Development and Implementation of an Institutional Repository within a Science, Engineering and Technology (SET) environment

A. van der Merwe
University of Pretoria
Department of Informatics
School of Information
+27 12 841 3944
avdmerwe@csir.co.za

Prof. J.H. Kroeze
University of Pretoria
Department of Informatics
School of Information
+27 (0)12 420 3341
jan.kroeze@up.ac.za

ABSTRACT
Parallel to the Open Source Software movement, there is an increased demand and need for free, open access to information resources. The Open Access initiative is characterized by two strategies: namely the promotion of self-archiving or, alternatively, publishing of research articles in open-access journals. The purpose of an Institutional Repository (IR) is to provide a suitable archival environment for the self-archiving of digital items.

This paper provides an understanding of the complexity surrounding the implementation of an IR. Issues discussed include software selection, as well as the development, implementation and marketing of an IR. Attention is given to the development of the policies that are required by an organization and its main stakeholders. Issues such as acceptance, usage, population, and management of the repository are reported.

The work that was done at the CSIR is used as a case study and the subsequent lessons learnt are used to highlight some of problems experienced and how these problems were solved. Issues that still need investigation, e.g. long-term preservation, are mentioned.

Categories and Subject Descriptors
H.2.7 [**Information Systems**]: Database Administration – *Data warehouse and repository*

General Terms
Management, Human Factors, Standardization, Legal Aspects.

Keywords
Institutional repositories, Open access; Research documentation; Research publications; Full text access.

SAICSIT 2008, 6 - 8 October 2008, Wilderness Beach Hotel, Wilderness, South Africa

1. INTRODUCTION
One of the tangible outputs generated by a research organization is explicit knowledge (i.e. research publications data sets and source code). The pressure on researchers to produce explicit knowledge artefacts contributes towards the plethora of explicit knowledge available in today's environment.

Knowledge only becomes valuable when it is shared. Researchers internationally often express the need to monitor and share in the work done by their peers. This trend is evident in the emergence of the *h-index* developed by Hirsch [2005]. The value itself is determined by peer recognition. In order to meet these demands, a reliable Information Systems (IS) for the efficient and functional storage and retrieval has to be in place. In addition, increasing pressure is place on the scientific community to make relevant items readily available to the scientific community. The Open Source Software (OSS) and Open Access movements both contributed to the demand for free access to scientific publications.

The emergence of IRs is the direct result of an attempt to address some of these demands. This is not a straightforward task as the structure and function of repositories are subjected to individual organizational cultures. It is therefore essential to identify and define all the relevant issues relating to the development of a repository.

1.1 Defining the concept 'Institutional Repository'
Rankin [2005:iii] refers to IRs as a '… a set of services for storing and making available digital research materials created by an institution.' The notion of a 'set of services' is reinforced by Lynch [2003:2] rather than the concept of physical storage space normally associated with the term repository. In February 2008, the concepts of 'digital repository' [McHugh et al. 2007] and 'trusted digital repositories' [Harmsen 2008] were reinforced, thereby placing a greater emphasis on the digital features of repositories.

Crow [2002:4] extends the definition by referring to IRs as '…digital collections capturing and preserving the intellectual output of a single or multi-university community'. According to Crow it has become the responsibility of institutions to take back and retains the ownership and control of scholarly

communications and to reduce the monopoly currently exercised by publishers.

The concept of an IR implies an internet based service that provides free, complete and perpetual access. However, it should be made clear that national and international copyright laws have to be observed and respected at all times. Furthermore the peer review process also is important. IRs should therefore not be equated to plagiarism nor with inferior quality.

1.2 Benefits and value of an IR

Allard et al [2005:170] explains the value of IRs as services that provides members with '…the ability to add, or self-archive, items they have authored … thereby facilitating instant access to their work''. As confirmed by Anuradha [2005] the success of an IR is dependent on the collaboration and cooperation between the generators of the knowledge and the expertise obtained from librarians, archivists, record managers, policy makers and ICT staff. In view of the high costs associated with the implementation of an IR (ranging from R567 531.00 for developing countries to R12 347 990.00 for developed countries) (based on [Bailey et al. 2006], a clear benefit statement is required. In general, the benefits include preservation and dissemination of scholarly communication, increased and perceptual access to free or affordable information. However, the greatest value lies in the creation of awareness. Researchers benefit from having their work accessible by their peers, especially in terms of peer recognition. A clear and supported benefit statement contributes towards the sustainability of a repository.

1.3 Generic features of IRs

The features of an IR are influenced by the organization definition. However, as evident in the available literature [Devakos 2006; Anuradha 2005; Barton and Waters 2004] the challenge is to ensure that the right technology and the required policies are in place to warrant the effective long-term access and distribution of information in a digital format. The authors point out that ultimately IRs are used for the same purpose as commercial publications, namely for promoting scholarly communication by means of preservation and dissemination. Although current literature focuses on tertiary institutions, the same basic criteria are also applicable to any organization that generates research outputs of value that requires long-term preservation.

