Project no. 507618

DELOS

A Network of Excellence on Digital Libraries

Instrument: Network of Excellence

Thematic Priority: IST-2002-2.3.1.12
Technology-enhanced Learning and Access to Cultural Heritage

DELIVERABLE 5.1.1
Report on Digital Repositories:
An evaluation study on the development and implementation of community repositories which support research (and learning and teaching)

Due date of deliverable: June 2006
Actual submission date: August 2006

Start Date of Project: 01 January 2004
Duration: 48 Months

Organisation Name of Lead Contractor for this Deliverable: UKOLN, University of Bath

Revision: 0.0

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)

Dissemination Level: PU

An evaluation study on the development and implementation of community repositories to support research and learning and teaching

Authors:
Rachel Heery, UKOLN, University of Bath (Chapters 1 and 6)
Liz Lyon UKOLN, University of Bath (Chapter 2)
Chrisa Tsinaraki, Technical University of Crete (Chapter 3)
Tim Brody, University of Southampton (Chapter 4)
Traugott Koch, UKOLN, University of Bath (Chapter 5.1)
Martin Doerr, Institute of Computer Science of the **Foundation for Research and Technology, Crete (Chapter 5.2)**

WP5: Knowledge Extraction and Semantic Interoperability
DELOS2 Network of Excellence in Digital Libraries

Table of Contents

1. Introduction (Rachel Heery)

Over the past few years the word ‘repository’ has entered the vocabulary of those engaged in the management and use of recorded knowledge, standing alongside more established concepts such as libraries, archives, catalogues, storage; it joins a set of other less well established concepts such as content management, records management, knowledge management. The context of the discussion about repositories is almost always in terms of digital information, whether born digital or digitised analogue materials.

This DELOS report on repositories first investigates the concept of digital repositories, and provides steps towards defining a typology of digital repositories, with notes towards establishing definitions (Chapter 1). The following sections investigate how repositories relate to the learning and research knowledge cycle (Chapter 2), considerations for their use for e-learning (Chapter 3), and the issues of multimedia content (Chapter 4). Finally a discussion is provided on semantic operability between repositories, including interoperable approach to subject access and integration using the CIDOC Conceptual Reference Model (chapter 5), closing with some conclusions and recommendations (chapter 6).

The growing interest in repositories has emerged from a number of drivers, including the open access movement; serials pricing crisis; preservation requirements; and not least the growing awareness of the importance of the intellectual assets of an institution to that institution's success, particularly in research, but also in teaching. There is a range of technical and organizational issues relevant to repositories, several of these are similar to the issues affecting ‘digital libraries’ and other content management systems. This report also includes consideration of repositories from across different cultures and across different educational and research structures. There are a number of perspectives on repositories in the report building up a snapshot of current interest in repositories from across Europe.

This report has been undertaken as part of the cluster activity of Work Package 5 of the DELOS2 network of excellence work, which focuses on knowledge extraction and semantic interoperability. The authors and contributors draw on the research expertise and experience of a number of organisations (UKOLN, ICS-FORTH, University of Southampton).

1.1 Defining repositories

Digital content collections within the information environment of research, learning and teaching are increasingly referred to as ‘repositories’. In order to encourage meaningful communication across activity areas, and promote interoperability, first we need to be able to define the characteristics of ‘repositories’ distinguishing them from ostensibly similar ways of organising recorded knowledge and, secondly, we need to seek the coherence of a common approach.

Often widespread use of a term goes hand in hand with diversity of meanings. Repositories are ‘collections of digital objects’ - but what makes repositories distinctive from other collections of digital objects such as directories, catalogues, databases? What are the defining characteristics of a ‘repository’? As with other terms that have been popularised in the digital world (e.g. portal, architecture, framework), some qualification is required: is the repository managed as an institutional repository? or a subject repository? What is the content of the repository? What is the underlying purpose of the repository, is it for preservation, access, or data management?

We propose that a digital repository is differentiated from other digital collections by the following characteristics:

- content is deposited in a repository, whether by the content creator, owner or third party
- the repository architecture manages content as well as metadata which describes or otherwise enhances the value of content
- the repository offers a minimum set of basic services such as deposit, fetch back, search, and access control
- the repository is sustainable and trusted, well-supported and well-managed.

Many repositories (though by no means all) support ‘open access’, at least in part. Open-access repositories can be distinguished by the following characteristics:

- the repository must provide open access to its content (except where there are legal or ethical constraints)
- the repository must provide open access to its metadata for harvesting.

The underlying motivations for establishing repositories also differentiate them from other collections. Repositories form an intersection of interest for different communities of practice: digital libraries, research, learning, e-science, publishing, records management, archiving, preservation. Within these communities motivation differs somewhat, and the key services that repositories might provide range over several functional areas:

- Enhanced access to resources
- New modes of publication and peer review
- Corporate information management (as in records management and content management systems)
- Data sharing (such as re-use of research data, re-use of learning objects)
- Preservation of digital resources.

If we consider Lynch's definition of repositories we see an emphasis on the significance of these services rather than on a particular software product or type of content: “a university-based institutional repository is a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members” [Lynch, 2003].

Various combinations of the areas of functionality listed above have been particularly significant for the growth of repositories, such as:

- Improved scholarly communication
- Open access
- Content management.

The intersection of interest across domains offers possibilities for various crossovers of technologies. There is also potential for sharing experience, sharing tools, and undertaking collaborative development work. It is important that there is co-ordination of this activity and that an appropriate level of interoperability is achieved, without placing barriers on innovative work.

1.2 A typology of repositories

The term "digital repository" may be interpreted in very many different ways. As noted interpretations and definitions will vary between communities, and can be viewed from varying stances, such as for functions provided by specific disciplines, organisational context or even by content format (as, for example, in an image repository). Given the heterogeneous nature of repositories noted above and discussed below, one is forced to the conclusion that it is usually necessary to qualify the term to make one's meaning clear, as in "institutional (digital) (image) repository".

This section, building on previous work [Heery, R. and Anderson, S. 2005], explores the question of definition further and sets out an initial typology of digital repositories, viewed from various perspectives. We examine classifications of repositories by:

- content type
- coverage (that is to say the organisational or geographical scope of the collection)
- functionality offered
- The primary target user group, and by
- The level of curation (as explained below).

1) Description by **content** type:

In an academic and research context we may have repositories which collect together one or more of the following:

- Raw research data (and supporting information (such as experimental protocols, process descriptions, intermediate derived data, calibration parameters, etc.)
- Derived research data
- Full-text pre-print scholarly papers
- Full-text peer-reviewed final drafts of scholarly papers
- e-theses
- Full-text original publications (including institutional or departmental technical reports)
- Learning and teaching objects
- Corporate records (staff and student records, licences etc)
- Patent information.

There is some content that currently appears to be largely missing within deployed repositories. For example, within repository deployment there is little evidence of awareness of connections with archival management of courses as opposed to learning objects, what Lynch refers to as "composite structures (such as entire courses

– in various sense, including both course "frameworks" and actual populated "instances" of courses within such frameworks – exported from learning management systems) ... " [Lynch, 2003].

A significant number of entries in institutional repositories for textual content are in fact “metadata only” - there is no link to the full text or other content. This appears to be due to caution regarding copyright and intellectual property rights (IPR). Repository administrators and authors are reluctant to come into conflict with publishers regarding copyright issues, and so will not include ‘full text’ when there is doubt about copyright, yet they wish to have a full record of their publications. In addition some repositories will only include links to full text for those entries published and/or authored whilst the author was employed by the institution. So for example only a percentage of entries within the Southampton University’s ECS e-prints repository (http://eprints.ecs.soton.ac.uk/) link to full text, and the UK’s CCLRC repository has a significant percentage of metadata-only records.

2) Description by **coverage,** in the sense of organisational or geographical scope:

Thus:

- Personal (an author's personal collections)
- Journal (output of a single journal or group of journals)
- Departmental
- Workgroup or project
- Institutional
- Inter-institutional (regional)
- National
- International.

We note that where there is coverage over institutional or legislative boundaries, issues arise concerning determining and agreeing ownership and the rights to access; there may also be questions of allocating responsibilities and liabilities when inappropriate use is made of content. Thus, in a medical research context, questions of access to sensitive information may be interpreted differently over national boundaries, and in repositories containing works authored over more than one centre, agreement must be in place as to who owns what and who can have access to what.

3) Description by primary **functionality** provided by the repository:

Examples are:

- Enhanced access to resources (resource discovery and location)
- Subject access to resources (resource discovery and location)
- Preservation of digital resources
- New modes of dissemination (including new modes of publication)
- Institutional asset management
- Sharing and re-use of resources

- Promotion of organisation (record of output from funding body; record of output from institution).

This dimension is closely allied to the way the repository is conceived to be used, and the degree of interaction and integration of content which is permitted. At one extreme, a repository my be intended as a relatively simple mechanism allowing contributing users to deposit discrete, isolated items (such as a text or image file) and for those accessing the repository to withdraw those items simply as deposited. At the other extreme are repositories which are actively curated, where deposited data is linked to, or otherwise integrated with, other contributions, and is annotated and enriched to various degrees. An example of such a repository is *Ensembl*, hosted by the European Bioinformatics Institute, which provides sophisticated but easy access – for layman to expert - to contributed and integrated genetic information on some two dozen complete genomes (see www.ensembl.org).

4) Description by target **user group**:

The major users fit into two broad categories – contributors and those who access to contributions (the same person can of course act in both roles). Thus we have as contributors and accessors:

- Learners
- Teachers
- Researchers (whether within educational or commercial sector).

Further user groups may include:

- Administrators
- Funding bodies
- The general public (*Ensembl* referred to above is freely available, world-wide, to anyone with a web browser),

Generally the latter may only require read-only access.

5) Other - by **content type** and **institutional context:**

Further dimensions of classification are possible, thus:

Format type: Clearly the technical foundations and functional needs of repositories for, say, video content, still images, genetics information, and texts of e-prints will vary markedly, as may the modes of access to their content.

Similarly the **institutional context** will generate different needs, functionalities and modes of use, reflecting the different roles and demands of, say, museums, libraries, archives, and research institutes.

1.3 Ecology of repositories

How do different types of repository interact? How far is interoperability achievable between repositories of such diverse types?

At present there is very little interoperability between repositories. For example, e-print institutional repositories are unlikely to be linked to or interact with repositories for teaching and learning. Software does not facilitate sharing services between repositories, or provide the full range of functionality that users might require - users in the broadest sense to mean those submitting content, those managing content, and those using content.

Interesting work is now emerging considering interaction between repositories in the context of the digital library and learning community [McLean and Lynch, 2004]. Recent work by Blinco and others seeks to map a repository landscape, and to place some order on the present somewhat confused and fragmented picture [Blinco et al., 2004]. McLean has put forward a tentative 'ecology' of repository services that may help to identify common services and to bring about a convergence of service domains [McLean, 2004].

Within the UK different players are emerging that need to work together, defining their different roles and responsibilities, and exploring ways to interoperate.

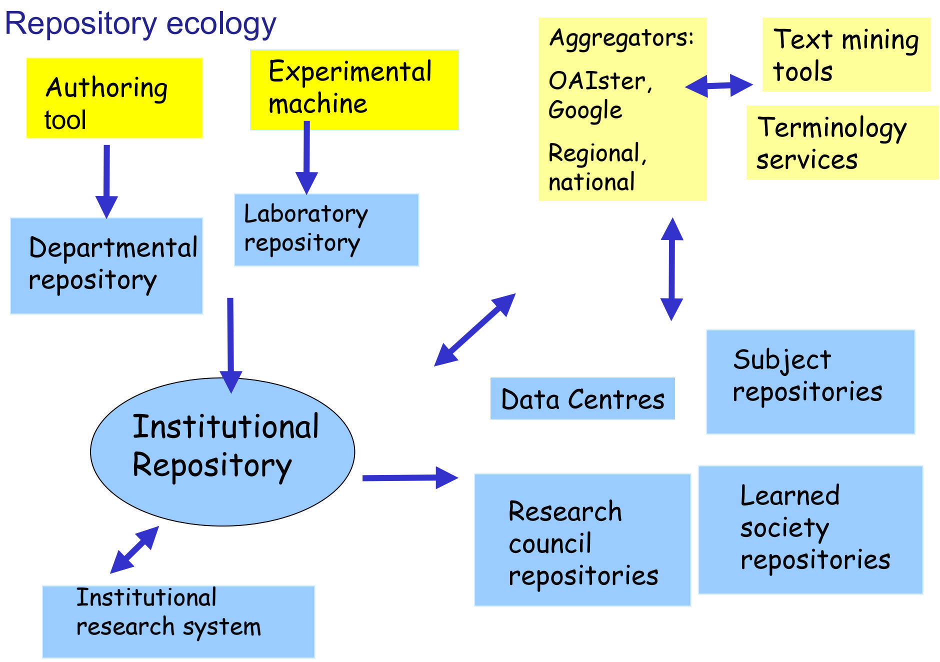


Figure 1: Repository ecology [Heery, JISC/CNI Conference 2006]

A framework needs to be established for repositories that would encompass:

- relation between repositories
- data flow between repositories
- workflow issues.

This would begin to address fundamental questions, such as how institutional repositories relate to thematic, subject repositories? Within institutions, how do repositories relate across the ‘service domains’ of research, learning, administration? A meeting point is required at various levels, both as regards service provision and technical infrastructure.

2. Digital repositories and scholarly communications (Liz Lyon)

2.1. Introduction

The complex process of scholarly communication has undergone some degree of change over the past hundreds of years since the time of the early scholars working with hand-illustrated manuscripts and vellum, but perhaps none as significant as the advent of the digital world, where the outputs of research can now be communicated and disseminated in seconds across the Internet in a highly effective manner. Digital repositories occupy a central position in today's scholarly communication process and this has been illustrated in the notion of a contemporary scholarly knowledge cycle, as shown in Figure 2.

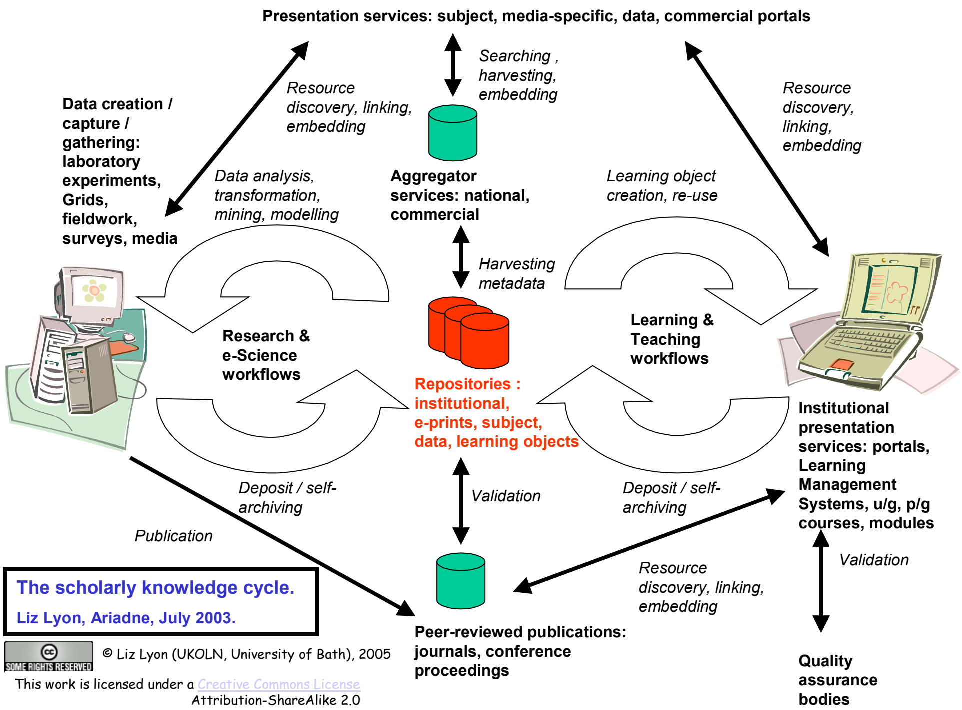


Figure 2: The scholarly knowledge cycle [Lyon, L., 2003]

This chapter of the Delos Digital Repositories Report will examine the provision of open-access (OA) repositories in more detail, describing both the political, organisational and socio-legal aspects of population and deployment, as well as looking at technical issues relating to their operation, management and development within the scholarly context. A case study gives a more detailed description of an OA implementation, which focuses on the dissemination and use of open data as a fundamental element of scholarly communications. This exemplar is particularly pertinent given the emerging trend towards data-intensive science: some fascinating visions of science and research in 2020 have been described in two recent publications [Nature, 2006; and Microsoft Research, 2006].

2.2. Building the infrastructure to support research and learning.

Digital repositories deliver a key part of the research infrastructure but also underpin learning and teaching, providing a source of materials for faculty staff and students. In many cases, repository software solutions based on the Open Archive Metadata Harvesting Protocol (OAI-PMH) have provided the technological foundation for facilitating wider dissemination of (primarily but not exclusively) research outputs through the OA movement, and we now see an increasingly "open" landscape in the scholarly world with some journal publishers such as BioMed Central also adopting the model. The issue of maintaining research quality achieved through a formal peer-review process has been raised by many as a possible down-side to the rapid communications that open-access digital repositories facilitate; however, there are new approaches to the peer-review mechanism [see Nature Publishing's "Nascent" blog page http://blogs.nature.com/wp/nascent/2006/06/natures_peer_review_debate_and.html and http://www.nature.com/nature/peerreview/debate/nature05031.html] that are being evaluated which may lead to more open review models being adopted in due course.

Digital repositories provide some components of the content provision layer within the JISC Information Environment technical architecture shown in Figure 3 below. There is a repository element in the draft Delos Digital Library Manifesto currently being produced by the DL Reference Model Task partners. Common repository services are being defined as part of the international e-Framework activity [www.e-framework.org/] and reference models such as OAIS[1] are being explored within the DLF Service Framework for Digital Libraries [http://www.diglib.org/architectures/serviceframe/dlfserviceframe1.htm] in the USA.

In the context of the deposit of eprints in institutional repositories, there is work ongoing to create a UK-wide e-prints Dublin Core application profile [http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile]. Adoption of this common schema will help to facilitate the provision of a cross-repository search service. Other infrastructure developments include the EU-funded DRIVER project [www.driver-repository.eu] which is a pan-European initiative to develop a network of repositories, the DAREnet [www.darenet.nl/en/page/language.view.home] exemplar from the Netherlands with the associated Cream of Science initiative, and the ARROW Project [www.arrow.edu.au/] in Australia.

