

# **High Level Group on Digital Libraries Position Paper on Digital Research Data Access and Preservation**

"Because of the critical importance of data and information in the global scientific enterprise, the international research community must address a series of new challenges if it is to take full advantage of the data and information resources available for research today. Equally, if not more important than its own data and information needs, today's research community must also assume responsibility for building a robust data and information infrastructure for the future". (International Council of Scientific Unions 2004)

## **1. Introduction**

Over the next 10 years the move to a digital knowledge economy should largely have been completed. Government, research, individuals, businesses, libraries, museums and other organisations will be dependent on digital information.

As a key component of this digital economy, the European Research Area requires ready and efficient access to digital information of all kinds such as experimental and observational data sets, databases, scientific journals, theses, conference proceedings and patents as well as unpublished material, called grey literature, such as theses, technical reports, pre-prints and non-refereed publications. The challenges this poses can be grouped in two main areas. One has to do with access to scientific journals, which has received much attention of scientists, publishers and policymakers. Open Access is the name of the debate. It is basically a debate about business models. The other area focuses on how the digital record of science is store efficiently and kept accessible for a long time, basically even permanently. The second area is about repositories, how to organise them, how to fund them, how to solve technical issues to do with long-term preservation, how to ensure quality, and so on and so forth. This paper concentrates on the second area. The conditions, emerging from the first debate, under which publications can be made available, are taken as given.

The records of science consist in essence of publications or documents on the one hand, and data on the other hand where 'data' refers to primary data (resulting from observations, surveys etc) and processed data (for example genome sequences or molecular structures). There are many important differences between documents and data but most of the organisational, funding and technical issues surrounding the creation of an infrastructure for storing and preserving these digital objects, as well as managing access to them (one more taken for granted the outcome of the OA debate) are quite similar.

Since less attention has been paid to all aspects surrounding research data, however, this paper will on many occasions explicitly refer to data, as the growing importance of data alongside electronic scientific journals and books form a new dimension of the problems involved. Indeed, the accessibility and availability of research data is paramount to the further development of e-science, as are a collaborative distributed computing infrastructure that provides shared access to large data collections, and advanced ICT tools for data analysis, large-scale computing resources and high performance visualisation. Increasingly data and

publications are linked providing direct access to the evidence on which reported results are based. In addition, new information services are emerging that build on the results of individual research projects, gathering together scientific literature and data, and using data and text mining techniques for academic and commercial purposes, for example to produce content for bio-databases.

While some research disciplines have developed widely different practices related to research data recording and sharing among researchers, there is an increasing awareness of the importance of devising sound policies, standards, guidelines and effective practice for the management, storage, and curation of research data and mechanisms for the involvement of key stakeholders in this arena. Particularly grave risks are the consequence of unresolved challenges in the long-term management, access and preservation of all sorts of digital records of science, and in particular those related to research data.

First of all, the differences between data and documents will be briefly considered, and the possible consequences for the repository infrastructure (section 2). In dealing with these issues some guiding principles are emerging which have to be translated into a framework of legal arrangements and policies (section 3). Secondly, the dimensions become visible of a sustainable virtual infrastructure that is required for permanent access to scientific information, including all disciplines from physical, biological, or environmental sciences, to social sciences and humanities (section 4). Thirdly, establishing such an infrastructure requires that a host of implementation issues are resolved involving collaborations between the various stakeholders in and around the different disciplines and communities of science. That is a mechanism is needed to foster these collaborations and introduce coordination (section 5). The national governments and funding agencies, but also the European Union will be key to turn these ideas into sustainable solutions, which is a fourth area to be discussed in section 6. The paper will conclude with some recommendations (section 7).

## 2. Data and documents

Some relevant differences between data and documents may be summarised in the following table.

