DARE to keep e-material ready for the future!

Paper presented at IATUL Conference, Ankara

June2, 2003

Maria Heijne, MA
Delft University of Technology
Librarian/Director

Contents

Digital provision of academic information

As a result of the rapid development of ICT (information and communications technology), the world of the provision of academic information has undergone radical change in recent years. Besides the traditional channels – academic associations, conferences, journals – a virtual world has come into being in which a large part of the exchange of academic information takes place. This is being done in new forms, and the division of roles between the parties involved is also changing.
The Netherlands has long played an important role in academic communications, and its academic output is impressive. Scholars, and the institutions with which they are connected, are keen to be part of the new developments. They want to continue to play their roles and keep research and development in the Netherlands up to the mark. The acceleration in the trends that are taking place and the pressure on financial resources is increasing the need for an efficient communications infrastructure.

The present system for handling academic output (in a nutshell)

Academic (scientific) output is, in this context, defined as written results of scientific research, generally in the form of text. The core of scientific output consists of journal articles and books written by scientists and published by commercial publishers, who (sometimes) pay royalties to the scientist for their copyright. These publications derive their quality and status for a large part from a peer review process, performed by colleagues in the academic environment. Publishers sell the publications to the public at large and to (academic) libraries. Libraries make the publications available to their circle of clients, consisting mostly of students and scientists who have an interest in the work of their peers. Universities use the number and the quality of publications as a measure of quality of individual scientists and of the University as a whole. Some universities construct collections of scientific documents (repositories) and grant access to Service Providers, to 'harvest' (selectively collect) information about scientific documents, for the purpose of constructing discipline-specific indexes and data collections.

Problems with present system

Following are some (selected) problems with the existing system:

- Scientific output, worth to be stored and preserved, is not strictly limited to (commercially) published text. Many other 'objects' are worth retaining together with or in addition to the final texts, e.g.
  - draft versions e.g. pre-print material
  - texts by others than scientists, e.g. students
  - texts that are not for publication e.g. teaching notes
  - supporting material, e.g. research datasets.
- The entire process is felt to be slow and cumbersome and scientists find that it gives them inadequate visibility.
- Scientists find it difficult to locate and retrieve suitable publications.
- Service Providers find it difficult to harvest information from a wide variety of scientific institutes.
- Universities find it difficult to collect a reliable overview of scientific output from their institute and spend increasing sums to construct systems to achieve this.
- There is much duplication in the storage of scientific output, leading to gross inefficiency.
- Individual scientists and an increasing number of universities wish to diminish the role of commercial publishers and to assume a greater role for themselves. At the same time they feel they can't do without those same publishers (love you, hate you dilemma).
- Universities find the cost of acquiring and maintaining the publications unbearable and mounting.

All the above problems appear to point in the direction of the need for a digital collection ('repository') of scientific output and related material, controlled by the Universities and freely accessible, through a digital network, to approved scientists, students and Service Providers: in other words a Networked Repository.

Improvements sought by the various parties

The SPARC document "Institutional Repository Checklist & Resource Guide" [reference 3] dated 2002, provides a good framework for the wants, needs and incentives of the various parties involved in a coalition to develop (a network of) digital repositories. We summarize the main points, by interested party:

- Scientists wishing to demonstrate the fruits of their work want:
  - greater ease in producing results and linking related material
  - faster dissemination of results
  - better bargaining position for copyright
  - easier access to quality processes (such as peer review)
  - assured archiving and retention of information

- Scientists involved in collaborative research and seeking related information need:
  - assured access to all related research results
  - fast and easy reconstruction of data into another format, e.g. for e-learning

- Teaching faculty have a need for:
  - a resource supporting classroom teaching
  - access to teaching material including visualizations, models, course videos, and the like

- Service Providers, involved in constructing and disseminating indexes and collections of scientific work, look for:
  - a fast and robust way of interfacing with all significant collections of scientific documents

- University administrators aim for:
  - less expenditure on books and journals
  - a comprehensive and authoritative collection of all research results of the institution, in order to apply scientific staff evaluation, promote the status of the university and support claims for funds
  - less costs of computer resources for storing and handling research data

