

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

Discussion paper – Lisbon workshop, 4th September 2007

This paper is provided to attendees at the final e-SciDR workshop hosted by Portugal's National Archives, at the "Torre do Tombo", on 4th September 2007. It sets out briefly the background to the e-SciDR study, study scope, method, and summarizes the study's findings. It then presents a draft vision, and draft recommended courses of action. The workshop is asked to consider these recommendations.

Background to the study

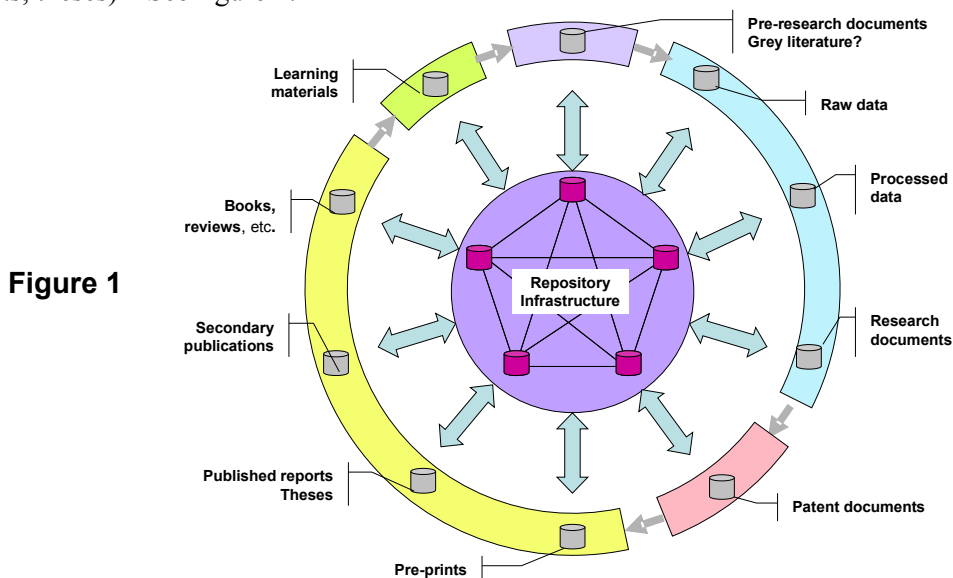
This short study was commissioned by the European Commission to provide an overview of the situation in Europe and recommendations in support of a definition of development scenarios for European-wide efforts to develop e-Science digital repositories for research and education.

Scientific digital repositories are of growing strategic relevance to Europe's objectives of establishing a Single Information Space. Establishing a strong and healthy base of scientific and educational digital repositories is a vital part of the European Research Infrastructure. In particular, the study should provide:

- Inputs to the policy initiatives on e-Science digital repositories
- Inputs to the i2010 Action Plan, especially addressing the objectives of building a European Information Space
- Inputs to the FP7 Capacity Programme.

Study scope

The study has had to cover a very large scope, covering materials held in repositories across the full breadth of the science and research processes, from research planning and administrative data, through raw and processed data to publications in various forms (patents, pre-prints, journal articles, post-prints, theses) – See figure 1.



It has covered many repository types (with different appellations), data repositories and publication repositories (including data libraries, community/discipline-related repositories with support services, institutional repositories, digital libraries and archives, e-learning repositories).

Science is interpreted in the broad sense (“Wissenschaft”), from the physical sciences, social sciences to the arts and humanities.

The stakeholder groups represented by repositories in total are multiple, and fall into cross-cutting categories:

- **Geography** - international grouping; country; region
- **Sector/discipline** – arts, astronomy, economics, genomics, etc
- **Nature of entity** - commercial entity, not-for-profit entity; unaffiliated individual, etc
- **Characteristics**: age; maturity; wealth; level, seniority; confidence; agility, size
- **Profession.**

E-science

The digital age enables new ways of working, new discovery and learning spaces.

“e-Science” is interpreted here as involving some or all of the following:

- Science (in the widest sense, as noted above) which uses computers
- Collaboration with others
- Powerful computation
- Large scale (either in terms of size/volume of data, computation, or collaboration).

