership, usage restrictions, financial arrangements, ethical rules, licensing terms, and liability.

**Professionalism:** building institutional rules for the management of digital research data based on the relevant professional standards and values embodied in the codes of conduct of the scientific communities involved.

**Protection of intellectual property:** describing ways to obtain open access under the different legal regimes of copyright or other intellectual property law applicable to databases as well as trade secrets.

**Interoperability:** paying due attention to the relevant international standard requirements for use in multiple ways, in co-operation with other international organisations.

**Quality and security:** describing good practices for methods, techniques and instruments employed in the collection, dissemination and accessible archiving of data to enable quality control by peer review and other means of safeguarding authenticity, originality, integrity, security and establishing liability.

**Efficiency:** promoting further cost effectiveness within the global science system by describing good practices in data management and specialised support services.

**Accountability:** evaluating the performance of data access regimes to maximise the support for open access among the scientific community and society at large.

Seek transparency in regulations and policies related to information, computer and communications services affecting international flows of data for research, and reducing unnecessary barriers to the international exchange of these data;


Contact: Peter Schröder, Ministry of Education, Culture and Science, The Netherlands, p.schroder@minocw.nl

Take the necessary steps to strengthen existing instruments and – where appropriate – create within the framework of international and national law, new mechanisms and practices supporting international collaboration in access to digital research data;

Support OECD initiatives to promote the development and harmonisation of approaches by governments adhering to this Declaration aimed at maximising the accessibility of digital research data;

Consider the possible implications for other countries, including developing countries and economies in transition, when dealing with issues of access to digital research data.

***INVITE THE OECD:***

To develop a set of OECD guidelines based on commonly agreed principles to facilitate optimal cost-effective access to digital research data from public funding, to be endorsed by the OECD Council at a later stage.


Contact: Peter Schröder, Ministry of Education, Culture and Science, The Netherlands, p.schroder@minocw.nl

# MINISTERIAL DECLARATION ON ACCESS TO RESEARCH DATA FROM PUBLIC FUNDING

## Explanatory Note

### CONSIDERATIONS

#### Updating scientific openness

The requirement for openness in scientific research has a long tradition. In the contemporary environment of Information and Communication Technologies (ICT), the consequences of the basic assumption of openness will be more profound. ICT have enlarged the scale, scope and pace of research and have enhanced its openness and accessibility. ICT have opened up new resources and methods for research. The use of enormous volumes of digital data in the natural and the social sciences has brought impressive new results that will strengthen the knowledge base of economies and the evidence base of policy making. By using research data in digital form, research is becoming more “data-driven” everyday. Digital research data and the accompanying hardware and software facilities are rapidly becoming a central part of research infrastructures. Open access to digital research data from public funding will be a key factor in the further progress of science. To fully realise the potential that open access to digital research data has to offer requires an updating of research practices, data management and science policies in order to even out unnecessary entry barriers. Science systems that succeed in establishing responsible and efficient data access regimes will be the first to benefit from the new potential of digital research data. Science systems that act less energetically will eventually lag behind in quality and productivity.

Open access means open international access to the digital data resources of the global science system. Open international access to digital research data can be realised in various ways, depending on the various national research practices and policies. Principles and Guidelines from OECD will contribute greatly to the co-ordination of the various national approaches. Strengthening international scientific openness Principles and Guidelines from OECD will create a level playing field and enhance the quality and productivity of the global science system.

Recent applications of information and communication technologies offer substantial new opportunities to expand the use of digital research data. ICT enable multiple use of research data for extended purposes that will transcend the traditional boundaries of individual research projects, institutes, disciplines, and nations. Therefore science systems as well as industry worldwide should seek to benefit as much as possible from these developments. These are the main reasons for proposing the Declaration.

Open access to, and unrestricted use of, data outside their initial use promotes open scientific inquiry, diversity of analysis and opinion, scrutiny and testing of hypotheses and results, methods and techniques of analysis and facilitating teaching of research. Non-exclusive access to data on

equal terms increases the efficiency of research by avoiding unnecessary duplication of data collection and permitting the creation of new data sets by combining data from multiple sources.

Broadening the use of expensive publicly funded collections of digital research data will enhance their scientific and societal value. Further use of research data will strengthen the knowledge base of the economies, and the evidence base of policy making in OECD. As the durable data resources can be used in a non-rival manner, *access* to data instead of ownership of data should be seen as decisive in getting a higher return on public investment. Facilitating international open access to digital research data will be a way to promote:

- Good stewardship of public knowledge;
- Strong value chains of innovation;
- The creation of value from international co-operation.

Timely realisation of internationally agreed data access regimes will bring a substantial increase of scientific and societal value to the participant countries of the global science system. Delay in the realisation of international open access will slow down the advancement of science, diminish its potential in quality and productivity and reduce the level playing field of the scientific arena.

**Access regimes**

Seizing the new opportunities offered by open access to digital research data calls for changes in the way research and data are managed. There will be data that are of little interest after their initial use, but in many cases data lend themselves to extended use for different purposes. In some fields of science researchers are already quite successful in making the new digital promises come true, but other fields experience serious obstacles in realising the digital potential. Restrictions to access for reasons of national security or privacy protection will be inevitable and have to be respected, but in many cases current obstacles can be avoided. On balance, it will be necessary to direct science policy and research management at the establishment of *data access regimes* that will help to overcome unnecessary barriers to multiple uses of digital data.

**The role of governments**

Adequate data access regimes require distributed as well as central responsibilities. Research agencies and institutes, data archives and libraries, publishing and software firms as well as governments should all play an active part in the establishment of data access regimes for scientific research. The impact of data access regimes will be highly dependent on the professional regulatory skills of research agencies, institutes and communities. But in the end the regimes should be based on governmental core responsibilities for legislation, budget and international relations that should underpin the domain of knowledge as a public good.


Contact: Peter Schröder, Ministry of Education, Culture and Science, The Netherlands, p.schroder@minocw.nl

### Access regimes to handle data content

Accessibility is highly dependent on the available (network) *infrastructure*. Infrastructure is not extensively addressed here. Rather, the focus lies on conditions for access to data *content*. Current discussions point to technical, institutional, financial, legal and cultural obstacles to a broader use of digital research data. Effective data access regimes could contribute greatly to the solution of most of these problems. They could include data services and facilities, some additional formal regulation as well as informal practical incentives and should allow for fine-tuning in the differing contextual specifics of the research practice. Data access regimes will help to lessen the administrative burden of individual researchers and increase the productivity of research. The declaration suggests a set of general *principles* that can be used in shaping the more specific data access regimes.

## DEFINITIONS AND SCOPE

### Definitions

*Research data*

In the context of this document data are defined as the factual records (numerical scores, textual records, images and sounds) used as sources, base material for scientific research. A scientific data set constitutes a systematic representation of the universe being researched. Direct, often automatic, collection produces “raw” digital data. To be usable for research, “raw” data are checked, cleaned, documented and further prepared into “final data”.