<b>Primary data</b>	<b>Documents</b>
No copyright, but data policy limitations for commercial use	Copyright
Highly non-standardised	Much more standardised
Need special Representation Information (structure, semantics, software) to be understood and processed, but machine-readable	Human readable when displayed, but not machine-understandable
Digital volume huge	Digital volume modest or even small

Metadata frameworks not easily translatable	Frameworks for metadata exist
Very dispersed at present	In journals or archives with publishers, libraries
Business models that include storage and preservation almost absent	Several models for storage, access and preservation

There is no need to go into the details of these differences. The question is to which extent they will influence the issues of storing, preserving and providing access to the digital objects that represent them. The following sections will consider this question and come to the conclusion that for both a coordinated infrastructure is needed, that these infrastructures will considerably overlap, and that their organizational, financial and operational basics are quite similar.

### **3. Guiding principles for the legal and policy framework**

#### **I. Wide accessibility and availability of primary research data**

Starting with the primary data gathered from publicly funded scientific research the basic principle is clear. Research data should be accessible and available to the research community with as little restrictions as possible.

Such access will not only ease the validation of published findings but will also permit reconstructing the research event; reanalysis, reducing potential costly duplication and opening new research opportunities.

This principle is shared by both the research organisations and the publishers. Some of the former called for access to raw data in the Berlin "Open Access" Declaration. Similarly, in a joint statement the International Association of Scientific, Technical and Medical Publishers and the Association of Learned, Professional and Society Publishers recommended that "research data should be as widely available as possible". Access to research data evidently requires its curation and its preservation, and hence the need for coordinated policies. Among various initiatives aimed at achieving this aim, the exercise of the OECD drafting a policy framework may be particularly relevant. Following the Ministerial Declaration on Access to Research Data from Public Funding (March 2004), the OECD followed up in December 2006, with a Recommendation of the Council concerning Access to Research Data from Public Funding. This document spells the out the principles OECD Member States nations will try to fulfil and the guidelines will serve as the framework in which the issues of preservation of research data and their access will be tackled.

In addition, appropriate incentives need to be identified by the scientific community for researchers that make available and allow access and re-use of the data they generate in their research.

#### **II. Access to publications**

Wide and quick access to scientific publications is of course vital to the scientific process. The principle of wide availability of publications however is balanced

against the principle of observing the commercial rights of publishers who add value. In the digital world we are now living in (90% of all science journals are now available on-line) new solutions have to be found to implement this balance. Wide-ranging discussions are ongoing, about the reach of institution's licenses, on-line access for registered users at National Deposit Libraries, or about the price of consulting publishers' archives over time. Finding business models that in some way or other embody the principles of Open Access and the (mandatory? with embargo periods?) depositing of peer-reviewed or not-yet-peer-reviewed articles in institutional or community-based public archives are central issues. This paper does not deal with these important questions, as this is being done elsewhere.

There is, however, an issue that deserves consideration. Apart from the formal publications in peer-reviewed journals, scientific material may only appear as grey literature which is still one of the most important sources for disseminating information, knowledge and expertise, enabling research processes and offering information that is not found in conventional sources. The new technological environment is increasing the amount of grey literature: slides presented at conference and seminars are for example an emerging source of information. In addition, grey literature might be indispensable to understand and therefore re-use data. And as new business models emerge, more and more of the records of science may have characteristics of 'grey literature'. Finding solutions for storing and keeping accessible the digital records of science therefore requires taking into full consideration the various categories of grey literature which must be harvested, curated and made discoverable, and possibly linked with the related journal publications and experimental data.

### **III. Full compliance with intellectual property rights; digital rights management**

If access to data and publications must be wider, faster and easier, we must ensure that the legislation and guidelines that have been created to balance the individual, commercial, and public interests, recognizes the increasing public interest in enabling access and digital preservation, but still is an appropriate balance.

Funding agencies, including governments, and scientists naturally are interested in making the results of research, whether data or publications, as widely and freely available as possible for advancing science, societal applications and innovation through combining, searching, mining or re-using data, for example. In the web-based world this is even more so.