1.4 Stakeholders

Based on the work Jones et al [2006] the following represents the current interest of stakeholders in IR within the SET environment:

- Authors:- especially in terms of peer recognition, acknowledgement of research-integrity and long-term preservation and accessibility of research outputs.
- Organizations:- focus on the ability to use and reuse contributions in similar environments but also on the national and international recognition of research and research integrity as well as long-term preservation and accessibility of the research outputs. By using IRs, the intellectual capital of an organization becomes a valuable asset that is safely stored and which will remain accessible, thereby increasing in value as a source of shared knowledge.
- Users:- the ability of the wider scientific community to use relevant research material in new research, re-engineering of existing research while enjoying affordable and easy access, even from remote locations.

2. RESEARCH METHODOLOGY

The hypothesis is that a well-planned and structured IR using OSS will enable the management, curation and retrieval of explicit knowledge artefacts with a SET environment. Because of the nature of IRs and the impact of open access publications, use has been made of web-based searches. The scholarly federated search engine of Google (http://scholar.google.com) has been used extensively but not exclusively. Subscription databases such as ISI's Web of Knowledge were also used.

An analysis of the exiting proprietary database, identified shortcomings and expressed improvement requirements were also used for the development of the repository. As a decision was made to utilize an OSS product, existing limitations within the software determined the final structure of the repository.

Interaction and discussions with stakeholders formed a crucial element of the project. Extended use was made of emails to ensure that observations, interpretations and conclusions were correct.

3. DEVELOPMENT AND PLANNING OF AN IR

There is an underlying assumption that the needs and expectations of institutions differ and that different approaches are thus required. Experience has shown this to be true as a university must make provision from users at an undergraduate level up to users at a post-doctoral level, as well as lectures. Content types at universities include peer reviewed items, course notes, theses and dissertations and conference papers. A research organization must make provision for national and international research with only a secondary interest in the general public. The content types at a research organization will also include peer reviewed items and conference papers but will have the added responsibility of managing general research outputs, e.g. research reports. Difference in approach will affect the final structure, the contents that will be included, the compliance management approach and the willingness to support the IR. With the existing publication model, it is extremely difficult to obtain a holistic view of the individual institution's research. With the emergence of IRs, otherwise scattered information can be grouped within a single information system. This is made possible by technological developments in terms of digital publishing, networking, open-source standards and a significant drop in on-line storage costs [Lynch 2003].

3.1 Challenges facing the implementation of an IR

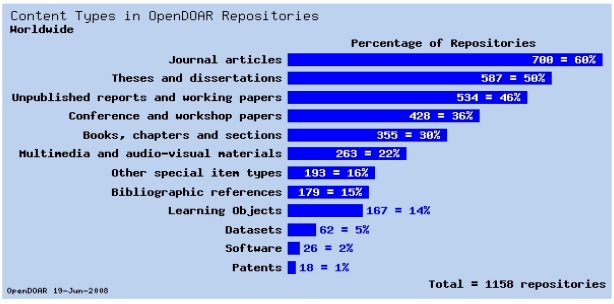
Typical barriers are the legal framework within which organizations function, existing ICT infrastructures, business models, recognition and awards, marketing, critical mass and the standard of indexing services. It is expected that an IR should provide a single access point and that it should also act as a self-evaluation tool as was the case with the PRABHAVIS system [Anuradha 2005]. Although not covered by this paper, it leads to

the issue of data mining and a theoretical comparison of existing databases vs. that which is required by the IR.

3.2 Core functions

During the development of the IR, some critical core functions should be planned and provided for. The following are loosely based on the work of Billings [2005]:

- Submission and editing of digital material by the author or other authorized individuals, including internal controls authorizing individual rights to add or edit items;
- Enhancement of metadata by implementing standards and guidelines by means of internal controls.
- Management of access rights as rights will change over time and detailed logs are required, including internal controls to prevent unauthorized changes.
- Registration of the IR with search engines and service providers such as Google, OAIster and DOAR demands that the IR manager keep abreast with developments in terms of suitable search engines and service providers.
- Long term preservation of records and upgrading of the digital formats with the assistance and support of the organization ICT group are critical as it is time-consuming, complex and labour-intensive.
- Defining the basic scholarly content that should be included. The general agreement is that scholarly content includes peer-reviewed journals, articles and conference papers, data sets, theses and dissertations, book chapters and research reports. Clear definitions of what are meant by the terms are essential prior to the implementation of the IR. The following figure illustrates the current distribution of content types contained in repositories worldwide [OpenDOAR 2007]



N.b. Most repositories hold several Content Types.

Figure 1: Content Types in OpenDOAR Repositories – Worldwide

Only after a clear understanding of the core functionality has been reached is it possible to select the most suitable application. The application must fit the need rather than the needs be changed to fit the product.

3.3 Proprietary vs OSS

The underlying assumption is that an IR should be developed by making use of an OSS application. The advantages of OSS vs. proprietary software are hotly debated. In the majority of cases, the difference between the functionality and the characteristics of the software is the greatest area of concern. With the move towards OSS by organizations such as the CSIR in South Africa, the challenge is to find a suitable OSS product. It must be accepted that some compromises will be required in terms of functionality. At the CSIR a comparison was made between an existing proprietary system and the envisaged OSS system. In the end it was necessary to sacrifice some of the functionalities of the proprietary system in order to stay within the national drive to move towards OSS systems.