Ian Foster neatly summarized typical activities in the pre-electronic and post-electronic ages:

- Pre-electronic:
 - Theorize and/or experiment, alone or in small teams; publish paper
- Post-electronic:
 - Construct and mine large databases of observation or simulation data
 - Develop computer simulations and analyses
 - Exchange information quasi-simultaneously within large, distributed, multi-disciplinary teams.

Defining digital repositories

If a digital repository is to be distinguished from a mere file store, further distinguishing characteristics need to be defined. Some of those identified during the study are:

- A concern for quality
- Forming part of an organisational system, thus with policy and requirements placed on the repository
- A concern for or commitment to sustainability
- Provision, in some way, of a user access view

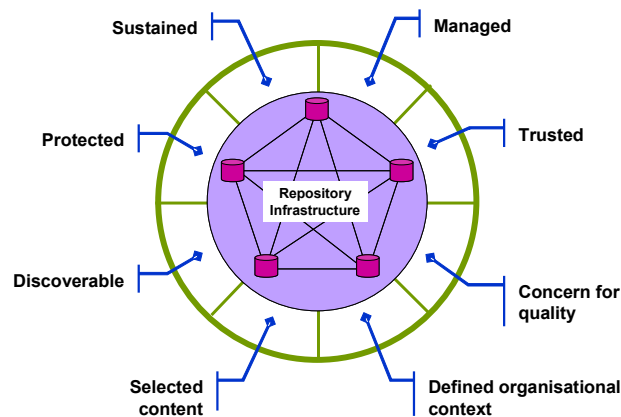


Figure 2

Figure 2 illustrates some of those qualities which contribute to the definition of a repository, in addition to the basic architectural requirements to store content and metadata, make them available and to provide services in some form to deposit, search, retrieve and impose access controls. Not one of these qualities, nor a particular subset of them, is necessary for the definition, but some subset has to be present.

Study method

The study began in January 2007 and is to submit its final report in September 2007. The final report consists of executive summary, body of report, and appendices. As well as the report's bibliography, references have been duplicated in Connotea, and study materials will be available on the study web site as well.

Three workshops were held in Brussels in the first three months of the study with invited experts and practitioners from Europe and beyond, looking at the overall landscape, standards, technologies, and legal and economic issues. We also held telephone and face-to-face with other key informants, aiming to cover the range of stakeholders involved. The team conducted desk research into the current position of digital repositories in Europe, with attention to stakeholder issues, identification of those groups looking at digital repositories, relevant technologies, standards, interoperability, and legal aspects.

Further input into our findings and recommendations was provided by an on-line public consultation, using a questionnaire mounted on the European Commission's IPM system.

Summary of study findings

The following is a summary of our findings, highly condensed for the purpose of this discussion paper. These (including case studies) are presented in detail in the final report and its appendices.

Data, data in science, and the frameworks and technical steps to support their use are subjects discussed and tackled at data user, provider and decision-maker level, across all disciplines and sectors. They have risen steeply up the agenda of governments and their agencies, multilateral and umbrella associations representing groups of interest. Access to data (data sharing) and availability to publications, through good access, are major and higher-profile areas of study and discussion.

At the same time, a huge amount of work is being done to create the rich information space enabled by information technology, by libraries, information scientists, and also by the commercial sector. Google is a major player - its stated ambition is "to organize the world's information and make it

universally accessible and useful". It is a fundamental reference point: researchers, students, teachers want the speed, power and ease of use it provides.

"e-Science digital repositories" covers a wide range of different resources, varying in type, age, size. Until recently there has generally been a marked division between those that contain data and those that contain documentation (bibliographic and published paper content).

If the repository landscape itself presents a complex picture to the average researcher, then it is also sitting in a complex matrix of technologies, facilities and unfamiliar and fuzzy terminologies: e-infrastructures, Grids, network technologies, web technologies such as Web 2.0, "SOA" (service-oriented architectures), the semantic web, and so on. Much of this will be obscure to users. Contrasting with this is the ease of use and intuitiveness of the Web, but in particular new resources such as YouTube and Del.icio.us.