Several legal rights need to be observed, however, when providing access to scientific data or publications and procedures must be in place for dealing with the legal issues arising in relation to accessing and re-using research data. In the first place there are the traditional intellectual property rights such as copyright and patents or licenses. Copyright law, while protecting original selection, sequence and arrangement, has treated research data as such as lacking originality and has therefore classified it as non-protectable. By adopting Directive 96/9/EC on the legal protection of databases, the European Union has, however, introduced specific protection rights —a sui generis right — on databases against unauthorised reproduction, to safeguard substantial investment in the obtaining, verification or presentation of the database contents.

Certain members of academia and the scientific community, however, have

expressed concerns that the exceptions to the “sui generis” right might fall short of what is required with regard to the access to and use of data and information for scientific and educational purposes. This will need further consideration.

In some disciplines, the principle of “wide access to research data” may collide with the principle of confidentiality under which those data are collected, for example, patient records or other individually attributable demographic or economic records. In these cases data protection legislation may apply but other safeguard mechanisms will need to be put in place, such as making survey data anonymous for scientific use. Similarly, data not gathered with the express purpose of scientific research – for example government statistics – may pose some form of restrictions which need to be applied through a licensing regime. Paradoxically also uncertainty about ownership of rights relating to data deposited by researchers into a database or digital repository prevents access to these data.

Technical tools and various new legal mechanisms are being created to cope with the problem of asserting and safeguarding rights while maintaining an appropriate balance between different interests in the new digital circumstances. Rights management tools (the technical tools) are one important example. Science commons licenses another.

Such digital tools for managing rights and access should stimulate ease of access and expressions of rights and permissions. At the same time they must not be a barrier to preservation activities as may easily happen when for example technical protection measures are applied to deposited publications in legal deposit libraries.

Similarly, when it comes to science commons licences or similar approaches to enable data sharing, these should incorporate appropriate safeguards for reserved rights and citation for data creators. More generally, new legal mechanisms or changes in existing ones must be analysed on their impact on digital preservation. If that is the case their formulation and implementation should take into account preservation requirements to the maximum extent possible. A particular case arises when research data are used across various research sectors e.g. agriculture, biotechnology, health and medical research. Different regimes for accessing data may exist or be considered for come to exist such sectors, and further policies and good practice guidelines for data then need to be developed.

#### **IV. Preservation; selection; quality control**

Long-term preservation needs to be stated as a principle on its own. Maintaining access to older data and publications is increasingly of importance. More than 30% of requests for data from ESA/ESRIN on earth observation relate to old data and the number is rising. There are many aspects to this that have to be full addressed when implementing repositories and regimes governing their operations. One has already been mentioned: technical protection measures applied to deposited publications will become rapidly obsolete and then increase the cost of preservation or even prevent required preservation actions. Another aspect is that access can have a significant part in the preservation process by allowing independent validation and review over time by the user community of preserved material and the effect of preservation actions. For this reason “dark archives” i.e. without any form of access or with very restrictive access policies are seen as undesirable for long-term preservation by most experts.

Quality standards for content in terms of authenticity and provenance, and quality standards for services are critical to both access and preservation of research data. In recent years, much effort has gone into developing a consensus on criteria which might be used to assess those archives which reach a sufficiently high standard to qualify for formal certification as trusted digital repositories for long-term preservation. This so far applies to document repositories but much of the work is highly relevant for data repositories as well. A Draft Audit Checklist is now undergoing practical testing on four archiving repositories.