[1] Open Archival Information System: developed by the Consultative Committee for Space Data Systems and adopted as an ISO standard, ISO 14721

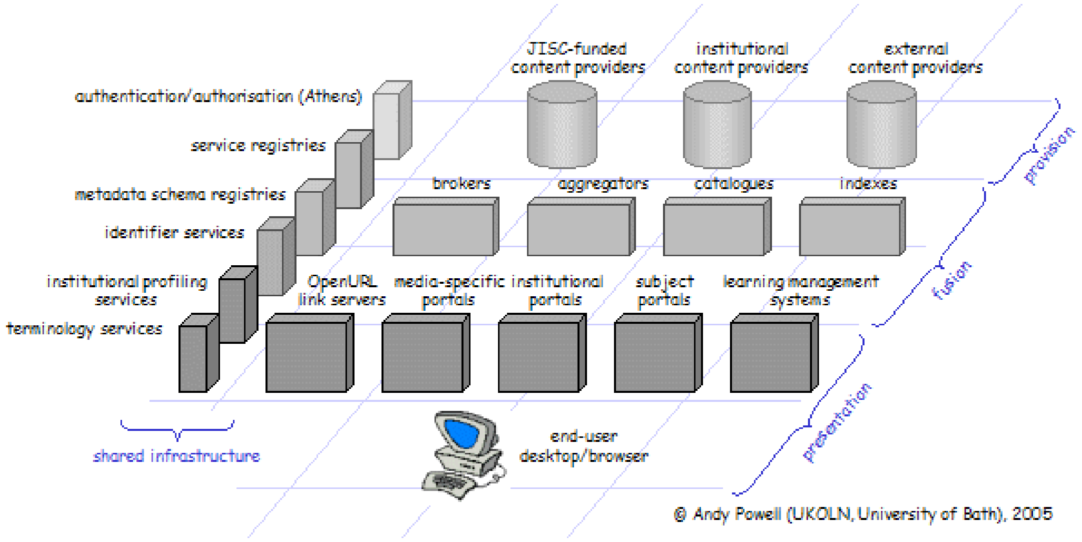


Figure 3: JISC Information Environment Architecture. [Powell, A., 2005]

The case for institutional repositories providing a rapid dissemination mechanism for research outputs has been well-documented in a number of recent reports [Heery, R., and Powell, A, 2006, and Swan, A. and Awre, C., 2006]. The case for institutional repositories as a sustainable mechanism for longer-term access to the scientific record has been less well-rehearsed and the concept of a trusted digital repository with appropriate certification is a current research area. To assist practitioners, the draft RLG/NARA Audit Checklist for Certifying Digital Repositories [http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf] was published in 2005. Following an initial consultation period, revisions are currently underway. The checklist is also being evaluated by test audit projects led by the US Center for Research Libraries and the UK Digital Curation Centre (DCC) [http://dcc.ac.uk] against various repository testbeds. The DCC is also developing a Representation Information Registry [http://registry.dcc.ac.uk/omar] to capture representation information about diverse data types used in different disciplines. The Registry will initially be populated with space science, engineering and crystallographic data types. This work is complemented by the EU-funded CASPAR Project [http://www.casparpreserves.eu/] which will be building a European cross-sectoral framework for digital preservation activity with community exemplars in science, cultural heritage and performing arts. CASPAR has strategic links with the European Task Force on Permanent Access [http://tfpa.kb.nl/], which seeks to bring together national bodies such as the British Library and the KB, the National Library of the Netherlands (Koninklijke Biblioteek), who are leading this initiative. This area is covered in more depth by the Tasks in Delos Work Package 6: Digital Preservation.

2.3. The Open Access landscape: political, socio-legal and cultural issues

Over the past five years there has been significant momentum building behind the OA philosophy, demonstrated by growth in the number of OA journals [Directory of Open Access Journals] (2,293 as at 3 July 2006) and a parallel increase in the number of digital repositories containing OA materials. Currently (July 2006) there are 379 repositories listed in the OpenDOAR Directory maintained by the Universities of Nottingham and Lund [Directory of Open Access Repositories] and 704 archives listed in the ROAR Registry

[Registry of Open Access Repositories]. The content of the repositories ranges from research outputs such as e-prints to resources for learning and teaching. There are disciplinary repositories covering a wide range of subjects across both sciences, social sciences and arts & humanities, and some of these open archives are fully-embedded within the scholarly communication process within that particular community e.g. arXiv [http://arxiv.org/], originally based at Los Alamos National Laboratory, now at Cornell University, serving parts of the Physics, Mathematics, Computer Science and Quantitative Biology communities. There are national repositories such as the JORUM service [http://www.jorum.ac.uk/] funded by the Joint Information Systems Committee (JISC) in the UK and being developed by EDINA and MIMAS, to provide shared resources for learning and teaching support staff. There are also media-centric projects (such as MIDESS [http://www.leeds.ac.uk/library/midess/] -images, and GRADE [http://edina.ac.uk/projects/grade/] – geospatial resources) which are investigating how particular data types can be shared and distributed in open repositories within an institutional setting. The development of open archives for images is particularly interesting given the growth of new services such as Flickr [http://www.flickr.com/] based on Web 2.0 principles and the observation that the current generation of "Net Gen" students are multi-media-based [Lippincott, J., 2006] in comparison with the traditional text-based generation.

The scholarly OA movement is highly complementary to the rapidly-evolving growth in social computing and social software, promoting the principle of a "two-way Web" which facilitates the sharing, use and re-use of freely available content in a variety of ways such as the "mash-up"[2] exemplar of the Avian flu Google Earth prototype developed by Declan Butler at Nature Publishing [See http://www.nature.com/news/2006/060105/full/060105-1.html - doi:10.1038/news060105-1 - and the mash-up itself at http://www.nature.com/nature/googleearth/avianflu1.kml]. The provision of greater volumes of OA content both in digital repositories and in mass digitisation projects (Google, British Library/Microsoft, Open Content Alliance) is building a rich foundation for the development of innovative, "added-value" services and opportunities for the extraction of new information and knowledge from existing distributed corpora. These aspects will be considered in more depth in Section 5 below.

Whilst the continuing growth in global numbers of digital repositories is acknowledged, there is still some considerable reluctance amongst the research community to deposit research outputs into these repositories, particularly in institutional settings: this applies both to e-prints but also to primary data. Policy makers and research funding organisations have begun to address this apparent barrier to OA content provision through a variety of mechanisms. In the UK, in June 2005, the Research Councils (RCUK) published a Position Statement on "access to research outputs" which was open for consultation to the community. There was much input to this process, and one year later RCUK released an updated Position Paper [http://www.rcuk.ac.uk/access/2006statement.pdf] which, whilst supporting the principle of self-archiving, leaves some ambiguity in practical direction, since the individual Research Councils have not adopted a uniform position on the topic. These positions and policies are described as part of the SHERPA JULIET service [http://www.sherpa.ac.uk/juliet/]. The Australian Research Council is also considering a requirement for grantees to deposit their published articles arising from grants in an OAI-PMH repository. In the USA, the House of

[2] A "mash-up" is a Web page or application that integrates elements from two or more sources.

Representatives has directed the National Institutes for Health to adopt an OA mandate, and support from the Senate Committee is now being sought to expedite the process, thereby opening the possibility of other publicly funded research following the OA route.

At the institutional level, five organisations have mandated self-archiving in institutional repositories with a published policy listed on ROARMAP: the Universities of Southampton, Minho and Zurich, Queensland University of Technology, and CERN [http://www.eprints.org/openaccess/policysignup/sign.php]. In addition, model policies ("strong and weak") are posted on ROARMAP which have been drafted collaboratively for national and private research funding organisations and are based on the Wellcome Trust Self-Archiving Policy [http://www.wellcome.ac.uk/doc_WTD002766.html]. These political incentives are supplemented by the recent JISC announcement in March 2006 of a £14 million Capital Funding Programme on Digital Repositories and Preservation which will include an allocation against matched funding for UK institutions wishing to set up and develop institutional repositories.

Aside from basic awareness of the OA philosophy, one of the main barriers to the deposit of research papers in institutional repositories is concern about legal and intellectual property rights (IPR) issues. The RoMEo Project [http://www.lboro.ac.uk/departments/ls/disresearch/romeo/index.html] investigated appropriate rights metadata to describe research outputs in repositories and collated journal publisher copyright and self-archiving policies in partnership with SHERPA [http://www.sherpa.ac.uk/romeo.php]. There is clearly still much awareness and advocacy work to be carried out working with the research community to increase the volume of deposited materials. The Rights & Rewards Project [http://rightsandrewards.lboro.ac.uk/index.php?section=1] is exploring motivational incentives and support issues in the learning and teaching context and is gathering user requirements information and community cultural views on sharing resources in a major national survey. Open licensing initiatives such as the Science Commons and its funded Scholar's Copyright Project [http://sciencecommons.org/literature/scholars_copyright] are assisting faculty members by providing standard copyright agreements based on Creative Commons licences, which authors can use when making their work more openly available.

2.4. Institutional repositories: examining their role in the scholarly workflow

In order to engage the research community – and indeed the teaching and learning communities - and encourage deposit, we need to acquire a much better understanding of the research process within institutional faculties and departments. We need to gain more detailed information about the specific steps, tasks and activities undertaken by scholars when initiating and carrying out their research: we must be able to describe and document workflows and be aware of inter-disciplinary differences [Lyon, L., March 2006]. Traditionally scholars in the arts and humanities have tended to work in a relatively independent fashion - the "lone scholar" approach, whilst in the pure and applied sciences for example, there has been a more collaborative mode of working. This latter mode is now becoming the norm across most disciplines supported by collaborative technologies, grid computing and a focus on interdisciplinary research themes. Three JISC-funded projects, StoRE [http://jiscstore.jot.com/WikiHome], RepoMMan [http://www.hull.ac.uk/esig/repomman/] and R4L [http://r4l.eprints.org/about.html], have begun to gather information through surveys and questionnaires and are developing use-cases built upon user experience. RepoMMan is seeking to build a workflow tool based on BPEL (business process execution language) and Web Services, for the institutional repository at the University of Hull, which has implemented the FEDORA platform. The initial outcomes of the RepoMMan survey suggest that there are a wide range of file and format types used by the research community, there are storage, access, management and preservation aspects to the workflow and that the development of automated tools is required to manage these file-handling processes. A three-tier technical model is forming the basis for software development. The Repository for the Laboratory (R4L) Project is considering scientific workflow in the domain of crystallography where highly complex laboratory processes must be identified, mapped and described in order to be able to build repository services to assist deposit. In the case of R4L, primary research data (molecular structures of crystals) are being deposited in an open access institutional repository. This is described in more detail in the case study in the next section.

Other aspects of the scholarly workflow which are being explored include managing versions of documents (e-prints). A research paper develops through multiple versions, which undergo revisions, sometimes many. This part of the workflow needs to be effectively managed when digital repositories are used as the container for one or more versions of the pre-print. Some recent data from arXiv [Simeon Warner, JISC Repositories discussion list, 16 June 2006 posting] show that more than 30% submissions have more than one version; all versions have been stored since 1997 and the majority of accesses are to the latest version. Initial (and incomplete) results from a survey carried out by the Versions Project [http://www.lse.ac.uk/library/versions/] suggest that more than 90% researchers at some time find multiple versions of articles online, but almost 40% users do not find it easy to identify which version to read. Aside from the usability issues, there are implications for the ability to demonstrate provenance and assure longer-term preservation of the scientific record as tracked through the intellectual development of a research article.

One further aspect of research workflow relates to managing administrative information about grants and contracts, funding, project details and associated research outputs. Some of this information resides in digital repositories but much of the data currently resides in management information systems (MIS) or in dedicated Current Research Information Systems (CRIS). The euroCRIS [http://www.eurocris.org/en/] network is keen to explore how repositories and CRIS can be inter-linked in a way exemplified by CCLRC in the UK with their corporate data repository based on the CERIF Common European Research Information Format standard.

2.5. Data repositories: a case study in crystallography

Whilst there is a growing body of work relating to experience with institutional e-print repositories, there is rather less evidence describing practical examples of institutional data repositories and their role in scholarly communications. One initiative which has a working repository, together with an added-value aggregation service running across the repository, is eBank [http://www.ukoln.ac.uk/projects/ebank-uk/], a JISC-funded project led by UKOLN at the University of Bath with the Universities of Southampton and Manchester in the UK. The eBank activity complements the R4L Project mentioned earlier and also the Smart Tea initiative [http://www.smarttea.org/], which is exploring workflow in a "smart lab" in the chemistry domain.

The evolving e-research environment is characterised by Grid-enabled applications where there may be automatic data capture by analytical or other observational equipment together with appropriate metadata. This is supplemented in chemistry by the use of electronic laboratory notebooks capturing other experimental parameters, metrics and environmental data, by the scientist. In the crystallography workflow, the chemical sample to be synthesised is assigned a chemical identifier, a record is created in the laboratory management system, spectral and other analyses are run on the sample capturing relevant experimental metadata, outputs (spectra) are captured, deposited and stored in the laboratory repository. When all analyses are complete, data is collected, corrected, processed and reduced to a standard format (a Crystallographic Information File or CIF), to inform the creation of a human-readable crystal structure report which is deposited in the institutional eCrystals repository [http://ecrystals.chem.soton.ac.uk/] at Southampton. The crystallography workflow will be described in full in a forthcoming paper to be presented in the Digital Curation Centre/eBank Data Curation Mini Workshop to be held at the UK eScience All Hands Meeting, Nottingham, September 2006 [http://www.allhands.org.uk/programme/index.html].

The eCrystals repository uses the eprints.org software which has been modified to support data. Metadata describing the crystal structures is exposed by OAI-PMH as Dublin Core records for harvesting by the prototype aggregator service [http://eprints-uk.rdn.ac.uk/ebank-demo/] developed and located at UKOLN. The aggregator is based on the e-Prints UK service which was developed using the Cheshire software platform. The aggregator service has been embedded in a learning portal, PSIgate, which is part of Intute, the Resource Discovery Network, enabling PSIgate to be used by students to discover digital resources. A pedagogical evaluation study is underway to gain some preliminary evidence about the potential application and benefits of exposing students to primary research data as part of their online learning course.

Each crystal has been assigned a persistent identifier (a DOI issued by the DOI registry for datasets at the National Library for Science & Technology TIB, University of Hanover, Germany), and a domain identifier (the International Chemical Identifier or InChI). With this functionality, a dataset can be cited using a standard format - *http://dx.doi.org/10.1594/ecrystals.chem.soton.ac.uk/145* - (eBank has a provisional data citation policy), and can be discovered using Google [Coles, S.J., Day, N.E., Murray-Rust, P. et al.]. Datasets also have links to derived research publications to demonstrate proof-of-concept, added-value linking capability. Underpinning the data repository there are an eBank data model and a metadata schema application profile for describing crystal structures, which have been published on the project Web site.

Semantic interoperability within the crystallography domain has also been investigated as part of the eBank work and the issues will be documented in a forthcoming report. Given the e-research tenet of enabling inter-disciplinary science, there are significant challenges associated with the best use of name-authority files, keywords, cataloguing terms and descriptors. Future technical challenges will include comparing data outputs from different laboratory workflows, consideration of the metadata schema used to describe crystal data in the distributed repositories, and requirements for metadata normalisation services to facilitate harvesting by aggregator services.

Whilst there is much further work to pursue, eBank has begun to demonstrate the full scholarly knowledge cycle through linking the primary data outputs from the research process with the activities and tasks carried out during the learning and teaching experience.

2.6. The native digital scholar: developing repository services for added value and knowledge extraction

The eBank project has begun to demonstrate how value can be added to repository content through building innovative services. If we look to the future, where the "native digital scholar" is working in a complex landscape of federated heterogeneous content repositories, discovering, using and re-combining data and information from a range of resources, we can start to identify the types of repository services we might require in order to satisfy the demanding workflow requirements of the e-researcher.

An ongoing study investigating resource discovery services [Ferguson, N., 2006] has indicated that the majority of scholars use Google to discover and locate resources: we therefore need to ensure that scholarly content within digital repositories is exposed to search engines like Google Scholar. We also need to consider assigning tags to digital repository content to ensure discovery through services such as Technorati. The Connotea TaggingTool [http://www.connotea.org/taggingtool] with eprints.org software facilitates the tagging of repository content by any bookmarking or tagging service using the Connotea code: this includes the del.icio.us service (a book-marking service for Web users at http://del.icio.us/).

In the case study, additional services might include the ability to annotate published crystallographic data and 3D structures, the application of visualisation tools to facilitate the presentation of large and complex data sets, and text mining tools to enable the mining of the significantly large chemistry corpora such as Chemical Abstracts, to extract connections and associations between the data and the intellectual interpretations of experiments and simulations. The UK Text Mining Centre NaCTeM [http://www.nactem.ac.uk/] is developing a range of tools for annotation, term mapping and analysis, and information extraction, which may be applied in this context.

In other disciplines such as astronomy where datasets are extremely large and homogeneous, data mining algorithms will operate across distributed repository content to identify anomalies within the data which might lead to the discovery of new stars and other astronomical features [Gray, J., 2002, on data mining for the Sloan Digital Sky Survey http://research.microsoft.com/research/pubs/view.aspx?msr_tr_id=MSR-TR-2002-01]. Tools to support these transformations and analyses will need to be part of the scientist's workbench. The eScience myGrid [http://www.mygrid.org.uk/] project is developing workflow-based processes and services in the biosciences domain using the Taverna workflow tool: there is clearly much to be gained by bringing the "Grid" community, digital library and domain scientists together to explore synergies and common development paths, and to facilitate knowledge transfer.

Returning to the crystallography exemplar, it is planned to extend the eBank work in a third phase which will begin to scope a federated landscape where a global network of heterogeneous distributed repositories stores crystallographic primary-research data captured from "smart" laboratories. The repositories will include services from institutions, national initiatives, commercial publishing organisations and learned societies. In this complex landscape, there will be a range of aggregator services (some generic, some domain-centric), harvesting metadata describing repository content. There will be various third-party portal services providing wider access to crystallographic data. Data centres and data banks will offer data curation and preservation services to the community. This federation model is shown in Figure 4.

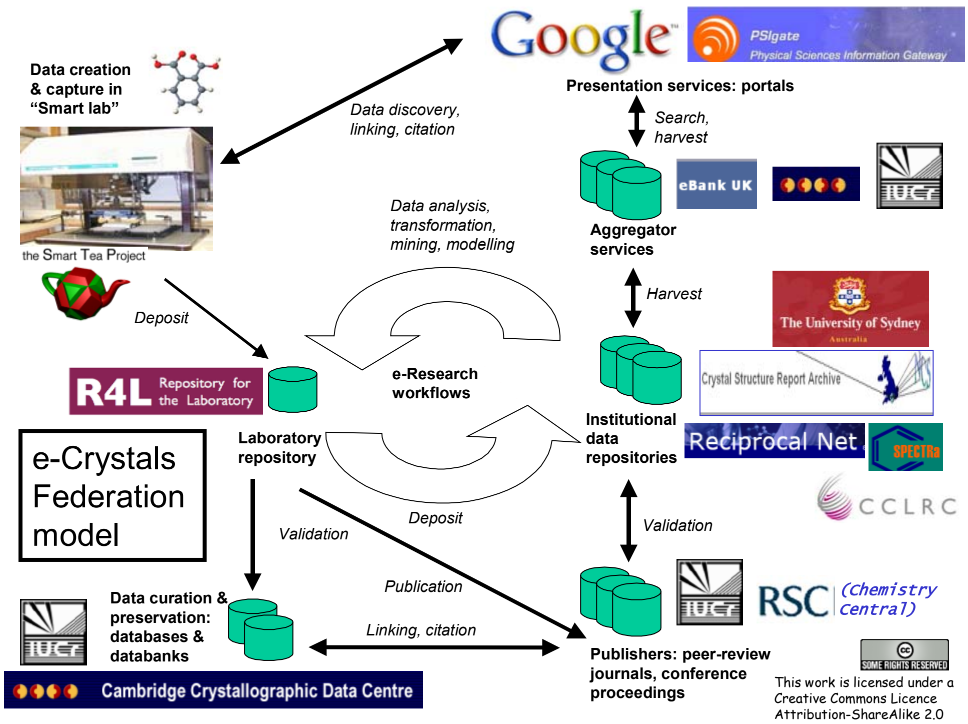


Figure 4. The e-Crystals Federation Model [Lyon, L. 2006 SCONUL]

A range of publishing organisations will make secondary, peer-reviewed content available: some open access, some based on the author-pays model. Data will be integrated with textual interpretations in new ways and in new types of publication. This is already evident with journals such as the Molecular Systems Biology journal, where structured data is published based on an open standard (Systems Biology Markup Language, SBML), and is integrated with textual interpretation. This publication is the result of a partnership between EMBO, the European Molecular Biology Organisation, and the Nature Publishing Group.