Anything imaginable can be used as data for scientific research. This document limits its attention as much as possible to source data as distinct from bibliographical data. There are valuable irreplaceable data of unique observations, data generated by instruments that no longer exist and data from experiments that can easily be reproduced. Data from experiments in some fields are often used only once, re-analysis of other data sets can lead to valuable new results. In social sciences and meteorology for example, some data from single events may be of limited use for future research, but data from time series of comparable surveys will often be of timeless value. The National Institutes of Health define final research data as “the recorded factual material commonly accepted in the scientific community as necessary to validate research findings”.

*Open access*

Open access means easy, timely, user-friendly, Web-based, access on equal terms to the international research community as well as industry, at the lowest possible cost, on a non-profit basis, resulting in maximum use.

## Scope

Access to digital data from public funding is determined by the availability of technical (network) infrastructure on the one hand, and conditions on the use of the data content for research on the other. The access regimes discussed here are limited to the organisational, technical (software tools) and procedural aspects of using data 'content'.

Access regimes for digital data to be used for scientific research should cover:

- Access to existing data sets from publicly funded research (primary and secondary use).
- Access to existing data sets from other public data collecting agencies.
- Access to administrative data from public organisations
- Participation in the set-up of the publicly funded collection of new data.

The intended access regimes should in an equal way apply to all interested parties, regardless of the institute, firm or nationality involved.

The intended access regimes will in principle be suitable as well for the access to data from other, private, sources for the purpose of public research.

## PRINCIPLES

The following principles are proposed to lay down rules for data access regimes:

### ** Openness*

The core principle of the data access regimes should be *open access* as the default: publicly funded research data should be openly available to society subject only to legitimate restrictions. Open access has its price and requires solid funding arrangements


Contact: Peter Schröder, Ministry of Education, Culture and Science, The Netherlands, p.schroder@minocw.nl

* ***Transparency***

Lack of visibility of the existing digital data resources and future data collection poses serious obstacles to access. Information on data producing organisations and their supply, documentation on data sets available and conditions of use should be easy to find on the Web. Research organisations and governmental data collecting agencies should adhere to policies of active dissemination of research data.

**** Legal conformity***

Access to and use of certain data will be limited by legal restrictions. Mention should be made of restrictions for reasons of:

***National security***

Data pertaining to intelligence or political decision making (for example from atmospheric or geological surveys) may be classified and non-accessible.

***Privacy***

Data from human subjects are vulnerable to breaches of confidentiality and privacy and therefore should only be obtained by fair and lawful means, with the knowledge or consent of the data subjects.

***Trade secrets***

Data on business containing confidential information may be non-accessible for research.

In addition to these restrictions, (governmental, company) information under consideration in legal action (*sub judice*) will not be accessible. Restrictions will apply primarily to 'secondary' use of existing data sets collected for other purposes than scientific research. In some cases the sensitive parts of data sets can be deleted without making the whole set useless. Elimination of identifiers from personal data is common practice in the social sciences. When collecting new data, taking the right precautions (for instance asking explicit informed consent from human subjects) may make less severe access restrictions possible. Subscribing to professional codes of conduct may facilitate meeting legal requirements. Responsibility for compliance should rest with institutes (legal persons), not individual researchers.

**** Formal responsibility***

Many of the problems related to access to and sharing of data result from lack of explicit institutional agreements on the terms of access and use. With data management becoming ever more complex and expensive, many traditional informal arrangements between researchers are no longer adequate and need to be complemented by formal rules. Responsibility for the various aspects of data supply such as authorship, producer credits, ownership, sustainable archiving, rights of disposal, financial arrangements, licensing terms, restrictions on use and liability should be formally laid down in the relevant documents such as the formal tasks of institutes, grant applications and research contracts.

**** Professionalism***

Making use of codes of conduct for professional scientists and their communities could help to promote subsidiarity and simplify the regulatory part of access regimes. Mutual trust between researchers and trust between researchers, their institutes and other organisations can play an important role in their establishment and maintenance. In some cases adherence to professional codes can lead to preferential treatment concerning regulatory restrictions and tariffs. On the other hand, in current research practice, the initial data collecting researcher (and/or his institute) is sometimes rewarded with temporary exclusive use of his data. However, such incentive arrangements are not self-evident and should be formally accounted for. Currently, one of the most serious entry barriers to data may be their inaccessibility as a result from ignorance and neglect (ending up in shoe-boxes)

**** Intellectual Property***

The basic rule should be: publicly funded data are public property and should be publicly accessible. In principle research data as such do not qualify for intellectual property rights. Where copyright or database law applies, the publicly funded parties responsible for agreements and contracts concerning access to data should take the relevant implications of the existing legal framework into account to allow for open access. As public/private partnerships in the funding of research and data collection for research are increasing, balanced public/private arrangements should facilitate broad access. Access for science, as well as for business, to publicly funded data should not be impeded by prohibitive costs based on intellectual property rights.

**** Interoperability***

Although science is a highly globalised endeavour, incompatibility of technical and procedural standards used can be a most serious barrier to multiple uses of data sets. A first requirement for interoperability would be the explicit mentioning of the standards employed. It may not always be easy to reach agreement on (international) standards for classification and documentation of data and technical standards for storage and retrieval. There are indications that in some disciplines standardisation has rather low priority. Adopting the practices of disciplines more advanced in this respect (for example physics and astronomy) should be promoted. There is certainly an important role to be played by the international professional organisations, by the International Council of Scientific Unions (ICSU) and its Committee on Data for Science and Technology (CODATA). Concerning the more general ICT standards, the work of an organisation like the International Organization for Standardization (ISO) could be considered.

**** Quality and Security***

Problems of interoperability overlap with issues of quality control. The value of research data depends on their compliance with explicit quality standards. Data should be documented on the methods and techniques employed in their collection and archiving, and the measuring instruments and calibration.


Contact: Peter Schröder, Ministry of Education, Culture and Science, The Netherlands, p.schroder@minocw.nl

Specific attention should be paid to:

***Authenticity***

Origin of sources and authors should be documented and specified in a verifiable way.

***Integrity***

There should be guarantees for the completeness of data and absence of errors.

***Security***

Data should be protected against loss, destruction, modification and unauthorised access in conformity to explicit security protocols.

***Liability***

The responsibility for potential damages following errors in the data released (e.g. damage to health or environment) should be taken into account. There could be use for traditional peer review along professional scientific lines, quality control of a more technical character could sometimes be another option.

**** Efficiency***

Collection of the right data for research often entails substantial effort. Multiple use of the same data can avoid unnecessary duplication of data collection. On the other hand many archived data sets available for secondary use remain untouched after their initial use. Use of accepted retention protocols and thorough documentation of data will help reduce unnecessary duplication of effort as well as to establish the necessary selectivity in preservation. Cost effective collection, use, management and archiving of data require specialised support services. Insufficient incentives to researchers or database producers may lessen their efforts on data related activities resulting in a sub-optimal return on public investments. Adaptation of current reward structures could be a way to avoid this. Other options would be the employment of non-academic specialists or outsourcing this task to specialist firms.

**** Accountability***

Gaining a clear insight in cost, benefit and performance of data access regimes will not always be easy for the public and not even for people closer to research. To ensure sustained support in science and society, those in charge of data access regimes should put considerable effort into showing the benefits of open data access.