In recent years Dutch academic institutions have, of course, responded also to the developments surrounding the digital provision of academic information by means of diverse experimental projects. Each did so in its own way or in smaller partnerships, taking as their starting point the realisation that not only the creation of knowledge but also its communication was part of their tasks. Delft University started the project Roquade, in collaboration with Utrecht University [reference 10]. This project has been a subject for the IATUL conference in earlier years. Another example is the E-Archive project that is discussed later on in this presentation.
The diversity and local character of the various experiments created a situation in which the rich source of academic knowledge and information in the Netherlands is still not properly accessible; improvement is desirable and possible. At the same time that diversity and local character create the impression that the academic institutions have a fragmented image of the (possible and desirable) direction development should take. This is not the case, however, it is clear that broadly speaking all the institutions are working on similar questions.

DARE: solutions through collaboration!

So it is that the boards of the Dutch universities recently decided to combine their efforts in a single new initiative, the establishment of a digital platform for academic information: DARE (Digital Academic Repositories). The Royal Library, the Royal Netherlands Academy of Arts and Sciences and the Netherlands Organisation for Scientific Research (NWO) are collaborating in this initiative. SURF, a foundation aimed at supporting ICT needs in higher education, is coordinating the project, which covers a four-year period (2003-2006).
The DARE programme received a government grant of Euro 2 million for the period 2003-2006. With this award the Dutch government is giving a strong boost to innovation in the provision of academic information in the Netherlands. Another 4 M will be generated by the participating partners.

The joint approach is yielding significant benefits such as standardisation, national exchange and linking of files, the bringing together of scarce expertise, and cost efficiencies. As a result of the integrated approach, aspects such as long-term preservation and coordination with digital learning environments are also being taken into account.
With this joint initiative, an important step is being taken in the development of the national knowledge infrastructure. It is a good example of operational knowledge management in the academic field. The project is of great importance in the national and international knowledge economy.

DARE: aim

The aim of the DARE project is to modernize the facilities for Dutch scientific information by realizing an infrastructure and services to store, preserve, provide access to and distribute the scientific output of the Netherlands.
DARE should benefit the scholar on the one hand by the efficiency and speed of the system, which enables him to supply his material once only for the various forms of use and re-use. In addition the system enhances the visibility of his work and, as a result of this, contributes towards the building up of his academic reputation and career.
On the other hand DARE should:

- lead to improved access to scientific output
- leave responsibility for the required facilities with the universities
- avoid duplication of storage and of data management effort
- ensure interoperability
- use standards that are forward looking and aligned with international developments.

DARE: Approach

DARE provides a distributed network of 'institutional repositories'. An institutional repository is a facility, consisting of hardware, software, data and procedures, that

- contains digital objects representing all scientific output such as working papers/pre-prints, theses, research reports, data sets, conference contributions, multimedia presentations, etc from one or more universities, Research Organizations and Scientific Institutes
- insures adequate identification of the digital objects by means of metadata and a unique digital identifier
- provides facilities for management functions and archival of digital objects to a Repository Manager, representing the management of the university(s) and academic staff
- provides easy and standardized access to digital objects and metadata to approved scientists and Service Providers, thus enhancing visibility and interoperability
- provides adequate security for digital objects and metadata
- delivers digital objects and metadata to the National Library, for preservation in the national digital depot.

Appendix 1 shows a functional model of a repository.

The institutions themselves will look after the recording, management and in due course the communication of their own research results. Simultaneously, for the first time a clear and coherent insight into the (results of the) research efforts at the Dutch academic institutions is being created for all those concerned. This comprehensibility is of great importance not only for the development, planning and management of the research itself, but also for teaching.
Digital availability, based on open, international standards, like OAI-PMH (Open Archives Initiative-Protocol for Metadata Harvesting), Dublin Core and DOI (Digital Object Identifier) simplifies the further use of the information for various purposes: examples are publication in traditional or new journals (including electronic ones), long-term preservation in the e-Repository of the Royal Library and incorporation in digital learning environments for teaching.