The landscape envisaged makes a step towards a global research infrastructure of repositories supported by a suite of shared services and sophisticated data manipulation tools and "utilities" available to the "native digital scholar". The content of data repositories will be heterogeneous, with certain data collections aspiring to become "reference collections" in the nomenclature described in the National Science Board Report on long-lived data collections [2005]. In the future, we can envisage these data collections embedded within our notions of infrastructure, i.e. reference data sets are key elements of the research infrastructure, and they become the accepted starting point for new scientific research, discovery and innovation.

3. e-Learning repositories (Chrisa Tsinaraki)

The crystallography data case study presented in the previous section highlights the enormous widening of research horizons thanks to the availability of the data repository, together with tools and utilities. But digital repositories are also highly relevant for teaching and learning, to support the re-use of materials, increasing teaching efficiency, and also to support new means of pedagogic delivery.

One manifestation of the widened form of delivery is the use of podcasts to deliver lectures, highlighted in an article earlier this year in the Financial Times [Knight, R., 10 February 2006] on the use of podcasts in universities, enabling lecturers to provide a podcast version of lectures, with the time released made available to students on a one-to-one basis.

This application area – teaching and learning - has been developing its own standards and approaches on the e-learning side for some time, and some of these are introduced below. The objects that are managed in learning environments - "learning objects" - are complex; a large and increasing proportion of learning objects are multimedia files, with specific issues which we discuss in chapter 4 below. Once again, a point of crucial importance is interoperability: that items held in digital libraries and in digital repositories can be interoperable with learning applications and standards and that the libraries and repositories can smoothly integrate complex objects.

As practice and understanding evolve, a number of different definitions have emerged for learning objects. This section first explores these definitions, to derive key features of learning objects for their accessibility, usefulness and for their management in a digital repository. We then touch on curation and rights management, before considering the architectures of learning systems, again with a view to considering the accessibility and retrievability of learning objects. Appendix I lists the major educational digital repositories containing collections of learning objects. As Casey et al. point out [Casey, J., Proven, J., Dripps, D., 2005], there are few centralised, shared collections of teaching materials within institutions. One exception is the UK's Jorum repository initiative, which we present briefly in 3.2 below.

3.1 Definitions: e-learning repositories and learning objects

According to the IEEE-LTSC[3] learning object metadata standard [http://ltsc.ieee.org/wg12/], a learning object (LO) can be defined as "any entity, digital or non-digital, that may be used for learning, education or training". This rather broad definition has been actively criticized by at least one important author, David Wiley from the Utah State University, who believes that learning objects are elements of a new type of computer-based instruction grounded in the object-oriented paradigm of computer science [Wiley, D.A., 2001]. Object-orientation values highly the creation of components (called "objects") that can be re-used in multiple contexts. This is also the fundamental idea behind LOs: instructional designers can build small (relative to the size of an entire course) instructional components that can be re-used a number of times in different learning contexts. In addition, learning objects are generally understood to be digital entities stored in some repository system and deliverable over an intranet or the Internet, meaning that a number of people can access and use them

[3] Institute of Electrical and Electronics Engineers Learning Technology Standards Committee

simultaneously (as opposed to traditional instructional media, such as an overhead or video tape, which can only exist in one place at a time). Moreover, the users of learning objects can collaborate with and immediately benefit from new versions. These are the significant differences between learning objects and other instructional media.

Vladiou [Vladiou, M., 2003] believes that learning objects also have to be stand-alone, discoverable, able to be aggregated, interoperable, based on a clear instructional strategy, be interactive and tagged with metadata. Moreover, they are expected to provide a learning experience which is just enough, just in time and personalized ("just for you"). This view was also expressed on the website of the project Reusable Learning Objects (RLO), a project at the Tropical North Queensland (TNQ), Technical and Further Education (TAFE) Institute, Australia. For TNQ-TAFE team members learning objects are:

- **A new way of thinking about learning content** – traditionally, content comes in a several hour chunks called "a course". Learning Objects are much smaller units of learning, ranging for example, from 2 to 15 minutes.
- **Small, independent chunks of knowledge or interactions** – stored in a database can be presented as units of instruction or information.
- **Based on a clear instructional strategy** – intended to cause learning through internal processing and/or action. In order to be defined as a Learning Object there must be some intrinsic instructional value related to a knowledge or information object.
- **Self-contained** (stand-alone) – each learning object can be taken independently. Like the Lego toy, RLOs are small stand-alone, re-usable components – video, demonstrations, tutorials, procedures, stories, assessments, simulations, case studies, HTML/text pages etc. that can be assembled to provide resources for education and training. Objects have a defined level of granularity which means they can stand alone as single items or be combined (aggregated) with other objects to form larger instructional units. Some metadata standards, such as SCORM (Sharable Object Content Reference Model) 0 and IMS Learning Resource Metadata Specification IMS [http://www.imsproject.org/metadata/ and http://www.adlnet.gov/scorm/index.cfm , and see examples in 3.2 below] allow for metadata description at various levels of aggregation such as the resource, item and LO level.
- **Discoverable – tagged with metadata** – Learning objects must be easily discoverable. LOs must be tagged with appropriate descriptive metadata in order to be identified for the purpose they are to be put. Whilst the technical description of objects is being addressed via the various metadata schemas that have been developed, there are increasing calls from educators for the metadata schemas to improve the pedagogic description of learning objects.
- **Interactive** – each learning object requires that students view, listen, respond or interact with the content in some way.
- **Re-usable** – a single learning object may be used in multiple contexts for multiple purposes. The main idea of Re-usable Learning Objects (RLOs) is to break educational content down into small stand-alone chunks that can be re-used in various learning environments, in the spirit of object-oriented programming.

- **Able to be aggregated** – learning objects can be grouped into larger collections of content, including traditional course structures. We might add that objects can be linked together.
- **Interoperable:** - learning objects must be interoperable. That is, content from multiple sources must work with different learning systems. In order to do this they must be designed according to appropriate standards.
- Meant to let learners **have learning facilitites** that are:
  - **Just enough** – if learners need only part of a course, they can use only the LOs they need.
  - **Just in time** – LOs are searchable, thus learners can instantly find and take the content they need, when they need it.
  - **Just for you** – learning objects allow for easy customization of courses for a whole organization or even for each individual.

The use of learning objects has the potential "to provide learning customized for each specific learner at a specific time, taking into account, their learning styles, experience, knowledge and learning goals" [Schatz, S., 2000]. In addition, learning objects may "offer great value in terms of saving time and money in course development, increasing the re-usability of content, enhancing students' learning environment, sharing knowledge within and across disciplines, and engaging faculty in a dynamic community of practice" [Metros, S. and Bennett, K., 2002].

3.2 Curation and rights management

The entry in the list above on the need for metadata-tagging for discoverability of learning also flags the increasing calls from educators for metadata schemas to improve the pedagogic description of learning objects. Further issues arise here. The enhanced metadata schemas may provide for some automated provision of pedagogic description, but human input is still likely to be needed. This enhancement of metadata description is a form of data curation, as illustrated for instance in the examples in section 2.6 above, and initiatives in the area of research data curation may be usefully explored in the context of repositories holding e-learning materials, for models of provision and pointers to tools and utilities to support this curation.

This curation mark-up is likely to be increasingly important over time. The logical assumption is that re-use of a learning object at a different juncture will be in a different context – perhaps only slightly different initially, but as the time distance increases, so the context around the object will have changed. The way the object is re-used is likely to be of pedagogic value to future users, and the repository and its tools should support the means of capturing and making that information available.

One element of curation identified in the UK's e-Science curation report [Lord, Macdonald 2003] is that of deletion of materials. This is a more complex and sensitive area than first meets the eye – for instance, who selects? What processes are there to ensure that repository materials are not deleted prematurely? The life-cycle of e-learning objects in repositories is a likely area for future research, with cost/benefit issues likely to be a major area for consideration.

A further point here is that research data objects in repositories can be used in teaching as well, and thus become e-learning objects as well as research objects. This is an area for further research, to examine the inter-operability of the metadata mark-up structures and tools of research-focused and learning/teaching focused repositories.

The features listed in 3.1 above do not include information about the rights relating to learning objects. A recent study on data sharing [Lord, Macdonald et al.] showed that recognition of rights (attribution of authorship for example) was one of the critical factors for people to make their data (here, learning objects) available to others.

This aspect is addressed in the UK's JORUM service [www.jorum.ac.uk/index]. Jorum "is a free online repository service for teaching and support staff in UK Further and Higher Education Institutions, helping to build a community for the sharing, re-use and re-purposing of learning and teaching materials". As with many other types of organisation (including companies) the question of ownership of materials created by one or more individuals when working for an organisation is complex, and this is reflected in the current arrangement for deposit of materials in the Jorum repository: Where a material is definitely owned by the institution, deposit of material into the Jorum repository is straightforward, covered by the Jorum Deposit Licence Agreement. However, where the material is definitely owned by the individual (rather than by the institution), the individual is asked to licence the material to his/her institution, and ask the institution to deposit the materials in Jorum. Though cumbersome, this brings the owner/licensor the benefit of the indemnity provided by the depositing institution. JISC is currently working on a model licence to support individual deposit of materials. Institutions sign a separate, user licence for access to materials held in Jorum.

The Financial Times article referenced above [Knight, R., 2006] raises another IPR question: who owns "classroom content" (now captured in podcasts, for example). In this context, and also with reference to the Apple Learning Interchange presented in Appendix 1, Apple Computer introduced a service at the beginning of 2006 that enables select universities to put course lectures and other educational material online, using Apple's iTunes software. In the case of those lecturers at Harvard University who use the service, their lectures are immediately available to their own students via a password-protected link and later released to the public, free of charge.

3.3 Architectures of learning systems

The architecture of a learning system is a major issue affecting the spectrum of offered services, the performance, the adaptability and the usability of the system. The objective of this subsection is to present major architectural frameworks that are investigated by state of the art learning systems. These architectural approaches, fall into two types: Peer to Peer (P2P) and Service Oriented Architectures (SOAs). The architectures enable re-usability of educational material and are concerned about making learning objects (LOM) available in repositories on the Web, in a way that they can be accessed and used by other systems.

The Learning Framework (Naeve et al., 2001)

A learning framework was suggested in the proceedings of the second European Web-Based Learning Environments Conference [Palmir, M., Naeve, A., and Nilsson, M., 2001]; in this framework, services can be developed and exchanged between, as well as within, systems.

An important part of this framework is the semantic layer, which builds on the structure of the Semantic Web. Hence they do not regard metadata as something "objective" that one has to download from a central server. On the contrary, metadata should be allowed to consist of subjective views of resources that are distributed and shared in contexts that can evolve dynamically.

In order to support such requirements, the proposed learning framework consists of a combination of Semantic Web techniques and peer-to-peer services for search, retrieval, publication, replication and mapping of metadata. The framework developers focus on interoperability support that is data-based integration using XML files or records in relational databases to share information. They also think of standards (e.g. SCORM) as either schemas in isolation or schemas together with the services that make use of them in order to accomplish different tasks. Moreover, they believe that a learning framework should be layered but with well-defined interfaces only for the lowest (physical and transport) layers. The semantic layer, which comes next, should serve as the "middle ground" for the different sub-systems in the service and application layers above. Figure 5 depicts some of the suggested layers in this learning framework. The service layer includes services that solve problems such as profiling and resolution of resources. Several more advanced services will need active cooperation from other peers in order to work properly, e.g. search for metadata, definition of mappings between schemas and publication of content/metadata. Since not all the services and the applications will be able to use RDF and schemas to express themselves, there will always be a need for export information in XML or special application programming interfaces (APIs). Some services will deal with this explicitly.

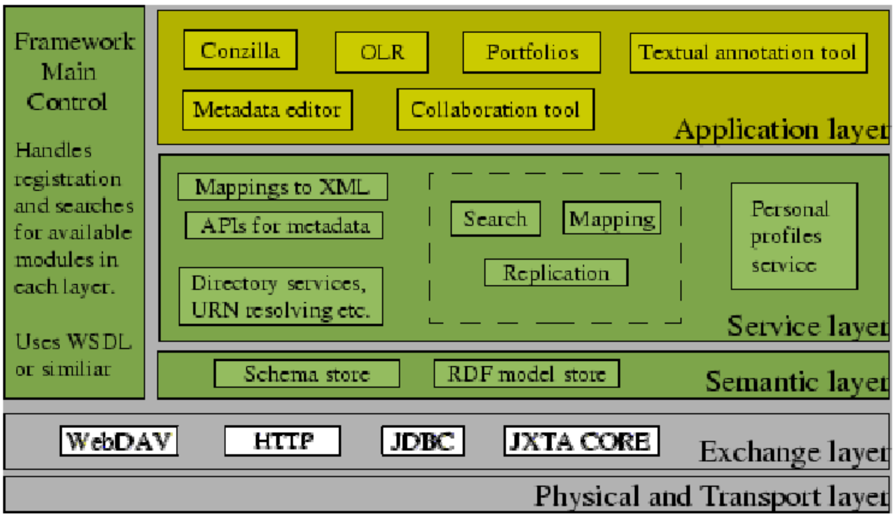


Figure 5: Suggested layers of the Learning Framework (Naeve et al. 0)

The Courseware Watchdog [Wetterling, 2003]

The Courseware Watchdog [Tane, J., Schmitz, C., et al., 2003] is an integrated architecture where ontologies are used to help find and organize distributed courseware resources by offering a common framework for the retrieval and organization of courseware material. The diverse tasks that are supported by a courseware watchdog are:

- Understanding the ontology of the content and browsing the content
- Querying semantically annotated resource repositories
- Retrieving relevant material through crawling other other LO systems
- Organizing the documents according to the ontology
- Updating the ontology and the knowledge base according to the current data.

The Courseware Watchdog addresses these requirements using a comprehensive approach which exploits concepts from the Semantic Web, such as ontologies, in an E-Learning scenario. It is part of the PADLR framework (Personalized Access to Distributed Learning Repositories) which builds upon a peer-to-peer approach for supporting personalized access to learning material. The Courseware Watchdog components are illustrated in Figure 6. All the modules presented in this figure are built on top of an ontology framework named KAON4. As also illustrated in the figure, the Courseware Watchdog includes the possibility to access the Edutella [peer-to-peer network (Edutella is presented below). The Courseware Watchdog addresses the different needs of teachers and students to organize their learning material. It integrates the Semantic Web vision using ontologies and peer-to-peer networks of semantically annotated learning material. Moreover, it addresses the important problems of finding and organizing material using semantic information. Finally, it offers primitive solutions to the problem of evolving ontologies.

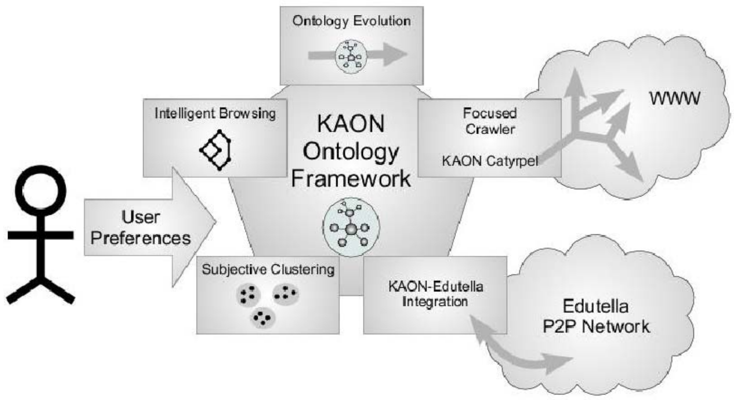


Figure 6: Components of the Courseware Watchdog (Tane, J. et al 0)

The ELENA Project

The work of the ELENA Project is described in papers presented by Simon and others at the World-Wide Web conference in Budapest in 2003 [Simon, B, et al., 2003, and Simon, B.,

[4] KAON is an open-source ontology management infrastructure targeted for business applications. It includes a comprehensive tool suite allowing easy ontology creation and management and provides a framework for building ontology-based applications. An important focus of KAON is scalable and efficient reasoning with ontologies. For more information see: http://kaon.semanticweb.org.

Miklos, Z. et al., 2003]. The design of a learning services[5] mediation infrastructure is discussed, called Smart Spaces for Learning. The architecture of Smart Spaces for Learning is based on P2P and Semantic Web technologies. Smart Spaces for Learning are defined as service mediators, which support the personalized consumption of heterogeneous educational services[6] provided via assessment tools, learning management systems, (meta) repositories of educational material and live delivery systems such as video conferencing systems. Smart Spaces for Learning are built on top of learning management networks. Learning management networks connect systems like the ones mentioned above and provide an infrastructure for the provision, booking and consumption of educational services.

In Smart Spaces for Learning, learners can choose learning services from heterogeneous sources. Users maintain their own learning profiles, where learner background data such as job titles, hobbies, previous learning events as well as demographic data such as location are stored. Learning profiles are also dynamically updated with the learning progress. Learning services are complete entities designed for a specific purpose and targeted at a specific audience. Mediation of learning services requires matching the goals of the prospective learner with the educational objective addressed by a learning service. Ontologies have been identified as one of the most important ingredients in distributed, heterogeneous (and, especially, Semantic Web) applications. They enable the various actors in these applications to communicate with each other on a high level of abstraction. In ELENA, these actors are learning service providers which are realized as peers in a P2P network. The ontologies are represented with the help of Semantic Web ontology languages like RDFS, DAML+OIL, and OWL and are based on existing or upcoming standards like Dublin Core, IEEE LOM, WSDL, DAML-S etc. Figure 7 depicts the various components of a Smart Space for Learning. The Personal Learning Assistant performs the search for suitable learning services based on the learner's individual profile, processes the selected services and supports the evaluation of a learner. In a Smart Space for Learning, providers of educational services are connected within the learning management network, which is based on Edutella [Nejdl, W. et al., 2002].