---

# BACKGROUND / Contents

# I. Summary

Openness has always been a central characteristic of the societies of OECD countries. The democratic political systems, the market economies and the education and science systems of OECD countries all are based on the body of thought of openness guaranteeing a free exchange of ideas, information, commodities, services and knowledge. Openness is a *sine qua non* for science. Progress of science as a collective, "communal" and cumulative endeavour has been the result of a free flow of ideas, information and knowledge, scrutinised and debated by an open forum of peers. OECD countries fund open science as a public good. Individual citizens and their institutions and industry do have open access to the results of science in publications. The openness of science facilitates accountability: quality, productivity and efficiency are open to public review.

The use of contemporary Information and Communication Technologies (ICT) in principle enhances the value of openness. Use of ICT has opened up new connections between society, economy and science, strengthening the networks of the knowledge-based society and tapping new resources for prosperity and welfare. ICT have led to closer interaction between science, technology and industry and have strengthened the links between social sciences and the institutions of the societies of OECD countries. Use of ICT has connected life sciences tighter to public health.

The requirement of openness in science is not new, but the confrontation with the basic assumption of openness will be more intense in the ICT environment. ICT have opened up new fields for application of the *sources* of research: the base material of *research data*. Digital data from research are increasingly used for other purposes in other research fields and in industry. Digital administrative data from the institutions of OECD societies have found extended use in social sciences as well as policy making, strengthening the evidence base of policy making.[2] Digital data from public health play a growing role in life sciences. Use of ICT has made collections of scientific data in many respects comparable to musical scores: to be used time and again for a diversity of performances from a diversity of artists.

Effective *open access* to research data in a responsible and efficient way is required to reap the benefits of the new opportunities ICT offer. As the Arzberger Report[3] concluded, "Open access" is to be taken as the easiest access on equal terms at the lowest possible cost resulting in

[2] Important collaborations between the social science research community, the statistical office and the administration on the access to data are have been worked out in Germany and France. See "*Les sciences sociales et leurs données*" (1999 Paris) de Roxane Silberman (http://www.iresco.fr/labos/lasmas/document.htm) for France and for Germany "*Wege zu einer besseren informationellen Infrastruktur*", Baden Baden 2001, from the Kommission zur Verbesserung der Informationellen Infrastruktur zwischen Wisschenschaft und Statistiek (hrsg), (http://www.bmbf.de/presse01/338.html). Also "International Social Science Data Service": Scope and Accessibility", Report for the ISSC, Ekkehard Mochmann, Cologne 2002.

[3] See the report "Promoting access to Public Research Data for Scientific, Economic and Social Development" from the CSTP Follow-Up Group led by Peter Arzberger, DSTI/STP(2003)20, also at:http://dataaccess.ucsd


Contact: Peter Schröder, Ministry of Education, Culture and Science, The Netherlands, p.schroder@minocw.nl

maximum use. The authors stated earlier that open access to research data is an important condition to:

- Good Stewardship of public knowledge;
- Strong value chains of innovation;
- The creation of value from international co-operation.

The international *Human Genome Project* is a good example of a large research endeavour in which an open accessible repository of data was used successfully by many different researchers all over the world for different purposes at different places and times. Scientific databases are rapidly becoming a decisive part of the infrastructure of the global science system. The data sharing policies of the US National Institutes of Health[4], accountable for one of the largest research budgets world-wide, are a good example of regulation intended at meeting the new challenges.

Paul Uhlir points out[5] the central role of databases: *"…the great and continually increasing importance of databases in the national and international research enterprise, on an equal level with major research facilities, the scientific literature, and technological innovations. Databases are a fundamental infrastructure component of modern science. Not only are most of the journal articles and patented technologies dependent on the production and use of S & T data from public research, but databases are increasingly the source of major discoveries and innovations themselves, through data mining techniques and the integration and transformation of multiple data streams into new databases and knowledge. Making those factual inputs freely and openly available on digital networks makes them a "force multiplier" and vastly increases S&T progress through the network effects of the Internet."*

The governments of OECD countries have all introduced additional *information policies* to meet the challenges of the emerging Information Society in general. These policies promote the benefits of openness to citizens but at the same time take into account the new risks of openness for legitimate interests. A closer policy look had to be taken, for example, at the protection of national security, privacy and intellectual property.

Additional attention to the ICT environment has also been included in general *science policy*, responsible for the efficient spending of taxpayer money on research and the functioning of the science system. Capitalising on the new ways of ICT facilitated openness is a central issue in this. Additional attention to the conditions of 'external' use of research data and the quality of data

[4] US National Institutes of Health (NIH) Final Statement on Sharing Research Data (2003): (see http://grants2.nih.gov/grants/policy/data_sharing/index.htm )

[5] Institutional and Legal Strategies for Promoting Open Access to Scientific Data Resources in Developing Countries," in International Public Goods and Transfer of Technology Under a Globalized Intellectual Property Regime, Keith Maskus and J.H. Reichman, eds., Cambridge University Press (forthcoming, 2004).


Contact: Peter Schröder, Ministry of Education, Culture and Science, The Netherlands, p.schroder@minocw.nl

(e.g. liability in case of irresponsible errors) will be necessary. The US adopted the new Data Quality Act[6] to this purpose. In the end, an appropriate mix of policy measures, incentives and regulation tailored to the ICT environment will be necessary for stable *data access regimes*.

In the current research practice there is often uncertainty about who is responsible for what kind of access to which kind of database. This can lead to irresponsible functionaries into sending valuable data resources, at the click of a mouse, to parties largely unknown – without realising the possible negative consequences. More often the uncertainty will give law abiding functionaries an excuse for not releasing any data for use outside the institute – without realising the loss this means for the progress of the research the institute actually collected the data for. Uncertainty and unfamiliarity about the *do's* and the *don'ts* lead, at its best, to time-consuming attempt to meet the relevant requirements or, at worst, if possible, to the expensive collection of existing data all over again. Science systems that succeed in establishing responsible and efficient data access refimes will be the first to profit from the new potential of digital research data. Science systems that act less energetically will eventually lag behind in quality and productivity.

To quote Paul Uhlir again[7]: *"The adoption of a presumption of openness, coupled with suitable higher-level principles and guidelines, will provide great impetus to scientific progress and help unleash the limitless potential of those resources in the hands of those who can do the most with them. Clearly the value of data lies in their use"*.

From an OECD perspective, the timely realisation of a global level playing field for all concerned will be the logical aim, and the establishment of Principles and Guidelines will be a logical instrument. The formulation of OECD Principles and Guidelines for access to publicly funded research data will greatly contribute to a coherent and consistent science policy approach at the international level.

---

[6] DQA, aiming at "Maximising the quality, objectivity, utility and integrity of disseminated by federal agencies", setting much of the standard for the relevant research institutes. *See* www.whitehouse.gov/omb/inforeg/agency_info_quality_links.html.

[7] Op cit. in footnote 4.

# II. The Policy Context

## General information policies and scientific information policies of governments

The availability of the current powerful ICT facilities has opened up a broad range of information sources to new audiences at relatively little cost. The collection and ensuing preparation, documentation and editing of data can be very expensive, but the cost of additional delivery in a network environment can be considered nil. Consequently, governments of most OECD countries have formulated general policy principles in a way that promotes optimum Internet access to "Public Domain Information". This concerns administrative and other taxpayer funded information sources, primarily for the purpose of democratic transparency and accountability. Data and information from publicly funded research usually fall into this category.

