

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

Project reference no: 2006 S88-092641

Interim report 2



for
DG Information Society and Media
Unit F – GÉANT and e-Infrastructure

Date: 21st February 2008

Document status: Final

Prepared by:

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

www.e-scidr.eu

Interim report 2

Contents

Section 1: Further analysis of the situation in Europe.....	4
Section 2: Standards landscape	13
Section 3: Legal implications of open access.....	23
Section 4: Results of the Public Consultation	27
Section 5: Study Workshop Arrangements	61
Recommended courses of action	75
Appendices	82



e-SciDR: Interim report 2

Introduction

This is the second interim report for the e-SciDR study (“IR2”). It complements the earlier First Interim Report (“IR1”), which provided an overview of the situation in Europe concerning e-Science digital repositories.

This second report sets out:

- A further review of the situation in Europe concerning e-Science digital repositories
- An analysis of the standards landscape for digital repositories
- Legal implications of open access in the context of e-Science digital repositories

In addition it presents:

- The results of the public consultation held during July and August 2007, and
- The plan for the Study Workshop held in Lisbon in September 2007.

References

The bibliography for this report accompanies the final report, with the exception of a bibliography of more specialist, technical literature on the standards, incorporated into this text.

The reports’ bibliography is also hosted on the e-SciDR web site, with active links to referenced materials.



Section 1: Further analysis of the situation in Europe

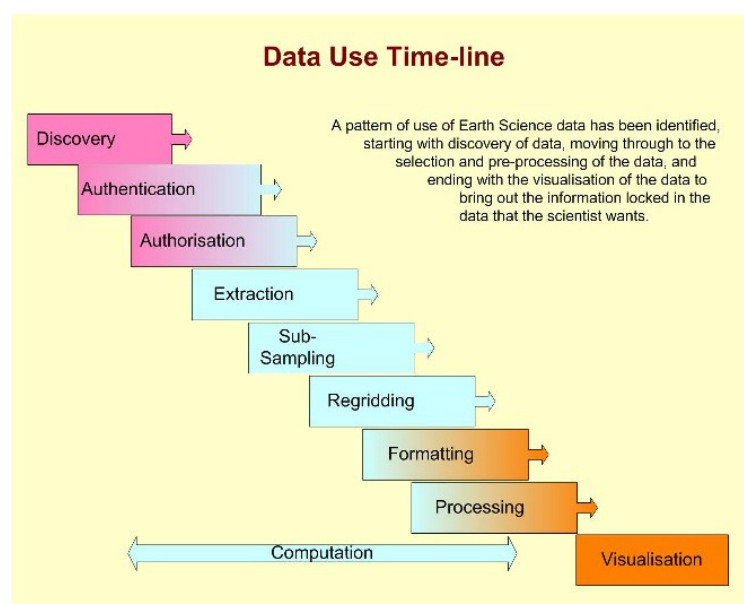
Overview

This section sets out a brief review of the landscape of digital repositories in the context of e-Science. We do not repeat areas covered in the first interim report.

A digital repository is the construct which holds digital data or information. As discussed in our first interim report, this construct can have various services associated with it. From the perspective of a digital repository, its minimal core functions centre around taking in material, housing it, and enabling access to the materials. From the perspective of the owner or other stakeholder in digital materials, additional core functions are stewardship and maintenance of the material, but these functions are not necessarily attributed to the repository housing the material.

Basic repository functions - ingest, storage, access – are already demanding

Simply fulfilling the three core functions of ingest, housing and access is not a simple matter where digital data is concerned, in particular scientific (in the broadest sense of the word). Typically scientific data is heterogeneous, often complex. The following diagram, prepared by the NERC DataGrid¹ project, illustrates the multiple steps that a user may have to go through in order to load and run some data, and that the repository holding the data must be able to support:



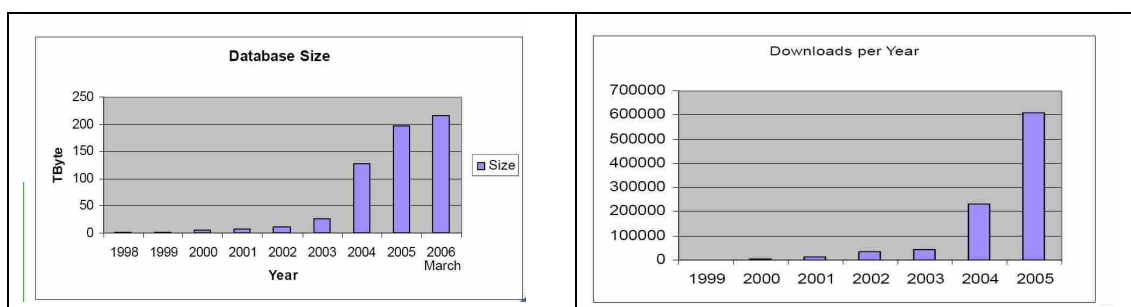
NERC DataGrid diagram, 2003

Optimizing a very large database to run efficiently can require weeks of collaborative effort, work which may have to be repeated every time there is a new release of the underlying database software.

¹ <http://ndg.nerc.ac.uk/> This project aims to facilitate data discovery, delivery and use.



This and other work has had to be done at the same time as supporting enormous increases in volumes, in terms of new material into the repository, numbers of users and numbers of downloads.



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

A few more statistics illustrating different types of growth: The Nucleic Acids Review, a major journal for genomics (inter alia), each year produces a database issue, in January. The editorial of the issue reports on a count by the journal of databases which qualify for coverage. This number has risen from 226 databases in 2000 to 968 in 2007.

The UK Data Archive, founded in 1967 (celebrating its 40th anniversary in 2008), houses thousands of datasets from the social sciences and humanities in the UK. In 2003-04 it recorded 588, 193 separate visitors to its web site, but in 2006-07 this figure was more than seven times greater, at 4,349,052. Its annual reports show that in 2004-05 it processed 164 new datasets for online delivery, compared with 114 in 2003-04 (processing new datasets is far more resource-intensive than processing updates or new editions of datasets). At the same time, however, the UKDA had to do this with no increase in human resource base.

Little heralded work by repositories and associated providers

Many repositories also provide services and tools to make acquisition of, access to and use of their holdings easier and richer. These range from optimization of storage, servers that are powerful enough to support the weight of usage (in and out), to participation in standards development, development of pipelines to support easy, efficient submission of well-prepared data, specialist search and query tools, visualization tools, and semantic tools and services, to support discovery of items mis-spelt items, homonyms and so on.

Many of these tools are developed by providers outside the repository itself. Often, in the past this activity was funded by finite, short-term project funding, and the developers also moved on, and documentation on the tools' technical development was often found to be wanting. The primary point here, however, is that the tools were developed externally to the repository (examples are posted on the e-SciDR web site).

These additional services and tools are predominantly at scientific and computational levels; they are less active work at repository level as regards administrative dimensions (for example, relating to identity management, systems to support rights information). This also reflects the repository's expertise.



The top page of the EMBL submission portal (www.ebi.ac.uk/embl/Submission/index.html), “provides a single point of entry for submitters of all ENA (Ensembl Trace Archive) data types”.

What’s in a name?

Our first interim report noted that members of the “digital repository” constituency bear many different names – for example, data centres, archives, data warehouses, facilities, data libraries, data sets, databases – or indeed they have a designation which does not include any related word or synonym – Ensembl, for example.

There is also a question of division, or size: for example, one team member pointed out that, while the European Bioinformatics Institute houses multiple repositories, it is itself a repository. The same applies to many other institutions; indeed, another team member suggested that the Internet as a whole might be classed as a repository.

Defining and identifying digital repositories

Our definition of a digital repository, as the construct which takes in, houses and provides access to digital materials, is broadly drawn. In Interim Report 1 we also noted the minimum qualities which should characterize a digital repository as being a subset of the following:

- A concern for quality
- Forming part of an organisational system, with policy and requirements placed on the repository
- A concern for or commitment to sustainability
- Provision, in some way, of a user access view
- Some degree of interoperability,
- Having some of the following characteristics: sustained, managed, trusted, discoverable, protected, having selected content.

The resources covered by this rather fuzzy term are vast. They cover the full range of data formats and types, from texts and images (moving, still), to many forms of databases, from all types of sources – from instruments, sensors; processed data, simulations; analyses, to texts



and publications, from the entire spectrum of subject disciplines and sub-disciplines, across Europe and more widely.

There is no single register or catalogue of these resources.

The DRIVER project (Digital Repository Infrastructure Vision for European Research) carried out a survey of repositories compliant with OAI, a study published in 2007. By its nature, it focused predominantly on text-based repositories.

Types of repositories

Many repositories are continuations, or digital arms, of existing archives. Countries' national archives, for instance, are now also digital repositories, extending their fiduciary role to the digital format. Public broadcasters' digital resources also belong to our subject area, and providers of "social networks" (such as SlideShare and YouTube) also fall into our category.

The "Long-lived data collections" study identified three core categories of data collections: research data collections (the product of one or more focused research projects); resource or community data collections (serving a single community); and reference data collections (intended to serve large segments of the scientific and user community). Of course, a collection can change category over time. (The e-SciDR web site has posted some examples.)

In our first interim report we stressed that a repository is not the same as a collection. A repository can contain multiple collections, or part of a collection.

Reviewing the landscape, we identified several other categories into which repositories can be classified, for example, by age, size, location, sector. Some hold collections which are closed when they acquire them, others remain open. One useful distinction, we believe, is between repositories holding material which is generated *in situ* or *ex situ*. Where the repository takes in materials generated externally, it has less direct control over the format and condition of the incoming material, and therefore must do more communication and co-ordination to endeavour to receive materials in appropriate formats and with adequate metadata of good quality.

Evolution of digital repositories

The history of digital repositories goes back over half a century, to the earliest days of computing. In 1955 The International Council of Scientific Unions (now the International Council for Science) recommended the creation of World Data Centers ahead of the International Geophysical Year of 1957-58. Multiple Data Centers were established around the world, to protect data held and for the convenience of users. As the WDC web pages note, "the 1955 recommendation mentioned that Data Centers should be prepared to handle data in machine-readable form, which at that time meant punched cards and punched tape". There are now some 50 World Data Centers, based around the world, acquiring, storing and providing access to their holdings, though with funding which they sometimes have to struggle to retain.

Key milestones in the advance towards today's data resources include the advent of the personal computer, the handheld computer (enabling scientists to gather data during their work), and the increase in processing power of computers, increase in storage capacities, and in the infrastructure, power and speed of telecommunications and networks. Particularly important was the arrival of the Internet, underpinned by the simplicity and lightness of the



HTTP protocol, making the Internet accessible to all, subject to having access to a computer and adequate communications bandwidth.

The Web and the Grid

The turn of the century and first years of the 21st century saw a rapid shift to web-based access and downloads, widening as Internet access extended. This posed a substantial burden on existing repositories, where access had previously been based on FTP (file transfer protocol) or indeed the mailing of CD-Roms in the post. Users wanted web-based access. In some disciplines, some journals initially refused to acknowledge scientific work supporting web-based access, penalizing those who pioneered web-based querying tools, or database structures to support web-based access. This was soon reversed.

At the high end, the Grid was also developed, enabling Grid users to access, share, analyze and process extremely large data sets, or bring together multiple sources – enabling the high end of e-Science. As we will see in the following example, data Grids can come with data services, which of course need to be developed and maintained.

The power of instruments

The other major driver is the ever-increasing power of instruments, and the increase in availability of powerful instruments, as they become cheaper.

With instruments increasingly generating their output in digital form, this needed to be housed in digital repositories. But scientific and technological advance also meant (and still does) that we are able to monitor more and more than ever before (and also enable access to observational data in real time).

Data sharing

There was also the recognition that publicly funded research (and indeed education generally) generated data as well as studies, and that this data was (a) expensively generated, (b) had been paid for by taxpayers and (c) could be re-used.

We live in an age of data deluge; a large proportion of this data was expensively created (some of it extremely so, for example space data), some of it unique observational data. The return on cost of investment would be multiplied by wider and deeper use of the materials.



The Eagle Nebula: “This image was taken by the European Space Agency's Infrared Space Observatory, ISO, which operated until May 1998 [...]. As an infrared telescope ISO had the ability to see objects and material that other telescopes cannot see, for example, cold dust*. The dust in the Eagle, seen in the picture as a 'bluish fog', is at about minus 1000C. Although perhaps difficult to believe, it is inside freezing dust like this that new, hot stars are born. In this image ISO has captured a view of the ice enshrouding the fire. The Eagle nebula is an active 'star nursery' located 7000 light-years away, in the constellation Serpens. It is a huge cloud composed mainly of gas with

microscopic particles of dust. Surprising as it may seem, its cold temperature is a key requirement for star-birth to actually occur.”

Text and image harvested from

<http://sci.esa.int/science-e/www/object/index.cfm?fobjectid=28114>



As more people could access publicly available data, and share data, more people encountered difficulties in using other people's data, heterogeneity being one of the main problems - a mass of different database schemas and structures, using different field names, labels, different terminologies; updates to databases were not notified, and there was insufficient documentation to help users load and run the data. So, bottom up, in the range of different scientific disciplines (from humanities, social sciences to the life sciences), communities got together to work on facilitating interoperability and standards.

These drivers together spurred activity and initiatives around the theme of data sharing, peaking in the years 2003-2005 in the work of groups such as CODATA, and the publication of several influential multilateral reports. Obviously, repositories facilitate data sharing, providing a single point of access to materials.

Interestingly, an important term used in data sharing context is “community resource”, and repository is sometimes synonymous with this, a phrase implying community governance.

Umbrella access

A next step was to link digital repositories, to enable users to search across and access materials from a potentially global source of materials. Several communities and groups have begun this endeavour, setting up a facility through which users can browse, search, query, access and retrieve materials from multiple repositories.

Again, these facilities carry different names, such as “facility”, “portal”; some add the adjective “virtual” to a community resource name.

In the case of astronomy, the name used is “virtual observatory”, and it was one of the earliest examples of establishment of grouped access to multiple repositories.

Example of umbrella facility: IVOA

Astronomy's virtual observatory framework provides an excellent illustration of what is needed for this type of capability. Effectively, this capability is a type of research infrastructure.

Astronomy has always been rich in data, and was amongst the earliest in setting up data centres, supporting electronic publishing and also building links between distributed systems, thanks to exchange standards such as FITS (Flexible Image Transport System) and the Bibcode for describing bibliographic references (now used more widely, for example by NASA's Astrophysics Data System). The FITS format was proposed in 1981 and adopted widely; it enabled data to be interchanged easily between astronomy's sub-disciplines, and helping accelerate discoveries about the origin and evolution of the universe.

The aim of the Virtual Observatory (VO) is to make access to astronomy databases as seamless and transparent as possible, federating data flows from astronomy facilities, surveys, computational resources and tools to use these. The IVOA – International Virtual Observatory Alliance – is the body which co-ordinates the work of the various VO projects worldwide, and agrees on technical standards.

One of IVOA's earliest actions was the creation and maintenance of a new astronomical data format, VOTable, which uses XML (thus supporting automated exchange of data). The Grid provided power for transfer of very large volumes of data, and IVOA set up technical working



groups for service registries, content description, data access, data models and query languages.

The VO aimed to be transparent and seamless, like the WorldWide Web; a concomitant of this is independence of location. Another aim was to support “collaboratories”, distributed research teams who share data, workflows, and results. The VO offers the ability to perform operations on the data and return results, with the Grid a key enabler given the sheer size of the datasets involved.

To achieve their vision, the IVOA identified five areas of work:

- Development of, agreement on and adoption of standards and protocols
- Development of “glue” software components: portal, registry, workflow, user authentication, virtual storage
- Adoption by data centres, who need to publish to the system (who need to write – code - data services which comply with the VO system)
- Develop and maintain tools to work with the data
- Establishment and maintenance of resource registries and user support systems.

Of course, this means that repositories must regularly update their systems to support changes in web browser technologies and systems.

In January 2003 the IVOA decided on six major technical initiatives needed to achieve the international virtual observatory:

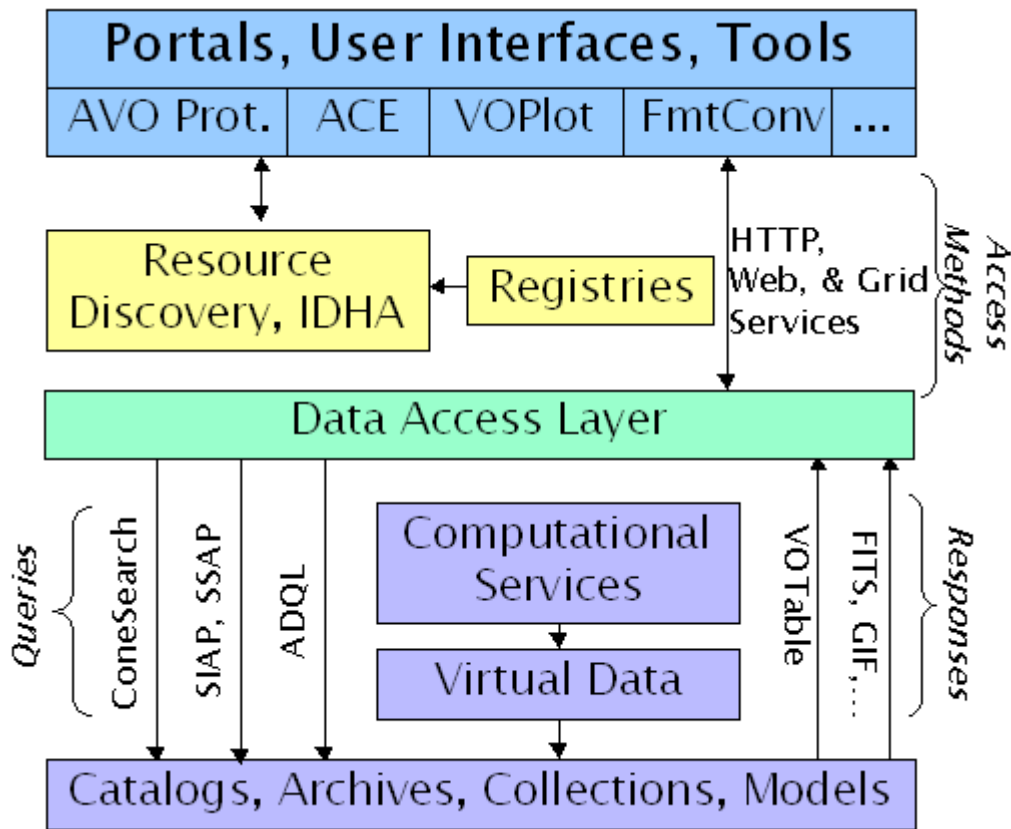
- Registries – lists with metadata about data resources, information services, gathered into a database which can be queried, and which can be distributed; the community sought to apply a registry standard, the Open Archive Initiative Protocol for Metadata Harvesting
- Data models – the FITS standard allowed many variations in the way metadata could be encoded; the data models aimed to define common elements of astronomical data structures and provide a framework for describing their relationships
- Uniform Content Descriptors
- Data Access Layer – provides standardized access mechanisms to distributed data objects.
- VO Query Language – a standard query language was needed, to work with the many, distributed VO databases
- Grid and web services – The VO is effectively a service Grid, with nodes where tools are located with data collections, needing standard web service interfaces; also needed are single sign-on for authentication, and workflow management.
- VOTable – the astronomy XML mark-up standard.

There are two points to stress here, common to most other federated or linked resources: firstly, at least five of these initiatives need to be maintained and supported over time; secondly, while the VO registries can be queried using OAI-PMH (thereby enabling ready



federation with other OAI-PMH-compliant resources), the querying *within* the VOs themselves, at content level, cannot use OAI-PMH.

The building blocks of the VO infrastructure. It is not a linear text.



Credits for diagram: Szalay, A., Williams, Hanisch, et al., 2004. From: http://www.euro-vo.org/cgi-bin/twiki/bin/view/Avo/PublishDataToVO#Data_Access_Layer

On a final note on this example, interestingly, the astronomy community also agreed at its 2003 annual general meeting on open access principles, asserting open access to publicly funded materials after a period of privileged access.

Curation and annotation

We noted above that repositories can hold open or closed collections. There can also be more than one type of “open” content, as for example in proteomics, where the rate of new discovery is high, and existing content often needs to be corrected or re-worded for the new discovery or annotated. Given that these resources often contain hundreds of millions of entries, this curation is a huge challenge: how to cope with the volumes, and at the same time ensure quality of curation.

Community axes

Again and again in our survey, the role played by subject communities was clear and key. These are the fora in which scientists, researchers, sector computer scientists generally decide on data management needs, directions and roadmaps (data management invariably including data centres or repositories).



Similarly, care and advance of data repositories is generally entrusted to subject specialists, at science, computational and informatics levels; specialist expertise is a sine qua non.

The most common exception comes in institutional repositories, by their nature usually having to cover a wide spectrum of disciplines, but working predominantly with text-based materials.

Users and access

The user base for digital repositories is potentially comprehensive, ranging from members of the relevant scientific community/ies (from researchers to students), special interest groups, journalists, the commercial sector, to individuals unaffiliated to any institution. These users can also be data generators, and increasingly important groups are members of the public and unaffiliated researchers or students.

Weather data is obviously a matter of interest to everyone, in all walks of life, but it is still revealing to see the long list of categories of customers for weather data:

Users of Denmark's Meteorological Institute include:

Newspapers	The armed forces	The offshore industry
The building industry	Railways	The police
Electronic media	Municipal and county authorities	Radio and television
The energy sector	Cultural arrangements	Travel industry
Contractors	Agriculture	Shipping
Insurance companies	Road traffic	Schools
Research institutions	Aviation	Sporting events

Many repositories support entirely open access by users. Others, because of bandwidth and support pressures, have to prioritize support. A few repositories ask users to register, but do not track or control usage in any way; registration enables the repository to keep an accurate count of number of users.



Section 2: Standards landscape

The vision of an e-Infrastructure for European e-Science Digital Repositories is a compelling one. Ideally this e-Infrastructure should allow researchers to find and access a rich range of information² held in data repositories determined by their end user research interests and their associated privileges. Targeted tools and community services supporting discovery of, access to and analysis of the contents of these repositories is essential. Common standards underpin the successful deployment of advanced digital repositories that span discipline specific silos of information and facilitate inter-disciplinary sharing of data resources.

In this section we use the term “standards” to cover standards developed by formal standards-setting bodies such as NISO³, ITU-T⁴ and ISO⁵ (which provide detailed specifications against which implementations of the standard are to be conformant, compliant or consistent) through to community agreed recommendations/standards produced without formal standardisation processes. The latter are often very effective due to the speed at which they can be produced and adopted by the research community at large, even though they are not formalised.

The purpose of this section is to summarise the standards situation today with respect to digital repositories to support the widespread dissemination and use of information of all sorts across many platforms and disciplines across Europe.

We do not discuss here the taxonomy of standards and frameworks in which they are developed. We also recall a point made in Interim Report 1, that the boundary between standards and technologies is not always clear cut.

2.1 Why standards for repositories

Standards are a necessary component in the development of repositories because they can:

- Ensure interoperability of tools and linkage of data across repositories:
 - Enabling interoperability between repositories at semantic and syntactic levels (that is, so that repositories can communicate with each other consistently and with facility)
 - Enabling interoperability between repositories and users, both human and computer, at semantic and syntactic levels (where these users may be either information producers or consumers).
- Provide better exploitation of resources by different communities:
 - Help reduce effort by forestalling the need to re-work data, build plethora of interfaces and data conversions
 - Provide consistency and ease of use for users
 - Enhance reliability.