In the research cycle shown in Figure 1 on page 1, each new item of content is to some degree supported by what goes before in the chain; in this sense there is a continuum of digital content, which (if recorded) would provide a chain of provenance and authenticity. The pipeline absorbs digital content as well as producing it, as well as a flow along/within the chain itself. The fragmented external repository landscape does not mirror what is in reality a process continuum and actually imposes a set of artificial barriers.

We also noted in the introduction that e-Science is (generally) an international activity – the practitioners do not want to be concerned with international and supranational boundaries (raising pressing legal issues), though successful science bolsters economies, the environment and the social fabric at local, national and regional levels as well as international.

Drivers and rationale for e-Science digital repositories

e-Science enables new ways of working. Experiments can be conducted "*in silico*", simulations run, massive comparisons, analyses – activities opened up to groups hitherto unable to participate in science. Inevitably, these changes impose strains – organisational, cultural, technical, legal – on traditional frameworks.

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). Some of this is unique observational data.

The sheer volume of data is a major operational and cost pressure, for managers and administrators. What do you keep? Where do you put it?

For users, there is a huge problem of handling the information, and the problem of finding useful information in the first place. Our ability to generate and collect information continues to grow more quickly than our means to organize, manage and use the information effectively; our ability to do so is of extremely high strategic importance. Thus efforts to create enduring digitally based tools and resources which enable us to organize, manage and use data and information are particularly important – such as taxonomies, indexing tools, ontologies such as GO (Gene Ontology).

Re-use and re-purposing of data are a benefit, as well as being a fundamental driver to activity (which comes up against cultural, organisational, technical obstacles). The benefits of access to data are summarized in Box 1, quoted from the OECD Principles and Guidelines for Access to Research Data from Public Funding [2007].

“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.

Summary of issues and themes

Core themes and issues identified:

1. As well as sheer scale, **complexity, heterogeneity and dispersion** are identified as major challenges. Scientific data are often highly **specialist** and only understood by experts; several studies raise the question whether institutional repositories are equipped to manage discipline-specific data in which the repository has no scientific expertise. There is also heterogeneity in metadata, in structures and between and even within disciplines. These are not just management issues, they affect the amount of materials discoverable and the ease with which they can be located and accessed.
Some materials are also dynamic (and the traditional concept of a repository risks locking data into a static representation, as Jürgen Renn and Malcolm Hyman have pointed out).
2. There are also worries about a coming **deluge of metadata**, dwarfing the data deluge.
3. The axes of **communication and professional incentive** are community and discipline based.
4. There are **differences across Europe** in the level of use and penetration of repository technologies.
5. **Inappropriate funding models** apply to the maintenance of repositories, for their own efficiency, sustainability and the preservation of content. Funding is also inadequate.
6. **Harmonised and simplified authentication and authorisation** mechanisms are needed across Europe to gain access to e-Science resources in general and repositories specifically
7. There are a few digital repository **notification services, registries** of data, and registries of repositories, but no one reliable, single pointer