The comparison revealed that functionalities such as unique fields, easily accessible log files and effective Boolean searching including truncation were some of the issues that would require additional development [Van der Merwe 2008]. A decision had to be made regarding the essentiality of required functions and whether an alternative was readily available should the function not exist in the OSS system.

The following is based on a comparison done by OSI [Open Society Institute 2004] and was used to compare two products, namely EPrints [EPrints 2007] and DSpace [DSpace Foundation 2007]:

- technical specifications including support for both hardware and software
- repository and system administration including user registration and authentication.
- content management especially in terms of metadata standards and search capabilities
- archiving with a focus on persistent identifiers and data preservation and curation support
- systems maintenance with a definite focus on a service level agreement specifying minimum requirements.

4. POLICY AND PLANNING ISSUES

4.1 Policies

Those directly affected generally regard policies as dictatorial, especially if they are not aware of the bigger picture. The ideal solution is to involve representatives from all areas during the finalization of the policy. Policies should include the provision of relevant system-based training programs and identify gatekeepers and role players. The policy will be heavily influenced by the philosophy of the organization and the ultimate purpose of the IR.

Essential elements that should be addressed in the policy includes the definition of content types, format, compliance and legal implications. Other issues will include: a) definition of whose work will be included and who will be allowed to submit items. b) The structure of the repository and future maintenance. c) will access be free or fee based or a hybrid system defined by the potential end users. d) Preservation, backups, down time and general SLA issues. e) Authorisation regarding the withdrawal of items.

4.2 Cost models

The policy should also include the cost model as this directly impacts on the sustainability of the IR. Both direct costs, i.e. system development, resources, scale and service maturity, and indirect costs such as office space and utilities training and

marketing and support, should be addressed. Should outsourcing be considered, additional allowance must be made for the monitoring and management of the service. The long-term expenses associated with digital preservation must be planned both in terms of direct costs and indirect costs.

5. Preservation issues

Goh et al. [2006] mention that preservation '… refers to the preservation of metadata and quality control measures to ensure integrity, and persistent documentation identification for migration purposes.' Problems and costs associated with changing formats and versions, technology becoming obsolete or outdated, software and hardware becoming incompatible with each other all impact on the preservation of the digital format. Smith [2002] rightly warns that digital formats do not survive or remain accessible by change. Although technical issues are regarded as the most pressing, priorities should also include conceptual requirements, standards, security and Intellectual Property Rights (IPR) protection, risk management and the testing of the proposed models. Experience with the migration from Corel to Microsoft and subsequently to OpenOffice has shown that any conversion should be approached with caution. Conversion is expensive in terms of real costs, person-hours and the potential corruption of data.

Although the emergence of the IR is addressing one aspect of preservation, it is threatened by a lack of preservation of the full text files linked to the repository. The INFORM methodology discussed by Stanescu [2005] provides six classes of risk associated with digital preservation. These are:

- Digital object format: Risks introduced by the format specification itself and by dependent specifications of compression algorithms, proprietary vs. open formats, digital rights management, encryption and digital signatures. Examples of risks include royalties or licence fees, incompatibility between different versions, lack of expertise of existing staff and complex or poorly documented specifications.
- Software: Risks introduced by all essential software components, e.g. operating systems, applications, library dependencies, archive implementations, migrations programs, implementations of compression algorithms and encryption and digital signatures. Examples include: unavailability of the source codes and incompatibility between versions
- Hardware: Risks introduced by necessary hardware components, including media type (such as CD, DVD, magnetic disk or tapes), CPU, I/O cards and peripherals. Examples are hardware interfaces that are very complex, large, ambiguous or poorly documented, as well as hardware interfaces that are not widely accepted and that might be unique in their class and therefore cannot be mapped to other systems.
- Associated organizations: Risks related to the organizations supporting the classes identified above to some extent, including beneficiary communities, content owners, vendors and open source communities. Examples are the following: High staff turnover and an associated lack of continuity, inability to obtain support from other organizations – also due to a lack of competitors, insufficient budgets and to because the user community might not be effectively involved in preservation planning.
- Digital archive: Risk introduced by the digital archive itself (i.e. architecture, processes and organizational factors). For example: a) each time a digital object is transferred there is likelihood for corruption of the data to occur; b) access security is weak, allowing unauthorised or accidental alternation or deletion; and c) off-site storage of hardware, media, software, etc. does not conform to existing policies.
- Format migration preservation plans: Risks introduced by the migration process itself, not covered in any other category. Examples given include the following: a) difficulty in proving authenticity after name changes has taken place; b) the conversion program effecting unauthorised changes to original contents; c) possible need for additional skill sets; and d) unpredictability of transformation costs.

6. Compliance and content recruitment

Obtaining compliance has proved to be a challenge on an international level. As evident in the work of Lynch [2003] and Mackie [2004] voluntary compliance was slow to emerge. Other authors [Jenkins et al. 2005; Mark and Shearer 2006] explains the lack of participation at the hand of a fear for compromising the relationship between authors and the publishers of peer-reviewed journals and a fear for not receiving the same recognition/accreditation as items published in accredited peer-reviewed journals. Concerns about infringing copyright laws and losing ownership of their IRR are also mentioned. Another, perhaps less valid reason, is the prevalent reluctance to trust a third party to take care of the long-term viability and sustainability of digital content and formats. Underlying all of these is the age-old question of 'what is in it for me?' which will only be addressed by means of visible incentives.