² As in Interim Report 1 we do not make a formal distinction between information and data; **information** is understood here to mean **all** (digital) information collected over the research life-cycle, from raw data to final publication, and including research administration information and metadata.

³ The USA’s National Information Standards Organisation. See: <http://www.niso.org/>

⁴ International Telecommunications Union (Standards). See: <http://www.itu.int/net/home/index.aspx>

⁵ International Standards Organisation. See <http://www.iso.org/iso/en/ISOOnline.frontpage>

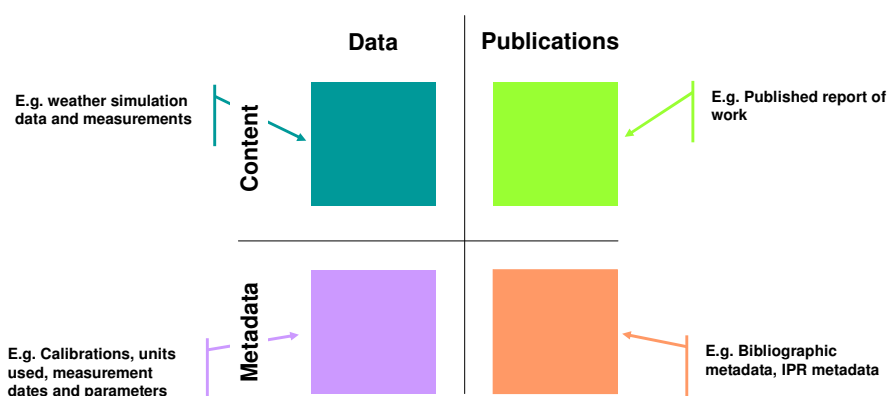


- Define, and ensure compliance to, agreed conditions of use of information in repositories
- Promote efficiency and cost savings in the research environment
- Facilitate migrating systems forward as technologies and needs change.

Achieving these capabilities brings benefits to knowledge, to learning and ultimately to the economy.

2.2 Standards and domains of data and knowledge

We noted in Interim Report 1 (Section 5.2, Technologies) the divide between worlds of data repositories and worlds of repositories of publication information which includes published reports, pre-prints, reprints, theses, patents. This divide is also seen in the standards arena too: there is less domain specificity in the publication world: a (digital) journal article on performance practice in, say, 19th century theatre has much in common with one on an engineering experiment in terms of data types (e.g. texts, images), structure (e.g. title, abstract, body, etc.), if not intellectual content. Data and metadata standards in the “data” world are, in general more diverse than that in the “publication” world due to the diversity between the different data domains.



In this report we call data collected during the research process and the subsequent publications derived from it **Content**, and note that both data and publications can be described by **Metadata** (see figure above). Each of the four components in this diagram may be standardised at both syntactic and semantic levels.

We note that the vast majority of **data** is not held in publicly accessible repositories. Rather, 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, if at all, once they have published papers. Standards that allow for researchers to provide secure access to their data within the context of a “Virtual Organisation” (VO) where local policies on access and usage of data can be enforced are thus needed. This idea of a VO is at the heart of Grid Computing and collaborative e-Research. The Open Grid Service Architecture standards that are being developed by bodies such as the Open Grid Forum (www.ogf.org) are outlining the framework that should eventually provide the standards, technologies and guidelines in their application for all stakeholders involved in the digital repository space.



We note too that some disciplines have put forward their own domain-specific standards. Examples of these might be the various flavours of XML-based mark-up languages used by different research communities. Thus within the life sciences domain, the bioinformatics community has defined Systems Biology Markup Language (SBML), a software-independent language for describing and exchanging models among different systems biology tools; Gene Expression Markup Language (GEML) for storing DNA and microarray data; Microarray and Gene Expression – Markup Language for managing microarray experiment results and supporting their future use through agreed annotations and capture of necessary metadata describing experiments. There are many other examples of such languages that have been put forward within the life sciences by different researchers, communities and standards bodies working in that space.

The same phenomenon is also occurring across many other disciplines from mathematics, geographical information, chemistry, the clinical and healthcare domains, etc., etc.

2.3 Standards relevant to repositories

In the same way as we did not discuss base, or underlying, technologies in Interim Report 1, we do not discuss what may be called base, or underlying, standards here. They include XML and the standards that in turn underpin that, database standards such as SQL, and network protocol standards as examples. Neither in this document do we attempt to describe all the multitude of domain-specific standards. Our focus is to identify instead the generic approaches that have been adopted to underpin best practice in data sharing. We focus in particular on the need standards for:

- Security including authentication and authorization controls
- Rights assertion and management
- Information description at syntactic and semantic levels:
 - Data
 - Metadata, annotations
 - Tools for expressing semantics, such as ontologies
- Object identification and name resolution
- Information and metadata harvesting and capture
- Repositories as managed stores
 - Organisation
 - Standards for long-term data storage, archiving; and preservation
- Search and retrieval
- Standards for distributed data Grid architectures.

We also note standardisation for Current Research Information Systems, pipeline tools, and data protection. In some areas we felt it helpful to provide more detailed and technical notes and these are provided in Appendix A2, as referred to in the review notes that follow in section 2.4 below.

These standards cover a broad area and often overlap with other standards efforts in areas unrelated to e-Science digital repositories, such as in the area of semantic web. Where necessary we provide examples of solutions that the community has produced.



2.4 The standards landscape

In this section we list and comment briefly on the major standards across the various areas identified above.

Security including authentication and authorization controls

Robust standards in the area of security, authentication and authorisation are vital for interoperability to take place in a climate of trust. It thus forms 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 necessary for acceptability and take-up.

In Appendix A1, section A.1.5, we have provide an extensive overview of security standards for e-Science digital repositories with references; in our First Interim Report, section 4.2 we listed relevant technologies in this area.

Relevant standards are listed in the following table with some relevant references (further references are provided in Appendix A1.6):

Specific standards for security	
■ DyVOSE (including delegation of authorisation)	■ Shibboleth ⁸
■ Globus GSI	■ SOAP ⁹
■ LDAP	■ VOMS (Virtual Organization Membership Service)
■ OMII Europe standards and OMII Security Portlets (including attribute acceptance and release policies)	■ Web Services:
■ XACML ⁶	○ Includes: WS-Policy, WS-Trust, WS-Privacy, WS-SecureConversation, WS-Federation, WS-Authorisation, WS-Agreement.
■ OpenSSL	■ X509
■ PERMIS	■ X812
■ SAML ⁷	

Rights assertion and management

Intellectual Property Rights (IPR) technologies implement or display licensing policies and business models for digital resources distribution and usage. These Digital Rights Management systems (DRMs) can gather together several building blocks:

⁶ OASIS eXtensible Access Control Markup Language. See: <http://www.oasis-open.org/committees/download.php/2406/oasis-xacml-1.0.pdf>

⁷ Security Assertion Markup Language. See: http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=security

⁸ See: <http://shibboleth.internet2.edu/>

⁹ Simple Object Access Protocol, and lately also Service Oriented Architecture Protocol. See <http://www.w3.org/TR/soap/>



- Technical protection measures controlling access to, and usage of, digital data and other resources through cryptography and scrambling technologies
- Technical information stating licensing elements to be implemented by the protection mechanism or displayed to the end-user to inform them which actions they may perform on the data/resources and under which conditions
- Fingerprinting and watermarking, based on steganography and similar technologies, constitute an intermediate level between protection and providing information. Hidden information embedded in the data makes it possible to identify, track, and compare data.

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

- DCMI (Dublin Core Metadata Initiative, ISO 15836) provides elements to describe resource's content, rights and instantiation
- FRBR (Functional Requirements for Bibliographic Records) is a conceptual model describing the creation process for bibliographic structures
- METS (Metadata Encoding & Transmission Standard) is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library.
- PREMIS, which provides a data dictionary and XML schemes for long-term preservation metadata including action restrictions, time-stamped (based on the ISO 8601 standard for dates and time).
- OAI-PMH (Open Archives Initiative's Protocol for Metadata Harvesting) provides a mechanism for repository interoperability through structured metadata exchange between data and service provider.

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

- For geospatial information: Geospatial DRM¹⁰
- Open Digital Rights Language (ODRL)¹¹, from The Open Digital Rights Language (ODRL) Initiative
- eXtensible rights Mark-up Language (XrML)¹² developed by ContentGuard Inc.

For more on IPR see the section on legal issues in section 3 below.

Information description

Information needs to be communicated to and from repositories or between repositories, needs to be stored in some format, and when used, needs to be understood. Standards are essential to enable all of these.

Interoperability requires standards for both the syntax and semantics for both the content being communicated, and standards for descriptive information about it ("metadata"). Underlying these are standards for specifying both syntax and semantics independently of specific information types. Those for expressing syntax are well established (now mainly based on XML); we make some notes below on standards for specifying semantics independently of specific domains of knowledge.

¹⁰ See: <http://www.opengeospatial.org/>

¹¹ Open Digital Rights Language Initiative. See: <http://odrl.net/>

¹² See: <http://www.xrml.org/about.asp>



Content

The divide that separates the domains of data, and document information is quite sharply drawn in this area, and we reflect that in our notes on data, document information and metadata below.

When it is stored data may conform to a different standard from that used to communicate it in some exchange; we are more concerned with the latter – the “external” view of content. An example is a bibliographic database: when stored in some library management system it may conform to some relational database management standard; when communicated it may be expressed as, say, MARC21. However it is important to note that database schema need to match when content is exchanged at a database level – failure to do this can present considerable difficulties¹³.

Data is represented in many diverse formats depending on the domain that it applies to. It is almost an impossible task to list the standards for these, as they are so diverse reflecting the diverse needs of different communities and sciences. We note that XML is becoming more common way to express and communicate data content – just some examples: ChemML (chemistry), MathML (mathematics), GML (Geo-spatial content). For common data types there are of course well established standards such as the various MIME¹⁴ types, standards for images, text content, audio, etc, etc. Many of the standards in this area express both syntax and, to some extent, semantics.

Publications present a much more restricted class, being mainly variations on texts with some embedded objects, such as images. Significant standards for these at the syntactic level are the “base standards” of XML, PDF (and PDF/A, the archival version of PDF).

Metadata

Data: Clearly there are some metadata standards which are specific to metadata describing the multitude types discussed above. Again the field is vast and diverse – thus for microarray experiments (in genomic research) there is MIAME¹⁵. Another is the JCamp-DX standard¹⁶ that is used for exchanging infrared spectral data. We note that there is a blurring of the distinction between data and metadata for data – thus for example the MIAME standard contains within it much information which could be regarded as metadata alongside the raw data itself.

Publications: For bibliographic document metadata there are very many standards, of which the most ubiquitous is Dublin Core. The following table lists some of these standards, noting that each has its own specialist area of use, such as for libraries or archives:

¹³ See for example Large-scale sharing in the life sciences, Lord and Macdonald, 2005, p50. Available at http://www.nesc.ac.uk/technical_papers/UKeS-2006-02.pdf

¹⁴ Multipurpose Internet Mail Extensions. See <http://www.iana.org/assignments/media-types/>

¹⁵ Minimum Information About an Microarray Experiment. See: <http://www.mged.org/Workgroups/MIAME/miame.html>

¹⁶ From The International Union of Pure and Applied Chemistry (IUPAC). See: <http://www.jcamp-dx.org/>



Specific bibliographic standards and technologies	
<ul style="list-style-type: none"> ■ Dublin Core ■ MARC (including MARC21, MARCXML) ■ Electronic Archival Description (EAD) ■ General International Standard Archival Description (ISAD(G)) (and related archival description standards) ■ Metadata Authority Description Schema (MADS) 	<ul style="list-style-type: none"> ■ Metadata Encoding and Transmission Standard (METS) ■ Metadata Object Description Schema (MODS) ■ MPEG-21 (and Digital Item Declaration Language(DIDL)) ■ Preservation Metadata Implementation Strategies. (PREMIS)

Semantics

The Resource Description Framework (RDF)¹⁷ 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 OWL Web Ontology Language¹⁸ is designed for use by applications that need to process the content of information instead of just presenting information to humans. OWL facilitates greater machine interpretability of Web content than that supported by XML and RDF (and RDF Schema (RDF-S)) by providing additional vocabulary along with formal semantics. OWL has three increasingly-expressive sublanguages: OWL Lite, OWL DL, and OWL Full.

DAML¹⁹ 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.

Another approach is CIDOC Conceptual Reference Model (CIDOC CRM)²⁰, which has been an ISO standard since 2006 (ISO 21127:2006). CIDOC provides a common and extensible semantic framework that any cultural heritage information can be mapped to

Object identification and name resolution

The permanent digital object identifier (PDOI) is the Digital Object Identifier (DOI) from the International DOI Foundation (IDF)²¹. This is a system built upon web-addressing techniques and the Handle technology (see the First Interim Report, section 4.3). 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 is much used for citing bibliographic and published content. It is not unchallenged: rival proposals are the Archival

¹⁷ A family of World Wide Web Consortium (W3C) specifications. See: <http://www.w3.org/>

¹⁸ See: <http://www.w3.org/TR/owl-features/>

¹⁹ Darpa Agent Mark-up language: See: <http://www.daml.org/>

²⁰ See <http://cidoc.ics.forth.gr/>

²¹ See: <http://www.doi.org/>



Resource Key (ARK) and Persistent Uniform Resource Locator (PURL). ARKs are proposed by the California Digital Library (CDL)²², and PURLs by OCLC²³.

Appendix A1.3 discusses naming standards further in the context of Grid technologies.

Information/metadata harvesting/capture.

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) standard was described in the First Interim Report. An extension of this work is the Open Archives Initiative Object Reuse and Exchange²⁴ (OAI-ORE). ORE has developed specifications that allow distributed repositories to exchange information about their constituent digital objects.

Repository management

We note here standards which facilitate directly the management and sustainability of repositories as managed stores.

Organisation and sustainability

The Open Archival Information Systems (OAIS) reference model (ISO 14721:2003) presents a design and management framework for repositories which are used for long term retention of digital information of all kinds. This framework has reached a high level of acceptance (though penetration of its concepts is rather slow into the established archival community).

The DRAMBORA toolkit²⁵ for assessment of repositories was recently introduced and is still undergoing evaluation. It was developed by the Digital Curation Centre and the Digital Preservation Europe project.

Standards for preservation

Technologies and tools relevant to the conservation and preservation of digital materials were discussed in the First Interim Report (Section 4.2). In addition to the tools mentioned there, various information standards form a key part of the armoury for preservation – in particular PREMIS has been developed by an international panel²⁶ to address the needs of metadata to assist conservation and preservation.

Use of well established, non-proprietary standards for content and metadata provides a hedge against the inaccessibility of digital information as technologies change. However, robust solutions to the problem are still elusive.

Search and retrieval

Technologies and associated standards for these were covered in the First Interim Report, Section 4.2.

Standards for distributed data Grid architectures

Data Grids are arguably the most difficult Grids to establish and manage. This is due to a variety of reasons. Some of these include: the complexity of the data itself which can often be very domain specific and require expert interpretation; the evolutionary nature of research and

²² See: <http://www.cdlib.org/inside/diglib/ark/>

²³ See: <http://purl.oclc.org/>

²⁴ See: <http://www.openarchives.org/ore/>

²⁵ See: <http://www.repositoryaudit.eu/>

²⁶ See: <http://www.oclc.org/research/projects/pmwg/default.htm>



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, and perhaps above all, the amount of data that is being generated across all research disciplines. In this context the establishment of a data infrastructure for e-Science digital repositories is especially challenging both in scope and complexity.

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 Appendix A1 for further details.

2.5 Frameworks for standards

Standards are developed in many forums, 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. Of particular relevance in this regard are the standards being developed for the Web Services, Web2, and Service Orientated Architectures. Industry can also develop frameworks or architectures – higher level standards – which serve to integrate activities in a particular domain. A good example in this respect is CDISC (Clinical Data Interchange Standard Consortium) from the pharmaceuticals sector, which has developed and issued a wide range of standards for the drug development and testing process²⁷.

Standards rarely start from zero – they tend to be defined in hierarchies, one standard being based upon others at a lower level.

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

- Community participation in setting standards is essential for community buy-in
- Development of standards needs to be controlled to avoid overly-complex and large standards
- They need to be flexible to accommodate the current rapid rate of change of technologies
- Standards setting takes valuable time from experts in the field – due allowance needs to be made to allow individuals of the right calibre to attend to the task for the sake of progress.

Standards are generally developed to serve two purposes: to achieve interoperability and/or to provide a yardstick for measuring compliance. In the report “Large-scale data sharing in the life sciences”²⁸ the various types and levels of standards are discussed. Though focussing on the life-sciences, the conclusions of that study are widely applicable over a much wider domain.

²⁷ See: <http://www.cdisc.org/>

²⁸ Lord, P., Macdonald, A., Sinnott, R. et al., Large Scale Data Sharing in the Life Sciences, pp A111 & p33. Available at: http://www.nesc.ac.uk/technical_papers/UKeS-2006-02.pdf



Gaps in the standards landscape

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

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



Section 3: Legal implications of open access

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. It has not focused on the specific issues of open access, copyright, e-prints and publications extensively discussed in many other reports and articles²⁹.

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 GBIF³⁰, 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),

²⁹ Links and bibliography are set out on the www.e-scidr.eu web site.

³⁰ 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”



interpretations of existing data in textual analysis; annotations of data (tags, keywords, unique identifiers, comments, or 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 Commons³¹ 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.

³¹ www.sciencecommons.org



Section 4: Results of the Public Consultation

4.1 Methods

Over the period of 16th July to 31st August a public consultation was conducted to gather opinions which would inform the development of policy options. The questionnaire is provided in Addendum B to this section. The consultation was conducted using the European Commission's web site using their Interactive Policy Making (IPM) system (See http://ec.europa.eu/yourvoice/ipm/index_en.htm).

The questionnaire was publicised through e-mail notes to individuals on the EC's databases, and on the databases of the study team members. Postings were also made to relevant e-mail lists. Because of the way the e-mail list notifications are propagated it is impossible to know the exact size of the population invited to respond; an educated guess would put it at some thousands.

4.2 Questionnaire

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.

A combination of and free text and multiple choice questions were used to collect information. The questionnaire is reproduced here in Addendum B to this section.

4.3 Analysis of the responses

Given the methods employed the respondents do not form a randomised sample from a well defined population. In general terms we can assume that those responding to the questionnaire will have been those with an interest in repositories and related areas. This supposition is supported by the quality of the knowledgeable free text contributions made, and the by the responses received from senior individuals working in this field, both in Europe and beyond.

Analysis was conducted by creating simple statistics and graphics in Excel and the Minitab 15 statistical analysis program.

The final download of the database of results was made when the consultation period expired on 31 August 2007, by when 426 people had responded. The full data set is lodged with the commission.



The respondents

The 426 respondents were drawn from a world-wide population, but the bulk of the responses (335, 78.6%) were from Europe (including non-EU countries). Figure A1 (in Addendum A to this section) shows the distribution of responses by country (Question 1.2). In terms of raw counts the UK provided the most respondents, but adjusting for population size per country (Figure A2) shows that Switzerland was the most enthusiastic nation of responders, followed by Belgium and Greece, and then Estonia, the United Kingdom and Denmark. To some extent this distribution will reflect interest and penetration of repositories in various countries; another factor which may have affected the response rates across different might be levels of understanding of English, the sole language of the questionnaire.

Though Switzerland hosts CERN, a major repository centre, only two of the 16 Swiss respondents gave CERN as their affiliation. The numbers from some countries were small and not too much reliance can be placed on the detail of these rankings.

There was a good deal of interest from outside Europe, providing 91 responses (see Figures A1 and A2), notably from the USA (24) and Brazil (17). The number of responses from the Europe, split by EU members, EFTA members and others, and the rest of the world are shown in Figure 1.

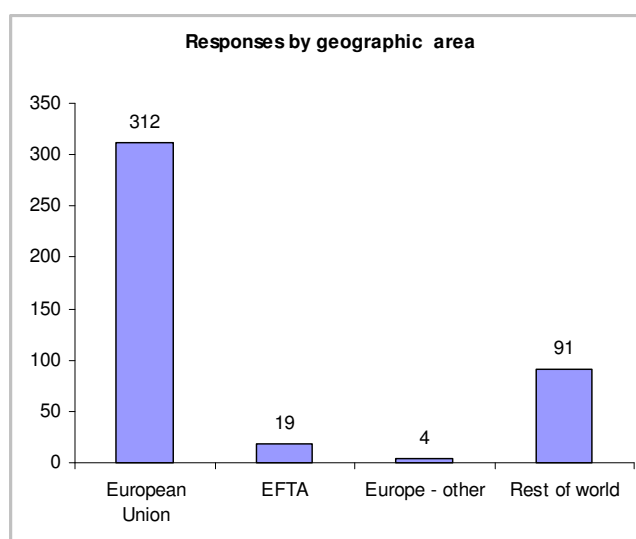


Figure 1. Number of responses over Europe and the rest of the world

Some 73% of the respondents were male and 27% were female (Question 1.8). Males and females did not appear to give significantly different replies to the questions. The age profile of respondents was flat over the range 25 to 65; only 20 responses were obtained from people outside this range (Question 1.7).

70% responded as individuals and 30% on behalf of their institution (Question 1.9); there were no significant differences between the replies given by these two groups. In what follows the responses of these two groups are aggregated.

From question 1.4.1, the organisational affiliation of most respondents was the academic sector (63%). (See Figure 2).



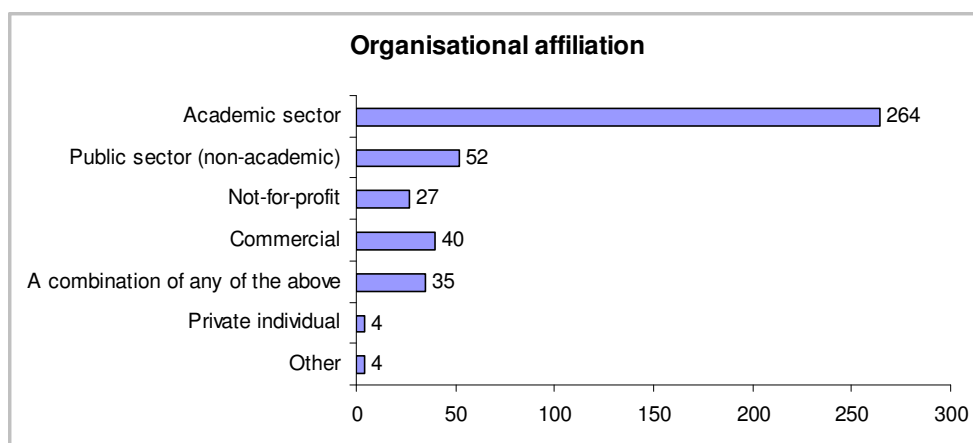


Figure 2: Respondents' organisational settings

Regarding the disciplines of respondents (Question 1.5), library and information science topped the list (just over one quarter), followed by astronomy and mathematics and physics (See Figure 3). There is a strong (data) repository movement within the astronomy community (also a much internationalised science), but it is a little surprising it ranked so high. We suspect that most of the mathematics and computer science respondents were computer scientists rather than mathematicians, but cannot confirm this. 67 people marked “Other” as their discipline, and the text responses to a question asking for clarification included telecommunications (8 people), 5 historians, 4 educationalists and one enologist.