[5] A **learning service** is defined here is an event that is provided by a learning service provider in order to support the accomplishment of a specific learning objective. This is achieved by creating a learning environment consisting of learning resources, Communication Infrastructure, Meeting Places, etc. Learning services are most likely concerned with various functions of instruction such as motivating learners, re-calling learners' pre-existing knowledge, conveying learning content, exercising, and learner assessment. Learning services can result in a predefined type of outcome (grade, certificate, degree, etc.) and sometimes require specific prerequisites to be fulfilled before a learner is allowed or recommended to consume a learning service. Examples for a learning service are the delivery of a course, the provision of a web-based training application or self-study material. Educational activities such as distributed classroom sessions, training units and tutoring sessions are by default considered as learning services, regardless of their level of IT-support. The provision of material that is not designed as self-study material does not meet the requirements of a Learning Service, but can be considered a Learning Resource.

[6] **Educational services** comprise learning services and services that supplement learning services. An educational service can be used in order to prepare or control learning services. For example, a brokerage service for educational material can be used for preparing the delivery of a course.

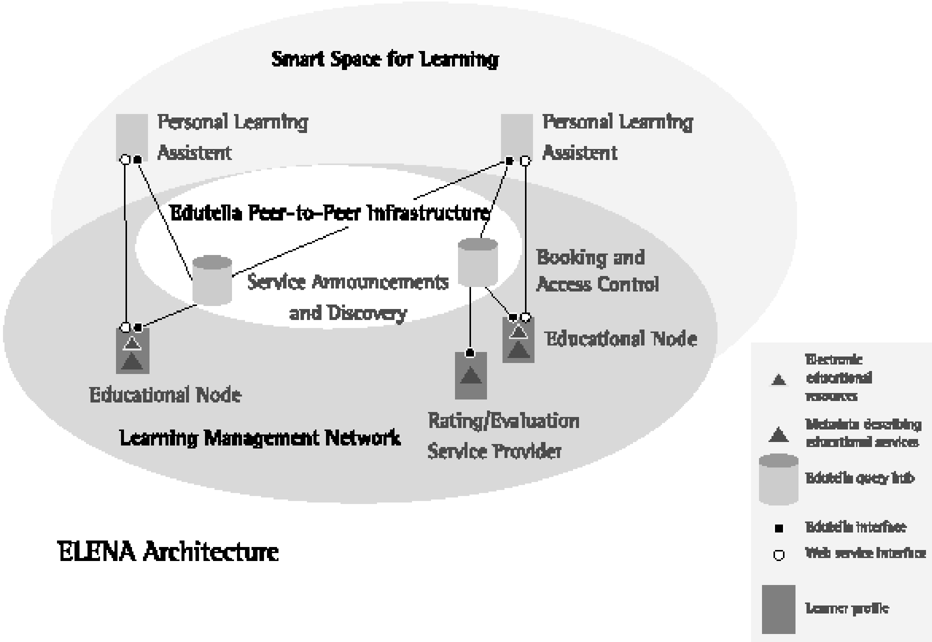


Figure 7: ELENA Architecture [Simon, B. et al]

Edutella

Edutella [Nejdl et al., 2002] is a peer-to-peer (P2P) infrastructure that aims at connecting highly heterogeneous educational peers with different types of repositories, query languages and different kinds of metadata schemas. It builds upon the metadata standards defined for the World Wide Web and aims to provide an RDF-based metadata infrastructure for P2P applications, building on the JXTA Framework (See: http://www.jxta.org/). Their vision is to provide the metadata services needed to enable interoperability between heterogeneous JXTA applications. The first application is focused on a P2P network for the exchange of educational resources using schemas like IEEE LOM [http://ltsc.ieee.org/wg12/20020612-Final-LOM-Draft.html], IMS [http://www.imsproject.org/metadata/], and ADL (Advanced Distributed Learning) SCORM [http://www.adlnet.gov/scorm/index.cfm] to describe course materials. Each Edutella peer is capable of performing a number of basic services such as querying, replication and mapping. Learning service providers either connect directly to the network as Edutella peers or use other Edutella peers.

The CANDLE Project

In the same spirit, the CANDLE Project (Collaborative and Network Distributed Learning Environment) [Wetterling, J., 2003] uses metadata to describe course material to make it more re-usable. CANDLE extends the metadata set defined by the IEEE LOM standard. These additional metadata describe learning objects by their purpose (learning goals, assessment methods), complexity level, type of learners (face-to-face, distance), setting (corporate, university), estimated time for completion, and others.

The Open Learning Repository

The Open Learning Repository (OLR) [Dhraeif, H. et al., 2001] aims at metadata-based course portals, which structure and connect modularized course materials over the Web. The modular content can be distributed anywhere on the Internet, and is integrated by explicit metadata information in order to build courses and connected sets of learning materials. Modules can be re-used for other courses and in other contexts. IEEE LOM metadata are used by authors to help them choose modules and to connect them into course structures.

The SeLeNe project

The SeLeNe (Self e-Learning Networks) project [http://www.dcs.bbk.ac.uk/selene/] offers advanced services for the discovery, sharing and collaborative creation of learning resources, as well as a personalized access to such resources.

The LOMster project [Ternier et al., 2002 and 2003]

LOMster, described in a paper by Ternier and others [Ternier, S., Duval, E, and Vendepitte, P., 2002] and on the IMS web site [http://www.imsglobal.org/profiles/index.cfm], is a project that addresses sharing of learning objects on a P2P base with the following functionalities:

1) Publishing: A user selects the learning objects he wants to share with other users, by dragging and dropping. When the user adds a file to the system, the tool should generate as much metadata as possible. Currently, only file size and file name are taken into account. The user can add additional metadata manually.

2) Searching: A user can formulate boolean combinations of search criteria over LOM fields. LOMster sends its query to all other known peers. Replies from peers include the full LOM description of the results. Relevant learning objects can then be downloaded in the next phase. Finally, a user can inspect the state of the current downloads and uploads. In 0, the design of an architecture that connects learning object repositories is also described which provides interoperability with both Client/Server systems and other P2P networks.

The Edutella, CANDLE, ORL and LOMster projects examine problems related to the re-usability of the educational material and are concerned about making learning objects (LOM) available in repositories on the Web, in a way that they can be accessed and used by other systems. The Selene and Elena projects are also concerned about the re-usability of the educational material, and, besides this, they offer personalized access to the educational material according to the learner model. Each project defines its own learner model based on the IMS-LIP 0 and PAPI 0 standards.

4. Multimedia content and its relation to digital repositories (Tim Brody)

4.1 Multimedia content

"Multimedia" can encompass a wide range of meanings. The word is often used in a loose sense to mean any format other than simple-text, but more correctly it is used in the sense of many media – that is, content that is comprised of more than one format. It thus encompasses formats which can be rendered into forms which can be apprehended by our senses, for instance, audio, video, animations, pictures or hypertext, but it also includes more specialised formats which cannot be directly apprehended. In the latter category we can include materials as diverse as genetic sequence information, mathematical simulations, chemical formulae, descriptions of the interations between sub-atomic particles. In this review we place more emphasis on the former category.

A multimedia object typically consists of one or more files that may contain one or more internal 'streams' of data representing different formats. A multi-file multimedia object generated, for example, from a live performance might contain a photograph, an audio track and a textual description. Most linear-video format files are actually container objects for streams of compressed visual and audio data. While a user might not care about the internal structure of the files that a repository contains, the internal formats chosen may have significant consequences for the management of the repository, the functionality it provides (particularly for seaching and access), performance and, not least, long-term accessibility of those files. The preservation of content for which there is no standard likely to survive for the longer term (say over 10 years) is still problematical and is the subject of ongoing research. Multi-file multi-format digital objects need to be stored by repository software, with the relationships between those files and the different formats captured and then presented in a meaningful way to users (for example, "this is a textual description of the attached image").

In the digital library sphere, multimedia will increasingly consist of data sets – homogeneous collections of material that may be analysed or processed into other data sets. As the volume of multimedia content increases, so the ability of users to search, navigate and discover successfully within that collection is tested – in effect the user is dealing with data sets that require domain-specific analytical tools. Purely data-driven collections (e.g. experimental results) may be interacted with by using visualisation tools, effectively a multimedia interface into the repository collection.

4.2 Multimedia issues

Because of the complexities introduced by mixing formats (often in intricate combination), multimedia content presents numerous issues for repository use and management. These include how to describe that material in such a way that enables search and discovery, the provision of long-term access and the ability to aggregate across collections. In METS terminology (See: http://www.loc.gov/standards/mets/METSOverview.html) both descriptive and structural metadata are required respectively to describe the object in order to support search and discovery and secondly to describe the relationship between the multiple files that make up a multimedia object.

The most widely used multimedia metadata format is the EXIF standard (Exchanged Image File - see http://www.exif.org/) used in digital photography. Digital camera manufacturers have added metadata to images that describes the type of camera used, exposure settings, time and date etc. The metadata are embedded in the image file (JPEG format). This type of technical data is intended to help identify camera settings or otherwise, but may in future provide more descriptive data, for example, the location where an image was taken.

The CIDOC Conceptual Reference Model (CRM) presented in section 5.2 below was developed to describe "the implicit and explicit concepts and relationships in cultural heritage documentation." For instance in the SCULPTEUR project a CRM ontology (a collection of concepts that describe a domain) was developed that allowed the project members to provide a method to navigate around the collections by concept (e.g. by the period an artwork depicts).

A multimedia repository needs to be more than just a collection of files that relies on the user's web browser to render downloaded material appropriately. Complex multimedia material cannot always be rendered directly by a web browser. One example is annotated video: In this case the repository interface needs either to provide suitable rendering tools, or provide – at the very least – links to tools that can render the multimedia content. Better still, a repository should provide a consistent mechanism for user access that does not distort or change the meaning of the original material (for example by avoiding 'lossy' formats that may reduce the size of material at the expense of quality). As an example, in order to improve interactive streaming of content such as audio or videos, facilities should be provided to allow a user to jump to a particular location in the content, that would otherwise require them to download the entire (large) file. Streaming of content requires an evaluation of the appropriate server technology, depending on user's preferences and the playback quality (and flexibility) afforded. Choosing proprietary formats may preclude users who cannot afford to purchase the necessary software; similarly choosing an uncommon (but 'open') format may preclude novice users who are unable or unwilling to install complex tools.

Choosing formats is a difficult decision (assuming a repository has the technical capability to convert or determine formats supplied). A multimedia repository needs to make its material accessible to users now (based on the rendering tools users are most likely to have) as well as future-proofing to enable long-term access. Many formats used for multimedia content – because of their sizes – use some form of lossy compression. "Lossy"[7] compression makes use of particular parts of a human's perception to discard detail that is not normally perceived (for example, slight gradients in colour blocks). However, repeated or overly aggressive lossy compression can lead to non-recoverable loss of information, and perceptible loss in quality. Therefore repositories have a choice to make between purchasing more storage (with all the consequent costs) and accepting some loss of quality. As costs are often limited by the available budget, a value judgement may be necessary to select some material for poorer quality, for example, material that is 'less important'.

[7] In lossy compression, not all original data is recovered when the file is uncompressed (in comparison to "lossless" compression).

4.3 Multimedia in common repository software

Widely used repository software systems such as Fedora, DSpace, eprints.org can be used to store multimedia objects (if defining a multimedia object as a collection of files). The user interaction, structural metadata and file formats supported are, however, dependent on site-specific customisations.

One typical multimedia user interaction is the addition of thumb-nail previews – useful for large images that would otherwise take a long time to download. Thumbnails can be integrated into the DSpace and eprints.org software through bespoke mechanisms, for instance, through using a 'magic' filename for thumbnails that, when the abstract page is shown, are shown immediately. Such bespoke solutions need to be standardised when exporting data from multimedia repositories (for example, so an external service can correctly associate a thumbnail image as being a low-resolution, small-volume version of the main image file). The METS metadata format and MPEG DIDL standard are early contenders for exporting such structural metadata for complex, multimedia objects. METS stands for the Metadata Encoding and Transmission Standard (http://www.loc.gov/standards/mets/), while the DIDL in MPEG DIDL stands for Digital Item Declaration Language; it is based on part 2 of the MPEG-21 standard (see for example the article by Bekaert, J., Hochstenbach, P., and Van de Sompel, H, in the November 2003 issue of D-Lib, http://www.dlib.org/dlib/november03/bekaert/11bekaert.html).

4.4 Exemplars of multimedia repositories

In July 2006 The Directory of Open Access Repositories (OpenDOAR[8]) listed 32 entries as containing 'multimedia' content (from a list of 379 entries). While these repositories contain multimedia content (mostly images), few provide customisations tailored towards multimedia-specific user interactions.

Some examples of repositories with multimedia content follow:

Electronic Cultural Heritage made Accessible for Sustainable Exploitation (eCHASE, at http://www.echase.org/)

This project seeks to demonstrate "…policies and processes for cultural heritage content holders to filter and make accessible their content according to the needs, prospects, rights and usage of different user-groups at a European level."

SCULPTEUR (http://www.sculpteurweb.org/)

"In the framework of the SCULPTEUR project Giunti Interactive Labs has developed an application that assists curators, instructional designers and educators to build 3D Virtual Learning Objects. These Learning Objects include virtual museums and galleries which offer an interactive learning experience following e-learning standards like IEEE LOM, IMS Content Package and ADL SCORM."

[8] 2006-06-22 OpenDOAR http://www.opendoar.org/

Demetrius repository at the Australian National University
(http://dspace.anu.edu.au/)

The Australian National University (ANU) institutional repository contains a large collection of digitised images in its Demetrius repository. These are stored as high-resolution TIFF-format files (about 15MB each). These images are not practically usable in a web browser – without additional plug-ins neither Internet Explorer nor Mozilla will render these TIFF files directly; usually they will also take a long time to download. To alleviate this the ANU provides 'thumb-nail' (small images) versions in the navigation system as well as a lower resolution version that is displayed on the abstract page (which links to the actual, full-resolution TIFF file). A thumb-nail view and the intermediate-level image (plus download reference to the full file) are shown in the illustrations below.

Demetrius is based on the DSpace software.

Thumbnail or Submission Date	Title	Authors
	arcv12345678.tif	Australian National University, Centre for New Media Arts CNMA; Eleanor Gates-Stuart
	arcv432.tif	Australian National University, Centre for New Media Arts CNMA; Eleanor Gates-Stuart

Figure 8: Demetrius – Image catalogue

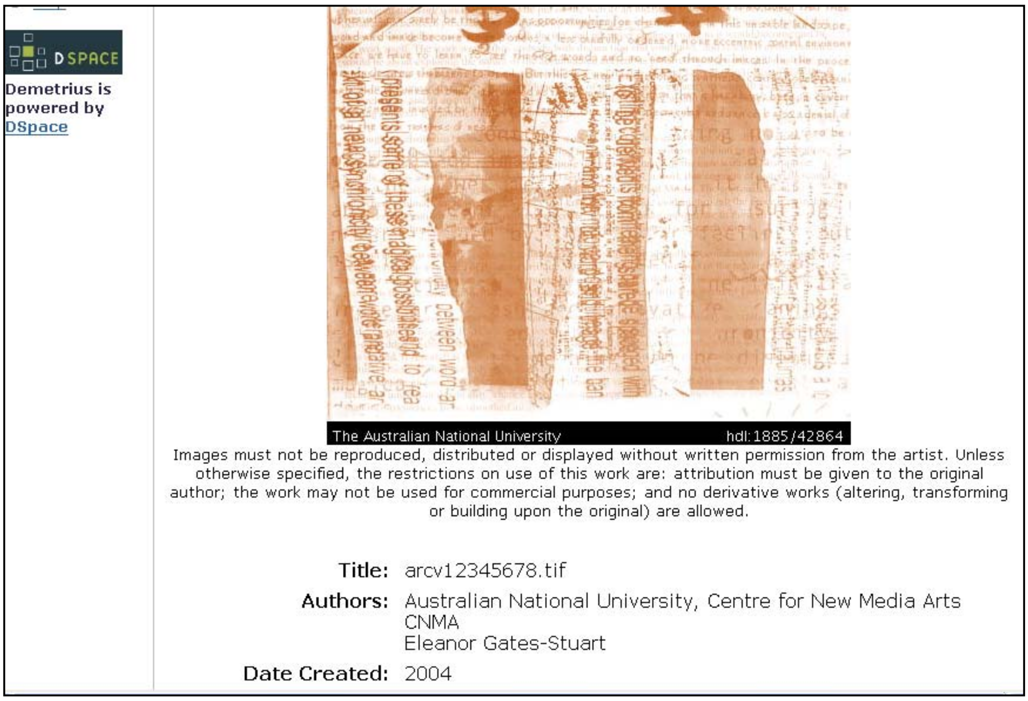


Figure 9: Demetrius – Image display at high resolution

CERN Document Server (see http://cds.cern.ch/)

The CERN document server (CDS) is a large repository of some 800,000 records (of which 750,000 are textual). CDS has a conversion service that can convert common document and image formats. For audio files (largely lectures or speeches) CDS provides the same recording in multiple formats, for instance the common MP3 format and the open format 'Ogg Vorbis' (see Figure 10 below).

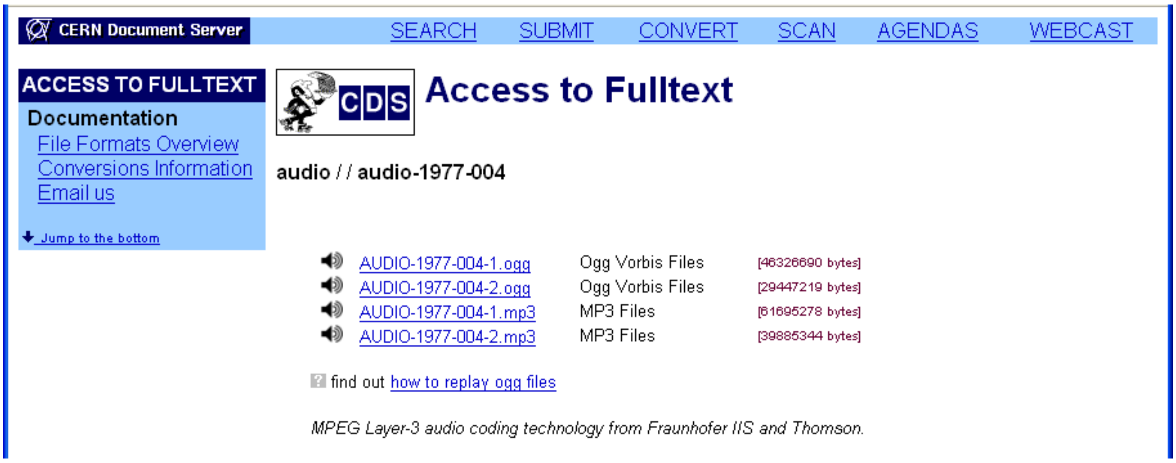


Figure 10: CERN document server

Scripps Institution of Oceanography (SIO) Expedition Discoveries (see http://nsdl.sdsc.edu/)

The SIO repository contains "data, documents and images from 822 expeditions by the Scripps Institution of Oceanography (SIO) since 1903". Of note is that the data sets stored in the SIO repository can be accessed and downloaded through a bespoke JAVA application. This is in addition to the SIO web site, as well as being aggregated into the US National Science Digital Library (NSDL). An example image is provided below (Figure 11), showing tracks of ocean exploration trips (commercial field oil exploration repositories show similar tracks, and linking them to extensive collections of scientific and geological data gathered along the tracks).