Foster and Gibbons [2005] provide advice in breaking down the resistance currently being experienced. The first step is to understand and respect the work practices of the contributors, e.g. knowledge workers, scientists and researchers. Secondly, it is essential to understand what the needs of the targeted contributors are. The third step is to enhance the IR so that it meets the needs of the potential stakeholders while accommodating existing work practices. Lastly it is essential that the stakeholders understand, on a personal level, the long term benefits of the IR.

An alternative to voluntary participation is the implementation of an organizational workflow 'impelling' the author to participate. By making use of a documentation workflow linked to the project workflow. Thereby an author is reminded of the existence of the repository and the potential benefits in participating and submitting publications to the IR. A well-designed workflow will contribute to alleviate concerns such as IPR, copyright implications and quality control..

6.1.1 Copyright and legal issues

The question of preservation and digital curation of research outputs requires understanding form all stakeholders. There are two issues involved, namely ownership and copyright. Achieving the delicate balance between acknowledgement and recognition on the one hand and ownership and copyright on the other hand, presents a challenge.

The SHERPA RoMEO (Publisher's copyright and archiving policies) project [SHERPA 2007] provides IR staff with an insight into existing policies, especially in terms of pre- and post print versions of articles published in specific journals. Currently there are four categories of publishers' archiving policies, indicated by a colour code.

Table 1. Archiving policies

RoMEO colour	Archiving policy
Green	Can archive pre-print and post-print
Blue	Can archive post-print (i.e. final draft post refereeing)
Yellows	Can archive pre-print (i.e. pre-refereeing)
White	Archiving not formally supported

In terms of items classified as yellow or white, negotiations with the publishers is required prior to inclusion in an archiving system such as an IR. At the moment negotiations are taking place on an *ad-hoc* basis with not clear guidelines. Best practices are still being developed and are influenced by the organization itself. Effective rights management is of the utmost importance and any deviation from this could potentially damage the reputation of the organization.

7. CASE STUDY

The case study used is the development of an IR at CSIR (Council for Scientific and Industrial Research), South Africa. The CSIR was constituted by parliament in 1945 and developed into a leading SET research and development organization on the African continent. Its mandate includes the fostering of research in the national interest and to contribute to the quality of life of the people of South Africa [Republic of South Africa 1988]. The CSIR is also required to adhere to the Access to Information Act [Republic of South Africa 2000]. In order to meet both these demands, it was decided that an IR is an appropriate tool to use, especially in terms of making relevant information readily available.

Known as Research Space, the CSIR's IR was launched on 1 August 2007. The IR was developed and is managed by the Information Services of the CSIR. It was also decided to deviate from the general academic institution approach and to make the repository a sub-set of an existing classified database. The IR therefore will contain all the publicly available publications and special collections identified as suitable for inclusion.

This decision resulted in a concerted effort by different departments to develop and populate the CSIR Research Space repository on a sound platform with quality data. The project team consisted of representatives from CSIR Information Services, the ICT services, Communications Group, R&D Outcomes and the EBAS (Enterprise Based Applications and Systems) groups. The project team mostly functioned in a virtual environment as conflicting schedules made it difficult for the team to meet on a regular basis.

The decision to select the DSpace platform as a suitable application was guided by the experiences of institutions such as the Universities of Pretoria and Glasgow. This decision was further supported by a study done by Jihyun. In his study, Jihyn [2005] proved three hypotheses namely:

- H1: users will spend less time completing the tasks in DSpace than in EPrints.
- H2: users will make fewer errors in DSpace than in EPrints.
- H3: Users' satisfaction with DSpace will be higher than with EPrints.

Although experience with the default search functionality of DSpace was disappointing it was still considered the best choice due to its compatibility with existing products and the infrastructure at the CSIR. [Open Society Institute 2004]

Hardware issues were resolved and a dedicated file server was purchased. Having recently implemented a repository at the university, the University of Pretoria's Academic Information Services provided support and constructive criticism,.

With the hardware and software issues resolved and a draft policy in place, it took approximately five months of dedicated team work to add some thousand full text items to the repository. These items consisted mainly of peer-reviewed publications published since 1999 for which copyright clearance was readily available. As the data was harvested from other sources, it was decided to capture the data manually, the logic being that a) quality needed to be monitored and verified; b) copyright issues had to be resolved and verified; and c) to ensure that the data is also reflected in the 'mother' database that the repository is a subset of. In the spirit of the IR principles, free and open access to the content of the IR is provided to all interested parties, nationally and internationally

The progress made was more rapid than initially anticipated. It was then decided to advance to the next phase and the CSIR's Annual Reports, CSIR's E-news and the journal CSIR ScienceScope were added. A collection of mining related reports was targeted next for inclusion as was a historical collection, the South African National Scientific Programme reports.