At the same time these information policies entailed additional regulation and legislation restricting access to certain information to protect the interests of the state (national security), its citizens (privacy), and industry (intellectual property rights) that might be harmed in the Internet environment.

The new electronic ways of accessing digital data promise large benefits to science, society and industry, particularly in the cases of scientific research and innovation. Access to and use of the original research data (i.e. apart from the value added to the original 'raw' data by analysis and processing into statistics and publications) in ICT powered databases enable multiple uses of the same data for different purposes. Extended use will transcend the traditional boundaries of individual research projects, institutes, disciplines and nations. A much higher return on public investment in expensive data collections for the global science system, as well as for industry worldwide, may be the result.

At the same time the additional, 'secondary', use of digital research data outside the confinement of the original data collecting institutes makes new demands on the quality and reliability of the new data supply. To make demand and supply meet more satisfactorily, the emerging market place of research data needs more transparency.

## The ICT interface: output and input

Increasing use of ICT has changed the science system and its place within society. Enhanced means of communication and information have enabled science to penetrate deeper into the life of citizens. The *output* of research, as published in electronic scientific journals, today reaches a larger and broader audience in the science communities, governmental organisations, industry, the public media and associations of citizens.
Faster and easier desktop access to research results in peer reviewed scientific journals has substantially strengthened the knowledge base of OECD countries.

On the *input* side of research continuously growing quantities of data from the most diverse parts of the universe around us are collected by research institutes, governmental agencies and industry.

Research data constitute the raw material to be digitised and processed for further research and information purposes. In digital form, research data are developing from a *base material* into a *half product*. In this way an increasing supply of data is becoming available for uses beyond the purposes of the original collection. Many of these digital data, for the most part publicly funded, can be of great value for additional use by a multitude of research institutes, governmental agencies, R&D departments in industry, public media and organisations of citizens. Data from very different sources from all over the world can be used in new combinations. Demand for these data is rising.
In OECD countries faster and easier desktop access to research data is opening up an expanding range of fields for further scientific enquiry.

**Access to electronic publications and digital data**

Among researchers there is a continuing debate on the changes E-publishing is bringing to the dissemination of scientific knowledge. The formation of the *Public Library of Science* [8] highlights the uneasiness that many researchers feel with the profits commercial publishers make on publicly funded content. There have been complaints about the subscription rates of the 'name journals', the negative effects of copyright transfer and the slow pace at which large scientific publishers have adapted to the opportunities ICT offered for cheaper and faster knowledge dissemination. Researchers in some disciplines (for instance in high energy physics[9]) pioneered new ways of very efficient Web-based publishing long before the established publishers followed. But in the end, no one will deny that access to research results published in scientific journals has never been as fast, efficient and user-friendly as today.

E-publishing can be seen as a modernisation of an established distribution system with a long and successful tradition of trustworthiness. The same commonly accepted standards of validation and quality control guarantee a high level of trustworthiness of the knowledge published by the established scientific journals. The ongoing debate tends to concentrate on the most efficient business models to be employed.

There have been large scale data collection efforts in a number of fields since about the middle of the nineteenth century – along with the rise of International Statistic Societies. From this time data were collected for example on diseases to enable epidemiological research. Longitudinal surveys of flora and fauna have run for over 100 years.
Following the *International Geophysical Year* (1957 – 1958) researchers from geophysical, atmospheric, marine and other relevant fields pioneered *World Data Centres* (WDC)[10] under the umbrella of the *International Council of Scientific* Unions (ICSU) from UNESCO. ICSU set up

[8] "The Public Library of Science (PLoS) is a non-profit organization of scientists and physicians committed to making the world's scientific and medical literature a freely available public resource", see www.publiclibraryofscience.org

[9] see e-print arXiv

[10] see World Data Center Homepage at www.ngdc.noaa.gov/wdc/wdcmain.html


Contact: Peter Schröder, Ministry of Education, Culture and Science, The Netherlands, p.schroder@minocw.nl

the CODATA committee on data for science and technology[11]. From the 1960s, social researchers from Germany (Zentralarchiv[12]) and the US (the International Inter-university Consortium for Political and Social Research (ICPSR[13]) set up the first data archives for the social sciences. In 1977 they joined forces in the *International Federation of Data Organisations* (IFDO), member of UNESCO's *International Social Science Council* (ISSC).
However, release and dissemination of research data as a commodity in its own right is a more recent phenomenon and has only really taken off with the large scale introduction of ICT in research, the contemporary PC-Internet networks and large databases in particular.
In many cases working procedures for quality control and technical standards for access to these databases are not yet fully established. The regulatory frameworks for access, as well as the business models, are largely in a developmental stage. The *Clinton-Blair statement* on the public ownership of data from the *Human Genome Project* [14] highlights part of the debate on the rights of access to research data. In the end, access to the new potential of digital research data is far from being fully realised.

**Treating research data as a force on its own**

Before the spread of PC-Internet networks, ('raw') data constituted the base material of the research process, usually collected by researchers themselves. In most cases data were inextricably linked to specific projects and of little use outside that realm and, consequently, accessibility for further, 'secondary' use was no issue. Apart from that, open accessibility of these data for review purposes has always been considered mandatory as "*Final research data comprise the factual material commonly accepted in the scientific community as necessary to validate research findings*" [15].

Outside the science community in the strict sense, data collection with comparable scientific methods and techniques has been the core business of many large national government agencies and services. Well-known examples are the national censuses, ordinance surveys and meteorological and geological institutes. There are many more public institutes delivering information for use by national government, the science community, industry and citizens. In these cases the base material, the 'raw' data, are processed (usually less exhaustive than in research) into statistical information products.
At the same time, digital administrative data from governmental organisations and other large organisations ranging from hospitals to social security services, insurance companies and educational institutes qualify for research purposes in the social sciences. Access to these sources

[11] see www.codata.org

[12] see www.gesis.org/ZA/

[13] see www.icpsr.umich.edu/

[14] http://www.whitehouse.gov/library/PressReleases.cgi?date=0&briefing=2

[15] see url in footnote 3

for research will contribute to the progress of social sciences. The results of this research will strengthen the evidence base of governmental policy making.

**E-Science**

The use of contemporary Information and Communication Technologies (ICT) has increased the scale of scientific research, transcending traditional institutional and national borders, promoting global research. Use of ICT has also broadened the scope of research, transcending traditional disciplinary boundaries and promoting multidisciplinary collaboration. Hitherto unconnected elements of the research process can be assembled into unexpected new configurations. The research strategy developed by Rita Colwell in her studies on cholera is a case in point[16]. 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, would have remained invisible.

Large collections of digital data will play a central part in the emerging Global Science System. The 21st Century holds the promise of E-Science:

"*E-Science will refer 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)[17]

E-Science requires a "Cyberinfrastructure". The US Blue Ribbon Advisory Panel on Cyberinfrastructure[18] anticipates *"…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."* The panel sees *"… significant risks and costs if we do not act quickly and at a sufficient level of investment"*. Risks include *"… adoption of incompatible data formats in different fields; permanent loss of observational data due to lack of well-curated, long-term archives…"*

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

[17] www.research-councils.ac.uk/escience/.

