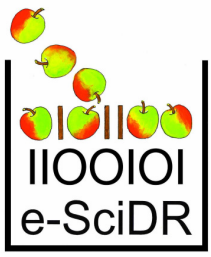
# Towards a European e-Infrastructure for **e-Sci**ence **D**igital **R**epositories

a report for the European Commission

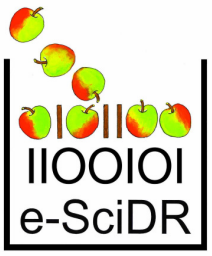



*Harvesting and seeding the fruits of e-Science*

e-infrastructure

The e-SciDR study is funded by the EU's Sixth Framework Programme and led in the Commission by GÉANT and eInfrastructures unit of DG INFSO.

European Commission
Information Society and Media




# Towards a European e-Infrastructure for **e-Sci**ence **D**igital **R**epositories

Project reference no: 2006 S88-092641

# Final Report

for

DG Information Society and Media

Unit F – GÉANT and e-Infrastructure

**Prepared by:**

The Digital Archiving Consultancy Limited
2 Wayside Court
TWICKENHAM
Middlesex
TW1 2BQ
United Kingdom
www.d-archiving.com www.e-scidr.eu

# Preface

Innovation, the foundation of economic development today, depends on rapid scientific advances. Science for its part has become increasingly based on open, cross border collaboration between researchers the whole world over. In addition, modern science is a heavy user of high capacity computing to model complex systems and to process experimental results.

The emergence of new research methods that exploit advanced computational resources, data collections and scientific instruments, in other words *e-Science*, is poised to revolutionise the future scientific discovery process, as the "Scientific Renaissance"[1] did in setting the base of modern science. It is crucial for Europe to embrace the underlying paradigm shift in order to keep its competitive position and to respond to societal expectations.

To enable the fast transition towards e-Science, the European Commission and Member States have made significant investments in *e-Infrastructures*, including the pan-European research network GÉANT, e-Science grids, data infrastructures and supercomputing.

Striving for world leadership in e-Science, establishing e-Infrastructures as a sustainable utility and exploiting them as a factor of innovation are the three vectors of a renewed European strategy supporting the ground breaking science of 2020 and beyond. This strategy requires a significant step forward in terms of type and intensity of investments, better linking of research and innovation policies and coordination of national and Community strategies.

Against this background, the present Communication has three main objectives – to *highlight* the strategic role that e-Infrastructures play in underpinning European research and innovation policies, to *call* on Member States, the European Commission and the scientific communities for a reinforced and coordinated effort that foster world class ICT infrastructures, and to *provide* for a renewed strategy against which specific actions and investments can be deployed.

DG Information Society and Media
Unit F – Géant and e-Infrastructure

[1] Marie Boas Hall, The scientific renaissance, 1450-1630 ISBN 0486281159

# Contents

# Acknowledgements

This study was made possible by the contributions of many people.

The e-SciDR team at the European Commission have been immensely supportive and helpful, for which we would like to thank them, in particular Carlos Morais-Pires, Mario Campolargo, Elina Zicmane, Krystyna Market of DG INFSO and Celina Ramjoué in DG Research.

We would also like to extend our sincere thanks to the National Archives of Portugal at the Torre do Tombo in Lisbon and their staff for hosting the study's closing workshop, held in Lisbon during Portugal's Presidency of the European Union, and for all their help in making the final study workshop such a rich and valuable day. In particular we would like to thank Dr Francisco Barbedo, Deputy Director of the National Archives, and his colleagues Mrs Lurdes Henriques and Mr. Miguel Veloso. Our very special thanks go to Maria Jordão at UMIC, Portugal's Knowledge Society Agency, for all her suggestions and support in co-ordinating the Lisbon workshop. The Lisbon workshop was opened by Dr. Pedro Ferreira of UMIC with an inspiring address of power and clarity on the importance of e-Science digital repositories in a European knowledge society. To Herbert Van de Sompel who kindly agreed to present the key-note address and to Jens Vigen who delivered the closing address go our grateful thanks for their presentations on their important work. (Links to the digital versions of their presentations are given in the bibliography in part 2.) Our thanks also go to Andrew McHugh of the Digital Curation Centre.

Lastly, we would like to express our deep thanks to all those experts, and their institutions and sponsors, who gave their time to attend the various workshops we conducted, and others whom we interviewed during the course of the study; we list them in part 2. We thank all those across Europe and beyond who contributed so fruitfully to the public consultation we conducted.

# The e-SciDR project team

The e-SciDR study was conducted by a consortium of organisations drawn on a pan-European basis; the lead organisation was The Digital Archiving Consultancy Limited (DAC), based in Twickenham, UK. The team leaders were Alison Macdonald and Philip Lord of the DAC.

The following lists all the consortium members and their main contributing staff:

| | | |
|---|---|---|
| Digital Archiving Consultancy | The Digital Archiving Consultancy Limited (DAC)<br>Twickenham, UK | Alison Macdonald<br>Philip Lord<br>Damian Counsell<br>Melanie Dulong du Rosnay<br>Isabel Galina |
| Charles Beagrie | Charles Beagrie Limited<br>Swindon, UK | Neil Beagrie<br>Daphne Charles |
| gridwise tech | GridwiseTech, Sp. z.o.o.<br>Krackow, Poland | Pawel Plaszczak<br>Krzysztof Wilk<br>Pawel Jarosz<br>Paul Pillar |
| University of Glasgow | University of Glasgow<br>National e-Science Centre<br>Glasgow, UK | Prof. Richard Sinnott |
| com'tou | Com'tou Sàrl<br>Paris, France. | Hanne Mostecky |
| Imperial College London | Imperial College London<br>London, UK | Prof. John Darlington<br>Dolores Iorizzo<br>Brian Fuchs |

# Presentation and other study documents

We begin the final report with a short presentation on core terms and a summary of the e-Science digital repositories landscape in Europe. We then lead straight into the study's recommendations, prefaced by a vision for these recommendations. These are followed by short sections on standards, technologies, the public consultation and a summary on legal issues relating to open access to e-Science digital repositories.

## Other e-SciDR materials

The e-SciDR study generated other materials, in particular two Interim Reports. These reported on core definitions, stakeholders, reflectors and studies on the area, relevant technologies and standards. They provide a full report on the study's three initial workshops, the public consultation, and presentation of the landscape of e-Science digital repositories in Europe, and discussed legal issues relating to open access and e-Science digital repositories.

These and other documents, including case studies and sets of links, are available on the project web site, www.e-scidr.eu, or from the Digital Archiving Consultancy.

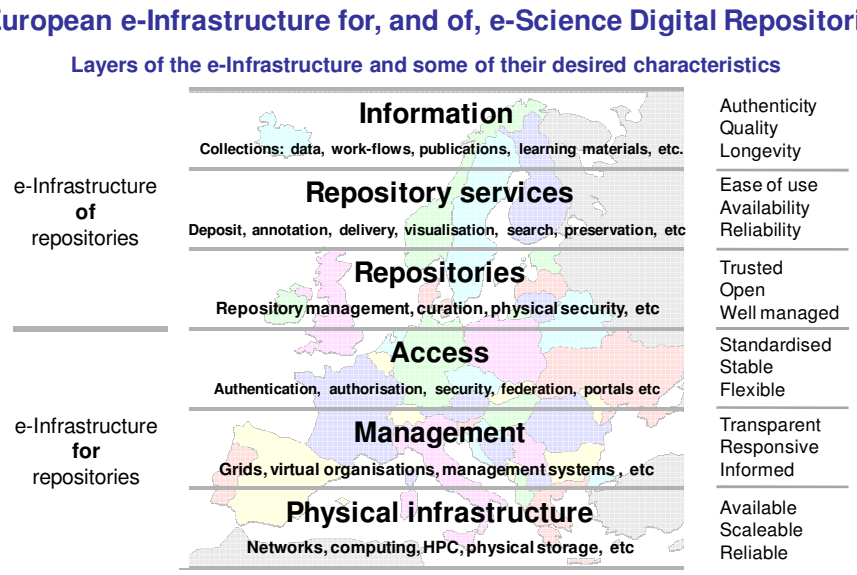
# Executive summary

The primary output of science is information. This information contributes to economic development by driving the development of products and services; it increases social welfare and improves public health. Scientific knowledge is itself a core part of our cultural heritage. Further, information is the vital link in a virtuous circle, being the feedstock of further research.

Historically, Europe has the finest tradition of innovation and discovery in science, continuing to the present day. This pre-eminence has been challenged over the last few decades. We must therefore look after our scientific information in Europe not only as a precious resource in itself, but also as a strategic and competitive resource.

Overwhelmingly information is now kept in digital form, and the care of this information is therefore entrusted to digital repositories of many kinds. Like the libraries and archives that traditionally care for paper-based records, these repositories need to offer a diverse and essential set of services beyond their basic remit of storage, such as providing deposit, access, searching and visualisation tools. These are supplemented with information-age infrastructure elements, such as semantic standards, specialist query and visualization tools, preservation services and elements which sustain critical characteristics of the repository materials: their integrity, authenticity, usability, and their ability to be understood and discovered.

To derive greatest benefit from the materials in repositories, a state-of-the-art ICT infrastructure is fundamental – including high-performance computing (HPC), fast networks, storage, access and management structures. Surveying repositories and infrastructure together, a wider vision emerges: a European e-Infrastructure **for**, and **of**, e-Science digital repositories:




Over the last few decades, the power of information and communications technologies has soared, vastly extending and accelerating reach and access to repositories and tools to use their contents. Over the same time, instruments and devices have proliferated, grown in power and many have become more affordable. So we have seen a vast increase in the amount of data generated and captured – raw data from sensors, instruments, surveys; processed data, analyses, information in the form of studies, articles, and data recording the management of the scientific process itself.

The combination of the power of ICT and the availability of repositories of vast quantities of information has had an enormous impact on the conduct of science and on scientific information. From such developments, "e-Science" has emerged, a term for new ways of conducting science: collaborative, computationally intensive, with the ability to work with massive volumes and data from different sources and diverse subject domains. New "computational laboratories" have been enabled, performing new science by working on existing data. Repositories are the constructs which hold the data, and around which supporting services and tools are provided.

### Background to the study, how conducted

The e-SciDR study was conducted by a consortium of expert organisations lead by the Digital Archiving Consultancy Limited, for DG Information Media and Society of the European Commission. The objectives of the study were, in brief:

A. To provide a reliable overview of the situation in Europe concerning e-Science digital repositories of e-Science information, data and knowledge.

B. To address policy options to encourage the development of e-Science digital repositories to provide low-cost open access to e-Science data and learning resources, considering multiple aspects: standards, technologies, stakeholders, previous work, and legal implications.

C. To provide recommendations and define development scenarios for European-wide efforts to develop e-Science digital repositories for research and education.

All data types and all scientific disciplines (in the wide sense, from the arts to physics) were considered. The programme of work undertaken was to:

- Conduct three workshops with European experts to consider different aspects of repositories.
- Undertake extensive research into the repository situation in Europe and the wider world from multiple perspectives, to draw a landscape of the repository situation in Europe.
- Conduct a public consultation to elicit the views of repository users and other stakeholders on the use made of digital repositories, the barriers to and enablers of their use.
- Conduct a final workshop of invited experts to consider findings and emerging policy directions, followed by the drafting of the final reports.

## Headline findings

The findings and recommendations from the study are set out in the final e-SciDR Report, and in supporting papers (Interim Reports 1 and 2). We summarise below the study's twelve mutually reinforcing **recommendation sets** for policy measures to drive towards a European e-infrastructure for and of e-science digital repositories.

The repository landscape is complex and diverse. Repositories come in many forms and sizes –data centres, archives, data warehouses, databases, and many more. There is diversity over many parameters: types of data; single and combined disciplines; organisational settings and structures; the scope, variety and sophistication of the tools, interfaces and services provided; commercial or open access. Some collections are distributed (accessed, for example, through portals), others are local. The landscape is confusing and obscure to the average user, and lies in a complex matrix of technologies, facilities and unfamiliar and fuzzy terminologies: e-Infrastructures, Grids, web technologies such as Web 2.0, "SOA" (service-oriented architectures), the semantic web, and so on. The scientific data held in the repositories is usually specialist, heterogeneous, complex, and difficult to use for the lay person. Europe boasts many e-Science digital repositories and services which make their holdings easy to use, provide search, query and visualization tools, and a range of supporting infrastructural services, from storage and database optimization, thesauri and curation to community

work establishing and maintaining standards (computational, semantic), for interoperability. A huge amount of work is being done to create a rich information space enabled by information technology, by libraries, information scientists, from all sectors. This work vastly increases users' productivity and the quality of the science.

## Recommendations

### R1. Funding reform

The most strongly expressed need was for funding for e-Science digital repositories which is aligned to their role as sustained digital custodians and providers of tools and services, efficiently managed.

- We recommend funding for e-Science digital repositories that is specific to their role and function as digital repositories. This funding should be stable and rolling, matching the duration of the repository's role or of its holdings.

- The funding should be sufficient to support and maintain the repository holdings, individually and as collections, to provide quality services and support to users, and provide good management at repository level, that can deliver continued, efficient, rich, easy-to-use access to trusted, quality materials. This will entail the provision of funding which extends beyond the repositories themselves.

### R2. A European e-Infrastructure *of* and *for* European e-Science digital repositories

The areas of e-Science, repositories, e-Infrastructures consist of multiple layers, domains, dimensions, as well as nations.

- We recommend considering a co-ordination framework at European level to bring together e-Science repository and services providers, users, experts from the different scientific disciplines, and from different areas of professional expertise, to identify commonalities, opportunities for sharing of expertise and synergies, in the area of e-Science digital repositories (and of e-Infrastructure).

There are opportunities for pooling expertise and facilities across Europe to support e-Science repositories and services, to strengthen European science, nationally and internationally and address obstacles to European e-Science – fundamentally collaborative in nature - caused by fragmentation. This co-ordination would strengthen European science, nationally and internationally and address obstacles to European e-Science – fundamentally collaborative - through fragmentation.

### R3. Support for data producers

The work of repositories and the quality of their holdings will be substantially eased and increased with the provision of good-quality data at the outset – that is, data that is well described, and conforms with relevant standards, supporting discoverability, interoperability and usability.

- It is important that institutions maintain policies and measures which insist on and support good data planning and management by data producers. These policies and measures should be accompanied by corresponding adjustments in funding.

### R4. Discovery and navigation

- Research and investment are needed into easy-to-use tools and frameworks for discovery of repositories, their holdings, collections, and for navigation between, within repositories; also between data and publication, both forwards and backwards along this information chain. There is a need for registries, and a single point of information and discovery.

- Research and investment are needed into tools and frameworks for information discovery methods and tools for exposing, searching for and harvesting data, metadata, tools, methods, workflows or information, within single and across federated environments.

Sufficient and sustained investment is needed in standards for data, formats and others, particularly those for expressing semantics such as thesauri and ontologies.

### R5. Open access to publicly funded data

The most frequently voiced opinion during all phases of the study was that publicly funded data should be free at the point of use. Publicly funded data should be free at the point of use to the user. It should be available for open access, except where required otherwise for confidential, ethical or security reasons or during a period of privileged use for the generator of the data.

### R6. Collections management, selection and appraisal for sustainability

Huge volumes of data are being generated and accumulating; not all of it needs to be kept indefinitely.

- Automated tools are needed for appraisal, particularly given the huge volumes. Research is needed into scientific information appraisal for selection into digital repositories and subsequent reappraisal (the criteria, processes, automated support tools, possibly even different approaches for the digital information age). More generally, management structures and automated tools needed in the future should be charted and planned for now, to cope with the increases in volume and for organizational stability.

### R7. Preservation of digital information

Much of the data generated and held in repositories is of long-term or indefinite value; a lot will need to be kept as part of the record of science. However, e-Science data are at the difficult end of the digital preservation spectrum: typically specialist, complex, heterogeneous.

We recommend increased investment into digital preservation research in the context of e-Science digital repositories.

### R8. Trust and recognition

Data will not be used unless it is trusted: the user needs to know how it was generated and that its integrity has been preserved. Better prepared data at deposit stage is important in this regard, and also substantially reduces repository costs.

- We recommend investigation into measures of review and recognition of data as well as publications, and also mechanisms of recognition for an individual's work in data management (whether directly in science, research or teaching, or in data management services), such as data citation with the aims of increasing trust and levels of use of repositories.

### 9. Governance and management

- Good governance is fundamental. Digital repositories should have a clearly defined remit and responsibilities, with matching policies (and target service levels for their customer groups) for users, data suppliers, and collection owners. Formal reporting by repositories to funders is important, for accountability and good communication, to inform sustained funding and opportunities for enhancing resource management (stressing that low use does not necessarily mean low resource value).

### 10. Training and awareness

Training multiplies the level and quality of use of repository holdings. It is an excellent conduit for feedback about services and tools. It fuels and strengthens the competency base, and will help ensure that the European Union maintains a leadership.

- We strongly recommend training in good data management practices at all levels, outreach by e-Science digital repositories and teaching of related skills at an early age, and aspects relating to the use of materials held in e-science digital repositories. This training should be conducted in the home

language, if possible. Working with scientific data calls for knowledge of science, computer science and information science, and we recommend cross-training between these areas.

### R11. Legal issues

Science and e-Science in particular work across national and administrative boundaries; working across heterogeneous legal, regulatory and administrative systems can slow the work of e-Science to a standstill. The lack of harmonisation in legal frameworks relating to intellectual property in general, and to copyright in particular, across the EU and the EEA, is a severe obstacle to e-Science and risk to repositories and users, and we endorse calls for a more fundamental review and analysis of the nature of intellectual property and copyright.

- Further research is needed into rights management and rights expression tools for rights relating to the use of data in e-Science contexts, which can be supported at very low transactional cost. We further recommend provision of a clear, simple multi-lingual information source, guidance and basic training to all repository providers, higher-education students, scientists, teachers and more widely on the basics of intellectual property, the different types of licences that can be used, and laws which might apply to e-Science repositories and their holdings and related tools.

### R12. International

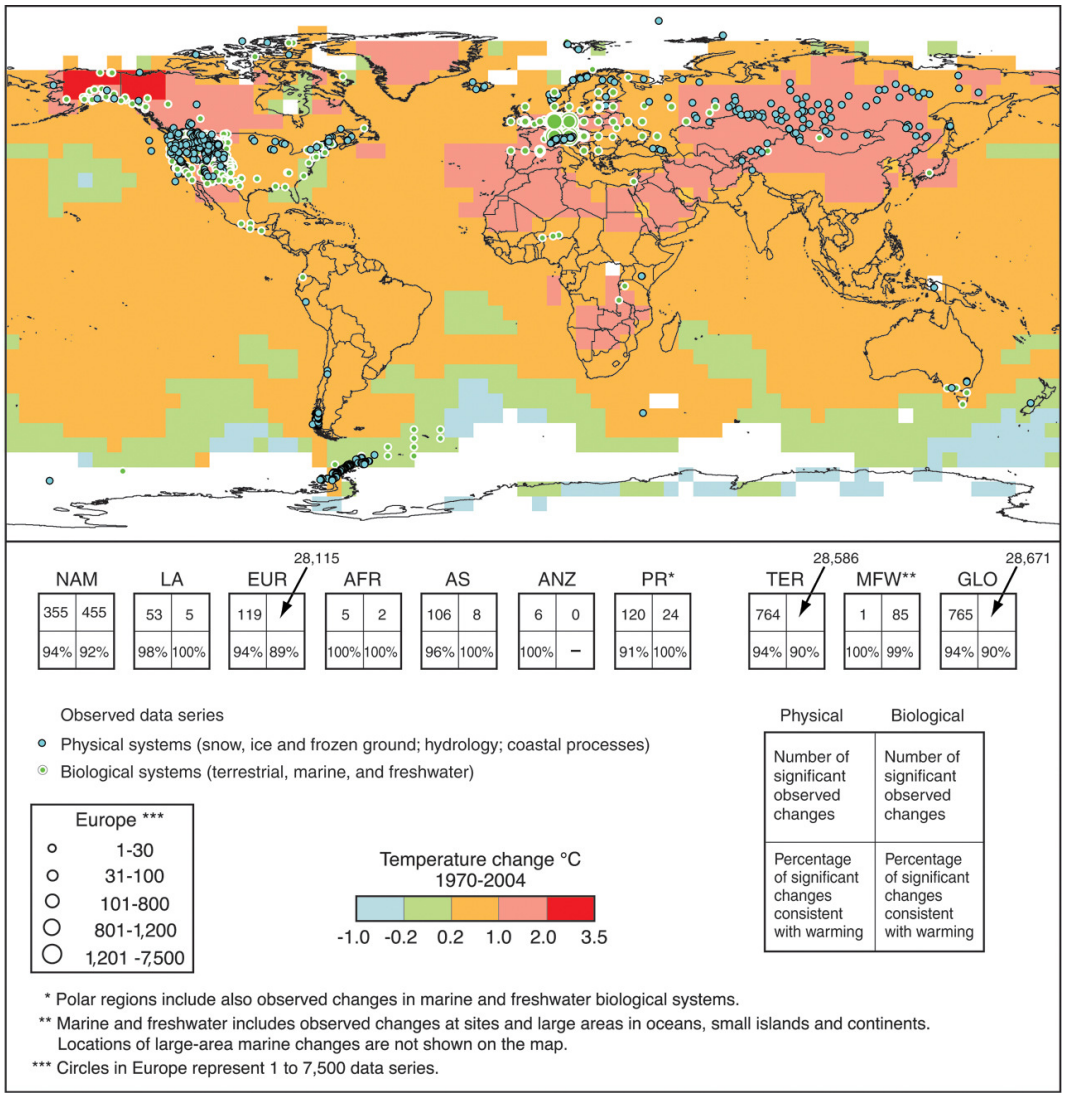
A significant proportion of the data held in e-Science digital repositories forms part of global collections, whose management is thus *per se* an international matter. Global, system-level, integrative research is at the forefront of science, but many developing nations have stretched resources for collecting and keeping data.

- Engagement with developing nations in this field of repositories and their supporting infrastructure is an important geo-political and strategic opportunity, as well as part of a global responsibility. Given the global nature of e-science collections, data and working, we suggest establishing a designation of World Heritage Data and conducting a review into how archives for these data might be supported.

### Conclusion

A strong repository infrastructure in Europe benefits science, scientific productivity and impact, supporting new scientific methods and paradigms. It improves the return on investments in science, and feeds improved economic performance, society and public health through the availability of key data and more productive research. Scientific heritage is better assured. Investment into the incorporation of e-Science digital repositories and their holdings into the information ecosystem, traditionally formed by the library framework, will deepen and broaden Europe's Single Information Space and Research Area of the 21st century information age.