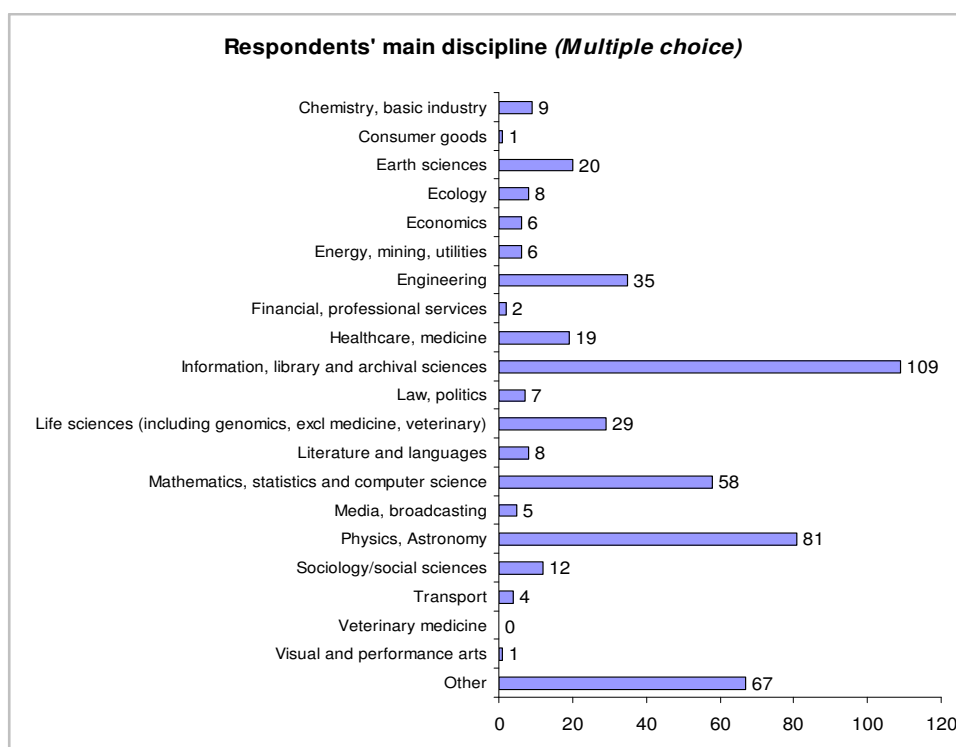


Figure 3: Disciplines of respondents

Whatever their discipline, the questionnaire respondents' roles (Question 1.6) showed there was a preponderance of researchers (Figure 4) – twice as many as the next highest response



(directors of institutions/companies). Librarians, administrators, directors, principal investigators each provided ca. 60 responses each.

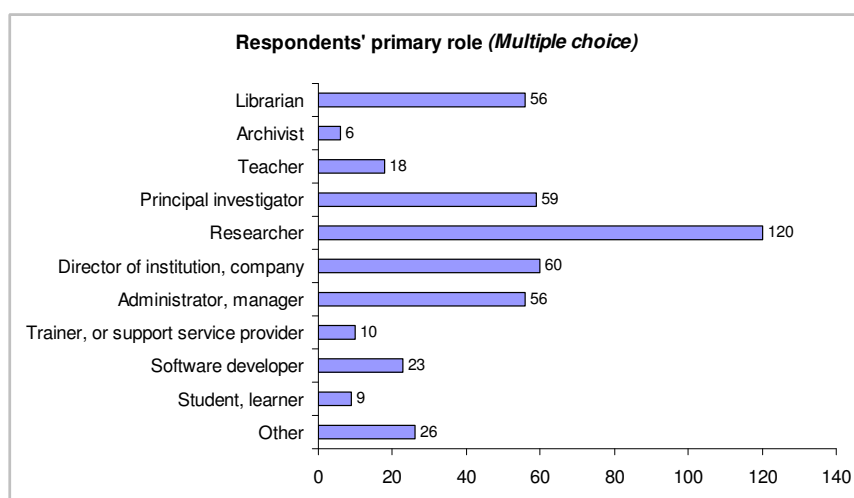


Figure 4: Respondents' reported primary roles³²

Use of repositories

About two thirds of the respondents (63%) reported that they had received no training for using repositories (Question 2.5), and 33% said they had (4% did not answer).

The respondents were heavy users of repositories (Question 2.1), most of them using them on either a daily or weekly basis (78%); less than one percent never used them (Figure 5).

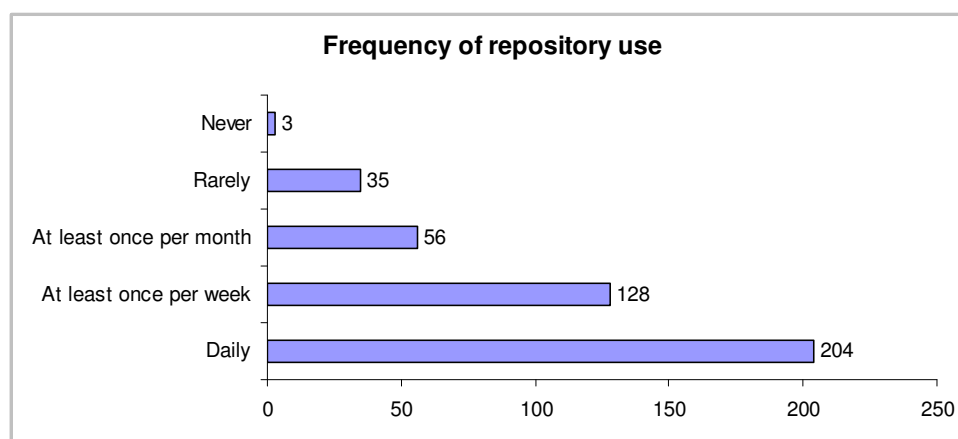


Figure 5: Frequency of use by respondents

Payment for use at the point of access (Question 2.3) is **not** the norm (75% never pay, or only sometimes). Figure 6.

³² Note that multiple answers allowed to this question.



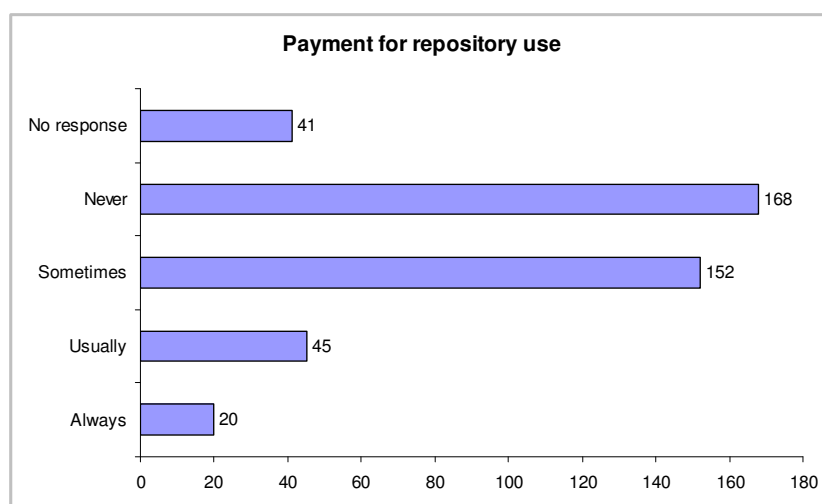


Figure 6: Payment for use, either by the respondent or their institution

Regarding respondents' roles vis-à-vis repositories and repository types they used, the responses are summarised in the following table from Question 2.2, also shown in graphical form as Figure A3. The respondents' roles vis-à-vis repositories showed that they were, in the main, users of information from repositories. Note that multiple answers were allowed to this question. To a much lesser degree the respondents were depositors into repositories (only half as many responses as for users). The role of repository managers was also quite well represented.

Repository role	Repository type							Row totals
	Digital library	Community repository	Discipline-related repository	Institutional repository	e-Learning repository	Commercial repository	Other	
Data user	294	196	237	228	120	95	18	1188
Depositor	82	91	125	144	56	23	3	524
Repository manager	37	48	62	79	26	11	8	271
Policy maker	49	41	57	92	29	6	6	280
Support role	34	37	58	59	30	9	1	228
Trainer	25	24	37	43	37	4	1	171
Funder	23	24	22	34	17	6	1	127
Other	9	10	13	13	11	10	6	72
Column totals:	553	471	611	692	326	164	44	

Table 1: Roles adopted against various repository types

The most frequently used repository types from which to get information were digital libraries, then 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

Table 2: Ratio those user a user role to deposit role 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.

Question 2.2.1 asked respondents to specify other roles they assumed: responses included curation, digital preservation, and research on open access policies. Of the 96 answers, 15 responses indicated work related to the design, creation or definition of repositories, and 5 reported an advocacy role for the repository they were associated with. Similarly question 2.2.2 explored other types of repository used, eliciting 68 replies: interesting responses here were national repositories (3 people), software repositories (3), personal repositories (one person), patent repositories (2) and data grids (1). One respondent asked the question “*what is a repository – is the internet a repository?*”

Question 3.1 explored the types of information which were both deposited and used by this group of users. The results are shown in Figure 7, with publication-type material (pre- and post-prints, publications, texts) and data types (raw and processed, and images) scoring high. Use and deposit rates for the various data types are very highly correlated.

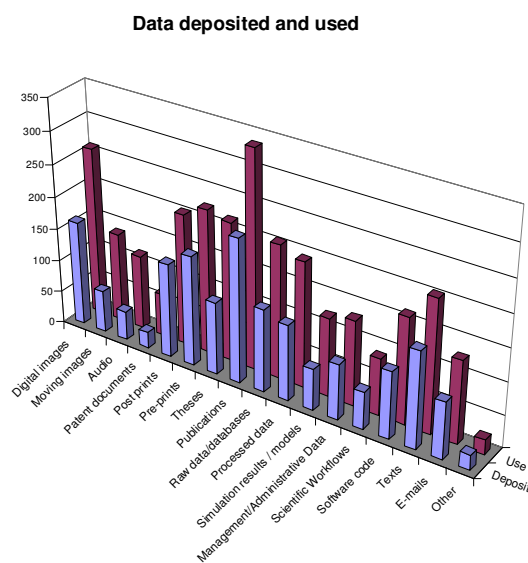


Figure 7: Information types deposited and used (Multiple choice)



Use (finding/extracting information) exceeded deposit overall by a factor of 1.6. Looking at individual data types we see the following pattern for the ratio between use and deposit:

Information type	Ratio use to deposit
Digital images	1.6
Moving images	2.1
Audio	2.6
Patent documents	2.6
Post prints	1.4
Pre-prints	1.3
Theses	1.9
Publications	1.5
Raw data/databases	1.6
Processed data	1.6
Simulation results / models	1.9
Management/Administrative Data	1.5
Scientific Workflows	1.5
Software code	1.6
Texts	1.4
E-mails	1.5
Other	1.0

Patents, audio and moving images appeared to have the highest use relative to deposit, followed by theses and simulation results. Presumably people find repositories particularly useful for these data types. Pre-prints, according to these data, have low re-use relative to deposit.

A supplementary question (3.1.1) inquired about other data types. The 32 answers included (PowerPoint) presentations (4 respondents), learning objects (3), webcrawls (1) and CAD/CAAD³³ (2).

Inhibitors to use and enablers

When asked about difficulties using information from repositories (Question 4.1) the chief obstacles mentioned 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 8.

³³ Computer-aided Design and Computer-aided Architectural Design.



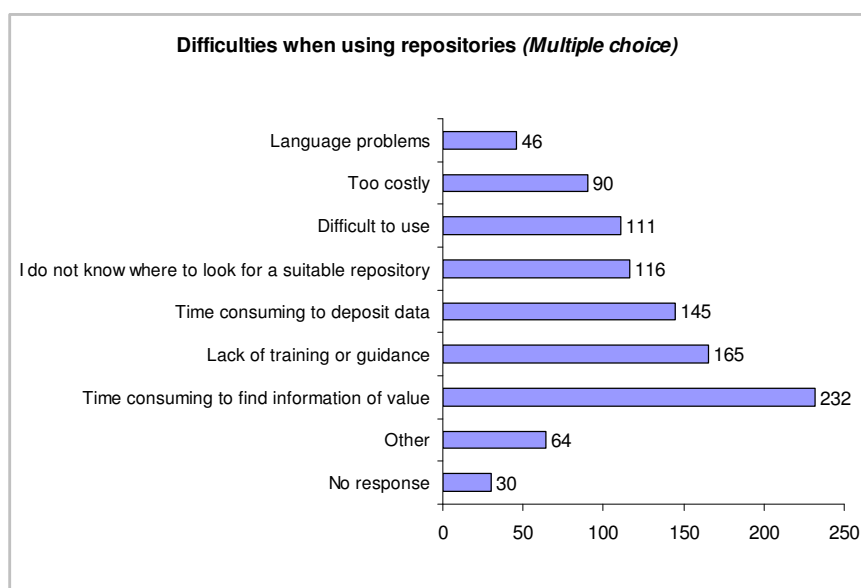


Figure 8: Types of difficulty encountered using repositories

The average number of difficulties mentioned per respondent was 2.3, which seems to be rather higher than one would wish; clearly there are inhibitors to use, reinforced by noting that all but 30 people answered this optional question.

Other difficulties which were cited in response to question 4.1.1 included copyright restrictions and data ownership issues and clear permissions. Lack of incentives to use repositories was also mentioned a number of times. Few technical difficulties were raised here, though one respondent cited unreliable host machines. A view was offered that *“The ease of use of a repository is directly related to the financial input into its creation and maintenance”*.

When asked for indicators of trust in repositories and their contents (Question 4.2) peer reviewed contents was the most cited option (by 73% of the respondents). This was followed by availability of clear policies (63%) and demonstrated awareness of the needs of users (55%). See figure 9. Payment was not seen as endowing any trust, and there was little support for help facilities (13%) or registration for use (21%). On average 3.4 suggestions were made by each respondent.



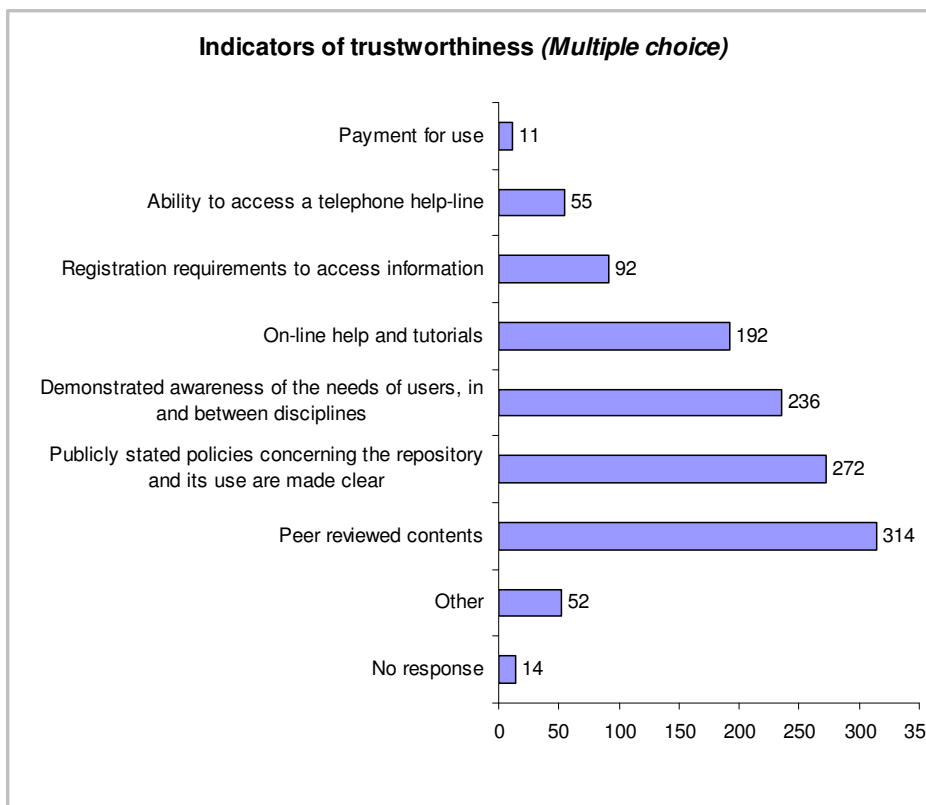


Figure 9: Indicators of trustworthiness

Other indicators of trust (Question 4.2.1) were solicited; noteworthy, among a great variety of suggestions were the reputation of the host institution (many responses), indicators of long-term sustainability, recommendation by peers, and “publicly accessible validations of collection properties (completeness, authority, integrity, authenticity)”.

Regarding whether certification of repositories (Question 2.4) would encourage use a majority of respondents said “yes”(54%) and a minority said it would not (18%), the others being undecided or gave no response. See Figure 10.

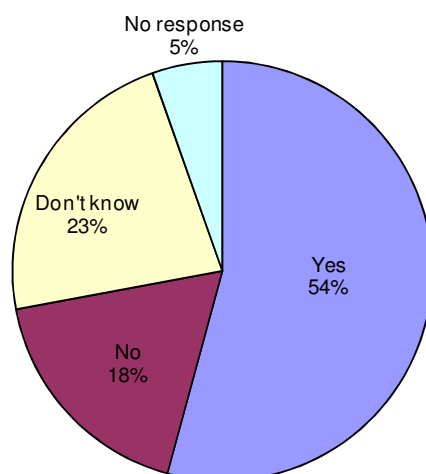


Figure 10: Would certification of repositories encourage use?



When asked what would make repositories easier to use (Question 8.1), making it easier to find information came out strongly, 65% citing the need for registries to find repositories and 76% of respondents citing the need for better searching tools within and across repositories. Also cited were tools to assign metadata automatically (69%). A substantial minority also chose faster networks as an enabler (44%). See Figure 11.

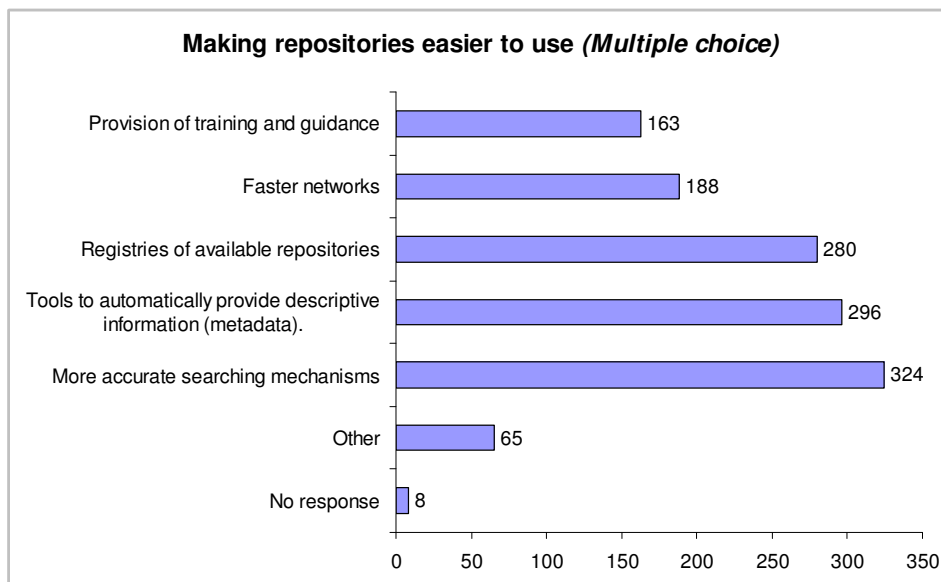


Figure 11: What would make repositories easier to use?

Responses to “Other” from question 8.1.1 included the adoption of standards, clear policies and codes of practice associated with the repository, faster, better and more intuitive user interfaces. Two interesting responses were expressed as “*be a non-redundant part of [the] research workflow*” and “*embedding repository deposit within institutional research workflows*”. A number of respondents mentioned integration into other services, such as Virtual Learning Environments (VLEs), Google, social networking systems. Others mentioned linking of repositories, cross-repository access/searching and federation of repositories. There were 71 replies to this question.

Question 5.1 asked for free text responses to identify inhibitors to **deposit** of materials into repositories. 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 were mentioned 35 times (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 some (12%)
- Costs or payment policies were mentioned 12 times
- Absence of a suitable repository, or awareness of one was also mentioned 12 times.



Other inhibitors which were mentioned included privacy concerns (7), concerns over ensuring longevity for data (6), lack of motivation (and lack of enforcement), unclear repository policies, fear of a lack of standards for interoperability, and fears about plagiarism (“being scooped”) or lack of attribution for the information. Interestingly few mentioned the burden of providing metadata or lack of peer review processes.

Question 5.2 asked the same question regarding **use** of information from repositories – 207 people answered this (49%). Again these were varied:

- Most concern was expressed in regard to payment and cost, mentioned by 48 respondents answering this question (23% of those answering). High costs were mentioned frequently
- 29 people referred to difficulties finding and retrieving information, or the complexity of finding information. Poor user interfaces were also mentioned in this regard
- 25 people referred to lack of trust in the information or repository, and in a few cases linked this to insecure hosting sites or unreliable hosting
- Again, 25 people mentioned quality issues regarding data – could they rely upon it? This was linked in some cases to lack of indicators of quality. Stability and lack of comprehensiveness were also mentioned in this context
- Lack of adequate metadata (14 responses)
- Copyright issues, legal restrictions and disadvantageous licensing terms were also cited as a concern (12 responses).

Other inhibitors to use mentioned were concerns over authenticity and provenance, difficulties finding information, problems due to unsuitable file formats. Only two people mentioned language as an issue.

Question 5.3 asked for specific frustrations encountered when using repositories. The 140 (33%) responses varied widely – one respondent listed frustrations succinctly as:

“Registrations, logins, payments, technological barriers, bad interfaces, broken links.”

There were many complaints about poor user interfaces, difficulties searching and inadequate search engines. A typical comment was:

“non-intuitive use of interfaces; insufficient on-line help; interfaces not sufficiently user-oriented (that's actually a key issue): they require me to learn their language before I can make use of the repository.”

Poor reliability, servers not available, poor implementation of embargo periods, payment difficulties (and expense – *“The frustration comes when you have to pay 30€ just to have a look at a paper”* and *“[you] get free access [but] when one reaches important information - being asked for money”*), IPR restrictions mean the information in question cannot be used. A few people mentioned slow networks (one of these respondents was in Nigeria).

Metadata-only records were also a source of frustration.

Just over a third of respondents were required to deposit the information they created in a repository of some kind (Question 6.1), shown in Figure 12.



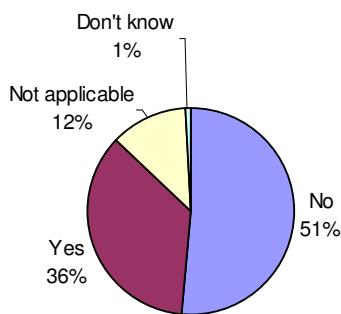


Figure 12: Are you required to deposit data?

28% of all respondents had to use a specified repository, 7% had a choice (Question 6.2).

