OZ 05.3601

Conference report
Final

31 August 2005

Making the strategic case for Institutional Repositories
A conference report

On 10-11 May, the international conference *Making the strategic case for institutional repositories* took place at the Royal Netherlands Academy of Arts and Sciences in Amsterdam. Seventy-four leading experts from 14 countries were in attendance, spending two eventful days discussing all aspects of Institutional Repositories (IR). The conference was organised jointly by SURF (the Dutch higher education and research partnership organisation for network services and information and communications technology), the UK's Joint Information Systems Committee (JISC), and the US-based Coalition for Networked Information (CNI). Most presentations can be downloaded from http://www.surf.nl/cni-jisc-surf-conference .

The aim of the conference was to provide an overview of the state of the art, and to discuss long-term strategic options. Repositories are a new phenomenon, driven by a complex mixture of the need for better research management, and the scholarly communication, e-science and cyberinfrastructure agendas, as well as technology. The evolution of the traditional publishing paradigm into a more integrated electronic information environment has just begun. Open access has become a major policy debate within the research and higher education worlds, and repositories can offer infrastructure to support a variety of open access strategies. The conference provided participants with an opportunity to share experiences and discuss scenarios.

The core of the meeting consisted of four topical sessions dealing with technical and policy issues, as well as matters involving faculty (see below, session two and three) The main outcome was a clearer understanding of the many implicit and intricate issues with which IRs have to deal. It was clear that participants from many different countries have confidence that IRs will be a vehicle for improving access to scholarly output. On the other hand, it was realised that the promotion and development of IRs, and the way funding is made available, varies from country to country.

In preparation for the meeting, SURF's Gerard van Westrienen organised the first detailed survey of academic IRs in 13 countries, the so-called 'country updates'. Though great progress is being made, many technical and social obstacles came to the fore, many of which were discussed at the topical sessions. The country updates, and the analyses of their content, will be published separately (http://www.surf.nl/download/country-update2005.pdf). In order to sharpen the discussion, SURF commissioned a discussion paper by Joost Kircz, which elaborated on a wide range of themes relevant to IRs. A second version, integrating the comments received during and after the conference, remains available as reference paper (http://www.surf.nl/download/Discussionpaper-institutional-repositories.pdf)

The meeting opened with thematic panel discussions on the country updates mentioned above. Topics included: repositories' criteria for accepting material, metadata, national policies regarding IRs, centralised versus decentralised models, and organisational and workflow issues. The panel reports have been integrated into an article of Clifford Lynch and Gerard van Van Westrienen (soon to be published in D-Lib Magazine).

Keynote speakers, historian Mark Kornbluh and computer scientist Tony Hey, offered a wide-ranging perspective of science and education in an integrated, interactive, multimedia, electronic environment. Reflecting the purpose of national initiatives such as the UK e-Science programme, they showed how scientists can do better, faster or different research. Advanced projects in astrophysics (http://www.astrogrid.ac.uk), crystallography (http://ecrystals.chem.soton.ac.uk) and cultural heritage (http://valley.vcdh.virginia.edu/) illustrate the complex and challenging issues faced in designing systems to support emerging research practice.

CNI's executive director Clifford Lynch stressed in his address that the structuring of information is now the key issue in a world with more available data than humankind has ever had before. Software will do most of the pre-selecting, ordering and presenting. This development means that we have to consider writing differently and ensure we have specialised mark-up languages fit for every discipline and able to communicate with each other.

SURF director Wim Liebrand presented the Dutch Dare initiative (http://www.darenet.nl/en/page/language.view/home): the first comprehensive national network of research repositories, supported by 18 academic institutions in the Netherlands.

A special feature of the conference was a national celebration that marked the threshold IRs are now crossing. Royal Academy president Frits van Oostrom formally opened a special repository called *Keur der Wetenschap* (Cream of Science) containing the work of more than 200 prominent contemporary Dutch scientists (http://www.darenet.nl/en/page/language.view/keur.page). This repository has been created to show that complete bodies of work can be made available as an archive and source for further research. The enthusiasm of the participating scientists demonstrates a clear benefit available to those academic institutions that make their scientific outputs widely accessible. The Cream of Science model was discussed extensively by the conference participants and lauded as a significant achievement by the Netherlands.