For the most part science is an international endeavour, but its management, policy and funding structures in contrast have been mainly conducted in the past at national or institutional levels. In the context of the recommendations from this study we stress the importance of a global perspective; without this, European scientists and science will drift to where this perspective is better acknowledged.




This illustration, from the International Panel on Climate Change Fourth Assessment report in 2007, is eloquent testimony to the huge achievements of e-Science digital repositories, scientists, data scientists, data management communities. The sizes of the blogs indicate the sheer scale of data sets used in the IPCC report.

# Introduction – Core definitions

> *"Repository. (1) A place where things are stored or may be found, especially a warehouse or museum. (2) A receptacle. (3) (often followed by 'of') A book, person, regarded as a store of information."*
>
> *Oxford Concise English Dictionary, 1996*

Before presenting our findings we need to define key terms and concepts associated with digital repositories.

## e-Science and e-Infrastructure

The focus of this study is "e-Science digital repositories". The repositories and collections under consideration therefore contain information originating from, involving, or potentially contributing to scientific processes over the research cycle or in education that apply **e-Science**—that is, science supported to a significant degree by digital information-processing and/or computational technologies, or wholly based on these. Note that such a definition is *functional*, not some intrinsic property of the science. **Data-based science,** that is science which is based wholly or in part on exploiting *existing* information, is included within this definition.

E-Science includes a very broad class of activities, as nearly all information gathering is computer-based, or uses information technologies for measuring, recording, reporting, analysing. [egee quote] E-Science often involves intensive use of such technologies: advanced in technique, collaborative or on a large scale (over various possible measures: volumes of information, computational intensity, extent of distribution, variety of information types handled). We stress that e-Science can be conducted equally by individuals and small units – in other words, e-Science is equally relevant to small science, and indeed e-Science brings big science within the grasp of less well-equipped – all you need is a computer.

Professor Tony Hey presented the term e-Science (in 2002) in terms of goals - "solving the new problems of science and engineering". To solve these problems, "we will need to be able to pool resources and access expertise distributed across the globe". Thus e-Science is a means and an enabler, a way of working in the activity of science, whatever the discipline.

A presentation by the EGEE project summarized the emergence of e-Science as the invention and exploitation of computational methods, and linked it to the need to curate data:

- "To generate, curate and analyse research data
  - From experiments, observations, simulations
  - Quality management, preservation and reliable evidence
- To develop and explore models and simulations
  - Computation and data at extreme scales
  - Trustworthy, economic, timely and relevant results
- To enable dynamic distributed virtual organisations
  - Facilitating collaboration with information and resource sharing
  - Security, reliability, accountability, manageability and agility."

Common themes from these definitions are the availability of, and access to, data as the product of science and as a source for further science, and an ability to share it within set security and access limits.

### e-Science spans all disciplines

As with the word "science", e-Science is relevant to all disciplines. The challenges imposed by the "hard" sciences in terms of technology needs, and the accumulation and use of data, are just as critical to the social sciences and humanities, and in some cases exceed them (for example, linguistic analysis, or the capturing and analysis of performance art in digital form).

e-Science increasingly takes place not in subject domain silos but across disciplines. e-Science opens new opportunities for interdisciplinary research and innovation. Enabling this inter/cross-disciplinary work substantially increases the inter-operability problems, at computational, semantic and also organisational and professional levels. For instance, an e-engineer may be needed in a performance art research project, but how is his/her contribution recognized? Is the research report in a humanities journal recognized by engineering faculty?

The study therefore draws the boundaries wide, covering the traditional sciences to the humanities. e-Research and e-Learning can also fall under the umbrella of e-science. There are also overlaps with other e- terms, notably e-Health and e-Government, a point we see as reinforcing the need for co-ordination.

An **e-Infrastructure** for e-Science digital repositories is taken to be the technical and administrative framework and facilities underlying e-Science digital repositories. Until recently, the concept of e-infrastructure has usually been defined minimally, to include networks, authentication and authorisation mechanisms, middleware, computational resources (in particular high-performance computers), and those which enable collaborative working, including Grid technologies. We adopt a wider interpretation and include technologies of various kinds for creating, collecting, annotating, manipulating, storing, finding and re-using information and services such as those to provide user support, and training, preservation. Further, we include information resources and associated tools such as vocabularies, ontologies, rights management and privacy protection systems, and curation. Several of these resources depend upon manual human input.

## Collections and repositories

To scientists, what is most important is the information itself; secondarily information should be accessible, usable and trustworthy, and there should be various tools and services which help them to assess, analyse or use information. The repository where information is held is of no significance to them, as long as the contents are reliable, available, accessible.

Reflecting this priority, we note first that repositories are not the same as collections of information. **Collections** of items of information are information items brought together for some specific purpose or with at least one feature in common. The purpose behind bringing information together maybe specific, or quite general – examples might be a collection of information generated by a given institution, or generated by a particular project, or gathered together according to relevance to some particular discipline, or gathered together by an individual. The reason may be explicitly stated or may be inferred by the context of the collection. Note that collections thus defined comprise information itself, not the mechanisms to store and manage it.

Collections may be permanent or brought together for a short time, to serve the needs, say, of a particular investigation, and disbanded when the investigation is complete (we discuss the implications of this characteristic below).

We define **repositories** as the constructs that hold collections and facilitate their use. This can be interpreted narrowly to mean storage equipment and supporting computer systems (as we see in the findings, such systems alone can demand levels of expertise outside those available to a digital repository). We use a wider definition, and include the management framework, services and tools associated with a repository as well as the storage machinery itself. Where the narrower interpretation is applied, we make that clear.

The word "repository" is sometimes used elsewhere as a synonym with "collections", that is: the information stored. The distinction between the two is very important in several respects (not least sustainability and governance). We study this distinction and highlight it in our recommendations.

This notion of a repository managing collections is a distinguishing characteristic from simple file stores, as is the presence of various services associated with it as described below.

A repository can contain one or more whole collections, or a collection can be distributed over more that one repository (the relationship is said to be an "n-to-n" one). Collections may of course be copied (replicated) to two or more repositories—and indeed, this is an operational need, for example for large-scale data sets, and for preservation.

The digital repositories we covered in our study bear a range of different designations—most often data centres, archives, data libraries. This makes it extremely difficult to review the territory, exacerbated by the lack of registries or catalogues of digital repositories. There are registries for institutional repositories, but these represent a fraction of the resources relevant to our study. There are domain-specific portals, providing a layer above resources, through which users have a single point of access. These layers can be said to be a form of e-infrastructure.

Several opportunities are missed because of the lack of overarching registries; above all, it means that those working in the area have little contact with their peers in other domains.

We encountered several categories of repositories:

- Institutional repositories—those set up for the use of a specific institution, mainly to hold the information outputs from the institution
- Community repositories—set up to manage the information of a community of interest
- Subject (discipline) repositories—as a community repository, the community defined by a scientific discipline
- E-learning repositories—containing pedagogical information and materials.

To which may be added others, including data repositories and publications repositories. Digital libraries are another form of repository, which we did not cover in depth in our survey, except where active work is being conducted to activate or implement "e-Science" capability.

Repositories are often part of wider institutions; frequently, there was a blurring between the services provided by the wider institution and those by the repository.

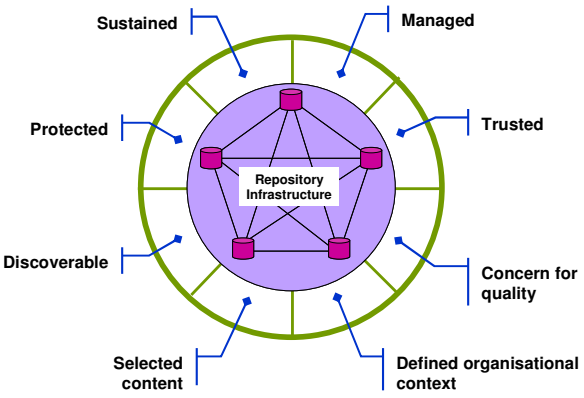
A further distinction is between **private** and **public** repositories – the former holding collections available only to a closed community (such as a commercial company's employees), the latter being available to a wider public than those concerned with the creation of the information (though constraints may apply to who has access and under what conditions). Private repositories include

information held in storage before being submitted to a repository. Another dimension of characterisation is payment: **open access** repositories make their contents available free of direct charges (at the point of use), whereas others (such as most commercially-run repositories) may demand a direct payment for information and/or services provided.

Repositories clearly vary to an enormous degree by their organisational setting, their audiences, the processes they support, the type of information they hold, the services provided with them, and more. Many putative defining qualities have been proposed and we encountered many suggestions during the study. These include:

- Content is deposited in the repository, whether by the content creator, owner or third party.
- The repository architecture manages content as well as metadata.
- The repository offers a minimum set of basic services *e.g.* put, get, search, access control.
- The repository must be sustainable and trusted, well-supported and well-managed.

On examination we could not find one single minimal subset of qualities which clearly characterized all e-Science digital repositories. We concluded that to qualify for the term at least some of the qualities listed should be present - as well as the basic idea of holding all or parts of collections. See figure 1.




**Characteristics of digital repositories**

Repositories (and collections) may be **federated** – that is joined together logically (if not physically), to serve some particular purpose such as for purposes of cross-repository/cross-collection searches or to perform some other function on them.

## Science disciplines the axes of communication and organisation

Our survey confirmed that organisation, services, tools, standards, data management in general, governance relating to digital repositories happens to a very large degree within scientific disciplines. These are the natural channels for communications for scientists; the data structures, formats, and so on are usually specialist and domain-specific. However, it is clear that there are substantial opportunities for sharing of good practice, techniques and know-how. Some fora for these exchanges are emerging, such as in the context of e-Science programmes and curation initiatives.

## Information types

So far we have been careful to use the term "information" when discussing the contents of collections and repositories. We use this general term to include more specific concepts. The information in repositories is usually differentiated into different, well-defined **items;** synonyms for "item" are record or entry. Items are the units of information as a whole which belong to a collection which is stored and managed in a repository. We can often distinguish two sorts of information in an item: its content and metadata. **Content** is what is considered to be the thing lodged in the repository and of primary interest for deposit and use—say a data file or a digital form of a publication. **Metadata** is descriptive and other information pertaining to that content. Metadata can be extensive, and varied – such as descriptive information, annotations, indexing, classifications, technical information about the content's file formats, and more. The distinction between metadata and content is not always clear-cut—we have more to say on this below.

In what follows it is useful to make a distinction between two types of item—data items and **publications**. The latter are those whose content is in the form of a publication of some sort—thus published article and reports, pre-prints and post-prints, theses, patent documents and similar. **Data items** are all other types of content—such as databases, images, video, and simulation results—and so on. The reason for making this distinction is the higher level of maturity and uniformity of publications repositories over data repositories, due to the lesser measure of format heterogeneity. (Below we also mention points raised by Dr. Peter Murray-Rust and others about the possible changing nature of the traditional publication.)


**Figure 2: Characterisation of information types in repositories**

### Other information types

Some other information types are relevant to repositories. These include thesauri, classification schemes, ontologies, indexes, registries and catalogues (and indeed simple word lists as authority files). Technologies to handle these and the standards they may conform to are discussed later in this report. These adjunct information types may themselves be kept in repositories, and will need maintenance within the repository structure. The same applies to tools and software.

### Drivers and rationale for e-Science digital repositories

In a sense, this is also an answer to the question, why are people reflecting on e-Science digital repositories?

As Tony Hey and Anne Trefethen noted in their "**Data Deluge**" paper in 2003, the quantities of data we are generating are enormous, thanks to technological advances not only inside IT, but outside IT

(through the development of new sensors, instrumentation and techniques). The sheer volume of data is an operational and cost pressure, for managers and administrators; for users, there is a huge problem of finding useful information, the needle in the haystack. Our ability to generate and collect information continues to grow more quickly than our means to organize, manage and use the information effectively.

Nevertheless, re-use and re-purposing of data remain both benefit and driver. The OECD Principles and Guidelines for Access to Research Data from Public Funding give examples of benefit, in the context of increasing the return on public investments in scientific research, and provides many of the reasons for and benefits of digital repositories:

> *"Accessibility to research data has become an important condition in:*
>
> **The good stewardship of the public investment in factual information;*
> ** The creation of strong value chains of innovation;*
> ** The enhancement of value from international co-operation.*
>
> *More specifically, improved access to, and sharing of, 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;*
> - *Permits the creation of new data sets when data from multiple sources are combined.*
>
> *Sharing and open access to publicly funded research data not only helps to maximise the research potential of new digital technologies and networks, but provides greater returns from the public investment in research.*

The Netherlands aim to reinforce research in the Netherlands by improving dissemination of knowledge and collaborating in an international context; they see networks of data repositories as playing an important part in the knowledge infrastructure.

Peter Murray-Rust makes the additional point that repositories enable experiments to be conducted outside a laboratory. It is effectively also the point that experiments can be conducted *in silico*, part of the new paradigm of science and research, enabling new models of working, in teaching , research, and business.

Some quantifications of the benefit of efficient re-use of data are given in the Joint Data Standards Study [Lord, Macdonald, 2005].

### Data and collection classifications

The Long-Lived Data Collections study [2005] distinguishes "three functional categories of data collections":

- Research database collections, which are specific to a single investigator or research project

- Resource or community database collections, which are intermediate in duration, standardization, and community of users, and
- Reference collections, which are managed for long-term use by many users.

As the report says, the distinctions between these categories are not always clear-cut. The categories are based on "functional attributes of the collection rather than location or size of the data set". In some ways, the distinctions above relate to different stages in the life-cycle of a data set.

Another key distinction is that some data is unique and non-reproducible. Our review also suggested that a big factor in practical management of repositories is whether data is generated *in situ* or *ex situ* – that is, whether data comes from external sources, outside the direct control of the repository or its parent facility.

### Linkage and enhanced publications

There is an increasing drive towards movement in the meaning and drive of the word "publication". Peter Murray-Rust believes the traditional (current) scientific publication comes in fragmented form, separating publication from underlying data. Initiatives and projects covered in the overview are working to achieve linkage of related objects (primary data, processed data, presentation, pre-print, publication, etc) and the development of text mining tools to support linkage between objects. The SURFshare programme in the Netherlands notes that it is "removing absolute distinction between research data and the traditional publication as research output". Several studies make the point that "enhanced publications" strengthen the quality and reliability of the publication.

In a sense, a publication is metadata (and that is certainly true for publications on simulations). Peter Murray-Rust points out that 20th century storage and management structures do not easily support holistic (non-fragmented) publications; in addition, commercial entities and structures have been developed on these structures. This does not necessarily mean a fundamental challenge to existing structures; ArXiv has existed alongside the formal publication system for two decades.

A further challenge on linkage relates to maintaining confidentiality at the same time as maintaining links to support efficient discovery.

### Types and nature of materials held

The NSF Cyberinfrastructure report defines "data are any and all complex data entities form observations, experiments, simulations, models, and higher-order assemblies, along with the associated documentation needed to describe and interpret the data".

Studies and workshops by the UK e-Infrastructure Roadmap working groups or the Knowledge Exchange (for example) see the materials held in e-Science digital repositories as from all formats (text, image, audio, video, combinations), and at all stages in the information and life cycle of objects, from primary research data, algorithms, models, visualisations to publications, to e-theses. Indeed, there are also repositories of software.

Malcolm Hyman and Jürgen Renn of the Max Planck Institute for the History of Science have pointed out on several occasions that much of the data to be held is dynamic. So the traditional concept of a repository risks locking data into a static representation. They extend this to note the risk of lack of structural correspondence between content and representation, the need to maintain links between media, related objects. All these points have operational and technical implications; web 3.0 they see as presenting an opportunity to address these issues.

### Linkage and enhanced publications

There is an increasing drive towards movement in the meaning and drive of the word "publication". Peter Murray-Rust believes the traditional (current) scientific publication comes in fragmented form, separating publication from underlying data. Initiatives and projects covered in the overview are working to achieve linkage of related objects (primary data, processed data, presentation, pre-print, publication, etc) and the development of text mining tools to support linkage between objects. The SURFshare programme in the Netherlands notes that it is "removing absolute distinction between research data and the traditional publication as research output". Several studies make the point that "enhanced publications" strengthen the quality and reliability of the publication.

In a sense, a publication is metadata (and that is certainly true for publications on simulations). Peter Murray-Rust points out that 20th century storage and management structures do not easily support holistic (non-fragmented) publications; in addition, commercial entities and structures have been developed on these structures. This does not necessarily mean a fundamental challenge to existing structures; ArXiv has existed alongside the formal publication system for two decades.

A further challenge on linkage relates to maintaining confidentiality at the same time as maintaining links to support efficient discovery.

### Types and nature of materials held

The NSF Cyberinfrastructure report defines "data are any and all complex data entities form observations, experiments, simulations, models, and higher-order assemblies, along with the associated documentation needed to describe and interpret the data".

Studies and workshops by the UK e-Infrastructure Roadmap working groups or the Knowledge Exchange (for example) see the materials held in e-Science digital repositories as from all formats (text, image, audio, video, combinations), and at all stages in the information and life cycle of objects, from primary research data, algorithms, models, visualisations to publications, to e-theses. Indeed, there are also repositories of software.

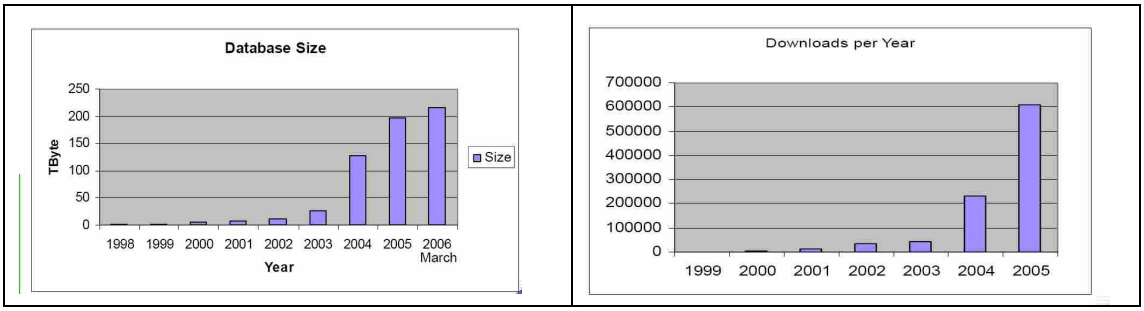
Malcolm Hyman and Jürgen Renn of the Max Planck Institute for the History of Science have pointed out on several occasions that much of the data to be held is dynamic. So the traditional concept of a repository risks locking data into a static representation. They extend this to note the risk of lack of structural correspondence between content and representation, the need to maintain links between media, related objects. All these points have operational and technical implications; web 3.0 they see as presenting an opportunity to address these issues.

### Key features and challenges

For brevity, we list a few key features and challenges facing repositories:

- Extremely rapid growth, and sheer volumes of data and metadata, in and out of the repository
- Digital repositories, scientific data communities, umbrella portals do a huge amount of little heralded work, developing and maintaining standards, database schemas, ontologies and so on, to improve interoperability
- With the existing infrastructure platforms of powerful networks and computers, repository frameworks, there are several major initiatives to create global information collections – information infrastructures – at semantic level, also enabling better automated search and discovery of materials.

- Difficulties relating to data deposit; these have been eased by publication requirements to deposit data in designated repositories and mandates from funders, but this remains a problem (behavioural), and also affects the quality of data deposited
- Repositories are under-staffed; the perception is all too easy that a repository is just a box attached to the Internet.




**Database growth and annual downloads at the World Data Centre for Climate**

## Study method

This study conducted the following programme of work:

1. Early in the project, we conducted a series of three workshops, each addressing different aspects of the repository movement:

   (i) Roles, functions and drivers,

   (ii) Interoperability, standards and technologies, and

   (iii) Legal, economic and sustainability aspects in the third.

   These workshops were attended by invited experts from across Europe and beyond, representing different, sometimes contrasting, perspectives on repositories, data management (data and publications, institutional repositories and community repositories, commercial and open/source resources, the vary varied user communities). Part 2 includes reports on these workshops.

2. We undertook desk research and analysed the situation in Europe and the wider world. We approached from multiple perspectives: technical, organisational, financial, legal, usage, distribution and repository penetration, Specific areas examined included repositories, stakeholders, those reflecting on repositories, technologies and standards, and legal aspects.

   During the course of our work many relevant meetings were held, at European and national levels, and which we were able to attend, including OAI5 in Geneva, the April and October e-IRG workshops in Heidelberg and Lisbon, the European Commission's conference on Scientific Publications in February 2007). We conducted a public consultation process using the Commission's website and IPM system. Our aim was to elicit the views of users and other stakeholders on the use made of digital repositories, the barriers and enablers to their use, adequacy of provision in Europe, and the need for longevity. The consultation also asked about stakeholders' vision for the future of repositories. Part 2 includes a report on the consultation and a summary of results from the 428 responses made by individuals from within Europe and beyond, and we include quotations from these responses throughout this report.

3. A final workshop of invited experts was held in Lisbon to discuss finding and emerging policy directions. The meeting reviewed, discussed and prioritized a list of 20 recommendations, which guided our distillation of the 12 recommendations set out in this report.

# e-Science Digital Repositories: Vision

To formulate policy options, we need a vision and objectives for those policies.

Before setting out our vision we stress that this vision does not mean that its elements do not exist; as our overview of the current situation shows, many elements of our vision exist or are developing.

In our vision, this e-Infrastructure should allow researchers, teachers, students and other stakeholders – and also machines - to find and access a rich range of data repositories and their contents determined by their research interests and their associated privileges. Targeted tools and community services supporting discovery of, access to and analysis of these repositories are essential.

Many of these repositories and associated tools relate to specific domains; however, there are needs and issues common to the different communities. To support the long-term vision of e-Science, co-ordination and management of data sets and associated services are essential. This is especially the case in supporting inter-disciplinary research. In other words, co-ordination is key to achieving the vision.

#### Multiple domain perspectives

By its very nature, e-Science is broad in scope, covering the complete spectrum of modern research and education from the arts to the traditional physical sciences, from the theoretical to the experimental, from the commercial to the academic, amongst numerous other dimensions. Each domain has its own perspectives and makes its own demands on e-Science digital repositories, how it accesses and uses them, and, ideally also relating to longer-term models for their sustainability. A vision here must address these different perspectives.

#### Part of a single information space

Overall the vision for e-Science digital repositories in Europe should encompass and be part of a single science information space that serves multiple stakeholders and permits multiple perspectives: for science, scientists, researchers, students, schools, the publishing community and industry. These repositories and their holdings should be recognized as part of the overall corpus and ecosystem of information and knowledge. The e-Infrastructure for e-Science digital repositories should embrace all the states of the EU and cater for the special needs of its new members, and it should bridge differences between the rich and poor.