[18] "Revolutioning Science and Engineering Trough Cyberinfrastructure: Report of the National Science Foundation Blue Ribbon Advisory Panel on Cyberinfrastructure, http://cise.nsf.gov/evnt/report/atkins_annc_020303.htm


Contact: Peter Schröder, Ministry of Education, Culture and Science, The Netherlands, p.schroder@minocw.nl

## Formal requirements

The realisation of the new knowledge potential requires additional institutional arrangements. Transparency and equality require access conditions of a more formalised character. A recent survey *"Policies on Digital Research Data: an international survey"*[19] found that a majority of research funding agencies in OECD countries expected policies on data access to become a major policy issue. The same report concluded that additional regulation from funding agencies was most advanced when new legislation[20] of information rights required explicit implementation.

When changes in research policies and data management are called for, an international level playing field for data access requires a co-ordinated international policy effort. Commonly accepted principles from OECD can play a decisive role in the establishment of efficient data access regimes.

## The role of national governments

National governments are responsible for the quality and efficiency of the national science and innovation systems. Use of ICT and digital data have increased the scale, scope and pace of research in such a way that changes in science policy and research management are called for. As data management plays an increasingly important part in this transition, specific science policy attention for this transition will be required. A role for governments is all the more advisable as the management of digital research data involves other societal interests that governments are responsible for, such as information rights, national security, privacy, intellectual property rights and foreign affairs. To strike the right balance between the various interests at stake, an active role of governments will be indispensable. Taking the required specific practical measures will be the task of the various scientific organisations and institutions. Governments can draft the framework by establishing the *Principles* of the course to be taken by the relevant parties. Agreement on principles will help to shortcut red tape and attempts to perpetually re-invent the wheel . OECD consensus on those principles will contribute to the consistency and coherence of the actions on the international level.

## The role of international organisations: OECD

There is a real risk that disparities in national practices, legislation and regulations will hamper the optimum use of publicly funded research data on the international scale. As science and technology are part of the larger Global Science System, international policy attention will be a necessary complement of the policies at the national levels. The *International Council of Scientific Unions* (ICSU) and its CODATA subsidiary have for a long time been active in developing international co-ordination and co-operation concerning access to research data. UNESCO itself is in the process of formulating *"Policy Guidelines for the development and promotion of Public Domain Information"*. [21]
As the OECD is the most influential and authoritative intergovernmental forum for international science and technology policies, a harmonising role for OECD becomes apparent. Co-ordination

[19] Paul Wouters: "Policies on Digital Research Data: An International Survey", NIWI-KNAW, Amsterdam 2002.

[20] For example the "Shelby amendment" to the Freedom of Information Act and the Data Quality Act.

[21] See www.icsu.org

with similar activities of UNESCO, ICSU and CODATA can strengthen the OECD initiative. OECD has come up with influential *Principles and Guidelines* for ICT related issues in the past. The "*Guidelines governing the Protection of Privacy and the Transborder flow of Personal Data*" [22]and the "*OECD Guidelines for the Security of Information Systems and Networks*"[23] are examples of successful policy action aiming at international consistency and coherence. The introduction of "OECD Principles and Guidelines for Access to Publicly Funded Research Data" would be a logical next step.

---

# III. Considerations

### A free flow of data, information and knowledge

Overall, success in science as well as innovation is based on an optimum exchange of ideas, information and knowledge. This applies to research data as well. There certainly are research data that have only a very limited, one-time use. Generally speaking the data collected in the 'Big Science' endeavours of physics and astronomy are systematically distributed for rather exhaustive use among the participating researchers. Secondary use of the raw data is limited.
Social and health sciences show quite another picture: that of an archipelago with a broad variety of differing data sources that are far from exhausted but need quite some registration, documentation and cataloguing. Many domains have gained new research centrality through the wide-scale distribution of their data in flexible, queriable electronic databases. At the same time, sciences which have been traditionally laboratory intensive and whose emphasis has been on the reproducibility of experiments have shifted to a more data driven model of enquiry, reliant on persistent databases. Over all, the emergence of large databases in practically all scientific disciplines demonstrates the value of many research data for extended use. The US National Institutes of Health [24]state:
*"Sharing data reinforces open scientific inquiry, encourages diversity of analysis and opinion, promotes new research, 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, and permits the creation of new data sets when data from multiple sources are combined. By avoiding the duplication of expensive data collection activities, the NIH is able to support more investigators than it could if similar data had to be collected de novo by each applicant".*

[22] Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, 23 September 1980, [C(80)58/FINAL]

[23] Guidelines for the Security of Information Systems and Networks [C(2002)188/FINAL]

[24] http://grants2.nih.gov/grants/policy/data_sharing/index.htm


Contact: Peter Schröder, Ministry of Education, Culture and Science, The Netherlands, p.schroder@minocw.nl

## The default - Open Access

> The legal and regulatory frameworks in all OECD countries acknowledge the public interest in a free flow of facts and therefore agree on the inadmissibility of the appropriation of facts alone. It will not be easy to find a reason why these data, when publicly funded, should not be openly available for public use in the Global Science System. "Open access" is to be taken as the easiest (timely, user-friendly, Web-based) access on equal terms (to the international research community as well as industry) at the lowest possible cost (on a non-profit basis) resulting in maximum use (to produce as much knowledge as possible).

Openness is a condition to scientific validation by review. Public accessibility has made science reliable and successful. The cornerstone of scientific validation consists of review by peers (not necessarily limited to those reviewing articles for scientific journals). The open forum of peer review presupposes the possibility of re-analysis of the data used so that publication and public access to those data will be the normal condition. It follows, therefore, that science policies and research management should consider restrictions on open access to data as unfavourable in principle.

> Generally speaking, science and innovation are best served by open access to research data. Knowledge creation is a cumulative process, the larger the data resources available, the more information and knowledge can be extracted. There are no substitutes for data on unique events that cannot be reproduced. Data also provide important baselines to track rates of change and computing the frequency of rare events. The re-analysis of existing data may lead to different conclusions. Thus data allow for the formulation of new hypotheses, which may unexpectedly change the relative importance of the data. Limits on access diminishes, open access increases their value.
>
> Although there is much debate on selfish researchers locking their colleagues out from use of data they collected, ignorance leading to insufficient data management is probably a more severe access problem. The proverbial unregistered shoebox repository containing undocumented files in outdated formats on obsolete media may be a far greater danger to access. A greater danger than the formalised temporary exclusive use based on explicit funding conditions that in the end are dependent on policy and regulation.

### *Open access for science in developing countries*

There is no reason why open access of publicly financed databases should be limited to the country of initial data collection. Sometimes risks of international free riding are mentioned, but the fact that eventually the original investor will only benefit from the added knowledge resulting from additional use sounds more convincing.
The benefits of international access to research data from public funding will *a fortiori* apply to the science in developing countries. If costs involved under standard regimes are prohibitive, preferential (free) rates could improve the situation. As lack of the research infrastructure to be able to use state of the art research databases will pose serious problems, making available additional facilities (hardware, software) and services (training, advice) should be considered.
Research funding agencies in OECD countries should be seen as the parties to facilitate these processes.