In the second half of 2003, various services will be created for the repositories. A tender procedure just ended and several projects are awarded a grant for developing these services. These are seen as ' quick wins' in order to show the participants the benefits of joining DARE. The projects will be in the following areas:

- copyright management
- online publishing of conference proceedings
- connection to national system for research results
- digital review process

The institutions can make use of these services or develop themselves more services with added value in these areas, whether or not in collaboration with other parties such as academic associations and publishers.

Keep *e-material ready: the E-archive project*

In the case of institutional repositories scientific information items should be archived with the purpose of accessibility and reusability, in a setting of reuse for education and research goals. Keeping information items over time should be as reliable as possible, but the repositories are meant to be working archives, the scientists need the informational value far more than the evidential value.
Of course there should also be a long-term preservation.

Delft University of Technology Library has been working on the e-archive aspects of institutional repositories in their e-archive project, carried out in partnership with Universities of Utrecht and of Maastricht.

In the last couple of years their have been quite some discussions about the way materials can be preserved for future use: Delft University of Technology became involve in these discussions through the 'E-archive' project and has been looking for solutions.

In the short history of the development of electronic archiving it seemed that there were already two "camps": the camp in favour of emulation (an original bit stream is converted by a sequence of emulators) and the one in favour of conversion (after emulation the resulting bit stream is stored until a new emulation is necessary). On a conceptual level the results of both strategies are the same. With emulation, conversion is executed at the time of request ('on the fly') and with conversion the intermediate results are stored. In both cases the same conversion programs will be used. The difference boils down to a trade-off between computing power and storage capacity. If conversion has been performed neatly, with respect to all characteristics of the information item, the level of authenticity of an information item might be acceptable. But the same condition applies to emulation. So, for the aspect of authenticity the difference between emulation and conversion is so small that it can nearly be neglected.
The choice between emulation and conversion will be more dependent on costs and other contingency factors. One can calculate (on the one hand) the cost of complete (sequenced) emulation from the original bit stream and (on the other hand) storage of intermediate results after any emulation for future use. The cost of each and estimated future use of an information item may be decisive. Only if and when emulation becomes doubtful, as no hardware may longer be available, conversion becomes the last solution.
The E-Archive project liked to prove that this is not necessary and that there is not a big difference between conversion and emulation and there is certainly no need to make this choice beforehand.

E-Archive is developed on the basis of 3 principles:

1. data and metadata of the document are inseparable linked to each other, making use of a so-called XML container. XML is self descriptive, so the contents of the container can be deciphered as long as the characters are recognised

2. in e-Archive the original document/data always is saved as a bit stream

3. The viewer, being the program that gives any significance to the data is also saved in an XML-container. A viewer container consists of the source code, the description and a description of way to compile the program. It also is possible to save several representations of the document, e.g. the Word version, but also PDF or html. This way the user will be offered flexibility.

In the case of the storage of data and metadata including viewers and representations in the containers, the program picks up a bit stream in a container, looks for the right viewer container to go with it and converts the bit stream instantaneously to a readable representation; looking at it this way it can be seen as a form of emulation.
Any time the result of the viewer can be stored as a converted version in a container again and then it can be seen as conversion. See appendix 2.

How does it work?

The producer supplies data to the archive, and different kinds of metadata are added: descriptive metadata (title, author, year et cetera) and metadata for sustainability (who is the owner, check-sums, the quality and size of the original, criteria for selection). These form altogether the submission-information-package (sip).
The maintainer of the archive forms all these data into an archival-information-package (aip), a file that contains all data. This is the basis of the XML container.
All containers are stored in the digital archive and the archive system is indexed via a catalogue. The user looks up a document in the index and the aip-container with the data and metadata and the connected container with the viewer both are retrieved. The viewer starts up at the server of the archive and the result (also called dip =dissemination information package) is presented to the user in the format required by the user. The programs that can provide the proper representation (PDF, Word html) are installed at servers of the library. The user himself does not have to bother about these. See appendix 3.

[Reference 11] shows you a demo site for the E-archive project.