#### Specific elements of the vision

The following sets out specific elements of a vision for an infrastructure for e-Science digital repositories in Europe, under the following headings:

- A reliable, sustainable e-Infrastructure for e-Science repositories
- A high-quality information space
- Information that is readily available
- A well-managed, accountable repository infrastructure

### A reliable, sustainable e-Infrastructure for repositories

- The e-Infrastructure should support **easy and reliable submission** of materials for science, research and learning into known, trusted repositories through the whole science, research and

education cycle, providing confidence that the materials will be well and **securely stored**, maintained, and not abused. Scientists with data to deposit should be able to do so with ease, supported by timely guidance and tools to help create sufficient quality metadata. They should be assured of their prior right to recognition and a defined period of privileged use (where applicable).

- The repository infrastructure should be **funded adequately** for service provision and sustainability.
- The various stakeholders - administrations, the scientific, education and learning communities, the private sector and the general public – should have well-founded confidence that the infrastructure is **reliable**, delivers value for money, can adapt to change as technologies and science move on and that it continues to collect and preserve securely Europe's great scientific heritage and core information (which plays an increasingly critical role this century, as we face critical threats such as climate change and biodiversity loss).
- The repositories should have a capacity or associated framework to support the long-term **sustainability** of collections, be trusted, and to guarantee the authenticity of stored materials and cope with changing levels in demand.
- Europe's infrastructure of repositories should deliver services **equally across the whole of Europe** and participate as (a) partner(s) in the wider global e-Science information infrastructure.

### A high-quality information space

- Peer review and similar mechanisms should be available to provide **quality assurance** for depositors and users of data, and also repositories.
- The e-Science repositories and their contents form part of the overall body and ecosystem of information and knowledge.
- Users should be able to gain **access and authorisation** with ease whilst privacy, property rights, copyright and ethical use are securely protected. Repository policies should be clearly identifiable and available.

### Information that is readily available

- The outputs of **publicly funded science should be made available** in open-access repositories, with strict but fair controls to protect privacy, property rights, security and ethical use. Ideally these outputs should be provided free at point of use.
- The repository infrastructure should be **transparent**, and it should be easy to find the collections which are of interest. The e-Infrastructure and repositories should support the scientist at all points in the science cycle, providing easy, cost-effective access in a joined-up fashion to materials of all types that are already available.
- **Minimal barriers** should be placed in the way of scientists wishing to use information free at the point of use. Reliable statements about data quality should be in place, and accompanying restrictions on use clearly stated. Easy-to-use search tools should be available, and reliable, supported and maintained tools should also be available to assist use and rendering of information. It should be possible to navigate seamlessly between resources and to move along the information chain of data to final publication.
- The structure of provision of repositories should be such that:
  - (i) it satisfies users' needs to be in contact with and visible to their discipline,
  - (ii) at the same time, it satisfies the desire of institutions to be associated with the scientific outputs whose creation they host, and
  - (iii) it provides opportunities for commercial and value-added providers.

- **Training** should be available at all levels on good data management and repository use, in one's own language so that users know how to use materials, understand the materials, rights issues.

### A well managed, accountable repository infrastructure

- Management of repositories should be responsive to funders and users and their needs; **clear objectives, policies, performance and timelines** should be articulated and available (for human **and machine** consultation). Policy makers at all levels should have a clear view of what is available in repositories and its (scientific) value, and thus also of the management of these collections over time.
- The collections in repositories should be **expertly maintained.**
- Fair and efficient methods should be available to **appraise information on accession** and **thereafter** to determine selection criteria, retention times and levels of support needed for individual datasets.

## Encouraging advanced architectures

Technology should be applied so that stakeholders [can find what resources are available, so that ownership and digital rights management are addressed; and so that a basic benefit of Grids is supported: a **single sign-on**, where users authenticate once and thereafter are able to move seamlessly across a range of distributed e-Resources. Some communities, such as that around high-energy physics, have a track record in building and using large-scale Grid infrastructures for managing large, heterogeneous data sets. Other domains, such as the arts, social sciences and biological sciences (amongst numerous others), require environments where services and data resources are offered in a coherent and user-driven environment. The focus should therefore be on creating environments that facilitate research, rather than providing Grid infrastructures *per se*. Furthermore, the domain knowledge needed has to be transferable across disciplines, and ideally the e-Infrastructure itself has to be seamless and transparent to the end users.

### Maximising discovery and use

To maximise the discovery and use of European digital resources, it is likely that domain-specific trusted portals will need to be established. These should support services searching across large numbers of existing diverse collections, and quickly provide comprehensive and reliable references to relevant material. Key to this will be the production of fast searchable indexes and support of advanced content-based queries. Access rights to these resources can take many different forms: free web access; restrictions for educational use only; institutional subscription; registration models; private subscriptions; pay-per-download models, or time-limited leasing. It is likely that a mix of models like this will need to be supported.

> *".. a distributed system of national and community-specific digital repositories linked semantically at the European level [with] more and more automated metadata annotation and extraction of important facts from unstructured text and images." (Response, e-SciDR public consultation)*

# Recommendations

We set out below 12 core **recommendation sets** for policy and measures to drive towards a European e-Infrastructure for and of e-Science digital repositories. We also hope the recommendation sets will also contribute to the building of a Single European Information Space, including the embedding of digital information into the body and ecosystem of information and knowledge.

These recommendations have been drafted in the light of discussions with experts, the study's sponsors, the e-Infrastructure unit of DG INFSO and colleagues in DG Research, our consultation with stakeholders and the Lisbon e-SciDR workshop, and extensive desk research. These recommendations are informed by policy at European, national, community and multilateral levels, and they have been cast using the vision and objectives set out above.

The policies and measures are mutually reinforcing. Key cross-references are flagged in the recommendations in square brackets, with the abbreviation "qv" followed by the recommendation number, in the right-hand column.

The 12 core recommendations sets are presented under the following headings:

1. Funding
2. An e-Infrastructure *of* and *for* European e-Science digital repositories
3. Support for data producers
4. Discovery and navigation
5. Open access to publicly funded data
6. Collections management, selection and appraisal for sustainability
7. Preservation of digital information
8. Trust and recognition
9. Governance and management
10. Training and awareness
11. Legal issues
12. International

# 1. Funding

The most urgent need expressed over the course of the study, in the workshops, interviews and consultation, is that funding of digital repositories (including services) needs reform.

**R1: We recommend that the European Commission urge member states and their agencies to provide funding for e-Science digital repositories, which is specific to their role and function as digital repositories.**

**This funding should be stable and rolling, and of a duration to match the duration of the repository's role or of its holdings.**

**The funding should be sufficient to support and maintain the repository holdings, individually and as collections, to provide quality services and support to users, and provide good management at repository level, which can deliver continued, efficient, rich, easy-to-use access to trusted, quality materials. This will entail the provision of funding at levels beyond the repositories themselves.**

This funding recommendation should extend to small, and medium-sized repositories as well as those holding large datasets, as in physics and astronomy.

Funding should be sufficient to enable efficient performance and service in distributed environments. Duplication of holdings is also an important security and integrity measure (at least three copies of information, located apart, are needed to ensure continued availability). Such duplication may be based on mutual exchange and cooperation agreements with peer organisations in other parts of the world (as exemplified by initiatives such as LOCKSS[2]). [qv R6: collections management, version management]

The funding should cover the development and maintenance of services and tools to support easy, rich use of the repository holdings. Easy-to-use, intuitive interfaces and tools generally take substantial resource – they need experts (computational and domain experts) to develop and maintain them, and it takes time to do so. As well as information-level services and tools, the services can involve providing computational resource at the repository itself. The various services may be discipline-specific or they may be generic to all types of repository or data. This will entail funding for things like development and maintenance of pipeline and ingest tools, retrieval and visualization methods, as well as metadata tools and curation of holdings. The multiplier benefits (and thus return on investment) in terms of productivity for users, quality of research and reach of the resource are massive. [qv R2: e-Infrastructure] [qv: R8: Trust]

Europe's e-Science digital repositories should be sufficiently resourced to enable them to improve and widen access to their holdings and services, in terms of quantity and quality – enabling more scientists, researchers, teachers, students, citizens, to access materials. More access means more demands on resources, so there are provisioning/forecasting and scalability implications. At the same time they must sustain performance and quality of service, across networks and the Grid. This requires investment in research into means of optimising access, storage and use, as well as

[2] http://www.lockss.org/lockss/Home LOCKSS: Lots of Copies Keeps Stuff Safe. An international community initiative that provides libraries with open-source software, decentralized preservation infrastructure and support to preserve authorized copies of e-content.

infrastructure in and outside the repository (storage, networks, and their management). [qv R2: e-Infrastructure]

**Funding for quality:** The funding should cover provision for curation of holdings, where appropriate, and other quality measures. [qv R8: Trust]

**Funding structures for services and shared facilities:** Services to support access to data, efficient retrieval and use of objects in repositories, may be built and/or maintained independently, in co-ordination with, or by, the repositories. In other words, the funding beneficiary might not be the digital repository itself, but an external expert and/or grouped supplier. The study's findings also show examples of and opportunities for grouped or centralized development and provision of services, where funding goes to a single unit, servicing many. [qv: R2: e-Infrastructure]

**Discipline-based repositories:** Very clearly in the study's consultations, the preference of users was for discipline-based repositories. Moreover, the authors found that discipline-based repositories are efficient in science, quality, risk and economic terms. They enable not only the availability of expertise in the first place, but a concentration of expertise, and also more efficient identification of user needs at science level. They are also significantly more efficient in drawing relevant material into repositories, because depositors gain personal advantage from doing so, in the form of scientific recognition through the visibility of their materials in discipline-specific repositories. We suggest that the focus of support for e-Science repositories should be towards community-based initiatives, while recognising that repositories based on multi-disciplinary institutions provide a framework of organisational stability as well as for their active participation in the "business" and science of repositories. Encouraging virtualisation of repository access and linking may be one way to obtain the best of both these approaches. We also note that it is easy to under-estimate the substantial difficulties in and expertise required to manage e-Science data.

A study into this funding reform may be useful. We note, however, that we believe the next few years will also see major developments in what we call the vertical information dimension (from data to publication), affecting the traditional boundaries of information objects and therefore with implications and opportunities for institutional roles (such as libraries) and, by extension, organisational, structural and economic opportunities. In the context of repositories, and the heritage of science, there are additional burdens for those with responsibility as long-term custodians [qv: R7].

The outcome and benefits of dedicated, adequate funding for e-Science digital repositories, their holdings and services, are reliable science, based on reliable materials, productivity for scientists, researchers, teachers and students, and longevity of information.

> *"A challenge [for digital repositories] is to store huge amounts of information – and this is very important – have the tools to "play" with it. So repositories are not only about information, but also about tools." (Response, e-SciDR public consultation)*

> *"It is not enough to build instruments, one needs also to invest in tools to manage, manipulate and analyze the data they capture. This often takes a team. … Skilled data scientists should be trained and have a chance for a career. These issues should be stressed nationally and by the EU, and a suggested solution or path for societal and scientific repositories should be agreed upon. Data and repositories represent the next generation in scientific computing."*
> *(Response, e-SciDR public consultation)*

## 2. An e-Infrastructure *of* and *for* European e-Science digital repositories

In the past, science has been viewed in terms of its physical processes, headline outcomes and the equipment that supports it; there has been less emphasis on the data/information output of science. These information outputs come from all stages in the process, from initial planning of investigations to final publication, and can take many varied forms. Repositories form the basis on which these outputs can be made accessible and can be sustained, and in this they may be said to form part of an e-Infrastructure **of** European e-Science digital repositories, within the wider e-Infrastructure.

An e-Infrastructure **for** e-Science digital repositories comprises services and infrastructure within and outside the digital repositories. Some are domain-specific, others are common to all fields. Where these elements are generic, e-Science often needs levels of service or resource hugely greater than average, but which are essential for the scientific endeavour. E-Science digital repositories, their frameworks and an underlying e-Infrastructure provide efficient, cost-effective means for supporting these demands; critical mass, economies of scale and cross-fertilization are more likely and greater at European scale.

The **support services** within and outside e-Science digital repositories are fundamental to the efficient, rich use of the materials held in repositories. Repositories and their wider community/data management and e-Science frameworks provide an efficient framework to identify commonalities, opportunities for interoperability, as well as to provide core and value-added services. They are also a source of or route to computational and informatics expertise to develop and maintain the services, expertise which would otherwise be hard to acquire.

Targeted tools and community services supporting discovery of, access to and analysis of these repositories and their holdings are essential. Many tools will be domain-specific, but several elements, themes and issues are common to all.

**R 2:** **We recommend that the European Commission, member states and their agencies promote the concept of a European Science Information Space, as part of the Single Information Space and the European Research Area. This European Science Information Space includes data collections, repositories, materials, services, tools and other supporting e-Infrastructure resources.**

**We recommend that the European Commission recommend member states and their agencies adopt policies which support identification and co-ordination of opportunities for pooling expertise across Europe to support e-Science repositories and services, to strengthen European science, nationally and internationally and address obstacles to European e-Science – fundamentally collaborative in nature – caused by fragmentation.**

**We further recommend that the European Commission consider establishing a co-ordination framework at European level to bring together e-Science repository and services providers, users, experts from the different scientific disciplines, and from different areas of professional expertise, to identify commonalities, opportunities for sharing of expertise and synergies, in the area of e-Science digital repositories (and thus by extension of e-Infrastructure), and beyond, with other arenas.**

This co-ordination framework should encompass vertical and horizontal dimensions, across disciplines, at science, computer science and information science levels; between communities, between service providers at community level and above, and vertically between data and information levels, across the EU and the EEA. [qv: R4: discovery and navigation].

The co-ordination framework can take advantage of programmes and projects funded under FP7 and FP6, and should work with ESFRI working groups and projects, e-IRG, as well as national programmes. It would also liaise with other frameworks - global, international, national, regional.

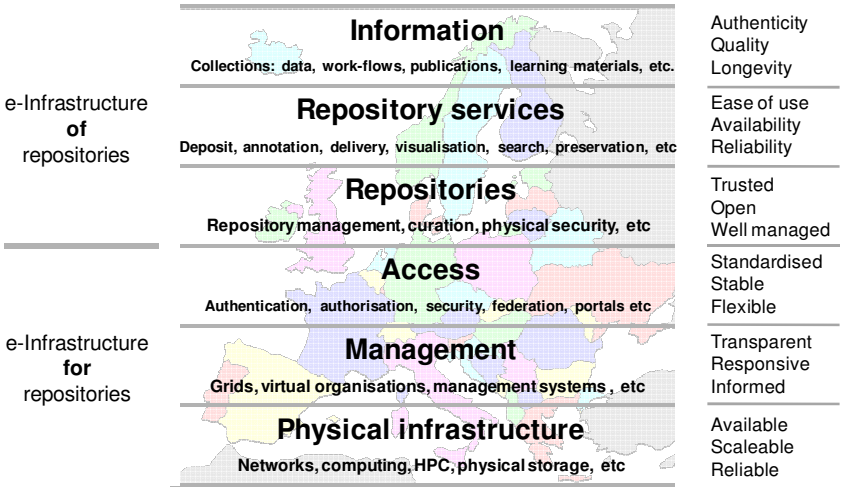
**Generic infrastructure:** Clearly, tangible elements of the infrastructure are necessary to support e-Science digital repositories (networks, computing capacity) and intangible elements and tools : support mechanisms for users, software tools and other enablers (including grid infrastructures). Of particular importance are those tools which ensure security and integrity of information whilst providing ease of access for users.

The primary example is authentication, authorisation and accounting ("AAA") tools – one of the most fundamental building blocks for collaborative e-Science. The ideal AAA tools give users access to the data they need, preventing access to data and services for which they have no rights, providing a single sign-on capability and being transparent to the user thereafter. This ideal has not yet been attained, and part of the effort of establishing the e-Infrastructure for digital repositories should be directed at this issue. [overlap]

An important point in this regard is that several elements of the e-Infrastructure for e-Science digital repositories are also elements of other e-Infrastructures – for research, for learning, for health. Many of these common elements are the most basic, such as AAA.

**A European e-Infrastructure for, and of, e-Science Digital Repositories**

**Layers of the e-Infrastructure and some of their desired characteristics**




This means that development and maintenance of these tools are conducted in other and wider contexts, and must meet wider requirements. It is important that the specific needs of e-Science repositories and their users are taken into account in the development and maintenance of these tools.

*The following paragraphs include further suggestions for actions that will contribute to the establishment of an identifiable infrastructure for e-Science digital repositories from an EU*

*perspective, recognising that in parallel member states and EU-funded science organisations and communities could be encouraged to contribute:*

**Single point of information:** For efficient use and creation of digital repositories, there should be one or more points of information, reference, materials (such as licence guidance and links) and generic support for those setting up and maintaining their own digital repository, data store or data collection. This resource could be virtual, provided by more than one supplier. The resource must be multi-lingual. This single point would help towards inter-operability at administrative level – particularly important as we move towards computer-computer interactions – as well as making for better use of people's time, with the greater familiarity with the frameworks that this would encourage.

We strongly endorse the work of the DRIVER project, and commend the initiatives of DRIVER members, and other national initiatives, in providing single-point of access information about digital repositories.

We recommend investigation into opportunities at national, regional, European and discipline level, for provision of **shared facilities**, such as back-up or archive storage. Digital repositories themselves aggregate and concentrate resources, enabling economies of scale, but there are opportunities for economies of scale between repositories (within, across and independent of disciplines) and also on wider dimensions. For small institutions or groups, secure, well-run storage or back-up will often be too expensive; this could be a source of facilities, their use be made conditional on meeting required standards of data integrity and conformance.

A further extension of the European dimension to repository provision is for the EU itself to set up a repository structure. Whilst the notion of an EU-level repository had much support from the study's survey, analysis of accompanying comments (supported by feedback during the workshops) suggested this would best be delivered on a federated basis, with repository provision distributed. Such a structure has appeal on a number of counts: for deep archival storage, as a place where "orphaned" data can be placed where no other repository "home" for it is available, and provision of local services in local languages but giving access to a wide pool of resources where none exist at the moment.

**Standards** (formats, metadata, ontologies, vocabularies, etc) are fundamental to efficient e-Science. We believe that they form part of the e-Infrastructure for e-Science digital repositories; again, however, they are not specific to e-Science repositories, and therefore their development and maintenance take place in wider dimensions. Nevertheless, it is important that e-Science repositories and service providers are adequately resourced to participate in standards work, which is usually international in scale and requires sustained input over many years.

There are some standards which are very widely used. One example is geospatial data. It is important that e-Science digital repositories and their stakeholders are adequately resourced to represent their needs in development of these standards (such as INSPIRE). It is also important that they are included in consultations. [qv R11 re legislation]

**Information gathering:** There is an opportunity for a body to monitor activity relating to e-Science digital repositories in particular in Europe, draw and report findings, maintain statistics, and support a meta-analysis layer above and across relevant projects and programmes. We believe this activity would also make a significant contribution to medium-term economic and business model analyses. [qv re reporting]

We close with a suggestion extending some of those made above, for establishing a **European Data Institute**. As well as a centre of expertise, this could be charged with overseeing specific EU actions, monitoring the situation in Europe (taking into account technological advances and the wider international needs), and providing advice on future policy directions to the organs of the European Union, national governments, and scientific bodies. It could promote initiatives into data citation and data quality assessment, represent the EC on relevant international bodies, and publishing standards, guides and catalogues, and more. It would also be an effective mechanism for disseminating and multiplying expertise throughout the EU, in the same way as, for example, the European Bioinformatics Institute.

> *"Data-driven scientific fields like the life sciences need a stable and robust infrastructure capturing the data generated in large-scale and high-throughput experiments (…). This is a very good investment since it allows scientists to use the data produced … to come up with new hypotheses and to plan their experiments much better and faster, based on the data available in databases and literature"*
> *(Response, e-SciDR public consultation)*

> *"They [digital repositories?] will come, but they can come faster, if activity is co-ordinated by [the] EU."*
> *(Response, e-SciDR public consultation)*

# 3. Support for data producers

Good data management and data longevity start with planning, before the point of data creation [ref jdss]. It ensures information is stored in the most appropriate formats, that it is well described, and has the appropriate features to ensure confidentiality and ethical constraints. (The same applies to a large extent to the development of software for and in research projects.)

**R 3: We recommend that the European Commission urge member states and their agencies to maintain policies and measures which insist on and support good data planning and management by data producers. These policies and measures should be accompanied by corresponding adjustments in funding.**

There are now many examples of good practice in this regard[3]. The European Commission also provides an exemplar to other funders by implementing a similar shift in its own funding activity for science and its funding of repository infrastructures.

Data planning and management needs to be underpinned by policies and data frameworks (such as storage, tools, standards, pipeline tools and designated repositories to hold the materials, if retained, after project end). They also need to be accompanied by training [qv]. A large part of these elements we define as e-Infrastructure elements. [qv R2, R1, R8, R11] This emphasis is also related to outreach and awareness activities, proposed in R10.

At data-producer level, there should be specific allocation within funding for data planning, curation and management by the producers of the data. To support good data production and curation, where the data generated is important, we recommend that funders consider strict policies such as withholding of a percentage of funding, until fulfilment of a project's data management plan. A side-effect of this policy is also likely to be the inclusion of data-management expertise in research boards, which we believe would contribute to competitive research and good asset management.

Data producers range from individuals, collaborative research projects, to the major research facilities, such as the Large Hadron Collider (LHC) at CERN, and those under ESFRI. Some illustrations of exemplary work in this respect are set out in the overview in the Second Interim Report.

Repositories should have the option to refuse submissions and have the power to require non-compliant data to be re-submitted. Repositories and depositors should be encouraged to use, where possible, open standards for data and metadata, including rights information.