8. A need to avoid **data loss** – valuable data is slipping away for lack of awareness or for a suitable place of deposit. For example, how can we capture grey literature?
9. At the same time there is some evidence that many repositories are **poorly supplied with content**. Several surveys point to disappointing levels of materials in institutional repositories. On the other hand, we also note that important and long-established digital resources took time to reach critical mass.
10. **Quality** of data and the need for good metadata. Without this, data will not be used. This is stressed as a key factor (if not **the key factor**) in success and thus sustainability of digital repositories.
11. Tools are needed to automate **metadata generation** and help users provide metadata.
12. **Incentives** are needed to encourage data generators to deposit (share) their data, and provide good-quality metadata. Incentives include **citation** and publication; this is almost non-existent for data, so this mechanism (and supporting framework) for professional recognition of work and expertise in data management is unexploited.
13. There is a substantial need for **training**, of those working in digital repositories, libraries, of users. The Association of Research Libraries for instance points to the need to train more information professionals able to discover, locate, reference, create, manage and present digital content, more information and library professionals who can work on data curation in research teams.
14. Following key studies by CODATA into **data sharing**, the lead of some institutions and disciplines, and studies into the utility of mandates, several funders have produced **data policies** and **data management guidelines**, and **mandates** for data submission to repositories.
15. **Roles and responsibilities** are shifting, and there is a need to identify the roles and interfaces involved at different stages in the digital repository chain.
16. **Cultural and behavioural issues** are frequently identified as major obstacles to the population of repositories (and creation of metadata) – such as fear of misuse of data or loss of ownership.
17. A closely related issue is that of the **period of privileged use** of an object. There needs to be a balance between public access to data and a researcher's right to privileged time for use.
18. Can we develop an understanding of **how data can be re-used**, re-purposed? This would be useful for collections management and preservation management, and would inform rights management tools and frameworks.
19. **Collections policy** and collections management: **Appraisal** is a major issue, particularly in the face of the huge and growing volumes of data. What does the repository keep (links to objects, whole objects, versions, annotations)? Is there **co-ordination** of holdings at national, regional, global level? If so, who keeps master copies of data?
20. Questions on **organisational structure** to support data curation in the context of digital repositories, organisational relationships are being examined in actual initiatives and testbeds.
21. **Preservation**: long-term access depends on preservation practices. Research is needed into good practice, technologies. What institutional frameworks might best support preservation – national and international repositories?
22. There is work on **certification** for repositories; this will encourage trust and usage and improve quality; it implies a need for organisational homes which will provide the framework for such a system.
23. **Permanent digital object identification methods** are needed, noting that the DOI (Digital Object Identifier) scheme may not be adequate for wider repository use.

Public consultation

The public consultation was held over six weeks from mid July to the end of August 2007. It harvested 426 responses, from users, repository managers, researchers, librarians, publishers,

students, commercial companies and service providers. The consultation is anonymised, but we can confirm that contributions came from leading figures in data management, repositories, libraries.

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, 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%). As might have been expected the most common forms of repository used were community and discipline-related repositories, digital libraries and institutional repositories.

Some of the headline statistics from the survey were:

- 39% never paid (directly) for repository use
- 63% had no training in repository use
- The main difficulties encountered during use were:
 - Finding it time-consuming to find information (55%)
 - Time-consuming to deposit data (34%)
 - Not knowing where to look for a suitable repository (27%)
 - Lack of training or guidance (39%)
 - Too costly and too difficult to use were each mentioned by just under 25% of respondents
 - NB while language was not seen as a barrier by most respondents (but the language of the questionnaire was only English), there were numerous free-text comments on language obstacles
- 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.

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

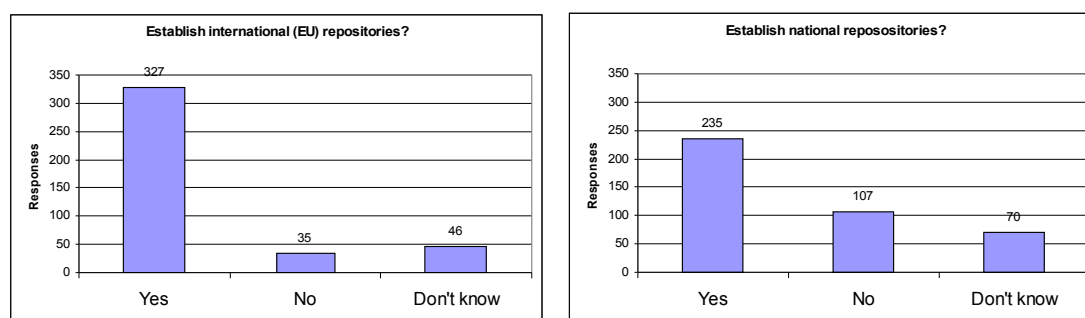


Figure 3: Support for international and national repositories

The following chart shows the wide range of content these respondents deposited into and used from repositories.