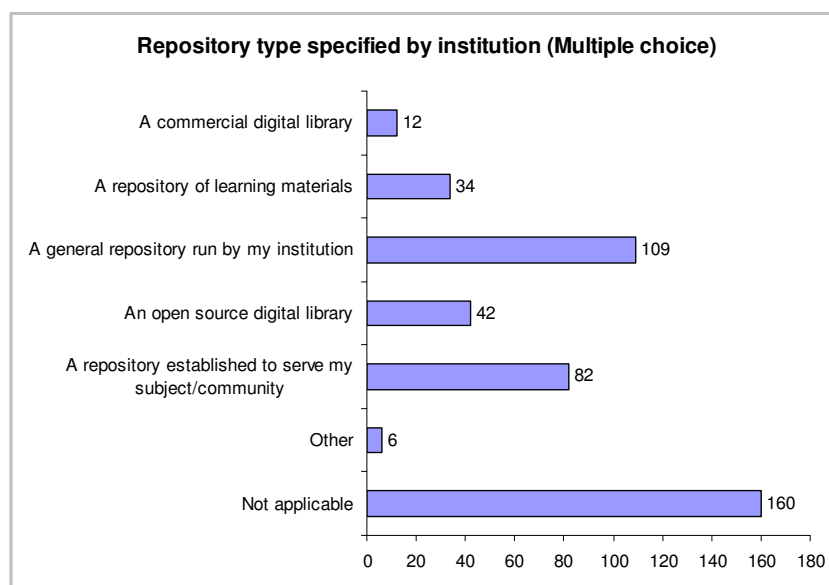


Figure 13: Preferred repository type specified by institutions

A supplementary question (Question 6.1.2) asked what sort of repository was specified by their institution or funder, if any. The survey software did not provide provisions to restrict answers to this question to those who answered “Yes” to question 6.1, and so it was answered by some other respondents. Figure 13 shows institutional repositories were most frequently specified, then a subject/community repositories.

In contrast question 7.1 asked for the preferences of the users, and the profile is shown in Figure 14. This shows a preference for a community/subject repository, followed, in order, by open source repositories and institutional repositories. On average respondents made two choices when answering this question.



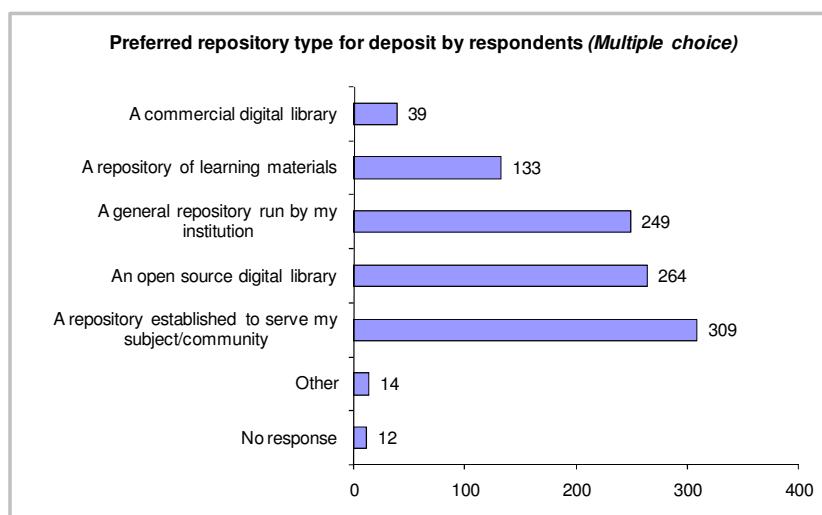


Figure 14: Preferences of respondents for depositing their data

Revealing comments were made as to why these preferences were expressed (Question 7.1.1). From the 144 replies, the reasons given were in summary:

■ For commercial digital libraries:

- Good management
- Clear policies
- Credibility
- Have financial resources to develop the repository well.

On the other hand people noted a fear of losing access to their materials after deposit, and through pricing adding to the digital divide. Perhaps it is significant that only one person indicated this type of repository preference and no other.

■ For institutional repositories:

- They indicated trust (by being attached to one's own institution)
- There is potential for direct interaction with the repository management
- They provide a resource where there is (as yet) no suitable community/discipline repository available for some disciplines
- Enhanced prospects of sustainability by being linked to a stable institution. (In this context one response drew attention to the UK's AHRC recently withdrawing support for the Arts and Humanities Data Service, a community repository resource for the arts and humanities in the UK).

■ For open source digital libraries:

- They are adapted to the specific needs of the community or discipline
- Respondents noted they had greater identification with their discipline than with their institution
- Availability of peer review mechanisms and scientific validation
- It is easier to locate information (the repository is known to the community)
- Driven by user needs
- Contributes to a higher research impact



- Respondents in medical research noted that confidentiality of personal information was more likely to be respected.
- For open access digital repositories:
 - Much research is publicly funded, and that it was appropriate to make the results publicly available at no further cost
 - A feeling expressed that one should not make a profit out of research information, particularly where it is publicly funded.

These sentiments were expressed frequently.

Access to help to deposit materials (Question 6.2) was reported available to some 46% of the respondents (Figure 15).

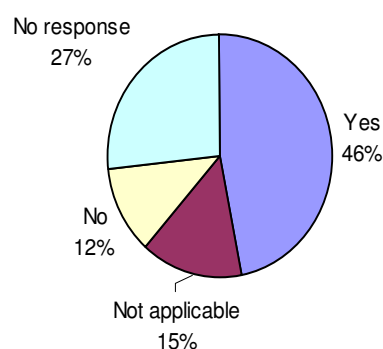


Figure 15: Access to help to deposit materials

An interesting comment on help was the following in answer to question 6.2.1:

“The requirements will be repository and discipline specific. The model is usually self deposit with and added QA process and help service. There is an issue about lack of funding and support for capacity building.”

To further explore views on enablers for repository use Questions 9.1 and 9.1.1 together asked where investments in repositories would best be made to enable science. These were free text questions, and 238 and 67 responded respectively answered these two questions. 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 out of the public purse should in principle be available free without further charges (see also responses to question 7.1.1). This theme recurred in the various free text 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 suggestions which are interesting, but which were mentioned less commonly were:

- Further development of data grids
- Well-established rights management requirements
- Provide help to specific communities
- Establish a “European Label” for repositories
- Establish a “Quality Stamp” mechanism for repositories (and data?)
- Invest in training (including *early* training in data management – perhaps in schools)
- Establish a database for EU-funded e-science outputs
- Reduce the burden of the European Directive 2001/20/EC, particularly as a as regards its effects on clinical research
- Distributed models (such as Lockss – Lots of Copies Keeps Stuff Safe³⁴) are more robust than isolated solutions
- IP rights tend to national, whereas science is global – the resulting tensions needs to be resolved.

Lastly, the view was expressed that repositories were essential for the establishment of the European Research Area.

Preferences and policies

Two questions (10.1, 10.2) asked respondents attitudes to establishing national and international (EU) repositories - these received considerable backing, 79% in the case of international repositories. See Figure 16.

³⁴ See: <http://www.lockss.org/lockss/Home> at the Stanford University Libraries



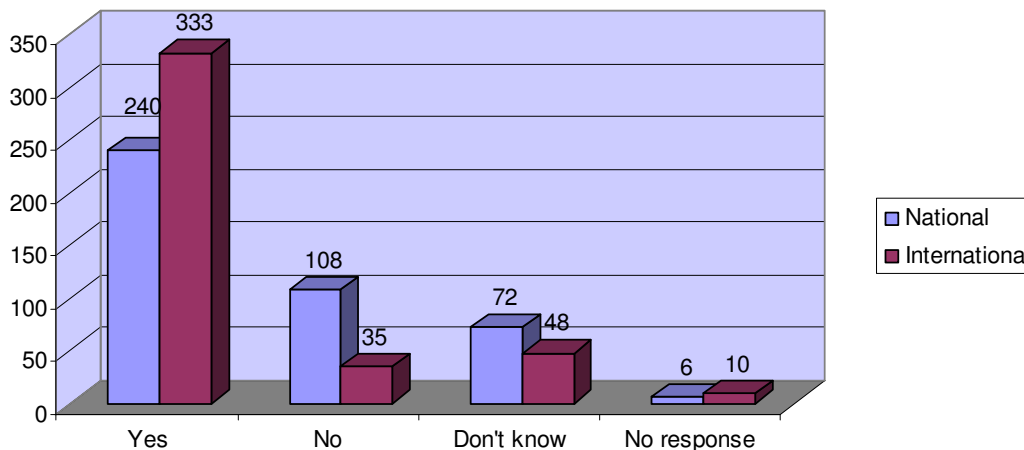
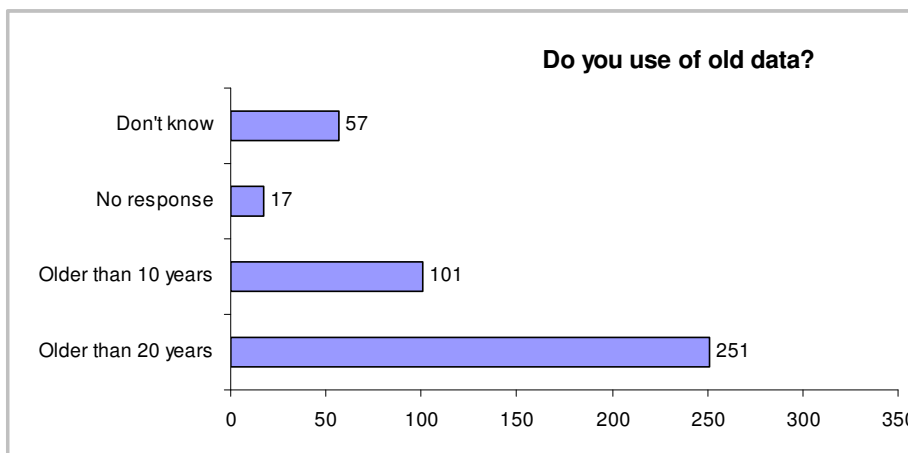


Figure 16: – Attitudes to establishing national and international repositories

Much comment was elicited in free text form to both these propositions (190 and 166 replies to Questions 10.1.1 and 10.2.1 respectively). There was support for both propositions, but many people made the point that science is not an activity constrained by national boundaries, and that setting up national (or indeed supranational) monolithic structures was not appropriate. Rather, any such structures should be of a federated or otherwise linked nature (“meta-repositories” to quote one phrase used).

Question 10.3 asked if older data was consulted – some 59% said yes to information over 20 years old. Interestingly, most people (73%) thought that the data they were producing would have a life span of over 10 years, the period beyond which the effect of digital data’s vulnerability to obsolescence starts to be felt.



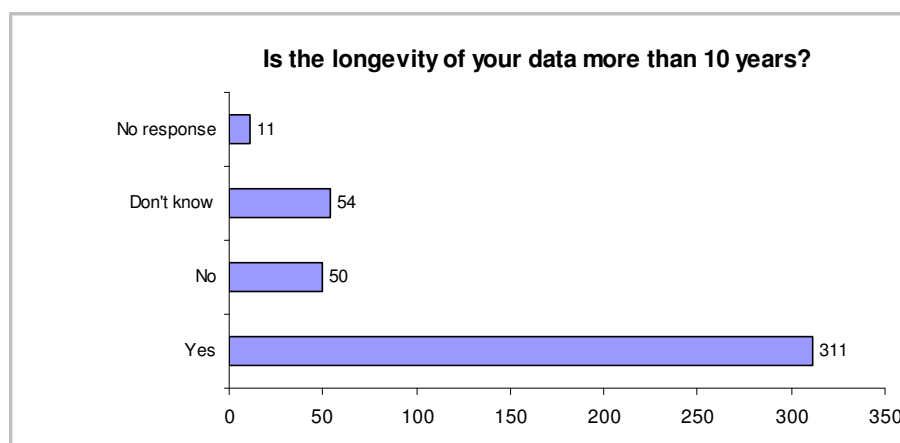


Figure 17: Use of old data and expected lifespan of data being produced

The number of responses given to questions 10.3 and 10.4 are shown in Figure 17 above.

Some 93 comments were made on the longevity issue (Question 10.5) – most of these supported the need for sustainability of information and its long-term value, but noted the difficulties it raised: technical, procedural, funding (including its adequacy) and organisational. The view was expressed that many copies were means of assisting longevity (cf the note about the Lockss model above).

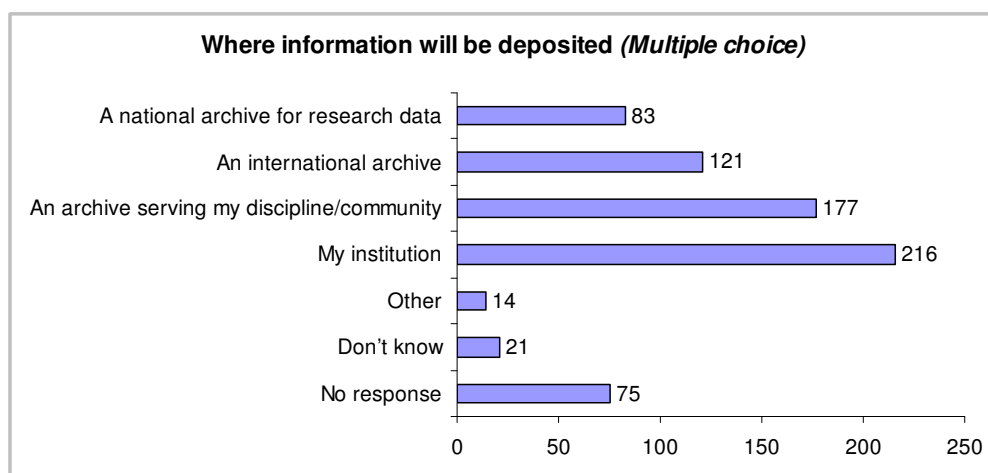


Figure 18: Where data produced is expected to be kept

Respondents expectations of where their data will reside after their use is shown in figure 18 (compare with figures 13 and 14 above).

Vision for repositories

The last group of questions asked for views on a vision for the future for repositories (Question 11.1), and any additional comments (Question 11.2) in free text form. 228 and 58 respondents respectively provided replies.

Replies varied widely as one would expect. We conclude this summary of results with a selection of some of the statements offered (NB Some of these are copy edited, marked []).



The question to which the comment formed a response is appended to each comment, and we have grouped them under broad topic headings:

Policy directions

“I believe that the EU should be working hard to make sure that research paid by governments around EU is also available for the public. I also believe that the EU could help making initiatives on national levels interoperable and supporting metadata exchange formats and international research infrastructure projects. Every European country [wants] to be the best globalized research based economy in the world ... We should instead work [to make] science open for the benefit of all.” (Qu. 11.1)

“All repository dreaming will remain dreaming unless proper mandatory policies are put in place to require deposit of content. Once that happens, all the rest - the services, the added value, the innovations -will follow naturally. Without the policies, money spent on repositories is money down the drain.” (Qu. 11.2)

“There are many technical problems, but I feel that the hardest problem is [a] policy one. It is necessary to make things beneficial for all parties involved: data producers, mediators, and consumers.” [copy edited from original] (Qu. 11.1)

“Public release and deposition of data is often regarded as a useless chore by data producers. Repositories have to work on making the deposition easier, but that does not solve the problem [completely], as data producers often have no interest in releasing the data, [so as] to keep a real or perceived advantage during the project. At the end of the project the data is then usually lost. A strong data release policy for EU projects is needed, with adequate provision for data management beyond the project lifetime, if any useful data is [to be] expected. Repositories should be freely accessible: attempted commercialisation of academic resources often leads to a lot of effort to collect small amounts of licence fees, and to a duplication of data generation efforts.” (Qu. 11.2)

“Digital repositories are very important for open access. [The] EU should issue guidelines so the repositories within the EU are interoperable otherwise the effort of creating repositories will be a waste a time and money. Also [it] should help institutions and scientists with copyright problems, again giving clear guidelines.” (Qu. 11.2)

“Popular science information is also important, such as our web site <http://web4health.info/>. Such web sites need a process of certification of quality.” (Qu. 11.2)

“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. That rarely exists outside large centers. 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.” [copy edited from original] (Qu. 11.1)

Effectiveness of past programmes

“[The] EU has worked with these kind of repositories for a long time in UN, DESIRE, DELOS, ETB, RENARDUS etc ... to just mention the ones I worked in myself... All have/had excellent ambitions, but none of them seems to have hit the audience!” (Qu. 11.2)

Embedding repositories into the scientific mainstream

“Digital repositories will become increasingly important for the development of science. In particular, fast access to raw, not processed, uninterpreted data will be crucial to enable researchers to truly exchange results and ideas by providing a common ground for discussion that is not biased by preconceptions.” (Qu. 11.1)

“Repositories need to become embedded into research life in such a way that it is the natural thing for an academic or researcher to send his/her publications/primary data into the repository for long-term curation.” (Qu. 11.1)

“Depositing data in a repository should be come part of the scientific workflow to allow verification of published works and avoid unnecessary duplication. Repositories could then hold data and the literature in which the data are interpreted, plus the communication on the interpretation of the data.” (Qu. 11.1)

“I see repositories as sources of data that can both generate scientific development and business opportunities. If data obtained by national institutions is made available for free, many companies would be able to produce high-quality data services that would, through taxes more than offset the cost of obtaining the original data. If scientific and cultural sources are freely available the chances of increasing the amount and quality of research will increase. So future repositories should be as common as general search engines or e-mail.” (Qu. 11.1)

“There are many initiatives in Europe. All these are assuring interoperability, but the use of different technologies could hamper this 'interoperability'. The definition of guidelines valid for all projects could be beneficial in order to have a European system able to provide access to all relevant information residing in different places. Federation and real interoperability should be [the basis] of the guidelines” (Qu. 11.2)

“Social science data repositories have a long track record for acquiring, preserving, and making available social science research data. Many of the practices are applicable and potentially transferable. As science repositories are envisioned, there should be extensive and intensive collaboration and cooperation.” (Qu. 11.2)



Vision for an e-infrastructure for e-science digital repositories

“Standardized, distributed, intelligent, searchable, ubiquitous” (Qu. 11.1)

*“A virtual place where both people and machines can share information.”
(Qu. 11.1)*

“[T]here is no distinction between a repository and a well managed Web site. The current focus on 'repositories' as some kind of distinct artefact on the network is unhelpful and holding us back.” (Qu. 11.1)

“I think it's important that the EU should seek to support and enhance existing good practice within its research community, rather than impose a solution. In my area of astronomy we have well established, professional data centres across Europe, which work pretty well together in the development of interoperability standards, etc. It would be a colossal mistake to try and replace that by national or international repositories staffed by people who do not have the expertise required for the active curation of data which is necessary in a discipline like astronomy. Instead, the EU should concentrate its efforts on supporting the existing data centres financially and on sorting out fundamental infrastructural issues, such as an authentication and authorisation system which can be of use within an inherently multinational community like astronomy.” (Qu. 11.2)

“It is essential that digital repositories serve their respective research communities. That is, there is no "one size fits all" approach to digital data repositories and management. The nature of data from one discipline to another, and the culture for using and sharing data, is very different. Thus I hesitate a bit about endorsing "national repositories", or in particular, a national repository, because it is unlikely to be responsive to the needs of such diverse communities.” (Qu. 11.2)

“A sustainable, fully user transparent infrastructure with powerful search engines and data curation tools[, with] long term guaranteed preservation of important data[, w]here importance is determined by the end user.” (Qu. 11.1)

“Certainly not a centralized organisation, but 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.” (Qu. 11.1)

“Comprehensive open access collections of publications at the institutional level with aggregation of metadata at the national and international level to facilitate discovery. Research data would be linked (via metadata) to the relevant publications in institutional repositories but files [might] be stored locally (same institutional repository), nationally or internationally.” (Qu. 11.1)



“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.” (Qu. 11.1)

The users' view

“Better linking, less fragmentation.” (Qu. 11.1)

“Digital libraries will play the role of information and cultural disseminators of the future. People will be able to access and visualise (including get[ting] immersed via visual, aural, haptic or even odour channels) any type of information from simple text to 3D artefacts from all domains of application (culture, science, art, etc.).” (Qu. 11.1)

For sustainability of data

“[The] long term security of digital assets can't be guaranteed by any single entity, even the EU. A distributed model is better. This will need to be well thought out and I believe that it can be largely based on existing standards, of which there are many. So there is no need to invent new standards or protocols. We just need to think cleverly and learn lessons from models found in nature, and in other areas of human endeavour.” (Qu. 11.2)

“It's almost certain that we'll have to migrate content from the current generation of repositories to the next generation of repositories. Eventually, this will seem perfectly natural.”