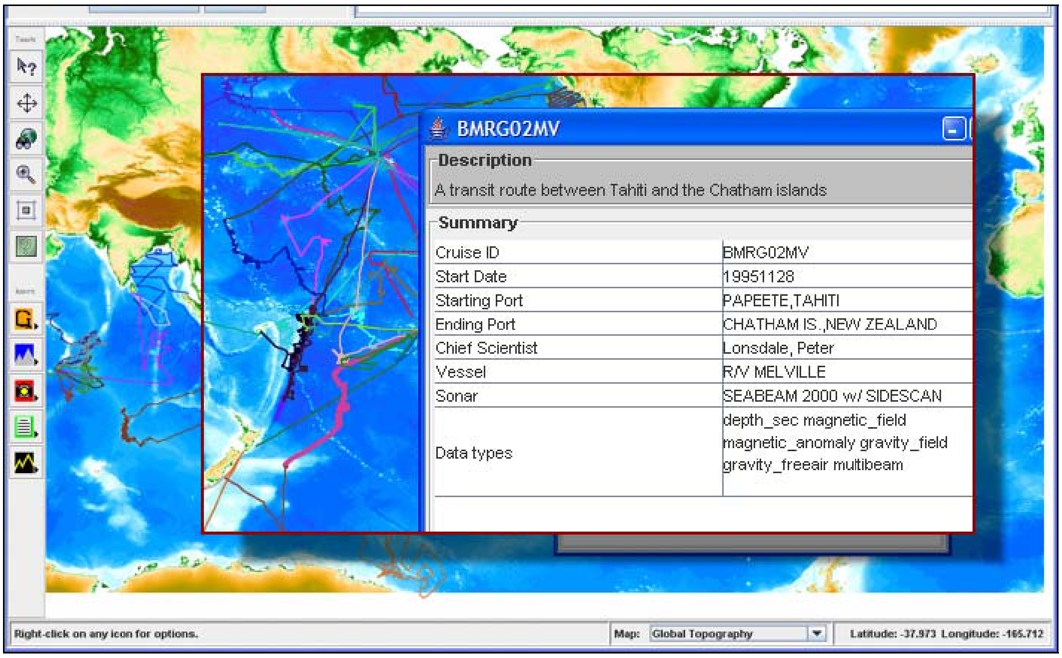


Figure 11: SIO repository – showing overlaid evploration data

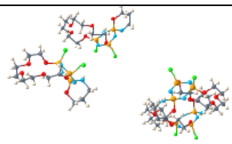
Crystal Structure Report Archive

The Crystal Structure Report Archive at Southampton University is a repository of crystal structures identified using X-Ray diffraction. This (semi-automated) process results in a number data files produced by machine and by expert refinement. The repository provides access to these data files, a metadata description and a 3D java-applet-based rendering of the crystal structure. The metadata includes the chemical name that can be searched for using text-based search engines, such as Google's web search facility. (See also the case study presented in section 2.5 above.)

Help

$C_{11}H_{23}Cl_2N_4O_6P_3$

InChI=1/4C11H23Cl2N4O6P3/c4*12-24-15-25(13,17-26(16-24)14-2-1-3-23-26)22-11-9-20-7-5-18-4-6-19-8-10-21-24/h4*14H,1-11H2



DOI: 10.1594/ecrystals.chem.soton.ac.uk/147
Compound Class: Inorganic
Keywords: cyclotriphosphazene
Creation Date: 18 October 2004
Deposited By: A.N. Admin
Deposited On: 14 March 2006

Data collection parameters

Chemical formula	C11 H23 Cl2 N4 O6 P3
Crystallisation Solvent	
Crystal morphology	Blade
Crystal system	Orthorhombic
Space group symbol	P2(1)2(1)2(1)
Cell length a	13.648(3)
Cell length b	17.348(3)
Cell length c	33.734(8)

Available Files

Final Result

04src0794.cif	40k
04src0794.cml	25k

Validation

04src0794_checkcif.htm	12k

Refinement

04src0794.res	22k
04src0794_xl.lst	109k

Figure 12: Crystal Structure Report – with chemical structure display

5. Subject access and semantic interoperability

This section is divided into two parts, the first discussing subject access and recommendations for improvement of semantic interoperability, the second section presenting the CIDOC Conceptual Reference Model to enable high level conceptual information integration.

5.1 Subject access and semantic interoperability in repositories (Traugott Koch)

5.1.1 Introduction

Definitions of a digital repository vary (see sections 1 and 2 above); so too does the level of value-adding services provided with a digital repository. Semantic interoperability and knowledge organisation in repository-related services, however, always depend on efforts at an individual repository level and at the service level.

The Delos report on semantic interoperability in digital library systems [Patel M. et al., 2005] published definitions of these various concepts and set out the prerequisites and methods for enhancing semantic interoperability in digital libraries. Though digital libraries represent a much broader type of service than digital repositories, much of the report is also relevant for digital repositories. This is particularly true given that repositories are often established at the outset as elements within larger or different types of services.

Subject portals, digital libraries, databases, “collaboratories[9]”, journals, virtual learning environments, personal web sites – all these can be built on top of several different repositories; alternatively, the repositories can form an integrated module within such services. Repositories, therefore, should be seen as one type of cell in a “recombinant”, modular library and information services world [Dempsey, L., 2003].

However, for recommendations on enhancing knowledge organisation and semantic interoperability it is important to differentiate between the different purposes and types of repositories, and their involvement in higher-level services (see section 5.1.3 and 5.1.4 below).

This section looks in particular at institutional, global and subject repositories, based on deposit/active submission and on OAI harvesting (OAI is the Open Archives Initiative, at www.openarchives.org). We examine selected repositories with regard to their provisions for subject access, and potential other semantic interoperability efforts, discuss some possible relevant methods and approaches for enhancing semantic interoperability (see also section 5.2), and present selected, existing best-practice guidelines in this respect from national support organisations.

9 A collaboratory is “an amalgamation of collaboration and laboratory, conveying the concept of a collective research organization where a high value and focus is placed on the sharing of effort and findings such that the quality and progress of the research is optimized and relevant, and every member of the research organization receives benefits greater than their actual investment”, date of harvest June 2006: http://www.ichnet.org/glossary.htm

5.1.2 Empirical findings regarding subject access in repositories

In July 2006 the OAI registry at the University of Illinois Urbana-Champaign lists 1,112 responding repositories, featuring 183 different metadata schemas [OAI Registry at UIUC; see: http://gita.grainger.uiuc.edu/registry/].

An examination of subject access based on knowledge organization and semantic interoperability measures in a few selected repositories reveals some of these repositories are based on OAI harvesting, others not; some are institutional, others are global, and others are subject repositories.

The majority, by far, do not seem to make any provision for improving semantic interoperability when gathering data or offering services. This is especially true for services based on metadata harvesting without prior agreements with data providers. Subject-based repositories and services offering templates for submission or applying "central" metadata generation feature richer subject access provision and sometimes apply standardisation to one classification or subject heading system.

Among the examples of different types of repositories where we found an acceptable level of subject access were the global repositories ArXiv (http://arxiv.org/) and InfoNet Eprints, DEF Denmark (http://preprints.cvt.dk/); the institutional repositories OCLC Research Publications Repository (http://www.oclc.org/research/publications/default.htm), e-Prints Southampton (http://eprints.soton.ac.uk/), and Humboldt University Berlin (http://edoc.hu-berlin.de); and the subject repositories E-LIS (http://eprints.rclis.org/), DLIST Arizona (http://dlist.sir.arizona.edu/) and Organic Eprints (http://orgprints.org/). Even among those good examples, however, important features such as systematic browsing, subjects as search filters, consistency of usage, standardisation of subject schemes or full exploitation of available subject information was missing.

The general impression is, that repositories and aggregation services only in exceptional cases so far have been taking any measures to harmonise the heterogeneity of subject information or to offer proper searching and browsing of documents by subject/topic. The necessity to do so might be limited in homogeneous small institutional repositories but this becomes dramatically different as soon as these documents or their metadata is exported, aggregated or re-used in a different context.

Using a recent review of repository software [Open Society Institute, 2004] we can see to what degree the software packages support or prevent the exploration of improved semantic interoperability efforts. All major packages support metadata review or enhancement (though there are perhaps some question-marks over Fedora), the addition of metadata fields, the setting of default values for metadata (with a question-mark with regard to ePrints) and facilities for metadata export. The user interfaces are all modifiable. The built-in search capabilities are often rather weak, but most major packages support metadata search and browse. Several systems (ePrints, Fedora) do not enable full-text search, while DSpace does not allow browsing by subject terms. Many ePrints installations have augmented the system's baseline capabilities - for example by integrating advanced search, extended metadata, and other features.

In general, the major packages appear to accommodate semantic interoperability and knowledge organization efforts and have basic functionality to offer the results to end-users.

5.1.3 Semantic interoperability and knowledge organization work

The Delos Semantic Interoperability report [Patel, M., et al., 2005] distinguishes semantic interoperability at three levels of abstraction:

1. **Data structures**, be it for metadata, content data, collection management data, service description data.
2. **Categorical data**, i.e. data that refer to universals, such as classification, typologies and general subjects.
3. **Factual data**, i.e. data that refer to particulars, such as people, items, places.

Most awareness in the field of repositories seems to be about efforts regarding the first item above, data structures. The OAI-PMH, on a protocol level, points to the need for this by making a few elements from unqualified Dublin Core the default for OAI compliance, thus enforcing data providers to perform the necessary data element mapping between their own structures and these common elements.

SVEP in Sweden (see below) establishes a recommended metadata profile; ePrints UK recommends establishing a common metadata application profile, and DINI (also see below) focuses on the presence of certain data elements such as subject, publication and document type. Creating a common metadata profile for services integrating distributed heterogeneous resources and databases is now a quite established method, see for example Renardus [http://www.renardus.org/]. Registries provide access to such mappings [Patel, M., et al., 2005, section 4.5.3] for reference and re-use.

Classifications and other knowledge organization systems, referring to categorical data, have been standardised or mapped to a much lesser degree as the basis for semantically improved access in the repository context. DINI and SVEP have made recommendations, and DINI in particular has moved towards one common classification system (the new mandatory requirement for certification, to classify all documents using the Dewey Decimal Classification, DDC), at least for set categorization of OAI-compliant document collections.

Among the aggregation services developed for the UK ePrint repositories (JISC ePrints UK project [ePrints UK]), OCLC Research has been experimenting with an automatic classification service to enable routing of metadata for ePrints to the discipline-focused hubs of the Resource Discovery Network [Dempsey, L. et al., 2004].

Treatment of heterogeneity when integrating data sources has been and still is researched in projects supported by the German Research Foundation, especially by the Social Sciences Information Centre, Informationszentrum Sozialwissenschaften in Bonn. [Hellweg, H. et al., 2001; and Krause, J., 2004]. In April 2005, the Berlin Open Forum on Metadata Registries discussed semantic registry support, especially in the language engineering field, an important prerequisite for semantic enhancements regarding categorical data.

The third level, factual data, has been addressed by efforts to standardise with the help of name-authority databases, gazetteers and other geographic name authorities. As a value-adding service OCLC Research developed name-authority control (a name look-up service) as metadata creation support to be hooked up with templates in the DSpace repository software package. Created as web services, this and other developments of the Metadata Switch umbrella project/Terminology services project at OCLC [http://www.oclc.org/research/projects/mswitch/default.htm], can provide remote semantic interoperability enhancing functionality which can be plugged into specific applications.

OCLC Pica has been commissioned to develop a "National Author Thesaurus" for the Dutch national repository initiative, DARE, and several European projects include further development and integration of name authority records emanating primarily from national libraries.

Metadata enhancement can take place at all three levels. Different strategies for metadata enhancement involve semantic interoperability efforts. Cornell University leads such developments for the National Science Digital Library (NSDL) project in the USA, together with partners such as INFOMINE, University of California Riverside and its iVia Virtual Library software.

Metadata augmentation can enrich metadata records with subject headings, for example, with Library of Congress Subject Headings, and keywords. Transformation services can correct degraded terms from controlled vocabularies and recognize values from recommended vocabularies and ascribe the appropriate vocabulary encoding scheme to the statements. New metadata values can be generated based on mappings between schemas or vocabularies [Hillmann, D. I. et al., 2004].

Guidelines published on the improvement of metadata quality in ePrint archives in the context of the ePrints UK project [Guy, M. et al., 2004] underline the importance of early decisions on the usage and granularity of controlled vocabularies, their consistent application and the importance of built-in support for them in metadata editing tools.

5.1.4 Recommendations

Recommendations regarding activities for enhancing semantic interoperability in digital repositories have to be differentiated according to the scope and ambition of each repository or service and to different types of repositories. The scopes of the recommendations can vary between relating to:

a) one repository, focusing on storage and simplified access
b) integration of heterogeneous data and sources into a common service
c) basic discovery and retrieval services, metadata enhancements
d) advanced services such as Digital Libraries; considerably improved semantic interoperability and knowledge organization for precise subject access

Different levels of coverage of repositories and services call for slightly different recommendations:

- A coherent individual, often institutional, repository should focus on standardising the description of its own environment and focus on inviting as rich semantics as possible. (cf. for instance, ePrints UK, cf. Guy, M. et al., 2004).
- Services integrating heterogeneous repositories should aim at trying to preserve the richness from the participating sources. They should allow at least search and browse and filtering by schema and display the multiplicity of vocabularies. For further improvement, one should engage in future semantic interoperability efforts, as recommended by DINI or SVEP, and try mapping and other heterogeneity treatment between the vocabularies used.
- For subject repositories there is an immediate and fundamental need to carry out integration or original subject indexing in one or several standardised systems. Heterogeneity treatment, and efforts on knowledge organization and semantic interoperability need to take place at an elaborate level, employing many of the methods described in Patel et al 2005, especially in section 5.

5.1.5 Existing national recommendations on good practice

Several national repository support organisations have started to formulate recommendations for best practice when creating institutional repositories, often in order to accomplish a good foundation for aggregating services.

DINI, Germany

DINI (German Initiative for Networked Information) and its working group OAI in Germany published 2003 a second version (first version March 2002) of recommendations for OAI data providers at German universities [DINI 2003a]. In Germany, more than 100 servers offer academic publications via the OAI protocol and this is the highest number known from any country.

In order to improve interoperability for the benefit of OAI services, DINI makes recommendations to service providers to assist in selective harvesting of metadata, and recommends a certain subject and formal structuring of the repositories, expressed in set categorizations.

Sets should be defined using a two-stage hierarchy in each of three classifications:

1) a subject classification, here ddc e.g. ddc:020 (setSpec)
2) a publication type pub-type e.g. pub-type:monograph (setSpec) and
3) a document type doc-type e.g. doc-type:text (setSpec)

English language should be used for the classes and the setName element of the protocol which is a verbal description of the set.
It is said that it would be desirable if each item would be classified in all three dimensions and carry at least one second level subject classification. The ListSets request should only list classes/sets actually used.
Reports of experiences communicated by responsible DINI colleagues suggest that the specific recommendations relating to the use of sets have not been widely adopted so far. The subject information contained in the local servers is still limited and very heterogeneous. In the absence of suitable classification data, institutional servers offer browsing by keyword

(Chemnitz) and harvest based on keywords, rather than on subject classes and subject-based sets.
For (subject) communities DINI recommends agreement on specialized metadata formats. Recommendations regarding the use of the Dublin Core's 15 metadata elements are issued by the DINI Dublin Core working group. Some of the German data providers have introduced these sets into their OAI interfaces.

In the context of the quality control through a certification process for repositories offered by DINI [DINI 2003b], the minimum requirement for a certificate (among others) is the availability of a well defined and written policy for subject indexing and verbal indexing of the documents with keywords or document classification. The newest version of the requirements for a certificate, to be released by the end of 2006, prescribes now the use of DDC for classification.

At this time, the OAI Search in German Open Access Repositories, covering 57 000 documents from 54 servers, does only provide searching based on keywords, but not on subject categories such as DDC nor does it offer any browsing [http://www.dini.de/oaisuche/oai.php].

Recommendations which might become future requirements are:

- At least one additional standardized system of verbal or classificatory indexing, general or subject-specific
- Keywords in English
- Abstracts in German and English

"In the long run, indexing not based on a controlled vocabulary will not be sufficient to guarantee good recall and precision in searches for publications. To permit quality searches across multiple document repositories, international indexing systems should be used."

Regarding metadata in general, Qualified Dublin Core is recommended. This has the potential to allow advanced semantic interoperability efforts as has the presence and continued activity of a strong national support organisation such as DINI. Several follow-up projects are expected to further improve national coordination and interoperability in the network of Open Access servers.

Among recent major political recommendations, DINI lists in September 2005: "Improve (national/international) networking of German institutional repositories including subject specific access options." "...it is an urgent requirement to implement institutional repositories according to standards to integrate them into subject based or global metalevel services (e.g. subject portals and databases, usage and impact services, or, search and harvesting services). ... The interaction between subject based and global services must be taken into account more strongly than is the case." [DINI Schriften: Building on e-Publications Infrastructure, 2005].

SVEP, Sweden

SVEP (Coordination of e-Publishing within Swedish Higher Education) [http://www.svep-projekt.se/english/] has issued recommendations on metadata format and subject categories to be used in Swedish academic repositories. They are only available in Swedish and not yet widely implemented. Some of them have also been adopted as Norwegian national recommendations for university repositories.

The recommendations comprise:

a.) A metadata profile for interoperability between Swedish institutional repositories at a OAI service level and for metadata reporting to the national library, encoded in XML. The subject element is made mandatory. A minimum level requires only unqualified DC simple oai-dc for the subject element and for SVEP services a subject classification according to point c) below. The vocabulary encoding schema information is not standard-based.

    Resource type information is not made mandatory and it proposes dc:type at minimum level and MODS TypeOfResource for SVEP services. Publication type is mandatory, requiring dc:type for minimum and inviting a Swedish list (with English terms from ePrint UK) for SVEP services for extended interoperability. See: http://www.ub.uu.se/epub/recommendations/rek1_0.pdf (in Swedish only)

b.) A metadata profile and a OAI-PMH set structure for a national search service of student and exam papers (Libris Upsoek).

c.) A subject classification for scientific publications, based on existing systems at the Central Statistical Office and the Science Council. It includes rough mappings to a few other commonly used knowledge organization systems. See: http://www.ub.uu.se/epub/categories/. These categories are also used to offer subject-based selective harvesting of metadata records via OAI sets. There are good experiences with this recommendation: all institutional repositories that follow SVEP recommendations use the common subject classification.

d.) A proposed metadata profile for the purpose of bibliographic description and analysis in common research reporting, encoded in XML.