7.1 Structure and features of Research Space

The structure of the IR was based on the research structure of the CSIR. A community was created for each of the research units. The research areas within each unit are reflected by means of collections linked to the communities. As the CSIR is a dynamic organization, it was essential that the platform and structure should be able to accommodate rapid changes without the danger of data corruption or loss.

A previous ill-thought through decision proved to be fortuitous. During the testing phase of DSpace, hosted by the University of Pretoria repository UPSpace (https://www.up.ac.za/dspace/), all items were placed in a single collection under a single community. At a later stage it was decided to import the data directly for the UPSpace rather than repeating the work. As insufficient investigation was done prior to this phase, it was discovered that moving the information into specific communities or collection would be very time consuming. As a result it was decided to continue with a single 'black bag' approach and to use mapping to link communities/collections with publications.

The advantage of this 'black bag' approach is its direct relationship with the dynamic nature of the organizations. Units can be dissolved and new ones formed as the need for research in a specific area diminishes or as new research areas emerges. However, the intellectual property of the organization will be

accessible in logical context rather than through long lists of communities. Items will not be deleted from the 'back bag' but rather suppressed until such time that any restriction is lifted.

The black bag approach also prevents accidental deletion should a unit be dissolved or be merged with another. The danger of data loss or corruption is minimized when maintaining only one central storage area. New communities and collections can be added as needs are identified and existing records, if suitable, can be mapped to the new categories. An additional need for versatility is the fact that the organization often works across silos and it is not always possible to slot information neatly into an existing category. Again the 'black bag' approach easily accommodates linking to various communities and collections.

7.2 Document management workflow

Although DSpace provides a built-in workflow it does not sufficiently address the needs of the CSIR. The repository is and will remain a small part of a much wider document management system. The IR forms have therefore become part of the organizational document management workflow process. The workflow allows the author to indicate suitability of inclusion in the repository. Additional ownership is given to the authors as they can select the applicable collection from a drop-down list. They also have the option to suggest specific keywords. Depending on the content/publication type, additional approval for inclusion within the IR is required in order to ensure that IPR is not infringed in any manner. The author is also requested to approve the quality of the indexing done and can monitor the submission process, ensuring that data is captured and disseminated in a timely manner.

As concerns regarding copyright issues are paramount for an organization such as the CSIR, the workflow also serves as a tool whereby copyright approval is indicated. The actual approval notice is filed in a supporting document management system for future reference should any disputes arise.

The workflow also helps to ensure that the quality of the metadata is on an acceptable level. Accuracy in terms of the author's details is ensured as deviations in the format of the author's name are eliminated - the data is obtained via the workflow, which is linked to the HR system. The author is also empowered to approve the items linked to his name and is responsible for ensuring that all items were correctly included.

7.3 CSIR IR Policy

The policy is managed by the IR administrator but approval from an executive team is required should any changes be necessary. The policy can be summarized as follows: a) free and open access is provided but only to selected items; b) all formal externally published materials, with the proviso that the required copyright clearance was obtained, are included; c) peer-reviewed publications will get preference although all suitable publications will be included; d) the repository will always form a subset of the restricted technical outputs database; e) authors do not retain personal copyright; f) information will be managed and curated in accordance with existing standards and policies; and g) only items generated by CSIR personnel will be included.

A policy decision was also made not to include historical items (created prior to 1995) immediately but rather to follow an 'on demand' stance. Suitable items and collections will be identified as the need arises and only then will a decision be made whether to digitize and include items.

7.4 Visibility and statistics

It is clearly not always straightforward to implement and, as already mentioned, it can be very expensive. The question is whether the costs were justified. System generated statistics kept since the launch in August 2007 up to the end of the financial year in March 2008 proved to be very satisfactory and exceeded all expectations in terms of access via Robots/Spiders, Search Engines and Referring Sites. What is especially valuable is the number of non-CSIR sites that are linking directly to CSIR Research Space. This increase is beneficial in the ranking of the site by search engines such as Google. Another satisfactory

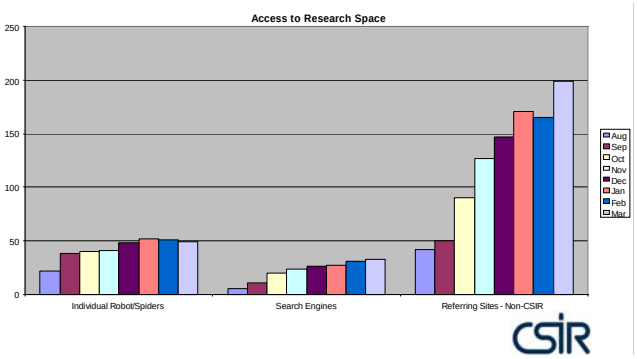


Figure 2. Access to Research Space Aug 2007- Mar 2008

statistic was the distribution of international visitors. Although the rate of monthly increases is slowing, the mere fact that there is still an increase is indicative that the repository is achieving its goal. Statistics for the first four months of the service indicated that visitors literally came from all over the globe as is indicated in Table 2.