The Topical Sessions

Session one, chaired by Don Waters of the Andrew W Mellon Foundation, dealt with the core issue of how to create new knowledge from existing data via the linking of repositories. The discussion began with a detailed overview from Liz Lyon, director of UKOLN (a UK national information management service), on the life cycle of data in scholarly communication. The huge influx of database-stored data of all kinds enables scientists to use and reuse, analyse, aggregate and compile it in a historically unprecedented way. Many examples of scientific collaboration between different groups in different locations and, most importantly, using different techniques and methodologies, show the potency of linked repositories. However, the fundamental traditional demands of quality assurance, open access, and integrity must now be cast in terms of workflow protocols to enable machine-to-machine (M2M) interactions, as humans are physically unable to handle the quantities of data involved. Lyon emphasised the symmetry in the knowledge cycles of research and educational activities (see also session four below). Finally, Lyon emphasised the great diversity of data collections, which highlights the importance of data structuring and standardisation.

The second speaker was Herbert van de Sompel (now at Los Alamos National Laboratory) who made the case for an underlying technology infrastructure for objects in repositories. Confronted with many different heterogeneous repositories, in the sense of techniques used as well as content stored, he presented a schema for object-level interoperability. An object can be seen as a unit of communication, which carries a unique resource identifier (URI) residing in a particular repository. Such an object could be merged with another object from another repository to help create a new unique information-object. Such a compound object must have its own persistent identifier, and will contain materials (including other compound objects) and metadata about those materials. In order to allow for repository interoperability we need XML-based representations. We also need an infrastructure that supports a federation of repositories. Van de Sompel showed that off-the-shelf software, based on open archive harvesting and Open URL tools, is insufficient to create standardisation at the repository level. A federated repositories infrastructure additionally requires repository and object registries. The issues of data authentication, integrity and usage rights have not yet been addressed. It is clear that these issues are pertinent for those IRs that are the primary store for scientific data, as the M2M tools must be informed of the status of an information object. A concern is that it is hard to find funding for the development of 'low-level plumbing', as Van de Sompel called this type of work. This kind of financial support is badly needed to enable large-scale e-science projects. Throughout the discussion, various participants suggested starting an international forum, as in the Internet community, to discuss and develop standards. The argument against this suggestion was that the technology is, in principle, already available. Concern was also expressed over who would organise such a forum, and who would follow it. Nevertheless, there is a compelling agenda for technical research and development.

The future will bring more diverse information objects, which may have multiple identifiers. This complexity demands clear choices on policies vis-à-vis versioning, authenticity, integrity, provenance, as well as a better understanding of the various relationships between information objects.

Session two, on the impact, citation and copyright issues in IRs, chaired by Chris Bailey from the University of Glasgow, dealt with the important issues of measuring usage, assessing quality and managing rights. Tim Brody from the University of Southampton and Johan Bollen from Los Alamos both presented research on ways to measure and assess the usage of documents in repositories. Usage and other metrics derived from log files can give clues to the structure and interrelationship of research programmes and the relative importance of works. Social and scientific networks can be identified, and research assessment enhanced, by using this information to supplement flat citation rates. Graphs and other representations of connectedness can play important roles, but valid values have to be assigned in them. It is important to develop sound and agreed metrics and, where multiple copies and versions of the same work exist, reliable and persistent identifiers are essential. In addition, impact analyses depend on reliable records for works, so tools have to be developed, and procedures and standards agreed, in order to make it easy for authors and others to create such records in IRs.

Charles Oppenheim of Loughborough University presented a detailed overview of the ways in which intellectual property rights are relevant to IRs. For many scholars, the differences among all the various rights are not always very clear, and some differ strongly by country. In the academic environment, the right to be identified as author, and guarantees that the work will not be altered or quoted out of context, are essential. On a more operational and commercial level, the distribution rights, the (European Union) database rights, and the rights related to the metadata (which are certainly not always added by the author), remain areas where various parties sometimes have conflicting interests. Oppenheim therefore recommends a focus on open access in general, rather than policies mandating deposit in specific named repositories (which may be considered restraints on trade). He also recommends that IRs work closely with Creative Commons (http://creativecommons.org) to develop consistent, academic-friendly, appropriate licences for the reuse of materials in IRs, and relevant groups should develop clear guidelines for IR copyright policies. Furthermore, IRs should closely monitor the diverse policies of commercial and not-for-profit society publishers, although, interestingly, a study by the Max Planck Institute shows that it is not the publishers' restrictions on copyright, but the researchers' inertia, that is the biggest challenge to open access initiatives (Science, 29 April 2005, p624).