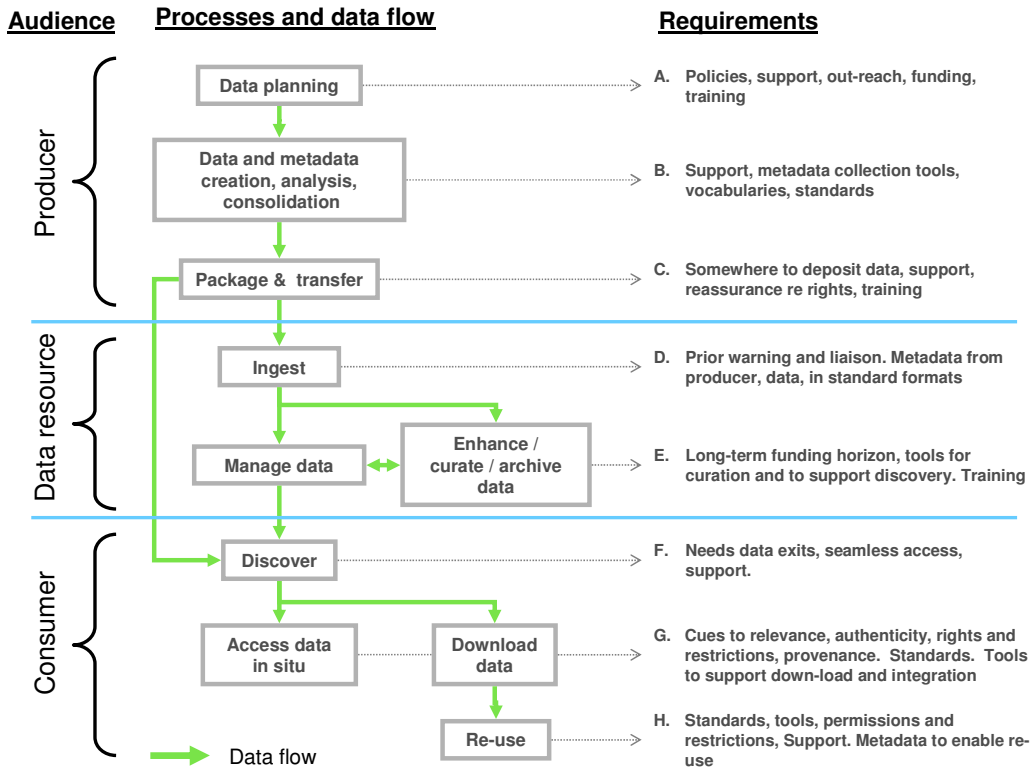
### Benefits

The benefits and outcomes of these measures include:

- An increase in data quality from a scientific viewpoint
- Increased interoperability
- Increased awareness on the part of data generators (scientists, teachers) of value-adding resources available to them

[3] One of the earliest examples is the Environmental Genomics Centre set up by the UK's Natural Environment Research Council (www.nerc.ac.uk), providing support, access to open-source software, guidelines for data planning and management, enabling submission of publicly funded data from research projects to data centres and archives.

- More data management skills in the form of larger pools and increased levels of data management skills
- Better planning for long-term management of data, both at project level, but also for e-Infrastructure suppliers and digital repositories
- Easier ingest into repositories, releasing resources for more value-added work (and thus services) elsewhere in the repositories; easier curation and preservation of the data in the repository
- More effective discovery and retrieval by future users
- An enrichment of data collections.




**A model data/information flow, from data generation to repository to re-use[4]**

[4] Lord, P., Macdonald, A. et al., The Joint Data Standards Study, 2004

# 4. Discovery and navigation

A clear finding from this study was the urgent need for better discovery tools. Their inadequacy or absence is a major hindrance to repository and data use and scientific productivity at multiple levels:

- Finding repositories (and being aware that they exist)
- Deficiencies when searching within repositories and collections, difficulty of use and poor interfaces (frequently cited in consultation responses)
- Poor facilities to search across and between repositories
- Navigating along information chains – particularly the chain of links between raw data as captured or generated and the published conclusions they support.

These difficulties are supplemented by issues of searching *within* items in repositories, though mining techniques are being developed for textual data and for some image forms and other resources. There are several examples of very good practice, for example at Germany's Fraunhofer Institute[5], or within resources at the European Bioinformatics Institute, such as Ensembl BioMart[6].

**R 4: Research and investment are needed into tools and frameworks for information discovery methods and tools for exposing, searching for and harvesting data, metadata, tools, methods, workflows or information, within single and across federated environments. Sufficient and sustained investment is needed in standards for data, formats and others, particularly those for expressing semantics such as thesauri and ontologies.**

The investment must be sufficient to develop and maintain tools and interfaces which are intuitive and easy to use. This is of critical importance. [qv: R1: funding]

The difficulties cited above are related to other aspects of information management: data standards, particularly those for expressing semantics such as thesauri and ontologies, and those for attaching a permanent digital identifier to an object. Search is often also related to visualisation as the next step in the discovery process [qv R2]. All these areas need to be supported.

*There are many paths to be explored in resolving difficulties in this area, and we highlight below some areas where action and resource might be fruitfully directed.*

There is a clear need for further reliable **registries and catalogues** of repository resources which are available. These registries and catalogues are elements of information management, and should be integrated into information management frameworks, and their establishment, design and maintenance should be informed by information management experts, at the least, and should be woven into wider information management frameworks. These catalogues and portals in themselves form a subclass of repository – and therefore appropriate funding needs to be made in them and in their maintenance as long as they are of value. A **European repository portal** should be considered, and would itself form part of the infrastructure of digital repositories for e-Science as discussed under recommendation 2.

[5] As reported by Prof. Dr. Martin Hofmann-Apitius in his presentation, "Paradigm changes affecting the practice of scientific communication in the life sciences", at the conference on Scientific Publishing in the European Research Area, 16-17 February 2007, and reported in the conference proceedings (p.16).
[6] See http://www.biomart.org/

The use of ontology-based systems is relatively new and could be further supported to improve the search capability of tools. In some areas the semantic level of the discipline has still to be developed as a tool, and this may be a fruitful area of investment.

Searching across repositories is facilitated by the encouragement of shared standards for search engines into repositories and the more formal federation of repositories and collections for the purpose of wider discovery. This might be a multi-step process, first looking at federation at community level, exploring models, issues, and then review of federation methods (including in the light of discovery/exposure tools).

Over the life of an information item it is likely that it will be migrated between different repositories (as might whole collections). There is also the question of discovery across different versions of an object, and also across copies held in different places (which may offer different support). Thus discovery needs to be supported on multiple dimensions, including moves in data location (and including from active data stores to archives). Locating resources which are not permanently attached to one storage location would be facilitated by developing permanent further digital object identifier technologies (PDOIs) and the means to track the changes in locations to which they refer. (There is also a granularity issue to be explored here, to develop identifiers which can point to different subparts of digital objects in a standardised way, and the question of feasibility of supporting identifiers to a high-level of granularity.)

Further research should be promoted into linking along the **data-to-information chain**, from raw data to final publication, in a seamless manner and regardless of where items may be stored, and how the linked navigation could enable not only validation of reported findings against source data, but also deeper or wider drill-down into the data, beyond that presented in the final publication. To date the final publication has been stored in a different repository to the supporting data, and these repositories are likely to be of very different natures. Related investigation and research into end-to-end data-information efficiency and how the whole information chain could be "bundled" into a virtual single digital object structure are extremely important. These developments have implications and opportunities for research libraries and publishers in particular, and for the economic and financial structures underpinning the repositories landscape [qv R1: funding; qv: R2, information space, statistics]. In this context we strongly commend existing initiatives, such as in crystallography, the e-Bank project and its successors in the United Kingdom, the work at the Fraunhofer Institut, and the worldwidescience portal, whose participants include the VTT Technical Research Centre of Finland, the Institut de l'Information Scientifique et Technique in France, Germany's National Library of Science and Technology and the British Library[7].

These developments also have implications for **document formats**. While recognizing the important role played by PDF (portable document format), and the search functionality it supports, the format is limiting and currently does not support strong navigation and re-use, nor is it currently compatible with an object bundle of a virtual single object. This format issue, while appearing to lie outside the field of e-Science digital repositories, is of critical importance to the value of the contents held in general and also specifically the level of e-Science use derived from items in digital repositories. We believe it is also an important point for the commercial scientific and educational publishing industry.

**Discovery through publications:** Publications, whatever their underlying model (be it "traditional" or for example as e-prints) and however they are distributed, play a huge role in the discovery and use

[7] http://worldwidescience.org/about.html

of e-Science data. The final publication contains information describing the science – in other words, provides key metadata which can easily be read. They can also be "scraped" (that is, computer "scripts" or applications can automatically collect structured information which can be used to support discovery). Both our desk research and the study's public consultation found that the cost of subscription journals meant that access was limited to those who could afford to pay.

**Scalability:** There two important points here. First, further challenges are to enable these discovery and navigation capabilities across the very large scales of today's and tomorrow's science. Secondly, better discovery, greater awareness of repositories and their holdings [qv R10] will increase demand for materials and support, which may place strain on repository resources. Similarly, science moves on, and heavily used resources in one decade, in terms of both the materials but also the people supporting those materials, may be little used the next. This is where the e-Infrastructure [qv R2] can play an extremely important role (particularly the infrastructure of shared or generic resources). Scalability up and down is important – demand for some datasets can grow to a peak and then shrink again; this has resource implications, in particular for staffing and professional development.

**Co-ordination:** To accomplish all of this will involve the engagement of multiple skills, areas of expertise across multiple various disciplines, interface development, systems architectures and software engineering. The development of a European layer discussed under R2 will be a major facilitator in bringing this all together [qv: R2, R6, R9 (governance), R12 (international)].

Work in this area should actively engage information scientists, publishers, and learned societies.

**Training** is of fundamental importance in these areas. It is particularly important that there is cross-training of librarians, repository managers, and data collection managers, to enable them to provide a better service to users. [qv R11, training on legal issues]

# 5. Open access to publicly funded data

The study's consultations, workshops and desk research affirmed the importance that data obtained from publicly funded science should be openly available for consultation and re-use. Digital repositories provide an efficient means for enabling this (and also provide added value).

**R 5: Publicly funded data should be free at the point of use to the user. It should be available for open access, except where required otherwise for confidential, ethical or security reasons or during a period of privileged use for the generator of the data.**

The application of this principle should be modified only in specific, well defined circumstances, including where there are concerns for the protection of individuals' privacy (as exemplified with medical records), where there are other ethical concerns (such as to protect fragile ecosystems), or security risks.

It is reasonable and important to allow a period of privileged use by producers of some researcher-generated data. This also encourages data producers to release their data.

Barriers in the past to open access to data included reluctance to make materials available for sharing or re-use. We believe recognition of data-related work [qv R8] and frameworks and tools to support recognition of rights relating to data are of fundamental importance in this regard [qv R11, legal issues].

**EU-funded materials:** The public consultation in particular stressed the importance of a single point of discovery of and efficient access to materials generated from EU-funded work (thus endorsing plans for an EC institutional repository). Discipline-specific or specialist materials should have specialist management and may thus be more appropriately held in the relevant repository on behalf of the Commission, while maintaining discoverability through the EC portal or equivalent discovery mechanism.

# 6. Collections management, selection and appraisal for sustainability

There is an exponential increase in the volumes of information being generated, stored and thus accumulated. We are now able to generate more, including new types of data. As well as accumulating a richer store of material and information, this trend also represents a risk for the management of information and its costs, for organisational stability and for the basic long-term sustainability of the information in economic terms.

Sustainability of repositories, data collections, indeed of our wider information infrastructure and heritage depends on good collection management and responsible stewardship. It will be impossible to keep everything. Collections policies, management of these policies, and co-ordination of policies and stewardship over the appropriate dimensions, and underpinned by far-sighted collections governance, are of critical importance, both for the quality of collections and for the sustainability of collections and repositories (in the short term, and over time). Aside from geography, both the compound nature of many e-Science objects and the increasing capability to span over the various stages of an information bundle, from data to processed data to publication, mean that collections policies and management will need to be conducted across what are at the moment multiple organisational and professional levels. [qv: R2 – co-ordination]

**R 6: Research is needed into scientific information appraisal for selection into digital repositories and subsequent reappraisal (the criteria, processes, automated support tools, possibly even different approaches for the digital information age). This should draw on an established body of archival expertise dating back hundreds of years for analogue documents.**

**More generally, management structures and automated tools needed in the future should be charted and planned for now, to cope with the increases in volume and for organizational stability.**

Such action needs to be sensitive to the varying needs of different communities and various repository institutions.

Managing and curating e-Science collections is particularly challenging, as their data are very likely to require other elements in order to be retrieved, consulted and used. These additional elements will also need to be managed so that they are discoverable, accessible, usable, in many cases including over time.

A collection is not the same as a repository. A repository might hold several data collections, or parts of one or several collections. This has important implications for management (and in turn also for funding [qv R1]). The roles of digital repository managers are not the same as collection managers (though one person may perform both functions). So there is a need for clear governance and communications frameworks between these functions, and between the providers of the various elements which enable retrieval and re-use of an item. [qv R2: co-ordination]

Many e-Science data collections are **global** in nature, so collection management and appraisal will need to be conducted in a global context (see Arabidopsis example overleaf).

Models of sustainability for collections (and therefore of the items within them) which employ distributed multiple copies of materials, as exemplified by the LOCKSS system (Lots of Copies Keeps Stuff Safe) and the Portico initiative are worthy of investigation and support.

The membership and governance of the collections management board (or equivalent) should take into account future stakeholders (not yet born), and the difficult question of which materials should be retained. Indeed, given the opening of repositories to form part of a multi-disciplinary fabric of resources from data to information, collections policy, strategy and management are community issues, at national, regional and global levels.

Research is needed into **automation of tools** to support collection management and the appraisal process, to offer ways of coping with the issue without the training and deployment of significant human resources (i.e. cost).

**The Arabidopsis Information Resource – a global resource**

Arabidopsis is a model organism used in plant biology and genetics;
it was the first plant genome to be sequenced

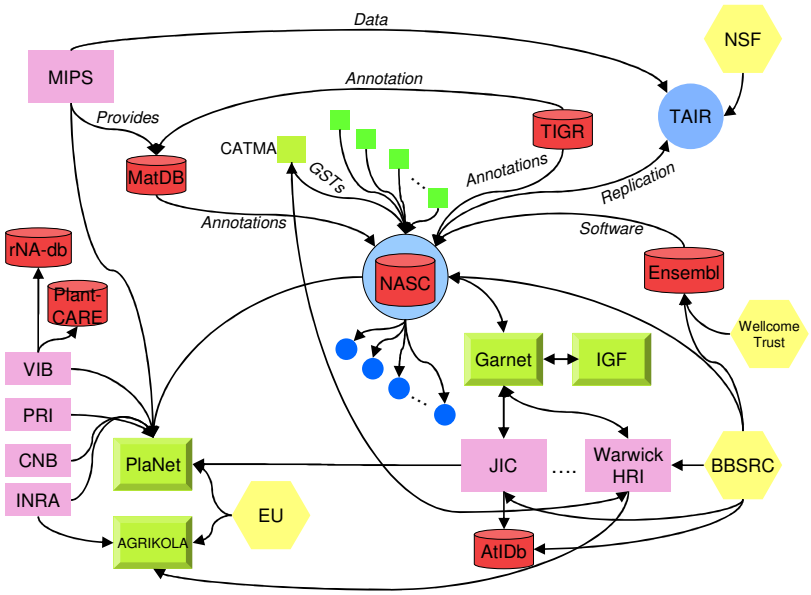



***Diagram from Lord, P., Macdonald, A. et al, Joint Data Standards Study***

***Acronyms:*** *Community resources / databases:*

| | |
|---|---|
| *AtIDb* | *Arabidopsis thaliana Insertion Database* |
| *MatDB* | *MIPS Arabidopsis thaliana Database* |
| *NASC* | *Nottingham Arabidopsis Stock Centre* |
| *Plant-CARE* | *European plant promoter database* |
| *rNA-db* | *European ribosomal RNA database* |
| *TIGR* | *The Institute for Genomic Research database* |

*Projects, joint initiatives, programmes:*

| | |
|---|---|
| *AGRIKOLA* | *Arabidopsis Genomic RNAi Knock-out Line Analysis* |
| *CATMA* | *Complete Arabidopsis Transcriptome MicroArray* |
| *GARNet* | *Genomic Arabidopsis Resource Network* |
| *IGF* | *Investigating Gene Function* |
| *PlaNet* | *Network of European Plant Databases* |

*Organisations:*

| | |
|---|---|
| *CNB* | *Centro Nacional de Biotecnologia* |
| *INRA* | *L'institut National de la Recherche Agronomique* |
| *JIC* | *John Innes Centre* |
| *MIPS* | *Munich Information centre for Protein Sequences* |
| *PRI* | *Plant Research International* |
| *TAIR* | *The Arabidopsis Information Resource* |
| *TIGR* | *The Institute for Genomic Research* |
| *VIB* | *Vlaams Interuniversitair Instituut voor Biotechnologie* |
| *Warwick HRI* | *Warwick Horticulture Research International* |

# 7. Preservation of digital information

Preservation is a key element of the durability of digital materials and their sustainability. In the context of this study, sustainability refers to the stability of repository frameworks which hold and manage information, and also to the sustainability of the digital collections they hold. Preservation refers both to the actions needed to ensure that information is not made inaccessible and useless by "digital decay" over time brought on by ever-changing technologies (even if held successfully in a sustained repository or collection) and to ensure the successful result of those actions.

Over the last decade, recognition of the digital preservation issue has steadily increased, and risen in the priorities of policy makers. Much work has been done, in particular through the auspices of the European Commission, but the problem is still unsolved at fundamental technical, organisational and financial levels.

**R 7: We recommend that the European Commission and member states increase their level of investment into digital preservation research in the context of e-Science digital repositories and the emerging frameworks around these.**

Preservation activities will also need to cover preservation[8] of tools to use data, and we commend here initiatives such as OMII and OMII-Europe[9].

Repositories are the point at which digital preservation failures are likely to manifest themselves. Furthermore, the preservation challenge is particularly great for scientific data (heterogeneous, complex, distributed, large). Those working in and with e-Science digital repositories have in-depth information and knowledge of dependencies which are critical to addressing the preservation issue. E-science digital repositories and their frameworks therefore represent an opportunity for access both to the problem and to relevant knowledge for digital preservation work.

Repositories should have sufficient funding for repository staff to participate in, be informed about and learn about digital preservation, relevant frameworks, tools, standards, and processes. [qv R1, Funding]

Increased support in this area will enable Europe to remain at the forefront of expertise in this area. Digital preservation expertise is critical to the short-term and long-term health of society and the economy, in enabling digital materials to be preserved in the first place, and at acceptable cost. If the cost of digital preservation is very high, the cost of preserving collections of digital materials risks becoming unaffordable.

Preserving the ability to use e-Science data will require the preservation of many elements beyond the data itself, such as project-specific software and calibration archives. There is substantial work being conducted already. The work done by projects such as Caspar and PLANETS, initiatives such as the Alliance for Permanent Access to the Records of Science, and institutes such as OMII-Europe are extremely important (so again we see the need for co-ordination [qv R2, e-Infrastructure]). More investment is needed, at discipline level and at generic level, to investigate how digital materials can be maintained over time.

[8] Which could include the enabling of emulation of software or functionality

[9] http://omii-europe.com/ and http://www.omii.ac.uk/wiki/AboutUs . OMII's mission is "to provide software, support and sustainability to the UK research community", keeping software and providing support for it.fs

Consideration should be given to supporting more than one preservation framework, to reduce risk. We also recommend that the European Commission support "blue-skies" thinking about digital preservation.

The specific community (discipline) context is important for successful digital preservation. Again, this has funding structure implications. Difficulties of access and re-use of old materials, because of digital preservation issues, will manifest themselves at the interface between user and the point of access to the material: in the past, national libraries have been the traditional home for heritage materials, and several national libraries in the EU are investigating how they can perform this role for data as well, and are participants in EU digital preservation projects.

#### Preservation and clarification of a repository's roles

Our overview of e-Science repositories revealed many instances where a repository's position and role vis-à-vis digital preservation was unclear. We strongly recommend that the responsibility of digital repositories, individually and at community or national levels, be clarified with regard to their digital preservation and archive roles. This information must be clearly stated in the repository's policies and description (whether and to what extent it forms part of the repository's function) [qv R9, governance, management]. This clarification will identify gaps in provision for the long-term archiving of datasets, which is a matter of governance and stewardship, and basic asset management.

# 8. Trust and recognition

Issues of trust and recognition are essential elements in increasing the use made of repositories. Depositors of information need to feel they can trust the institutions to which they commit their materials; users need to feel they can trust the materials that they find in repositories – that it is the information is not only of good quality, but that it is an also authentic, true record of that science and is passed on to them in an uncorrupted form. [qv: R7 preservation]

Professional and scientific recognition is also important for data producers, depositors and those working in data management. At the publication end of the science process, in the past recognition has been achieved by publication in peer-reviewed journals, and from citation of the publication. Data deposited in repositories has not had these incentives and quality controls.

**R 8: We recommend investigation into measures of review and recognition of data as well as publications, and also mechanisms of recognition for an individual's work in data management (whether directly in science, research or teaching, or in data management services), such as data citation with the aims of increasing trust and levels of use of repositories.**

Peer-review mechanisms for data (or similar) and data citation may help incentivize collection and preparation of good-quality data for collections. Data citation will require frameworks and mechanisms, set up in co-ordination between the relevant parties and stakeholders involved.

### Citation tools could carry administrative and rights information

We suggest that the research into data citation also investigate the possibility of citation tools that can also carry rights management and administrative information. The cost of such a mechanism is likely to be offset in good part by savings at the repository, for example, in seamless, rapid ingest of high-quality materials, and in gains in quality of science performed using the data, by the originator and downstream. Criteria for data citation should include use of community and open standards where possible.

Establishing such mechanisms would raise awareness of repositories, and the quality of repository provision and content.

**Data integrity:** We also recommend research into (i) mechanisms for checking for errors in data, and (ii) research into risks to and protection of the integrity of data as it is gathered from multiple or large-scale sources and when it travels across e-Science functions, and (iii) provenance tracking, as data passes from machine to machine, repository to repository. This is increasingly important given the growing and accumulating volumes of data and metadata.

**Certification:** The European Commission should continue support for work on methods and processes for repository certification. We note that important work exists and is underway into the requirements for trusted digital repositories, at international level and in Europe. We recommend that any certification processes adopted in Europe should be kept as light as possible: burdensome processes are counter-productive.

In due course, there will be an opportunity for the EU and/or member states to provide digital repository services with an EU Repository "CE" mark. Users and depositors will see the CE mark

and be more willing to deposit materials; they will also set greater store by the materials retrieved from the repository. The service could charge commercial institutions for certification.

The benefits of these measures include:

- Better understanding on the part of data producers of data science and data management issues
- Growth in data-management skills pool
- Greater retention of skilled staff
- Lower costs for repositories, releasing resources for more value-added work
- Higher quality of holdings in repositories.

# 9. Governance and management

Repositories need to be well managed and have appropriate governance regimes in order to be efficient and trusted custodians of the European scientific heritage and critical datasets and information, but also to sustain appropriate funding, and help funders to make well-informed decisions about forward support.