It is important to ensure that there are effective procedures for determining which data are of high value. It is clear that not all of the digital research data, and the same is true for publications, being generated will have long-term value and need to be preserved for the future. However a significant and growing proportion of it does. It is essential that decisions on selection for this preservation and curation form part of an organisational process and are not made on an ad hoc basis. Data and record management and information management policies are central to this and require co-ordination of policy at a high-level. This is increasingly being recognised and addressed by research funders, but as yet very few have fully formed data and information policies in place. Formal legislation, at EU or national level, of the requirement to curate and preserve data, whether primary or processed, in general would not seem the way to go. Universities, research organisations, research funding agencies and the large international science communities have, where appropriate together with major user organisations, to develop regulations, working guidelines and practices to address the different needs of distinct fields of science. Of course, in certain areas such a need for formal regulation, does exist and has already been recognised. One may think for example of data of clinical trials or related natural hazards. There is a need to verify, and possibly adapt, the existing regulations to fully take into account the requirements of a digital knowledge economy.

## **V. Sustainability**

Long term preservation requires sustainability of custody and business models. Long-term custody can be met in the public sector through repositories such as national libraries or national archives with legal mandates and established long-term roles. In the scientific research and commercial sectors organisations normally have renewable fixed term reviews or are subject to market forces. In such cases they may not be able to provide indefinite custody or access. Sustainability will require such organisations to consider arrangements for transfer of custody or access management to other approved organisations in appropriate circumstances, as for example happens in relation to electronic journals where some national libraries are backing up publishers on the basis of licence clauses providing customers assurance of continued access.

The issue of economic sustainability of digital preservation has been object of several studies aiming at designing suitable models for the estimation of the corresponding costs<sup>1</sup>.

The description of different models demonstrates that even in rather homogeneous

---

<sup>1</sup> Among these deserve a special reference the projects ESPIDA (<http://www.gla.ac.uk/espida>) and LIFE (<http://www.life.ac.uk>). Also noteworthy the cost modelling exercise developed by the NASA for their long term data management activities.

domains it is extremely difficult to establish the digital preservation costs simply as a percentage of the overall research budget. Digital preservation experts highlight values ranging from a few (1 or 2%) percentage points to even higher. The lower number corresponds to research activities which rely on very expensive infrastructures (particle accelerator, satellite), while the higher percentages refer to research activities which involve lower investments in infrastructures such as those in humanities and social sciences. These figures are substantially higher than the current investments in preserving research data.

## **VI. Designation**

Legal deposit, i.e. the obligation for content producers to make one or more copies of scientific materials available to a designated deposit body, is a central issue for the preservation of digital scientific publications. Member States have started to extend deposit arrangements to digital information, at different speeds and with different types of information covered.

Some national research funding bodies have also designated and funded national data archives and data curation centres, and require research data generated in projects they have funded to be offered for deposit to their designated repository. There are now well developed networks of such national data archives in Europe in areas such as the social sciences. Other data archives and data centres have been organised on a European or Global level. Designated places of deposit are also a feature of national archival legislation and can cover scientific datasets generated by government.

Designation can be a valuable tool in preservation in providing a statutory function for a repository and a funded mandate for its work. In some cases designation has no associated funding attached but the prestige and branding support the work of the repository.

### **4. The need for a common framework**

There is an increasing awareness throughout the scientific community, research organisations and policy makers on the importance of improved access to research data and publications, on curation and preservation, and on the need for coordinated policies. There is a need for an overall general and common framework sensitive to stakeholders needs. Organisational issues, technical needs and legal and policy issues could be the thereof main areas addressed by this framework.

Many different key organisations from the worlds of science and science information have been working to create such a general framework that is based on their own experience and of national coalitions that have been leading in setting up practical experiments<sup>2</sup>.

The key building blocks for laying this sustainable virtual infrastructure across Europe would consist of:

- establishing a network of repositories;

---

<sup>2</sup> Reference is made to the Strategy for a European Digital Information Infrastructure published by the European Task Force Permanent Access ([www.alliancepermanentaccess.eu](http://www.alliancepermanentaccess.eu))

- developing shared services and tools;
- promoting certification and standards;
- demonstrating increased usability and added value from long-term retention and preservation of research data;
- delivering collaboration between the various stakeholders;
- promoting development of relevant policies by funding agencies and research organisations;
- and fostering relevant education, and skills.