Table 2. International access - Status Nov 2007

Geographical area	Hits
Africa	34768
Asia	4638
Australasia	2741
Europe and the UK	18665
Middle East	1103
North America	8427
South America	1075
Islands and networks	22272
Origin suppressed	85791

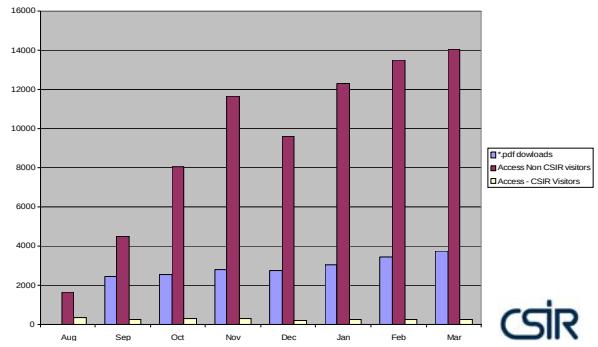


Figure 3. Research Space Usage Aug 2007- Mar 2008

Even more satifying is the steady increase in repeat and unique visitors to the repository since August 2007.

Although the statistics above are very relevant to the usage of the repository, it is no real indication of the research areas that are most topical in today's environment. Additional information is required. Once again, system generated statistics supply the analytical information that is required. Since the launch of the repository in August 2007, one item has been accessed 1755 times. It is anticipated that this type of information will assist the organization in identifying changes in research trends and focus areas.

Table 3. Top 10 items viewed

Item/Handle	Number of views
Math on MXit: using MXit as a medium for mathematics education (Butgereit, L) (10204/1614)	1,755
Expression of Rabies antibodies in tobacco and maize (Zungu, N et al) (10204/2001)	1,642
Crime and public transport: designing a safer journey (Kruger, T et al) (10204/1028)	1,592
Solar selective absorber functionality of carbon nanoparticles embedded in SiO2, NiO and ZnO matrices (Kalumba, G et al) (10204/1997)	1,429
Utilization of solar energy in South Africa (Whittier, A) (10204/999)	1,201
Model of the transverse modes of stable and unstable porro–prism resonators using symmetry considerations (Burger, L et al) (10204/1293)	1,119
Bessel–Gauss resonator with internal amplitude filter (Litvin, IA et al) (10204/2198)	1,069
An appetite suppressant from Hoodia species (Van Heerden, FR et al) (10204/765)	1,047
Acceptability of the integral solar water heater by householders in the low income urban community (Basson, FA et al) (10204/995)	959
Harnessing sorghum and millet biotechnology for food and health (O'Kennedy, MM et al) (10204/1040)	959

8. PROJECT SUMMARY

As part of the marketing and awareness process, a series of road shows was held to introduce the repository to CSIR personnel. These road shows proved to be valuable for monitoring existing perceptions. Several issues that kept on surfacing were identified and dealt with. The biggest concerns were linked to confidentiality, copyright and intellectual property rights. These could be addressed by assuring staff that repository personnel would not make decisions, regarding the inclusion of an item, in isolation. Authors who had already made use of the system were able to give valuable support and feedback, based on their perceptions and experiences of the repository, directly while in the company of their more concerned colleagues. An interesting comment made during one presentation: the speaker indicated that he was using the statistics generated as a marketing tool, thereby generating additional income for his current projects.

Lastly: the launch of the repository led to another unexpected event. The team leader's contact details were displayed on the repository's webpage. As a result, she started receiving calls from potential clients interested in additional research. As this reaction wasn't anticipated, the structure did not allow for the routing of the calls but it raised an interesting question about the value of the repository.

8.1 Lessons learnt from the CSIR's process

As mentioned in Section 5.1, the CSIR developed an experimental collection with UPSpace. During the transfer of the data from UPSpace to Research Space, some data was lost. As the log sheets of DSpace are not fully developed, it resulted in a manual verification process that proved to be very labour intensive. This way of doing is not routinely parte of the IR development process and it is not recommended that another institution follows a similar process.

Another problem experienced was that records that were suppressed did not appear in a readily accessible log. The implication is that one could forget about the existence of the item and therefore it necessitates a manual record which will have to be kept until such time that the logs are improved and more readily accessible.

Working within a virtual team proved to be both challenging and frustrating. The complexity of virtual teams and a lack of support within the teams did lead to some delays. The need for concise and clear communication became apparent as misunderstandings regarding deliverables and responsibilities caused further delays.

Delays prior to the actual development started, resulted in the loss of safety margins and resulted in a pragmatic acceptance of what could be done within the allotted time span. Although the deadline for the project was not changed, all non-crucial customization activities planned were shelved until further notice.

Some of the positive lessons include the positive reaction of the researchers and their eagerness to support the IR., The Executive, CSIR stakeholders such as the Board was highly positive in terms of what the project team was able to achieve in a very short period, as well as the positive statistics that is available. The project also presented a huge opportunity to develop new skills as the IR was an unfamiliar area for all team members. CSIR's research and collections are internationally acknowledged as valuable content. Lastly, the IR contributed towards Africa's contribution to the global knowledge pool.

Items on the CSIR's internet web site (http://www.csir.co.za), listed as 'Published research highlights', links directly to the IR, thereby providing quick and effective access to research publications. In addition, the link to the repository provided on the home page is an indication of the high regard that the organization is placing on the repository.