**R 9: We recommend that the European Commission encourage member states, through their institutions and agents, call for measures to promote and sustain the good governance and management of and reporting by repositories.**

Digital repositories should have a clearly defined remit and responsibilities, with matching policies (and target service levels for their customer groups) for performing their roles, vis-à-vis users, data suppliers, and collection owners. This includes clear identification of a repository's role with regard to (i) liability for the items they hold and for re-use by third parties, and (ii) preservation of its digital holdings (which theoretically might be nil) [qv R7, preservation].

To sustain their funding and for good resource management and planning, repositories should submit budgets and regular formal reports and accounts to their funders, using pre-agreed metrics which are not too onerous. Above all, we believe repositories or their umbrella frameworks should engage in active communication with their funders, as well as customers and stakeholders. They must articulate what they do (often not very visible) and the benefits of what they do. Low use does not mean low value.

Repositories should have a governance structure which reflects their responsibilities and activities, objectives, business plan, strategy and operations plan, and relevant wider contexts and frameworks.

All these requirements have yet to become commonplace. We therefore also recommend that the European Commission consider instituting the development of a **code of good repository practice**. [qv R2, e-Infrastructure and co-ordination, R8, Trust].

# 10. Training and awareness

We strongly recommend training, Europe-wide, and at multiple levels:

- Within repositories, at management, operational, technical and information levels (in software, tools, science, and in information science)
- For users, from scientists, teachers, students, to lay people
- Depositors (again, across the full spectrum of actual and potential depositor groups)
- Librarians and information managers

In addition, there should be education and awareness-raising in schools about e-Science data, data management, e-Science, e-Science repositories and their services; use of data, including e-Science use, should be introduced in schools. These will bring substantial long-term benefit to Europe. Teenagers with access to computers are already learning a lot about semantic tagging, software programming, and also taking part in an emerging culture of sharing. We believe training and awareness are core to the deepening and broadening of a European Research Area, where young people grow a sense of awareness of data curation, responsibilities and benefits, and interest in building and maintaining our corpus of information and knowledge. This training will grow good habits, in turn contributing to the sustainability of e-Science resources through decreased load on providers and curators and a larger and stronger competencies base.

**R 10: We recommend that the European Commission, member states and their agencies actively support and encourage training in good data management practices at all levels, including outreach by e-Science digital repositories and teaching of such skills at an early age, training in aspects relating to the use of materials held in e-Science digital repositories.**

**We strongly recommend that the European Commission and member states support cross-training between the professional areas of science, computer science and information science.**

Nowadays, training can be provided very cost-effectively by using the internet, through podcasts, webcasts, to a wide spectrum, whether unaffiliated, old or young, and emerging new types of stakeholder groups.

Of particular importance in the European context is training in the home language, and the exchange of ideas and materials across language borders. [qv R2, co-ordination]

There should be pan-European, multi-lingual actions to promote awareness and visibility of e-Science digital repositories, in science, education and to the wider public.

### Additional benefits

Training not only increases the quality and the level of use, but it also provides an excellent conduit for feedback to providers.

# 11. Legal issues

Laws and regulations can impede access to and use of data and tools, in multiple ways.

Legislative differences exist across and even within national borders. In contrast, scientists work globally. However, in today's information age, a large proportion of e-Science materials and activities work across national boundaries, the boundaries between different legal jurisdications and also administrative systems.

Researchers and research groups have to interrupt their core work and take multiple different legal requirements and frameworks into account when working. Planning research projects which have to straddle different regulations is onerous and time-consuming. Transfer of information and materials is substantially slowed by the lack of harmonisation across national and also local borders, in particular by diverse intellectual property frameworks (in particular copyright) and conditions relating to management and use of clinical data.

**R11: We recommend that the European Commission pursue review into harmonisation of legal frameworks relating to intellectual property and to copyright in particular across the EU and the EEA, and we also endorse calls for a more fundamental review and analysis of the nature of intellectual property and copyright in the digital age.**

**Secondly, we recommend that further research and international discussions are held on rights management expression tools for rights relating to the use of data in e-Science contexts, which can be supported at very low transactional cost.**

**We further recommend provision of a clear, simple multi-lingual information source and basic training to all repository providers, higher-education students, scientists, teachers and more widely on the basics of intellectual property, the different types of licences that can be used, and laws which might apply to e-Science repositories and their holdings.**

### Simple training in basic legal issues

The lack of basic understanding of intellectual property frameworks, rights management, and about the different licences commonly used in publicly-funded research and education was frequently raised in the study's public consultation and workshops.

We recommend that basic training in intellectual property and licences should be provided to all higher-education students, with simple, up-to-date supporting guidance materials readily available to all stakeholders, in their own language, not just on IP but also on other legal and regulatory issues that may arise (for example, data protection acts, freedom of information, cross-border transfer of materials, environmental information rules, etc.) [qv: R10, training]

### Consultation, co-ordination, clarification

We recommend that representatives of digital repositories and/or e-Science data communities should be consulted in the context of drafting of legislation which might affect access and use of publicly funded data. [qv R2, e-Infrastructure, co-ordination] The legal status of individual repositories should be clarified, including their position vis-à-vis liability for use of their holdings.

**Benefits:** These measures will help seamless, dispute-free cross-border collaborations, and will help support machine-to-machine working, and encourage greater willingness to submit data to repositories.

## 12. International

E-Science in the 21st century is global in nature. The tools and standards which support data use have been developed in a global context, and the e-Science conducted using these data and tools is frequently done in international collaborations.

A significant proportion of the data held in e-Science digital repositories forms part of global collections, whose management is thus per se an international matter. 80% of global biodiversity data lies outside Europe, North America.

Affirmation of the global nature of certain data collections and support for them is particularly important in the face of global challenges such as climate change.

Engagement with developing nations in this field of repositories and their supporting infrastructure is an important geo-political and strategic opportunity, as well as part of a global responsibility. Europe has many world-leading institutions and experts in e-Science and data management, and runs, supports and participates in such initiatives, large and small.

**R 12: We recommend that the European Commission urge member states and their agencies to adopt policies in favour of an active role in the stewardship of global data.**

**Given the global nature of e-Science collections, data and working, we recommend that the European Commission, member states and their agencies consider establishing a designation of World Heritage Data for data of particular significance and conduct a review into how archives for these data might be supported.**

Equally, trust and governance [qv R8, R9, R11] are key to participation by other countries in sharing of data.

Several key informants stressed the importance of international co-operation in addressing the problem of preserving scientific data [qv R7]:

> *"In my opinion international cooperation is the only way to save, collect and provide all past, present and coming scientific data."*
> *(Response, e-SciDR public consultation)*

## Priorities

We conclude this section with suggestions for the relative priorities to be adopted to implement these recommendations. Perceptions of priorities of course vary across different stakeholders – we adopt the viewpoint of policy makers, whether at the European level or within member states. We have mentioned on a number of occasions that for the most part science is an international endeavour, but that its management, policy and funding structures are in contrast mainly conducted at a national or institutional level. The need for international collaboration and co-ordination is reflected in the starting point for this study, its title and vision of a European e-Infrastructure.

We also note in the context of this study it behoves policy makers to think and act with a global perspective in mind; if they do not the scientists and science will drift to where this perspective is better acknowledged.

The downstream headline benefits of establishing a world-leading repository infrastructure in Europe are expected to be:

- Improved science and improved scientific productivity and impact
- Improved value for money from investments in science
- A securely preserved scientific heritage and better assured scientific validation
- Improved economic performance through the availability of key data and more productive research

# I. Relevant technologies – a summary

Quite clearly technologies are a central issue for digital repositories. The technical requirements for e-science digital repositories include:

1. Technologies to allow appropriate access to repositories, while respecting privacy and intellectual property rights;
2. Facilities to identify target repositories and to get information/data[10] and their metadata into repositories ("ingest");
3. Tools to store, manage and preserve information once it is deposited;
4. Methods to identify and locate repositories of value to users and methods to find, display and extract information/data within repositories/collections and then re-use them if appropriate. ("Dissemination".)

The various technologies studied and discussed here address these four functional areas, and have varying degrees of relevance. Thus ICT technologies such as storage and repository management systems are quite clearly of vital importance, while others have less or peripheral relevance. Those technologies also with a central role include networks, ingest (deposit) tools, search engines, authentication and authorisation systems, information packaging and presentation tools.

## Terminology

Many ambiguous terms are used in this arena and these we have avoided. For example, there are multiple, common but different definitions of the terms "middleware" and "grids" (see glossary).

When we use the word "**tools**" we mean software with a particular purpose related to use of a repository. Just one example is BLAST, which is software to find similarities between genes and gene products in biological repositories[11].

Technologies evolve. By "**migration**" we mean the succession of technologies, one following another over time, that are used to support digital repositories[12].

**Technical standards** are closely related to specific technologies. Standards are described below; we confine ourselves here strictly to the technologies *per se* as much as is possible while noting that it is sometimes difficult to differentiate a technology from a standard (XML is a good illustration – it is usually thought of as a technology, but is also a standard for mark-up.[13]).

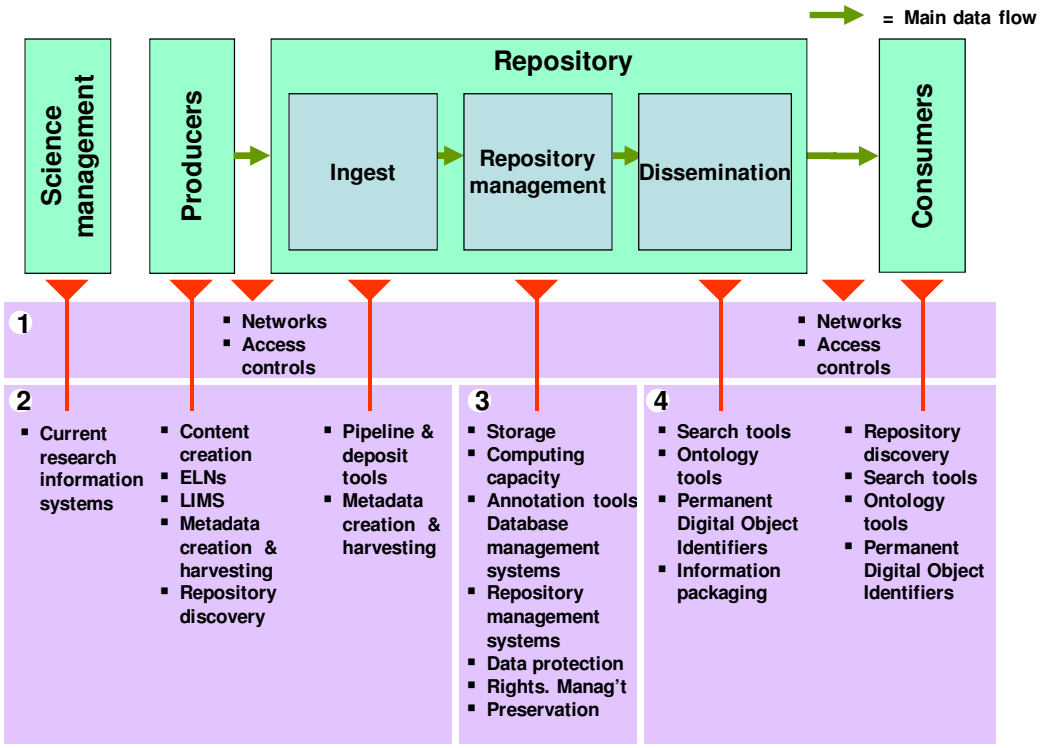
## The technologies

As noted above, we list relevant technologies in a functional sequence. We provide brief definitions and reviews, and some (non-exhaustive) examples. We comment, when relevant, on the situation in Europe, on interoperability and migration as defined above.

The summaries provided below are extracted from more extensive information provided in IR1.

---

[10] We do not make a strict distinction here between data and information.
[11] Basic Local Alignment Search Tool See: http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
[12] Note that the term migration is used by the digital preservation community to transform information content from an obsolete standard to a new standard as technologies change. We adopt a wider interpretation of the term here.
[13] See: http://www.w3.org/XML/




**Relevant technologies**

The figure above illustrates the classes of relevant technologies identified, placing them in the research information flow from producer to consumer, together with science research management functions[14]. In this diagram various technologies are listed in four numbered boxes (corresponding to the list of functional areas given above). These are shown below the relevant stages in the process from planning and research management, through creation of data, repository storage, to use.

## 1. To get access to digital repositories

Getting access to resources is common to all stages in the process, see box 1.

### 1.1 Networks

Networks are fundamental to repositories, providing access for users and enabling repository interoperability. In the publicly-funded research area in Europe network access is provided by the various National Research and Education Networks (NRENs)[15]. At a European level a network of these networks is provided by GÉANT2, managed by the DANTE organisation and is funded by the EU.

[14] This figure is based on the Open Archival Information System model (ISO 14721:2003) which adopts the term "ingest" for those functions which admit information into the repository and "dissemination" for those functions which supply information to users. The OAIS functions of Management (of the repository), Administration, Storage Management, Data Management and Preservation Planning are assumed included in the Repository management cell.

[15] In the Nordic countries access to NRENs is provided via NORDUnet, a regional network.

### 1.2 Access controls: Authentication, Authorisation and Auditing/Accounting ("AAA")

These commonly refer to the processes of

- securely identifying an individual user: *authentication*;
- defining what a user is allowed to do and enforcing this: *authorization*; and,
- recording, monitoring and checking what the user has done (*auditing/accounting*).

Security underpins e-science and is critical to ensuring that the intellectual and potential commercial property associated with digital repositories across Europe is protected to the satisfaction of users, data owners and data publishers, among other stakeholders. In this context, Europe and numerous international communities are currently in the process of deploying Shibboleth technologies (http://shibboleth.internet2.edu/) to support local (existing) methods of authentication for remote login to resources. Further developments aim to enable virtual organisations (VOs) to be established across institutions and national borders, either on a temporary or a more permanent basis.

Any security infrastructure for digital repositories in Europe must be easy for end users, data providers and data owners. An ideal e-Infrastructure would be as intuitive and non-intrusive for end-users as is the World Wide Web.

The proliferation of new web service security standards and technologies also needs to be reconciled with European needs for digital repositories. It is unlikely that all digital repositories will adopt the same security infrastructures.

## 2. To get information into digital repositories

### 2.1 Pre-deposit phase technologies

This represents a collection of technologies which create the information which is to be kept in repositories, and which manage the information production pipeline. While not directly related to the repository technologies themselves, they influence repository effectiveness by:

- Delivering content in formats suitable for repository management and for repository users, and for eventual preservation;
- Creating metadata of value to repositories for information management, sustainability of content and discoverability;
- Providing integration with downstream repository systems, thus easing the delivery of information.

Technologies which fall in this category include:

#### 2.1.1 Current Research Information Systems (CRISs)

CRISs are systems that manage the information that underpins research, bringing together information on the complete workflow, from funding application up to and including peer-reviewed publications. Some of these systems are themselves repository management systems (see below).

#### 2.1.2 Content creation systems

These are essential in the information flow, and can help by producing standards-compliant data. They include a myriad of software systems and instrumentation types, too numerous to list.

### 2.1.3 Electronic Laboratory Notebooks (ELNs)

While CRISs take a global view of the research process, ELNs are used to manage specific data capture processes in the laboratory—as their name implies they are attempts to replace the paper laboratory notebook with an electronic equivalent. They are relevant to repositories because they can provide at-source metadata, can link to related information, and can generate well-structured, consistent datasets to appropriate standards. They form local (private), temporary repositories which can pass on content and metadata to downstream repositories.

### 2.1.4 Laboratory Information Management Systems (LIMS)

LIMS serve a rather different function from ELNs (though they are closely related and are often confused). LIMS manage scientific data once it has been captured, rather than when it is collected. They coordinate and relate data gathered across processes and projects. Their significance is similar to that of ELNs; like them they are themselves repositories, but with a local, often temporary, scope.

In the longer term perhaps LIMS and ELNs will become integrated, and possibly merge with repository systems.

## 2.2 Metadata creation and harvesting tools

This represents a large class of diverse tools. Metadata can be gathered in four ways:

(i) generated by systems which create, or manage the creation, of data content;
(ii) be supplied manually;
(iii) extracted automatically from the content itself; or by
(iv) harvesting existing metadata from elsewhere.

For reliability and utility, it is best if metadata is generated at, or even before, the point of content creation by tools such as those described above. This is especially true for metadata that is critical to increase the longevity of the information. Metadata assignment is difficult and often labour intensive, and thus inhibits digital repository use; this must be must tackled at both generic and domain-specific levels

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[16] has gained wide acceptance as a means of harvesting metadata from existing repositories. The OAI-PMH system allows metadata from repositories to be queried and extracted, but presupposes the donating repositories provide OAI-PMH services to handle requests. In more specialist environments more specific tools will be needed. As one respondent stressed in a response to the study's public consultation, OAI-PMH is not adequate to support metadata-harvesting from the typically heterogeneous scientific data formats.
(See for example the work of the commercial OASIS[17] consortium, its XML Metadata Exchange[18] (XMI) and Web Services Metadata Exchange[19]).

## 2.3 Discovery or identification of digital repositories

Scientists need to identify the right repository for their data. Users need to know in which repository to look for it—see 4.1 below.

---

[16] See http://www.openarchives.org/OAI/openarchivesprotocol.html
[17] See: http://www.oasis-open.org/home/index.php
[18] See: http://xml.coverpages.org/xmi.html
[19] See: http://xml.coverpages.org/ni2004-03-05-a.html

### 2.4 Pipeline and deposit tools

These include tools to enable content and associated metadata to be transferred and incorporated into digital repositories, and tools to help with data entry through to deposit. These tools have been developed to improve information quality and consistency, and consequently reduce the effort, time and cost of deposit. An example of a tool developed with this in mind is MaxDLoad2 for depositing microarray data. Systems like this are likely to be dependent on the discipline context in which they are used.

## 3. Store and manage information once deposited

### 3.1 Storage

Mass storage technologies continue to shrink in size and cost for a given amount of storage capacity. At the same time the volume of information being generated is rising exponentially. These trends are likely to continue for years and have huge implications. Ever increasing amounts of information can now be deployed locally or even be carried by individuals. Once it becomes possible to carry the equivalent of the contents of the printed contents of the Library of Congress in one's pocket worlds of new possibilities unfold. New methods of managing mass storage are emerging, such as virtualisation, storage area networks and data grids and clouds. These trends of course bring attendant issues such as maintaining security, and maintaining control over the authenticity and validity of information.

### 3.2 Computing capacity

Moore's law is, for the moment, still providing increasing computing power for the same costs. Increasing computing power, and harnessing of supercomputers and compute grids offer greater powers of analysis of the information in repositories and for collaborations. These technologies can be applied to repositories, providing for example, more sophisticated searching capabilities, and further integration of repositories into data analysis activities.

### 3.3 Database management systems

Commercial and open source database management systems are of relevance as the data management foundations upon which some (but not all) repository management systems are built.

### 3.4 Repository management systems

Repository management technologies range from simple file system management tools to sophisticated inter-organisational data management systems. They manage the storage of digital content and its metadata and the services provided around the repository, such as deposit and access.

There are many types of repository management systems: at one end of the spectrum they are simple storage systems managed by simple tools supporting manually mediated processes. At the other end, they can be complex software for digital object management delivering many added-value functions. A subclass comprises Digital Library systems designed primarily for document-type data.

As noted, the terminology varies, and we include here systems which are designated as digital library systems, institutional repository systems, and object stores and similar. Some systems labelled with other names are also repository managers, such as the LIMS or CRIS systems referred to above.

Another subclass is the special systems which have been developed to manage specific classes of digital repositories or data types, such as the Ensembl[20] system at the EBI for

[20] See: http://www.ensembl.org/index.html

genomic data, DEAL[21] from Common Data Access, a system for oil well exploration data in the North Sea, and others.

There are many repository management systems available – the OpenDOAR project[22] (late 2007) lists 62 different repository software systems in use world-wide to manage 861 open access/OAI compliant repositories. Interestingly of the 1009 repositories listed, 259 (26%) did not record the system in use, and just two systems were used to manage 453 installations (45% of the sample), ePrints and DSpace. The list provided by OpenDOAR is by no means exhaustive, and in the main only covers repositories of documents (71% of those listed).

Europe is not lagging in this area—witness the high-take up of ePrints, developed at the University of Southampton in the UK.

We note that in some instances a repository can exist without a single, packaged specialised system to manage it. What is essential are to provide the core functions for content ingestion, digital object management, and access.

### 3.5 Annotation tools

We distinguish the (relatively) simple tools to update metadata in repositories from those which provide *post-hoc* annotation to content in repositories, either automatically or manually. Annotation can be thought of as a class of metadata.

### 3.6 Data protection tools

Though access to a resource may be granted through the AAA systems noted above, some data in repositories are particularly sensitive, and privacy and security must be assured to a high degree (for example, personal medical records and genetic information, or data on sensitive ecological locations). The price of easy network access to repositories is that they must be protected by firewall and anti-intrusion software.

Behind a firewall, the data that needs further protection for privacy, ethical or security reasons may require further measures. Common approaches to address this include encryption, anonymisation and pseudonymisation techniques, but note the latter are often limited.

### 3.7 Rights management tools

These enable the owners and managers to protect rights that may be attached to data, and where appropriate charge for access.

### 3.8 Preservation tools

Preservation of digital information is a growing concern, and there is now much activity in the area, especially in Australia, the USA, the Netherlands, Germany and UK. A lot of this activity centres on policy and procedural issues. Development of new techniques seems to have thinned over the last few years.

Textual objects are generally easier to preserve than the great variety of experimental data sets, but such differences are often lost in the literature by lumping data under the collective label of "databases". The current emphasis in preservation technologies is on data migration in the sense of copying information from an obsolete technology format to a current one and, more rarely, emulation. Longevity also depends upon information standards used, discussed in section xxx.

---

[21] See: http://www.cdal.com/HOME/DEAL/page34253.asp

[22] See http://www.opendoar.org/

## 4. To find, extract information from repositories and re-use it

These actions can take place at, or across, three information levels: to locate repositories or collections of interest, to find items of interest within repositories/collections, and then to find information within items. We note that users may need to find information across multiple repositories and items, requiring some form of federation or aggregation.

### 4.1 Discovery or identification of digital repositories

Various technologies can be used for the discovery or identification of digital repositories and their content (or, from the repository's perspective, expose the repository and/or its contents to the potential user). Perhaps the simplest way into a repository is via a search engine, but other approaches are the use of portals or *de facto* community web-based resources. Many subject and institutional gateways have been set up, for example PlaNET[23], for the *Arabidopsis thaliana* community. Registries, such as OpenDOAR[24] provide another means of locating repositories. Management approaches and promotional materials are as much part of discovery as pure technologies.