Findings and future steps

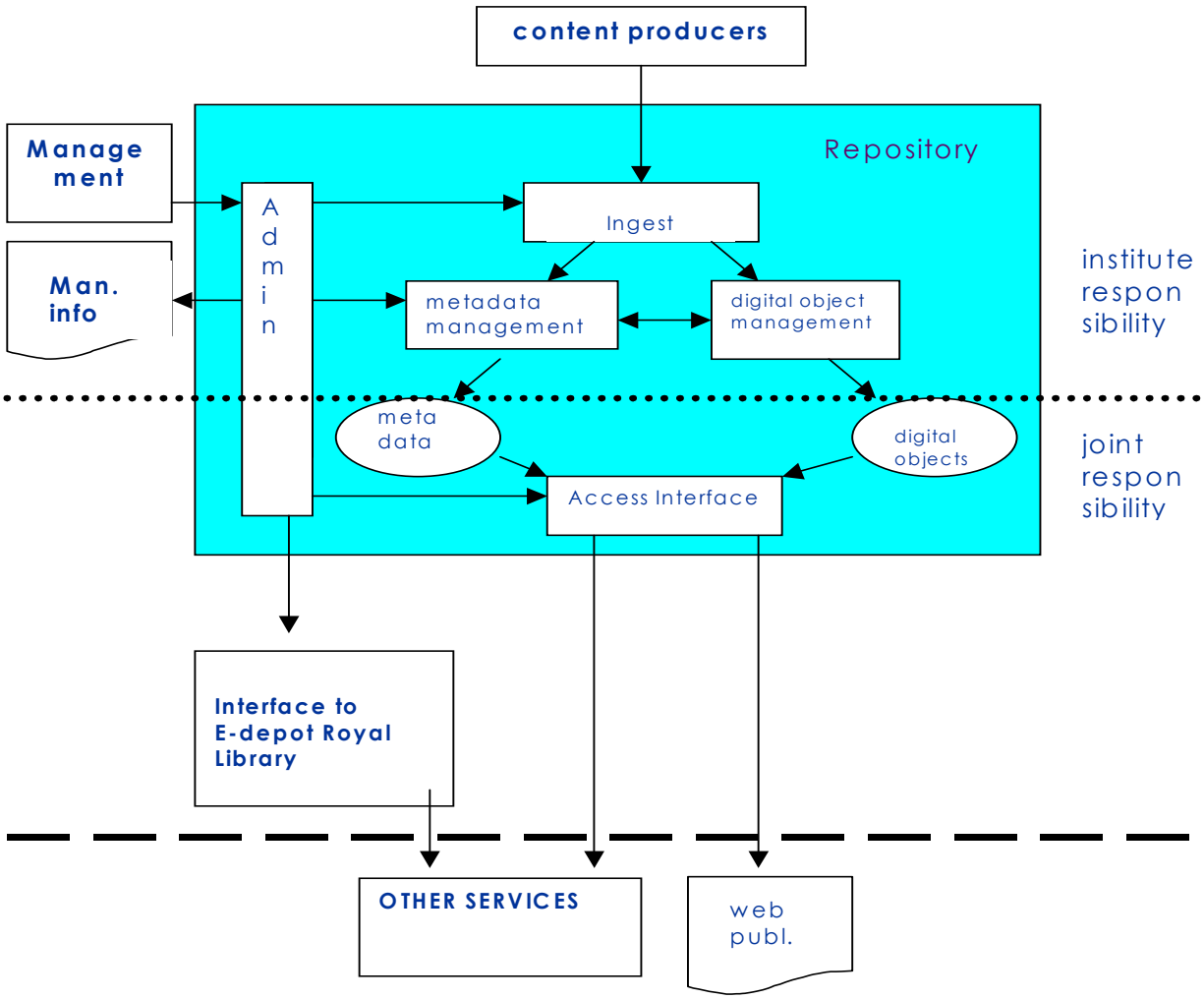
The project has supplied us with several conclusions, along 4 dimensions:

- the organisational dimension; i.e. (inter) national co-operation and standardisation
- the production dimension; implementing an organisational and technical infrastructure for the digital archive
- technology dimension; further research on preservation strategies and there implementation
- the business dimension; designing business models for the exploitation and economical viability of a digital archive.

