The International Journal of Digital Curation
Issue 1, Volume 1 | Autumn 2006

Key Stakeholders Pledge to a Strategic Approach to Preserve the Digital Records of Science

Peter Tindemans,

Chairman, Task Force Permanent Access

Abstract

An overview of how the hopes offered by digitisation turned into a problem when the explosion of formats and new technologies began to render material inaccessible and the way in which the Chairman sees the Task Force movng to address the challenge of preserving and accessing the record of science in the long term.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

How Do We Convert a Solution, Which Turned into a Problem, Back into a Solution?

As soon as computer scientists and electrical engineers started to create generation after generation of computer hardware and software, they realized that there was a problem in making the new generations compatible with previous ones. Indeed, it did not always work out as intended. The few scientists in those days who applied computers to their work ran into similar problems of maintaining access to their data and making their programmes work on new hardware. However, it was not something that featured high up on the agendas of either their institutions or indeed their governments. Preservation, however, did attract attention. Libraries, in those days still firmly with their feet in the paper world, grappled with acidification, and projects were devised both to retrieve their documents on microfiches and establish ways of treating the crumbling books, magazines or journals to stop the process of acidification. More generally, preservation of the world's cultural heritage did become an item on political agendas. The arrival of the digital world was therefore quickly hailed in the communities of libraries, archives and other cultural heritage organizations as manna from heaven, the solution to many problems. But not for long.

The massive use of digital storage and computing methods quickly led people to realize that they faced completely new challenges. The sheer volume of digital data, the bewildering variety of formats and digital objects, which sometimes turned up in fast-changing sequences of new versions, soon became a cause of great concern. Deposit libraries are often legally obliged to safeguard and keep accessible publications, be it on paper or born digital, for future generations, perennially. National archives face comparable requirements. Broadcasting companies, media production companies and museums have invested heavily in their digital collections and want to re-use them. So does the public at large. The story involves science so far purely through its output in articles, magazines or books; or the cultural heritage domain through the recording of its development as a historical, social process, again in documents or audiovisual media. But scientists and scholars, too, have for many years realized that the data they collect in experiments, observations, surveys, simulations and databases of historical or sociological events also need to be maintained, and not just in the processed form represented by their publications.

Vastly different as the scales may be, ranging from the famous OECD economist Madison's database of historical Gross Domestic Product data to the huge amounts of data generated in space science (next to the collections of images from space), earth-bound astronomy or particle physics, many of the problems encountered are similar. Neither are they limited to the natural sciences and engineering. Medical and biological sciences, environmental sciences and social sciences and humanities are equally affected. Pharmaceutical companies, oil and gas companies, meteorological and geological services are in no different a position. The last ten years or so have therefore seen many efforts - scattered efforts for the most part - to address the problem not only of building digital libraries, archives and repositories, but also of developing conceptual and technical tools to help ensure long-term preservation of and access to the digital data stored.

This is where people turn once more to the solution that originally had turned into a problem to find a way out: can we really find an 'architecture', a way of imagining the world, and the related ICT solutions to help us come to grips with the problem of preservation of and access to the digital records of science and the world's digital heritage in general?

The Task Force Permanent Access

The EU Conference "*Permanent Access to the Records of Science*" organised by the Dutch national library, the Koninklijke Bibliotheek KB, and the Netherlands government as President of the EU Council on 1st November, 2004 in The Hague, was a deliberate attempt to raise the awareness among stakeholders and politicians of the importance of the issue of permanent access. It also deliberately brought together the worlds of libraries, archives and science. The participants of this conference agreed that Europe should create an infrastructure for the long-term preservation of and permanent access to the record of science, and the KB was urged to take the initiative to create a European Task Force to drive things forward quickly. As a result, a high-level Task Force was established in early 2006 comprising persons, usually at board level, from major libraries (e.g. the British Library and the KB), science organizations (such as the European Science Foundation, CCLRC, ESA/ESRIN, Max Planck Institute, or the Hungarian Academy of Sciences), an archive (Sweden) and the Association of Scientific, Technical and Medical Publishers.