8.2 Recommendations

As a result of the project, several recommendations can be put forward. The first is to avoid any technical assumptions. For example, the project team assumed that the export of data from the experimental collection in UPSpace could easily be manipulated and moved to other communities and collections. The second is to accept that communication will always prove to be a challenge. As it was often difficult to get all the team members together on short-notice, extensive use had to be made of emails, which led to delays and miscommunications. The result

was conflict and stress that could have been avoided if effective communication systems were in place.

A third recommendation is that all branding issues should be resolved prior to the implementation of the systems. Delays in finalising the branding had the potential to delay the launch of the repository even though it did not affect the development work.

The fourth recommendation is that issues such as the structure of the repository be cleared with all the stakeholders prior to the development process but also to take note that a compromise will have to be taken at one stage or the other. It is not feasible to change the structure of the repository halfway through the development process.

The last recommendation is linked to proper project planning linked to human resources and capacity issues. The project owner needs to allow for staff turnover, potential expansion of the original scope of the project and for delays caused by the unavailability of critical personnel. SA also faces a unique situation where skills transfer is an essential part of any new project. Transferring skills often can lead to unnecessary frustrations for both parties – especially when the project starts running behind schedule.

8.3 Future research

There are two potential research areas, namely the use of IRs to determine the *h-index* and a way of accurately measuring the Return-on-Investment of the repository itself, thereby ensuring the sustainability of the repository. Calculating the *h-index* is a very complex and labour intensive exercise. It will be necessary for any system to source data from other databases and repositories, focussing on the citations within the document and then calculate the score as defined by Hirsch [2005].

Another problematic area is to accurately determine the Return-on-Investment for the organization. Pure financial ROI is not a suitable form of measurement and therefore some type of procedure should be developed to calculate the research requests and increase in funding that result primarily from the existence of the repository. At the moment, most of the information is open to interpretation. Research is required to identify what should be measured and how it can be measured. All biases should be removed from the calculations.

9. CONCLUSION

It has been argued that an IR provides an essential and valuable service in terms of the availability of information. The argument continues by implying that the benefits of a repository outweigh the time and costs invested in the development of the repository and that a repository is therefore a sustainable endeavour.

A concern highlighted in the literature is the issue of long-term preservation [Bullock 1999; Harmsen 2008; Stanescu 2005]. Preservation of digital formats is more complex than that of paper-based information, mainly due to the rapid advances in technology. Enabling effective long-term usage of digital formats there requires detailed planning and budgeting. Care should therefore be taken that digital formats do not end up in digital waste lands of inaccessible artefacts with a value that only future archaeologists might be interested in.

The literature [especially authors such as Barton and Waters 2004; Barton and Walker 2003; Crow 2002; Lynch 2003; Mackie 2004] provide valuable guidelines and information regarding the development of a repository. With the rapid development in ICT it is the responsibility of the IR administrator to implement only those changes that will be of value and improve the functionality of the repository. For the repository to be sustainable, it is essential that the stakeholders can place their trust in an efficient, well planned and well managed repository.

Although there is still a lack of documented proof regarding the sustainability and value of repositories, all indications are there that they will proof to be sustainable although perhaps not in the current format.

The time and effort in finalising the policies and obtaining the support from the stakeholders should not be underestimated. The role of individual organizational cultures can have either a positive or negative impact on the development process and therefore policies should guide rather than dictate.

During the development of CSIR's Research Space, it became clear that without the development of shared understanding, the project would not have been as successful as it was. Nor would it have been possible to complete the project in the short period available. Good teamwork, the ability to adapt as new information became available and the willingness to share experiences all contributed towards the ultimate success of the project. As a result of this, the departure of a key team member did not have a severely negatively impact on the project.

Finally, it became very clear that a service such as an IR could not – and should not – be implemented in isolation or as a 'library' project. Co-operation and collaboration within the organization play an essential, dynamic and ongoing role in ensuring the sustainability of the service. But just as important, collaboration between organizations proved to be of mutual benefit. The fact that the project team was able to share in the experiences of colleagues, greatly contributed towards completing the project in time. Lessons learnt during the individual development phases were freely shared as were the solutions to problems.

10. REFERENCES

[1] ALLARD, S. et al. 2005. The librarian's role in institutional repositories: a content analysis of the literature. *Reference Services Review* 33, 25 February 2007. DOI=http://www.emeraldinsight.com/0090-7324.htm.

[2] ANURADHA, K.T. 2005. Design and development of institutional repositories: A case study. *The International Information & Library Review* 37, 169-178. DOI=http://www.sciencedirect.com/science?_ob=ArticleListURL&_method=list&_ArticleListID=549863603&_sort=d&view=c&_acct=C000049363&_version=1&_urlVersion=0&_userid=958262&md5=8bb416c222bc740fba2012bbb899e030.

[3] BAILEY, C.W. et al. 2006. Institutional repositories. SPEC Kit 292

[4] BARTON, M.R. AND WATERS, M.M. 2004. Creating an institutional repository: LEARDIS Workbook. DOI=http://hdl.handle.net.1721.1/26698.

[5] BARTON, M.R. AND WALKER, J.H. 2003. Building a Business Plan for DSpace, MIT Libraries Digital Institutional Repository. *Journal of Digital Information* 4, 25 May 2007. DOI=http://dspace.mit.edu/handle/1721.1/26700.