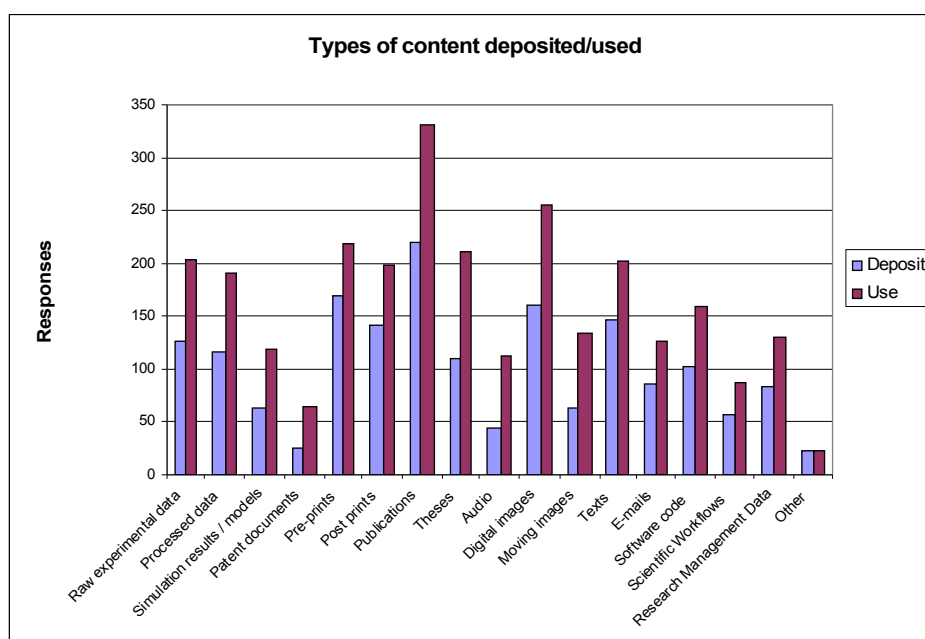


Figure 4: Types of content deposited and used

Vision

To formulate policy options we need a vision of what those policies need to achieve. In a little detail, vision for an infrastructure for e-Science digital repositories in Europe:

- It should support the scientist at all points in the research cycle by providing easy, cost-effective access in a joined-up fashion to materials of all types that are already available (subject to well understood precautions in respect of ownership, privacy and ethical use), thus supporting excellence in science and innovation
- Support easy and reliable deposit of materials for science, research and learning into known, trusted repositories through the whole research cycle, providing confidence that the materials will be well maintained, and not abused.
- The collections in repositories are expertly maintained
- The repositories should have a capacity or associated framework to support the long-term sustainability of information, be trusted, guarantee the authenticity of stored materials and cope with future demand
- The infrastructure delivers services equally across the whole of Europe and participates as leaders and partners in the wider global e-science information infrastructure
- The various stakeholders - administrations, the scientific community, the private sector and the public - 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.

Recommended courses of action

These are the major draft recommendations for action. There is also a substantial list of other recommendations, technical, legal, organisational and other. These suggestions are formulated as courses of action, and accompanied here by brief commentary; the final report will formulate the recommendations to the European Commission in the appropriate style.

The recommendations below are cross-cutting; some could be combined, depending on business model. We will present some scenarios for these recommendations at the workshop.

1. Digital repositories and related infrastructures should be funded on a rolling or long-term basis, under criteria aligned to the purpose of the repository.

As well as instantly improving sustainability, this change will also increase resource availability at no extra cost, enabling open allocation of resources to provision of service, and will release all the time otherwise taken to seek renewed funding. The additional resource availability will go to improved customer service, service and infrastructure development, thus better return on investment. An essential aspect of improved customer service must be sufficient funding to provide intuitive, easy-to-use user interfaces.

It will also help people express pride in their work, supporting quality and indirectly boosting the resource pool.

There are several ways of maintaining quality of service – salary bonus schemes; requirement to re-bid to be service provider every five years (say).

An important factor is that the governance structure and management of the resource takes into account the span of active use of the particular resource.

However, there should still be funding for the creation of new types of resources in research contexts.

We endorse the need for software repository and service (such as OMII), which should also have rolling funding.

- 2. Planning, reporting, recognition, awareness:** To help sustain appropriate infrastructure funding levels, repositories will need to continue to demonstrate the value and benefit of the work they do. This should be achieved by means of reporting, using objective, pre-set metrics, reported annually in a formal report. Repository entities should have a governance structure, objectives, business plan, strategy, and resource allocation.

The repository board and funders should recognize that the benefit generated by the repository's work will not accrue to the repository, but to others.