The Task Force set out to compile an outline for an R&D Programme in the field of long-term preservation of and access to the digital scientific record, to devise a framework, i.e. the common elements of an infrastructure for preserving and accessing digital data, and to propose concrete steps to start building this infrastructure. Its two outputs, the R&D Programme and the Strategic Action Programme 2006-2010 of December 2005, including its full composition, are to be found at http://tfpa.kb.nls

The Challenge of Preserving and Accessing the Record of Science in the Long Term

Record of science is shorthand for a very broad and complex notion. It covers in the first instance all fields of science and technology. It comprises the traditional ways of communication in science, that is books and articles, but it is increasingly important to include the less structured informal communication mechanisms such as websites and bulletin boards. The actual data need to be 'curated' too: these may be critical primary datasets as well as datasets aggregated into higher-level digital objects. For primary datasets to be re-used, to do secondary analysis or to check later results against previous ones, 'lab books' about the way the data have been obtained need to be preserved as well. But even then, of course, there is much contextual information, in the form of knowledge of the persons involved or otherwise which is very hard to document in such a way that experiments could really be repeated exactly. So choices have to be made partly on the basis of affordability, and these choices will probably vary for different fields of science. Observational data to be used primarily for operational services, such as meteorological data, would be included in the records of science. For practical reasons the Task Force decided to concentrate on public domain data and to exclude military-related data and medical data where ethical and privacy aspects create special circumstances.

The challenges for the long-term preservation of and access to this record of science are manifold. There are important *technical* ones. Partly it is the sheer volume of data to be stored and provided access to. Digital data are easy to alter, too – hence the issue of authenticity. Nowadays one needs to provide access to integrated repositories combining data, documents, multimedia presentations of very complex and increasingly dynamic datasets. Data are also perishable because of changes in hardware and software. Conversion or migration into a new format and onto new hardware is one way of addressing the latter problem; another is emulation of old technologies by present-generation software and hardware. But more approaches need to be investigated: the universal virtual computer is based on the assumption that there exists a number of elementary characteristics of future generations of software and hardware with which a sufficiently 'bare' virtual computer would be able to cope. For some the ultimate hope is to find universal and ultimately self-describing ways to explain formats.

The challenge is also *economic*. It is one thing to estimate the costs of preservation; quite another to estimate the value of preservation, based on the reasons why the records of science should be accessible, and then linking this to whom this value could be attributed in order to allocate costs. However, in many respects science remains what economists call a public good: everyone, not just its producer, reaps the benefits. Hence there must be a significant public role in creating and maintaining the infrastructure for long-term preservation of the records of science. That is, the costs will have to remain largely part of the normal funding models of science.

The management of rights and access is a third area where major challenges lay ahead. Although public accessibility of research data acquired through public funding has now been established as a principle, this does not mean that we will not need arrangements for rights and access management. Furthermore these arrangements eventually need to be implemented in a networked environment by developing technical, and not just physical solutions to a technical problem. The interests at stake are immense and go well beyond the boundaries of science, extending into the realms of entertainment, broadcasting and publishing in general. Ongoing discussions at national and EU level illustrate the complexity of the issues involved.

- To address these issues systematically the Task Force has outlined an R&D programme with six focused themes:
- Developing and deploying technical tools to support different preservation strategies;
- Representation information, registries and object identification;
- Managing complex dynamic datasets and databases;
- Developing distributed archives and network solutions from a preservation perspective;
- Devising new approaches to the development of IT solutions from the perspective of the durability of information;
- Developing life-cycle costings, value chain analyses and other economic models, as well as digital access management guidelines and tools to support sustainable long and very long-term preservation and access.

An Infrastructure for Long-term Preservation and Access to the Records of Science

Evidently creating one big repository with the best possible technical tools to preserve and provide access to the output of science will not do. Equally, relying on individuals, universities and research and other organizations to establish their own repositories, and just supporting the development of common tools, would represent, at the least, a gigantic duplication of effort and almost certainly fail to lead to the desired levels of accessibility. The clue as to a solution that warrants a large degree of commonality while at the same time allowing for diversity where necessary, lies in the way science is produced and communicated.

To a large extent this occurs in international communities, disciplinary or interdisciplinary, such as the particle and nuclear physics community, the astronomy and space science community, the fine analysis of matter community, the social sciences community, the life sciences community, the meteorology and climatology community, or the historical research community. More intra- than inter-community communications and the existence, in many communities, of strong focal institutions (research facilities, data archives, etc) are characteristics that are significant to the present discussion. CERN and the European Social Sciences Data Archives (ESSDA) are good examples, though totally different as yet in terms of resources and actual support to their communities. Data and information services for a community are generally provided by laboratories, specialised database providers, e.g. protein or gene information or specialised software for social science research, specialised publishers or web-based archives and specialised libraries, all moving away from the role individuals often played previously. There are 'horizontal' providers as well, cutting across communities: general scientific publishers, deposit libraries, university libraries, general archives and so on.).