A sustainable digital infrastructure is essential for transforming Europe into an information society: preserving and developing long-term access to this record can largely be viewed as part of the public infrastructure for science.

The network could consist of many types of repository. Some may be institutional, discipline, national, or international in their remit; they may have differing funding models and missions that may make some transient because they are project or business based, while a smaller number may be permanent repositories supported by government or other public bodies. Networks can promote agreements on long-term preservation and citation, replication, and inter-operability, between them and other stakeholders.

## **5. Implementation issues**

Implementation of such a framework faces many hurdles. They relate to organisational aspects, choices with respect to standards and technical solutions, capacity building and skills, legal aspects and not least financial aspects.

One key organisational issue mentioned in the previous section concerns the question who takes responsibility for data and publications: will it be national libraries, research institutes, subject repositories or institutional repositories. In other words is there going to be a centralised model, or widely distributed solutions with probably many overlapping collections? Following the way the process of science is organised would suggest a 'community-based' solution in many cases. Per community or main disciplinary area stakeholders (scientists, current research databases, funders, publishers, librarians,..) should set out to agree on the repository infrastructure, metadata and so on. In such an effort, international research organisations or professional associations such as ICSU, ESF, ISO etc should play a role along with information stakeholders such as libraries and publishers.

Important issues have to do with the information stored. For example, choices need to be made with regard to standards for metadata, storage and retrieval, and the appropriate roles for each of the stakeholders in this process. A different question is what links will need to be created back to the primary literature; who will do this; how will links be monitored and repaired. This is becoming rapidly an issue as more and more journals require or at least encourage authors of papers to deposit underlying data in a recognised appropriate repository e.g. the Protein Data Bank. A link between publication and data deposit is immensely helpful in streamlining workflows for the researcher and encouraging deposit.

Building digital repositories, developing tools to support different preservation strategies, managing complex and dynamic datasets or developing life-cycle

costings and value-chain analyses requires more research and development. But there is a need to develop a coordinated approach to such work which so far has been absent.

Capacity building should not be underestimated. Traditional libraries and archives undergo major transitions; the new community and institutional repositories are just being built up; working methods and technologies are in the process of being developed. As a consequence the need for training in these new skills is large, and not only for a small category of 'repository professionals': scientists at large need not only become aware of the importance of storing and preserving data, they will need some general skills as well.

Digital rights and access management tools need further work. Also extensions to the data realm of deposit obligations may be considered, though good practices rather than formal policies, and as consequence guidelines from research funders and universities and research institutes, would seem to be the way ahead. The exceptions apply to the same areas that were singled out in section 2 from the overall principle of unrestricted access. It is however, critical that the European scientists who are being funded by different sources are not confronted with different regulations and procedures regarding accessibility and preservation. Research funding agencies, research organisations and higher education institutions in Europe are slowly starting to coordinate their activities and (sometimes) policies in this regard. Examples include the networking of major national research organisations (both research funding and research performing) being facilitated by ESF. The EU could play a clear role in ensuring that there is sufficient coordination between these frameworks, starting by encouraging the beneficiaries of its research funding to actively engage in making sure that data is deposited in repositories or data centres in such a way that accessibility and preservation are maximally assured.

A very critical issue is who pays for it and through what business models (if any; for example for primary data no clear business model seems to exist for recouping the investment and running costs for a repository, other than the funding agencies coming in). The re-use of data, which requires of course data archiving, for which in e.g. the social sciences, environmental sciences or genomics practical implementations are already very advanced and of long-standing, may offer useful practice and indicators for other research areas, including shedding some light on cost issues. The three UK research councils who maintain data centres/services for archiving/re-use (the Arts and Humanities Research Council, the Natural Environment Research Council, and the Economic and Social Science Research Council, for example, all spend 1.4 -1.5% of their research budget on these services. This indicates that data archiving and access are not necessarily onerous costs in overall research budgets, but they do require permanent financing. Eventually, there will be no escaping the conclusion that storing data is part of the scientist's job, and that funding the storing and preservation of the records of science will be largely part and parcel of the regular funding mechanisms for science.