### 4.2 Search tools to find items and information within items

These are the means by which we find, access, and visualize information in repositories. They range from simple indexing tools to data mining suites. Google is of course now pre-eminent and is most users' first point of call; but there are many other search tools, including those specialized for images, for particular disciplines (like genomics), or for visualisation (Ensembl is a sophisticated example), and other domain-specific engines.

There are also many commercial search tools and data mining systems — a notable European success is Autonomy[25], based in Cambridge.

### 4.3 Ontology tools

These are tools that assist the search and retrieval process by providing semantic level assistance, such as widening search strategies, word disambiguation, *etc*. We make a distinction between ontology management systems and the ontologies themselves (their intellectual content) – though often the usages are confused.

Ontology management systems (from the relatively simple, such as thesauri, to the sophisticated such as those that support OWL, RDF and similar standards) could transform the ways we use repositories. But they depend upon the ontologies themselves—the information, concepts and their interrelationships they embody, and these need to come from, and be validated by, domain experts to ensure that they are accurate and trustworthy.

### 4.4 Permanent Digital Object Identifier (PDOI) technologies

These include technologies to support the **use** of PDOIs. The standards to which these conform (such as the DOI, ARKs, etc) are discussed below.

### 4.5 Information packaging

Technologies for the packaging of information from repositories vary from the minimal (such as simply providing a data file), to sophisticated presentation of results such as that seen with Ensembl from the EBI[26] or any one of a growing number of "mash-ups": data displays assembled from a combination of repositories, for example GPS displays of information from

[23] A network of European databases on plants. See: http://mips.gsf.de/projects/plants/PlaNetPortal/databases.html

[24] See http://www.opendoar.org/

[25] See http://www.autonomy.com/content/home/

[26] See: http://www.ensembl.org/index.html

public sector geographical resources. This diversity makes generalisation impossible, but such packaging will become more sophisticated as computing power and network speeds increase, and markets and imagination drive developments. Adherence to standards is a key component of success in this area.

## II. Standards

Standards underpin the successful deployment of advanced digital repositories that can span discipline-specific silos of information and facilitate inter-disciplinary sharing of data. They:

- Ensure interoperability of tools and linking of data across repositories:
  - Enable interoperability between repositories at semantic and syntactic levels
  - Enable interoperability between repositories and users, both human and computer, at semantic and syntactic levels, for users and producers
- Help different communities better to exploit resources by:
  - Reducing the need to re-work data, build new interfaces and make conversions;
  - Providing consistency and ease of use for user;
  - Enhancing reliability.
- Define, and ensure compliance with, agreed conditions of use of information in repositories
- Promote efficiency and cost savings in the research environment
- Facilitate migrating systems forward as technologies and needs change.

A more detailed review of repository standards appears in IR2.

### 4.2.2 Interoperability

Interoperability between repositories can take place at a number of levels:

1. **Data interoperability** exists when data is understood, syntactically and semantically, as it is exchanged;
2. **Technical interoperability** exists when technologies can work together—for example when repository management systems can share services;
3. **Management and governance level interoperability** exists when organisations can co-operate and/or exchange management information and policies.

The interoperability of digital repositories in general depends upon making linkages at all of these levels.

Two other aspects are important:

4. **Organisational** - allowing interoperability within and between communities or subject domains, and defining any constraints to be applied;
5. **Temporal** - new or rapidly changing domains may not be able to become equal partners in interoperability exchanges: timing is important.

#### Relevant standards

The areas to which standards are of relevance to e-science digital repositories are:

**For security, including authentication and authorization controls.**

Robust standards in the area of security, authentication and authorisation for access to repositories are vital for interoperability to take place in a climate of trust. They thus form a key element in promoting

repository use in Europe (or elsewhere). The issues to be resolved are technically complex, but from the user perspective a minimum of administration and effort (including single sign-on) is essential for acceptability and take-up. Developments of significance here include the spread of Shibboleth[27] in Europe and various Web Services standards.

However standards in this area are not yet settled nor universally available or easy for non-specialists to use. These shortcomings are a barrier to more widespread information sharing.

#### Rights assertion and management

Intellectual Property Rights (IPR) technologies enforce or display licensing policies and business models for digital resource distribution and use from repositories. These reassure depositors and help them make submit into repositories. From the information consumer's point of view they are often seen as a barrier. More therefore needs to be done to make them easier to understand and use.

Standards for expressing IPR can be found as elements of other standards, primarily designed for expressing metadata more generally in various bibliographic and description contexts (see below) and include:

- DCMI (Dublin Core Metadata Initiative, ISO 15836),
- FRBR (Functional Requirements for Bibliographic Records),
- METS (Metadata Encoding & Transmission Standard),
- PREMIS and
- OAI-PMH (Open Archives Initiative's Protocol for Metadata Harvesting).

For content (data) delivery further standards of relevance include examples such as:

- Geospatial DRM[28],
- Open Digital Rights Language (ODRL)[29],
- eXtensible rights Mark-up Language (XrML)[30].

The vast majority of **data** are not held in publicly accessible repositories. Studies have shown that the most important research data sets are often held by the researchers themselves and they are often only willing to make their data publicly available once they have published papers exploiting them.

#### Information description at syntactic and semantic levels for content and metadata

Information needs to be communicated to and from repositories or between repositories, stored in some format, and when used, be understood. Standards for both syntax and semantics are vital to all of these processes. Like so many of the standards mentioned here, these are now mainly based on XML.

The divide that separates the domains of data, and publication information is quite sharply drawn in this area, and we reflect that in our notes on data and metadata which follow (See also the diagram below).

---

[27] See: http://shibboleth.internet2.edu/
[28] See: http://www.opengeospatial.org/
[29] Open Digital Rights Language Initiative. See: http://odrl.net/
[30] See: http://www.xrml.org/about.asp


**Data content** is represented in many diverse formats depending on the domain that it applies to. There are too many to list. Their diversity reflects the diverse needs of different communities and sciences - just three examples: ChemML (chemistry), MathML (mathematics), GML (Geo-spatial content). For common data types there are of course well established standards such as the various MIME[31] types, standards for images, text content, audio, *etc*. Many of the standards in this area specify the expression of both syntax and semantics.

**Data metadata:** Clearly there are metadata standards which are very specific to particular data types. Such standards often blur the distinction between data and metadata - thus, for example, the MIAME standard for microarray experiments (in genomic research) contains within it much information which could be regarded as metadata alongside the raw data itself.

**Publication content** is a much less diverse information class. Significant standards for these are XML, PDF (and PDF/A).

**Publication metadata:** There are very many standards for bibliographic and document metadata, of which the most ubiquitous is Dublin Core. The following table lists some of these standards:

| Specific publication metadata standards | |
|---|---|
| ■ Dublin Core<br>■ Electronic Archival Description (EAD)<br>■ General International Standard Archival Description (ISAD(G))<br>■ Metadata Authority Description Schema (MADS)<br>■ MARC (including MARC21, MARCXML) | ■ Metadata Encoding and Transmission Standard (METS)<br>■ Metadata Object Description Schema (MODS)<br>■ MPEG-21 (and Digital Item Declaration Language(DIDL))<br>■ Preservation Metadata Implementation Strategies. (PREMIS) |

The domain-specificity of data content and metadata standards is inevitable and useful within

[31] Multipurpose Internet Mail Extensions. See http://www.iana.org/assignments/media-types/

domains. For wider sharing demanded by interdisciplinary work, new e-Science methods are needed to bridge interdisciplinary divides. New methods of expressing semantics independently of specific information types are being developed but are not yet in full mainstream use. The Resource Description Framework (RDF)[32] is a set of standards that bring together URI's and XML to in order to provide a way of expressing relationships and meanings of uniquely identified resources. The Web Ontology Language (OWL)[33] is designed for use by applications that need to process the content of information instead of just presenting information to humans. DAML is an extension to XML and RDF (the latest release is DAML + OIL 2006) which provides a semantically rich set of constructs with which to create ontologies and to mark-up information so that it is machine-readable, understandable and supports semantic interoperability.

- **Digital object identification and name resolution**

We need to reference (identify) objects within repositories to facilitate their retrieval and use, much in the way that we can unambiguously identify a book by its ISBN. Systems to make this possible need to be permanent - that is they must be independent of changes in technology with time. The prime example of a permanent digital object identifier is the DOI (Digital Object Identifier) from the International DOI Foundation (IDF)[34]. It is much used for citing bibliographic and published content and it has the weight of publishers' support behind it, but doubts have been expressed about the viability of the economic model it is based on and its true persistence. It was developed primarily with published content in mind, and lacks standardised methods for expressing different levels of granularity within documents, such as objects embedded in larger structures (for example a specific image in a multimedia presentation). It is not unchallenged: rival proposals are the Archival Resource Key (ARK) [35] and Persistent Uniform Resource Locator (PURL) [36].

- **Information/metadata harvesting/capture.**

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) standard was described above. It could be considered a standard rather than a technology (it uses the simple HTTP protocol), but it is mentioned here because, at least for institutional repositories and digital libraries, it has gained wide acceptance.

### Repositories as managed stores

The Open Archival Information Systems (OAIS) reference model (ISO 14721:2003) presents a design and management framework for repositories that are built to retain of digital information of all kinds for the long term. This framework has reached a high level of acceptance.

- **Standards for distributed data Grid architectures**

Data Grids are arguably the most difficult to establish and manage. Reasons for this include:

- the complexity of the data itself which can often be very domain specific and require expert interpretation;
- the evolutionary nature of research and changing nature of scientific and other data sets;
- the lack of foresight and/or education by the data creators on how best to annotate their data so that it might be found and subsequently used by others;

[32] See: http://www.w3.org/RDF/
[33] See: http://www.w3.org/TR/owl-guide/
[34] See: http://www.doi.org/
[35] See: http://www.cdlib.org/inside/diglib/ark/
[36] See: http://purl.oclc.org/

- and perhaps above all, the amount of data that is being generated across all research disciplines.

This aspect of the establishment of a data infrastructure for e-Science digital repositories is, therefore, especially challenging. The focal point of data standards within the Open Grid Forum OGSA community is the OGSA Data effort. This has associated with it, numerous working groups (see IR2 for further discussion).

- **Current Research Information Systems**

CRISs capture the frameworks - financial and organisational - within which research takes place and provide tools for managing the research process. In shared endeavours they can provide a mechanism for efficiency, comparability and management of collaboration. The CERIF2000 (Common European Research Information Format) is a standard (a data model) for representing research information[37].

### Frameworks for standards

Standards are developed in many forums, both formal and informal. Often those developed informally by a community effort, driven by a common perceived need, are the most successful.

Standards not only arise in an academic environment, but arise from the needs of industry and commerce - OASIS is particularly relevant and successful. Particularly relevant here are the standards being developed for the Web Services, Web2, Service Orientated Architectures. Industries can also develop frameworks or architectures - higher level standards - that serve to integrate activities in a particular domain. A good example in this respect is CDISC from the pharmaceuticals sector.

Standards rarely start from zero—they tend to be defined in hierarchies, one standard drawing upon another at a lower level.

This and other studies show that, to be effective, the standards setting process:

- Community participation - to ensure "buy-in" by its members,
- Control - to avoid overly-complex and large standards,
- Flexibility - to accommodate rapid changes in technologies,
- Expertise - individuals of the right calibre must be allowed to attend to the job of standards development for the sake of progress.

### Gaps in the standards landscape

All the areas we have examined have been subject to some standards process. The issue of coverage is generally not one of gaps but of maturity and take-up, but significant deficiencies lie in the following areas:

- Standards that help guarantee the longevity of information,
- Widely accepted and used standards for authentication and authorisation,
- Permanent digital object identifiers that address objects at all levels of granularity,
- Accepted standards for repository service levels.

[37] See: http://cordis.europa.eu/cerif/

## III. Workshops - summary

In the first three months of the study a series of three workshops were organised to solicit the repository community's views on the issues attending, and development of, repositories in Europe. The workshops were each of one day's duration, and each was attended by some 20 invited experts, EC representatives and members of the DAC's e-SciDR team. Depending on the theme of the workshop these were drawn from a range of disciplines, roles, and institutions. They mostly come from across Europe, but representation was also obtained from North America. The three workshops each dealt with a different theme:

**Workshop 1** was held on 7th February 2007. The workshop's objective was to discuss e-science digital repositories in their wider context, draw out high-level issues to be addressed by the study. Fundamental questions explored were what is meant by the term "digital repository", and how do repositories stand in relation to similar entities such as digital libraries or archives? The themes explored were:

- Defining digital repositories – what counts as a digital repository, what are the defining characteristics?
- Roles adopted by and within digital repositories
- Accessibility, interoperability and infrastructure issues
- Sustainability of repositories and their contents
- Use of repositories and their user communities
- Sociological and cultural issues.

**Workshop 2** was held on 5th March 2007, in Brussels. The workshop's objective was to discuss technical matters: interoperability, standards and technologies in the context of e-Science digital repositories. This workshop debated:

- Standards and the representation of information
- Provenance tracking as information moves from is source through successive repositories
- Metadata standards and its generation
- Sustainability and digital preservation.

Related to these it also touched on funding of repositories, the possibilities for refereeing repository contents, and organisational structures for optimal repository development.

**Workshop 3**. This was held in Brussels on 26th March 2007. The workshop's objective was to discuss legal, economic and sustainability aspects relating to e-Science digital repositories. This workshop discussed the following issues:

- Intellectual Property rights in relation to repositories, IPR awareness and training, and the public and commercial aspects of IPR
- Copyright, "copyleft", and open access
- Permanent identifiers for digital objects
- Permanence and quality control
- Behavioural factors and incentives to deposit into repositories.

Recommendations for EC activities were also identified.

For each workshop the outcomes and ideas arising were recorded; and opportunity was given for attendees to supply further thoughts and opinions after the workshops, and many replies were received in response.

A full report on the initial workshops and their outcomes is provided in IR1.

### The final study workshop

The final Study Workshop was held on the 4th September at the National Archives of Portugal (at their "Torre de Tombo" building in central Lisbon). The timing and venue of the workshop was arranged to coincide with the Portuguese Presidency of the EU.

The purpose of the Workshop was to review the findings of the study, and to give an opportunity for the invited members of the repository community, broadly drawn, to comment on draft proposals and recommendations for incorporation in the study's final report. The meeting was structured in the following way to achieve these objectives:

- Welcome to the National Archives/Torre do Tombo from Francisco Barbedo, Depty Director, Direcção Geral de Arquivo:
- Welcome to Lisbon and behalf of the Portuguese Presidency of the EU from Pedro Ferreira, Director of UMIC, Lisbon: .
- Presentation: "Towards a European e-Infrastructure": - orientation from the European Commission by Carlos Morais-Pires, European Commission:
- Keynote address from Herbert Van de Sompel, Los Alamos National laboratory, USA:
- A presentation giving an overview of the e-SciDR study and the draft recommendations, presented by the study team
- Three parallel breakout sessions structured to ask the participants to answer the following questions:
    - What are the three most important of the recommendations put forward?
    - What is missing from the recommendations?
    - What has the lowest priorities?
- Reporting back to the plenary meeting the results from the three break-out sessions and an open discussion in plenary session
- A closing address, by Jens Vigen of CERN.

Some 81 people signed up to attend the meeting, drawn from a widely drawn section of the community concerned with e-Science digital repositories, and representing a wide spectrum of interests: different disciplines, varied organisational contexts (academic, commercial, research institutions, libraries and archives, technologists), a geographical spread across Europe, and a range of user types (users of information, data suppliers and repository managers).

With such a wide variety of interests represented there were a wide range of views and priorities expressed. However, when analysed, it was possible to pick out common themes and these were incorporated into the final recommendations which are presented earlier in this report.

#### Summary of the discussion

The issue of funding (for the long term) came out strongest in the poll of top 3 concerns, and this also took in funding of core services. These should be seen as linked. The question arose of whether funding should be focussed on the repository, the service or the digital assets themselves, possibly so

that the asset might move with its funding to different repositories over time. It was suggested that as particular datasets became less current they might move from local to European repositories. The question of funding is also related to evaluation of digital assets; how can funding agencies judge what to fund?

A related theme was strong support for the notion that publically funded outputs from activities/research should mandate submittal to an approved digital repository – and thereafter be available publically, due regard being given to periods of exclusive use by depositors.

Preservation was also a significant concern, but there was less discussion on the subject; the lack of contention suggests that the importance of preservation is universally recognised. It was noted that this too is a question of funding, since preservation implies long-term funding. It is necessary to preserve the means of interpretation alongside the data itself otherwise the risk is that the data becomes unusable and meaningless. This probably means storing source code for the associated software.

After funding, recommendations suggesting more emphasis on the discovery of information were strongly supported. This was often raised in the context of promoting the development of methods and technologies to enable the linking of information and navigation through the research cycle – from (or before) data creation to the final publication of results.

Another theme of the discussion was on the subject of whether repositories should be subject (discipline)-based or institutional. Institutional repositories have a clearer legal, funding and ownership basis to both set up and maintain, and may be better for cross-disciplinary searching, but discipline-based repositories are more intuitive to use and much more likely to be more popular and useful to users. The use of views to mimic subject repositories could resolve this dilemma. One commentator noted "Discipline-based repositories are better for the researcher, but how do we get to this point? We need to plan ahead otherwise there will be a plethora of institutional repositories" and related this to the superior branding of disciplinary repositories. Another commentator noted that there could be no single model for organising repositories.

There was also a plea for repositories to be more user-centric. A separate problem is how to conduct searches across disciplines; differences in approach e.g. between the hard sciences and humanities make this difficult, but it is important to address this for some types of research such as environmental studies. It was noted that one class of users will be machines.

There was much discussion of certification and the need for metrics related to repositories.

The final group of issues revolves around incentivising researchers to deposit their raw data in repositories. There was a lot of support for mandating data deposit as part of the funding agreement, though some discussion on whether both carrot and stick should be necessary to motivate producers to deposit. Including data citation in the measurement of academic achievement might be sufficient carrot. All this is contingent on resolution of legal issues including IPR and machine-understandable levels of access. These mechanisms need to be secure and reliable so that researchers can trust that their careers will not be adversely affected by depositing data. A suggestion was made to attach rights to data – generally this does not happen – and it might incentivise deposit. The question of deposit is related to the award structures for research workers, and breaking the "publish or perish" cycle.

A full report of the Study Workshop is provided in IR2.

## IV. Public consultation - summary

The public consultation was held over six weeks from mid July to the end of August 2007 using a questionnaire delivered through the European Commission's website. The questionnaire was advertised via a wide range of relevant e-mail lists and through communications with key individuals and organisations, harvesting 426 responses from all over the world. These came from contributors to repositories, repository users, repository managers, researchers, librarians, publishers, students, commercial companies and service providers. An appreciable number of contributions came from leading figures in data management, repositories, libraries.

A full report on the consultation is provided in IR2.

A combination of and free text and multiple choice questions were used to collect information the questionnaire explored the following areas:

- The respondents' uses of digital repositories, and the type of information they deposited in, or used from, repositories
- Perceived barriers to use
- Views on the adequacy of provision of digital repositories
- Views about the enablers of the use of digital repositories and policy directions which should be taken to encourage repository development
- Information to guide the future sustainability of repositories and a vision for the future of repositories in Europe.

In addition material was collected on the profile of the respondents. Respondents were primarily users of repositories (78% using repositories at least once a week, and nearly half on a daily basis). Most were from Europe (77%), but nearly a quarter were from outside Europe; a quarter of them described their primary role as researcher. A wide range of disciplines were represented, but the top three were information and library science (26%) physics and astronomy (19%) and computing and mathematics (14%). The organisational affiliation of most respondents was the academic sector (63%, see Figure 2). As might have been expected the most common forms of repository used were community and discipline-related repositories, digital libraries and institutional repositories.

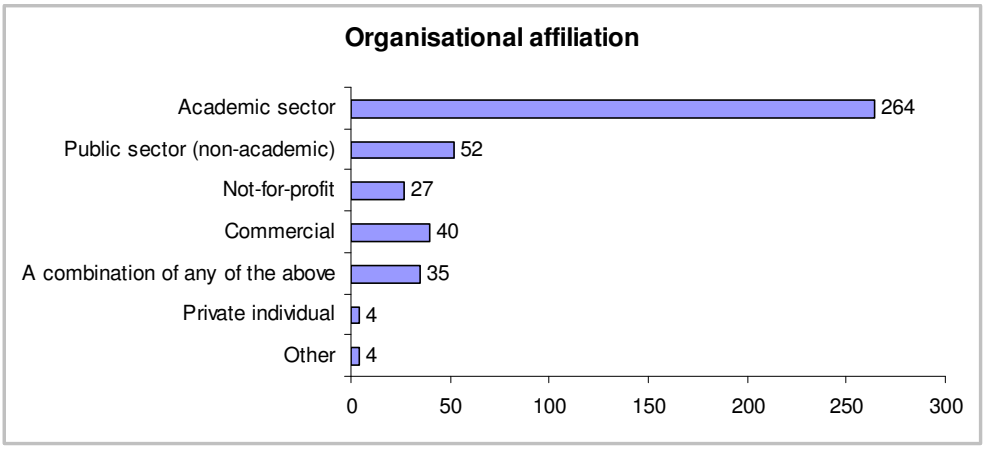



**Figure 2: Respondents' organisational settings**

When asked about difficulties encountered **using** repositories the chief obstacles specified were the difficulty and time taken to find the relevant repositories and to find information within them. A secondary concern was lack of training (about a quarter of respondents). Language barriers did not emerge as a major issue neither did cost of use (though as most people do not pay directly this is not surprising). See figure 3. All but 30 people answered this optional question

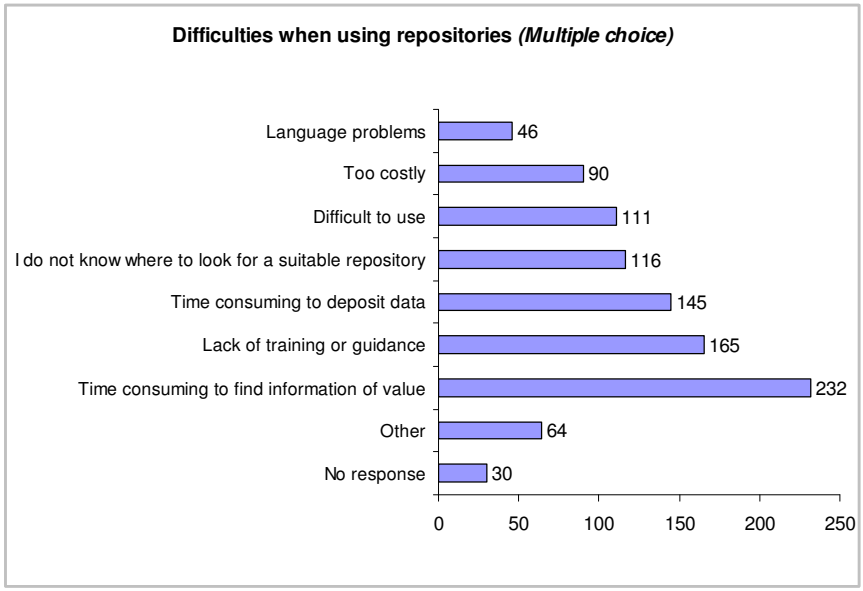



**Figure 3: Types of difficulty encountered using repositories**

Asked to identify inhibitors to **deposit** of materials into repositories as free text, 215 responses were made (50% of respondents). Summarising:

- A mix of intellectual property rights, copyright and contractual restrictions, and publishers embargoes were most frequently raised (some 25% of those answering). Respondents were concerned about lack of clarity about these as well as their loss of rights
- The time taken to deposit, and its complexity/difficulty (16%)
- Security concerns of various types were mentioned frequently, including loss of authenticity, data corruption, lack of control over use of information. This was raised by 12% of respondents
- Costs or payment policies were mentioned 12 times
- Absence of a suitable repository, or awareness of one, was also mentioned 12 times.