Session three dealt with the crucial issue of faculty awareness and the building of an IR as a research tool. Jean-Claude Guédon from the University of Montreal chaired this session. Certainly this was one of the liveliest discussions as it dealt with the key to success. After all, technology and infrastructure are only enabling a behavioural change in scientific communications. Alma Swan of Key Perspectives Ltd reported on a new survey of 1,300 researchers covering 12 geographical areas and 15 scientific disciplines. The full 100-page report by Alma Swan and Sheridan Brown is now available (http://www.jisc.ac.uk/uploaded_documents/Open Access Self Archiving-an author study.pdf). It reveals a steady trend toward a greater awareness and usage of open archive initiatives. Computer scientists are prime users, which is understandable as they work in a largely electronic environment. It is certainly a 'bootstrapping' problem, as a critical mass of the scientific output has to be in repositories before the great majority of researchers will use them as a standard tool. Around half of the respondents self-archived on an average of two documents. This sounds low but is also a reflection of the learning curve: self-archiving is a new phenomenon in most disciplines. Apart from the awareness question, the advantages of an open archive have to be developed and promoted further. To that end, usage measurements, including citation counts and scores, are obvious and necessary additions to the existing systems (see session two, above). It must also be straightforward to archive a work, and easy-to-use templates for all necessary bibliographic and discipline-dependent metadata can facilitate deposit and improve retrieval. Certainly, as long as depositing a work in an IR is not compulsory, the academic systems that administer the scientific output of faculty (papers, conference contributions, books, lectures, etc.) and the repository must work together closely.

Susan Gibbons of the University of Rochester in New York presented a study on the use of IRs as research tools based on anthropological participant observation, *in situ* interviews and videotapes. A multi-disciplinary team investigated the work practice of 25 faculty members in a variety of fields. The findings reveal that, typically, stakeholders have their own priorities. The institutional perspective emphasises the efficiencies in research management, and the showcase effect that helps to establish the prestige of the university. From the library side, archiving, the permanence of the collection and the proactive response to high serial pricing are key issues. Obviously, faculty call for communication between colleagues, reading and citation, and well-presented work. In research, writing and publishing cannot be pulled apart, so tools have to be developed to integrate the process, especially with respect to the creation of metadata, where the old saying 'garbage in, garbage out' is relevant. There are too many places to look for research material, so pertinent metadata (as strongly defended by Van de Sompel in session one) is essential to target searches. Faculty are not interested in how it works (with the exception of computer and information scientists), but an IR has to fulfil at least those functions made possible by the present system. In that sense, all information has to enter the repository, including monographs.

The **fourth session**, which was chaired by Joan Lippincott of CNI, dealt with the central issue of policy and business cases. After all, the present day experiments have to evolve into full-fledged services. A better understanding of what constitutes an IR, and its position in the institutional or national context, are preconditions for developing a future federation of IRs. Typically now repositories not only vary as a function of the administrative unit, but also as a function of the discipline or type of content.

Ann Wolpert, Director of Libraries at MIT, strongly emphasised in her presentation that policy is central and needs to be focused on primary beneficiaries. An IR is an enabling service and without a primary constituency it is doomed to failure. A service must address real needs and benefits. In many cases, the traditional library roles, such as collecting and purchasing, may not apply, once a great variety of content is deposited into repositories. Repositories now hold preprints, data sets, educational and teaching materials, images and even simulations or games. All of these different information objects need their own approach. Wolpert stressed that the business cases are local and are works in progress. No grand design will solve the problems. For example, ownership issues vary from country to country and even by institution. Her main theme was that evolutionary developments need time. Habits change only if new experiences give better results. The greater the variety of content, from student portfolios, to educational material and to case data sets, the greater the need for dedicated policies towards these different types of content and the greater the need to develop overarching procedures and technology.

Lorna Campbell from the Centre for Educational Technology Interoperability Standards (CETIS) presented an overview of the intricacies of dealing with teaching and learning materials. These are information objects that are often specific to institutions and even to a given year and to particular individuals. Furthermore, practitioners have a different relationship with their material from scientists, who want to publish swiftly, so ownership, versioning and updating or deleting are difficult issues. Some works are used for local classes, while others are for distance-learning and independent study, where the interaction between teacher and student can be distinctly different. The important conclusions are that each work or element thereof should be clearly identified, that its life cycle is known, and that the requirements of the stakeholders throughout that life cycle are understood. This means that staff need training and experience but also that, as made clear in session three, the promotion of the use of the material must be based on well-defined and clear advantages for all stakeholders. As with other material types, when learning and teaching material from many different IRs merge, the issues of ownership, quality assurance and proper metadata have to be resolved.