Contact: Peter Schröder, Ministry of Education, Culture and Science, The Netherlands, p.schroder@minocw.nl

## Research data

The proposed Principles and Guidelines are aimed at publicly funded research data. This target does not preclude any positive effects of these Principles and Guidelines for the use of research data funded from other sources. Therefore, aspects of access to data funded from non-public sources will also be discussed.

Research data can be defined as the factual records (numerical scores, textual records, images and sounds) used as sources for scientific research. *"Scientific data refers to the numerical quantities or other factual attributes generated by scientists and derived during the research process (through observations, experiments, calculations and analysis)"*.[25] A genomic sequence, the speed of subatomic particles, the orbit of the earth, the water temperature, a response in a social survey, the frequency of nouns in a text corpus, satellite images of the surface of the earth all are used as research data. They constitute the first stage of sorting out the part of nature and society around us that constitutes the universe to be studied. Practically anything imaginable can be used as data for research.

There are observational data (data recorded from premeditated observation of natural and social processes), process generated data (e.g. data stemming from relatively autonomous processes like administration). There are unique, irreproducible data from historical observation, there are data from reproducible experiments. There are quantitative data (numerical records) and qualitative data (records of participant observation, diaries, images). There are data collected for research purposes (methodological tailored to the research purpose) and data from other sources suitable for research (collected with a comparable methodology). If collected and prepared with a suitable scientific methodology, all of these types are relevant to data access regimes.

Databases contain scores of individual data that are systematically documented with a machine readable code-book, a description of variables, frequencies and methodological information. Open access largely depends on the accepted set of *metadata* being used.

## Data as the building bricks of science

Data are the building bricks of research findings on the one hand, and on the other hand, availability of the relevant data is indispensable for the validation of research results. As the US National Institutes of Health states: "*Final research data comprise the recorded factual material commonly accepted in the scientific community as necessary to validate research findings*".[26] "Final" means the final stage of routine processing according to established standards.

The course of data processing in 'traditional' research can in a simplified manner be described as: Collection of raw data>> data preparation/checking/editing >> documented final research data >> scientific analysis >> scientific conclusion.
Making the data systematically available in a data archive/repository opens up many alternative trajectories of simultaneous and subsequent research. Additional research may be carried out in

[25] CODATA Working Group on Archiving Scientific Data, www.nrf.ac.za/codata

[26] See NIH: http://grants2.nih.gov/grants/policy/data_sharing/index.htm


Contact: Peter Schröder, Ministry of Education, Culture and Science, The Netherlands, p.schroder@minocw.nl

the same discipline for the original purposes, but also in other disciplines, for other purposes. Combining different data sets will broaden the scope of research. Combining data from time series of comparable experiments and surveys will enlarge the time dimension.

Practicability of those additional trajectories is dependent on common standards of technical interoperability of the hardware and methodological standards and metadata.

### Refining raw data

> In the research practise it is often hard to draw the line between "raw", primary data, further structured information and the final knowledge of scientific findings. One man's data are another man's knowledge and of course the output from one research project can be used as data input for another. Further along the course of scientific research the data will be further analysed and modelled in the process of scientific research 'proper'. Here the contribution of creativity and originality will be growing and will end up in the research findings.
>
> At one end of the scale, research data start out as, "raw" data to be taken as routinely, often automatically collected, scores representing the first round of putting scientific order in the social and natural universes being studied. Once a standard design of data collection has been established, there will be not much originality or creativity involved. The effort will be primarily directed at precision and representativeness of samples drawn from nature and society to be studied further.

### Non disclosure

> For reasons of national security, protection of privacy and protection of trade secrets, data owners may be legally bound not, or only partially, to disclose their data. The same sort of excludability goes for data on matters *sub judice*. Parties collecting data for research could also decide not to disclose the data for other reasons than the above-mentioned legal restrictions. Any such decision will put the research involved outside the realm of standard science and its time proven quality control by the (international) peer review forum.

National security can be a legitimate reason not to disclose data and/or results from research. The prospect of commercial viable results (possibly qualifying for patents) can be another reason not to release detailed information (including data) on research. In the latter case the secrecy will usually be temporary, in the case of patenting the publication of detailed information in the publicly accessible patent register will eventually be a requirement. Exclusive rights on data and related research information will usually be of a partial and/or temporary character because commercialisation in general, and particularly patenting, eventually requires publication. Chances are that further monopolies on co-funded data will conflict with competition law.

### Restrictions on access

The legal and financial framework under which research databases can be compiled, exploited and accessed has a number of restrictions. In principle, only a few of these restrictions will, under certain circumstances, pose serious problems to access publicly funded research data. There probably will be no way to access (company) data under consideration in legal action. The same goes for (company) data considered trade secrets. Sometimes access may be difficult or expensive under these restrictions. There may be cases where existing practices based on traditional agreements no longer seem viable, but in the end, if the right terms in the relevant contracts are stipulated explicitly beforehand, most of the access problems posed by legislation can be solved.

For researchers it will take a learning process in which research policy and management should give guidance. The following aspects are mentioned here:

- *Privacy of human subjects*

Personal data play an important role in social and behavioural sciences, economics and epidemiology. This may concern data subjects, individuals related to the data subjects under study (doctors of patients, teachers of pupils) as well as researchers. Although in the research results usually only aggregate information is published and identification of individuals is normally impossible, unintended disclosure of personal information during and after the research process is recognised as a real risk.
Legal requirements of privacy and confidentiality must be taken very seriously. The emphasis should be put on security measures that include the deletion of identifiers and limit access to sensitive personal information to as few formally authorised researchers as possible.
If possible use should be made of volunteer respondents having agreed to co-operation on the basis of informed consent. When secondary use is made of existing personal records, arrangements comparable to informed consent are necessary.

- *National Security*

The case of restrictions of access and use of research data for reasons of national security presents a complex picture. Almost any data, information and scientific finding can play an important part in warfare and/or terrorism. There is enough knowledge and information accessible trough the Web to build your own nuclear weapons or anthrax spray on some scale. But it would be hard to imagine an access strategy that eliminates 100% of the risks, without serious damage to the social and economic life of the countries of OECD. The same is true for information on populations and economies: it would be impossible to govern civil society without responsible access to it. National security clearly might demand the non-disclosure of factual data but also withholding information on the process of decision making and classifying procedures. Experts from governments should take up emerging issues with the experts of the relevant funding agencies and research institutes and governmental responsibilities should prevail.

- *Openness and exclusivity*

> Openness presupposes more than passive availability. Active dissemination of systematic releases of research data in a user-friendly way will be a necessary condition for optimum use.
>
> Science and innovation are best served by non-exclusive open access to research data. However, sometimes it may be helpful to allow for temporary monopolies. Temporary (partial) exclusive use of data may be an incentive to invest in data collection by rewarding the original data collector/investigator with a right of “primogeniture” to the first results of further analysis of the data. In publicly funded research this reward will be of a scientific nature: scientific discovery. In the case of data collection and ensuing patenting in industry, the reward will be financial: the opportunity to be the first to file a patent and make profit out of a commercially viable invention.