It is clear that creating a digital information infrastructure for the records of science requires the active involvement and collaboration of large research organisations, science funding agencies, science publishers and for example national deposit libraries. And it must be done in concertation with the development of European and national policies for access and preservation. It is

worth noting the recent establishment of the Alliance for Permanent Access, which comprises precisely the sort of organisations mentioned. As it supersedes particular disciplines or areas of science it may very well provide a useful way to coordinate the various efforts in more dedicated projects and plans.

It is important to note that coordinating the efforts of key stakeholders and national and European policymakers is becoming urgent: the US first took the lead towards more coordinated approaches with the National Digital Information Infrastructure and Preservation Programme led by the Library of Congress. More recently an Interagency Working Group on Digital Data has been established to create a robust infrastructure for preserving the data of science, and the National Science Foundation has made the creation of this DataNet a cornerstone of its March 2007 Cyberinfrastructure Vision for 21<sup>st</sup> Century Discovery. A major funding initiative has now been launched, posing a challenge for Europe not so much to compete but to organise itself and join forces because eventually the data infrastructure will be a global one.

## **6. Areas of current and possible future funding at EU level**

The EU funding instruments will accommodate funding a series of important activities concerning access and preservation. In the ICT theme of FP7 there is some scope under "Digital Libraries and Content".

The Research Infrastructures part of the Capacities will explicitly support in a coordinated way scientific digital repositories, curation etc to organize the data repositories for science<sup>3</sup>, in addition to the previous incarnations of ICT based e-infrastructures (data communication, high performance computing, GRIDs). Thirdly, the 2010 Digital Libraries initiative provides for some funding under the eContentplus programme.

Yet, there is a need to enhance the coordination between the activities which will be funded under these various schemes.

Neither is the spectrum of activities needed for a well coordinated approach completely covered. Several areas have been suggested for complementary funding to ensure effective coordination in addition to R&D activities. They concern prototype services, projects and supporting actions, such as:

- There is a need to work with a number of pilot communities to identify the key repositories, both for documents and for data, for these communities, and to identify which common metadata would be essential for inter-operability and appropriate for the characteristics of the scientific fields involved. Also, some validation activities need to be funded for these communities.
- Standards need to be promoted, and a network of providers of certification services needs to be built up. On top of this a common European accreditation organization or mechanism needs to be established.
- A few large scale pilots need to be carried out, and for these benchmarking and evaluation studies are necessary to demonstrate increased usability and added value from long-term retention and preservation of selected materials.

---

<sup>3</sup> See for example DRIVER. (<http://www.driver-repository.eu/>)

- Outreach activities consisting of training, information exchange and knowledge transfer to member states where so far little activities in the area of preservation of scientific information are taking place are crucial.

## **7. Conclusions and Recommendations**

- I. The digital repositories in which the record of science is collected and managed encompasses publications but increasingly primary and processed data. Indeed, linkages between publications and data are proliferating. The ability to use and re-use these data as freely as possible is key to innovation and the further advancement of science.
- II. A general (policy) framework, including sustainable custody and funding/business models, needs to be established by the key stakeholders in science and science information and national and EU policymakers to establish the roles and responsibilities of these in building a European Digital Information Infrastructure that allows the access and re-use of research data and ensures their long term preservation.
- III. Considerable efforts and money will be required to build up appropriate infrastructure. Given the increasingly importance of the preservation of research data, research funding organisations at national and European level, should consider dedicating increasing investment of research budgets to research data preservation efforts.
- IV. We believe that the EC should build on existing initiatives. The Alliance for Permanent Access may be a good starting point.