An experimental search service across digital scientific publications from 25 Swedish university repositories [http://svep.epc.ub.uu.se/testbed/?lang=en] illustrates the feasibility and the interoperability effect of the recommendations. It includes browsing by categories, a rather rare feature in other repository based services. A decent level of support of the SVEP recommendations and solutions has been accomplished [Andersson, S. , 2006].

Australia

ARROW (Australian Research Repositories Online to the World) [http://www.arrow.edu.au] is a national resource discovery portal for higher education research. It daily harvests metadata records from institutional repositories using the OAI protocol and makes them available for searching, browsing, citing or statistics. There seem to be metadata about 70 000 documents from about 30 repositories.

ARROW recommends the Dublin Core metadata schema and the use of ASRC for subject classification. It invites other subject indexing such as LCSH or free keywords. The subject metadata element is optional, however. ASRC, the Australian Standard Research Classification list (last updated 1998), is a required coding for Research fields, Courses and Disciplines in the governmental reporting system and will be used in the developing Research Quality Framework, equivalent to the UK Research Assessment Exercise RAE. Some Australian colleagues claim a rather low level of use of the ASRC in searching by end-users. A few universities are using the Library of Congress Classification (LCC) in addition to ASRC and accomplish thus some international interoperability.

To support harvesting of subsets of repository documents, the use of OAI sets is recommended, "based on the values in the type element and/or the subject element." [ARROW Discovery]

Metadata records can be searched and browsed using ASRC in the ARROW Discovery Service [http://search.arrow.edu.au/apps/ArrowUI/?adapter=ViewASRCListAdapter]. AUSESEARCH [http://rollyo.com/arthursale/ausesearch/] offers free text search (in metadata records) only.

UK

At this time there are no national recommendations, similar to the ones in Germany and Sweden, in the UK. The main recommendations made in JISC's Digital Repositories Review [Heery, Anderson 2005] include, however, the funding of smart tools for automatic data extraction, automatic classification etc. The need for repository user interfaces directed towards communities of practice is also mentioned. A focus group report mentions tools such as clustering and synonym services and tools to facilitate interdisciplinary exploitation. A software developers' survey adds a demand for the use of qualified Dublin Core metadata and even more sophisticated metadata handling "to allow subject aggregators to deliver real value" (see below for progress on developing a DC Application profile for ePrints). Elsewhere in the report other semantic interoperability efforts are mentioned as desiderata, e.g. metadata enhancement in subject repositories, subject interfaces to institutional repositories, application of name authority services, use of common schemas, semantically aware annotation services, quality control of metadata and interoperability between different types of repositories in general.

Currently within the UK a JISC working group is developing a Dublin Core Application Profile for describing scholarly publications (ePrints) held in institutional repositories (see http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile). This work is being undertaken within the JISC Digital Repositories. The objectives of the working group are to develop:

- a Dublin Core application profile for eprints;
- any implementation / cataloguing rules, that might be necessary to support functionality offered by the search service, such as fielded searches of the metadata or indexing the full-text of the research paper;
- a plan for early community acceptance and take-up, bearing in mind current practice.

In the context of this work an ePrint is defined to be a *scientific or scholarly research text* for example a peer-reviewed journal article, a preprint, a working paper, a thesis, a book chapter, a report, etc. Whilst this work specifies how an ePrint should be described, it is not prescriptive regarding which subject scheme (if any) should be used to indicate the topic of the ePrint.

5.1.6 Summary

National recommendations or requirements are certainly highly beneficial for accomplishing some semantic interoperability and for allowing cross-searching and cross-browsing in national all-subject aggregation services. To make documents available for topical searching and download on an international scale or in discipline specific services, national vocabulary schemes are not sufficient, however. Standardised international subject schemes and international standard schemes with broad subject coverage are necessary if anything more than known item searching (author names, document titles, institutions known) is to be accomplished. If a national system is already in wide use, mapping to a broad international scheme might be an option. The needs of subject-specific services will, however, require re-classification and re-indexing of the documents for that specific purpose, since broad national schemes tend to be very shallow.

This section (5.1) could only touch on the issues of subject access to and semantic interoperability in digital repositories by looking at existing solutions and recommendations. The author of this section and colleagues have recently completed a much more thorough state-of-the-art review on terminology services and technology for JISC, including detailed recommendations. This study is not limited to this area, but is highly relevant for potential interoperability improvements in the area of digital repositories [Tudhope, Koch and Heery, 2006].

5.2 CIDOC Conceptual Reference Model (Martin Doerr)

Semantic interoperability between repositories requires compatibility between both the knowledge organization systems used (such as classification systems, terminologies, authority files, gazetteers) and the data and metadata schemata employed. Currently, the scope of repositories covers not only traditional publications but also scientific and cultural heritage data. The difference between traditional publication in text form and provision of structured data is becoming more and more blurred: databases now contain texts in XML form, and texts and multimedia are being described by structured metadata records. Natural-language processing techniques are increasingly used to extract structured information from free texts. The grand vision is to see all these data integrated so that users are effectively supported in searching for and analyzing data across all domains.

Even though the Dublin Core metadata element set is well accepted as a general solution, it fails to describe more complex information assets and their cross-correlations. This deficiency is seen across data from political history, history of the arts and sciences, archaeology or observational data from natural history or geosciences, and indeed in other disciplines.

Core ontologies describing the semantics of metadata schemata are the most effective tool to drive global schema and information integration [Patel, M., Koch, T., et al. 2005], and provide a more robust, scalable solution than tailored ‘cross-walks’ between individual schemata. Information and queries can be mapped to and from a core ontology, which serves as a virtual global schema and has the capability to integrate complementary information from more restricted schemata. Many scientists question the feasibility of such a global ontology across domains. On the other side, schemata like Dublin Core reveal the existence of overarching concepts. Ideally, the European Digital Library which has been envisaged would be based on one sufficiently expressive core ontology by harmonization and integration of the relevant alternatives. The challenge is to explore at a practical level the limits to harmonizing conceptualizations over different domains.

5.2.1 The CIDOC CRM – engineering core information structures

The International Committee for Documentation (CIDOC) has developed a formal ontology called the Conceptual Reference Model (CRM) which is intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information. The primary role of the CRM is to enable information exchange and integration between heterogeneous sources of cultural heritage information. It aims to provide the semantic definitions and clarifications needed to transform disparate, localised information sources into a coherent global resource, be it within a larger institution, an intranet or on the Internet. Its perspective was to be supra-institutional and abstracted from any specific local context. This characteristic determined the constructs and level of detail of the CRM [CRM version 4.2].

The CRM is the culmination of more than a decade of standards development work by CIDOC of the International Council of Museums (ICOM). Work on the CRM itself began in 1996 under the auspices of the ICOM-CIDOC Documentation Standards Working Group, applying precursors and methods developed at ICS-FORTH since 1992. Since 2000, development of the CRM has been officially delegated by ICOM-CIDOC to the CIDOC CRM Special Interest Group, and this collaborates with the International Standards Organisation working group ISO/TC46/SC4/WG9 to bring the CRM to the form and status of an international standard.

The CRM and its precursors were developed by interdisciplinary teams of experts, from fields such as computer science, archaeology, museum curation, history of arts, natural history, library science, physics and philosophy. It started “bottom up”, by re-engineering and integrating the semantic contents of more and more database schemata and documentation structures from all kinds of museum disciplines, archives and (recently) libraries. The development team applied strict principles to admit only concepts that serve the functionality of global information integration, and other, more philosophical restrictions about the kind of discourse to be supported.

More specifically, the CRM defines, and is restricted to the underlying semantics of database schemata and document structures used in cultural heritage and museum documentation in terms of a formal ontology. It does not define any of the terminology appearing typically as data in the respective data structures. Likewise it does not aim to propose what cultural institutions should document, but rather to describe the logic of what they actually document, and thereby to enable semantic interoperability.

It is intended that it provide an optimal description of the intellectual structure of cultural documentation in logical terms. As such, it is not optimised for implementation-specific storage and processing aspects. Rather, it provides the means to understand the effects of such optimisations on the semantic accessibility of repository respective contents.

In dozens of applications and schema mapping exercises, real limits to the applicability of the CRM - slight extensions notwithstanding - have not yet been found. It seems that, beyond its original limited scope, the CRM formulates concepts which are nearly generic. Three ideas are central to the CRM:

1. Once data are registered in an information system such as a repository, the relationships between entities and the identifiers used to refer to them are part of the discourse about them and part of their historical reality. In simple words, within information assets there are references to entities by multiple names which cannot be associated unambiguously and therefore the ambiguity has to be properly documented.
2. Types and classification systems are not only a means to structure information about reality from an external point of view, but are also part of described reality as human inventions. Similarly, all documentation is also seen as part of the reality, and may be described together with the documented.
3. The way humans normally analyse the past is to split up time into discrete events, therefore information about the past can be formulated as events involving "persistent items" ("continuants" or "endurants"). Their involvement can be of quite different nature, but it implies in any case the presence of the respective things. Even immaterial items can be present in events through their carriers.

From this view, a picture of history emerges as a network of "lifelines" of continuants meeting in events in space-time. This abstraction turns out to be extraordinary powerful, and of value far beyond consideration of cultural assets. With a minimal schema, it allows for a surprising wealth of queries to be made. For instance, complex genetic family relations can be circumscribed by birth events associated with a father and a mother. Contribution and influence on human achievements can be traced in the meetings of people and ideas leading to them. Relative chronologies can be justified with causal ordering of events. This historical perspective is implied by and part of all metadata for stored information in libraries and digital libraries, independent of the application domain. Descriptive sciences, like geosciences and biodiversity studies, collect knowledge in immense numbers of observations, which can be described as events of human scale correlated by people and ideas and hence by the CRM.

5.2.2 The CRM - FRBR harmonization project

Quite independently of the development of the CIDOC CRM, the Functional Requirements for Bibliographic Records (FRBR) model was designed as an entity-relationship model by a study group appointed by the International Federation of Library Associations and Institutions (IFLA) during the period 1991-1997. It was published in 1998. Its innovation is to cluster publications and other items around the notion of a common conceptual origin – the 'Work' - in order to support information retrieval. It distinguishes four levels of abstraction from conception to the book in my hands: The Work, Expression, Manifestation, Item. Its focus is domain-independent and can be regarded as the most advanced formulation of library conceptualization [see Leboeuf, Patrick (ed.), 2005, and the IFLA study group on functional requirements for bibliographic records, 1998].

Initial contacts in 2000 between the two communities eventually led to the formation in 2003 of the International Working Group on FRBR/CIDOC CRM Harmonisation. The common goals are to express the IFLA FRBR model with the concepts, ontological methodology and notation conventions provided by the CIDOC CRM, and to merge the two object-oriented models thus obtained. This working group has published the first complete draft of the object-oriented version of FRBR (FRBRoo), in June 2006. This formal ontology is intended to capture and represent the underlying semantics of bibliographic information and to facilitate the integration, mediation and interchange of bibliographic and museum information.

5.2.3 Conclusions and future work

The potential impact of the combined models can be very high and the domains explicitly covered by the models are already immense. Further, they seem to be applicable to the experimental and observational scientific record for e-science applications. From a methodological perspective, the endeavour of core ontology harmonization experimentally proves the feasibility of finding viable common conceptual grounds, even if the initial conceptualizations seem incompatible [LeBoeuf, P., 2005, and Doerr, M., Hunter, J., Lagoze, C., 2003]. Even though this process is intellectually demanding and time-consuming, it is to be hoped the tremendous benefits of nearly global models will encourage more integration work on the core-ontology level.

6. Summary and Recommendations (Rachel Heery)

At various points, this report has had to go back to first principles by addressing definitions – see for example the typology of digital repositories in section 1 and exploring the definition of content types (e-learning material) in section 3. This is symptomatic of the immaturity of reporitory technologies and services. As a relatively new arrival in the landscape, the digital repositories field is fragmented and the role of repositories is still fluid. Within the UK, different players are emerging, beginning to define roles and responsibilities, and exploring ways to interoperate (such as illustrated in the repository ecology in Figure 1 and the e-Crystals Federation Model, in Figure 4). But fundamental questions remain: how do repositories relate to digital libraries? how do institutional repositories relate to national data centres, and to community subject-based repositories (which are often international in nature)?

As set out in the opening chapter, repositories form an intersection of interest for different communities of practice, with differing motivations and service needs that range over a number of areas. The intersection of interest across domains (domains of responsibility, activities, disciplines) offers possibilities for crossovers of technologies, and there is a real potential for adoption of a common architecture and data model amongst reporitories. Co-ordination of the implications of this crossover is important: at repeated points, the authors in this study point to the need for co-ordination and collaboration at an international level. For example, standardised metadata schemas, development of interoperable data exchange and deposit, agreements to use international subject schemes with broad subject coverage, are needed if a rich information environment is wanted – that is, something more than basic item searching (by author name, document title, institution).

Specific areas identified for further consideration include:

- Common approach to rights management in repositories
- Investigation of interoperability of learning and research objects in repositories, with particular consideration of the role of curation (provenance, identifiers)
- A life-cycle review of digital objects in repositories, considering inter alia versions, selection and deletion
- A detailed study of multimedia objects in repositories, examining the extent of the objects affected by multimedia issues, as well as retrievability, representation, and preservation implications
- Investigation of the potential for a common approach to modelling research data in repositories across eScience and GRID

In order to underpin more detailed investigations there is a need to develop a framework for repositories that would encompass:

- interfaces between repositories
- data flow between repositories

- integrated of repositories into the personal workflow of users (researchers, authors and information seekers), and the workflow of institutions (integration with research information systems, with funders research outputs systems etc)

This report touches on a variety of aspects of repositories. Raising awareness of the repository research agenda within European countries is beneficial for the European digital library community, and for the international research community. Collaboration and co-ordination of research efforts is essential to achieve interoperable solutions.

7. References

(Web pages valid as at June-July 2006)

Abecker, A., Bernardi, A., Hinkelmann, K., Kuhn, O., Sintek, M. "Towards a Technology for Organisational Memories", IEEE Intelligent Systems & Their Applications (May/June 1998)

Ackermann, M. S. "Definitional and Contextual Issues in Organizational and Group Memories", Proceedings of 27th Hawaii Int'1 Conference on System Sciences (1994)

ADL Co-Laboratories: Repositories Listing, http://projects.aadlcolab.org/repository-directory/repository_listing.asp

ADL Co-Laboratory, 2001, The SCORM Content Aggregation Model, http://www.adlnet.org/index.cfm?fuseaction=scormdown&cfid=220791&cftoken=29020644

Advanced Distributed Learning (ADL): SCORM, http://www.adlnet.org/

Agostini, A., De Michelis, G., Grasso, M.A., Prinz, W. and Syri, A. "Contexts, Work Processes, and Workspaces". Computer Supported Cooperative Work: The Journal of Collaborative Computing 5, 2-3 (1996). pp. 223-250

Analyti, A. and Christodoulakis, S., Multimedia Object Modeling and Content-Based Querying, Proceedings of Advanced Course – Multimedia databases in Perspective, Netherlands (1995)

Andersson, S. Testbed foer interoperabilitet. (2006) http://svep.epc.ub.uu.se/pub/testbed_about.html

Ariadne Project, http://www.ariadne-eu.org/index.html

ARROW Project http://arrow.edu.au/

ARROW Discovery Service harvesting guide (2005). http://www.arrow.edu.au/docs/files/harvesting.pdf

arXiv http://arxiv.org/

Asher, N., "Reference to Abstract Objects in Discourse. Studies in Linguistics and Philosophy". Studies in Linguistics and Philosophy. Kluwer Academic Publishers (1993)

Baru, C., Chu, V., Gupta, A., Ludaescher, B., Marciano, R., Papakonstantinou, Y. and Velikhov, P.. 1999, XML-based information mediation for digital libraries. In Proceedings of the fourth ACM conference on Digital libraries, pages 214-215, ACM Press (1999)

Bateman. J.A. “Enabling technology for multilingual natural language generation: the KPML development environment“, Natural Language Engineering 3(1) (March 1997)

Bekaert, J., Hochstenbach, P., and Van de Sompel, H. D-Lib (November 2003) http://www.dlib.org/dlib/november03/bekaert/11bekaert.html).

Belloti, R., Decurtins, C., Grossniklaus, M., Norrie, M.C., Palinginis, A. “Modeling Context for Information Environments”, In Proceedings of Ubiquitous Mobile Information and Collaboration Systems, Riga, Latvia (2004)

Berthouzoz, C., 1999, “A Model of Context Adapted to Domain-Independent Machine Translation”. Proceedings of the Modeling and Using Context, Second International and Interdisciplinary Conference, CONTEXT'99, Trento, Italy, (September 1999)

Blinco, K., Campbell, L.M., and Mason, J., 2004, Repository management and implementation: a White Paper for alt-i-lab 2004: Prepared on behalf of DEST (Australia) and JISC-CETIS (2004).

Bouquet, P., Giunchiglia, F., van Harmelen, F., Serafini, L. and Stuckenschmidt, H., “C-OWL: Contextualizing Ontologies”, Proc. of the 2nd International Semantic Web Conference, Springer. (2003)

Buckingham Shum, S. “Negotiating the Construction and Reconstruction of Organizational Memories”, Journal of Universal Computer Science, 3(8), Springer (1997)

Butler, D., Avian flu mash-up (2006) http://www.nature.com/nature/googleearth/avianflu1.kml

CanCore Profile, http://www.cancore.ca/indexen.html, http://www.cancore.ca/guidelines/CanCore_Guidelines_Introduction_2.0.pdf

Casey, J., Proven, J., Dripps, D., Geronimo’s Cadillac: Lessons for Learning Object Repositories (2005). http://www.uhi.ac.uk/lis/projects/trustdr/outputs/Geronimo_casey_et_al.doc

CASPAR Project http://www.casparpreserves.eu/

CIDOC: Definition of the CIDOC CRM: http://cidoc.ics.forth.gr.

Coles, S.J., Day, N.E., Murray-Rust, P., Rzepa, H.S., Zhang, Y., Org. Biomol. Chem., (2005) (10),1832-1834. **DOI:** 10.1039/b502828k

Connotea tagging tool http://www.connotea.org/taggingtool

CRM Core definition: http://cidoc.ics.forth.gr/working_editions_cidoc.html

DAREnet http://www.darenet.nl/en/page/language.view/home

DCC Representation Information Registry http://registry.dcc.ac.uk/omar/

Decker, S., Erdmann, M., Fensel, D. and Studer, R. “Ontobroker: Ontology based access to distributed and semi-structured information”, in Semantic Issues in Multimedia Systems, R. M. et al., Ed. Boston/USA: Kluwer Academic Publisher, (1999) pp. 351-369.