Wider public awareness and thus recognition of the work, expertise and importance of digital repositories, their services, is also fundamental to sustainability: there will be greater willingness to maintain funding levels, and more people will be attracted to the profession.

This will require communications activity.

A possible benefit of more active communication and awareness might be more easily programmed meetings and actions at international level, as well as motivating users (scientists, teachers, students etc).

3. **This funding should also be sufficient to support the creation and maintenance of core services and tools both at community and generic level (e.g. controlled vocabularies, ontologies, checklists, ingest tools).**
4. **At data producer level: There should be specific allocation (and monitoring) within funding of research and teaching, for good data management by data producers from before point of creation through to deposit.**

This will require data policies and support for the data producers, for example in the form of advice on software programmes, database schemas, semantic conventions, vocabularies, etc, and training in their use. This support framework should consult, liaise and co-ordinate with the repository providers, relevant data science.

The direct and indirect benefits of this will be vast: greater awareness amongst data producers (also users) of the reasons for good data management; better-quality data (accompanied by the requisite, and more accurate metadata), so (a) lower costs at repository “ingest” stage and (b) better-quality data for downstream use. The improved-quality data, accompanied by the planning information, can also feed into preservation planning.

At data-producer institution and higher levels, the data planning will increase interoperability, enable identification of economies of scale, needs and opportunities at scientific, administrative and financial levels.

The planning and digital resource information can be collected and provided in advance to the downstream repositories, for their resource planning (also further down the line, at preservation level).

This implies close co-ordination and good communications frameworks between data providers, repositories and preservation layers.

5. **Publicly funded activity should mandate that digital output is submitted to a repository (designated or approved); the repository does not have to accept the item. This output must conform to specific criteria, including data integrity, and ready, equitable accessibility.**

This will require data management policies, support and co-ordination frameworks and information flows between funders, data providers and repositories. There should be workflows and automated pipelines to help compliance and reduce costs.

6. **A European-level multi-lingual gateway providing comprehensive, concise, clear registers of repositories, services, and resources.**

This should include libraries of information, for example libraries on repository policies; off-the-shelf governance structures, etc. Behind the single gateway this would be a federated resource, as the expertise to maintain it would be dispersed. The resource’s contents would be exposed to search engines, and it would provide direct links to the repositories.

This is at meta-repository level, above resources and gateways such as the EBI.

The DRIVER project has already begun building a resource along these lines.

Such a resource would need (a) the type of funding recommended in 1, and (b) more funding, in particular to ensure ease of use: unless the interfaces are intuitive and easy to use, it will not be used. This requires expertise.

7. **Centres of repository excellence:** community-based and generic: these would provide support to users, and also some of the support entailed in recommendation 3.

The centre(s) could also work on interoperability and repository federation issues.

8. European-level repository facility, available to eligible entities.

This would provide e-Science repository facilities (with easy to use, multi-lingual interfaces) for those without access to or unable to afford suitable storage or repository resources. It could also provide a home for orphan data. The facility could also be leased out to commercial customers.

Economies of scale could be available with areas outside e-science digital repositories.

(Conversely, the storage space might be provided by a wider European data storage layer, which might provide storage to institutional repositories.)

This facility might also support a repository for EU-funded output. Currently much of the output (including web pages) generated in EU-funded research disappears, for want of a mandate for its submission to a designated or approved repository. Some materials will need to go to specialist repositories.

9. A range of **preservation-related** activities needs to be funded. One of these should be to establish representation information registries.

e-Science digital repository holdings pose particularly difficult preservation challenges, and will need to draw on preservation services and advice over the life of the objects concerned. However, this need is common to all other digital objects, and a corresponding, over-arching provision layer may be more appropriate.

10. **Selection and appraisal:** research is needed into data appraisal (criteria, processes, support tools, possibly even different approaches for the digital information age). There is an established body of expertise dating back hundreds of years for documents. An equivalent needs to be established for data, at generic level and taking into account the needs of the different communities.

Selection must be underpinned by repository and/or collection policy.

11. **Discovery:** This was possibly the most frequently and vehemently raised need in the public consultation. More research is needed into searching and harvesting methods and tools.

Support is needed for ontologies, vocabularies, user interfaces and querying, text and other format mining.