## Intellectual Property aspects

> Intellectual property laws are meant to protect the interests of those who create original works (of arts and science) by assigning them (temporary) exclusive rights to the use of these works. The rights on the

exploitation of the works offer the opportunity to recoup the investments made in creating the work. In this way intellectual property rights promote the creation and the dissemination of original works (of arts and science). National regimes of intellectual property rights show differences, on the international level, treaties of the World Intellectual Property Organisation are authoritative. The use of contemporary ICT, with its possibilities of easy and endless copying and distribution, has entailed the updating of regulation and legislation that, generally speaking, puts additional limits to access to information.

As far as access to research data from public funding is concerned (something quite different from the patenting of scientific inventions), intellectual property rights in principle need not play more than a marginal role. Provided that the (publicly funded) parties responsible for agreements and contracts take the relevant implications of the existing legal framework into account beforehand, open access can also be realised under database law.

- *Intellectual property rights*

Once in the open, claims of ownership of data will not be supported by law. Although research data in general cannot be considered intellectual property, there are cases in which intellectual property rights or copyrights that might limit access can be exercised. Commercial exploitation of a database can be based on database rights. In that case market rate prices should be the rule (and appeal to competition law will be possible when pricing is prohibitive or when unfair preferential pricing is practised.) Apart from database rights, use of database software may involve patents and/or copyright. This can be relevant for the access to formats and/or the questionnaires used to collect the data and the tools used to process the data.

Furthermore, along the research course the process of "refining" the "raw" data by researchers will add original, creative value. This could justify IPR or copyright on the "refined" data.

As long as it concerns publicly funded research and publicly funded data, possible problems of accessibility will have to be solved by funding conditions of the public funding agencies.

- *Copyright law*

Important research results end up in scientific journals of publishers who organise the peer review and editing processes and usually acquire the copyright for the articles. In a growing number of cases, simultaneous publication of the data used, although often in a condensed form, is a requirement for the publication of an article in a scientific journal. In the current research context the interests of the various partners involved, be it researchers or publishers, put such a high premium on prompt and accessible publishing that in such cases IPR pose no barriers to data access. In the case of public funding, conditions in grants and employment contracts should mention terms of access to the data.
Important aspects of copyright and intellectual property law may differ from country to country. However, most legal regimes have exceptions for uses in education and science. Exempted uses in the US fall under "fair use" clauses. If copyright law applies, either private individuals, private institutes or public institutes can be proprietors. US regulation forbids private copyrights on publicly funded information. Under UK and Commonwealth rules, publicly funded information may rest with the Crown (Crown Copyright). If proper care is taken in research funding conditions, effective open access can be arranged under all of these regimes.

- *Database law*

As the primary data get processed and put in databases, the software added for storage and retrieval represents value protected under software intellectual property law.
In a growing number of cases[27] a database producer can claim *sui generis* intellectual property rights to the database, even if the database content does not qualify for such protection. The EU Database Directive is primarily targeted at databases for commercial services (financial and legal data) but does not exclude research data nor data from public funding. Like the US Digital Millennium Copyright Act, the terms of recent database legislation in principle extend the protection of information at the expense of the public domain. The DMCA can make bypassing protective cryptography (potentially making also 'un-copyrightable' material inaccessible) illegal. The Database Directive can make it difficult and/or expensive to make use of non-protected information that is only available as part of databases. Scientists are sometimes apprehensive of additional barriers to the use of data for research. Until now no cases of conflict on accessibility of publicly funded research data under the EU Database Directive (that has been implemented in different European countries in slightly differing ways) have been documented.

On the international level, negotiations are taking place to enter into a treaty to harmonise the various legal database regimes within the framework of UNESCO' s World Intellectual Property Organisation (WIPO). In this process governments should pay due attention to the requirements of open science.

Protagonists of database protection are convinced that private database owners have every interest in setting reasonable market prices for access in order to make a profit. In the end those who exploit databases containing data from public funding should be bound by funding conditions stipulating public access rather than database law.

## Authorship

Within the framework of funding conditions and employment, research findings can be attributed to individual scientists and institutes entitled to intellectual property rights also relevant to access and uses by other parties. Authorship of databases presupposes publishing and therefore a degree of accessibility. Explicit authorship of databases will be important in guaranteeing authenticity, integrity and quality in general. In certain cases (usually small-scale, non-routine research) authors may legitimately wish to impose certain restrictions on the use of their data. They may wish to be assured that the data will not be (wilfully) misinterpreted, that reference be made of the context in which the data were collected and may wish to discuss proposed usage/interpretation with the author prior to publication (cf. the moral rights in copyright law).

Database authorship credits in published scientific articles will help to make research institutes and research teams aware of the importance of data related activities. In the case of public funding, conditions in grants and employment contracts could stipulate authorship.

## Funding and exploitation of research databases

[27] The legislation stemming from the EU Database Directive, the WIPO Database Treaty currently being negotiated.

Publicly funded science, its data, information and knowledge are public goods ; non-rival, non-excludable goods. The long term and the broad scope of its return on investments most often make it hard to adapt the activities involved to market conditions. Although the collection of raw data and their subsequent preparation, editing and documenting may be very expensive, the market value of research data will usually be rather limited. Considering the 'public good' character of public science and the premium of network effects, rigid recovery of the initial costs of data collection will not often contribute to an adequate exploitation.

Digital research data as the base material of research are an intermediary to produce knowledge, the end product. In a technical sense, digital data do not wear out in use. In most scientific fields, the content of data can also hardly be exhausted. As a consequence, sharing data will produce more knowledge from the same investment in data. In the context of public research, where the produced knowledge will be accessible to everyone, even the party that invested in the original data collection will end up with additional knowledge about the phenomena studied, at no extra costs.

The impression should be avoided that "open access" should mean access without costs or access to each and every data set imaginable. Even when a positive cost/benefit ratio seems obvious, the economic models employed should give an indication of costs involved in the implementation of data proposals. Transparency of cost effectiveness will be a strong incentive for setting the necessary priorities in data policies and data management. There will be enormous amounts of data that will not be of interest outside the realm of the initial use. Selectivity in the trajectories of data collection, classification, documentation, searchable storage, dissemination and sustainable archiving will be imperative.

**Costs of implementation of the proposed principles**

As stated above, costs of additional (Internet) access and delivery will be small compared to the investments in data collection and preparation already paid for by the taxpayer, as well as the expenditure legally required for sustainable archiving. Rigid cost recovery may end up in substantial transaction costs. Open access in most cases will short circuit administrative measures entailing superfluous "red tape" and rising handling costs.
Transparent regular access regimes replacing poorly organised incidental requests will save on costs of delivery.

Currently, in some cases research data are delivered on the basis of subscription fees, in other cases prices are differentiated according to academic on profit and non-academic commercial uses. These pricing regimes may be at odds with legal requirements of equal access to public domain information. Open access regimes will seldom entail additional costs for compliance or enforcement compared to current regulation and legislation.

**Hidden costs**

Currently a clear picture of the costs involved is lacking. Much of data related costs could be considered "hidden costs" that are not booked as a distinct entry in research budgets. Depending on the activities included, quantitative estimations of the proportion of data related expenditure range from about 3-to-6 % in overall research costs.