The infrastructure the Task Force has in mind would in the first place consist of a (small) number of core community subject repositories, to be identified in discussions with the particular community. In addition there is a limited number of 'horizontal', cross-cutting repositories (as an example one may refer to the KB with its back-up arrangements with several large scientific publishers). Of course, institutional and national repositories do exist but usually for different purposes: accountability, public relations, output metrics, education to name but a few. To ensure interoperability, long-term preservation and access, quality of service, cost-effectiveness and reduced investments in resources and time, a number of common enabling mechanisms and conditions need to be established as part of the infrastructure. Repositories need to be Open Archival Information Systems (OAIS)-compliant. A common framework for metadata (i.e. a common basis allowing for variations per community), for persistent identifiers and a number of registries are required. A set of cost-effective preservation methods and services need to be developed and made available. Next is a common framework of principles and guidelines for managing rights and access. There should be a mechanism, which is very largely a financial incentive mechanism, to develop and test implementation tools, techniques and services for registries, preservation, authentication and authorisation, enforcing rights management etc. Finally, a crucial role has to be played by certification service providers to test tools and audit repositories. Hence a common European accreditation mechanism must be developed.

A strong basis for such an infrastructure, conceptually and practically, is now available. Examples are the OAIS reference model, the Dublin Metadata Core Initiative and the MARC21 metadata scheme of the Library of Congress, or the Draft Audit Checklist for Certification of Trusted Digital Repositories which has been developed by the US Research Libraries Group, NARA, plus a number of European experts. As regards practical developments, one may mention the work of national coalitions, such as the Digital Preservation Coalition in the UK or Nestor in Germany or the DARE project in the Netherlands. Several EU-funded projects have been carried out, though scattered and with a rather large focus on co-ordination. Some important new ones are on the horizon such as DRIVER and CASPAR. The Audit and Certification of Digital Archives Project of the Council for Research Libraries to test-audit three archives, is another.

An Alliance for Permanent Access

Addressing these digital challenges will require new collaborations and co-ordinated action by a range of European and national organizations. To ensure a genuine co-ordinated approach resulting in a sustainable virtual infrastructure commitment at 'board level' by several major stakeholders from the worlds of science, libraries and archives, publishing and communications will be essential. In addition governments at national and EU level must understand that vital European economic, scientific and cultural interests are at stake. Working with committed key stakeholders is the best way to avoid the dangers of either too political an approach or one which will be too diffuse by focusing on a multiplicity of nationally or EU-funded projects. Indeed, the example of the US where Congress made available $100 million to create a National Digital Information Infrastructure and Preservation Programme co-ordinated by the Library of Congress which collaborates with a number of strategic partners, should point the way ahead for Europe. Agreeing on a common Europe-wide framework and committing resources to accelerate and catalyse in a relatively short period of time collaborations in and between a number of key communities are the pillars of what now needs to be done.

Task Force member organizations therefore intend to establish an Alliance for Permanent Access to the Digital Records that aims to:

- establish a wide consensus on a strategic level among the major stakeholders as to the main characteristics of a European virtual 'infrastructure' for long-term preservation of and access to digital data, with an initial focus on the records of science;
- accelerate significantly the creation of the main building blocks of this 'infrastructure';
- work with industry vendors, national competence networks and coalitions, and international partners in science to ensure the technology, skills, and standards to underpin this infrastructure are in place;
- work with national governments and the EU to strengthen European strategies and policies and their implementation in the area of long-term preservation of and access to digital data, and thereby contribute to Europe as an Information Society;
- strengthen the role of European parties in the world-wide efforts in the area of long-term preservation and access;
- build, articulate and maintain a continuing R&D programme.

Its founding partners should consist of a critical mass of major stakeholders among national and European research organizations, libraries, publishers and archives. A work programme has been drawn up that focuses on:

- establishing a network of repositories;
- developing shared services and tools;
- promoting certification and standards;
- demonstrating increased usability and added value;
- delivering collaboration between the various stakeholders;
- and fostering relevant policy, education, and skills.

The Alliance will set up a small office to co-ordinate and act as a catalyst for collaboration and advocacy on behalf of the partners. It will not be a “talking shop” but a practical catalyst and tool for its member organizations and stakeholders to achieve shared aims. It will work closely with user communities to help them establish the physical repositories they need and the mechanisms for these to interoperate in a Europe-wide infrastructure. It will not seek itself to act as a funding body nor to be a distributor of funds (the Task Force seeks to establish a €100 million fund, partly from FP7 for the proposed research agenda, partly from other EU funds to leverage national funding), but its partners will work together, and with national, European, and international programmes. The Alliance will be established in principle for a 3-year period within which the key building blocks for the infrastructure will have to be put in place and sufficient momentum and mechanisms generated for further activities to become self-sustaining.

Circumstances in Europe are beginning to look more propitious than in the past now that the activities of the European Strategy Forum on Research Infrastructures (ESFRI) will probably lead to a more strategic and structured approach towards research infrastructures in Europe. That should be the ideal starting point for engendering the political and financial support, both at national and EU level, for the work the Alliance would like to generate for providing and keeping access to the records of science and digital heritage more generally.