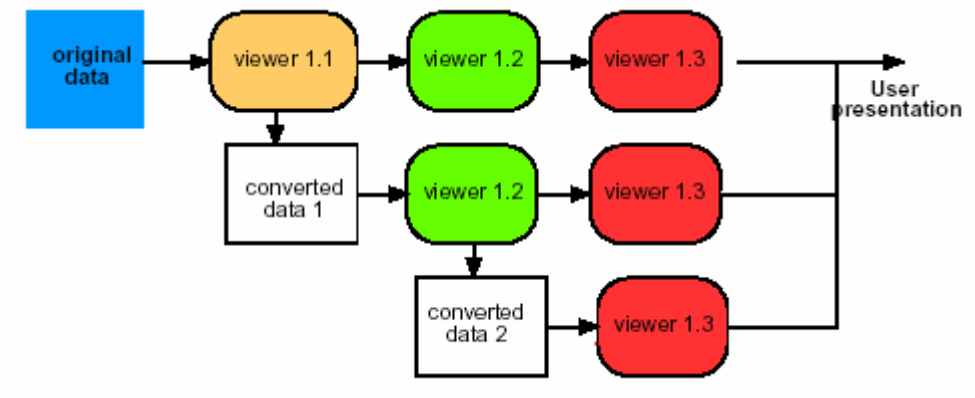
To work on all these dimensions and in line with the expectations within DARE the Royal Library (being the DARE partner with their e-depot for long term preservation and well known because Elsevier is making use of this depot) and Delft have been working towards a European project on long term preservation, where all the existing developments and techniques will be bundled and made operational. The project is called PATCH (Permanent Access Toolkit) and all European parties are involved that have contributed so far to the development of knowledge about permanent, long term access to digital materials.

