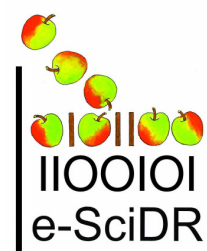


Towards a European e-Infrastructure for **e-Science Digital Repositories**

a report for the European Commission

Executive summary



Harvesting and seeding the fruits of e-Science

Project reference no: 2006 S88-092641

Prepared by:

e-SciDR consortium, led by
The Digital Archiving Consultancy Limited
2 Wayside Court
TWICKENHAM
Middlesex
TW1 2BQ
United Kingdom
www.d-archiving.com

www.e-scidr.eu

for

DG Information Society and Media
Unit F – GÉANT and 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.



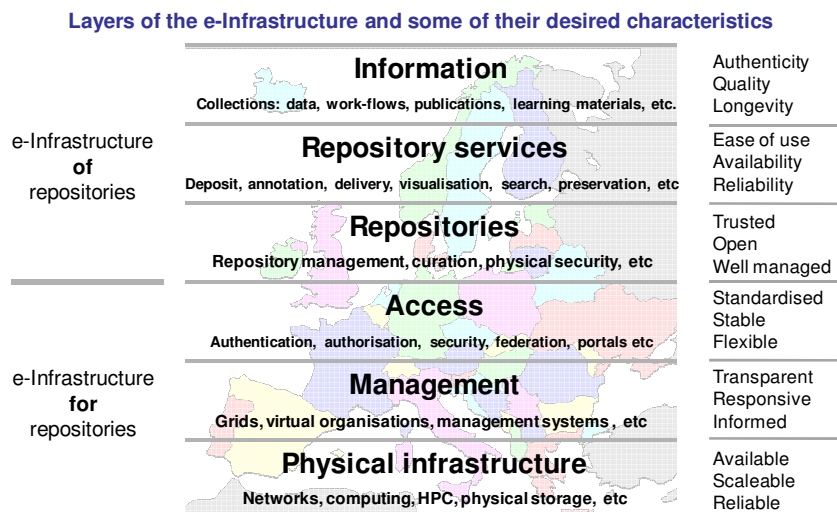
Executive summary

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. However, 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 discovered and understood.

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

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.

Study remit

The e-SciDR study was conducted by a consortium of expert organisations led 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). In summary:

The repository landscape is complex and diverse. Repositories come in many forms, sizes and ages – 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. However, there is a mismatch between availability of resources to develop and maintain the repositories, tools and many elements of the e-infrastructure, and also a mismatch in incentives and career paths relating to data preparation and collection. A further issue of consultation

and ultimately strategic co-ordination is that elements of the e-infrastructure for e-Science repositories (such as access management) overlap or should be co-ordinated with other areas and perspectives.

Recommendations

We summarize 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 main report also groups recommendations according to perspective and level: A: Building an e-infrastructure for research continuity; B: Engaging users and service providers; C: Providing access to researchers, educators, students, unaffiliated stakeholders and interest groups; and D: Maintaining and preserving information. We indicate these groups in square brackets in the recommendation title.

R1. Funding reform [A]

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 [A]

The areas of e-Science, repositories, e-Infrastructures consist of multiple layers, domains, dimensions. 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 [B]

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 [C]

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 [C]

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 [D]

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 [D]

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 [B]

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 for 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.

R9. Governance and management [A]

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).

R10. Training and awareness [B]

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 leadership in e-Science data use and repository management.

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, including 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 [C]

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 a 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 [C]

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 the 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 have been mainly conducted in the past at national or institutional levels. Today's elements for e-Science – data, information in repositories, tools, services and other e-infrastructure elements – are new, numerous, multi-layered; the landscape is dynamic and vast. Co-ordination and strong communication are critically important to driving and sustaining the endeavour.