There should be research into cross-repository searching.

More research is needed into persistent unique digital object identifiers; these will need to be more granular than the DOI system.

Identifiers will also be needed for repositories (there is current work on this in NISO, but this will need review).

12. **Fund or otherwise promote further research into how to link data along the information chain, from raw data to final publication, in a seamless manner and regardless of where items may be stored.**

As part of this, try to establish standards of demonstrating the chain of validity and provenance from raw data to publication. Possibly set up test beds to show proof of concept.

This linkage captures workflow of the *whole* research process, rather than parts.

It also bridges the divide between data and publication.

13. **Establish data citation:** This incentivizes deposit data into repositories. For citation data will have to meet specific criteria with regard to quality and interoperability, and this will be a major contributor to sustainability: repositories will need to do less work rescuing data on ingest, users will trust repository contents more; users will also be emboldened to release their data because

they will know they get recognition for their data, so volumes will rise. Data citation will need frameworks and mechanisms. Data journals will help (and also contribute to awareness in 5 above).

14. **Harmonisation of access and authorisation methods and techniques across Europe** to provide single sign-on. It must be simple (if not invisible) for the users. There should also be sufficient security, with levels to meet the needs of commercial collaboration.

15. **Training:** There must be training at multiple points and levels, for users, repository workers, managers, curators. Training is one of the top priority actions. A framework of training programmes will be needed. We also recommend training and awareness in schools. Training results in immediate increase in levels and quality of use. It also provides communications channels for suggestions for improvements in tools and resources, and identification of needs.

There must also be cross-training between information scientists, computer scientists, librarians, in what they do. This will contribute to resource discovery, and also help individuals in the shift in skills sets entailed in the digital information age.

16. **Career structures** need to be established for people working in repositories and curation. Data citation, data journals, repository reporting will help maintain a resource pool. In due course, professional qualifications might be established. In the near term, career paths must be established and communicated within the relevant institutional frameworks.

17. **Certification of repositories:** Repositories designated as deposit repositories should have certification as trusted digital repositories. This requires not only the certification standard(s), but also a certification framework, including issuing body and training.

Certification standards have been and are being developed. However, there is little implementation experience and information as yet. This also represents an opportunity for the certification body and/or European and national resources (such as under 6, 7 or 8 above) to gather information, to inform updates to standards.

Training will be needed for applicant repositories. Certification should include a requirement for regular renewal of certification status.

Certification should also extend to the commercial (on a fee-paying basis) and unaffiliated sectors; services could also be provided, as available. Certification need not be provided to European countries only.

18. **Networking, co-ordination:** There should be a resource which provides for networking, contact, information, co-ordination and research between all types of digital repositories, and between repositories and other parts of the e-Science, science, learning and in particular **information chain. This networking resource (which could be combined with 6 above)** would play an important information and research role, channelling, identifying and co-ordinating participation in cross-repository initiatives – for instance, taskforces to identify generic repository elements, operational costs and opportunities for cost sharing, standards development, research.

It could also provide a professional association.

19. **Legal and regulatory:** e-Science works across boundaries; however, at these boundaries, laws often change to a greater or lesser degree. New legal infrastructures are being developed which take this into account (Creative Commons, Science Commons), and these should be supported. Harmonisation of copyright legislation, cross-border data exchange, and clinical confidentiality regulations across Europe would contribute substantially to efficient e-Science activities.

There have been several instances in recent years where legislation has been drafted which makes e-Science more difficult or more costly at multiple levels. The increasing profile of e-Science digital repositories should help ensure that they are consulted during legislative or regulatory drafting or review.

Clear information about intellectual property rights should be provided (accessible at 6). Students, scientists, researchers, teachers should be provided with basic training in IP rights, so that they understand what is entailed. They should be encouraged to pass a basic certificate in IP.

Accessible at (and perhaps co-ordinated by), there should be guides for researchers and institutions on the legal framework for creation, deposit, access and re-use.

The legal and liability status of repositories should be clarified, at general and specific levels.

20. There should be support for good-quality data collection, maintenance and curation in the **developing world**, and access to advice and, where appropriate, repository facilities and services.