Future role of the published article

“Publish or perish [should] be forgotten; open access or no success [should be a] reality” (Qu. 11.1)



Addendum A: Additional graphs

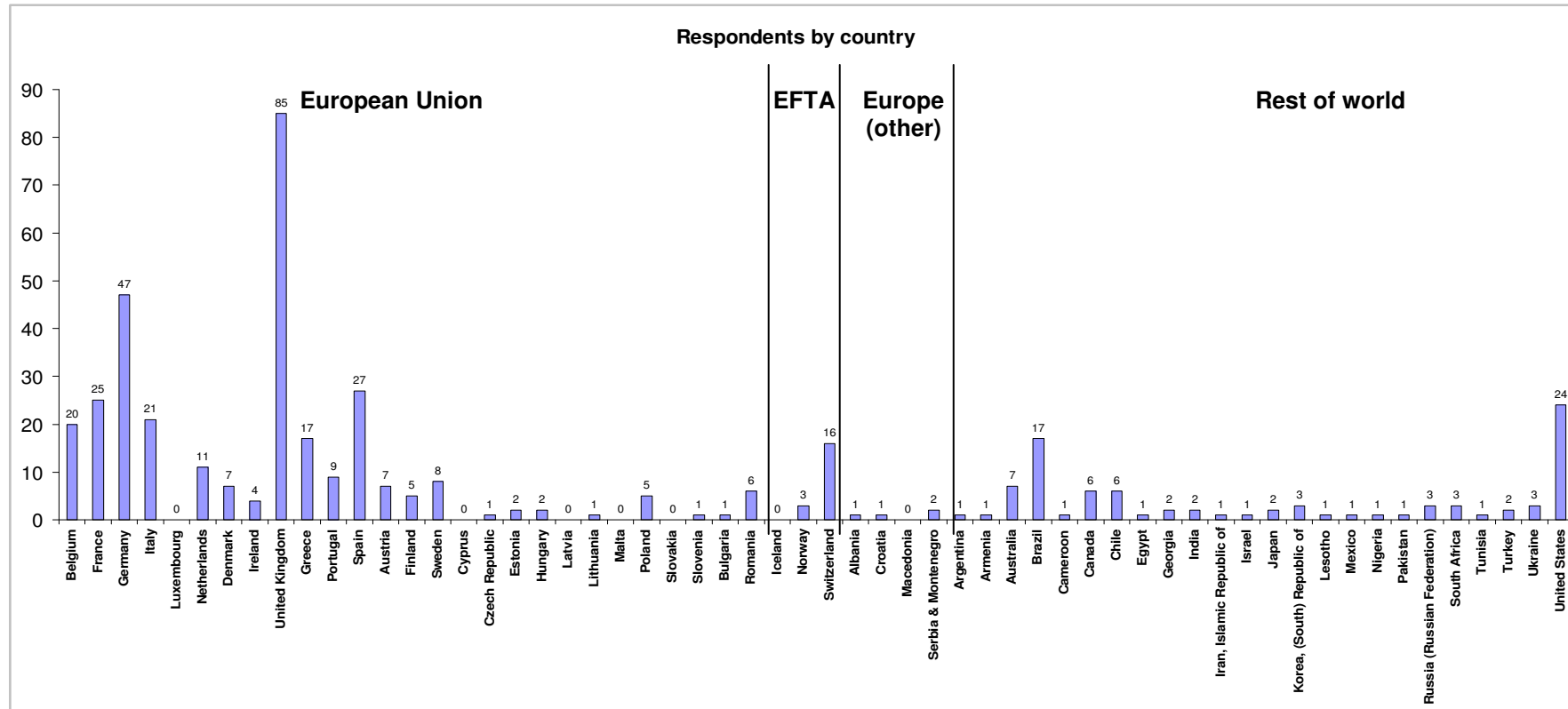


Figure A1: Respondents - raw counts per country



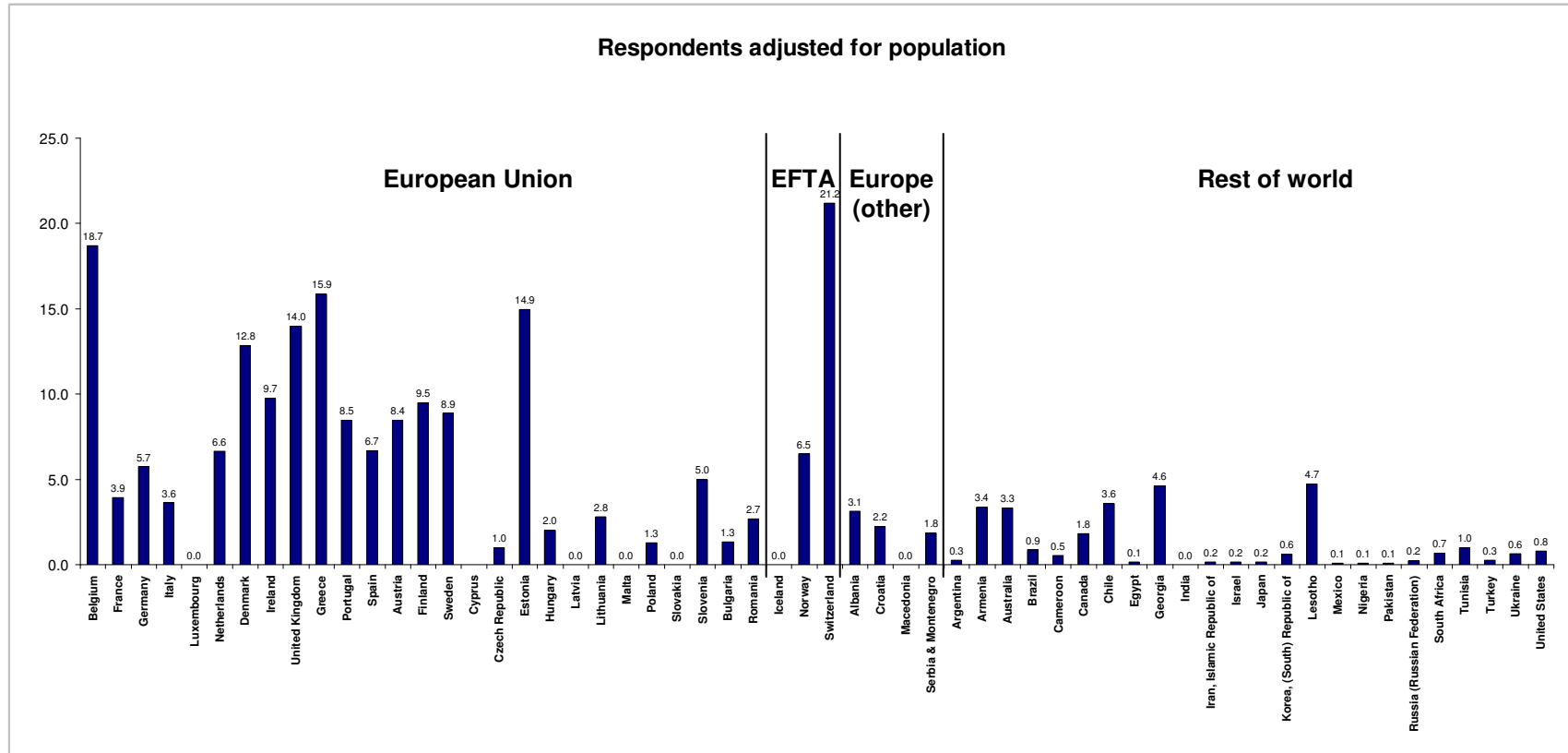


Figure A2: Respondents adjusted for country population (respondents per million people)



Use of repositories

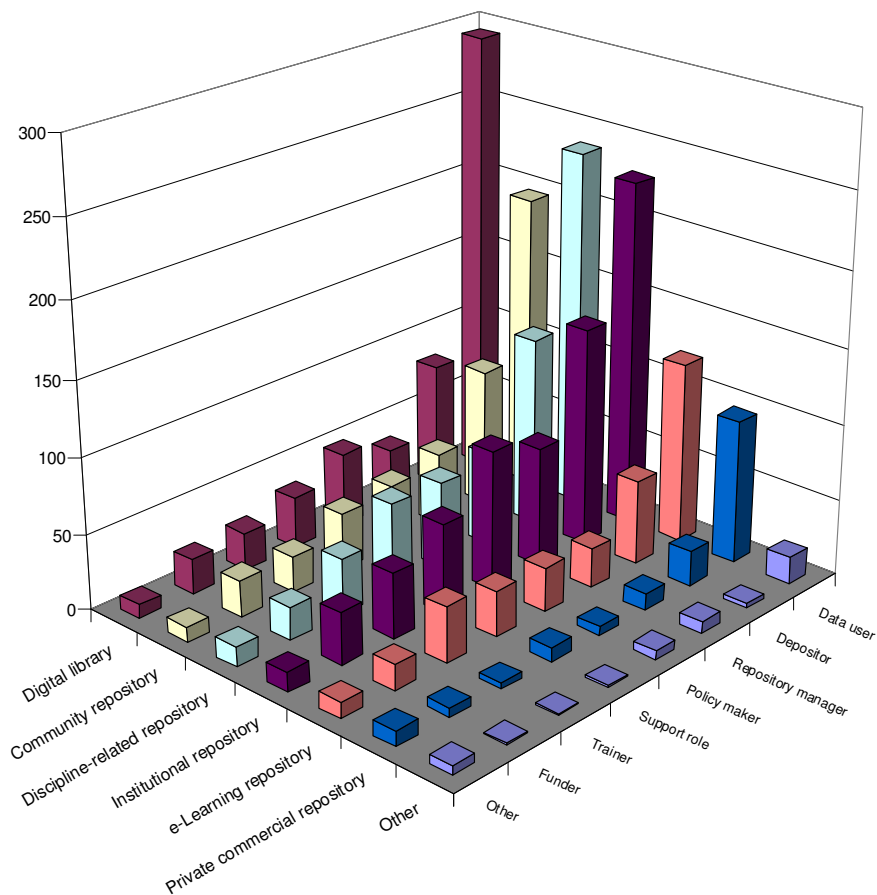


Figure A3: Use made of repositories –response counts by user type and repository type (Multiple choices available)



Addendum B: The public consultation questionnaire

The questionnaire used is reproduced here in Word format.

Three symbols are listed after each question thus: $[x, y, z]$ where
 $x = 1$ for a single response only allowed, $x = N$ where multiple responses allowed
 $y = O$ for an optional question and $y = M$ where a response was mandatory
 $z = T$ for a free text response, $z = C$ for pre-coded tick(s)

A. Information about the respondent and their preferences

Please enter details about yourself and your institution here.

Question group 1. Respondents' demographics

- 1.1 Your name:
 $[1, M, T]$
- 1.2 Country
 $[1, M, C]$ *[Standard drop-down country list]*
- 1.3 e-mail address
 $[1, M, T]$
- 1.4 Organisational affiliation (or equivalent)
 $[1, M, T]$
- 1.4.1 In what sector is this organisation?
 $[N, M, C]$
- Public sector (non-academic)
 - Academic sector
 - Not-for-profit
 - Commercial
 - A combination of any of the above
 - Private individual
 - Other
- 1.5 What is your main discipline/sector?
 $[N, M, C]$
- Chemistry, basic industry
 - Consumer goods
 - Earth sciences
 - Ecology
 - Economics
 - Energy, mining, utilities
 - Engineering
 - Financial, professional services
 - Healthcare, medicine
 - Information, library and archival sciences
 - Law, politics
 - Life sciences (including genomics, excl



- medicine, veterinary)
- Literature and languages
- Mathematics, statistics and computer science
- Media, broadcasting
- Physics, Astronomy
- Sociology/social sciences
- Transport
- Veterinary medicine
- Visual and performance arts
- Other

1.5.1 If “Other” please specify

[I,O,T]

- 1.6 What is your primary role? Director of an institution, company
- [N,M,C]* Principal investigator
- Researcher
- Software developer
- Librarian
- Archivist
- Research administrator
- Student
- Teacher
- Trainer, or support service provider
- Other

1.6.1 If “Other” please specify

[I,O,T]

- 1.7 Please indicate your age range. Below 25
- [I,O,C]* 25 - 40
- 40 - 65
- Above 65

1.8 Are you male/female Male Female

[I,M,C]

1.9 Are you answering on behalf of an institution? Yes No

[I,M,C]

1.10 I am willing to be contacted concerning this questionnaire Yes No

[I,M,C]



B. Establishing respondents’ use of digital repositories

The following questions ask you about your use of digital repositories.

Question group 2 – Use of repositories

- 2.1 How often do you use digital repositories?
[1,M,C]
- Daily
 - At least once per week
 - At least once per month
 - Rarely
 - Never

2.2 How do you work with digital repositories? Please tick all that apply across the repository type(s) and role(s) you assume.
[N,O,C]

	Data user	Depositor	Funder	Policy maker	Repository manager	Service provider	Support role	Trainer	Other
Community repository	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Digital library	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Discipline-related repository	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
e-Learning repository	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Institutional repository	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Private commercial repository	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Other	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

2.2.1 Do you have any other roles(s) vis-à-vis repositories?
[1,O,T]

2.2.2 Please specify other repository type(s) you use or work with.
[1,O,T]

- 2.3 If you use information from repositories, how often do you (or your institution) pay for access at the point of use?
- Never Sometimes Usually Always



[1,O,C]

2.4 If repositories had quality certification, would that encourage you to use them more? Yes No Don't know

[1,O,C]

2.5 Have you ever had any training / guidance on using repositories? Yes No

[1,O,C]

The following question asks about the types of data you either deposit into repositories or use from them:

Question group 3. What type of information do you deposit/use?

3.1	What types of information do you use from, or deposit into, digital repositories? Please tick all that apply.	<u>Deposit</u>	<u>Use</u>
		<input type="checkbox"/>	<input type="checkbox"/> Digital images
		<input type="checkbox"/>	<input type="checkbox"/> Moving images
		<input type="checkbox"/>	<input type="checkbox"/> Audio
		<input type="checkbox"/>	<input type="checkbox"/> Patent documents
		<input type="checkbox"/>	<input type="checkbox"/> Post prints
		<input type="checkbox"/>	<input type="checkbox"/> Pre-prints
		<input type="checkbox"/>	<input type="checkbox"/> Theses
		<input type="checkbox"/>	<input type="checkbox"/> Publications
		<input type="checkbox"/>	<input type="checkbox"/> Raw experimental data, or databases.
		<input type="checkbox"/>	<input type="checkbox"/> Processed data
		<input type="checkbox"/>	<input type="checkbox"/> Simulation results/models
		<input type="checkbox"/>	<input type="checkbox"/> (Research) Management/administrative data
		<input type="checkbox"/>	<input type="checkbox"/> Scientific workflows
		<input type="checkbox"/>	<input type="checkbox"/> Software code
		<input type="checkbox"/>	<input type="checkbox"/> Texts
		<input type="checkbox"/>	<input type="checkbox"/> E-mails
		<input type="checkbox"/>	<input type="checkbox"/> Other

3.1.1 If you ticked “Other”, please specify:

[1,O,T]



C. Exploring barriers to the use of digital repositories

To maximise the value of repositories to e-science any difficulties or barriers to their use need to be removed or diminished. The following questions ask what you think barriers (if any) might be, so that policies can be directed to removing them.

Question group 4: Difficulties and barriers to use of digital repositories?

- 4.1 What do you think are main difficulties when using repositories? Please tick all that apply. *[N,O,C]*
- Lack of training or guidance
 - Too costly
 - Difficult to use
 - Time consuming to deposit data
 - Time consuming to find information of value
 - Language problems
 - I do not know where to look for a suitable repository
 - Other

4.1.1 If you ticked “Other”, please specify: *[I,O,T]*

- 4.2 Which of the following do you think would indicate that a high level of trust can be placed on a repository and its contents? Tick all that apply. *[N,O,C]*
- Demonstrated awareness of the needs of users, in and between disciplines
 - Peer reviewed contents
 - Registration requirements to access information
 - On-line help and tutorials
 - Ability to access a telephone help-line
 - Publicly stated policies concerning the repository and its use are made clear
 - Payment for use
 - Other

4.2.1 If you ticked “Other”, please specify: *[I,O,T]*

Question group 5. What would (or does) prevent you from placing your information in a digital repository

5.1 What would (or does) prevent you from depositing your material into a digital repository? *[I,O,T]*

5.2 What would (or does) prevent you from using



information from a digital repository?

[1,O,T]

5.3 Have you encountered any specific frustrations using digital repositories?

Please describe briefly:

[1,O,T]



D. Exploring adequacy of current provision

In the following questions we ask your opinions about adequacy of provision of digital repositories, and your preferences for access to them.

Question group 6. Exploring adequacy of provision

- 6.1 Are you required to deposit your data in a repository?
[1,M,C] Yes No Don't know Not applicable
- 6.1.1 If yes, is the repository specified?
[1,O,C] Yes I have a choice No
- 6.1.2 If yes, what kind of repository is specified?
[N,O,C]
- A commercial digital library
 - A open access digital library
 - A general repository run by my institution
 - A repository established to serve my subject/community
 - A repository of learning materials
 - Other
- 6.2 Do you have access to any help or resources to deposit the materials?
[1,O,C] Yes No Not applicable
- 6.2.1 Please provide further details if you wish:
[1,O,T]
-

Question group 7. Question on what institutional setting is preferred

- 7.1 In what kind of digital repository would you feel most comfortable placing/accessing information? Tick all that apply.
[N,O,C]
- A commercial digital library
 - An open source digital library
 - A general repository run by my institution
 - A repository established to serve my subject/community
 - A repository of learning materials
 - Other
- 7.1.1 Is there any reason for your preference?
[1,O,T]
-



E. Views about enablers for digital repositories and policy directions

Software tools and advances in technologies can make digital repositories easier to use and even more useful to users. The following questions ask for your views on where such developments should be directed.

Question group 8. Exploring what advances would increase use and effectiveness/influence of repositories

- 8.1 What tools, resources, or organisational changes would make digital repositories easier to use? Tick all that apply. *[N,O,C]*
- Faster networks
 - More accurate searching mechanisms
 - Provision of training and guidance
 - Registries of available repositories(e.g. on-line lists with descriptive information about specific repository resources.)
 - Tools to automatically provide descriptive information (metadata)
 - Other

- 8.1.1 If you ticked “Other”, please specify: *[I,O,T]*

Question group 9. Exploring where investments would best be made over the next 5-10 years?

- 9.1 What steps at European level would add value to science, through use of repositories and their materials? *[I,O,T]*

- 9.1.1 Please add any other comments relevant to this issue you would like to make: *[I,O,T]*



F. Exploring a vision for the future of digital repositories in Europe

In the last set of questions we are asking for your views on a vision for digital repositories for e-science in the future.

Question group 10. Further work to be done

10.1 Would establishing national data repositories be a good idea?
Yes No Don't know

[1,O,C]

10.1.1 Please add any supporting comments you would like to make:

[1,O,T]

10.2 Would establishing international (e.g. EU) data repositories be a good idea?
Yes No Don't know

[1,O,T]

10.2.1 Please add any supporting comments you would like to make:

[1,O,T]

10.3 Do you ever, or would you like to, use any old digital materials?
Older than 10 years

Older than 20 years

Don't know

[1,O,C]

10.4 Do you generate material that needs to be kept for more than 10 years?
Yes No Don't know

[1,O,C]

10.4.1 If so, where will they be held
 My institution
 An archive serving my discipline/community

[N,O,C]



- A national archive for research data
- An international archive
- Other
- Don't know

10.5 Please add any another other comments you may have on sustainability of digital materials and the role of digital repositories?
[1,O,T]

Question group 11. Opinions on a vision for the future?

11.1 What would be your vision for digital repositories in the future?
[1,O,T]

11.2 Please add any further comments you feel are important:
[1,O,T]

11.3 Please tick if you would like to be sent a summary of the survey results.
[1,M,C]

Yes No



Section 5: Study Workshop Arrangements

This section describes the arrangements for the final study workshop, and provides a brief overview of the proceedings. The outcomes of the meeting were recorded and incorporated into the final report.

5.1 The study workshop agenda

The purpose of the Workshop was to review the findings of the study, and to give an opportunity to the repository community, broadly drawn, to comment on draft proposals for recommendations in the study's final report.

An overview of the study findings and the draft recommendations were provided to attendees in advance of the meeting (see Addendum B to this section). The meeting was structured in the following way to achieve the objectives sought:

- Introductory welcomes orientation by the Commission and a keynote address
- An overview of the study and the draft recommendations was presented by the DAC team
- Three parallel breakout sessions of about 20 people, chaired respectively by Alison Macdonald, Philip Lord and Neil Beagrie. These were structured to ask each of the participants to answer, with reasoning, 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?

In general a round-table format was adopted. Remarks were recorded by rapporteurs.

- Reporting back to the plenary meeting the results from the three break-out sessions
- An open discussion of the plenary meeting
- A closing address

The formal agenda for the meeting is provided as Addendum A to this section.

The meeting concluded with a cocktail reception. Plenty of time was provided during the breaks and reception for attendees to exchange views.

5.2 Practical arrangements for the Study Workshop

The initial planning for this event as discussed with the Commission and agreed; these arrangements are summarised below, updated to reflect the final logistics of the meeting.

Advice on the logistics of the meeting and on its organisation were provided by Com'tou in Paris, a partner in the e-SciDR project. Com'tou also supplied information and preferential rates for the hotels in Lisbon.

The timing and venue

The final Study Workshop was held on the 4th September at the National Archives of Portugal, at their premises in Lisbon (the “Torre de Tombo” building in central Lisbon).

The Workshop was held in Lisbon in view of the then presidency of the EU by Portugal. Both the National Archives of Portugal and the Gulbenkian Foundation in Lisbon were approached as venues, from which the National Archives was selected on the grounds of (a) their having availability for the



selected date and (b) their also hosting meetings dealing with matters related to repositories in the days following the Workshop³⁵, thus potentially providing for travel economies by attendees.

Three visits were made to the National Archives at their Torre do Tombo site in Lisbon in advance of the meeting to inspect the facilities and to discuss arrangements with staff of the archives.

The National Archives was able to supply a large, well equipped theatre (maximum 300 people) for plenary meetings and four smaller meeting rooms for break-out sessions (two adjacent to the auditorium and 2 in the second level of the building. Ample space was provided for the coffee breaks, lunch and the cocktail reception. Also provided was a cloakroom, space for attendees to meet and hold discussions, space for the display of materials, and a reception desk. Technical support was provided by the National Archives staff (projectors, recording apparatus, roaming microphones, extra furniture, charts, flags for the main stage). Interpretation facilities were available, but it was decided that they would not be needed for this meeting.

There was a large area used for poster displays and stands.

The Torre do Tombo is situated in northern Lisbon, in the University area, fairly close to the airport (about 15 – 20 minutes away). A number of hotels were situated within walking distance of the building. It is in a parkland area, close to two metro stations providing rapid access to the old centre of the city (about 20 minutes).

Accommodation

The two nearest hotels are the NH Campo Grande and Radisson, both within walking distance of the venue, and both were near bus stops and metro stations. Advantageous rates were negotiated with both hotels for attendees. The DAC also negotiated for other hotels for attendees asking for special arrangements.

Refreshments

Local caterers (Atelier Gastronomico) were hired for the occasion, and provided:

- Coffee/tea on arrival, mid morning and mid afternoon
- A buffet lunch (alcohol free so as to avoid too much of an “afternoon dip” in attendees’ concentration.)
- Cocktails at the end of the meeting; this featured a variety of port wines and Portuguese delicacies.

Running the logistics for the meeting

Staff from the DAC checked and set up the meeting area on the afternoon of the 3rd August:

- Checked that all the necessary equipment is available and working
- Pre-load presentations
- Set-up sign posts to rooms
- Ensure furniture was in place (tables, stands etc.)
- Checked security arrangements

³⁵ The DPE, PLANETS, and CASPAR projects held their second annual joint conference at the Torre do Tombo on the following two days



- Got acquainted with the layout of the facilities.

On the day of the Workshop (4th September) the DAC staff arrived early with name badges, documentation, and conference packs, etc for the attendees. They also ensured the catering arrangements were running smoothly, and that the arrangements for the rooms and presentations were in order with the Torre do Tombo staff. The rapporteurs for the break-out sessions were briefed.

For the whole day two Portuguese-speaking hosting staff (from Hospedeiras de Portugal in Lisbon) were available and on duty at the main reception/registration desk to assist attendees and to liaise with the Torre do Tombo staff. The Torre do Tombo also made a member of staff available for the whole day to help with the facilities, and they supervised the cloakroom area.

Security staff were arranged through the Torre do Tombo, and were on duty 8am to 8pm. Insurance arrangements were made by the Torre do Tombo.

Chairmen and rapporteurs

Philip Lord and Alison Macdonald alternated as chairs for the plenary sessions. Chairmen/leaders for the breakout sessions were Philip Lord, Alison Macdonald (DAC), Neil Beagrie (Charles Beagrie). Rapporteurs were Brian Fuchs (Imperial College), Isobel Galina (DAC), Daphne Charles (Charles Beagrie), Melanie Dulong du Rosnay, Damian Counsell (DAC), Pawel Plaszczyk (GridWise Tech).

Invited Speakers

The following people were invited to address the plenary sessions of the Workshop:

Francisco Barbedo, Depty Director, Direção Geral de Arquivo: Welcome to the National Archives/Torre do Tombo

Pedro Ferreira, Director of UMIC: Welcome to Lisbon and behalf of the Portuguese Presidency of the EU.

Carlos Morais-Pires, European Commission: The context of the meeting: “Towards a European e-Infrastructure”

Herbert Van de Sompel, Los Alamos National laboratory, USA: Keynote address

Jens Vigen, CERN: Closing address

Conference packs

A conference pack was provided for the attendees, consisting of:

- Name badge, with e-SciDR logo
- Smart carrying bag from the Torre do Tombo, and conference folder
- Workshop agenda (See Addendum A to this section)
- Handouts of slides of the DAC’s introductory briefing
- A full briefing paper (See Addendum B to this section)
- Attendee list
- Small gift from the Torre do Tombo



■ Lisbon tourist information

Recruitment of attendees and communication with them

Attendance at the conference was by invitation only. People with an active interest in repository issues from multiple perspectives were sent invitations by e-mail and in some cases by personal or telephone contact. The selection of people to invite was done on the basis of representing a wide spectrum of interests – over different disciplines, varied organisational contexts (academic, commercial, research institutions, libraries and archives, technologists), a geographic spread across Europe, a range of user types (users of information, data suppliers and repository managers).