To further explore views on enablers for repository use questions asked where investments in repositories would best be made to enable science. These were free text questions, and 305 people responded, and the following summarizes the main points made.

The issue mentioned most frequently was that publicly funded research outputs should be placed in an open access repository, reflecting the view that outputs paid for from the public purse should in principle be available free without further charges. This theme recurred in answers to other questions. Going further than this, quite a few respondents were of the view that open access repositories should be positively promoted. Other suggestions were:

- Promotion of interoperability between repositories, federation and cross-repository searching
- Promotion of common standards for data (and metadata) formats and structures
- Better, more intuitive searching interfaces
- Establishment of registries of repositories (sometimes expressed as portals, directories or catalogues)
- Establishing adequate, stable, long-term funding for the curation of information; this was linked in some cases to promoting the establishment of infrastructures to support long-term preservation of information. The point was also made that making repositories compete for research funding was inappropriate.
- Establishing peer review mechanisms for data and e-publications
- (Further) investments in network infrastructures
- Establishing better tools for metadata generation and structuring.

Some of the other headline statistics from the survey were:

- 39% never paid (directly) for repository use
- 63% had no training in repository use
- 76% selected more accurate searching mechanisms as a way of making use easier, followed by tools to automatically generate metadata (70%) and provision of registries of repositories (58%)
- 62% of respondents said they need access to materials which were more than 20 years old – well beyond the boundary where preservation of digital resources becomes problematical.

The most frequently used repository types from which to get information were digital libraries, discipline-related repositories, institutional repositories and community repositories; combining community repositories with discipline-related repositories however gives these a combined lead. Deposit favours institutional repositories. Examining the ratio of use to deposit activity gives the pattern shown in the following table.

| | Repository type | | | | | | |
|---|---|---|---|---|---|---|---|
| | Digital library | Community repository | Discipline-related repository | Institutional repository | e-Learning repository | Commercial repository | Other |
| Ratio use:deposit: | 3.6 | 2.2 | 1.9 | 1.6 | 2.1 | 4.1 | 6.0 |

**Ratio: use to deposit by repository type**

It is not clear how to interpret this, but there is an indication that while deposit into institutional repositories may be relatively high, use of materials from them is relatively less frequent. The high ratio for commercial repositories is probably due their being mainly publication repositories, as are digital libraries.

Lastly, a striking finding was strong agreement with the notion of establishing international (EU-level) repositories (78%); there was also fair support for national repositories (56%).

> *"Digital repositories should be as important as libraries are. A challenge is to store huge amounts of information and---this is very important---have the tools to "play" with it. So repositories are not only about information, but also about tools."*
> *(Response, e-SciDR public consultation, summer 2007)*

# V. Legal issues and open access to e-Science repositories

Part of the study's remit was to look at legal issues relating to open access in the context of e – Science digital repositories.

Our review worked primarily outwards from the perspective of digital repositories and focused on non-text data. Its has not focused on the specific issues of open access, copyright, e-prints and publications extensively discussed in many other reports and articles[38].

The very definition of the term e-Science, coupled with the diversity of materials held in digital repositories, suggests vast and potentially highly complex legal territory: e-Science is collaborative; it works with many digital objects (themselves often compound, or part of larger objects), in different formats, and works across different jurisdictions, across different sectors, in a variety of different locations, contexts and types of interaction.

### Open access and use

"Open access" can mean different things with regard to use, but does not imply comprehensive rights per se relating to re-use. The level of right can vary: You can access a resource, but without any right to transform it, or conduct further research on it (more common in educational contexts). You can access the resource or item, and have the right to transform it and develop new work using it, but this does not necessarily automatically entitle re-use of an item or resource for commercial purposes.

A further point is that with several umbrella portals, such as GBIF[39], which provide a portal with access to multiple resources, the terms of use can vary from one resource to another, or participants can agree to apply the same terms of access. Either way, participation needs to be negotiated and agreements signed between participant and umbrella provider, and possibly further parties as well.

It is therefore extremely important that the user is aware of, understands and respects the rights and restrictions which apply to data she accesses.

A digital repository will also need the right to manage a resource or item, either as directed by the resource owner, or to decide itself on management.

### Diversity of types of scientific data, rights applicable to scientific data

Several rights are applicable to scientific digital content or digital objects. These rights evolve along the object's life-cycle, and ownership is defined according to the position in this life-cycle and the kind of interaction of each actor.

Any one repository could hold or enable use of digital objects from multiple points along the life cycle: objects created from scratch (for example, where a repository is also a facility generating data), objects created from pre-existing data (for example, normalized brain scans), interpretations of existing data in textual analysis; annotations of data (tags, keywords, unique identifiers, comments, or

[38] Links and bibliography are set out on the www.e-scidr.eu web site.

[39] Global Biodiversity Information Facility: www.gbif.org: "a coordinated international scientific effort to enable users throughout the world to discover and put to use vast quantities of global biodiversity data, thereby advancing scientific research in many disciplines, promoting technological and sustainable development, facilitating the equitable sharing of the benefits of biodiversity, and enhancing the quality of life of members of society"

indeed corrections of annotations); compilations of existing data (for example, mash-ups, simulations), re-purposing of data in *in silico* experiments …

It would be beneficial to take a systemic approach, looking at data as part of a system which is subject to many factors and requirements - legal, administrative, technical, economic, preservation, evaluation. This would also be valuable for the design of automated systems to support rights expression and management in the context of e-Science digital repositories, and e-Science more generally, as it would help limit proliferation of systems which address only parts of the cycle or process, which then need to be interoperable, ideally seamlessly.

#### Open access, digital repositories and controlled access

The digital repository is one of the main points at which third parties access materials. The premise and success of e-Science (and indeed science more generally) are underpinned by access to as full a breadth and depth of materials as needed – if a researcher can only check a small proportion of relevant materials (possibly of different types and formats), it will be much more difficult for her to assert validity of analysis and findings.

For the digital repository to function as such, it must know the legal status (one might say, the conditions of openness) of each item it holds and makes available to others, and it must be in control of that access in accordance with the conditions of openness. This may sound a contradiction in terms, but its success in continuing to function as a trusted repository (and thus its ability to attract materials or retain custody thereof) is predicated on its ability to manage access to comply with the applicable terms of access. Thus in the first place the digital repository must understand the legal aspects relating to its activity and the items it holds and act in accordance with agreed policies.

Difficulties arise where items are subject to conflicting or ambiguous legal and regulatory requirements.

Difficulties also arise when items come with no or restrictive usage terms.

Here, traditional archiving practice is pertinent: at ingest (when the item is ingested into the archive or repository), the archivist agrees with the depositor the terms on which the item is deposited and the terms on which the item may be used. This agreement is recorded, and the archive applies the terms of the agreement. For e-Science repositories, these transactions (agreements relating to rights), the process and the mechanism supporting the recording and transmission of rights information, must be as simple, clear, automated and generic as possible.

#### Diversity of types of interaction using repository data

Parties to agreements relating to use of pre-existing content need to be aware of all types of interaction which might be targeted for that use (these are listed in IR2). Note that uses also include management of data within the repository, for example migration of format of the object, for example for access efficiency or preservation.

#### Examples of awkward areas – fair use and warranties

Fair use, fair dealing, or exceptions and limitations to exclusive rights in civil law statutes are often not familiar or unclear to researchers, teachers, librarians, and particularly so when having to deal with different rules from several jurisdictions. These prerogatives allow them to use or re-use material without prior authorization or payment. Guidance on these issues should be provided by

independent third parties rather than rights owners, who may not be in the best position to give neutral advice.

The prerogatives under these headings include citation, exceptions for teaching, research, libraries and archiving. There is a lack of harmonization in these prerogatives within EC member states. The lack of harmonization and their frequent narrow scope are obstacles to easy access to and sharing of data and works.

Compulsory and voluntary licences make it possible to use data without authorization, after a fee (typically annual). However, these licences carry consequences of which non-lawyers are likely to be unaware. Researchers are unlikely to be able to distinguish between fair use covered by a statutory licence paid annually by their institution and paid-per-fee commercial databases. These differences also mean additional arrangements and work to enable automated or seamless access across these different systems.

**Warranties** on data accuracy and quality can be negotiated by the transferring or acquiring party when negotiating a transfer or access contract. Warranting that data which is to be re-used, modified, re-distributed are not constitutive of a prior rights infringement is useful, as a secondary distributor might be held liable for re-distributing data which had not been cleared of such a warranty, even if done in good faith. Again, there is a lack of harmonization in this regard; this creates uncertainty for technical intermediaries – digital repositories, but also providers of services used in the digital repository/transfer process; it also has implications for publishers and editorial responsibility, also when providing links to data. Another area where liability might be invoked relates to search engines.

### Cross-border issues

Science and e-Science work across borders, ideally at speed. It is commonplace to talk of obstacles to seamless working (and indeed basic deposit of materials in repositories) arising from lack of harmonization, but this is one of the major areas affecting open access and e-Science activities generally.

To mention just a few examples: there is lack of harmonization within the EU among limitations, lack of transparency relating to royalties, collective management for compulsory/statutory licences (where the research institution pays a collective society in relation to compensation for fair use); lack of harmonization regarding public-order provisions statute and contractual overridability (can exceptions and limitations to copyright and database *sui generis* right be cancelled by a contract or database access licence?) There is lack of harmonization on technical measures relating to anti-circumvention legislation: factors, infringement, intention, commercial purpose, indirect circumvention (which can arise in bug-fixing in software programs).

Regional and local administrative regulations can vary, imposing local requirements on the release of data across borders, in addition to some national restrictions.

### Wish list

Key informants and respondents highlighted the need for awareness-raising, education and guidance for all actors working with e-Science repositories.

Guides should be published for scientists, researchers, teachers, and unaffiliated individuals, institutions, including digital repositories and related providers (eg of tools) on the legal framework for creation, deposit, access and re-use of digital materials, on fair use, the public domain, liability,

privacy and confidentiality, and so on. These guides should be available in the home language of the reader and the presentation should take into account specific legal perspectives and features relating to the range of EC member state jurisdictions and others likely to be important. There are several examples of excellent practice in this regard, such as the work by The Netherlands' Surf Foundation and DARE, the Dutch Network of Digital Academic Repositories.

The legal status of digital repositories should be clarified, and clearly set out for stakeholders and users. It would also be very helpful to have mandatory disclosure of rights policy by publishers and institutions, for transparency and efficiency; it is important that the information about rights policy is kept available and up to date.

Rights management automation: there should be research into the development of automated rights expression and rights management tools which work along the whole life cycle of an object. This should also take into account metadata format tagging, platforms and tools. Standardized rights expression languages and rights data dictionaries which work with scientific digital objects, processes and practices.

Science Commons[40] provides licences which can be adapted to a range of scientific needs: biological Material Transfer Agreements, licences for open data, databases, author's addenda standard side contracts to publishing agreements.

There is a need for citation systems which embed, forward and possibly also track attribution and other relevant information for links between primary research data, publications and other communications.

Scientific communities are effective arenas for working on licences, national jurisdictions, thanks to the strong communication achieved within disciplines. Again, co-ordination between disciplines will be of critical importance, to ensure that inter-disciplinary research is not impaired by over-specific, discipline-based approaches. Semantic interoperability of metadata is also important.

[40] www.sciencecommons.org

# Glossary of technical terms and acronyms

| Term | Definition |
|---|---|
| ARK | Archival Resource Key. An identifier scheme created to allow persistent references to digital objects |
| Bioinformatics | Science concerned with the use of techniques from applied mathematics, statistics and computer science to solve biological problems. |
| CDISC | Clinical Data Interchange Standards Consortium. A non-profit organization that has established standards to support the acquisition, exchange, submission and archive of clinical research data and metadata. |
| CERIF2000 | Common European Research Information Format. CERIF 2000 is a set of guidelines for research information systems (CRIS, qv). |
| ChemML | Chemical Mark-up Language. An XML standard for representing chemical structures. |
| CIDOC | International Committee for Museum Documentation. Also refers the CIDOC-CRM, an international standard for museum documentation, ISO 21127-2006. |
| Collection | Information items brought together for some specific purpose or with at least one feature in common |
| Content | The thing lodged in the repository and of primary interest for deposit and use—say a data file or a digital form of a publication. Distinguished here from Metadata (qv). |
| CRIS | Current Research Information System |
| DAML | DARPA Agent Markup Language (DAML). The goal of the DAML effort is to develop a language and tools to facilitate the concept of the Semantic Web |
| DAML+OIL | DAML and OIL that combines features of both. Superseded by Web Ontology Language (OWL). |

| Term | Definition |
|---|---|
| Data | (Digital) content, such as databases, images, video, and simulation results, and so on. Distinguished here from Publication content. |
| Data Grid | A grid computing system that deals with data — the controlled sharing and management of large amounts of distributed data. |
| DCMI | The Dublin Core Metadata Initiative. An open organization engaged in the development of interoperable online metadata standards for documentation. |
| DG INFSO | Information Society and Media Directorate of the European Commission |
| DIDL | Digital Item Declaration Language |
| Digital Repositories | The constructs that hold digital collections or items and facilitate their use. |
| Distributed computing | Computing coordinated and spread over a number of different systems, which may be geographically separted. |
| DOI | Digital Object Identifier. A system is for identifying content objects in a digital environment. The system is managed by the International DOI Foundation (See: http://www.doi.org/) |
| Dublin Core | A simple and standardised "core" set of metadata for describing objects in ways that make them easier to find. Dublin Core is defined by NISO Standard Z39.85-2007. See also DCMI. |
| DyVOSE | A project funded by the Joint Information Systems Committee (JISC) in the UK. Its intention is to explore the establishment of scalable Virtual Organisations. |
| EAD | Encoded Archival Description. A standard for computer-based description of archival collections. |
| EC | European Commission |

| Term | Definition |
|---|---|
| e-Infrastructure | In the context of digital repositories this is taken to be the technical and administrative framework and facilities underlying e-Science digital repositories. |
| e-IRG | e-Infrastructure Reflection Group |
| ERA | European Research Area |
| e-SciDR | e-Science Digital Repositories, the acronym for this study |
| e-Science | Science supported to a significant degree by digital information-processing and/or computational technologies, or wholly based on these. |
| ESFRI | European Strategy Forum on Research Infrastructures |
| euroCRIS | A not-for-profit association which acts as an internationally recognised point of reference for all matters relating to Current Research Information Systems. |
| Federated | In the context of data processing, computing regimes which involves a number of cooperating computing resources. |
| FLOPS | Floating Point Operations per Second. A measure of the speed of processing. |
| FRBR | Functional Requirements for Bibliographic Records. From the International Federation of Library Associations (IFLA) |
| Gbps. | Gigabits per second |
| GÉANT2 | GÉANT2 is the pan-European research and education network. GÉANT2 is co-funded by the European Commission and Europe's national research and education networks (NRENs), and is managed by DANTE. |

| Term | Definition |
|---|---|
| GEML | Gene Expression Markup Language, used for storing DNA and microarray data |
| Globus | The Globus Alliance is an international collaboration that conducts research and development to create basic Grid technologies. |
| Globus GSI | Globus Grid Security Infrastructure |
| GML | Geography Mark-up Language An XML grammar defined by the Open Geospatial Consortium (OGC) to express geographical features. |
| Granularity (of access) | The fineness of differentiation of access permission to data or to computing resources to perform given tasks. |
| Grid Computing | Can be defined as "the technology that enables computer resource virtualization, on-demand provisioning, and service (resource) sharing between organizations" (Plaszczak, Pawel; Rich Wellner, Jr. *Grid Computing "The Savvy Manager's Guide"*. Morgan Kaufmann Publishers. ISBN 0-12-742503-9. ). More informally the use of many separate computers to solve computing problems – usually of a large scale. |
| HPC | High Performance Computing. Computing demanding a capacity processing power, usually in the teraFLOP range and above. |
| iDGL Grid Operations Center (GOC) | International Virtual Data Grid Laboratory Grid Operations Center. See http://igoc.ivdgl.indiana.edu/ |
| Information | Used in this report to cover all different information types. |
| IPR | Intellectual Property Rights |
| ISAD(G) | International Standard Archival Description (General). A standard from the International Council on Archives for describing archival collections and objects. |

| Term | Definition |
|---|---|
| ISBN | International Standard Book Number. There are now two forms of ISBN: the original 10-digit ISBN-10, and the ISBN-13 compatible with the EAN-13's bar coding standard. |
| ISO | International Standards Organisation. See http://www.iso.org/iso/en/ISOOnline.frontpage |
| ITU-T | International Telecommunications Union (Standards). See: http://www.itu.int/net/home/index.aspx |
| LDAP | Lightweight Directory Access Protocol. An application protocol for querying and modifying directory services running over TCP/IP. |
| LHC | Large Hadron Collider. A new, powerful particle accelerator located at CERN in Geneva, due to become operational in 2009. |
| MADS | Metadata Authority Description Schema |
| MARC | Machine-Readable Cataloging. Standard formats for the representation and communication of bibliographic and related information in machine-readable form. They including MARC21, MARCXML, an XML based version of MARC21 |
| MathML | An XML-based coding standard for mathematical text |
| Metadata | Descriptive and other information pertaining to content (qv), from which it is distinguished here. See the discussion in Part 1, Information Types |
| METS | Metadata Encoding and Transmission Standard |
| MGED | Microarray and Gene Expression Data Society. An international organization of biologists, computer scientists. Maintains the MIAME standard. See: http://www.mged.org/ |
| MIAME | Minimum Information About a Microarray Experiment. A standard for reporting microarray experiments |

| Term | Definition |
|---|---|
| MIME types | Multipurpose Internet Mail Extensions. A set of file standards supported by e-mail. |
| MODS | Metadata Object Description Schema |
| Moore's Law | A rule of thumb which state that the power of computers increases exponentially, doubling every 12 to 18 months. |
| MPEG-21 | A standard, from the Moving Picture Experts Group aims at defining an open framework for multimedia applications. International standard ISO 21000. |
| NISO | The USA's National Information Standards Organisation. See: http://www.niso.org/ |
| NORDUnet | The Nordic Internet highway to research and education networks (NREN) in Denmark, Finland, Iceland, Norway and Sweden, |
| NRENs | National Research and Education Networks |
| OAI-ORE | Open Archives Initiative - Object Reuse and Exchange. An extension of the OAI-PMH initiative to specify how distributed repositories may exchange information about their constituent digital objects. |
| OAI-PMH | Open Archives Initiative Protocol for Metadata Harvesting. |
| OAIS | Open Archival Information System. A standard reference model for archival systems, ISO 14721:2003 |
| OASIS | Organization for the Advancement of Structured Information Standards. A not-for-profit consortium for the development of open standards for the global information society. |
| ODRL | Open Digital Rights Language. A proposed language for expressing rights information over content. See: http://www.w3.org/TR/odrl/ |

| Term | Definition |
|---|---|
| OGF | Open Grid Forum. an open forum for grid innovation, developing open standards for grid software interoperability. See: http://www.ogf.org/ |
| OGSA | Open Grid Services Architecture. An architecture for a service-oriented grid computing environment for business and scientific use, developed within the Global Grid Forum (GGF). |
| OIL | Ontology Inference Layer or Ontology Interchange Language. An ontology infrastructure for the semantic web. |
| omics | -omics is a suffix attached to the names of biological subfields where computing and mathematical techniques are applied. |
| OMII Europe | Open Middleware Infrastructure Institute. An EU-funded project which has been established to source key software components that can interoperate across several heterogeneous Grid middleware platforms. See: http://omii-europe.org/OMII-Europe/ |
| Open access | Access to information free online access which is free at the point of delivery |
| OpenSSL | Open Secure Sockets Layer. The OpenSSL project an effort to develop a robust, commercial-grade, full-featured, and open source toolkit implementing the SSL. |
| Orphan data | Data for which there is no repository system available for its longer term storage and management. |
| OWL | Web Ontology Language. See: http://www.w3.org/TR/owl-features/ |
| PDF | Portable Document Format |
| PDF/A | Portable Document Format / Archival. ISO 19005-1:2005 A standard which defines a format for the long-term archiving of electronic documents based on the PDF Reference Version 1.4 from Adobe Systems Inc. |

| Term | Definition |
|---|---|
| PDOI | Permanent Digital Object Identifier. A class of technologies to identify uniquely, and permanently digital objects. |
| PERMIS | PrivilEge and Role Management Infrastructure Standards Validation. A project to authorisation and authentication systems. See: http://www.permis.org/en/index.html |
| PREMIS | PREservation Metadata: Implementation Strategies. A standard for specifying digital preservation metadata. See: http://www.oclc.org/research/projects/pmwg/ |
| Publications | A form of Content (qv) in published form. Thus, for example, published article and reports, pre-prints and post-prints, theses, patent documents and similar. Distinguished here from Data content. |
| PURL | Persistent Uniform Resource Locator. A form of PDOI (qv). See: http://www.purl.org/ |
| RDF | Resource Description Framework. A general-purpose language for representing information in the Web. See: http://www.w3.org/RDF/ |
| Repository | See: Digital Repository |
| SAML | Security Assertion Markup Language. See: http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=security |
| SBML | Systems Biology Mark-up Language. A software-independent language for describing and exchanging models among different tools. |
| Semantic web | An evolving extension of the World Wide Web in which the semantics of information and services on the web is defined, making it possible for the web to understand and satisfy the requests of people and machines to use the web content. |

| Term | Definition |
|---|---|
| Shibboleth | A standards-based, open source middleware software which provides Web Single Sign On (SSO) across or within organizational boundaries. See: http://shibboleth.internet2.edu / |
| SOA | Service Oriented Architecture. A computer systems architecture for creating and using business processes, packaged as services. |
| SOA | Service Oriented Architecture. Computing architectures which provide methods for integrating computer systems which group functionality around business processes and package these as interoperable services. |
| SOAP | Simple Object Access Protocol, and lately also Service Oriented Architecture Protocol. See http://www.w3.org/TR/soap/ |
| Systems biology | Integration of different levels of information in order to understand how biological systems functions. |
| teraFLOP | A trillion floating point operations per second ($10^{12}$ FLOPS). |
| URI | Uniform Resource Identifier. A string of characters used to identify or name a resource on the internet. |
| URL | Uniform Resource Locator. A type of URI that specifies where an identified resource is available and the mechanism for retrieving it. |
| URN | Uniform Resource Name. A URI (qv) that uses the URN scheme (see: http://en.wikipedia.org/wiki/URI_scheme ). Both URNs (names) and URLs (locators) are URIs, and a particular URI may at the same time be a name and a locator. |
| VO | Virtual Organisation. An organizational entity that uses telecommunication tools to enable, maintain and sustain member relationships in distributed work environments. |

| Term | Definition |
|---|---|
| VOMS | Virtual Organization Membership Service. A system for managing authorization data within multi-institutional collaborations. See: http://www.globus.org/grid_software/security/voms.php |
| WDC | World Data Center. See http://www.ngdc.noaa.gov/wdc/wdcmain.html |
| Web 2.0 | World Wide Web technologies supporting web-based communities and hosted services such as social-networking sites, wikis, blogs, and folksonomies, which aim to facilitate creativity, collaboration, and sharing among users. |
| Web Services | A series of web-based standards, including: WS - Includes: WS-Policy, WS-Trust, WS-Privacy, WS-SecureConversation, WS-Federation, WS-Authorisation, WS-Agreement. |
| Workflow | A sequence of processes to accomplish some task. |
| X509 | An ITU-T standard for public key infrastructure (PKI) |
| XACML | eXtensible Access Control Mark-up Language. From OASIS (qv) See: http://www.oasis-open.org/committees/download.php/2406/oasis-xacml-1.0.pdf |
| XML | eXtensible Mark-up Language |
| XrML | eXtensible rights Mark-up Language. An XML-based standard for securely specifying and managing rights and conditions associated with resources, including digital content and services. See: http://www.xrml.org/ |