In general, the different components of an IR, and the interrelationship between those components, need further scrutinising. After all, a comprehensive vision would see the IR holding not just traditional published research material, but the entire information output of academia.

Participants in this session suggested that some areas for follow-up collaborative action could be found in, for example:

- Surveys of the types of repositories being deployed, the services being offered, and how they can be federated
- Initiatives to document the characteristics and needs of the different stakeholders and the audiences (or constituencies) of the IR
- Making the case by gathering statistics of use and by demonstrating the benefits of IRs
- Further studies of the costs of IRs and the costs of not implementing them
- Further modelling of IRs tailored to the particular kind of information they hold
- The building of discovery services that are discipline-based and might be the outcome of joint efforts with search engine developers

In the **final plenary session**, the following concluding remarks were presented as guidelines for the future:

- We need a better understanding of the trends, as well as a clearer picture of where we stand now. Questionnaires such as the country updates need regular follow-ups. We need snapshots and surveys. It is clear that we will have local-, national- and (sub)discipline-based teaching and learning repositories with a great variety of kinds of information objects
- It is time to go beyond the first step in building repositories and look at the particular ecosystem of a discipline. Federated IRs, as a basis for e-science, demand clear policies on metadata, validation of the content and validation of the connections between IRs
- The low-level interoperability issue, and the tools to address it, are high on the agenda
- Given the growing awareness of the critical role higher education institutions play in disseminating knowledge, the issues of curation, data integrity and rights require development of processes and policies
- The Dutch showcase *Keur der Wetenschap* may not be exactly replicable in other countries given the different cultures; nevertheless it is an excellent showcase to point to

- It is of the greatest importance to promote IRs to faculty and ensure that IRs are richer than pre-prints or published articles only
- The development of a forum like the Internet Engineering Task Force, for the development of standards for federated IRs, is probably not necessary, as most tools are, in principle, available. However, standardisation is a big issue requiring continuous collaboration among developers
- The IR as a tool for science demands a wide range of experimentation, from the low level of object-interoperability to dedicated, discipline-dependent systems of metadata, and intuitive interfaces

In conclusion, the development of IRs is evolutionary, and it is clear that there is no one-size-fits-all approach. Different kinds of information objects have their own life cycle and workflows. We do not yet know how to handle all of the diverse information objects in constant mutual interaction. Re-use and re-contextualisation of content will be increasingly important and increasingly demanded by users. For example, digital images are no longer just illustrations in a text, but might develop their own grammar, for example, in comparative studies in clinical medicine or art history. In this evolutionary process we have to direct the course pragmatically, depending on the local situation. In some universities the emphasis will be on data because of their involvement in Grid experiments; in others there might be an emphasis on learning and teaching material because there is an extensive programme of distance learning. Every institution must decide for itself what the priorities are and to what extent it is a collaboration or federation with other initiatives. Nevertheless, to enable institutions to gain maximum value from information objects held in their repositories, work is needed in all of the areas described above (technical, interoperability, metadata, rights, versions, identifiers, and so on).

On a financial level, it is not easy to determine a library's value in relation to its maintenance and acquisition costs. It is therefore very difficult to speak yet of a business model for IRs, as more information is needed concerning the financial viability and sustainability of a repository. However, models for IRs are being developed. Building and maintenance costs can be shared by consortia and by the development of common software. The impact of open access on the overall cost of access to information should be taken into account. Exchange of information on usage, performance and costs will be useful in establishing common approaches.

On a technical level, already much can be done regarding low-level object interoperability. As every type of information demands its own type of metadata (potentially in the context of discipline-specific ontologies), it is clear that the Dublin Core standards are too simple and need to be supplemented in applications, with reference to other open standards such as LOM. Other open standards may have to be developed and agreed. The German DINI approach is a good example (see: http://www.dini.de/dini/dokumente/dokumente.php).

In general, one can say that, slowly, standards will be developed at all levels, such as the technical architecture, the descriptive languages, the interchange and web services software, the curation and digital preservation processes, the acceptance and quality control procedures, and digital rights management. In other words, serious work is under way and the conference highlighted all of the aspects we will face in the near future. The institutional repository has started its evolution into a central metabolic organ for academic knowledge.

The invaluable help received in preparing this report from the JISC staff, in particular Neil Jacobs, the CNI executive directors Joan K. Lippincott and Clifford Lynch and SURF's Gerard van Westrienen and Leo Waaijers are gratefully acknowledged.

Joost Kircz
KRA Publishing Research
Amsterdam, Holland
www.kra.nl