Recruitment was started at the end of July 2007 by e-mail, and included a flyer explaining the purpose and background to the meeting (See Addendum C to this section) and the draft agenda. Further invitation notes were sent out during the early August, including those respondents to the public consultation who had indicated a willingness to be contacted. This recruitment was supplemented by telephone calls to key individuals. A total 81 people signed-up to attend, excluding the nine e-SciDR team members present. This figure includes two representatives from the European Commission (Carlos Morais-Pires and Elina Zicmane).

Further briefing information, travel and hotel advice was sent to attendees a few weeks before the meeting. Travel and hotel expenses were not, in general, subsidised or reimbursed by the project.

Summary of the discussion

Discussion centred on a number of themes, involving related to then draft recommendations presented to attendees (see Addendum B to this section).

Numerically, 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.



Addendum A: Workshop agenda



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

Review of draft study findings and recommendations

Tuesday, 4th September 2007, Torre do Tombo, Lisbon

Agenda

- 08.30 – 09.00 hrs: Coffee and registration
- 09.00 – 10.30 hrs: **Francisco Barbedo**, Deputy Director, Direcção Geral de Arquivo: Welcome
Pedro Ferreira, Director, UMIC: Welcome
Carlos Morais-Pires, European Commission: Towards a European e-Infrastructure
Herbert Van de Sompel, Los Alamos National Laboratory, USA: Keynote address
e-SciDR study team: Vision, findings and recommendations.
- 10.30 – 11.00 hrs: Coffee
- 11.00 – 12.45 hrs: Parallel sessions – discussion of recommendations
- 12.45 – 13.45 hrs: Buffet lunch
- 13.45 – 15.00 hrs: Reports from parallel sessions: questions and answers
- 15.00 – 15.15 hrs: Coffee and tea
- 15.15 – 16.30 hrs: Plenary session: open discussion
- 16.30 – 16.50 hrs: **Jens Vigen**, CERN: Closing address
- 16:50 - Cocktails



The e-SciDR team wish to thank the National Archives of Portugal for hosting the workshop.



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



Addendum B: 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

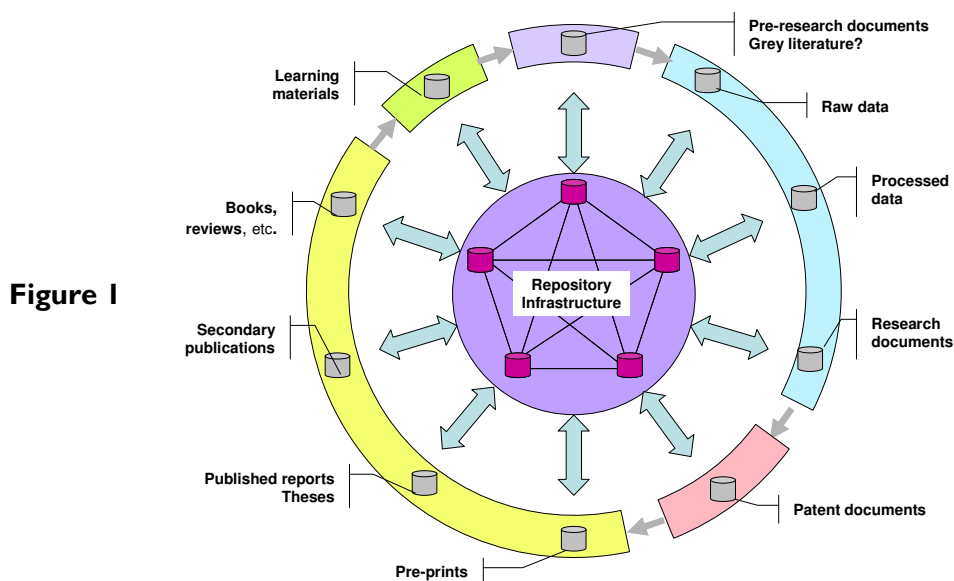
This short study was commissioned by the European Commission to provide an overview of the situation in Europe and recommendations in support of a definition of development scenarios for European-wide efforts to develop e-Science digital repositories for research and education.

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



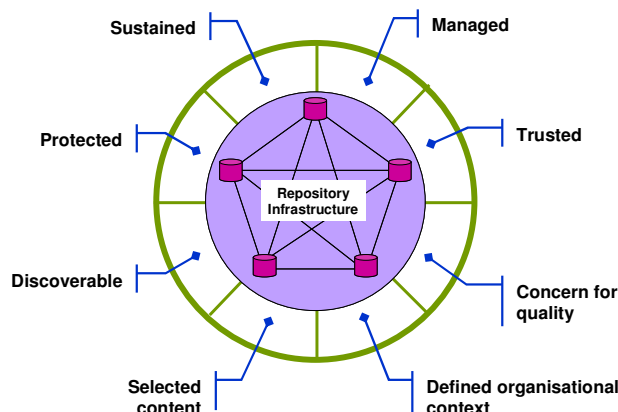


Figure 2

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:

- 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).
- There are also worries about a coming **deluge of metadata**, dwarfing the data deluge.
- The axes of **communication and professional incentive** are community and discipline based.
- There are **differences across Europe** in the level of use and penetration of repository technologies.
- **Inappropriate funding models** apply to the maintenance of repositories, for their own efficiency, sustainability and the preservation of content. Funding is also inadequate.



- **Harmonised and simplified authentication and authorisation** mechanisms are needed across Europe to gain access to e-Science resources in general and repositories specifically
- There are a few digital repository **notification services, registries** of data, and registries of repositories, but no one reliable, single pointer
- 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?
- 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.
- **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.
- Tools are needed to automate **metadata generation** and help users provide metadata.
- **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.
- 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.
- 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.
- **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.
- **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.
- 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.
- 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.
- **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?
- Questions on **organisational structure** to support data curation in the context of digital repositories, organisational relationships are being examined in actual initiatives and testbeds.
- **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?
- 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.
- **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%).

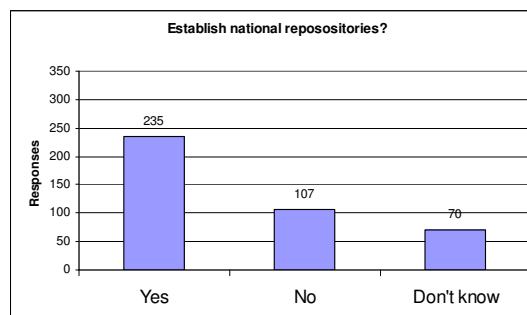
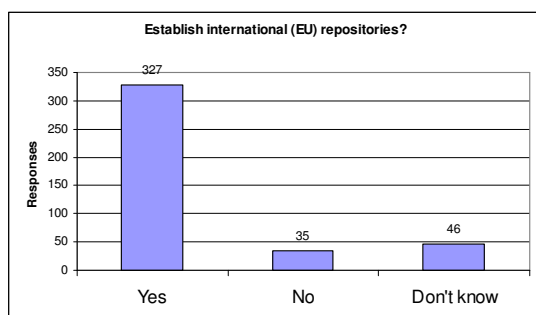


Figure 3: Support for international and national repositories

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

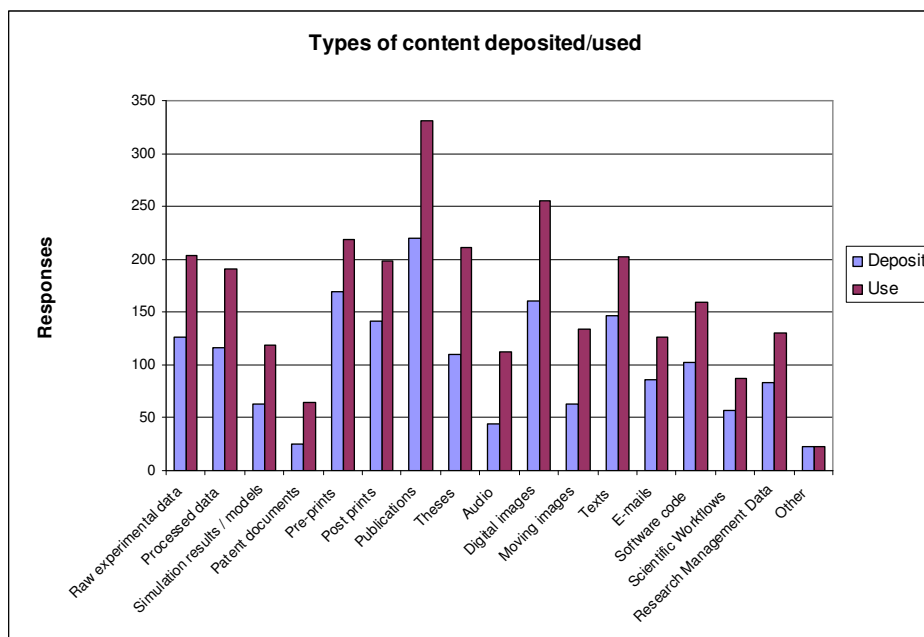


Figure 4: Types of content deposited and used

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.

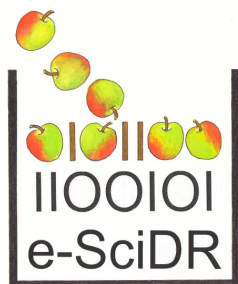
Accessible at (and perhaps co-ordinated by), there should be guides for researchers and institutions on the legal framework for creation, deposit, access and re-use.

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.



Addendum C: e-SciDR workshop flyer



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

Harvesting digital output for future use and distillation – a study for the European Commission

Ready and efficient access to digital materials and information of all kinds – experimental data sets, observation data, theses, publications, patents – is the life blood of new research and innovation. A robust data and information structure is critical for the future, not just for research, but education, the environment, the economy and society.

To give further impetus to the development and use of e-Science digital repositories, the European Commission's Information Society and Media DG has commissioned the study, "Towards a European e-Infrastructure for e-Science Digital Repositories" – "e-SciDR" for short.

■ Study objectives and programme

The e-SciDR study is to provide the European Commission with an overview of the situation in Europe regarding e-Science digital repositories (taking "e-Science" in the widest sense of science) and to identify an e-infrastructure for these repositories.

Using the overview and discussions with experts and practitioners as a baseline, the study will formulate policy recommendations for the Commission, in particular in the context of the Commission's FP7 activities, with the aim of actively driving forward the development and use of repositories in the EU.

After a public consultation and a workshop in September 2007, a final report with policy recommendations will be finalized in the autumn of 2007.

Further details and materials will be published on the study's web site at www.e-scidr.eu as the study progresses.

■ Study partners

This study is being carried out by the Digital Archiving Consultancy Limited (DAC), who are leading a team comprising GridwiseTech (Poland), University of Glasgow (UK), Charles Beagrie Limited (UK), Imperial College Internet Centre (UK) and Com'tou (France).



For further information, please contact the Digital Archiving Consultancy Limited:

2 Wayside Court Tel: +44 (0) 208 607 9102
TWICKENHAM Fax: +44 (0) 208 744 9322
Middlesex, UK info@d-archiving.com



The study is funded by the EU's Sixth Framework Programme and led in the Commission by the Emerging Technologies and Infrastructures unit of DG INFSO.



Appendices

Appendix A.1 Detailed notes on standards

A1.1 Introduction

In this appendix we present more extensive overviews of some of the standards areas summarised in section 1.4 of this report. These are:

- A1.2 Data Grid standards
- A1.3 Naming and name resolution
- A1.4 Data and metadata standards
- A1.5 Security including:
 - A1.4.1 Authentication
 - A1.4.2 Authorisation
 - A1.4.3 Web Services Security
 - A1.4.4 Security Assertion Mark-up Language (SAML)
 - A1.4.5 Extensible Access Control Mark-up Language (XACML)
 - A1.4.6 Liberty Alliance
 - A1.4.7 Standardised Risk Assessment
- A1.6 References **for this appendix**

Note: In this Appendix references in the text in square brackets (e.g. [REFERENCE]) refer to the list provided in section A1.6.

A1.2. Data Grid Architectures Standards for e-Science Digital Repositories

The vision of the Grid in supporting seamless access to a wide range of heterogeneous digital resources is a compelling one. Allowing end users to find, query and access data in flat files, spreadsheets or XML or relational databases is at the heart of data Grid vision. The standards associated with the establishment and long term management of Grids are still being defined. One of the main focuses of work within the Grid standardisation community (as represented through the Open Grid Forum) is the Open Grid Services Architecture (OGSA). This is intended to define service-oriented architecture comprising loosely coupled web services that can be applied across a variety of domains to produce a variety of Grids.

Despite the considerable effort that has already been invested, the OGSA is still very much a vision with much work still required for its successful completion and roll-out across the e-Research communities. One of the primary causes of this is the scope and application of Grid technologies. Providing a generic infrastructure that supports researchers from arbitrary disciplines with a plethora of, often orthogonal, requirements and usage expectations is especially challenging. Nevertheless the work on OGSA is putting together the core infrastructure needed for future Grids. It is fair to say however that the establishment of the OGSA has been delayed to some extent by the different flavours of technologies that have been put forward by the Grid and web service communities. Thus initial focus was on supporting Open Grid Service Infrastructure (OGSI) based web/Grid service (www.ibm.com/developerworks/grid/library/gr-ogsi/) with later efforts moving towards Web Service Resource Framework (WSRF - www.globus.org/wsrp/) and Web Service Interoperability (WS-I) (<http://www.ws-i.org>). As a result of this, many projects have developed their own solutions for managing the data sets that they are generating and have regarded the Grid standards development activities as non-critical to their own projects success – an all to common phenomenon



with projects of given fixed duration. It is the case however those truly successful projects produce software and data that are transferable to others after the project has completed.

It is generally acknowledged that establishment and management of compute Grids is far easier to achieve than data Grids. Providing access to a shared compute cluster for running simulations is a largely well understood problem with bodies of work on how to support secure access to such HPC resources and providing job scheduling across such resources etc. Many Grid efforts such as the UK National Grid Service (www.ngs.ac.uk) predominantly focus upon this kind of computational Grid. Many other kinds of Grid also exist however, enterprise Grids, campus Grids, semantic Grids and of concern here, information/data Grids. See the e-Science Gap Analysis (http://www.nesc.ac.uk/technical_papers/UKeS-2003-01/) for a survey of these different Grid classifications.

Data Grids are arguably the most difficult Grid to establish and manage. This is due to a variety of reasons. Some of these 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 in how best to annotate their data so that it might be found and subsequently used by others, and perhaps above all, the amount of data that is being generated across all research disciplines. In this context the establishment of a data infrastructure for e-Science digital repositories is especially challenging both in scope and complexity.

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 including:

- OGSA-Data working group are producing the overall data architecture as part of the larger OGSA effort. This data architecture describes the data services in the OGSA architecture and explains how they can be orchestrated to implement a range of data-oriented capabilities. Where possible, the group will use existing specifications to form appropriate parts of this architecture, liaising with the other groups defining those specifications to encourage them to fit into the overall OGSA picture.
- OGSA ByteIO working group are defining a minimal web Service interface providing "POSIX-like" file functionality allowing any service which implements the interface to be accessed in a file-like way.
- Database Access and Integration Services (DAIS) working group are developing standards for grid data services, focusing principally on providing consistent access to existing, autonomously managed databases through web services. This group has predominantly been working on the development of a family of data access and integration specifications. These specifications define data model independent properties and operations that are shared by interfaces to different kinds of data resource. To date, the group has focused on supporting access to relational and XML data resources.
- Data Format Description Language working group are defining an XML-based language for describing the structure of binary and character encoded files and data streams so that their format, structure, and metadata can be exposed.
- Grid File System working group are defining a standard service interface and architecture of a logical file system that can be used for management systems of both inter- and intra-enterprise grids.
- GridFTP working group is focused on improvements of FTP and GridFTP protocols with the goal to produce bulk file transfer protocol suitable for grid applications.



- Grid Storage Management working group is focused on the definition of the functionality of a standard storage resource manager interface offering dynamic space allocation and file management of shared storage components on the Grid.
- OGSA Data Movement Interface working group is focused on addressing the problems of discovery of data transport protocols available at a data's source and destination locations and agreeing on one of them, and the actual invocation of the agreed data movement. This includes direct data movements and 3rd party data movements.

Collectively all of these groups are producing specifications which ultimately once implemented will form a standards based Grid-based technology platform for access to and usage of e-Science digital repositories across Europe. These efforts also impact on other elements of OGSA however. Data replication, security, coupling of data Grids with computational Grids (moving data to HPC resources for processing or to specialized hardware for visualization etc) are just some of the many other factors that need to be incorporated to successfully establish data Grids serving the European e-Research community.

We note that outside of the Grid community efforts, commercial providers have also been developing technologies to support the establishment and management of data Grids. An analysis of one such technology and its comparison with the Grid communities OGSA-DAI technology is described in [SinAHM05].

It is the case that technology driven standards efforts need to be augmented by domain specific standards community efforts. Providing secure access to a range of data repositories from a technological perspective is useful, however supporting e-Research depends upon the contents of those repositories being accessible and of course useful. One of the challenges of supporting infrastructures for digital repositories is the naming of data and the meta-data describing the data itself. This is further complicated by the interdisciplinary nature of e-Research and especially by the length of time it takes for consensus and agreement being made to agree upon standards, often being far slower than the pace of scientific data production and research progress.

A1.3. Naming and Name Resolution Standards for e-Science Digital Repositories

In the future e-Science digital repository infrastructure across Europe it is to be expected that a multitude of repositories will exist across a variety of disciplines. These repositories will not be a closed or fixed set, but rather they will evolve with new repositories being established, old repositories either becoming obsolete or migrating for example to newer technology bases. Furthermore these repositories may themselves be complex entities offering a variety of different access and usage methods for different end users, for data providers, for data owners etc. In this environment, it is essential that the e-Infrastructure for digital repositories is designed to be extensible and scalable to accommodate change. Many domains also require that the data established in such repositories is maintained for long periods after initial deposition.

In this environment it is essential that the repositories can be accessed and used in a dynamic manner. Building on the distinguishing features of distributed systems such as dynamic discovery and late binding to address location and fault transparency, finding data and importantly services through which such data can be accessed and used is essential. To support such scenarios it is essential that these repositories and importantly the access points, potentially through a variety of different services, have names and addresses that can be resolved. Names in distributed systems are simply strings of bits or characters used to refer to entities where an entity can be almost anything such as a computer, a file or database. It is important to recognize that entities can be operated on, e.g. to upload data into a database.



The web service and Grid community have been involved in the establishment of a range of standards activities for naming and name resolution. The most developed of these efforts that has been refined by the Grid community is Web Services Addressing [WSAddr]. The WS-Addressing specification defines two interoperable constructs: endpoint references and message information headers. These constructs can be used to convey information typically required by transport protocols and messaging systems. In particular, these constructs normalize this information into a uniform format that can be processed independently of transport or application.

Rather than proposing changes or extensions to the WS-Addressing specification itself, the Grid standards community within OGF has chosen to define WS-Naming [WS-Naming] as a profile on top of the WS-Addressing specification. Neither web service clients nor web service endpoints need to be aware of this profile and either is free to fail to understand the WS-Naming elements described within. In such a case, the normal WS-Addressing behavior works exactly as described in the WS-Addressing specification. However, should a client, which is aware of the WS-Naming profile, encounter WS-Naming elements, it will have the opportunity to take additional actions with its communication to that web service endpoint in the event of certain communication failures or for the purposes of more robust or efficient communication.

These efforts are predominantly focused upon the naming of end point references whereby a specific service can be found and invoked however. Addressing finer grained data management naming and name resolution can be a much more involved process at the data level. Nevertheless fundamental to European digital repositories is that the capabilities they offer can be found and invoked in a consistent and standardized manner.

A1.4. Data and Metadata Standards for e-Science Digital Repositories

Accessing a single database and running a single query requires that knowledge of the underlying data model exists, i.e. the database schema. Knowing what tables exist, how they are structured (their rows and column names) is essential if meaningful results are to be returned. For small scale data models this is ordinarily not a major issue, however for e-Science digital repositories in the dynamic environment in which e-Research is conducted today, this can be (is!) an especially fraught process. To understand why data and metadata management is essential for e-Researchers we consider examples from the life science domain.

A1.4.1 Examples of Need for Standardisation of Domain Specific Data and Metadata: Bioinformatics Domain

One of the primary challenges that must be overcome in supporting life science research is naming resolution of gene identifiers and the associated experimental and array information. Being able to compare the results of different experiments fundamentally depends at the very least upon being able to assert a relation between the gene names or platform specific information between the experiments. Unfortunately repositories and individual sites typically use different naming conventions such as entrez and unigene. Accession numbers have also been introduced as a mechanism to uniquely identify genes and establish correspondences between information stored in different or in some case the same repository. For example, the NCBI's Gene Expression Omnibus [NCBIgeo] data set is available in both MINiML and SOFT formats but the two are not equivalent. There are many more SOFT files than MINiML but not all of the entries are available in one format or another.

Furthermore there are several large scale repositories that exist specifically for storage of microarray data sets. Some of these include Gene Expression Omnibus (GEO) at NCBI [NCBIgeo], ArrayExpress [ArrExpr] and CIBEX [CIBEX]. As well as storing microarray data sets, these



repositories also provide various kinds of services through which the repositories themselves might be searched or mined. These repositories typically require data sets to be MIAME compliant.

The stated goal of MIAME is to outline the minimum information required to interpret unambiguously and potentially reproduce and verify an array based gene expression monitoring experiment [MIAME]. Whilst the details of particular experiments themselves may be different, it is the intention of MIAME to define a core that is common to most experiments. It should be noted that MIAME is not a formal specification, but rather a set of guidelines which concentrate on the content of information. It is not in itself a data format but provides a conceptual structure for capturing the metadata associated with microarray experiment descriptions. A MIAME description will typically describe the design of the array platform and of the gene expression experiment. The array design specification consists of the description of the common features of the array as the whole, and the description of each array design elements, e.g. each spot. The gene expression experiment description includes a description of the overall experimental design; the samples used; how extracts were prepared; which hybridisation procedures were followed and ultimately what data was measured and how it was analysed and normalised.

MIAME compliance is not prescriptive in the sense that all or a given subset of the various sections that might be associated with a given experiment must be given. These sections are usually provided in free text format, along with recommendations requiring maximum use of controlled vocabularies or external ontologies. MIAME recognises that few controlled vocabularies have been fully developed, hence it encourages users to provide their own qualifiers and values identifying the source of the terminology. Of those that are available, the Microarray Gene Expression Data Society (MGED) [MGED] is one of the more established ontologies for microarray experiment description.

Several data formats have been defined and applied across different sites and with different user communities. These include: MAGE-ML [MAGEml], SOFTtext [SOFTt], MINiML [MINiML] and SOFTmatrix [SOFTm].

MAGE-ML is part of the MGED family of standards and is MIAME compliant and XML based. Libraries for handling MAGE-ML exist for Java, C# and Perl with a python version in development. Many major repositories, such as GEO, ArrayExpress and CIBEX support results being deposited in MAGE-ML as well as supplying data in that format.