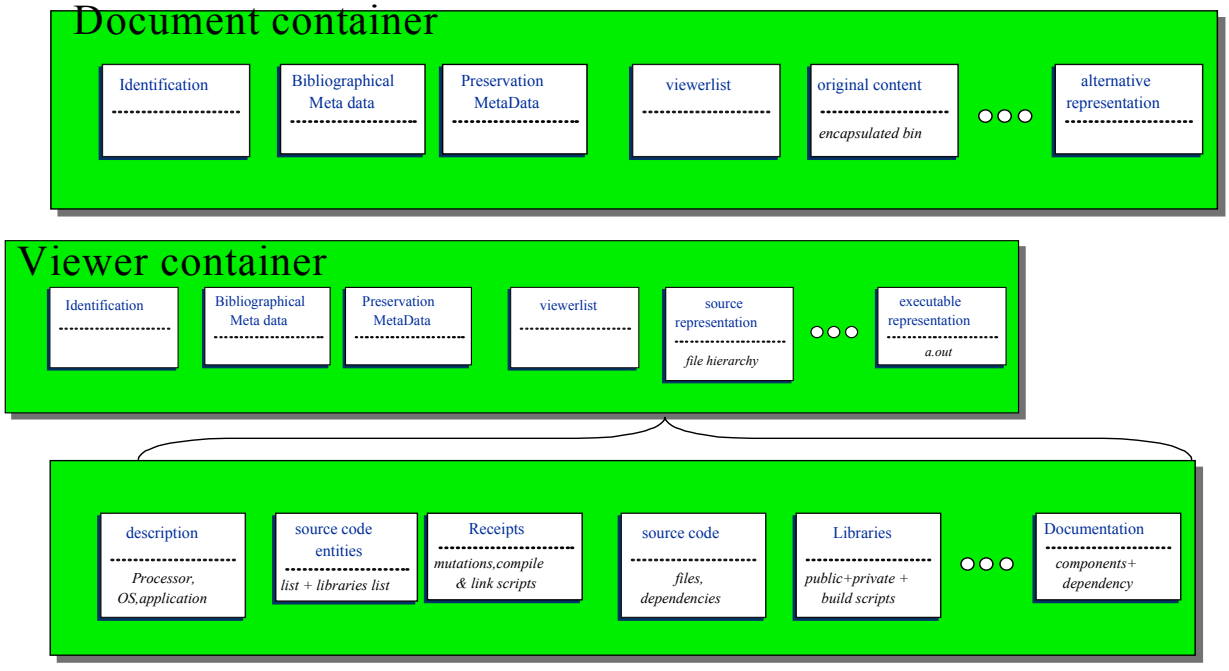
Appendix 1

Functional model of a repository

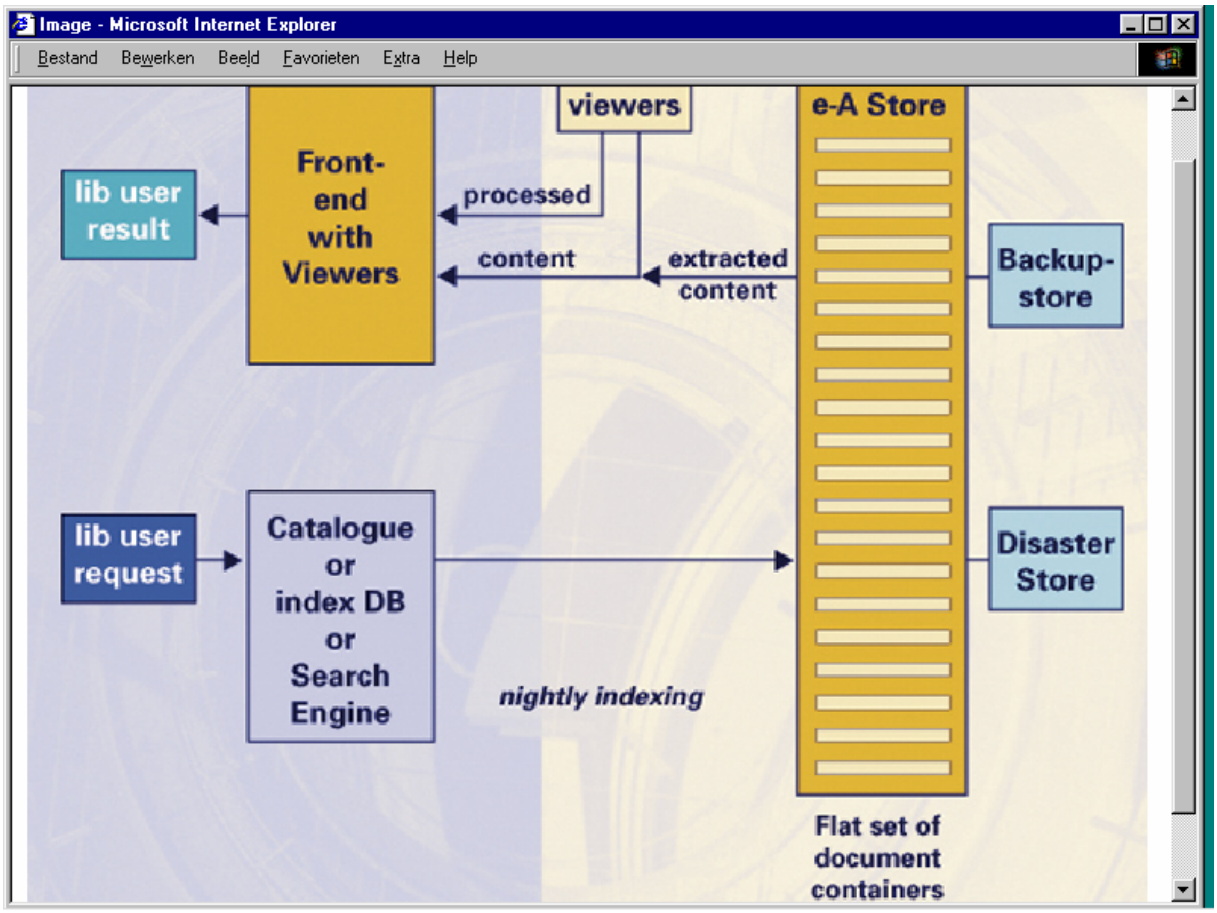


Appendix 2 Conversion/Emulation and XML Containers





Appendix 3 Delft Implementation of e-archive



Appendix 4 International Initiatives

There are several other initiatives in the international scientific community. The most notable initiatives are:

CCSDS (Consultative Committee for Space Data Systems – a committee founded by NASA) - has published a Reference Model for an Open Archival Information System (OAIS). See Reference [1]. This model contains Functional Specifications for the construction of a Repository, not limited to scientific output.

RLG (Research Library Group) has published recommendations for Attributes and Responsibilities with respect to digital repositories. The first recommendation stipulates compliance with the CCSDS - OAIS model. See reference [2].

SPARC (The Scholarly Publishing and Academic Research Coalition – *with the battle cry: Returning Science to the Scientists) is* an international conglomerate of scientific institutes and libraries (with backing of the ARL – Association of Research Libraries) with the intention to promote modern ways of disseminating scientific information and streamlining the publication process. They are strong proponents of digital repositories and have published some seminal papers on the purpose, structure and benefits of repositories, notably the SPARC Checklist and Resource Guide. See Reference [3]

DCMI (Dublin Core Metadata Initiative) is an international forum created to support the development of interoperable online metadata standards, with intention to converge to worldwide standards. The Dublin Core specification originally contained 15 elements and has subsequently been extended to 48 elements. See Reference [4].

OAI (Open Archives Initiative) is an organization, originated in the USA, set up to promote and develop interoperability standards for the efficient dissemination of information about scientific work. They have published a protocol (the OAI- PMH) for the harvesting of metadata, based on the Dublin Core format. The DC format is therefore part of the OAI standards.
See reference [5].

DNER (Distributed National Electronic Resource) is a (UK based and funded by the JISC – Joint Information Systems Committee) programme to create a managed network of cooperative repositories to be used by institutes of higher education. An important part is the development and adoption of community-wide open standards for description, access and use of scientific information. Standards that are currently adopted are: TCP/IP, HTTP, Z39.50, IP Authentication, and URI.

DOI (Digital Objects Identifier Foundation) has produced a proposal for the introduction of a standardized DOI, similar to the ISBN. Universities may issue DOI numbers, but international registration of these numbers costs money. See reference [6].

NEDLIB (Networked European Deposit Libraries) have promulgated recommendations for standardization and interoperability between national electronic depots. See reference [7].

DSpace. This is an operational software package, developed jointly by MIT and Hewlett Packard. See reference [8].
There are at present 4 official users; more are to be added soon; DSpace implements the OAI-PMH protocol for interoperability and includes DC metadata. The basis of DSpace is a PostgresSQL DBMS. It appears to be strong in both back-end and front-end functionalities. DSpace is available through BSF open source licensing

Eprints This is a (British) initiative promoting the creation of electronic on-line collections of academic research papers (in other words: repositories). Eprints supports the OAI-PMH and DC standards. Software has been produced based on the MySQL DBMS and has been made available as Open Source. See reference [9].

References

[1]. Consultative Committee for Space Data Systems (CCSDS): Reference Model for an Open Archival Information System, 1998.

[2] Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report RLG Mountain View, CA May 2002. www.rlg.org/longterm/repositories.pdf

[3] SPARC Institutional Repository Checklist and Resource Guide. http://www.arl.org/sparc/IR/IR_Guide.html

[4] http://www.dublincore.org/about/

[5] Open Archives Initiative (OAI): Protocol for Metadata Harvesting, V 2.0, 2002. http://www.openarchives.org.

[6] http://www.doi.org/

[7] http://www.kb.nl/coop/nedlib/index.html

[8] http://www.dspace.org/

[9] http://www.eprints.org/

[10] http://www.roquade.nl

[11] Demonstration site e-Archive project Delft: http://obelix.library.uu.nl:8030/IWIDEMO/index.html