Dempsey, L. "The Recombinant Library: Portals and People." Co-published simultaneously in Journal of Library Administration, 39,4; and in Improved Access to Information: Portals, Content Selection, and Digital Information, edited by Sul H. Lee, 103-136. Binghamton, NY: Haworth. (2003)

Dempsey, Lorcan, Eric Childress, Carol Jean Godby, Thomas B. Hickey, Andrew Houghton, Diane Vizine-Goetz, and Jeff Young. "Metadata Switch: Thinking About Some Metadata Management and Knowledge Organization Issues in the Changing Research and Learning Landscape." In *E-Scholarship: A LITA Guide*, ed. Debra Shapiro. (2005) Preprint available at http://www.oclc.org/research/publications/archive/2004/dempsey-mslitaguide.pdf

Dey, A.K. and Abowd, G.D. "Toward a better understanding of context and context-awareness." GVU Technical Report GIT-GVU-99-22, College of Computing, Georgia Institute of Technology (1999)

Dhraief, H. et al.: Open Learning Repositories and Metadata, In Proc. International Semantic Web Working Symposium, California, USA (2001)

Digital Curation Centre http://dcc.ac.uk/

Digital Library Federation, DLF Service Framework for Digital Libraries (2005) report prepared by Dempsey, L. and Lavoie, B. http://www.diglib.org/architectures/serviceframe/dlfserviceframe1.htm

DINI (German Initiative for Networked Information) (2003b). DINI-Certificate. Document and Publication Repositories. October 2003 http://www.dini.de/documents/Zertifikat-en.pdf

DINI (German Inititave for Networked Information) (2003a). Electronic Publishing in Higher Education. How to design OAI Interfaces -Recommendations- October 2003 http://www.dini.de/documents/OAI-Empfehlungen-Okt2003-en.pdf

DINI Schriften 7, 2005: Building on E-Publications Infrastructure (2005). DINI Schriften 7. DINI Working Group "Electronic Publishing". p. 22-23 http://www.dini.de/documents/DINI_Schriften_7-en.pdf

Directory of Open Access Journals DOAJ Lund University http://www.doaj.org/

Directory of Open Access Repositories University of Nottingham and Lund University http://www.opendoar.org/

Doerr, M., "The CIDOC CRM – An Ontological Approach to Semantic Interoperability of Metadata", AI Magazine, 4(1), (2003)

Doerr, M., Hunter, J., Lagoze, C., "Towards a Core Ontology for Information Integration", 2003, In *Journal of Digital information*, volume 4 issue 1, (April 2003)

DRIVER project http://driver-repository.eu/

Dublin Core Metadata Initiative, http://www.dublincore.org/

eBank aggregator service http://eprints-uk.rdn.ac.uk/ebank-demo/

eBank Project http://www.ukoln.ac.uk/projects/ebank-uk/

eCrystals repository http://ecrystals.chem.soton.ac.uk/

e-Framework http://www.e-framework.org/

ePrints UK. http://www.rdn.ac.uk/projects/eprints-uk/

e-Prints Application Profile
http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile
euroCRIS http://www.eurocris.org/en/

European Task Force on Permanent Access http://tfpa.kb.nl/

Ferguson, N., Presentation at JISC-CNI Conference (June 2006)
http://www.ukoln.ac.uk/events/jisc-cni-2006/programme.html

Fischer, G., Ostwald, J., Stahl, G. "Conceptual Frameworks and Computational Support for Organizational Memories and Organizational Learning", Project Proposal, (1997)

Flickr http://www.flickr.com/

Gaines, B.R., Shaw, M.L.G., "Knowledge Management for Research Communities" Proceedings of AAAI Spring Symp. Artificial Intelligence in Knowledge Management, Stanford, (1997)

Goker, A. "User Context Learning for Intelligent Information Retrieval", project proposal at http://www.scms.rgu.ac.uk/staff/asga , Robert Gordon University (1999)

GRADE Project http://edina.ac.uk/projects/grade/

Gray, J. et al. Data Mining the SDSS SkyServer Database. (2002)
http://research.microsoft.com/research/pubs/view.aspx?msr_tr_id=MSR-TR-2002-01

Gross, T., Klemke, R. "Context Modeling for Information Retrieval – Requirements and Approaches". Proceedings of the IADIS International Conference WWW/Internet 2002, ICWI 2002, Lisbon, Portugal, November 13-15, (2002)

Guy, M., Powell, A. and Day, M. (2004). Improving the Quality of Metadata in Eprint Archives. In: Ariadne 38, (Jan. 2004). http://www.ariadne.ac.uk/issue38/guy/

Heery, R. Digital Repositories Roadmap: looking forward, presentation at JISC/CNI Conference (2006) http://www.ukoln.ac.uk/events/jisc-cni-2006/programme.html

Heery, R. and Anderson, S. Digital Repositories Review. Report for the JISC. (2005)
www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf

Heery, R. and Powell, A. Digital Repositories Roadmap: Looking forward. (2006)
http://www.ukoln.ac.uk/repositories/publications/roadmap-200604/

Hellweg, H., Krause, J., Mandl, T., Marx, J., Müller, M. N. O., Mutschke, P., and Strötgen, R. (2001). Treatment of Semantic Heterogeneity in Information Retrieval. IZ-Arbeitsbericht Nr. 23, (May 2001) 47 pp.
http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab_23.pdf

Hillmann, D. I., Dushay, N. and Phipps, J. (2004). Improving Metadata Quality: Augmentation and Recombination. DC2004 Conference, Shanghai. (2004)
http://metamanagement.comm.nsdl.org/Metadata_Augmentation--DC2004.html

Hunter, J. Proposal for the Integration of DublinCore and MPEG-7, DSTC Pty Ltd, (2000). http://www.egov.vic.gov.au/Documents/DCforMPeg7.doc

IEEE Standard for Learning Object Metadata. http://www.ieee.org/

IFLA Study Group on the functional requirements for bibliographic records. "Functional requirements for bibliographic records : final report", volume 19 of UBCIM Publications : New Series. K. G. Saur, Munich, (1998)

IMS Learning Information Package Specification, http://www.imsglobal.org/profiles/index.cfm

IMS Learning Resource Meta-data Specification, http://www.imsproject.org/metadata

Institutional self-archiving policies in ROARMAP http://www.eprints.org/openaccess/policysignup/sign.php

JORUM http://www.jorum.ac.uk/

Kantor, M., Zimmermann, B., Redmiles, D. "From Group Memory to Project Awareness through use of the Knowledge Depot", Proceedings of the California Software Symposium (CSS'97), California (1997).

Karanastasi, A., Kazasis, F.G., Christodoulakis, S., "A Natural Language Model and a System for Managing TV-Anytime Information from Mobile Devices" , In Proc of the 9th International Conference on Applications of Natural Language to Information Systems (NLDB), Manchester, United Kingdom, (June 2004).

Kim, H.R., Chan, P. K. "Learning Implicit User Interest Hierarchy for Context in Personalization", In Proceedings of IUI03, (2003).

Kimbrough, S.O., Oliver, J. R. "On Relevance and Two Aspects of the Organizational Memory Problem", Proceedings of the 4th Workshop on Information Technology and Systems, (1994).

Klamma, R., Schlaphor, S. "Rapid Knowledge Deployment in an Organizational-Memory-Based Workflow Environment", 8th European Conference on Information Systems, ECIS 2000, Vienna, (July 2000).

Klemke, R., "Context Framework – an Open Approach to Enhance Organizational Memory Systems with Context Modeling Techniques". Proceedings of the Third International Conference on Practical Aspects of Knowledge Management, (2000).

Klemke, R., Koenemann, J., "Supporting Information Brokers with an Organizational Memory", in: XPS-99, 5th German Conference on Knowledge-Based Systems, Workshop on Knowledge Management, Organizational Memory and Reuse, Internal Report, Wurzburg University, (March 1999).

Knight, R., 2006, Podcast pedagogy divides opinion at US universities, in: Financial Times, (10 February 2006).

Krause, J., 2004. Standardization, Heterogeneity and the Quality of Content Analysis: a key conflict of digital libraries and its solution. IFLA Journal 4/2004 (2004), pp.310-318 http://www.ifla.org/V/iflaj/lj-4-2004.pdf

Learning Federation, http://www.thelearningfederation.edu.au/tlf/

LeBoeuf, P. (Editor), "Functional Requirements for Bibliographic Records (Frbr): Hype or Cure-All?", Haworth Press, (January 2005) ISBN: 0789027984
Lenat, D., The Dimensions of Context-Space. http://www.cyc.com/publications.html, (1998)

Lippincott, J., 2006, Presentation "Net Gen students: learning, technology and libraries", Presented at JISC-CNI Conference, York, (June 2006). http://www.ukoln.ac.uk/events/jisc-cni-2006/presentations/ppt-2000-html/j-lippincott-2_files/v3_document.htm

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

Lord, P., Macdonald, A., Sinnott, R., et al. Large-scale data sharing in the life sciences: data standards, incentives, barriers and funding models. (2005)

Lynch, C., 2003, "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age," *ARL Bimonthly Report* 226 (February 2003), 1-7. Online at http://www.arl.org/newsltr/226/ir.html. Reprinted in Portal: Libraries and the Academy 3:2 (2003), pp. 327-336.

Lyon, L, 2006, Digital Libraries and the business process: Presentation at British Library/JISC/UKOLN workshop. (March 2006). http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/presentations.html#2006-03-london

Lyon, L., 2003, eBank UK: Building the links between research data, scholarly communication and learning. In Ariadne, Issue 36, (July 2003) http://www.ariadne.ac.uk/issue36/lyon/

Lyon, L., 2006, Digital libraries and digital scholarship: changing roles and responsibilities? Presentation at SCONUL Conference 2006. http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/ppt-2000-html/sconul-june06_files/v3_document.htm

Jarke, M., Klemke, R.., Nick, A. "Broker's Lounge - An Environment for Multi-Dimensional User-Adaptive Knowledge Management", 34th Annual Hawaii International Conference on System Sciences (HICSS-34)-Volume 3

Ma, Jacky C.K. and Lyu, Michael R. Design and Implementation of XML-Based Digital Video Library System. In Proceedings of the 11th International World Wide Web Conference, 2002. (http://www2002.org/CDROM/poster/184.pdf)

Mach, M., Sabol, T., Paralic, J., Kende, R. "Knowledge Modeling in Support of Knowledge Management", Proceedings of Conference on Research Information Systems, CRIS 2000

Maurer, F., Dellen, B. "A Concept for an Internet-based Process-oriented Knowledge Management Environment", 11th Workshop on Knowledge Acquisition, Modeling and Management KAW'98, Banff, 1998

McLean, N. and Lynch, C., Interoperability between Information and Learning Environments--Bridging the Gaps. White paper produced jointly by IMS Global Learning Consortium and the Coalition for Networked Information www.imsproject.org/digitalrepositories/CNIandIMS_2004.pdf

Metros, S. and Bennett, K.: Learning Objects in Higher Education. Educause Center for Applied Research (2002), http://www.educause.edu/asp/doclib/abstract.asp?ID=ERB0219

METS (Metadata Encoding and Transmission Standard): http://www.loc.gov/standards/mets

Microsoft Research, 2020 Science (2006)
http://research.microsoft.com/towards2020science/downloads.htm

http://research.microsoft.com/towards2020science/background_overview.htm

MIDESS Project http://www.leeds.ac.uk/library/midess/

myGrid Project http://www.mygrid.org.uk/

NaCTeM http://www.nactem.ac.uk/

National Science Board/National Science Foundation (NSB/NSF), 2005, Long-lived digital data collections: enabling research and education in the 21st Century
http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf

Nature, 2020 Future of Computing, Nature (23 February 2006)
http://www.nature.com/nature/focus/futurecomputing/index.html

Nature Publishing "Nascent" blog entry (June 2006)
http://blogs.nature.com/wp/nascent/2006/06/natures_peer_review_debate_and.html

Nejdl, W. et al.: EDUTELLA: A P2P Networking Infrastructure Based on RDF, In Proceedings of International WWW Conference(11), Honolulu, Hawaii, USA (2002).

Ngai, C.H., Chan, P.W., Yau, E., and Lyu, M.R. XVIP: An XML-Based Video Information Processing System. In: Proceedings of the 26th Annual International Computer Software and Applications Conference, page 173, 2002.

NISO Standard (Z39-85), http://www.niso.org/Z3985.html

Open Society Institute (OSI), 2004 See:
http://www.soros.org/openaccess/pdf/OSI_Guide_to_IR_Software_v3.pdf

Palmir, M., Naeve, A., Nilsson, M.: E-learning in the Semantic Age. In Proceedings of the 2nd European Web-based Learning Environments Conference (WBLE 2001), Lund, Sweden. [Web Page]. URL: http://kmr.nada.kth.se/papers/SemanticWeb/e-Learning-in-The-SA.doc

Papadakis, I. and Chrissikopoulos, V, 2000, A Digital Library Framework based on XML. In Proceedings of the 3rd International Conference of Asian Digital Library (ICADL'00), pages 81-88, 2000

PAPI Learner Specification, http://edutool.com/papi/

Patel, M., Koch, T., Doerr, M. and Tsinaraki, C., 2005, Semantic Interoperability in Digital Library Systems. DELOS Network of Excellence on Digital Libraries, European Union, Sixth Framework Programme. Deliverable D5.3.1 June 2005.
http://www.ukoln.ac.uk/ukoln/staff/t.koch/publ/SI-in-DLs.doc

Porzel, R., Strube, M. "Towards Context-adaptive Natural Language Processing Systems", Computational Linguistics for the New Millenium: Divergence or Synergy. Lang. Frankfurt am Main, (2002).

Powell, A., JISC Information Environment Architecture http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/

Prinz, W., "TOSCA – Providing organizational information to CSCW Applications", Proceedings of Int'1 Conference on Computer supported Cooperative Work (ECSCW'93) Kluwer Press, (1993).

R4L Project http://r4l.eprints.org/about.html

RCUK Updated Position Statement on access to research outputs. (June 2006). http://www.rcuk.ac.uk/access/2006statement.pdf

Registry of Open Access Repositories University of Southampton http://archives.eprints.org/

Reimer, U. "Knowledge Integration for Building Organizational Memories", Proceedings of 11th Workshop on Knowledge Acquisition, Modeling and Management KAW'98, Banff, (1998).

Renardus. See http://www.renardus.org/

RepoMMan Project http://www.hull.ac.uk/esig/repomman/

RLG/NARA Task Force on Repository Certification. Audit Checklist for Certifying Digital Repositories, 2005 http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf

Reusable Learning Objects (RLOs) – a project at Tropical North Queensland (TNQ), Technical and Further Education (TAFE) Institute, Australia (2003)

Riche, S., and Brebner, G. "Storing and Accessing User Context", Proceedings of the 4th International Conference on Mobile Data Management 2003

Rights & Rewards Project http://rightsandrewards.lboro.ac.uk/index.php?section=1

RoMEo Project http://www.lboro.ac.uk/departments/ls/disresearch/romeo/index.html

Schatz, S. Paradigm Shifts and Challenges for Instructional Designers. IMS Global Learning Constortium, Inc. (2000), http://www.imsproject.org/feature/kb/knowledgebits.html

Scholar's Copyright Project http://sciencecommons.org/literature/scholars_copyright

Schwartz, D.G., "Towards the use of user-centric meta-knowledge in applying organizational memory to email communications", Proceedings of European Conference on Artificial Intelligence (ECAI'98)

SeLeNe Project, http://www.dcs.bbk.ac.uk/selene/

SHERPA/JULIET page http://www.sherpa.ac.uk/juliet/

SHERPA/RoMEo page http://www.sherpa.ac.uk/romeo.php

Simon, B. et al.: Elena: A Mediation Infrastructure for Educational Services. In Proc. of WWW Conference, Budapest, Hungary (2003)

Simon, B., Miklos, Z., Nejdl, W., Sintek, M., Salvachua, J.: Smart Space for Learning: A Mediation Infrastructure for Learning Services. In Proceedings of the Twelfth International Conference on World Wide Web, Budapest, Hungary (2003)

Sinclair, P., Addis, M., Choi, F., Doerr, M., Lewis, P., and Martinez, K., "The use of CRM Core in Multimedia Annotation", to appear in: Proceedings of First International Workshop on Semantic Web Annotations for Multimedia (SWAMM 2006), part of the 15th World Wide Web Conference (22-26 May 2006, Edinburgh, Scotland

Smart Tea Project http://www.smarttea.org/

StORe Project http://jiscstore.jot.com/WikiHome

Strang, T. and Linnhoff-Popien, C., "Service interoperability on context level in ubiquitous computing environments", in Proceedings of International Conference on Advances in Infrastructure for Electronic Business, Education, Science, Medicine, and Mobile Technologies on the Internet (SSGRR2003w), L'Aquila/Italy, January 2003.

Strang, T., Linnhoff-Popien, C., and Frank, K., "Applications of a Context Ontology Language", in Proc. of the International Conference on Software Telecommunications and Computer Networks (SoftCOM 2003)

Swan, A. and Awre, C. Linking UK Repositories 2006 http://www.jisc.ac.uk/uploaded_documents/Linking_UK_repositories_report.pdf

Tane, J., Schmitz, C., Stumme, G., Staab S., Studer, R.: The Courseware Watchdog: an Ontology-based tool for finding and organizing learning material. In: Mobiles Lernen und Forschen Beiträge der Fachtagung an der Universität Kassel, Kassel university press , ISBN 3-89958-521-6 (2003)

TEI (Text Encoding Initiative): http://www.tei-c.org

Ternier, S., Duval, E. and Neven, F.: Using a P2P architecture to provide interoperability between Learning Objects. EdMedia 2003 - World Conference on Educational Multimedia, Hypermedia & Telecommunications, Honolulu, Hawaii, USA (2003)

Ternier, S., Duval, E., and Vandepitte P. LOMster: Peer-to-Peer Learning Object Metadata. Proceedings of EdMedia-2002 (pp. 1942-1943), Denver, Co.

Tudhope, D., Koch, T. and Heery, R. (2006, in publication). Terminology Services and Technology. JISC state of the art review. 93pp.

UK Metadata for Education Group http://www.ukoln.ac.uk/metadata/education/

Van Heijst, G., Van der Spek, R., Kruizinga, E., "Organizing Corporate Memories", Proceedings of 10th Knowledge Acquisition for Knowledge-Based Systems Workshop, 1996

Versions Project http://www.lse.ac.uk/library/versions/

Vladoiu, M.: Learning objects need badly instructional digital libraries support. International Journal Informatics in Education, Institute of Mathematics and Informatics, Lithuanian Academy of Sciences, Lithuania, Vol.2, No.3, pp. 1-25, September, 2003

Wagealla, W., Terzis, S., English, C., "True-based model for privacy control in context-aware systems". Proceedings of 2nd Workshop on Security in Ubiquitous Computing (Ubicomp2003)

Wargitsch, C., Wewers, T., Theisinger, F. "An Organizational-Memory-Based Approach for Evolutionary Workflow Management System – Concepts and Implementation", Proceedings of the 31st Annual Conference on System Sciences, Vol.1, Los Alamitos 1998

Warner, Simeon : Posting on JISC-repositories discussion list 16 June 2006

Wellcome Trust Self-archiving Policy http://www.wellcome.ac.uk/doc_WTD002766.html

Wetterling, J.: Re-Usability of Telematics Course Material within a European Network: The CANDLE Project. In. Proc. EDEN Annual Conference (2003)

Wiley, D.A. (2001b): Connecting learning objects to instructional design theory: a definition a metaphor, and a taxonomy. In D.A. Wiley (Ed.), The Instructional Use of Learning Objects. Association for Educational Communications and Technology, Bloomington (2001), http://reusability.org/read/chapters/wiley.doc

Wolverton, W. "Exploiting Enterprise Models for the Automatic Distribution of Corporate Information". Proceedings of the 6th Int'1 Conference on Information and Knowledge Management (CIKM'97), Las Vegas, 1997

Zadrozny, S., Kacprzyk, J. "FQUERY for Access: towards human consistent querying user interface", Proceedings of the 1996 ACM symposium on Applied Computing

Appendix I: Educational Digital Libraries (DLs) and Courseware Repositories

In this appendix we list the major educational digital libraries, as made by the ADL (Advanced Distributed Learning) Co-Laboratories 0:

Apple Learning Interchange (ALI)

URL: http://newali.apple.com/ali_sites/ali/nav5.shtml

Sponsored by Apple Computer, Inc. the Apple Learning Interchange describes itself as an "online magazine where educators using Apple and other related software and hardware shared resources that support quality educational practice". The site was launched in 1998, delivering text-based units of practice and maintaining a collection of those practices and other teacher resources. Recently the collection has showcased media-based exhibits.