SOFTtext is a simple text based format designed by GEO. Unlike MAGE-ML, SOFTtext is not XML based using instead keywords for describing platform, sample and results. It has fewer fields than MAGE-ML yet is still MIAME compliant. GEO supports submissions in this format and makes results available in it as well. Since SOFTtext is based around a simple format it is easy to parse and use.

MIAME Notation in Markup Language (MINiML) is an XML based format used by GEO and is equivalent to SOFT. The NCBI accepts data deposited in MINiML format and makes records available in this format. MINiML can be considered an XML equivalent to SOFTtext as it provides the same properties, however in XML form. NCBI has made a schema for MINiML available allowing a validating parser to confirm that a MINiML file is well formed. This is a distinct advantage over SOFTtext where there is no formal definition of how the files should be formatted. As with the other SOFT formats MINiML is MIAME compliant yet has fewer fields than MAGE-ML. The relative simplicity of MINiML when compared to MAGE-ML has direct advantages for usability and associated learning curve.



SOFTmatrix is a new format based on a spreadsheet. Like SOFTtext it was developed by the NCBI based on MIAME. The format uses Microsoft's Excel .xsl files as a base and consists of a simple template. Given the extensive use of Excel in processing microarray results by the biological community, using it as a form of exchange format was arguably inevitable. It should be noted that the .xsl format is proprietary and its format is not officially published in the public domain. As a result, long term usage may be a potential issue due to potential licensing issues. As seen a multitude of on-going efforts in how to describe and annotate the data and metadata associated with microarray experiments and results exist.

As a result the e-Research projects typically resort to developing services that allow for correspondences to be established between gene names. One effort that was established to address this issue was the Life Science Identifier (LSid) initiative [LSid]. LSids are designed as a Uniform Resource Name based identifier which itself is a form of Uniform Resource Identifier. LSids themselves are written in the form: urn:lsid:<authority>:<database>:<object>:<version> where <authority> is the name of the authority who issued the LSid, <database> is the name of the authority's database the LSid is stored in and <object>:<version> identifies the object within the database and its revision.

LSids are intended to serve as persistent identifiers allowing them to be used without later being reassigned. They allow mapping to exactly the same set of bytes permanently. This means that an LSid, once assigned, is permanently attached to a specific encoding of its data which cannot be updated or corrected. An immediate advantage of this is that makes LSids usable as references.

LSids also support attaching metadata, in a variety of forms, allowing an automated parser to discover for instance, synonyms, creation information and alternate versions of the LSid. The versioning field at the end of the LSid is optional but can be used to differentiate between revisions of the object or different representations as well. When there is a mapping from an existing dataset's accession number to an LSid it is possible for previous accession systems to generate an LSid for their data making any program that uses LSid able to access a wider range of data. No standard mechanism for performing this transform is defined however hence this makes the use of automatically generated LSids by a program risky until a recognised authority formally assigns them.

The LSid specification suggests using an LSid proxy, e.g. lsid.biopathways.org, to resolve LSids. The biopathways resolver provides LSids for many existing data sets such as the NCBI databases, ArrayExpress and SwissProt for example. However relying on a sole point of access is dangerous as in the event of its failure, all of the data sets accessed through the proxy will become unavailable. A model with independent authorities is more robust as the loss of one authority results in a smaller loss. Conversely, having a great many authorities ensures that, at any given time, some of the authorities will be unavailable. Whilst there is no mechanism for reserving LSids, there are mechanisms for requesting that valid LSids exist.

At the time of writing, it is unclear whether LSids will solve the problems arising in uniquely identifying information in the life science domain. For example, the closure of the Interoperable Informatics Infrastructure Consortium (i3c) means the loss of RDF metadata associated with LSids. References to this data still appear in examples and tutorials but the i3c itself website no longer exists. The only implementations of the LSid stack found are from the IBM LSid project on sourceforge. There are two implementations available one in Java the other Perl. The logs of the source repository reveal little activity with the majority of the code remaining untouched since 2004.



As a result researchers are often left adopting more pragmatic solutions based upon for example, local hash tables and schemas for cross referencing gene expression naming information. Whilst suitable for demonstration and prototype production within the lifetime of their given projects, this will ultimately be a short term solution. A common standard and agreement adopted by the life science community is urgently required.

This phenomenon is not unique to the bioinformatics community. In the clinical sciences domain there are numerous developments in standards for the description of data and meta-data used in the clinical trials domain. However, this can be an involved process depending on standards groups developing and acting on strategies put together through major initiatives such as Health-Level 7 (HL7) [HL7], SNOMED-CT [SNOCT] and OpenEHR [OEHR]. There are often a wide range of legacy data sets and naming conventions which impact upon standardisation processes and their acceptance. The International Statistical Classification of Disease and Related Health Problems version 10 (ICD-10) [ICD10] is used for the recording of diseases and health related problems and is supported by the World Health Organisation. In Scotland, ICD-10 is used within the NHS along with ICD-9 and Read codes in the SMR data sets for example. ICD-10 was introduced in 1993, but the ICD classifications themselves have evolved since the 17th century [ICDb]. Global Grid frameworks that incorporate appropriate meta-data identifying the different local data classifications are needed to address such discrepancies.

The standardisation process itself may influence how readily any given standard is adopted. For example, standards developed to specific deadlines during the standardisation-making process, and standards bodies producing regular updates with solutions readily available for implementation are more likely to gain acceptance. This is also the case within the Grid community. Linking standardised data descriptions between domains so that entities and relationships within one organisational hierarchy can be mapped or understood within the context of another domain is fundamental to the development of the Grid applications. Once it has been established how meaningful comparisons can be made between the schemata of differing domains, this knowledge can be applied to a generic clinical trial that could run queries across heterogeneous domains, bringing back generic results, richer in scope and information than if single local sites had been independently queried.

To address these issues a variety of standardisation approaches have been adopted including:

- Community efforts
- Ontologies
- Semantic web
- Structured vocabularies, data dictionaries

Many of these kinds of approaches and their pro's and con's are described in detail in JDSS.

A1.5. Security Standards for e-Science Digital Repositories

The development of robust Grid security infrastructures for e-Science Digital Repositories across Europe is very much dependent upon agreements on technologies and practices. Standardisation plays an extremely important role in this regard. From the end user perspective, e-Infrastructures should provide for simple, secure single sign-on to a multitude of e-Research resources. That is, having authenticated once, they should be able to access a range of distributed resources without the need for further authentication. The privileges that the end users have (or do not have) should then be transparently used by resource providers to make their own local authorisation decisions on access and usage requests. End users will see digital repositories personalised to their research



interests and their associated privileges. This personalisation should be cognisant of the data owners, data providers and other key stakeholders involved in supporting and managing these repositories and services associated with them.

Whilst some communities such as 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 is thus on having environments that facilitate research and not in 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. The ideal scenario is that users are unaware that they are accessing Grid resources or rather these resources should be provided in a manner aligned with the way the internet is accessed and used more generally. Existing Grid models for the most part are not yet aligned with this modus operandi. Thus for example, many users' initial explorations into e-Research begin with having to acquire an X.509 digital certificate [X509] from a national certificate authority which is subsequently used for the Public Key Infrastructure (PKI) [PKI] based approaches to single sign-on (see section 8.1). It is now recognised however that many end users are put off by such certificates, hence other models of access to e-Infrastructures are required.

One such model that is being explored by many sites is to provide Shibboleth [SATO,SAPP] access to and usage of a range of resources. With this model, when a user attempts to access a Shibboleth protected service or Service Provider (SP) more generally, e.g. a European Digital Repository, they are typically redirected to a WAYF server that asks the user to pick their home Identity Provider (IdP) from a list of known and trusted sites. In the UK a single federation has been established [UKfed]. Other international federations have also been put forward and established [SWISSfed, FinFed, AusFed, USfed].

After the user has picked their home site, their browser is redirected to their site's authentication server, and the user is invited to log in. After successful authentication, the home site redirects the user back to the SP and the message carries a digitally signed SAML [SAMLv1] authentication assertion message from the home site, asserting that the user has been successfully authenticated (or not!) by a particular means using an authentication mechanism specific to the IdP and potentially containing one or more attributes defining the privileges that this user possesses. Assuming the digital signature on the assertion is verified and the user has successfully authenticated themselves at their home site the SP may decide to allow access.

This security model offers several direct benefits over PKIs for dynamic establishment of VOs in that users are no longer trusted to manage their X509 certificates and remember complex passwords. Instead institutions within a federation have a degree of trust with one another. Sites/IdPs and SPs are still autonomous and are able to decide for themselves whether the provided attributes are sufficient for access to the resources and which attributes they are prepared to release to which SP. Another key benefit of Shibboleth for VO establishment and management is that users are only required to remember their own usernames and passwords at their home institutions. Provided a common understanding of the roles and security attributes across the sites comprising the federation exists, single sign-on can be achieved. Thus if a SP trusts a given site for authenticating a user requesting access to its own resource, and also an agreement on the attributes which are to be exchanged between the sites exists, then the SP can authorize/restrict access to its resources from those sites that are within the federation provided the necessary attributes and values are presented by the IdP.



Within the UK federation, a small set of security attributes based upon a subset of the eduPerson specification is being adopted. European-wide federations needed for single sign-on across European e-Infrastructures may adopt similar sets of attributes. They may also require particular roles or licenses to be presented by end users. Technologies that allow assigning or revoking these privileges across multi-institutional, international settings are thus needed. Ultimately every resource provider (digital repository) will be autonomous and decide for itself whether the information that was supplied by a particular identity provider is sufficient to allow access to a particular resource or not.

There has been much work undertaken already by standards bodies in this area and a multitude of standards efforts are currently being pursued by the web service community. We provide an overview of these standards and related efforts here.

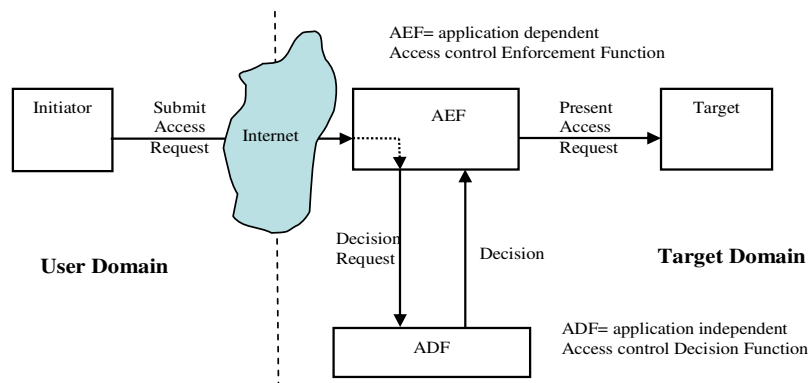
A1.5.1 Standards for Authentication

Authentication is a key element of security. Knowing the identity of the individuals attempting to access and use a given resource is essential to resource providers and digital repository stakeholders. The de facto way in which this is achieved in the e-Research community is through digital certificates used to support Public Key Infrastructures (PKI). The X509 standard [X509] has been widely accepted across the community for its support of PKI. In principle, through establishing the identity of the end user through the Distinguished Name (DN) of the X509 certificate and knowing/trusting the issuer of the certificate, a user can access a range of resources recognising that issuer.

Crucial to authentication and security more generally is trust. Knowing who issued a certificate to an end user is as important (if not more so!) as knowing the identity of the end user with that certificate. The establishment of trusted certification authorities and their associated policies is essential for the success of a European e-Infrastructure for digital repositories. This is one area that the e-Research community has focused upon with commonly accepted policies on cross-grid authentication. Within the Grid standards community this is the focus of the work of the Certification Authority Operations (CAOPS) Working Group (<https://forge.gridforum.org/projects/caops-wg>). This group maintains strong links with the International Grid Trust Federation (www.gridpma.org) and is concerned with the actual implementation of guidelines and accreditation of authentication providers.

A1.7.2 Standards for Authorisation

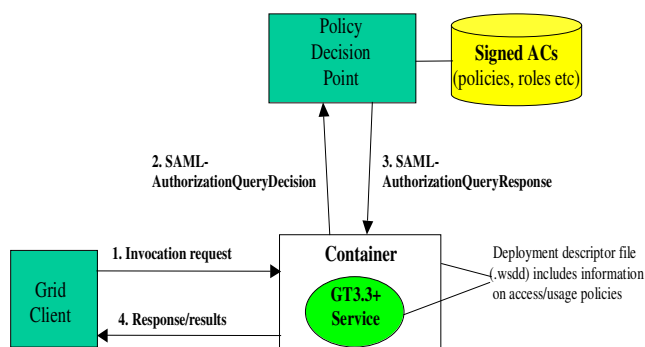
There has been much work already undertaken on standards for security that go beyond existing X509 based PKI authentication-only based models. One of the main authorisation standards in this area has been the X.812 | ISO 10181-3 Access Control Framework standard [X812]. This standard defines a generic framework to support the generic process of authorisation as depicted in Figure 1.



X.812 Access Control Framework

In this model, the initiator attempts to access a target in a remote domain, e.g. a protected e-Science digital repository. Two key components support authorised access to the target: a Policy Enforcement Point (PEP), described in the figure as the Access control Enforcement Point (AEF), and a Policy Decision Point (PDP), described as the Access control Decision Function (ADF). The PEP ensures that all requests to access the target are run through the PDP and the PDP casts the authorisation decision on the request based on a collection of rules (policies). To make this structure scalable and easily applicable within a Grid environment, a generic application programming interface (API) to model the PEP has been proposed and created by the Authorisation Working Group of the Open Grid Forum (OGF) (www.ogf.org) – previously known as the Global Grid Forum. This generic PEP can be associated with arbitrary authorisation infrastructures. The specification for Grid technologies is an enhanced profile of the OASIS [OASIS] Security Assertion Markup Language (SAML) v1.1 [SAML1-1].

The OASIS SAML AuthZ specification defines a message exchange between a PEP and PDP consisting of an AuthorizationDecisionQuery (which contains a subject, a resource and an action) going from PEP to PDP, and an assertion returned containing a number of AuthorizationDecisionStatements. The OGF SAML AuthZ specification [WSCMP] defines a SimpleAuthorizationDecisionStatement (a boolean stating “granted/denied”) and an ExtendedAuthorisationDecisionQuery that allows the PEP to specify whether the simple or full authorisation decision is to be returned. Figure 2 shows the interactions supported by this API.



Open Grid Forum SAML AuthZ API

Through this SAML AuthZ API, a generic PEP can be achieved which can be associated with arbitrary Grid services. Thus rather than developers having to explicitly engineer a PEP on a per application basis, the information contained within the deployment descriptor file (.wsdd) when the service is deployed within the container, is used. Authorisation checks on users attempting to invoke “methods” associated with this service are then made using the information in the .wsdd file and the contents of the LDAP repository (PDP) together with the DN of the user themselves. Releases of the Globus software since GT3.3 have supported this API.

We note that one issue that has been encountered with the SAML AuthZ profile, which has a direct consequence on its use for e-Science Digital Repositories in Europe, is the lack of granularity in how users might invoke actions [SC]. For example, different actions may or may not be allowed depending upon the data that they wish to access and potentially change. The SAML AuthZ profile does not currently allow actions to be distinguished based upon the parameters that might be associated with them. As a result, a query service cannot easily (at least in a manner that easily scales) be restricted to query those data sets in a given set of federated databases that are appropriate



to the invoker. Instead, the SAML AuthZ specification supports either a secure query service or a non-secure query service. The OGF AuthZ working group is now working on a new version of this API (to support parameters). This work is still under development within the OGF Authorisation working group. Once implemented, it will potentially offer an important component needed for digital repository providers to facilitate secure access to their data sets.

With the move of the Grid community towards web services and service-oriented architectures, web service security standards and their associated implementations are crucial to support future European e-Science Digital Repositories.

A1.5.3 Web Service Security Standards

It is the case that a multitude of specifications and proposals for web service standards have been promised and put forward, or merely promised. There are often cases of web service standards covering similar topics resulting in multiple competing specifications such as WS-Notifications [WS-N] and WS-Eventing [WS-E]; WS-ReliableMessaging [WS-RM] and WS-Reliability [WS-R]; WS-Orchestration [WS-O], WS-Co-ordination [WS-Co] and WS-Choreography [WS-Ch], along with the many varieties of workflow or business process languages that have been put forward to name but a few examples of the issues in the proliferation of web service standards. It is also the case that at the time of writing, many web services standards are only in working draft or draft status, often with no associated implementations or acknowledged conformance or interoperability definitions. Claiming conformance or compliance to a particular web service standard is thus often not possible (or meaningful!).

It is also apparent that although many standards use the common prefix “WS-”, this does not mean that there is an agreed WS-Architecture. This stems from a variety of reasons: vendor and commercial issues; political aspects and also the different bodies involved. For example the Internet Engineering Task Force (IETF) (www.ietf.org); the World Wide Web Consortium (W3C) (www.w3.org); the Organization for the Advancement of Structured Information Standards (OASIS) (www.oasis-open.org); and the Web Services Interoperability Organization (WS-I) (www.ws-i.org) are some of the most prominent bodies. The consequence of this profusion of standards and standards making bodies, and the lack of consensus on the core web service architecture, impacts directly upon development of Grid standards, architectures and associated implementations and middleware – and in turn on support for the infrastructure for e-Science Digital Repositories across Europe.

With this complexity in mind, several key standards have nevertheless been identified by the web service security. We provide a brief overview of these security standards. All of these standards build upon the basic SOAP foundations which include XML Signature [XMLSig] and Encryption [XMLEncrypt] for ensuring the security of messages.

A1.7.3.1 WS-Security

WS-Security describes enhancements to SOAP messaging to provide security enhancements for message integrity and message confidentiality. WS-Security also defines a general purpose mechanism for how to attach and include security tokens within SOAP messages including binary encoded security tokens such as X.509 certificates. These mechanisms can be used independently or in combination to accommodate a wide variety of security models and encryption technologies.

Message integrity is provided by leveraging XML Signature in conjunction with security tokens to ensure that messages are transmitted and received without modifications. The integrity mechanisms



are designed to support multiple signatures, potentially by multiple actors, and to be extensible to support additional signature formats. The signatures may themselves reference security tokens.

Message confidentiality is provided by leveraging XML Encryption in conjunction with security tokens to keep portions of SOAP messages confidential. Any portions of SOAP messages, including headers, body blocks, and substructures, may be encrypted. It should be noted that the encryption mechanisms of XML Encryption are designed to support additional encryption technologies, processes, and operations by multiple actors. The encryption itself can be realized using either symmetric keys shared by the sender and the receiver of the message or a key carried in the message in an encrypted form.

WS-Security defines a framework for securing SOAP messages, with the specifics being defined in profiles determined by the nature of the security token used to carry identity information. There are for example different profiles of WS-Security for various different security token formats such as X.509 certificates and Kerberos tickets. There is also a SAML token profile of WS-Security that specifies how SAML assertions can be used to provide message security. Additionally, SAML itself points to WS-Security as an approved mechanism for securing SOAP messages carrying SAML protocol messages and assertions.

WS-Security has now been fully implemented by several web service providers and the Grid middleware community. For example, the OMII server and client software stacks provide an implementation of WS-Security based upon Axis and WSS4J [WSS4J].

A1.7.3.2 WS-Policy

WS-Policy [WS-Policy] describes how senders and receivers can specify their security requirements and capabilities. WS-Policy has been designed to be extensible and does not place limits on the types of requirements and capabilities that may be described. However, the specification does identify several basic attributes including privacy attributes, encoding formats, security token requirements, and supported algorithms. WS-Policy thus provides a flexible and extensible grammar for expressing the capabilities, requirements, and general characteristics of web service-based systems. WS-Policy also defines a framework and a model for the expression of these properties as policies. Policy expressions can include both simple declarative assertions as well as more sophisticated assertions. A policy itself can be regarded as a collection of one or more policy assertions. These assertions might include for example the authentication scheme, transport protocol selection, privacy policy, or quality of service characteristics. WS-Policy provides a single policy grammar to allow for such kinds of assertions to be reasoned about in a consistent manner.

It should be noted that WS-Policy stops short of explicitly specifying how policies are discovered or attached to a web service. It is envisaged that subsequent specifications will provide profiles on WS-Policy usage within given web services technologies and domains. For example, specifications for WS-PolicyAttachments, WS-PolicyAssertions, WS-SecureConversation have been put forward already as have various domain-specific assertions such as WS-SecurityPolicy and WS-ReliableMessagingPolicy. (See [WS-Policy] for further information).

A1.7.3.3 WS-Trust

The goal of WS-Trust [WS-Trust] is to enable applications to construct trusted SOAP message exchanges. WS-Trust uses the basic mechanisms for secure messaging from WS-Security and defines additional primitives and extensions for security token exchange to enable the issuance and dissemination of credentials within and between different trust domains. Thus for example, to secure a communication between two parties, the two parties must exchange security credentials (either



directly or indirectly). However, each party needs to determine if they can trust the asserted credentials of the other party. To support such situations, WS-Trust has defined extensions to WS-Security that provide methods for issuing, renewing, and validating security tokens; and ways to establish, assess the presence of, and broker trust relationships. Through these extensions, applications can engage in secure communication designed to work with the general web services framework including WSDL service descriptions, UDDI and SOAP messages.

A1.7.3.4 WS-Privacy

The WS-Privacy specification was outlined in a joint white paper from IBM and Microsoft [WSW]. Here it was presented how the WS-Privacy specification could address how privacy practices could be stated and subsequently implemented and enforced by web services. By using a combination of WS-Policy, WS-Security and WS-Trust, organizations should be able to state and indicate conformance to stated privacy policies. The specification would describe a model for how a privacy language could be embedded into WS-Policy descriptions and how WS-Security may be used to associate privacy claims with a message. In addition, the WS-Privacy specification would describe how WS-Trust mechanisms could be used to evaluate these privacy claims for both user preferences and organizational practice claims.

At the time of writing, the WS-Privacy specification and associated implementation(s) have not materialised, nor is it clear when they will appear.

A1.7.3.5 WS-SecureConversation

The Web Services Secure Conversation Language (WS-SecureConversation) [WS-SC] allows clients and web services to establish a token-based, secure conversation for the duration of a given session. The secure conversation itself is based on security tokens that are procured from a service token provider. Once obtained and a secure channel established, the client and service exchange a lightweight, signed security context token, which optimizes message delivery time compared with using regular security tokens. The security context token enables the same signing and encryption features as other security tokens such as X509 security tokens.

WS-SecureConversation itself is built on top of the WS-Security and WS-Policy models to provide secure communication between services. WS-Security focuses on the message authentication model but not a security context, and thus is subject several forms of security attacks. WS-SecureConversation defines mechanisms for establishing and sharing security contexts, and deriving keys from security contexts, to enable a secure conversation.

It should be noted that WS-SecureConversation by itself does not provide a complete security solution rather WS-SecureConversation is a building block that is used in conjunction with other web service and application-specific protocols such as WS-Security to accommodate a wide variety of security models and technologies. It should also be noted that WS-SecureConversation is designed to operate at the SOAP message layer so that the messages may traverse a variety of transports and intermediaries. This does not preclude its use within other messaging frameworks. In order to further increase the security of the systems, transport level security may be used in conjunction with both WS-Security and WS-SecureConversation across selected links.

Several implementations of WS-SecureConversation are now available for example within Microsoft Web Service Enhancements for the .NET platform [WSE].