# References

Allen, R.E. (ed.), The Concise Oxford Dictionary of Current English. 8th Edition., 1990

ALT-SURF Seminar, ePortfolios and Digital Repositories. ALT-SURF seminar, 22 and 23 April 2004, Edinburgh UK, 2004

Arms, W Y., Repositories for large-scale digital libraries. Cornell University, NSF/JISC Repositories Workshop, March 27, 2007

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, p., Wouters, P. An international framework to promote access to data. Science Vol 303, Issue 5665, 1777-1778, 19 March, 2004

Ashley, K., et al. National Council on Archives (NCA): Interoperability Protocol. NCA, 2003

Association of Research Libraries. To Stand the Test of Time. Long-term Stewardship of Digital Data Sets in Science and Engineering. A Report to the National Science Foundation. From the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe, 2006

Ball, C. A., Sherlock, G., Brazma, A. Funding high-throughput data sharing. Nature Biotechnology 2004;22(9):1179-83. 2004

Beagrie, N. e-Infrastructure strategy for research: Final report from the OSI preservation and curation working group. National *e*-Science Centre, 2007

BELIEF: Bringing Europe's eLectronic Infrastructures to Expanding Frontiers, Research and Industry Handbook. First edition, 2006.

Berman, H.M., et al., The Protein Data Bank, 2003

Berners-Lee, T., Hendler, J., Lassila, O. The semantic web: a new form of web content that is meaningful to computer will unleash a revolution of new possibilities. Scientific American, May 2001.

Bradley, K. APSR Sustainability Issues Discussion Paper. Australian Partnership for Sustainable Repositories, National Library of Australia, 2005.

Brain Information Group White Paper, The Information Infrastructure Needs of Neuroscience Research: Opportunities and Issues of Implementation. Society for Neuroscience, 2007.

Brazma, A. Editorial: On the importance of standardisation in the life sciences. Bioinformatics. 17(2):113-114, February 2001.

Brooksbank, C. and Quackenbush, J. Data Standards: A Call to Action. OMICS. 2006 Summer;10(2):94-9

Brooksbank, C., Cameron, G., Thornton, J. The European Bioinformatics Institute's data resources: towards systems biology. Nucleic Acids Research, Vol. 33, Database issue D46-D53, 2005.

Buchhorn, M., McNamara, P. Sustainability Issues for Australian Research Data - The Report of the Australian e-Research Sustainability Survey Project, Document Version 1.1. October 2006

Buckingham, S. Data's future shock. . Nature 428, 774-777, 15 April 2004

Burton, A. Functional Checklist for Digital Repositories in the Research Quality Framework (RQF). Australian Partnership for Sustainable Repositories, 2006.

Campbell, L., and Blinco, K. Repository management and implementation, 2004. White Paper for alt-i-lab 2004 on behalf of DEST (Aus) and JISC-CETIS(UK)

Campbell, L.M. Repository issues from a teaching and learning perspective. Annex 3 to the JISC Digital Repositories Review, February 2005.

Catalyurek, U., Hastings, S., Huang, K., Kumar, V.S., Kurc, T., Langella, S., Narayanan, S., Oster, S., Pan, T., Rutt, B.. Zhang, X., Saltz, J. Supporting Large Scale Medical and Scientific Datasets, in *Parallel Computing:Current & Future Issues of High-End Computing*, Proceedings of the International Conference ParCo, John von Neumann Institute for Computing, Julich, NIC Series, Vol. 33, ISBN 3-00-017352-8, pp. 3-14, 2006. 2005

Casey, J., Proven, J., and Dripps, D. Managing IPR in Digital Learning Materials: A Development Pack for Institutional Repositories. 2006. Available at: www.trustdr.ulster.ac.uk

Choudhury, S. The Relationship between Data and Scholarly Communication. NSF/JISC Repositories Workshop, 2007

Coleman, A.S. Commons-based digital libraries, Interrogating the social realities of information and communications systems. Pre-conference workshop, ASIST AM 2006

Commission of the European Communities. i2010 eGovernment Action Plan: Accelerating eGovernment in Europe for the Benefit of All, 2006

Commission of the European Communities. Communication from the Commission to the European Parliament, The Council and The European Economic and Social Committee on scientific information in the digital age: access, dissemination and preservation, 2007.

Committee on Responsibilities of Authorship in the Life Sciences. Sharing publication-related data and materials - responsibilities of authorship in the life sciences. The National Academies Press, 2003

Conway, P. Institutional repositories: is there anything left to say? Presentation within OCLC Distinguished Seminar Series, October 7 2004.

Cordewener, B. Institutional Repositories in the Netherlands, a national and international perspective. NSF/JISC Repositories Workshop, 2007

Crane, G. Repositories, Cyberinfrastructure and the Humanities. NSF/JISC Repositories Workshop April 16, 2007

Crow, R. The Case for Institutional Repositories: A SPARC Position Paper. The Scholarly Publishing & Academic Resources Coalition, 2002

Crow, R. A Guide to Institutional Repository Software. Open Society Institute, 2004

Crow, R. SPARC Institutional Repository Checklist & Resource Guide. The Scholarly Publishing & Academic Resources Coalition. 2002

David, P.A. Towards a cyberinfrastructure for enhanced scientific collaboration: Providing its 'soft' foundations may be the hardest part. In *Advancing Knowledge and the Knowledge Economy*, eds. D. Foray and B. Kahin, Cambridge: MIT Press. 2005 [Available as Oxford Internet Institute Research Report No 4]

Dawyndt, P., Dedeurwaerdere, T., and Swings, J. Contributions of bioinformatics and intellectual property rights in sharing biological information. International Social Science Journal, Volume 58 Issue 188, Pages 249 – 258, 2006

Day, M. Prospects for institutional e-print repositories in the United Kingdom. ePrints UK supporting study, no. 1, University of Bath, 2003

de Vries, H., Verheul, H., Willemse, H. Stakeholder Identification in IT Standardization Processes. MIS Quarterly, Special Issue: *Workshop on Standard Making: A Critical Research Frontier for Information Systems*, ICIS workshop, 2003.

Dedeurwaerdere, T. The institutional economics of sharing biological information. In: International Social Science Journal 188. 2006

Dewatripont, M., Ginsburgh, V., Legros, P. and Walckiers, A. [for other authors, see notes]. Study on the economic and technical evolution of the scientific publication markets in Europe. Final Report ECARES, Université libre de Bruxelles; January 2006

Digital Curation Center & DigitalPreservationEurope. Digital Repository Audit Method based no risk assessment. DRAMBORA, Version 1.0. Release for public testing and comment., 2007

DSpace architecture review group. Toward the next generation: Recommendations for the next DSpace Architecture. DSpace, MIT, 2007

Eckersley, P., Egan, G.F., Amari, S., et al. Neuroscience data and tool sharing: a legal and policy framework for neuroinformatics. Neuroinformatics Journal, 1, 149—165, 2003

EDUCAUSE Evolving Technologies Committee, Institutional Repositories: Enhancing Teaching, Learning, and Research. Lawrence Technological University / University of Michigan, , 2003

Esanu & Uhlir (editors). The role of scientific and technical data and information in the public domain: proceedings of a symposium, National Academies Press, 2003

ESFRI, ESFRI. European Roadmap for Research Infrastructures - Report 2006

European Commission. Staff Working Document, Proposal for a Directive of the European Parliament and of the Council establishing an infrastructure for spatial information in the Community (INSPIRE), 2004

European Task Force Permanent Access. Permanent access to the records of science, Strategic Action Programme 2006-2010, 2005

Fitzgerald, A., and Pappalardo, K. Building the infrastructure for data access and reuse in collaborative research. An analysis of the legal context. 2007

Frueh, L., Internet Archive, Access Tools: Bridging Individuals to Information. 2007

Fusco, L. and van Bemmelen, J. Earth observation archives in digital library and grid infrastructures. Data Science Journal, 3, 222-226, 2004

Gardner, D., and Shepherd, G.M. A gateway to the future of neuroinformatics. Neuroinformatics. 2004;2(3):271-4, 2004

Gray, J., Liu, D.T., Nieto-Santisteban, M. Szalay, A.S., DeWitt, D. Scientific data management in the coming decade. *MSR-TR-2005-10*. Microsoft Technical Report, Redmond, WA, 2005

Greening, O. Information Environment Metadata Schema Registry: Market Proposition. UKOLN, 2006

Hamidzadeh, B. Scale: A repository challenge. Library of Congress, 2007

Heery, R., Anderson, S., Digital Repositories Review, UKOLN, 2005

Heery, R. Duke, M., Day M., Lyon, L., S Coles, Frey,J., Hursthouse, M., L Carr, L. and Gutteridge, C. Integrating research data into the publication workflow: eBank experience. In: Proceedings PV-2004: *Ensuring the Long-Term Preservation and Adding Value to the Scientific and Technical Data,* 5-7 October 2004, ESA/ESRIN, Frascati, Italy, (ESA WPP 232), 2004

Henty, M.. Ten Major Issues in Providing a Repository Service in Australian Universities. D-Lib Magazine, May/June 2007, Volume 13 Number 5/6, 2007

Hey, A., and Trefethen, A. The UK e-Science Core Programme and the Grid. Future Generation Computer Systems, 18 (8). pp. 1017-1031, 2002

Hey, A. Towards an e-Infrastructure for Research and Innovation: A Progress Report on e-Science. UK e-Science Core Programme, 2004

Hey, A., and Trefethen, A. The data deluge: an e-Science perspective., In: *Grid Computing - Making the Global Infrastructure a Reality*, pp. 809-824, Wiley, 2003

HM Treasury, Department of Trade and Industry, Department for Education and Skills. Science and innovation: working towards a ten-year investment framework, 2004

Hoorn, E. and van der Graaf, M. Towards good practices of copyright in Open Access Journals. A study among authors of articles in Open Access journals. Pleiade Management & Consultancy, 2005

Hunter, J. Scientific publication packages – A selective approach to the communication and archival of scientific output. In The International Journal of Digital Curation, Issue 1, Volume 1, 2006.

ICSU. Report of the CSPR Assessment Panel on Scientific Data and Information, International Council for Science, ISBN 0-930357-60-4, 2004

ICSU World Data Centers Panel. Meeting Summary, Bremen, Germany, February 1-2, 2005,

IITA Digital Libraries Workshop. Interoperability, Scaling, and the Digital Libraries Research Agenda: A Report on the May 18-19, 1995 - IITA Digital Libraries Workshop, August 22, 1995, 1995

International Panel on Climate Change, IPCC Fourth Assessment Report: Climate Change 2007. Available from: www.ipcc.ch

Jagadish, H.V., Oklen, F. Data management for the biosciences: Report of the NSF/NLM workshop on data management for molecular and cell biology, Berkeley National Laboratory, Feb 2-3, 2003, 2003

JCSR VRE Working Group, Roadmap for a UK Virtual Research Environment. Report of the JCSR VRE Working Group, 2004

Jerez, H. N., Ziamoming, L., Hochstenbach, P., Van de Sompel, H. Repository architectures: the multi-faceted use of the OAI-PMH in the LANL repository. LANL 2004

Koslow, S H., Hirsch, M D. Celebrating a decade of neuroscience databases: looking to the future of high-throughput data analysis, data integration, and discovery neuroscience. Neuroinformatics. 2(3):267-70, 2004

Kush, R D., Hardison, C. How necessary are data standards? Scrip Magazine, May 2004

Law, M. Reduce, Reuse, Recycle: Issues in the Secondary Use of Research Data. IASSIST Quarterly, Spring 2005

Loeffler, T., Race, P. Great Project? Perceptions of Project Success by Different Stakeholder Groups, 2007

Lord, P., Macdonald, A. e-Science Curation Report: Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision. JISC, 2003

Lord, P., Macdonald, A., Sinnott, R. et al. Large-Scale Data Sharing in the Life Sciences : Data Standards, Incentives, Barriers and Funding Models: the Joint Data Standards Study. Medical Research Council, 2005

Lowrance, W.W. Learning from experience: privacy and the secondary use of data in health research. The Nuffield Trust, November 2002

Lynch, C. A. Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. portal: Libraries and the Academy, Volume 3, Number 2, April 2003

Lynch, C., and Lippincott, J. Institutional Repository Deployment in the United States as of Early 2005. D-Lib Magazine, 11, (9) 2005

Lyon, E. Dealing with data: Roles, rights, responsibilities and relationships. UKOLN, 2007

Markey, K., Rieh, S., Young, St. Jean, B., Kim, J. and Yakel, E. for the Council on Library and Information Resources. Census of Institutional Repositories in the United States: MIRACLE Project Research Findings, 2007

Mathieson, S.A. and Cross, M. Ordnance Survey challenged to open up: The inventor of the world wide web wants access to Ordnance Survey data - and the freedom to manipulate it as he sees fit. The Guardian, Thursday 23 March, 2006

McDonald, J. and Shearer, K. Toward a Canadian Digital Information Strategy: Mapping the Current Situation in Canada, Version 2.0. Library and Archives Canada, 2006

McLean, N. An ecology of Repository Services: A Cosmic View. Presentation given at ECDL, 2004

MS Global Learning Consortium, Inc. IMS Digital Repositories Interoperability - Core Functions Information Model. IMS, January 2003

Murray-Rust, P. Data-driven science - a scientist's view. NSF/JISC. Repositories Workshop, April 10, 2007

Nass, S.J., and Stillman, B.W. Large-scale biomedical science: exploring strategies for future research. National Academies Press, 2003

National Cancer Research Institute. Strategic Framework for the Development of Cancer Research in the UK. National Cancer Research Institute, 2003

National Research Council. Bits of Power: Issues in global access to scientific data, Committee on Issues in the Trans-border Flow of Scientific Data. National Academies Press 1997

National Science Board. Long-lived digital data collections: enabling research and education in the

21st century. National Science Foundation, NSB-05-40, 2005

National Science Board, Science and Engineering Infrastructure For the 21st Century: The Role of the National Science Foundation, 2002

NDAC Working Group to the Social Sciences and Humanities Research Council of Canada and the National Archivist of Canada. National Data Archive Consultation: Building Infrastructure for Access to and Preservation of Research Data, 2002

NESTOR–Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung, Kriterienkatalog vertrauenswürdige digitale LangzeitarchiveVersion 1 (Entwurf zur öffentlichen Kommentierung), 2006

Neumann, E.K., Miller, E., and Wilbanks, J. What the semantic web could do for the life sciences. Drug Discovery Today: BioSilico 2(6): 228-236 2004

Newhouse, S.,, Schopf, J.M., Richards, A., Atkinson, M. Study of User Priorities for e-Infrastructure for e-Research (SUPER). Tech Report UKeS-*2007*-01, April 2007

Norris, R. Report to the 24th CODATA General Assembly on data activities in the International Astronomical Union. CODATA, 2004

National Science Foundation Cyberinfrastructure Council, USA. Cyberinfrastructure Vision for 21st Century Discovery. National Science Foundation 2007

OECD. OECD follow-up group on issues of access to publicly funded data - Interim Report. OECD, 2002

OECD. Main points from OECD workshop on human genetic research databases - issues of privacy and security, 26-27 February 2004, 2004

OECD, OECD Principles and Guidelines for Access to Research Data from Public Funding. OECD, 2007

Payette, S. Interoperability for Digital Objects and Repositories: the Cornell/CNRI experiments. D-Lib Magazine, May 1999, Volume 5 Issue 5, 1999

Peters, T. A. Digital repositories: individual, discipline-based, institutional, consortial, or national? The Journal of Academic Librarianship, Volume 28, Issue 6, 414-417 November-December 2002

Pickton, M. and McKnight, C., Research students and the Loughborough institutional repository. Journal of Librarianship and Information Science, 38 (4), pp 203-219, 2006

PREMIS Working Group. A Working Group Jointly Sponsored By OCLC and RLG: Implementing Preservation Repositories For Digital Materials: Current Practice And Emerging Trends In The Cultural Heritage Community. A Report by the PREMIS Working Group, 2004

Prudlo, M. E-Archiving: An Overview of some Repository Management Software Tools. Ariadne, Issue 43, 2005

Quackenbush, J. Data standards for 'omic' science. Nature Biotechnology, Volume 22, Number 5, May 2004

Reichman, J.H., Uhlir, P.F., "A contractually reconstructed research commons for scientific data in a highly protectionist intellectual property environment", Duke Law & Contemporary Problems 66, Winter/Spring 2003, pp 315-462

Repositorios colectivos de e-información. Towards a New World Data Center System: Meeting Global Needs. Report of the WDC Modernization Task Team to the Panel on World Data Centers, 2003

Repository Interoperability Working Group. Digital library repositories and instructional support systems, Coalition for Networked Information. University of Virginia Library, 2004

Research Information Network. Stewardship of digital research data: a framework of principles and guidelines. Responsibilities of research institutions and funders, data managers, learned societies and publishers. Draft for consultation, 2007

Richardson, J. Survey: Integration of OARs with Research Management Systems, Griffith University, August 2006

Rightscom. Research Funders' Policies for the management of information outputs. Research Information Network, January 2007

Royal Statistical Society and UK Data Archive - Working Group on the Preservation and Sharing of statistical Material. Information for Data Producers, Preserving and sharing statistical material, Royal Statistical Society, 2002

Smith, A. Thoughts on Scale and Complexity. NSF/JISC Repositories Workshop, 2007

Stabenau, A., McVicker, G., Melsopp, C., Proctor, G., Clamp, M., Birney, E. The Ensembl Core Software Libraries. Genome Research,14(5):929-33, 2004

Stoeckert, C.J., Jr., Quackenbush, J., Brazma, A., Ball, C.A. Minimum information about a functional genomics experiment: the state of microarray standards and their extension to other technologies. Drug Discovery Today: Targets, 2004, Vol.3, No.4: 159-164, 2004

Strong, D.F., Leach, P.B. National Consultation on Access to Scientific Research Data. Final Report. January 31, 2005, 2005

Swan, A., Awre, C. Linking UK repositories. University of Southampton, 2006

Tansley, R. Building a Distributed, Standards-based Repository Federation. The China Digital Museum Project D-Lib Magazine, Volume 12 Number 7/8, July/August 2006

The Sheridan Libraries at John Hopkins University. A Technology Analysis of Repositories and Services, Final Report, John Hopkins University 2006

The Wellcome Trust. Report of a meeting organized by the Wellcome Trust, held 14-15 January 2003, Fort Lauderdale, USA, Sharing data from large-scale biological research projects: a system of tripartite responsibility, 2003

UNESCO. Symposium: Open access and the public domain in digital data and information for science: Proceedings of an international symposium [held UNESCO, Paris in May 2003]. UNESCO, 2004

van Eijndhoven, K., and van der Graaf, M. DRIVER - Inventory study into the present type and level of OAI compliant Digital Repository activities in the EU. OZ 07.1301, 2007

Van Horn, J.D., Gazzaniga, M.S. Databasing fMRI studies - towards a 'discovery science' of brain function. Nat Rev Neurosci.;3(4):314-8, April 2002

Van Westrienen, G. Making the strategic case for institutional repositories - CNI-JISC-SURF Conference; Amsterdam, 10-11 May 2005: Completed Country Questionnaires. See: D-Lib Magazine, Volume 11 Number 9, September 2005

Van Westrienen, G. and Lynch, C. Academic institutional repositories. Deployment status in 13 nations as of mid 2005. D-Lib Magazine, 11:9. 2005

Warner, S., Bekaert, J., Van de Sompel, H. et al. Pathways augmenting interoperability across scholarly repositories. International Journal on Digital Libraries, 2006

West, L.J., Stark, J . Killworth, P.D. et al. Remote Visualisation of Large Oceanic Datasets, 2003

West, M. Some industrial experiences in the development and use of ontologies. EKAW04 Workshop on Core Ontologies, 2004

Wheatley, P. Institutional repositories in the context of digital preservation. Microform and Imaging Review, Vol. 33 No.3, pp.135-46, 2004

Wilbanks, J., and Boyle, J. Introduction to Science Commons, 2006. Available at: www.sciencecommons.org

Wouters, P. Data Sharing Policies, Networked Research and Digital Information. NIWI-KNAW, 2002

Wouters, P., and Schröder, P. (ed.). Promise and practice in data sharing, The Public Domain of Digital Research Data. NIWI-KNAW, Amsterdam, ISBN 90 6472, 2003

Yuille, M., Korn, B., Moore, T., Farmer, A.A., Carrino, J., Prange, C., Hayashizaki, Y. The responsibility to share: sharing the responsibility. Genome Research;14(10B):2015-9, 2004.