Costs of the setting up and exploiting research databases should be part of research project budgets, either as a separate database entry, or as part of the budget of specific projects and programmes. Autonomous data service institutes, at some distance from the actual research, are usually in a good position to implement robust quality control procedures, but they may run a greater risk of missing out on changes in demand from the research shop floor. Meeting demand in a flexible and efficient way will need permanent management attention. Earmarking part of project and program funding may put the demand closer to the supply, but may make it hard to achieve the necessary critical mass. In many cases, current research funding arrangements are not yet particularly well tailored to stable long-term funding of data services.

**Supply and demand**

Open access to publicly funded data in principle means public access. Public access may be laid down in constitutional information rights. The facilities and human resources required to make public access meaningful will entail costs for intermediary services. Availability of 'public use' files and/or professional intermediaries will be needed to pay more than lip service to public access.
In practice, supply-wise, open access could mean access for (the widest range of) scientific research, regardless of its purpose, be it non-profit or commercial research. Looked at from the perspective of demand, open access to data in the widest range of detail could be the starting point. Limits to detail of data (e.g. for reasons of privacy protection) may be acceptable to certain research purposes. Alternatively, a system of controlled circulation of data to certified institutes could be the optimum way of realising open access. Considering the possibilities of creating safeguards for compliance with legal and regulatory requirements in certification, (preferential) access for certified research institutes would be an interesting option for open access. Protection from misuse of data (terrorists, criminals) would certainly be easier to manage under a certification regime.

**Access to publicly funded information services**

In the case of the publicly financed governmental data collecting agencies that deliver public information products (for example census, meteorological, geological and topographical information), digital data (in addition to the traditional information products) have developed into a new outlet that can play an important role in many fields of scientific research. OECD countries currently try out a diversity of business models for access to those data, varying from free Internet delivery to substantial fees based on a cost recovery model.

The two principles of 'open public access to information paid by the taxpayer' and 'government should stick to its core business and leave as much additional activities as possible to private parties' have often led to a satisfactory accessibility of those data for business and academe. Part of the premise being that the taxes on downstream generated activities will recoup the investments upstream best.


Contact: Peter Schröder, Ministry of Education, Culture and Science, The Netherlands, p.schroder@minocw.nl

## Different business models

In some cases a model has been chosen based on cost recovery of the expenditure of the agencies. This approach is based on the idea that to match demand with supply as closely as possible, costs of additional services from public agencies should be passed on to the actual clients and not to the general public. This may prompt the agencies concerned to extend their product assortment with (sometimes expensive) data sets. The (unspoken) assumption being that there probably is insufficient outlet for data sets to allow for more than one efficient producer (and insufficient tax income from extended downstream activity). The official EU policy aims at a level playing field for all and prohibits unfair competition with publicly funded assets from the agencies.

Satisfactory accessibility of the data resources from the agencies can be realised in different ways. The ultimate proof of success lies in the actual use that is being made of the agency data for research and business purposes. Research databases that are underused are certainly the most expensive databases.

## Public / Private research

Scientific research and innovation are interdependent processes. Consequently, public and private co-operation on data access will be important. Private parties benefit increasingly from the use of openly available research data from public funding.
Publicly funded research and its data collection stand to benefit from additional funding from partnerships with private firms. The contracts involved may include terms that, within the existing regulatory framework of the publicly funded partner, grant the private partner exclusive rights to the commercialisation of research results. The purpose of making a profit on investment in research may lead to legitimate temporary/partial non-disclosure of the data involved or cost of access for other parties. Well-balanced arrangements will contribute not only to the profit of private parties, but to societal benefits in scientific progress and innovation as well.

## Public / Private databases

> The funding conditions of public/private partnerships in research concern primarily the research results. Participation from commercial parties usually brings along a transfer of patentable knowledge. In the regulation of the European Union, patentable knowledge generated by the public partner should be transferred at market prices to avoid conflict with competition law (illegal preferential treatment). In the case of patents, detailed information that may include essential data, are openly accessible in public registers and available for non- commercial purposes such as research.
>
> In the field of molecular biology there are examples of freely accessible publicly funded databases that have an extensive clientele from industry (e.g. the Protein Data Bank), as well as successful databases from mixed funding (Swiss Prot). Differentiated pricing for public and private institutes can be a way of making financial ends meet. The policy debate on “unfair” barriers to data access often focuses on industrial commercial interests. There is no reason, however, why non-profit and not for profit institutions would not get involved in consortia with an IPR portfolio and exclusive data collections that could limit the level playing field of public research.

## Transborder aspects of access to research data

Apart from regulatory aspects, international data access presupposes some linguistic uniformity, if not for the *lingua franca* of the bits and bytes, then for the natural languages of the accompanying documentation and metadata.

Conditions to access of research data in global 'Big Science', be it in national (e.g. NASA) or international (e.g. CERN) organisations, are usually explicitly formulated in the relevant regulation ranging from formal international treaties or less formal institutional rules. Terms of access may be tailored to the demands of the participants, but as far as data are concerned, openness prevails. As mentioned above, The Human Genome Project is a good example of the positive impact of open access to research data on a global scale. Access to the very large meteorological databases may be based on open regimes or bilateral contracts. Reciprocity may play a role. As far as data access is concerned, fear of international 'free riding' seems absent.

The content of trans-border flows of social and medical science data may restricted by differing national legislation on privacy protection. The international social science community has solved conflicts between UE and US privacy legislation by creating 'safe havens' in the US that meet the requirements of the EU Directive on the protection of personal data.

In a more practical sense, access will be dependent on the available computer hardware and network infrastructures.

## Responsibility for archiving

The value of publicly funded research data, be it small data sets from brain research, large astronomical databases or valuable data from governmental (geological, census, meteorological or epidemiological) agencies lies in their use. There will be a broad range of options for access regimes, from profit making to (partial) cost recovery schemes to differential pricing and free access. The more the available data are used for further analysis in research, the better the data collecting institutions meet their public purposes.

As stated above, there is usually a clear governmental responsibility for sustainable archiving of research data based on information law. In the digital environment the line between exploitation and archiving of data will be increasingly hard to draw. Compared to the traditional use of *archivalia*, a more intensive use of archived digital data is to be expected.

### The importance of sustainable access

The exploitation of scientific databases requires solid funding arrangements. Further value can be added by industry, but when it comes to guarantees for sustainable access in data depositories or archives, commercial parties do not seem to be interested. In the end, the complex and expensive tasks of archiving digital research data will remain with publicly funded institutions. As the boundaries between exploitation and archiving become increasingly blurred, sustainable archiving should be included in data access regimes from the start.
Sustainable e-archiving guaranteeing permanent access to research data may turn out to be the most expensive part of data management. There are socio-economic reasons for the long term

archiving of scientific data in addition to historical and scientific reasons. Scientific data have many industrial uses and other practical applications. The costs of preserving and archiving are relatively small in comparison with the costs of acquiring scientific records anew through additional observation. Governmental responsibility for preserving and archiving scientific data is sufficiently justified.

---


Contact: Peter Schröder, Ministry of Education, Culture and Science, The Netherlands, p.schroder@minocw.nl