A1.7.3.6 WS-Federation

The Web Service Federation Language (WS-Federation) [WS-Fed] defines how to construct federated trust scenarios using the WS-Security, WS-Policy, WS-Trust, and WS-SecureConversation specifications. For example, WS-Federation describes how to federate between Kerberos and PKI infrastructures. The WS-Federation specification defines the model and framework for federation between security domains. Subsequent documents define profiles which detail different ways that the WS-Federation language can be applied.

WS-Federation supports specification of a trust policy to identify and constrain the type of trust that is being brokered. Through this, different security realms are able to federate by supporting the brokerage of trust of identities, attributes, and authentication information between participating web services.

Various implementations of WS-Federation have been put forward. For example, Microsoft, IBM, RSA Security Inc. and various other vendors have implemented this specification and demonstrated interoperability between their implementations through passing a particular identity between six exemplar portals at a workshop organised in May 2004 [WS-FW].

A1.7.3.7 WS-Authorization

A standard for authorization does not exist for web services. In the Microsoft/IBM roadmap for web services security white paper [WSW], an outline for WS-Authorization was loosely described. This document outlined how the WS-Authorization specification would “describe how access policies for a web service are specified and managed. In particular it will describe how claims may be specified within security tokens and how these claims will be interpreted at the endpoint. This specification will be designed to be flexible and extensible with respect to both authorization format and authorization language. This enables the widest range of scenarios and ensures the long-term viability of the security framework”.

However, the WS-Authorization specification has not (yet?) been published. Since this roadmap document was published, developments within the Grid community regarding authorisation and how such infrastructures can be seamlessly linked to Grid services have matured however. As such, from a Grid community perspective, the question may well be asked, what would a WS-Authorization specification offer that can not yet be supported by Grid based solutions and existing authorisation infrastructures?

A1.5.4 Security Assertion Mark-up Language (SAML)

The OASIS SAML specification [SAML1-1] is an XML-based framework for communicating user authentication, entitlement, and attribute information. SAML allows making assertions regarding the identity, attributes, and entitlements of a subject to other entities. SAML has been designed to be a flexible and extensible protocol which can be customised by other standards. For example, the Liberty Alliance, the Internet2 Shibboleth project, and the OASIS Web Services Security committee have all adopted SAML for various purposes.

SAMLv1.0 became an OASIS standard in November 2002. SAMLv1.1 followed in September 2003 and has seen significant success, gaining acceptance across a wide range of domains and is supported by numerous security technology providers.

SAML is defined in terms of assertions, protocols, bindings, and profiles. An assertion is a package of information that supplies one or more statements made by a SAML authority. SAML defines three different kinds of assertion statement that can be created by a SAML authority:



- Authentication: which indicates that the specified subject was authenticated by an identity provider through some means at some given time;
- Attribute: the specified subject is associated with the supplied attributes;
- Authorization Decision: a request to allow the specified subject to access the specified resource has been granted or denied.

SAML defines a number of request/response protocols that allow service providers to request various things. For example, to request one or more assertions from given SAML authorities, or to request that an identity provider authenticate a principal and return the corresponding assertion.

Mappings from SAML request-response message exchanges into standard messaging or communication protocols are called SAML protocol bindings. A SAML SOAP Binding has been defined which outlines how SAML protocol messages can be communicated within SOAP messages. A profile of SAML typically defines constraints and/or extensions in support of the usage of SAML for a particular application. For instance, the Web Browser Single Sign On [WebSSO] profile specifies how SAML authentication assertions are communicated between an identity provider and service provider to enable single sign-on for a browser user. This profile details how to use the SAML Authentication Request/Response protocol in conjunction with different combinations of the HTTP Redirect, HTTP POST, HTTP Artefact, and SOAP bindings.

Other SAML profiles also exist such as attribute profiles which provide specific rules for interpretation of attributes in SAML attribute assertions. For example the X.500/LDAP profile, describing how to carry X.500/LDAP attributes within SAML attribute assertions.

SAMLv2.0 unifies the building blocks of federated identity in SAMLv1.1 with input from the Internet2 Shibboleth initiative and the Liberty Alliance's Identity Federation Framework [LA-IFF]. SAMLv2.0 includes numerous additional features from v1.1 including support for: opaque pseudo-random identifiers (pseudonyms) which can be used between providers to represent principals; identifier management allowing providers to establish and subsequently manage the pseudonym(s) for the principals for whom they are operating; metadata defining how to express configuration and trust related data to make deployment of SAML systems easier; how attribute statements, name identifiers, or entire assertions may be encrypted in SAMLv2.0; attribute profiles which simplify the configuration and deployment of systems that exchange attribute data; support for situations where authenticated users can be automatically logged out of all service providers in the session at the request of the identity provider, and provides mechanisms that allow providers to communicate privacy policy and settings.

The SAMLv2.0 specification was release at the end of September 2005.

A1.5.5 Extensible Access Control Mark-up Language (XACML)

XACML [XACML] is an OASIS [OASIS] standard that describes both a policy language and an access control decision request/response language (both written in XML). XACML version 2.0 was published in February 2005. The policy language associated with XACML is used to describe general access control requirements, and has standard extension points for defining new functions, data types, combining logic, etc. The request/response language allows formation of queries to ask whether or not a given action should be allowed, and interpret the result. The response always includes one of four values: Permit, Deny, Indeterminate (an error occurred or some required value was missing, so a decision cannot be made) or Not Applicable (the request can't be answered by this service).



The typical setup is that someone wants to take some action on a resource. They will make a request to a PEP protecting a resource. The PEP will form a request based on the requester's attributes, the resource in question, the action, and other information pertaining to the request. The PEP will then send this request to a PDP, which will look at the request and some policy that applies to the request, and come up with an answer about whether access should be granted. That answer is returned to the PEP, which can then allow or deny access to the requester. In addition to providing request/response and policy languages, XACML also supports finding policies that apply to a given request and subsequent evaluation of requests against that policy. XACML also allows for generic, distributed policies to be supported. Thus a policy can be written which refers to other policies kept in various remote locations. Hence rather than having to manage a single monolithic policy, different people or groups can manage sub-pieces of policies as appropriate, and XACML supports combination of the results from these different policies into one decision.

XACML comes with a core base language which can be extended. The standard language supports a wide variety of data types, functions, and rules about combining the results of different policies. In addition to this, standards groups are working on extensions and profiles that will hook XACML into other standards like SAML and LDAP, which will increase the number of ways that XACML can be used.

XACML 2.0 and all the associated profiles were approved as OASIS Standards in February 2005.

A1.5.6 Liberty Alliance

The Liberty Alliance [LibAll] is an industry consortium defining standards for federated identity – including enabling simplified sign-on through federated network identification, as well as supporting and promoting permission-based attribute sharing to enable a user's choice and control over the use and disclosure of their personal identification information.

The Liberty Alliance Identity Federation Framework (ID-FF) [LA-IFF] is based on SAML. Recognising the value of a single standard for federated single sign on, the Liberty Alliance submitted their Identity Federation Framework to the OASIS Security Services Technical Committee as input to SAMLv2.0. It intends to use the new version of SAML in concert with its own technical and business guidelines for identity federation going forward.

Liberty's Identity Web Services Framework (ID-WSF) [LA-WSF] provides a platform for communicating identity information among web services and continues to be developed within the Liberty Alliance. The latest version of Liberty ID-WSF now uses SAMLv2.0 assertions as the security token format for communicating authentication and authorization information amongst web service actors.

SAML assertions can be used within SOAP messages in order to convey security and identity information between actors in web service interactions. The SAML Token Profile produced by OASIS specifies how SAML assertions should be used for this purpose with the WS-Security framework. The Liberty ID-WSF builds on these specifications to use SAML assertions for enabling secure and privacy-respecting access to web services.

A1.7.7 Standardised Risk Assessment

Many of the issues of security go beyond security technologies that are adopted and rolled out. Highly secure Grid middleware solutions can easily be made redundant from poorly configured firewalls, web services or general practices. Ultimately no open system (such as those making use of the Grid) can be guaranteed to be secure. There is always the potential knock-on effect when a site is



compromised to collaborators (and collaborators of collaborators) as well as the risk to immediate projects that a given site might be involved in. Risk analysis can be used to better understand, protect and prepare sites for potential security breaches. A risk analysis will normally involve several stages:

- identify all information and resources that needs to be protected;
- identify all sources of risk;
- determine the probability of occurrence of each risk item on each protected item;
- quantitatively and qualitatively assess the likely impact on the sites' business of the occurrence of each risk item on each protected item;
- identify actions that can mitigate the effects of each risk item;
- quantify the cost of implementing mitigating actions.

Once all of these stages have been documented, informed decisions about which mitigating actions to implement for each protected item can be made. For European e-Science digital repositories, standardising such risk assessment procedures is crucial to limit the potential dangers from deliberate or unintentional compromises from users and data providers. This should incorporate both physical security and general working practices to prevent potential theft of equipment/data etc.

It is likely that a body (or bodies) needs to be established that can check and validate that sites meet all appropriate security requirements. Given the divergence of research across the European spectrum it is likely that a variety of domain specific bodies should be established to ensure that appropriate domain specific security measures are taken. Thus post-genomic personalised e-Health related research will have different (more stringent) security requirements than those in the high energy physics domain for example.

A1.6. References to appendix A1

- [SATO] Shibboleth Architecture Technical Overview, <http://shibboleth.internet2.edu/docs/draft-mace-shibboleth-tech-overview-latest.pdf>
- [SAPP] Shibboleth Architecture Protocols and Profiles, <http://shibboleth.internet2.edu/docs/draft-mace-shibboleth-arch-protocols-latest.pdf>
- [UKfed] UK Federation, www.sdss.ac.uk
- [SWISSfed] SWISS SWITCHaai federation, <http://www.switch.ch/aai/>
- [FinnFed] Finnish HAKA federation, <http://www.csc.fi/suomi/funet/middleware/english/>
- [AusFed] Australian Meta Access Management System (MAMS), <https://mams.melcoe.mq.edu.au/zope/mams/kb/shibboleth/>
- [USFed] US InCommon federation, <http://www.incommonfederation.org>
- [PKI] R. Housley, T. Polk, *Planning for PKI: Best Practices Guide for Deploying Public Key Infrastructures*, Wiley Computer Publishing, 2001.
- [X509] ITU-T Recommendation X.509 (2001) | ISO/IEC 9594-8: 2001, Information technology – Open Systems Interconnection – Public-Key and Attribute Certificate Frameworks.
- [TIES] JISC Authentication, Authorisation and Accounting (AAA) Programme Technologies for Information Environment Security (TIES), http://www.edina.ac.uk/projects/ties/ties_23-9.pdf
- [ESP] ESP-Grid project, e-science.ox.ac.uk/oesc/projects/index.xml.ID=body.1_div.1
- [PH] W. T. Polk and N. E. Hastings, *Bridge Certification Authorities: Connecting B2B Public Key Infrastructures*, <http://csrc.nist.gov/pki/documents/B2B-article.doc>
- [JBH] J. Jokl, J. Basney and M. Humphrey, Experiences using Bridge CAs for Grids, Proceedings of UK Workshop on Grid Security Practice - Oxford, July 2004
- [LRA] J. Liddell, K. V. Renaud, and A. De Angeli, *Authenticating users using a combination of sound and images*. HCI 2003, Bath, UK, September 2003.
- [KR] K. Renaud, *Quantifying the quality of web authentication mechanisms: a usability perspective*. Journal of Web Engineering, 3(2):95–123, 2004.



- [SB] S. Booth, *Grid Firewall Recommendations*, <http://www.grid-support.ac.uk/etf/firewalls/Firewalls.html>
- [RAH] A. Richards, R. Allan, D. Hanlon, *Globus Toolkit Firewall Port Selection*, <http://www.grid-support.ac.uk/etf/firewalls/FirewallPortSelection.pdf>
- [BOS] M. Baker, H. Ong, G. Smith, *A Report on Experiences Operating the Globus Toolkit through a Firewall*, <http://esc.dl.ac.uk/Papers/firewalls/globus-firewall-experiences.pdf>
- [MS] M. Surridge, *Rough Guide to Grid Security*, http://www.nesc.ac.uk/technical_papers/RoughGuidetoGridSecurityV1_1a.pdf
- [X812] ITU-T Rec X.812 (1995) | ISO/IEC 10181-3:1996, Security Frameworks for open systems: Access control framework.
- [WSCMP] V. Welch, F. Siebenlist, D. Chadwick, S. Meder, L. Pearlman, Use of SAML for OGSA Authorization, June 2004, <https://forge.gridforum.org/projects/ogsa-authz>
- [OASIS] Organization for the Advancement of Structured Information Standards (OASIS), <http://www.oasis-open.org>
- [XACML] eXtensible Access Control Markup Language TC v2.0 (XACML), <http://www.oasis-open.org/specs/index.php#xacmlv2.0>
- [WS-S] Web Services Security (WS-Security), version 1.0 5th April 2002, www-106.ibm.com/developerworks/webservices/library/ws-secure
- [WS-E] Web Services Eventing (WS-Eventing), www-128.ibm.com/developerworks/webservices/library/specification/ws-eventing
- [WS-N] Web Service Notifications (WS-Notifications), <http://www-128.ibm.com/developerworks/library/specification/ws-notification/>
- [WS-RM] Web Services Reliable Messaging (WS-ReliableMessaging), <http://www-128.ibm.com/developerworks/library/specification/ws-rm/>
- [WS-R] Web Services Reliability (WS-Reliability), http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsm
- [WS-C] Web Services Co-ordination (WS-Co-ordination), <http://www-128.ibm.com/developerworks/library/ws-coor/>
- [WS-Ch] Web Services Choreography (WS-Choreography), <http://www.w3.org/TR/ws-chor-model/>
- [WS-O] Web Services Orchestration (WS-Orchestration), <http://www-128.ibm.com/developerworks/webservices/library/ws-bpelcol2/>
- [WSS4J] Apache WSS4J, <http://www.ws.apache.org/axis>.
- [WS-Policy] Web Services Policy Framework, September 2004, <http://www-128.ibm.com/developerworks/library/specification/ws-polfram/>
- [WS-Trust] Web Services Trust Language, February 2005, <http://www-128.ibm.com/developerworks/library/specification/ws-trust/>
- [WS-Fed] Web Service Federation Language (WS-Federation), <http://www-128.ibm.com/developerworks/webservices/library/ws-fed/>
- [WS-FW] WS-Federation Passive Requester Profile Interoperability Workshop, <http://msdn.microsoft.com/webservices/community/workshops/wsfedprmar2004.aspx>
- [WS-SC] Web Services Secure Conversation Language, <http://www-128.ibm.com/developerworks/library/specification/ws-secon/>
- [WSE] Microsoft Web Service Enhancements (WSE), <http://msdn.microsoft.com/webservices/webservices/building/wse/>
- [WSW] *Security in a Web Services World: A Proposed Architecture and Roadmap*, A Joint White Paper from IBM Corporation and Microsoft Corporation, April 7, 2002, Version 1.0.
- [XMLSig] IETF/W3C XML DSIG Working Group, <http://www.w3.org/Signature/>
- [XMLEnc] W3C XML Encryption Syntax and Processing, W3C Recommendation, December 2002 <http://www.w3.org/TR/2002/REC-xmlenc-core-20021210/>
- [SAML1-1] OASIS, Assertions and Protocol for the OASIS Security Assertion Markup Language (SAML) v1.1, 2 September 2003, <http://www.oasis-open.org/committees/security/>
- [SAML2] Security Assertion Markup Language (SAML) version 2.0, March 2005, <http://www.oasis-open.org/specs/index.php#samlv2.0>



- [LibAll] Liberty Alliance, www.projectliberty.org
- [LA-IFF] Liberty Alliance Identity Federation Framework, <https://www.projectliberty.org/resources/specifications.php>
- [LA-WSF] Liberty Alliance Identity Web Service Framework version 1.1., <https://www.projectliberty.org/resources/specifications.php#box2a>
- [COB] D.W.Chadwick, A. Otenko, E.Ball, *Role-based Access Control with X.509 Attribute Certificates*, IEEE Internet Computing, March-April 2003, pp. 62-69.
- [CO] D.W.Chadwick, A. Otenko, *The PERMIS X.509 Role Based Privilege Management Infrastructure*, Future Generation Computer Systems, 936 (2002) 1–13, December 2002. Elsevier Science BV.
- [OpenSSL] OpenSSL to create certificates, <http://www.flatmtn.com/computer/Linux-SSLCertificates.html>
- [ShibA] Shibboleth Architecture Technical Overview, <http://shibboleth.internet2.edu/docs/draft-mace-shibboleth-tech-overview-latest.pdf>
- [ShibP] Shibboleth Architecture Protocols and Profiles, <http://shibboleth.internet2.edu/docs/draft-mace-shibboleth-arch-protocols-latest.pdf>
- [GT2] Globus toolkit version 2, <http://www.globus.org/toolkit/downloads/2.4.3/>
- [GT4] Globus toolkit version 4, <http://www.globus.org/toolkit/downloads/4.0.1/>
- [EGEE] Enabling Grids for E-science (EGEE) project, public.eu-egee.org
- [gLite] gLite software, glite.web.cern.ch/glite
- [OMII] Open Middleware Infrastructure Institute (OMII), www.omii.ac.uk
- [CROWN] China Research and Development environment over Wide Area Network (CROWN), www.crown.org.cn
- [Condor] Condor software, www.cs.wisc.edu/condor
- [Unicore] UNICORE Forum, www.unicore.org
- [RM] A. Robiette, T. Morrow, *Blueprint for a JISC Production Federation*, JISC Development Group, Version 1.1: issued 27 May 2005, http://www.jisc.ac.uk/index.cfm?name=middleware_documents
- [GridShib] GridShib project, <http://grid.ncsa.uiuc.edu/GridShib/>
- [SSCO] R.O. Sinnott, A.J. Stell, D.W. Chadwick, O.Otenko, Experiences of Applying Advanced Grid Authorisation Infrastructures, Proceedings of European Grid Conference (EGC), pages 265-275, Vol. editors: P.M.A. Sloot, et al June 2005, Amsterdam, Holland.
- [SSW] R.O. Sinnott, A.J. Stell, J. Watt, Comparison of Advanced Authorisation Infrastructures for Grid Computing, Proceedings of International Conference on High Performance Computing Systems and Applications, May 2005, Guelph, Canada.
- [TG] TeraGrid attack, <http://www.washingtonpost.com/ac2/wp-dyn/A8995-2004Apr13>
- [SC] R.O. Sinnott, D.W. Chadwick, *Experiences of Using the GGF SAML AuthZ Interface*, Proceedings of UK e-Science All Hands Meeting, September 2004, Nottingham, England.
- [CHAD] D.W Chadwick, *An Authorisation Interface for the Grid*, Proceedings of UK e-Science All Hands Meeting, September 2003, Nottingham, England.
- [MyProxy] MyProxy Credential Management Service, myproxy.ncsa.uiuc.edu
- [XCO] W. Xu, D. Chadwick, A. Otenko, “Development of a Flexible PERMIS Authorisation Module for Shibboleth and Apache Server”, 2nd European PKI Workshop, University of Kent, July 2005.
- [eduPerson] eduPerson Specification, <http://www.educause.edu/eduperson/>
- [AuthZ2] Prof David Chadwick, JISC proposal, Authorisation Interface V2 for the Grid, June 2005 – accepted for funding.
- [CAS] Community Authorisation Server – <http://www.lesc.ic.ac.uk/projects/cas.html>
- [CAS2] L Pearlman, et al., A Community Authorisation Service for Group Collaboration, in Proceedings of the IEEE 3rd International Workshop on Policies for Distributed Systems and Networks. 2002.
- [GSI] Globus Grid Security Infrastructure (GSI), <http://www.globus.org/toolkit/docs/4.0/security>
- [VOMS] R. Alfieri, et al, *Managing Dynamic User Communities in a Grid of Autonomous Resources*, CHEP 2003, La Jolla, San Diego, March, 2003;



- [STELL] A.J. Stell, *Grid Security: An Evaluation of Authorisation Infrastructures for Grid Computing*, MSc Dissertation, University of Glasgow, 2004.
- [NHSDD] NHS Data Dictionary – www.isdscotland.org
- [HL7] Health-Level 7 (HL7) - <http://www.hl7.org/>
- [SNOCT] SNOMED-CT - <http://www.snomed.org/snomedct/>
- [OEHR] OpenEHR - <http://www.openehr.org/>
- [ICD10] International Statistical Classification of Disease and Related Health Problems (ICD-10), http://www.connectingforhealth.nhs.uk/clinicalcoding/classifications/icd_10
- [ICDb] ICD background, <http://www.connectingforhealth.nhs.uk/clinicalcoding/faqs/>
- [SinnAHM05] R.O. Sinnott, D. Houghton, *Comparison of Data Access and Integration Technologies in the Life Science Domain*, Proceedings of UK e-Science All Hands Meeting, September 2005, Nottingham, England.
- [JDSS] P. Lord, A. MacDonald, et al, *Large-scale data sharing in the life sciences: Data standards, incentives, barriers and funding models (The “Joint Data Standards Study”)*, prepared for The Biotechnology and Biological Sciences Research Council, The Department of Trade and Industry, The Joint Information Systems Committee for Support for Research, The Medical Research Council, The Natural Environment Research Council and The Wellcome Trust.
- [MIAME] Minimal Information About a Microarray Experiment (MIAME), <http://www.mged.org/Workgroups/MIAME>
- [NCBIgeo] Gene Expression Omnibus (GEO), www.ncbi.nlm.nih.gov/geo/
- [ArrExpr] P. Rocca-Serra, A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, S. Contrino, J. Vilo, N. Abeygunawardena, G. Mukherjee, E. Holloway, M. Kapushesky, P. Kemmeren, G. Garcia Lara, A. Oezcimen, S.-Assunta Sansone. *ArrayExpress: a public database of gene expression data at EBI*. *C R Biol*, 326(10-11):1075–1078, Oct 2003.
- [CIBEX] K. Ikeo, J. Ishi-i, T. Tamura, T. Gojobori, and Y. Tateno. *CIBEX: center for information biology gene expression database*. *C R Biol*, 326(10-11):1079–1082, Oct 2003.
- [MGED] Microarray Gene Expression Data Society (MGED) Ontology Working Group, <http://www.mged.org/ontology>
- [MAGEML] MicroArray and Gene Expression Markup Language (MAGE-ML), <http://www.mged.org/Workgroups/MAGE/mage-ml.html>
- [SOFTt] Simple Omnibus Format in Text (SOFTtext), <http://www.ncbi.nlm.nih.gov/projects/geo/info/soft2.html>
- [MINiML] MIAME Notation in Markup Language (MINiML), <http://www.ncbi.nlm.nih.gov/projects/geo/info/MINiML.html>
- [SOFTm] Simple Omnibus Format in Matrix (SOFTmatrix), <http://www.ncbi.nlm.nih.gov/projects/geo/info/soft2.html>
- [LSId] Life Science Identifiers, lsid.sourceforge.net/



Appendix A2: Bibliography

Note: References relating to detailed technical issues and standards are provided in Appendix A1, section A1.6 above.

Lord, P., Macdonald, A., Sinnott, R. et al., Large Scale Data Sharing in the Life Sciences, 2005. Available at: http://www.nesc.ac.uk/technical_papers/UKeS-2006-02.pdf