The ALI Web site provides access to a collection of resources developed by educators and by ALI itself to encourage the use of information technologies in education and to improve educational practice generally.

The ALI collection is made up of "Exhibits", units intended for educators interested in improving their classroom practice. Some of the Exhibits are of teaching techniques requiring the use of computer technologies, but many can be practiced without using such technologies in the classroom.

Blue Web'N

URL: http://www.kn.pacbell.com/wired/bluewebn/

Part of the SBC Knowledge Network, Blue Web'N provides users with access to 1200 educational sites covering all ages and subject matters. Metadata for materials include Dewey Decimal Numbers. Materials are rated by onsite staff, and users may suggest sites but do not supply metadata.

Canada's SchoolNet

URL: http://www.schoolnet.ca/home/e/

Mandated by the Canadian government to work in partnership with the provincial and territorial government to connect Canadian schools and libraries, SchoolNet provides users with access to over 5,000 teacher-approved learning resources for teachers, students, and parents. The resources are searchable with basic keyword searches and advanced searches. Users suggest resources and approval of resources as well as metadata creation are performed by the site staff.

CAREO (Campus Alberta Repository of Educational Materials): Learning Commons Educational Object Repository

URL: http://careo.ucalgary.ca/cgibin/WebObjects/Repository.woa?theme=default

CAREO has developed a prototype repository of over 3,500 educational materials on various subjects and interactivity levels as part of a larger project to create "a searchable, Web-based collection of multidisciplinary teaching materials for educators" for Alberta and elsewhere.

The prototype repository is part of an ongoing research project that places it under revision occasionally. The repository may at times be offline for updates and upgrades as the project evolves.

Currently the CAREO collection is made up of materials submitted by users who have signed on as members of the project. Submitted resources are reviewed by an editorial review board. Each item may be discussed, allowing users to describe a resource or add comments to the resource record. CAREO does not have a permanent protocol for metadata creation and import but it does and will use CanCore to describe its materials.

Computer Science Teaching Center (CSTC)

URL: http://www.cstc.org

Sponsored by the National Science Foundation and the Association for Computing Machinery Education Board, the CSTC collection consists of roughly over one hundred materials useful in college and graduate level computer science education. Much of the material is not browser based but is downloaded directly and run off the user's system. Metadata for the materials are created by the users.

Computing and Information Technology Interactive Technology Interactive Digital Education Library (CITIDEL)

URL: http://www.citidel.org

CITIDEL serves to "establish, operate, and maintain a part of the NSDL that will serve the computing education community in all its diversity and at all levels". CITIDEL activities include community development, expanding through workshops knowledge and skills regarding the development and use of online educational content. CITIDEL currently contains 445699 resources from 9 source collections. Most of these resources are articles and technical reports but the collection does contain some materials created for the educational setting.

Connexions

URL: http://cnx.rice.edu

The Connexions project at the Rice University has created an open repository of educational materials and tools to promote sharing and exploration of knowledge as a dynamic continuum of interrelated concepts. Available free of charge to anyone under open-content and open-source licences, Connexions offers high-quality, custom-tailored electronic course material, is adaptable to a wide range of learning styles, and encourages students to explore the links among concepts, courses, and disciplines. Connexions fosters worldwide, cross-institution communities of authors, instructors, and students, who collaborate on the creation of knowledge building blocks from which courses are constructed. The ideas and philosophy embodied by Connexions have the potential to change the very nature of teaching and learning, producing a dynamic, interconnected educational environment that is pedagogically sound, both time and cost efficient, and engaging.

Materials within the Connexions collection can be created and joined together using authoring tools provided by the project. Using these tools creators can create or modify knowledge modules representing individual concepts, identify links to other knowledge modules or Internet resources, annotate module content, share a working space for working drafts while the content is under development, search the repository for other relevant modules, and perform many other functions prior to committing the knowledge building block to the repository.

Co-operative Learning Object Exchange (CLOE)

URL: http://lt3.uwaterloo.ca/CLOE/

A collaborative project of fifteen Ontario universities aims to "develop an innovative infrastructure for joint development of multimedia-rich learning resources". CLOE is developing a "virtual market economy" of learning objects. This economy of learning materials will encourage users to create materials and submit them to the collection using access to other learning materials as an incentive. The collection, which can be accessed only by member institutions, contains over 60 materials, all of which are interactive browser based materials combining various assets into the learning experience.

Metadata are created by users who submit them. CLOE's metadata schema is drawn from IEEE LOM. Most of the current fields are not required for resource submission.

CLOE is gathering extensive information on the re-use of learning materials, is applying it to community development and encourages participation in the learning materials economy.

Digital Library for Earth System Education (DLESE)

URL: http://www.dlese.org

DLESE, funded by the NSF, provides users with access to over 3,500 materials of interest to educators and learners to support Earth system science education. The collection includes resources such as lesson plans, maps, images, data sets, visualizations, assessment activities, curricula, online courses and other materials. DLESE prioritizes community building, which is reflected in its Web page, containing sections on using online resources in education, news from the fields of earth science and science education and information on resources for professionals within the earth sciences.

EducaNext (UNIVERSAL)

URL: http://www.educanext.org/ubp

EducaNext provides access to thousands of resources on various subjects of relevance to teaching and learning at the college level and higher. Member institutions list their materials on the site as well as fees for those materials. The site provides a forum for the creation of a market for learning materials. Member institutions are solely responsible for the creation and accuracy of Metadata, which is based upon Dublin Core, and IEEE LOM.

Education Network Australia (EdNA)

URL: http://www.edna.edu.au/go/browse/

Supported by education.au, a non-profit company limited by guarantee and owned by the Australian education and training Ministers, EdNA provides users with access to over 16,000 materials of interest to teachers and learners of all levels, covering a wide range of subjects.

Resources are suggested to EdNA by users. If these materials are accepted into the collection, metadata are applied onsite according to the EdNA metadata schema, which draws on the Dublin Core. EdNA draws on thesauri to apply natural language terms to materials and thus encourages resource discovery.

Metadata that conform to the EdNA metadata standard or Dublin Core can be harvested from other collections. EdNA harvests from selected collections within Australia. This harvesting is part of a movement on EdNA's part to develop a distributed repository of learning materials for Australia.

Eisenhower National Clearinghouse for Mathematics and Science Education (ENC)

URL: http://www.enc.org/resources/collect

Initially developed in 1992 as a collection of K-12 teaching materials within mathematics and the sciences, and information concerning federal funding for education, ENC has come to deliver a wide range of digital content to educators. Online lessons and activities are listed within the ENC online catalog of educational materials.

Located on The Ohio State University campus and funded through a contract with the U.S. Department of Education, the ENC has a huge collection of materials either of use in the classroom itself or that can supplement and direct lessons or pedagogical methods. ENC describes the collection as "the best selection of math and science education resources on the Internet". Its collection of online materials numbers in the tens of thousands. ENC both links to sites containing educational content and holds materials at its own site.

Enhanced and Evaluated Virtual Library (EEVL)

URL: http://www.eevl.ac.uk/index.htm

Based at the Hariot Watt University in Edinburgh and funded by the UK's Joint Information Systems Committee (JISC) as part of the Resource Discovery Network (RDN), EEVL provides users with access to over 10,000 resources useful to teachers and learners within engineering, mathematics, and computing. Originally a virtual library of engineering materials, this service expanded in 2001 to including mathematics and computing. Users may recommend material, submitting descriptions and keywords. Other metadata, including EEVL's subject classifications, are applied to the materials onsite.

European Knowledge Pool System (ARIADNE)

URL: http://rubens.cs.kuleuven.ac.be:8989/silo/

Developed to deliver educational content throughout Europe, the KPS facilitates the sharing and re-use of educational resources. The encouragement of the discovery and re-use of these materials encourages an increasing recognition that learning object production is a valid field of activity for academics. The collection contains materials of a wide variety of interactivity levels in many European languages, primarily English, French, Italian, German and Dutch.

The collection contains a variety of materials, primarily text documents, followed in order of frequency by hypertext, slide sets, video clips and interactive educational objects. Interactive objects include documents like multiple-choice questionnaires, quizzes, auto-evaluations, and simulations.

Exploratories

URL: http://www.cs.brown.edu/exploratories/home.html

This project, dedicated to producing electronic materials for use within courses has developed 71 Java Applets demonstrating concepts in science and mathematics. The materials are for college and graduate teaching. Materials have no searchable metadata but can be accessed by browsing the site of the project.

Fathom Knowledge Network Inc.

URL: http://www.fathom.com

Fathom provides users with access to around 300 free "courses", each requiring about 2 hours to complete. The "courses" cover a variety of topics and are meant for learners in higher and continuing education. Users must sign up for an account with Fathom to use these "courses", but use is free. Materials are submitted by a variety of member institutions including the British Museum, the American Film Institute and RAND, each of which holds the copyright to the materials they submit to the Fathom collection.

Gateway to Educational Materials (GEM)

URL: http://www.thegateway.org

Established in 1997 and funded by the U.S. Dept. of Education, GEM gives users access to a collection of over 31,000 materials, both educational for students and for the aid of educators. Materials within the GEM collection range over all educational levels and subjects.

The contents of the GEM collection are made up of "quality lesson plans, curriculum units and other educational resources on the Internet". Its resources are useful to educators with various levels of access to technology in the classroom. Some materials are browser based, interactive learning materials, others are lesson plans or class materials that are either for teacher use or must be printed out to be used by students.

GEM metadata records are almost always complete and include useful description fields that explain the contents and uses of a resource. Other fields specify the use of a material for teachers and/or students, and the instructional strategy the resource has been developed for.

Harvey Project

URL: http://harveyproject.org

This site provides users with access to 636 interactive materials within the domain of Physiology. The materials are created for college and medical students. Materials can be browsed by subject. Materials are also classified according to domains within physiology.

Health Education Assets Library (HEAL)

URL: http://www.healcentral.org/index.htm

Conceived in 1998, and having begun its collection development in 2002, HEAL provides "building-block multimedia items such as images, videos, and animations, and textual materials such as cases and quiz questions". Its current prototype collection of over 1,000 materials contains resources useful to medical students and medical professionals. The collection will eventually contain materials of use to all educational levels. The collection currently contains images and interactive tutorials.

Submitters of materials make use of a cataloging interface that enforces compliance with the HEAL metadata standard, which builds on the IMS Metadata specification. HEAL's metadata also includes fields relating to resource attributes of interest to persons in the medical professions or education, such as SpecimenType and Organ. HEAL provides visitors to its site with examples of the XML format of its metadata records.

Humbul Humanities Hub

URL: http://www.humbul.ac.uk/

Developed in 1999 by the University of Oxford's Joint Information Systems Committee – Arts and Humanities Research Board, the Humbul Humanities Hub is a catalog of humanities resources online. The project hopes that its site will be the "first choice for accessing online humanities resources".

Humbul refers to a variety of materials through its repository. Collecting materials to be used primarily by educators and students, Humbul has put together a collection that includes educational materials and links to academic institutions, as well as academic research projects and various resources as the Web pages of companies producing educational software.

Materials within the Humbul collection are submitted by users. Users submit a URL, which is then reviewed by on-site staff, who then passes materials on to a subject specialist cataloger.

Humbul metadata are based on the Dublin Core metadata element set. Humbul's guide for catalogers of materials includes a description of various metadata fields as well as rules for recording author and institutional names. Dublin Core controlled vocabularies are used for resource types.

iLumina

URL: http://www.iLumina-dlib.org

Funded by the NSF, iLumina provides access to nearly 1,500 materials useful for undergraduate teaching of Chemistry, Biology, Physics, Mathematics, and Computer Science. After submission, materials and metadata are reviewed by an editor.

Metadata is assigned using a series of forms and drop down menus that enforce standards compliance. Repository metadata are based upon the IMS Metadata specification and Dublin Core. iLumina classifies materials using a "modified" set of Library of Congress subject headings for Chemistry, ACM/IEEE Computing Curricula for Computer Science, and its own system for Mathematics and Physics.

Learning Matrix

URL: http://thelearningmatrix.enc.org

Funded by the Eisenhower National Clearinghouse and the National Science Foundation the Learning Matrix collection provides access to over 4,000 online resources useful to teachers of science and mathematics, or providing instructional and pedagogical training. Learning Matrix resources promote "inquiry and problem based learning in college mathematics, science, and technology classes". According to the project's own literature the collection contents range from "simulations and tutorials to research articles and video footage illustrating excellent teaching techniques".

Metadata for the Learning Matrix use Dublin Core and IEEE LOM elements. The IMS Learning Resource Metadata Specification is also drawn on. To better describe multimedia objects indexing protocols follow a modified POOL-IMS Version 1.0 that is itself based on the IMS Specification.

Learning Object Repository, University of Mauritius

URL: http://vcampus.uom.ac.mu/lor/index.php?menu=1&cat=10

Established to collect learning resources from the University of Mauritius and elsewhere, this English language site, containing materials in both English and French, provides users with access to over 300 learning materials on various topics for students at college level and higher. Submitters, the population of which is restricted to faculty and invited “guests”, submit resources and create descriptive metadata.

LearningLanguages.net

URL: http://www.learninglanguages.net

LearningLanguages.net is a portal that brings together online foreign language resources for English-speaking K-12 students and teachers of French, Spanish and Japanese. The project was created and is maintained and enhanced by staff at the Internet Scout Project. The project was initially funded by the Claire Giannini Hoffman Fund. “Content experts scour the web every week for the best resources, and then describe and catalogue them in order to make them accessible”. Production of Metadata follows a metadata schema based on the Dublin Core Metadata Element Set. The project uses the Scout Portal Toolkit to create a searchable portal to selected resources.

Maricopa Learning Exchange (MLX)

URL: http://www.mcli.dist.maricopa.edu/mlx

Established by the Maricopa Community Colleges in Arizona, the Maricopa Learning Exchange provides users with access to over 500 materials on various subjects. Materials, submitted by users within the Maricopa community college network, range from teaching materials and pedagogical tips to links to sites containing educational content. The site has an RSS feed to allow communication of new records. Users are permitted to comment on **resources.**

MIT OpenCourseWares

URL: http://ocw.mit.edu/index.html

MIT’s Open CourseWare Initiative is a “large-scale, Web-based electronic publishing initiative” funded by The William and Flora Hewlett Foundation, the Andrew W. Mellon Foundation, and MIT itself. It has been created in order to “Provide free, searchable, access to MIT’s course materials for educators, students, and learners around the world”. The project has placed materials for 500 classes, including Syllabi, Course Calendars, Readings, Lecture Notes, Assignments, and Exams, online. The OCW’s metadata is based upon the IMS Metadata Specification.

Multimedia Educational Resource for Learning and On-Line Teaching (MERLOT)

URL: http://www.merlot.org

Developed in 1997 by the California State University Center for Distributed Learning, MERLOT provides users with access to online learning materials through a collection of records that now totals over 8,000. Many materials described within the collection are peer reviewed by onsite staff and rated by users. Materials can also be collected by users into collections of recommended resources, or listed along with assignment or project concepts that use the listed resources.

MERLOT metadata records, the structure and fields of which are developed from the IMS Learning Resource Metadata Specification, are created by users submitting resources. MERLOT has developed a one page form for assigning metadata. The form enforces compliance with controlled vocabularies using drop down boxes and check boxes.

National Engineering Education Delivery System (NEEDS)

URL: http://www.needs.org

NEEDS is a digital library of resources for educators and learners within education. Records are drawn from various smaller collections. Cataloguing records contain multiple searchable fields including type of learning resource and publication year. Metadata is based on IEEE LOM. The collection will soon be searchable by user education level and for the existence of peer reviews.

National Learning Network (NLN): Materials

URL: http://www.nln.ac.uk/Materials/default.asp

The National Learning Network, a government-supported project based in the UK, was inaugurated in 1999 to encourage the adoption of Information and Learning Technology in post-16 education. The NLN collection, which was launched in 2001, exists in order to provide high quality materials with which to encourage this adoption. This collection is intended to eventually contain a large variety of materials, including lesson plans.

The NLN collection currently contains roughly 500 interactive browser-based resources for online learning. User capacity to search resources is limited. Resources are best accessed through a subject list that can be browsed or users can perform keyword searches for resources.

NLN uses the BECTA (British Educational Communications and Technology Agency) Metadata Requirements, based on the IMS Metadata specification, to apply metadata to its resources. Subject classifications for materials use the LearnDirect classification system, developed from the Superclass II classification system.

Scottish Electronic Staff Development Library (SeSDL)

URL: http://www.sesdl.scotcit.ac.uk

SeSDL is a resource center “designed to encourage the sharing and re-use of staff development materials”. The contents are intended to encourage the use of communications and information technologies in education and individual learning. The materials in the current collection relate to the use of technology in teaching.

The collection is made up of “granules”, the “smallest individual components into which educational materials can be divided”. As the project staff writes, “Granules may be composed of a single file, or a collection of files including text, diagrams, video sequences, interactions, etc.”. Granules can be drawn together using a Course Design tool provided by the site itself to create lessons. The collection currently contains 8 lessons made up of 96 granules. Use of the materials is restricted to individuals within member Scottish institutions of Higher education.

Guides are provided for the creation and design of granules, presenting rules regarding granules referring to other granules, rules for navigation among granules, and the bundling of files within them. Users may update materials themselves, making versioning the responsibility of the creator. The project metadata are developed from the IMS metadata specification. A subject taxonomy and a controlled vocabulary have been developed, drawing from the British Education Thesaurus.

Telecampus Online Course Directory

URL: http://courses.telecampus.edu/about/index.cfm?fuseaction=introduction

Organized by NBDEN Inc., Telecampus is a directory of over 60,000 online courses and resources available from pay and free collections. Materials range in length from one hour modules to full semester or year-long courses. Materials cover a wide range of subjects and educational levels. Institutions with collections of online courses and materials apply for inclusion in the Telecampus directory. All materials are from accredited institutions. Metadata applied to materials build off of IMS metadata.