[6] BILLINGS, M.S. 2005. Institutional repositories: sabbatical report January 30 - July 8, 2005. DOI=http://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1000&context=marilyn_billings.

[7] BULLOCK, A. 1999. Preservation of Digital Information: Issues and Current Status. 27 May 2007. DOI=http://epe.lac-bac.gc.ca/100/202/301/netnotes/netnotes-h/notes60.htm.

[8] CROW, R. 2002. The case for institutional repositories: a SPARC position paper. DOI=http://www.arl.org/sparc/bm~doc/ir_final_release_102.pdf.

[9] DEVAKOS, R. 2006. Towards user responsive institutional repositories: a case study. *Library Hi Tech* 24, 25 April 2007. DOI=http://www.emeraldinsight.com/0737-8831.htm

[10] DSPACE FOUNDATION. 2007. DSpace. DOI=http://www.dspace.org.

[11] EPRINTS. 2007. EPrints for digital repositories. DOI=http://www.eprints.org.

[12] FOSTER, N.F. AND GIBBONS, S. 2005. Understanding faculty to improve content requirement for institutional repositories. *D-Lib magazine* 11, 1 June 2006. DOI=http://www.dlib.org/dlib/january05/foster/01foster.html.

[13] GOH, D.H., CHUA, A., KHOO, D.A., KHOO, E.B. AND MAK, N., M.W. 2006. A checklist for evaluating open source digital library software. *Online Information Review* 30, 360-379. DOI=http://www.emeraldinsight.com/1468-4527.htm.

[14] HARMSEN, H. 2008. The final seal of approval: Directives for data producers/researchers, digital consumers and digital archives. In *African digital curation conference (1st : 2008 : Pretoria)*, Anonymous National Research Foundation, Pretoria. DOI=http://stardata.nrf.ac.za/nadicc/presentations/harmsen_henk.ppt.

[15] HIRSCH, J.E. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences* 102, 16569-16572. DOI=http://www.pnas.org/cgi/content/abstract/102/46/16569.

[16] JENKINS, B. et al. 2005. Content in, content out: the dual roles of the reference librarian in institutional repositories. *Reference Service Review* 33, 09 May 2007. DOI=http://www.emeraldinsigh.com/0090-7324.htm.

[17] JIHYUN, K. 2005. Finding documents in a digital institutional repository: DSpace and Eprints. *Proceedings of the American Society for Information Science and Technology* 42, DOI=http://dx.doi.org/10.1002/meet.1450420173.

[18] JONES, R. et al. 2006. Institutional repository. Chandos Publishing, Oxford.

[19] LYNCH, C.A. 2003. Institutional repositories: essential infrastructure for scholarships in the digital age. ARL Bimonthly Report 226, DOI=http://www.arl.org/resources/pubs/br/br226/br226ir.shtml.

[20] MACKIE, M. 2004. Filling institutional repositories: practical strategies from the DAEDALUS project. *Ariadne* 2 June 2006. DOI=http://www.ariadne.ac.uk/issue39/mackie/.

[21] MARK, T. AND SHEARER, K. 2006. Institutional repositories: a review of content recruitment strategies. In *World Library and Information Congress: 72nd general conference and council*, Seoul, Korea. DOI=http://www.ifla.org/IV/ifla72/papers/155-Mark_Shreaer-en.pdf.

[22] MCHUGH, A. et al. 2007. Digital repository audit method based on risk assessment (DRAMBORA); Version 1.0 (draft). <http://www.repositoryaudit.eu/download>.

[23] OPEN SOCIETY INSTITUTE. 2004. A guide to institutional repository software; 3rd edition. 2006. DOI=http://www.soros.org/openaccess/pdf/OSI_Guide_to_IR_Software_v3.pdf.

[24] OPENDOAR. 2007. Statistics. DOI=http://www.opendoar.org/find.php?format=charts.

[25] RANKIN, J. 2005. Institutional repositories for the research sector. National Library of New Zealand, Wellington.

[26] REPUBLIC OF SOUTH AFRICA. 2000. Promotion of Access to Information Act. No 2 of 2000, DOI=http://www.info.gov.za/gazette/acts/2000/a2-00.pdf.

[27] REPUBLIC OF SOUTH AFRICA. 1988. Scientific Research Council Act. No 46 of 1988; as amended by Scientific Research Amendment Act, no 71 of 1990, DOI=http://www.info.gov.za/docs/legilation_compliance/scientific_research_act.doc.

[28] SHERPA. 2007. SHERPA RoMEO project. 2007, DOI=http://www.sherpa.ac.uk.

[29] SMITH, B. 2002. Preserving tomorrow's memory: preserving digital content for future generations. *Information Services and Use* 22, 133-139. DOI=http://www.ebsco.com.

[30] STANESCU, A. 2005. Assessing the durability of formats in a digital preservation environment; the INFORM methodology. *International Digital Library Perspectives* 21, 61-81. DOI=http://www.emeraldinsight.com/1065-075x.htm.

[31] VAN DER MERWE, A. 2008. Development and implementation of an institutional repository within a science, engineering and technology environment. University of Pretoria.